OIKOS

Research article

Probing variation in reaction norms in wild populations: the importance of reliable environmental proxies

Jip J. C. Ramakers[1], **Thomas E. Reed**[2,3], **Michael P. Harris**[3] **and Phillip Gienapp**[4,5]

[1]Biometris, Wageningen University and Research, Wageningen, the Netherlands
[2]School of Biological, Earth and Environmental Sciences, University College Cork, Distillery Fields, Cork, Ireland
[3]Environmental Research Institute, University College Cork, Cork, Ireland
[4]Netherlands Institute of Ecology (NIOO-KNAW), Department of Animal Ecology, Wageningen, the Netherlands
[5]Michael-Otto-Institut im NABU, Bergenhusen, Germany

Correspondence: Jip J. C. Ramakers (jip.ramakers@gmail.com)

Oikos
2023: e09592
doi: 10.1111/oik.09592
Subject Editor: Denis Réale
Editor-in-Chief: Pedro Peres-Neto
Accepted 16 August 2023

Many traits are phenotypically plastic, i.e. the same genotype expresses different phenotypes depending on the environment. Genotypes and individuals can vary in their response to the environment and this genetic (G × E) and individual (I × E) variation in reaction-norm slopes can have important ecological or evolutionary consequences. Studies on I × E/G × E often fail to show slope variation, potentially due to the choice of the environmental covariate. Identifying the genuine environmental driver of phenotypic plasticity (the cue) is practically impossible and hence only proxies can be used. If the proxy is too weakly correlated with the cue, this may lead researchers to conclude there is little or no (variation in) plasticity, and hence lead to downwardly biased estimates of the potential for plastic responses (or evolutionary change in the slope) in response to environmental change. Alternatively, the environment-specific mean phenotype (ESM) across individuals – which captures all environmental effects on the phenotype – as covariate should be less prone to such bias. We showed by simulation – after verifying the concept analytically – that using weakly correlated proxies indeed biased estimates of slope variation vis-à-vis the true cue downward but that ESM as a covariate held up well, even when multiple sources of I × E or an interaction between environments (I × E × E) existed in the data. Analysis of two real datasets revealed that estimated I × E and G × E, respectively, were more sizeable and precise when using ESM as opposed to reasonably informative environmental proxies. We argue that the ESM approach should be adopted by biologists as a yardstick in the study of (variation in) plasticity in the wild and that it may serve as a useful starting point for the search of better environmental proxies and unravelling complex I × E or G × E patterns.

Keywords: covariate, ecology, environment, Finlay–Wilkinson, G × E, I × E, phenotypic plasticity, random regression

NORDIC SOCIETY OIKOS

www.oikosjournal.org

Introduction

Many traits are phenotypically plastic (Pigliucci 2005), i.e. individuals or genotypes express different phenotypes in different environments. Phenotypic plasticity is generally described by a "reaction norm"' (Woltereck 1909), which is defined by its intercept or elevation, commonly defined as the phenotype in the average environment, and its slope, i.e. the "sensitivity" to the environment, in the case of a simple linear reaction norm. Genotypes and individuals can differ in their reaction-norm slopes, i.e. populations can exhibit genotype-by-environment (G × E) and individual-by-environment interactions (I × E) (Nussey et al. 2007). Phenotypic plasticity enables populations to cope with changing environments (Yeh and Price 2004) but in the long term such plastic responses may be unlikely to be sufficient and the reaction norms themselves (their elevation) will have to evolve (Gienapp et al. 2014, Ramakers et al. 2018a).

Many studies have explored I × E and/or G × E in behaviour, phenology or physiology (Reed et al. 2009, Hau and Goymann 2015, Stedman et al. 2017), finding statistically significant variation in reaction-norm slopes in some (Nussey et al. 2005, Wilson et al. 2006, Brommer et al. 2008, Reed et al. 2009, Stedman et al. 2017) but not all studies (Reed et al. 2006, Charmantier et al. 2008, Ramakers et al. 2018a, Froy et al. 2019). Negative results always raise the question of whether it is a genuine absence of the sought effect or whether the study did not have enough statistical power. Simulation studies showed that the power to detect individual variation in slopes depends on both the number of observations per individual and the number of individuals sampled (Martin et al. 2011, van de Pol 2012), or on the presence of heteroscedasticity (Ramakers et al. 2019a), but the choice of the environmental variable in the analysis could also play a role. The majority of studies on I × E and G × E use so-called random-regression models (Henderson 1982, Kirkpatrick 1989, Morrissey and Liefting 2016). In random regression, the expected value of the trait and the (co)variance in slopes and elevations are modelled as a continuous (linear or non-linear) function of the environment. Individual elevations and slopes are modelled as random effects, i.e. assumed random draws from a normal distribution with mean zero and (co)variances to be estimated. This approach means that a continuous environmental variable is needed as covariate against which the phenotypes are regressed. Using a covariate that correlates weakly with the true cue can lead to incorrect conclusions regarding the presence or absence (or magnitude) of I × E and G × E with respect to the true cue, i.e. the actual environmental variable that individual organisms pay attention to in order to anticipate abiotic or biotic fluctuations (see Fig. 1 for a conceptual illustration of this principle). This relates to a general statistical phenomenon termed "attenuation" (Spearman 1904). Moreover, predictions of phenotypic and evolutionary change in response to directional environmental change will be biased downwards if based on the 'wrong' environmental variable in a reaction-norm analysis. It is not 'wrong' per se to study plasticity with respect to a proxy environmental variable that is imperfectly correlated with the true cue; estimates of reaction-norm parameters with respect to that proxy variable will not be biased. But such an analysis would fail to reveal the true potential for environment-driven phenotypic variation and scope for plastic responses to environmental change. In the case of G × E, an analysis based on a poor proxy would also underestimate the evolutionary potential of plasticity itself in the face of directional environmental change.

In observational studies of natural populations, many potential candidate covariates could be used in the analysis of phenotypic plasticity, and biological knowledge of the system would help in choosing the best. For example, phenological traits such as flowering, migration or breeding time generally depend on ambient temperatures (Sparks and Carey 1995, Visser et al. 2009) and hence local temperatures would be an obvious choice as covariate. More generally across traits, however, there are many different ways to quantify what constitutes 'the environment'. To complicate matters further, traits may depend on more than one variable. For example, it has been shown that breeding time is affected by spring temperature and population density in tree swallows *Tachycineta bicolor* (Bourret et al. 2015) or spring temperature in interaction with day length in great tits *Parus major* (Gienapp et al. 2005). Body size, growth rate and fecundity in seed beetles *Callosobruchus maculatus* depend on both ambient temperature and the type and size of rearing resource (see Stillwell et al. 2007, Westneat et al. 2019, Rodrigues and Beldade 2020 for more examples of environmental interactions). Consequently, it will very rarely be possible to identify the real driver of plasticity, simply because it is unknown, unmeasurable or a composite of several variables. Instead, a proxy that correlates with the real causal driver(s) of plasticity has to be used (Buoro et al. 2012) and the amount of I × E/G × E revealed will depend on the choice of this proxy. For example, in a study of I × E in breeding time in collared flycatchers *Ficedula hypoleuca*, only two out of three main environmental drivers (temperature, rainfall and NAO) revealed significant I × E (Brommer et al. 2005). Similarly, detection of statistically significant I × E in breeding time in the Wytham Wood great tit *P. major* population differed between studies depending on how the environmental proxy (spring temperature) was summarized (Charmantier et al. 2008, Husby et al. 2010). In such analyses it will be virtually impossible to ascertain that the chosen proxy is correlated closely enough with the true driver of plasticity to allow the detection of I × E.

At this point we need to stress that the estimation of variation in slopes (I × E and/or G × E) is biologically relevant even if identifying the true environmental driver of plasticity is difficult. For example, the presence of slope variation may explain why populations are not responding evolutionarily to selection; when selection acts in the same direction across environments, low consistency of individual phenotypes due to crossing reaction norms – regardless of the underlying environmental driver – means that there may be no net selection on the elevation (Turelli and Barton 2004, Kokko and Heubel 2008). Related to this, when investigating the

