Received: 29 February 2024 | Accepted: 13 June 2024

DOI: 10.1111/2041-210X.14379

Methods in Ecology and Evolution

REVIEW

Principles in experimental design for evaluating genomic forecasts

Katie E. Lotterhos

Northeastern University Marine Science Center, Nahant, Massachusetts, USA

Correspondence
Katie E. Lotterhos
Email: k.lotterhos@northeastern.edu

Funding information
Division of Ocean Sciences, Grant/Award Number: 2043905

Handling Editor: Itay Mayrose

Abstract

1. Over the past decade, there has been a rapid increase in the development of predictive models at the intersection of molecular ecology, genomics, and global change. The common goal of these 'genomic forecasting' models is to integrate genomic data with environmental and ecological data in a model to make quantitative predictions about the vulnerability of populations to climate change.
2. Despite rapid methodological development and the growing number of systems in which genomic forecasts are made, the forecasts themselves are rarely evaluated in a rigorous manner with ground-truth experiments. This study reviews the evaluation experiments that have been done, introduces important terminology regarding the evaluation of genomic forecasting models, and discusses important elements in the design and reporting of ground-truth experiments.
3. To date, experimental evaluations of genomic forecasts have found high variation in the accuracy of forecasts, but it is difficult to compare studies on a common ground due to different approaches and experimental designs. Additionally, some evaluations may be biased toward higher performance because training data and testing data are not independent. In addition to independence between training data and testing data, important elements in the design of an evaluation experiment include the construction and parameterization of the forecasting model, the choice of fitness proxies to measure for test data, the construction of the evaluation model, the choice of evaluation metric(s), the degree of extrapolation to novel environments or genotypes, and the sensitivity, uncertainty and reproducbility of forecasts.
4. Although genomic forecasting methods are becoming more accessible, evaluating their limitations in a particular study system requires careful planning and experimentation. Meticulously designed evaluation experiments can clarify the robustness of the forecasts for application in management. Clear reporting of basic elements of experimental design will improve the rigour of evaluations, and in turn our understanding of why models work in some cases and not others.

KEYWORDS
ecological genetics, genomic offset, novel climate, validation, vulnerability

1 | INTRODUCTION

Multivariate environmental change presents many challenges for predictive modelling. Genomic forecasting is a rapidly growing subfield that integrates genomic data with environmental and ecological data for prediction. The types of projections that genomic forecasting might be used to make range from estimating population maladaptation to an environmental change to choosing genotypes for a restoration project (Láruson et al., 2022; Rellstab et al., 2021). A *forecast* is a set of quantitative estimates that are often based on a theoretical model, statistical model or time series data and differs from a 'prediction' that can be more qualitative and based on intuition (Burford Reiskind et al., 2021). Here, the term *genomic forecasting model* refers to any kind of statistical or mathematical method that incorporates genomic data to predict fitness of a genotype in a given environment (e.g., the 'forecast').

A group of methods that are currently central to genomic forecasting models are known as genetic offset or genomic offset methods. A *genomic offset* is typically defined as the instantaneous degree of maladaptation of a genome in a new environment (Fitzpatrick & Keller, 2015; Láruson et al., 2022; Rellstab et al., 2021), although this definition has been questioned (Lotterhos, 2024). Genomic offsets are often conceptualized as the amount of genetic change that would be required for the population to be optimally adapted to a new environment (Figure 1a). They are measured as the amount of cumulative allele frequency change or turnover across an environmental gradient (Figure 1b), and calculations are often based on putatively adaptive loci identified as outliers in *genotype–environment associations (GEAs)* (Rellstab et al., 2021). GEAs are statistical methods that identify loci with associations between allele frequencies and environmental gradient(s) (Rellstab et al., 2015). Although some investigators have also referred to genomic offset as a 'genomic vulnerability', the use of the term has been debated because it is not consistent with established definitions of vulnerability (Foden et al., 2019), and simulations show that genomic offsets do not always estimate population vulnerability (Fitzpatrick et al., 2018; Láruson et al., 2022; Lind & Lotterhos, 2024; Lotterhos, 2024).