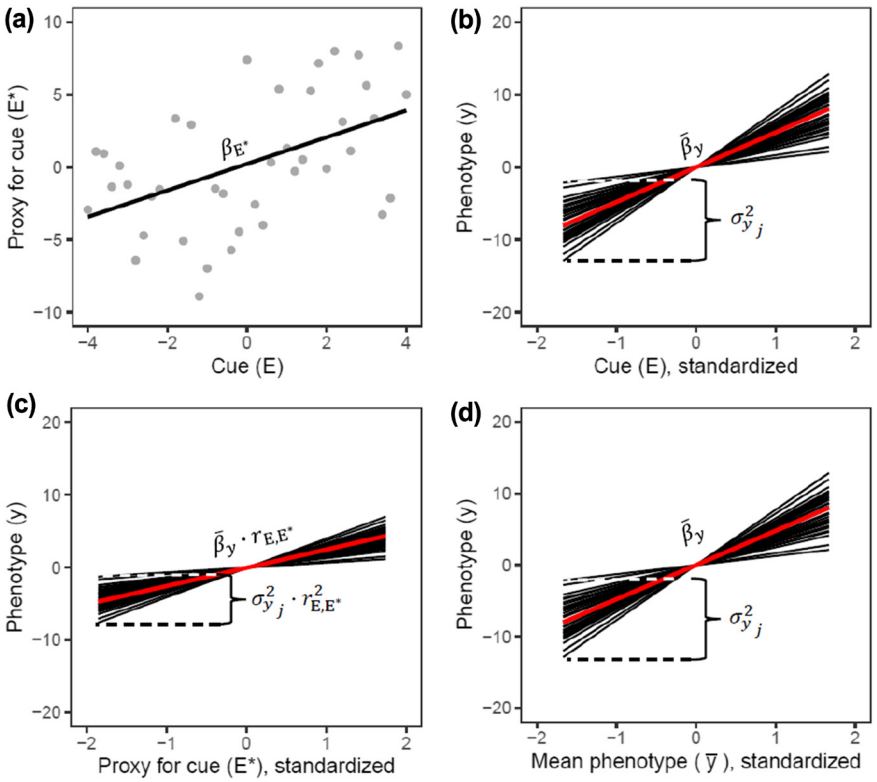

Figure 1. Conceptual illustration of the relationship between the environmental cue E, its proxy E^* (i.e. something we measure about the environment), and the phenotype *y*, assuming environmental stationarity (i.e. distributions of each environmental variable, and relationship between them, remain constant over time). We further assume an imperfect (simulated) linear correlation between E and E^* with slope $\beta_{E^*} = 1$ and correlation coefficient $r_{E,E^*} = 0.5$ (a) and, for simplicity, no residual variation (stochasticity) in individual phenotype over-and-above the effect of the true cue. Keeping intercepts at zero, let the mean linear reaction norm of the population with respect to the cue be drawn by $\bar{y}_j = 2E_j$ and that of 30 individuals $y_{ij} = (2 + \delta_i)E_j$, where $\bar{y}_j$ is the mean phenotype in environment *j* and δ_i is some random individual deviation from the mean. If we standardize the cue (E′), the equation of the mean reaction norm becomes $\bar{y}_j = \bar{\beta}_y E'_j$, where in this case $\bar{\beta}_y = 4.8$ and where we have environment-specific, between-individual variance $\sigma^2_{y_j}$ (b). Regressing the same reaction norms onto the standardized proxy variable $E^{*\prime}$ (through least-squares estimation) gives us $\bar{y}_j = (\bar{\beta}_y \times r_{E,E^*})E^{*\prime}_j$ for the mean reaction norm (i.e. in this case the slope is halved to 2.4) and the between-individual variance is $\sigma^2_{y_j} \times r^2_{E,E^*}$ (i.e. in this case quartered) (c). Replacing the proxy on the x-axis by the (standardized) mean phenotype in each environment ($\bar{y}'_j$ = ESM) retrieves the original response to the real cue, $\bar{y}_j = \bar{\beta}_y \bar{y}'_j$ (d), where $\bar{y}'_j = E'_j$ and $\sigma^2_{y_j}$ matches that observed in (b), given that, by definition, the unstandardized mean is expected to correlate perfectly with itself, with $\bar{y}_j = \bar{y}_j$. In (b)–(d), red lines indicate the mean reaction norm, whereas black lines indicate individual reaction norms; in (c), lines are least-squares regression lines.

evolvability in a population we are typically interested in quantifying the phenotypic and genotypic variation in the trait per environment (the scale at which selection takes place). Due to sample-size constraints precluding analysis per environment, this is most easily obtained by using a regression-type approach across environments, for which the identification of an environmental covariate is required (discussion in Ramakers et al. 2018b). Regardless of the environmental proxy used, it is important to realize that the inability to find statistically significant slope variation with respect to that proxy does not mean that I × E or G × E is absent in the population. To probe this 'hidden' I × E or G × E variation, we need an unbiased proxy that captures all elements constituting the environment driving the phenotype.

When we do not have access to a measurable, accurate environmental proxy, we may instead regress the phenotype on the environment-specific mean (ESM), i.e. the mean of the phenotypes of all individuals/genotypes in an environment (where the environment is a year if the relevant phenotypic variation is among years in a given location, a habitat type if the study system comprises several habitat types, or even 'year within a location or habitat type' – whichever suits the study system best). This approach, based on the 'Finlay–Wilkinson' (FW) regression (Yates and Cochran 1938, Finlay

and Wilkinson 1963), has been widely applied in animal and plant breeding to test whether the (relative) merit of cultivars/genotypes is constant across environments (Lynch and Walsh 1998, James 2009, Malosetti et al. 2013). Since ESM incorporates all plastic effects of environmental drivers, it would be correlated closely with the true – but unknown – driver of plasticity (Fig. 1). A prerequisite is that the reaction norm not be inherently strongly non-linear, as this may lead to similar mean values in different environments, and that the mean be based on an adequate and representative (random) sample of phenotypes. We have found that, although the ESM approach is nothing new, its use in analysis of plasticity – and by extension in probing the degree of I × E/G × E – in wild populations remains underappreciated among ecologists (discussions in Ramakers et al. 2018b, 2019b, Brommer 2019). We will argue in this paper that ESM should be tested alongside other environmental proxies whenever one is interested in uncovering variation in plasticity. Furthermore, beyond uncovering I × E or G × E per se, ESM allows for the fitting of individual reaction norms more accurately than weakly informative proxies can, and these estimated reaction norms can act as a yardstick for how accurate phenotypic predictions in future environments can be when using environmental proxies.

Here we explore how the correlation of a proxy environmental variable with the real driver of plasticity can bias estimates of mean plasticity and individual variation therein and how the ESM approach compares to using these various proxies. We do this to drive the point home that unravelling (variation in) plasticity – and by extension prediction of phenotypes under changing environmental conditions – requires careful consideration of which environmental covariate to use, even if this proxy has been established by convention. Using random regression, we regress simulated phenotypes against ESMs and against nine different environmental variables that correlate with the real driver to varying degrees but do not causally affect the phenotypes themselves (i.e. proxies). In the simulated data, individual variation in slopes (I × E) with respect to the true cue is present and the key point of interest is to determine with which proxy it can be detected accurately. Furthermore, we benchmark our environmental proxies by 1) introducing different sources of variation in the data (age structure, habitat structure, and a time trend in the phenotype) and 2) by adding complexity in the I × E structure through a second environmental driver ($I \times E_1 + I \times E_2$), an I × E × E effect, and an effect of sampling individuals within a limited range of environments. As practical, real-life examples of the concept we reanalyse I × E in phenology in a population of common guillemots *Uria aalge* (Reed et al. 2006), and reinterpret data on a previous study on fledgling weight in great tits *P. major* (Mulder et al. 2016a, b) to study G × E. We will argue that ESMs are useful for probing the extent of I × E (or potentially G × E) in the data and can serve as a starting point for the estimation of environment-specific (genetic) variances and accurate reaction norms, as well as the search for informative environmental proxies.

Material and methods

Here, we outline several simulated scenarios and analyse two real-life datasets to show that ESMs can be used to probe the extent of I × E (or G × E) in a population. For proof of how the ESM approach may outperform other environmental proxies (besides conceptual Fig. 1), we provide the mathematical derivation in Appendix A. Simulation and data analysis was done in R ver. 3.5.3 (www.r-project.org) using the 'lme4' package (Bates et al. 2018) and ASReml-R ver. 4 (Butler et al. 2017). The syntax is provided as Supporting information.

Simulation scenario (i): single environmental driver of I × E

Individual phenotypes were simulated with an 'individual-based model'. Individual reaction-norm elevations and slopes were drawn from normal distributions with means of zero and variances of 5 and 0.5, respectively, with a small covariance of 0.1 between them. The true driver of plasticity, environmental variable E1, was simulated for 20 years (i.e. one value per year) by drawing values from a normal distribution with a mean of zero and a variance of one. Nine other environmental variables (E2–E10) that correlated with E1 by 0.9–0.1 were simulated using the following equation:

$$\mathbf{E}_n = r_n \mathbf{E1} + \sqrt{1 - r_n^2}\,\boldsymbol{\mu} \quad (1)$$

with $\mathbf{E}_n$ being a vector containing the nth environmental variable, r_n its correlation with **E1**, and $\boldsymbol{\varepsilon}$ a vector of random errors drawn from a normal distribution with mean 0 and variance 1. Realised r_n will vary across simulations, but the average across simulations will reflect the input value.

Phenotypes of 500 individuals were simulated by first determining a first year that was drawn randomly from the years but ensuring that the four observations of each individual would fit within its 'lifetime'. Phenotypes were simulated as follows:

$$y_{ij} = \mu + a_i + \left(b_{\text{pop}} + b_i\right)\text{E1}_j + \varepsilon_{ij} \quad (2)$$

with y_{ij} being the phenotype of individual i in year j, μ the population-level elevation (intercept), a_i the individual elevation deviation ($a_i \sim N(0,5)$), where the latter value is expressed as variance throughout, b_{pop} the population-level slope ($b_{\text{pop}} = 0.5$), b_i the individual slope deviation ($b_i \sim N(0,0.5)$), E1_j the value of environmental variable E1 (i.e. the true causal driver of plasticity) in year j, and ε_{ij} a random error ($\varepsilon_{ij} \sim N(0,5)$). With residual variance being equal to variance in elevation, this means a repeatability of $r^2 = 0.5$ in the average environment (and an expected correlation of $r = 0.71$ between E1 and the ESM); other values were not explored to keep the length of this paper manageable, but we believe these values are relatively conservative yet representative for a

trait measured in wild, uncontrolled conditions. We did not model life-history variation, e.g. longer or shorter individual lifespan. The number of observations per individual can affect the power of I × E analyses and the accuracy of the estimates (Martin et al. 2011, van de Pol 2012) but we were not interested in general aspects of sample size. Furthermore, accuracy and precision may be impacted by heterogeneous residual variances known to occur in wild populations; this point has been addressed in detail elsewhere (Ramakers et al. 2019a) and we therefore do not elaborate here, although we have added one scenario addressing this point in the Supporting information.

Since the environmental variables varied 'annually', meaningful ESMs would be annual means of the studied trait and were hence calculated by averaging the phenotypes over all individuals within years. We stress that using annual means may not be appropriate in all systems and that the temporal (or spatial) scale needs to be sensibly adjusted to the system. Phenotypes were regressed against all environmental variables (E1–E10) and ESMs (all standardised before analysis) in separate analyses using random regression corresponding to model 2 to estimate individual variation in elevations and slopes. Individual identity was fitted as a random effect and the environmental variable or ESM as a continuous fixed effect, interacting with the random effect of individual identity. The significance of variation in slopes was tested with a likelihood-ratio test that compared a model with individual variation in elevations and slopes against a model with only variation in elevations. Heterogeneous residual variation can lead to spurious individual variation in slopes (Ramakers et al. 2019a) but since all potential variation in phenotypes across a given environmental variable was simulated to be driven by variation in slopes, residual variation was homogeneous across environments and was modelled as such (see the Supporting information for an example of heteroscedasticity in the data). The difference in model likelihoods is approximately χ^2-distributed with two degrees of freedom. For each model with a different environmental covariate (E1–E10 and ESM) the estimated variances for elevation and slope, the significance for variation in slopes (I × E) as well as the estimate for the population-level slope were stored. The simulation was iterated 1000 times.

To test whether additional but unidentified variables can bias results from an analysis with ESM more than results from analyses with environmental variables E1–E10, three such variables were simulated in the data: age, habitat effects, and a systematic time trend. Age effects were –0.5, 0, 0.5 and 0.3 for ages of 1–4, roughly simulating an increase with an optimum at age 3 and weak senescence for age 4. Habitat effects were inserted by randomly distributing individuals over 5 different habitats with habitat-specific effects of –0.1, –0.05, 0.0, 0.05 and 0.1. Individuals could disperse to different habitats throughout their lifetime. The time trend was realised by letting average phenotypes increase by 0.01 per year. In every class for each of these effects (age, habitat, time) a small variance (0.001) was assumed. The analyses described previously were repeated 1000 times on the three additional datasets containing either age, habitat effects, or the time trend.

Simulation scenarios (ii–iv): multiple environmental drivers of I × E

To test how the random-regression models performed under more complex I × E interaction, we simulated data where I × E was driven by two (weakly correlated) environmental variables. Specifically, we extended the data generated by model 2 to include the comparatively small I × E effect of a secondary environmental covariate, E10 (correlated with the environmental driver E1 by r = 0.1), such that

$$y_{ij} = \mu + a_i + \left(b_{\text{pop(E1)}} + b_{i(\text{E1})}\right)\text{E1}_j + \left(b_{\text{pop(E10)}} + b_{i(\text{E10})}\right)\text{E10}_j + \varepsilon_{ij}. \quad (3)$$

We compared three scenarios (scenarios ii–iv): in each scenario, $b_{i(\text{E1})}$ was the main driver of I × E such that $b_{i(\text{E1})} \sim N(0,0.5)$ and $b_{i(\text{E10})} \sim N(0,0.1)$), but the population-mean slope differed per scenario such that (ii) $b_{\text{pop(E1)}} = b_{\text{pop}(E10)} = 0.5$ (i.e. of similar size), (iii) $b_{\text{pop(E1)}} = 0.25$ and $b_{\text{pop(E10)}} = 0.5$ (i.e. a larger mean effect of the secondary environmental driver) and (iv) $b_{\text{pop(E1)}} = 0.1$ and $b_{\text{pop(E10)}} = 0.5$ (i.e. a much larger mean effect of the secondary environmental driver). Note that with the larger mean effect of E10 compared to E1, the former effectively becomes the main driver for plasticity. For simplicity, we did not investigate inverse slopes here. The performance of the random-regression model was assessed as above, using one of the environmental proxies (E1–E10 and ESM) as covariate. As in the previous scenario, we additionally evaluated the effects of age, habitat structure and time trends in the data.

Simulation scenario (v): I × E × E

To test the performance of the random-regression of model 2 in the presence of an (unobserved) interaction between environments (i.e. an I × E × E effect in the data), we adapted the data-generating model 2 to include an interaction between the environmental driver (E1) and a categorical environmental effect c, comprising three categories ($t \in \{A, B, C\}$) randomly distributed over the 20 years with equal probability (⅓ each). As before, $b_{\text{pop}} = 0.5$ and each individual's response to E1 was drawn as $b_i \sim N(0,0.5)$, but was made dependent on t such that $b_{iA} = b_i - q_i$, $b_{iB} = b_i$ and $b_{iC} = b_i + q_i$, where $q_i \sim N(0.25,0.25)$, i.e. a decrease ($b_{\text{pop},A} = 0.25$), no change ($b_{\text{pop},B} = 0.5$), and an increase ($b_{\text{pop},C} = 0.75$) in the mean slope, respectively. We also tested a scenario where $q_i \sim N(0.25,0)$ (i.e. no variation in the change in slope) but since the results of interest were very similar we do not report them here. The statistical model for this data is

$$y_{ijt} = c_t + a_i + \left(b_{\text{pop},t} + b_{it}\right)\text{E1}_{j(t)} + \varepsilon_{ijt}, \quad (4)$$

where c_t is the intercept for each environmental category and $b_{\text{pop},t}$ and b_{it} are the (fixed) population and (random) individual slopes with respect to environment E1 in year j specific to environmental category t (i.e. an interaction effect both at the population and individual level). The random elevations and slopes were jointly fitted such that

$$\begin{bmatrix} a \\ b_A \\ b_B \\ b_C \end{bmatrix}_i \sim N\left(\begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_a^2 & \sigma_{a,bA} & \sigma_{a,bB} & \sigma_{a,bC} \\ \sigma_{bA,a} & \sigma_{bA}^2 & \sigma_{bA,bB} & \sigma_{bA,bC} \\ \sigma_{bB,a} & \sigma_{bB,bA} & \sigma_{bB}^2 & \sigma_{bB,bC} \\ \sigma_{bC,a} & \sigma_{bC,bA} & \sigma_{bC,bB} & \sigma_{bC}^2 \end{bmatrix} \right)$$

(note that an unstructured matrix was assumed even though covariances were not explicitly simulated).

Model 4 served as the benchmark model to which we compared the random regression for model 2 for each of the environmental variables (E1–E10 and ESM). For each model we looked at 1) the fit according to AIC, 2) the root mean squared error (RMSE), averaged over all individuals and 3) the slope variance (for model 2 only). To calculate RMSE, predicted values across observed environments were computed for each individual as $\hat{y}_{ij} = \boldsymbol{\eta}_j \times \left(\hat{\boldsymbol{b}}_{(t)} + \hat{\boldsymbol{u}}_{i(t)} \right)$, where is $\hat{y}_{ij}$ is the predicted value for individual i in environment j, $\boldsymbol{\eta}_j$ is the jth row vector of the model matrix ([1, x_j]), $\hat{\boldsymbol{b}}_{(t)}$ contains the fixed coefficients and $\hat{\boldsymbol{u}}_{i(t)}$ contains the random elevation and slope coefficients (with t in parentheses indicating the category-specific coefficients when applicable). RMSE for each individual can then be calculated as $\text{RMSE}_i = \sqrt{\frac{1}{J}\sum_{j=1}^{J}(y_{ij} - \hat{y}_{ij})^2}$. The entire procedure was iterated 1000 times.

Simulation scenario (vi): mean, linear slope changing with the environment

The last scenario investigated here is where the mean slope of the population changed with the environment (i.e. the slope of the mean, linear reaction norm is steeper in later than in earlier years). In theory, this could occur when the environment changes over time and the population evolves to a steeper/shallower mean reaction norm. To examine this scenario, a generous correlation of r = 0.8 between 'year' and the cue (E1) was generated. The twenty environments were equally divided into three environmental categories ($t \in$ {A, B, C}), where each group contained six or seven consecutive values of the environment E1. Similarly to previous scenarios, individuals were randomly assigned a first year, but due to the correlation between year and E1, observations of the same individual were likely to be in similar environments. When observations of an individual spanned multiple categories (e.g. A and B), the one that occurred most often was assigned as this individual's category; with equal occurrence (e.g. A, A, B, B), the category was chosen randomly. Phenotypes were drawn as before, with $b_i \sim N(0,0.5)$ throughout, but with $b_{\text{pop}} \in \{0.5,1,1.5\}$, depending on the environmental category t. Therefore, an individual's phenotype was described as

$$y_{ijt} = \mu + a_i + \left(b_{\text{pop},t} + b_i\right)\text{E1}_{j(t)} + \varepsilon_{ijt}, \quad (5)$$

where the mean slope depended on t. The model in Eq. 2 was run as before on these data, using either of the 10 environmental proxies (E1–E10) or ESM as covariate.

Two practical examples: I × E in common guillemots and G × E in great tits

To exemplify the use of the ESM approach with real data, we reanalysed two real datasets. The first one concerns I × E in breeding time in the common guillemot *U. aalge* on the Isle of May, Scotland (Reed et al. 2006). The second example involves G × E in fledgling weights of great tits *P. major* from the Hoge Veluwe, the Netherlands, obtained from an online repository (Mulder et al. 2016a, b). The purpose of these reanalyses is to show how the ESM approach can aid in unveiling the magnitude of I × E/G × E in the data and how ESM-based estimates of reaction norms can actually be more accurate than a coarse environmental proxy. For full details of the analysis, we refer the reader to the Supporting information.

The breeding dataset for the common guillemot contains 2593 egg-laying dates of 245 females guillemots, recorded over the period 1983–2005 in five sub-colonies ranging from 23 to 69 individuals each. The median number of breeding attempts per female is 10, with a range of 4–22. Annual mean laying date in this population correlates negatively ($r^2 = 0.24$) with the North Atlantic Oscillation (NAO) index, which was found to be the strongest environmental predictor for guillemot phenology among local sea-surface temperatures and number of days with easterly winds across a 21 year study (Frederiksen et al. 2004), but individual variation in reaction-norm slopes (I × E) was reported to be small (slope variance: 0.01) and statistically insignificant (Reed et al. 2006). The guillemot data are therefore a good candidate for investigating the presence of I × E despite the apparent absence of NAO-driven I × E. To assess the presence of I × E in the guillemot population, we regressed laying date against NAO or ESM (mean laying date within a year) using an adjusted version of model 2 that accounted for different mean responses to the environment within the different subpopulations and for year effects. In addition, we calculated for each individual the fit of each line through individual-specific root mean-squared error (RMSE) values (Supporting information).

For the great tit fledgling-weight dataset, the original analysis by Mulder et al. (2016b) aimed at analysing the genetics of within-brood variability of fledgling weights. The dataset contains all broods from 1973 to 2012 with ≥ 5 nestlings that were weighed shortly before fledgling age at 15 days (n = 17 535 nestlings from 2175 broods). We note,

therefore, it is not a completely random subset as it excludes nests with smaller clutch sizes or those where part of the nestlings died before weighing (see the Supporting information for a histogram of clutch sizes and number of nestlings). We used the dataset to estimate G × E, but not I × E, in fledgling weight across different years, since each nestling had only one observation and pedigree information is available (see Mulder et al. 2016a, b for details). We chose the number of nests in a particular year – a proxy for breeding-pair density and thus food availability through between-brood competition – as the main covariate for our models and contrasted it to a model where the mean fledgling weight of the population (ESM) was the main covariate. Again, the model used was an adapted version of model 2 that accounted for pedigree information to be able to estimate G × E, as well as for within-brood competition and hatch date and for effects of year, mother and brood. To assess the model fit using either metric (breeding density versus ESM), RMSEs were calculated, but as this could not be done per individual we fitted year-specific RMSEs on the fitted values (Supporting information).

Results

Simulation models (i–iv): single and double sources of I × E

Unsurprisingly, in all I × E scenarios (i–iv) using the environmental variable that was used to simulate the phenotypes (E1) as a covariate in the analysis recovered the input values for reaction-norm slopes and variances in elevations and slopes closely (Fig. 2; elevations not shown). The estimates for population-level slopes and among-individual variation in slopes clearly declined with a decreasing correlation of the used covariate with the 'real driver' (E1). As the estimated individual variation in slopes declined, the power, i.e. the proportion of replicates in which I × E was statistically significant, decreased, indicating that it would be increasingly unlikely to detect statistically significant individual variation in slopes as the correlation with E1 decreased (Supporting information).