Recently, there have been a number of genomic forecasting methods published (Brauer et al., 2023; Capblancq & Forester, 2021; Fitzpatrick & Keller, 2015; Gain et al., 2023; Gain & François, 2021; Rellstab et al., 2016; Rochat & Joost, 2021) and with them the number of high-profile empirical studies that make forecasts about the maladaptation of populations across a species range has increased (Bay et al., 2018; Brauer et al., 2023; Chen et al., 2022; Exposito-Alonso et al., 2019; Fitzpatrick et al., 2021; Gain et al., 2023; Ingvarsson & Bernhardsson, 2020; Lind et al., 2024; Rhoné et al., 2020; Ruegg et al., 2018; Sang et al., 2022). However, only a few of these studies have actually evaluated those forecasts in experiments with ground-truth data (Exposito-Alonso et al., 2019; Fitzpatrick et al., 2021; Gain et al., 2023; Lind et al., 2024; Rhoné et al., 2020). The definition of *ground-truth data* is data that is known to be real or true, provided by direct and accurate measurement (in situ). Within the field of genomic forecasting, the ground-truth data should estimate the fitness of a genotype in a specific environment or a fitness offset between two environments. In practice, fitness can be estimated from the product of viability (survival) and fecundity (Conner & Hartl, 2004), but these can be difficult to measure. Instead, proxies of fitness such as growth rate or size are often used as ground-truth data (see Element 3 for discussion). For a study to qualify as a *ground-truth* or *evaluation experiment*, the genomic forecasts are evaluated against ground-truth proxies of fitness collected from multiple individuals or populations in one or more common garden environments (see Elements 1 and 2 below).

For the few studies that have evaluated genomic forecasts against ground-truth data, the degree of predictive performance has varied (Fitzpatrick et al., 2021; Gain et al., 2023; Lind et al., 2024; Rhoné et al., 2020). *Predictive performance* is generally measured as a correlation or the coefficient of determination between the forecast and a ground-truth fitness proxy (see Elements 3 and 4 below). In the tree Balsam poplar, Fitzpatrick et al. (2021) found that a genomic offset model explained ~60% of the variance in tree height in experimental common gardens, but did not assess the ability of the model to predict mortality. In a widespread conifer jack pine, Lind et al. (2024) found that offset models were worse at predicting mortality than tree height. In general, the studies that have compared multiple methods have found a wide variation in predictive performance among different algorithms (Gain et al., 2023; Lind et al., 2024). However, even different studies using the *same method* on the *same data* find different measures of predictive performance: For seed weight of pearl millet landraces in a single common garden, the predictive performance of a genetic offset calculated using the same algorithm (called gradient forests) varied from an R^2 of ~17% in one study (Rhoné et al., 2020, estimated as the square of Pearson's correlation) to ~49% in another study (Gain et al., 2023). Such variation highlights that there are nuances in how investigators build a forecast from the same method (e.g., the selection and spatial resolution of environmental variables, individuals and single nucleotide polymorphisms—SNPs). Indeed, for tree height in jack pine and varieties of Douglas fir, Lind et al. (2024) showed that performance of offset forecasts were highly sensitive to the set of individuals used in training, varying from no predictive performance (measured as the correlation between a forecast and tree height) to very high predictive performance.

Although this variation in performance of forecasts among study systems may be due to differences in the underlying biology of the system, it could also be due to variation in the nuanced decisions investigators make when designing experiments and building, training, and testing models. If such variation in model performance is due to nuanced decisions, this creates substantial hurdles to the field in terms of understanding the utility of genomic forecasting. For this reason, explicit reporting of key design elements would advance our understanding of why performance is high in some scenarios and low in others, but these are rarely reported.

The goal of this review is to enumerate the elements of experimental design and reporting that are necessary for the rigorous evaluation and interpretation of genomic forecasting models. These