The estimates for individual variation in slopes from the model using ESM as covariate closely matched, on average, the input value of 0.5 caused by E1 in the first scenario (Fig. 2a, b). However, when an additional (small) I × E variance of 0.1 (caused by E10) was present in the data (scenarios ii–iv), ESMs performed less well compared to the first scenario, particularly when the mean slope for E10 was higher than for E1 (Fig. 2c, d versus e, f and g, h). Nevertheless, in the second scenario (with equal mean slopes related to E1 and E10), ESM still outperformed, on average, most other proxies with the exception of E1 and E2 (which have a $r \geq 0.9$ with the real driver (E1); Fig. 2d). In the third scenario (with unequal mean slopes between E1 and E10), ESM performed approximately equally well as proxy E4 ($r = 0.7$ with real driver; Fig. 2f), whereas in the fourth scenario, ESM performed approximately equally well as proxy E5 ($r = 0.6$ with real driver; Fig. 2h). In scenarios (i) and (ii), the power to detect I × E at $p < 0.05$ using ESM was high (~ 0.8; Supporting information), but moderate in scenarios iii and iv. Thus, both in the absence and presence of a secondary (small) source of I × E (E10), ESM performed well to moderately in recovering the primary source of I × E (E1), but as the mean slope effect of E10 increased relatively to that of E1, recovered slope variances increasingly reflected E10- rather than E1-induced I × E (hence the reduction in power). This is to be expected since E10 is now in fact the main environmental driver for (mean) plasticity.

Other factors that affect the phenotype but are unaccounted for will alter the ESM phenotypes and can thereby potentially affect the results. Here, age, habitat structure, and a systematic time trend were tested. None of them led to any systematic bias in the probability to detect I × E and its estimates in the first scenario (a single environmental driver of I × E; see the Supporting information). However, in the second (ii) and third (iii) scenarios (with multiple environmental drivers of I × E) the presence of these factors improved how well the ESM-based estimates of mean slope and among-individual variance in slopes did in comparison to E1–E10 (Supporting information). Thus, additional sources of variation in the data partly negated the detrimental effect of E10 on the accuracy with which ESM could estimate E1-induced I × E variance (cf. Fig. 2). In scenario (iv), however, where the mean slope of E10 was much higher than that of E1, estimation of I × E variance by ESM that reflect E1 was clearly compromised (Supporting information). Again, however, this must be viewed in the light that with an increasing mean slope with respect to E10, the latter actually becomes the main environmental driver.

Simulation model (v): I × E × E

In the scenario where datasets contained an interaction effect between environments (E1 × environmental category), the I × E × E random regression of model 4 (conforming to the data) naturally performed best based on AIC as well as RMSE (Fig. 3a–b). In the random regressions ignoring the interaction (model 2), proxy E1 (the driver) performed best based on AIC and was directly followed by ESM; all other proxies performed less well than ESM (Fig. 3a). Based on average RMSE, the I × E × E model performed best as expected and was followed by random regressions fitting proxies E1–E3; ESM performed as well as E4 ($r = 0.7$ with real driver), although differences between the proxies were small (Fig. 3b). Standard deviations in RMSE for ESM, however, were on par with that for E1, indicating a higher consistency in prediction errors across individuals. Estimated slope variances decreased as the correlation with E1 decreased, but the variance for ESM was on par with E2/E3 (Fig. 3c). Combined, these results suggest that ESM performed well in the presence of an I × E × E interaction.

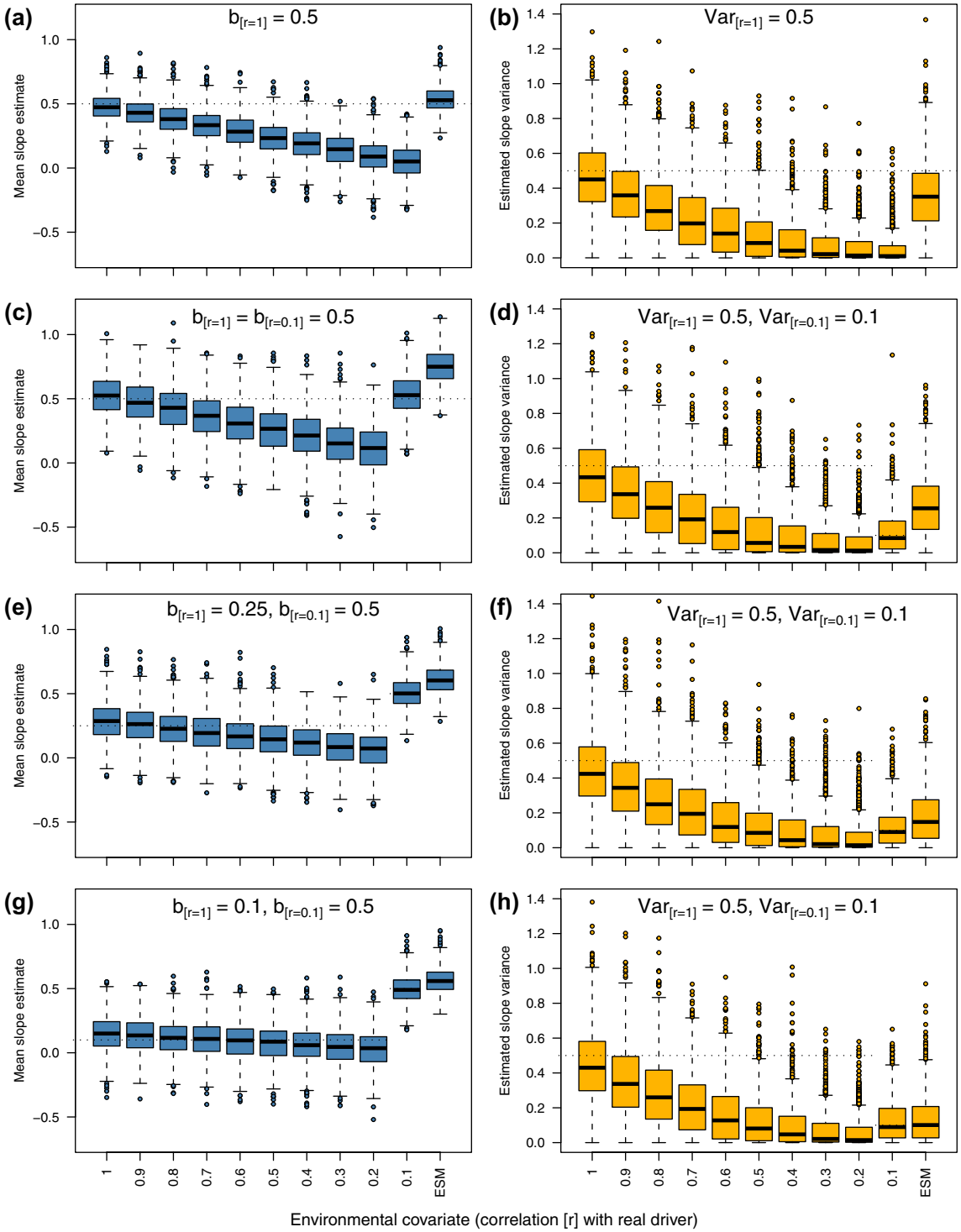

Figure 2. Boxplots of estimates for the population-level slope (a, c, e, g) and individual variation in slopes (b, d, f, h) depending on the covariate included in the random-regression model for simulation scenarios (i–iv) (scenarios did not include any effects of age, habitat, or time trend). The covariates included in the model are indicated by their correlation with environment E1; ESM indicates environment-specific mean phenotypes as covariate. (a, b) Only a single environmental driver (E1 [r = 1]) is responsible for I × E in the data; (c, d) two environmental drivers (E1 [r = 1] and E10 [r = 0.1]) for I × E in the data, with similar mean slopes but different slope variances; (e, f) two environmental drivers (E1 [r = 1] and E10 [r = 0.1]) for I × E in the data with different mean slopes and different slope variances (with E10 now effectively being the main driver of the phenotype); (g, h) two environmental drivers (E1 [r = 1] and E10 [r = 0.1]) for I × E in the data with an even greater difference between mean slopes. The dashed lines indicate the input values for the mean slope and variation in slopes. Bold line indicates median, box margins 1st and 3rd quartile and whiskers 1.5 × inter-quartile range.

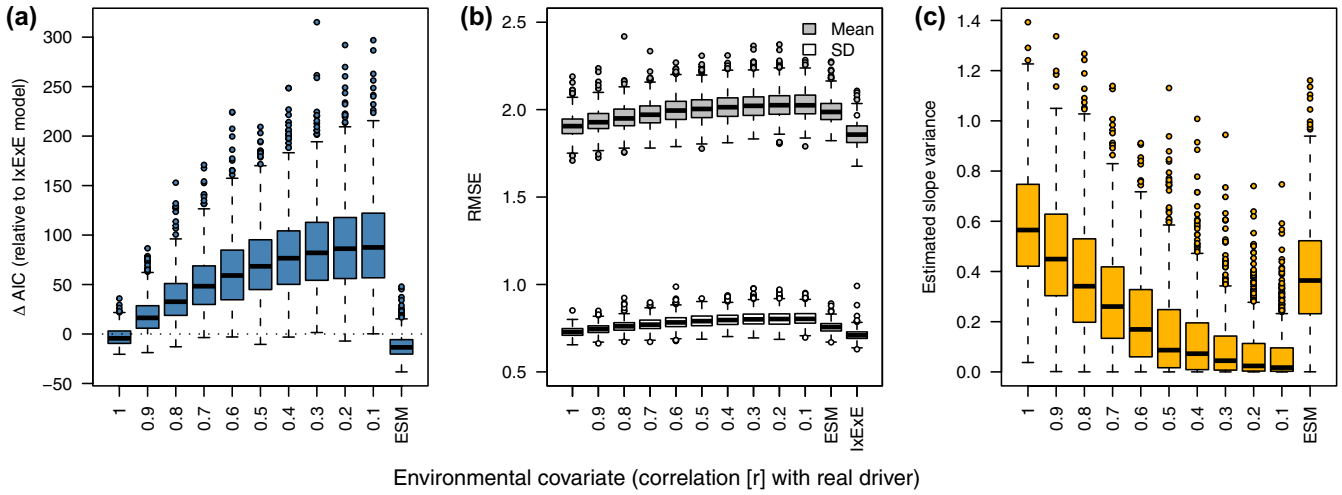

Figure 3. Results for simulation scenario (v), where an I × E × E effect was present in the data. ΔAIC values (a) are referenced to the benchmark random-regression model that included an I × E × E effect (lower is better). RMSE (b) are averages (grey) and standard deviations (no colour) across individuals. Slope variances (c) are estimates from random regressions using one of the environmental proxies as covariate. The covariates included in the random regressions are indicated by their correlation with environment E1; ESM indicates environment-specific mean phenotypes as covariate; I × E × E (in b only) represents the benchmark interaction random-regression model conforming to the data structure. Bold lines in boxplots indicate median, box margins 1st and 3rd quartile and whiskers 1.5 × inter-quartile range.

Simulation model (vi): change in mean, linear slope across environmental gradient

In the scenario where the mean slope changed with the environment (with $b_{pop} = 0.5$, 1 or 1.5 across three environmental categories), the estimated population slope from the model with the driver (E1) as a covariate was estimated, on average, to be around 0.9 (we would expect it to be around 1, i.e. the average of the three input values) (Fig. 4a). For the other proxies, the estimated slope was lower except for ESM, which averaged at 1, following expectations. slope variation was slightly above the input value of 0.5 for E1, gradually decreasing for other proxies; for ESM, the estimated slope variance ranked between that of E2 and E3 (i.e. r = 0.9 and r = 0.8 with E1, respectively), making ESM a better covariate than most other proxies.

Common guillemot analysis

As shown previously (Reed et al. 2006), there was little evidence for I × E variance in the guillemot population using NAO as the covariate (Fig. 5a, c; estimated variance 0.043 ± 0.271; $\chi^2 = 0.078$, df = 2, p = 0.63). The ESM model, however, provided evidence for significant I × E variance (Fig. 5b,

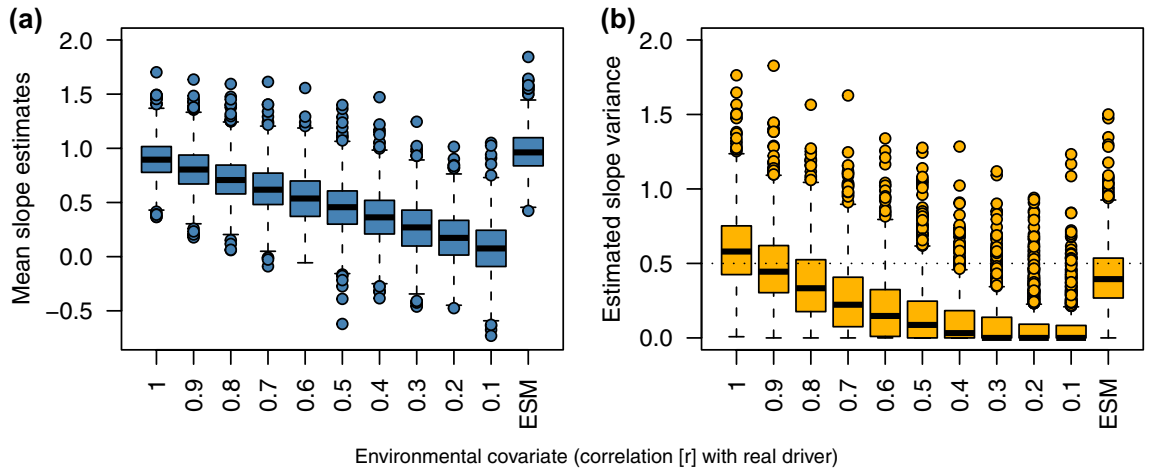

Figure 4. Boxplots of estimates for the population-level slope (a) and individual variation in slopes (b) depending on the covariate included in the random-regression models for simulation scenario (vi), i.e. with a change in mean slopes in the data being dependent on the environment E1. The covariates included in the model are indicated by their correlation with environment E1; ESM indicates environment-specific mean phenotypes as covariate. The input mean slope value was 0.5, 1 or 1.5, increasing in steps with E1 (Methods); its estimates with respect to the cue (E1) are therefore expected to average at around 1.

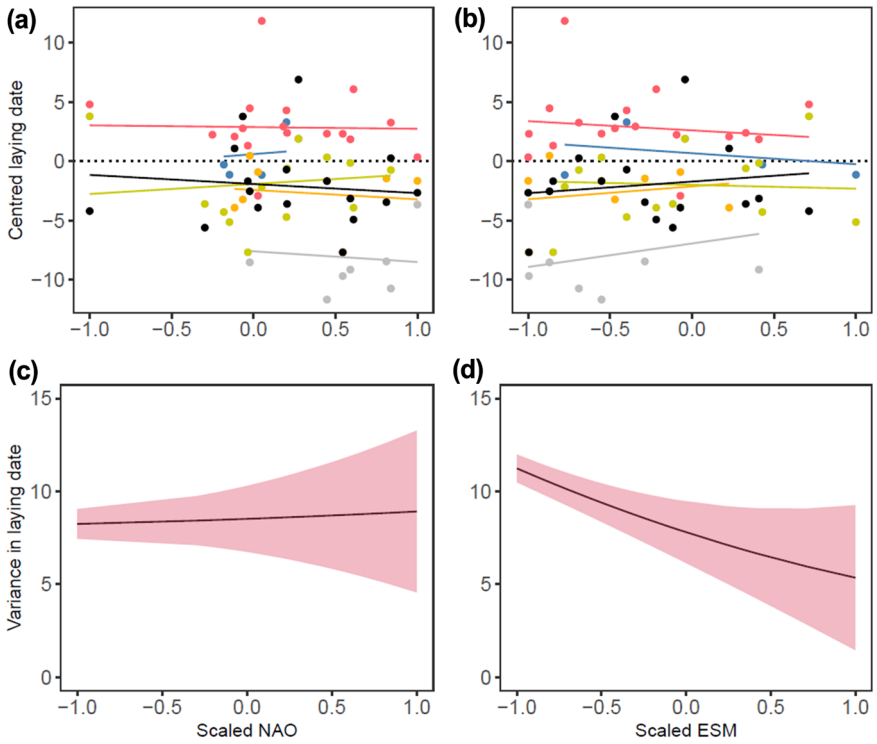

Figure 5. Fitted reaction norms (deviations from the mean trend) for six randomly selected female guillemots (a, b) as well as the phenotypic variance for laying date across the environment (c, d), resulting from a random-regression model with NAO (a, c) or ESM (b, d) as the environmental covariate (both covariates scaled between –1 and 1). Lines in (c) and (d) represent estimates with approximate 95% CIs.

d; $\chi^2 = 16.421$, df = 2, p = 0.0001), although variance in slopes was modest (0.338 ± 0.227). Predicted reaction norms (Fig. 5a, b) were more accurate in the ESM model than the NAO model (difference average $RMSE_i$ across individuals: 2.86–2.82 = 0.04; bootstrapped 95% CI of difference: 0.01–0.07). Overall, I × E was thus modest but clearly present in the guillemot population but could only be detected statistically when using ESM as the covariate.

Great tit analysis

Great tit fledgling weight responded quadratically, though weakly, to standardized breeding density at the population level (linear and quadratic slope 0.803 ± 0.229 and –0.508 ± 0.183, respectively). Additive genetic variance varied significantly with breeding density, indicating G × E (Fig. 6a; estimated variance: 0.179 ± 0.106; $\chi^2 = 43.76$, df = 2, p < 0.0001). Replacing breeding density with ESM (average weight) increased estimated G × E variance (1.112 ± 0.119; $\chi^2 = 144.73$, df = 2, p < 0.0001; although note that G × E variances are not directly comparable between methods as ESM and breeding density do not correlate perfectly) and considerably increased the range of estimated genetic variance across different years (Fig. 6b). Annual RMSE was, on average, lower in the ESM compared to the breeding-density model (difference average $RMSE_i$: 0.705 – 0.600 = 0.105; 95% CI of difference: 0.014–0.195). Overall, G × E in fledgling weight was present in the data and was most pronounced when using ESM as the covariate.

Discussion

Performance of ESMs as an environmental covariate in simulations

As has been found previously (Brommer et al. 2005, Husby et al. 2010) and systematically explored here, the choice of the environmental variable that affects the phenotype in the reaction norm can affect the probability to detect variation in slopes. The less related the included environmental variable is to the real driver of plasticity, the

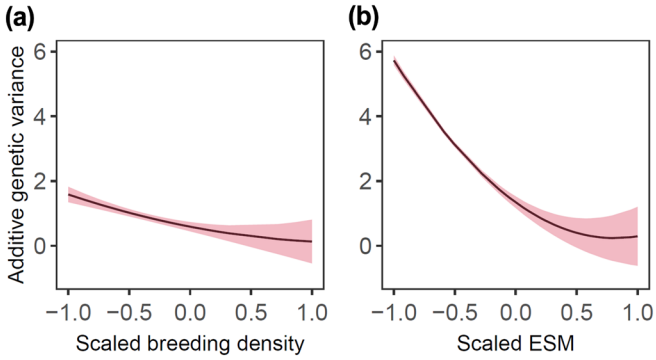

Figure 6. Estimated additive genetic variances for fledgling weight in great tits as a function of breeding density (a) and ESM, i.e. average weight (b) (both covariates scaled between –1 and 1). Lines represent estimates with approximate 95% CIs.

lower the probability to detect statistically significant I × E (Supporting information). This may lead the researcher to (wrongfully) dismiss the potential for I × E or G × E (henceforth collectively termed slope variation) based on testing a single covariate, or a few covariates. Using environment-specific population means (ESM) as a covariate, inspired by the 'Finlay–Wilkinson' regression (Yates and Cochran 1938, Finlay and Wilkinson 1963), could provide an alternative means to explore slope variation in the absence or presence of alternative proxies, even in the presence of additional sources of (small) I × E (Fig. 2), I × E × E (Fig. 3) or a change in mean slope across the environmental gradient (Fig. 4).

It could be shown (Appendix A) that the correlation between the true driver of plasticity and the ESM approaches one with decreasing 'sampling variation' in the ESM. This sampling variation should decrease with increasing sample size. However, since it also incorporates variation due to individual (or genetic) variation in reaction norms, it will also decrease with a decrease in reaction-norm variation. This means the smaller the effect of interest becomes, the more precise the estimated ESMs become and the better their suitability to detect variation in reaction norms. If the presence of slope variation has been shown using ESM, further analyses can aim to identify the meaningful biotic or abiotic environmental variables that drive slope variation in the trait of interest (i.e. back-tracing the results obtained with ESM). If slope variation of similar magnitude to that with ESM can be found using an environmental variable, it can then be concluded that this variable is reasonably closely correlated with the environmental variable that causally affects the trait.