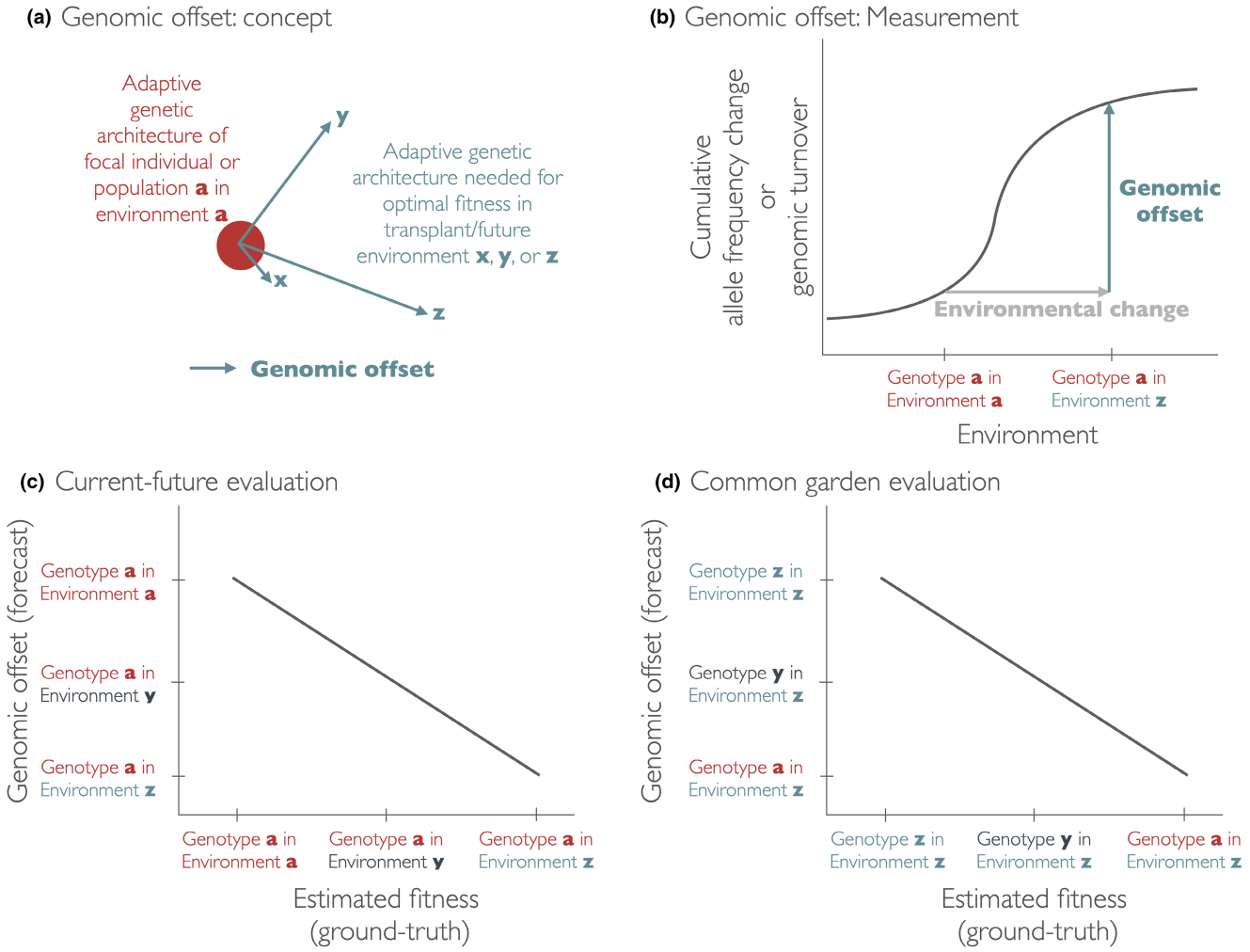

FIGURE 1 Concept, measurement and evaluation of genomic offsets. (a) A genomic offset is conceptualized as the amount of change in the adaptive genomic architecture needed for the individual or population to be optimally adapted to a new environment (visualized in multivariate genomic space). (b) For a given amount of environmental change, a genomic offset is typically measured as a cumulative change in allele frequency or turnover in allele frequency, amalgamated across loci. A genomic offset is inferred to be directly related to the degree of maladaptation to that environmental change, although this inference has been debated. (c) A *current-future evaluation* tests whether the genomic offsets are correlated with the estimated fitness of a particular genotype in different environments. (d) A *common garden evaluation* tests whether genomic offsets are correlated with the estimated fitness of different genotypes grown in the same common garden environment. See Lotterhos (2024) for more discussion on why the two evaluations in (c) and (d) are not equivalent.

elements include design principles such as independence of training and testing data, and reporting the relatedness among individuals in these two groups. These elements are germane to both evaluation and validation of methods. *Evaluation* is a quantitative comparison of how well one or more methods perform on one or more data sets (Lotterhos et al., 2022). If one quantifies the predictive performance of one or many methods, then it is an evaluation. However, even the 'best' method in an evaluation may have low predictive performance and be unsuitable for any kind of real-world application.

For any real-world application, a method must first be validated. Although the term *validation* is not used consistently (Augusiak et al., 2014), here the term refers to the process of comparing the model forecasts to a body of evidence obtained from many evaluations, and then determining if the model meets a set of a priori criteria for a specific real-world application (Lotterhos et al., 2022). For forecasting models, validation would be the process of consulting with stakeholders to develop consensus on the criteria a model should meet for incorporation into a management plan, and then determining if the model meets those criteria. Concluding that a forecast is 'valid' is not necessarily a binary outcome; models lie on a continuum of usefulness in which the overall credibility of a forecast is gradually built upon (Augusiak et al., 2014).