A caveat is that when multiple sources of slope variation exist in the data, this variation will in principle be harder to detect using ESM, provided that the population harbours strong mean plasticity with respect to the 'secondary' compared to the 'primary' environmental driver. Quotation marks are in order, since in this case the former would in fact be the primary driver, for which I × E variation can be recovered very well with ESM (Fig. 2). Furthermore, estimation of slope variation with ESM may be impaired when multiple environmental drivers collectively explain more variation in the trait than the main driver. However, in this case it will also be substantially more complicated – if not impossible – to correctly identify all these environmental drivers. Our simulations show, however, in the former scenario (two sources of I × E) that ESM may still perform equally well as other environmental proxies that correlate with the 'primary' driver E1 by $r \geq 0.7$ (but note, again, that in scenarios (iii) and (iv) the E1 de facto becomes the secondary driver); in the latter scenario (v) (I × E × E), ESM performed as well as the proxies that correlate by $r \geq 0.7$ up to 0.9 (whether it be based on RMSE or AIC, respectively). In this sense, a 'Finlay–Wilkinson'-derived approach can still serve as a 'yardstick' with which models featuring other environmental covariables can be compared.

Relevance of the ESM approach in real data

Why should we be bothered to look at ESM when studying (variation in) plasticity in the wild? There are a number of reasons why we should aspire to study I × E (and G × E) in general, even if we do not have an accurate proxy for E, as well as why we should consider ESM even when we do have reasonable proxies, which we outline (non-exhaustively) below.

First, whether or not I × E, and also genetic variation in slopes (G × E), is present in a population or species is biologically relevant, even if the true environmental driver of plasticity cannot be identified. For example, I × E can increase or decrease the amount of among-individual variation in novel environments, which affects the opportunity for selection and thereby potential evolutionary change, even in the absence of G × E (Hoffman and Merilä 1999, Lédon-Rettig et al. 2014, Ramakers et al. 2018b). The absence of I × E can also be biologically interesting; for example, Reed et al. (2006) found no I × E in breeding time of common guillemots *U. aalge* in response to the North Atlantic Oscillation and argued that this lack of I × E was a consequence of the benefits of coordinated breeding in colonial birds, such as the common guillemot. However, in such cases it would be desirable to be sure that this absence of I × E is not caused by an unsuitable choice of the environmental variable affecting the trait. In fact, when re-analysing this data using ESM – in this case annual mean egg-laying dates – as environmental variable, statistically significant I × E was found, although the magnitude of I × E was low, conforming to expectation considering the biology of the species (discussion in Reed et al. 2006). When comparing the fit of individual reaction norms, the root mean-squared error was lower (on average) for ESM than for NAO, which inspires confidence that the ESM approach provided a better fit and hence made for better predictions in particular environments.

Second, ESM can be used to give a meaningful, biological interpretation of the effect of environmental proxies on phenotypic responses. In experimental research, for example, manipulating one aspect of the environment in order to elicit a phenotypic response may inadvertently lead to changes in other aspects of the environment, obscuring the true effect of the manipulation. Food supplementation, for example, has been a popular method of studying the timing of avian breeding, more specifically why birds breed at the time that they do (Verhulst and Nilsson 2008, Ruffino et al. 2014). Leaving aside for the moment the issues regarding the validity of such experiments in answering the specific question at hand, the food supplementation itself may lead to an increase in the population density, may alter the dynamics of competition, and may attract predators, all of which may impact the expression of the studied phenotype alongside the food manipulation itself. Hence, in the study of plasticity, it is uncertain whether the environmental proxy termed 'food availability' is solely responsible for the observed phenotypic responses. Comparing the results to a model using the ESM will shed light on this.

Third, while the ESM approach per se does not give biological insights into the environmental variable underlying phenotypic plasticity and slope variation in the analysed trait, it does have a methodological application, related to how important evolutionary parameters are estimated. For example, we may be interested in estimating, retrospectively, the evolutionary potential (and response) of the trait over time or across locations (discussion in Ramakers et al. 2018b). For this purpose, evolutionary parameters (selection, genetic variance, breeding values) need to be computed at the level of the environment (year/location), but this will in most practical cases be impossible due to sample size. For example, estimating the evolutionary response in a given environment at the genetic level directly (Morrissey et al. 2012) equates to estimating an additive genetic covariance between fitness and the focal trait in that environment; to pull this off statistically, we would need unrealistically rich relatedness data for each environment (Gienapp et al. 2017). Random-regression models get around this by estimating (genetic) reaction-norm parameters (elevation, slope and their covariance), from which per-environment (genetic) variances can be estimated directly (Kirkpatrick et al. 1990, Meyer 1998). To do a random-regression analysis, however, we need a decent covariate; if a such an informative proxy cannot be identified from available environmental data, the ESM provides a solid – and likely a less biased – alternative (Ramakers et al. 2018b).

Fourth, if we are specifically interested in predicting the phenotypic or evolutionary consequences of (directional) environmental change, the ESM approach can be of aid here as well. Naturally, unravelling slope variation per se is uninformative in this regard but the ESM approach can help 'benchmark' the used proxy. For example, in the guillemot and great tit data, the best conceivable proxies (NAO and breeding density, respectively) were outperformed by ESM (Fig. 5, 6). This means that any predictions based on these proxies, such as predicting future changes in guillemot breeding time based on NAO indices derived from climate models, will likely be unreliable and that it would be worth identifying other, better performing, proxies. An important caveat here – which is not just limited to the ESM approach – is that the underlying relationships between different environmental cues and the trait should remain (largely) unchanged. This is because in case of an environmental shift, even when the cue and environmental proxy shift at the same rate, the prediction of the shift in phenotypes using environmental proxies will tend to be under- or overestimated (see Appendix B for a brief numerical elaboration on this matter).

Finally, the relative size of I × E interaction with respect to ESM versus an environmental proxy may hint toward the presence of complex I × E in the trait. Under 'simple' I × E, one may expect ESM to outperform proxies that correlate even relatively strongly with the true environmental driver (Appendix A, Fig. 2a–b). However, when estimated 'I × ESM' variance is smaller than 'I × proxy' variance, this indicates that 1) the chosen proxy is a particularly relevant one for the organism and that 2) there may be some underlying complexity (e.g. I × E × E) that ESM is not able to pick up. The nature and extent of this complexity will not be revealed by merely comparing the estimated slope variances, but it will hopefully spark a quest to tease apart I × E further and ultimately fit better models.

Practical considerations when using ESM approach

An important assumption underlying the 'Finlay–Wilkinson' regression is that the response to the 'true' driver of plasticity is (approximately) linear, as strongly non-linear reaction norms (e.g. quadratic or sigmoidal) could lead to identical mean phenotypes in different environments, which would interfere with reliably estimating slope variation. It is also important that the ESM be based on adequate sample sizes (e.g. ≥ 50), which so happens to be a requirement for reliable estimation of G × E per se (Calus et al. 2004). Moreover, the sample on which ESM is based should be reasonably assumed to be random; non-random missingness of phenotypes, for example due to different between-individual variation in phenotypes between environments, should lead to biased estimation of slope variance with respect to ESM. To ensure a good fit of the data, it is good practice to compare the accuracy of fitted individual reaction norms in observed environments between different covariates (in the case of I × E interactions), as we did in simulation scenario (v) and in the guillemot example, for example by computing the root mean squared error on each reaction norm. This will, however, be less straightforward to do for G × E interactions (which essentially entail family reaction norms), although model fits can still be compared, e.g. through environment-specific RMSE and information criteria. Additionally, directly comparing estimated residual variances between models may reveal how well a particular environmental covariate explains G × E variation, as larger residual variances would indicate a greater amount of G × E left unexplained.

Trait expression is not only affected by environmental variables but also by individual 'state' variables, such as age or physical condition and potentially additional environmental variables, for example habitat in addition to temperature. Not accounting for such variables can bias ESM phenotypes. For example, if age structure in the population varies from year to year, this could potentially lead to specific biases in estimates from 'Finlay–Wilkinson'-inspired regressions. This potential bias may be especially problematic when there is a consistent time trend in phenotypes, as caused, for example, by a genetic selection response. Such potential biases would also affect the phenotypes, albeit not the covariate, of analyses based on environmental variables. The results here show that random-regression models with ESM are not biased in this respect. In fact, under more complex I × E patterns (i.e. two sources of I × E), the use of ESMs as a covariate led to better estimates of I × E by negating the secondary I × E effect at least in scenarios (ii) and (iii) (Supporting information). Moreover, ESM performed admirably well in the presence of I × E × E (Fig. 3). It is, however, always preferable to include and test all variables potentially affecting the trait to obtain more accurate estimates of (variation in) reaction norms and also to gain a better understanding of which abiotic and biotic variables affect the trait.

Genetic variation in slopes (G × E) is biologically relevant as environmental change likely leads to selection on both elevation and slope (Gienapp et al. 2014). How the power to detect G × E, rather than I × E, depends on the choice of the covariate included in the analysis was not addressed in the simulations because of the specific issues with the power of quantitative genetic analyses. The sampling variation in heritability estimates depends on the sample size but also on the variation of relatedness within the population (Visscher and Goddard 2015). The variation in relatedness depends on a number of species- or population-specific ecological parameters such as dispersal or mating system. It would have been possible to simulate G × E but the obtained results would have been difficult to generalise. However, I × E is generally regarded as an 'upper limit' for G × E. Having found no I × E with sufficient sample size and using the 'Finlay–Wilkinson'-inspired regression it would hence be highly unlikely to find G × E in the same population. It should, however, be noted that this 'yardstick' cannot universally be applied, as in short-lived species that cannot be sampled repeatedly the necessary repeat observations of individuals may never be achieved. Hence, I × E may not be well estimable but – given suitable relatedness information – quantitative genetic analyses could be possible and statistically significant G × E could be found. A good case in point is the great tit fledgling weight data presented in the present study. For this type of data, repeat observations do not exist (as individuals are fledglings only once in their lives) but our data did show that G × E in fledgling weight was substantially more sizeable with respect to ESM than to breeding-pair density. This is likely because ESM is a more accurate description of the general environment (including, for example, food availability) than breeding-pair density alone. The difference in G × E magnitude (range of estimated additive genetic variance) means potentially a difference in environment-specific heritability and hence estimates of evolutionary potential.

Concluding remarks

In animal and plant breeding the 'Finlay–Wilkinson' regression has long been used in the context of 'genotypic stability' analysis, but very rarely outside this field (James 2009). We argued here that it can be usefully extended to a random-regression framework and that it has its biological merits in the study plasticity in wild populations. It can be used as a 'yardstick' in analyses exploring individual (and genetic) variation in slopes as its results are unbiased by the correlation between it and the environmental variable causally affecting the trait. This can be especially relevant for studies not finding statistically significant slope variation using known environmental proxies and could therefore give us a better understanding of how prevalent (or not) I × E interactions really are. Importantly, we argued that the environment-specific trait means (ESM) can be a useful starting point for the search for quantifiable environmental proxies. We caution, however, that in the presence of multiple sources of I × E, this ESM approach may lead to too conservative estimates of I × E (and by extension G × E), and we would therefore encourage researchers to think critically about the mechanisms that could give rise to plasticity in the first place.

Acknowledgements – M. E. Visser, H. A. Mulder, T. Van Dooren, D. Réale and J. Martin commented on earlier versions of the manuscript and provided useful feedback.

Author contributions

Jip J. C. Ramakers: Conceptualization (equal); Formal analysis (lead); Methodology (lead); Writing – original draft (equal); Writing – review and editing (equal). **Thomas E. Reed**: Conceptualization (equal); Data curation (equal); Writing – review and editing (equal); **Michael P. Harris**: Data curation (equal); Writing – review and editing (supporting). **Phillip Gienapp**: Conceptualization (lead); Formal analysis (equal); Methodology (equal); Writing – original draft (equal); Writing – review and editing (equal).

Data availability statement

The great tit data used in this article are available online (Mulder et al. 2016a). The guillemot data are available from the Dryad Digital Repository: https://doi.org/10.5061/dryad.zcrjdfnjr (Ramakers et al. 2023). R code (www.r-project.org) used for the simulations are included in the online Supporting information.

Supporting information

The Supporting information associated with this article is available with the online version.

References

Bates, D. et al. 2018. Package 'lme4': linear mixed-effects models using 'Eigen' and S4. – CRAN. https://cran.r-project.org/web/packages/lme4/lme4.pdf.

Bourret, A., Bélisle, M., Pelletier, F. and Garant, D. 2015. Multidimensional environmental influences on timing of breeding in a tree swallow population facing climate change. – Evol. Appl. 8: 933–944.

Brommer, J. E. 2019. More evidence is needed to show that heritability and selection are not associated. – Nat. Ecol. Evol. 3: 1407–1407.

Brommer, J. E., Merilä, J., Sheldon, B. C. and Gustafsson, L. 2005. Natural selection and genetic variation for reproductive reaction norms in a wild bird population. – Evolution 59: 1362–1372.

Brommer, J. E., Rattiste, K. and Wilson, A. J. 2008. Exploring plasticity in the wild: laying date-temperature reaction norms in the common gull *Larus canus*. – Proc. R. Soc. B 275: 687–693.

Buoro, M., Gimenez, O. and Prévost, E. 2012. Assessing adaptive phenotypic plasticity by means of conditional strategies from

empirical data: the latent environmental threshold model. – Evolution 66: 996–1009.
Butler, D., Cullis, B. R., Gilmour, A. R., Gogel, B. J. and Thompson, R. 2017. ASReml-R reference manual, ver. 4. – VSN International Ltd.
Calus, M. P., Bijma, P. and Veerkamp, R. F. 2004. Effects of data structure on the estimation of covariance functions to describe genotype by environment interactions in a reaction norm model. – Genet. Select. Evol. 36: 489.
Charmantier, A., McCleery, R. H., Cole, L. R., Perrins, C., Kruuk, L. E. and Sheldon, B. C. 2008. Adaptive phenotypic plasticity in response to climate change in a wild bird population. – Science 320: 800–803.
Finlay, K. W. and Wilkinson, G. N. 1963. The analysis of adaptation in a plant-breeding programme. – Aust. J. Agric. Res. 14: 742–754.
Frederiksen, M., Harris, M. P., Daunt, F., Rothery, P. and Wanless, S. 2004. Scale-dependent climate signals drive breeding phenology of three seabird species. – Global Change Biol. 10: 1214–1221.
Froy, H., Martin, J., Stopher, K. V., Morris, A., Morris, S., Clutton-Brock, T. H., Pemberton, J. M. and Kruuk, L. E. B. 2019. Consistent within-individual plasticity is sufficient to explain temperature responses in red deer reproductive traits. – J. Evol. Biol. 32: 1194–1206.
Gienapp, P., Hemerik, L. and Visser, M. E. 2005. A new statistical tool to predict phenology under climate change scenarios. – Global Change Biol. 11: 600–606.
Gienapp, P., Reed, T. E. and Visser, M. E. 2014. Why climate change will invariably lead to selection on phenology. – Proc. R. Soc. B 281: 20141611.
Gienapp, P., Fior, S., Guillaume, F., Lasky, J. R., Sork, V. L. and Csilléry, K. 2017. Genomic quantitative genetics to study evolution in the wild. – Trends Ecol. Evol. 32: 897–908.
Hau, M. and Goymann, W. 2015. Endocrine mechanisms, behavioral phenotypes and plasticity: known relationships and open questions. – Front. Zool. 12: S7.
Henderson, C. R. 1982. Analysis of covariance in the mixed model: higher-level, non-homogeneous, and random regressions. – Biometrics 38: 623–640.
Hoffman, A. A. and Merilä, J. 1999. Heritable variation and evolution under favourable and unfavourable conditions. – Trends Ecol. Evol. 14: 96–101.
Husby, A., Nussey, D. H., Visser, M. E., Wilson, A. J., Sheldon, B. C. and Kruuk, L. E. 2010. Contrasting patterns pf phenotypic plasticity in reproductive traits in two great tit (*Parus major*) populations. – Evolution 64: 2221–2237.
James, J. W. 2009. Genotype by envionment interaction in farm animals. – In: van der Werf, J., Graser, H.-U., Frankham, R. and Gondro, C. (eds), Adaptation and fitness in animal populations: evolutionary and breeding perspectives on genetic resource management. Springer, pp. 151–167.
Kirkpatrick, M. 1989. A quantitative genetic model for growth, shape, reaction norms, and other infinite-dimensional characters. – J. Math. Biol. 27: 429–450.
Kirkpatrick, M., Lofsvold, D. and Bulmer, M. 1990. Analysis of the inheritance, selection and evolution of growth trajectories. – Genetics 124: 979–993.
Kokko, H. and Heubel, K. 2008. Condition-dependence, genotype-by-environment interactions and the lek paradox. – Genetica 134: 55–62.
Lédon-Rettig, C. C., Pfennig, D. W., Chunco, A. J. and Dworkin, I. 2014. Cryptic genetic variation in natural populations: a predictive framework. – Integr. Comp. Biol. 54: 783–793.
Lynch, M. and Walsh, B. 1998. Genetics and analysis of quantitative traits. – Sinauer Assoc.
Malosetti, M., Ribaut, J. M. and van Eeuwijk, F. A. 2013. The statistical analysis of multi-environment data: modeling genotype-by-environment interaction and its genetic basis. – Front. Physiol. 4: 44.
Martin, J. G. A., Nussey, D. H., Wilson, A. J. and Réale, D. 2011. Measuring individual differences in reaction norms in field and experimental studies: a power analysis of random regression models. – Methods Ecol. Evol. 2: 362–374.
Meyer, K. 1998. Estimating covariance functions for longitudinal data using a random regression model. – Genet. Select. Evol. 30: 221.
Morrissey, M. B. and Liefting, M. 2016. Variation in reaction norms: statistical considerations and biological interpretation. – Evolution 70: 1944–1959.
Morrissey, M. B., Parker, D. J., Korsten, P., Pemberton, J. M., Kruuk, L. E. and Wilson, A. J. 2012. The prediction of adaptive evolution: empirical application of the secondary theorem of selection and comparison to the Breeder's equation. – Evolution 66: 2399–2410.
Mulder, H. A., Gienapp, P. and Visser, M. E. 2016a. Data from: Genetic variation in variability: phenotypic variability of fledging weight and its evolution in a songbird population. – Dryad Digital Repository, http://dx.doi.org/10.5061/dryad.2qv8n.
Mulder, H. A., Gienapp, P. and Visser, M. E. 2016b. Phenotypic variability of fledging weight and its evolution in a songbird: do families differ genetically? – Evolution 70: 2004–2016.
Nussey, D. H., Postma, E., Gienapp, P. and Visser, M. E. 2005. Selection on heritable phenotypic plasticity in a wild bird population. – Science 310: 304–306.
Nussey, D. H., Wilson, A. J. and Brommer, J. E. 2007. The evolutionary ecology of individual plasticity in wild populations. – J. Evol. Biol. 20: 831–844.
Pigliucci, M. 2005. Evolution of phenotypic plasticity: where are we going now? – Trends Ecol. Evol. 20: 481–486.
Ramakers, J. J. C., Gienapp, P. and Visser, M. E. 2018a. Phenological mismatch drives selection on elevation, but not on slope, of breeding time plasticity in a wild songbird. – Evolution 73: 175–187.
Ramakers, J. J. C., Culina, A., Visser, M. E. and Gienapp, P. 2018b. Environmental coupling of heritability and selection is rare and of minor evolutionary significance in wild populations. – Nat. Ecol. Evol. 2: 1093–1103.
Ramakers, J. J. C., Visser, M. E. and Gienapp, P. 2019a. Quantifying individual variation in reaction norms: mind the residual. – J. Evol. Biol. 33: 352–366.
Ramakers, J. J. C., Culina, A., Visser, M. E. and Gienapp, P. 2019b. Reply to: more evidence is needed to show that heritability and selection are not associated. – Nat. Ecol. Evol. 3: 1408.
Ramakers, J. J. C., Reed, T. E., Harris, M. P. and Gienapp, P. 2023. Data from: Probing variation in reaction norms in wild populations: the importance of reliable environmental proxies. – Dryad Digital Repository, https://doi.org/10.5061/dryad.zcrjdfnjr.

Reed, T. E., Wanless, S., Harris, M. P., Frederiksen, M., Kruuk, L. E. and Cunningham, E. J. 2006. Responding to environmental change: plastic responses vary little in a synchronous breeder. – Proc. R. Soc. B 273: 2713–2719.

Reed, T. E., Warzybok, P., Wilson, A. J., Bradley, R. W., Wanless, S. and Sydeman, W. J. 2009. Timing is everything: flexible phenology and shifting selection in a colonial seabird. – J. Anim. Ecol. 78: 376–387.

Rodrigues, Y. K. and Beldade,P. 2020. Thermal plasticity in insects' response to climate change and to multifactorial environments. – Front. Ecol. Evol. 8: 271. https://doi.org/10.3389/fevo.2020.00271

Ruffino, L., Salo, P., Koivisto, E., Banks, P. B. and Korpimäki, E. 2014. Reproductive responses of birds to experimental food supplementation: a meta-analysis. – Front. Zool. 11: 80.

Sparks, T. H. and Carey, P. D. 1995. The responses of species to climate over 2 centuries – an analysis of the Marsham Phenological record, 1736–1947. – J. Ecol. 83: 321–329.

Spearman, C. 1904. General intelligence, objectively determined and measured. – Am. J. Psychol. 15: 72–101.

Stedman, J. M., Hallinger, K. K., Winkler, D. W. and Vitousek, M. N. 2017. Heritable variation in circulating glucocorticoids and endocrine flexibility in a free-living songbird. – J. Evol. Biol. 30: 1724–1735.

Stillwell, R. C., Wallin, W. G., Hitchcock, L. J. and Fox, C. W. 2007. Phenotypic plasticity in a complex world: interactive effects of food and temperature on fitness components of a seed beetle. – Oecologia 153: 309–321.

Turelli, M. and Barton, N. H. 2004. Polygenic variation maintained by balancing selection: pleiotropy, sex-dependent allelic effects and G × E interactions. – Genetics 166: 1053–1079.

van de Pol, M. 2012. Quantifying individual variation in reaction norms: how study design affects the accuracy, precision and power of random regression models. – Methods Ecol. Evol. 3: 268–280.

Verhulst, S. and Nilsson, J.-Å. 2008. The timing of birds' breeding seasons: a review of experiments that manipulated timing of breeding. – Phil. Trans. R. Soc. B 363: 399–410.

Visscher, P. M. and Goddard, M. E. 2015. A general unified framework to assess the sampling variance of heritability estimates using pedigree or marker-based relationships. – Genetics 199: 223–232.

Visser, M. E., Holleman, L. J. and Caro, S. P. 2009. Temperature has a causal effect on avian timing of reproduction. – Proc. R. Soc. B 276: 2323–2331.

Westneat, D. F., Potts, L. J., Sasser, K. L. and Shaffer, J. D. 2019. Causes and consequences of phenotypic plasticity in complex environments. – Trends Ecol. Evol. 34: 555-568. https://doi.org/10.1016/j.tree.2019.02.010

Wilson, A. J., Pemberton, J. M., Pilkington, J. G., Coltman, D. W., Mifsud, D. V., Clutton-Brock, T. H. and Kruuk, L. E. 2006. Environmental coupling of selection and heritability limits evolution. – PLoS Biol. 4: e216.

Woltereck, R. 1909. Weitere experimentelle Untersuchungen über Artveränderung, speziel über das Wesen quantitativer Artunterschiede bei Daphniden. – Verh. Dtsch. Zool. Ges. 19: 110–173.

Yates, F. and Cochran, W. G. 1938. The analysis of groups of experiments. – J. Agric. Sci. 28: 556–580.

Yeh, P. J. and Price, T. D. 2004. Adaptive phenotypic plasticity and the successful colonization of a novel environment. – Am. Nat. 164: 531–542.

Appendix A. Mathematical derivation of ESM approach

Here the correlation between the true driver of plasticity (E1) and ESM is derived for the case that E1 is the main driver of plasticity. The ESM is the average phenotype in any given environment:

$$\mathrm{ESM}_j = \frac{1}{n}\sum_{i=1}^{n}\left(\mathrm{int}_i + \mathrm{slp}_i \times \mathrm{E1}_j + \varepsilon_{i,j}\right)$$

with int_i and slp_i being the intercept and the slope of the reaction norm of individual i, $\mathrm{E1}_j$ the 'real driver' in environment j, $\varepsilon_{i,j}$ a random error term, and n sample size. int_i and slp_i are the sum of the population-mean values and random deviations from them and for infinite sample sizes the averages over all individual intercepts and slopes, $\overline{\mathrm{int}_i}$ and $\overline{\mathrm{slp}_i}$, become equal to the population mean values and hence the average of the ESM in any environment E1 simply would equal the expectation

$$\mathrm{E}\left(\mathrm{ESM}_j\right) = \mathrm{int} + \mathrm{slp} \times \mathrm{E1}_j$$

with int and slp being the population mean values of the reaction norm intercept and slope. However, for limited sample sizes the individual deviations from the population reaction norm plus the additional error term imply that the average becomes

$$\mathrm{ESM}_j = \mathrm{int} + \mathrm{slp} \times \mathrm{E1}_j + u_j$$

with u_j being the 'sampling error' incorporating individual deviations from the population reaction norm and summed individual error terms. The contribution of u_j to ESM_j will tend to decrease with sample size n.

Now we can write the correlation between E1 and ESM across environments j as

$$r_{\mathrm{E1,ESM}} = \frac{\mathrm{cov}\left(\mathrm{E1}, \mathrm{int} + \mathrm{slp} \times \mathrm{E1} + u\right)}{sd\left(\mathrm{E1}\right) \times sd\left(\mathrm{int} + \mathrm{slp} \times \mathrm{E1} + u\right)}$$

which can be simplified to

$$r_{\mathrm{E1,ESM}} = \frac{\mathrm{cov}\left(\mathrm{E1}, \mathrm{slp} \times \mathrm{E1} + u\right)}{sd\left(\mathrm{E1}\right) \times sd\left(\mathrm{slp} \times \mathrm{E1} + u\right)}$$

Since $\mathrm{cov}(aX + bY, cW + dV) = ac\ \mathrm{cov}(X,W) + ad\ \mathrm{cov}(X,V) + bc\ \mathrm{cov}(Y,W) + bd\ \mathrm{cov}(Y,V)$ and $\mathrm{var} = (aX + bY) = a^2\mathrm{var}(X) + b^2\mathrm{var}(Y) + 2ab\ \mathrm{cov}(X,Y)$, the above equation can be re-written as:

$$r_{\mathrm{E1,ESM}} = \frac{\mathrm{slp} \times \mathrm{cov}\left(\mathrm{E1}, \mathrm{E1}\right) + \mathrm{cov}\left(\mathrm{E1}, u\right)}{\sqrt{\mathrm{var}\left(\mathrm{E1}\right) \times \left[\mathrm{slp}^2\mathrm{var}\left(\mathrm{E1}\right) + \mathrm{var}\left(u\right) + 2\mathrm{slp}\ \mathrm{cov}\left(\mathrm{E1}, u\right)\right]}}$$

Assuming unbiased sampling such that $\mathrm{cov}(\mathrm{E1},u)$ equals zero and using $\mathrm{cov}(\mathrm{E1},\mathrm{E1}) = \mathrm{var}(\mathrm{E1})$, this can be simplified to:

$$r_{\mathrm{E1,ESM}} = \frac{\mathrm{slp} \times \mathrm{var}\left(\mathrm{E1}\right)}{sd\left(\mathrm{E1}\right) \times \sqrt{\mathrm{slp}^2\mathrm{var}\left(\mathrm{E1}\right) + \mathrm{var}\left(u\right)}}$$

and further to:

$$r_{\mathrm{E1,ESM}} = \frac{\mathrm{slp} \times sd\left(E1\right)}{\sqrt{\mathrm{slp}^2\mathrm{var}\left(\mathrm{E1}\right) + \mathrm{var}\left(u\right)}}.$$

If we now express the ratio of var(u) and var(E1) as x, then we can re-write the above equation as:

$$r_{\mathrm{E1,ESM}} = \frac{\mathrm{slp} \times sd\left(\mathrm{E1}\right)}{\sqrt{\left(\mathrm{slp}^2 + x\right) \times \mathrm{var}\left(\mathrm{E1}\right)}} = \frac{\mathrm{slp} \times sd\left(\mathrm{E1}\right)}{\sqrt{\mathrm{slp}^2 + x} \times sd\left(\mathrm{E1}\right)}$$

which can be simplified to:

$$r_{\mathrm{E1,ESM}} = \frac{\mathrm{slp}}{\sqrt{\mathrm{slp}^2 + x}}$$

With decreasing var(u), for example due to increasing sample size n, x will approach zero and the above correlation will approach a value of one for positive values of slope slp, which means that the ESM would then outperform any other proxy.

Appendix B. Environmental proxies will under- or overestimate phenotypic shifts under environmental change

When the true environmental cue changes (e.g. through a shift to a new mean of environmental values), resultant phenotypic change will be under- or overestimated if we use an environmental proxy to predict it. This is the case whenever the slope of the cue on the proxy < 1 or > 1, respectively. In this section we briefly work out this principle using a numerical example.

Let the slope of the environmental proxy E^* against the cue E be $\beta_{E^* \sim E} = 2$ and let the slope of the phenotype y on the cue be $\beta_{y \sim E} = 4$. The slope of y on the proxy then becomes $\beta_{y \sim E^*} = \beta_{y \sim E} \times \frac{1}{\beta_{E^* \sim E}} = 4 \times \frac{1}{2} = 2$ (where $\frac{1}{\beta_{E^* \sim E}}$ equals the slope of E against E^*). Now suppose cue and proxy (both in the same units) change by 1 unit each year ($\Delta E = \Delta E^* = 1$; i.e. a directional environmental change). The predicted rate of change per year in the mean trait value (Δy_E) owing to plasticity is $\Delta y_E = \beta_{y \sim E} \times \Delta_E = 4 \times 1 = 4$; prediction based on the proxy (Δy_{E^*}) gives us $\Delta y_{E^*} = \beta_{y \sim E^*} \times \Delta E^* = 2 \times 1 = 2$.

So, the phenotypic shift is underestimated when $\frac{1}{\beta_{E^* \sim E}} < 1$ (and overestimated when $\frac{1}{\beta_{E^* \sim E}} > 1$). The only circumstance where we should find an equal shift in the phenotype is when $\Delta E^* = \Delta E / \frac{1}{\beta_{E^* \sim E}}$. In this example, we thus have $\Delta E^* = \frac{1}{1/2} = 2$ and consequently $\Delta y_{E^*} = \beta_{y \sim E^*} \times \Delta E^* = 2 \times 2 = 4$. In practice, this means that in case of an environmental shift (e.g. due to climate change), prediction of phenotypes based on the proxy estimated before the shift will almost invariably be biased.