Regardless of whether the goal is evaluation or validation, the outcome can inform the domain of applicability of a forecasting model. The *domain of applicability* of a model describes the set of conditions under which the model predictions are valid (Lotterhos et al., 2022). The following sections outline considerations for the most constructive use, publication and archiving of genomic forecasting models. Careful design and reporting will lead naturally to meta-analyses and syntheses that can help further define the

domain of applicability of these methods across systems. To date, most genomic forecasting methods models make predictions at the level of the population. Although individual-level predictive models have not yet been widely developed or tested, this article takes a prospective view with an eye towards designing experiments that would be able to evaluate individual-level predictive models when they become available.

2 | COMMON ASSUMPTIONS OF GENOMIC FORECASTING MODELS

All forecasting models make assumptions either directly or implicitly. An implicit assumption is one that is not always stated, but is a basic requirement for the model to produce accurate results. Not all genomic forecasting models share the same assumptions, but common assumptions include:

1. *The population is locally adapted to the environment and genotype–climate relationships reflect local adaptation to the environment*. *Local adaptation* is a property of a metapopulation in which subpopulations have higher fitness in sympatry than in allopatry (Blanquart et al., 2013). This can lead to a pattern in which locally adapted alleles have higher frequency in sympatry than in allopatry. Motivated by this, current genomic offset models are trained on contemporary spatial relationships between genotype and climate, and thus rely on the assumptions that (i) populations are locally adapted and (ii) the allele frequency at a particular environment on the landscape provides some information about the local fitness of that allele. Because the assumption of local adaptation is integral, investigators should establish proof of local adaptation in the metapopulation with common garden and/or reciprocal transplant experiments (for a quantitative method see Blanquart et al., 2013) prior to investing resources into building a forecasting model, as has been recommended to do prior to conducting genotype–environment and genotype–phenotype associations (Barrett & Hoekstra, 2011). Even if local adaptation exists, non-monotonic allele frequency patterns can evolve under evolution to multivariate environments (Lotterhos, 2023) and these patterns may confound forecasting methods that assume linear genotype–climate relationships.
2. *The environmental change is instantaneous*. Current genomic offset models predict the fitness of a genotype under an instantaneous environmental change (Fitzpatrick & Keller, 2015; Láruson et al., 2022). Due to this assumption that is specific to offset models, the forecasts can be evaluated by testing if they are significantly correlated with ground-truth fitness proxies measured on a set of individuals after moving them to a new environment (see Elements 2 and 3 for discussion). Current offset models do not incorporate evolutionary processes that happen over multiple generations, such as gene flow, inheritance, drift, recombination, mutation, and adaptive evolution. However, these processes may be included into eco-evolutionary models, which can also be used to make forecasts (Waldvogel et al., 2020).
3. *A model built on current genetic and environmental variables can be extrapolated without reduction in predictive performance*. Current genomic forecasting models assume that the genotype–climate relationship will remain the same under future environmental change. *Extrapolation* is the action of estimating a variable's value beyond the range of initial values by using relationships between this and other variable(s). For instance, projecting a genomic forecast trained on a dataset of genetic and environmental data to a novel genotype or novel environment (novel relative to the training dataset) is an extrapolation (see Element 6 for more discussion). Such extrapolations can fail if they do not account for evolutionary processes over multiple generations or the fixation of climate-adapted alleles (Hoffmann et al., 2021; Jordan et al., 2017).

   Even when a forecast is used to choose genotypes for immediate transplantation as in restoration (e.g., the assumption of instantaneous environmental change is met), extrapolations to novel environments could fail because of statistical reasons (e.g., the model was biased by historical evolutionary process such as drift: Láruson et al., 2022; or the model is not accurate in the extrapolated region: Lind & Lotterhos, 2024) or intrinsic biological reasons (e.g., cryptic genetic variation that is only expressed in a new environment: Bitter et al., 2021). Recently, DeSaix et al. (2022) showed that large portions of the breeding range of an alpine songbird will shift to novel climatic conditions, highlighting that substantial portions of a species' range can be subject to uncertainty in genomic forecasts. This could be an issue for genomic forecasts in many species, since novel climates without precedent in recent history are emerging around terrestrial and marine habitats and will become more widespread in the future (Lotterhos et al., 2021; Williams et al., 2007). Thus, estimating the degree that a model is extrapolated to novel climates or genotypes is an important element of evaluation (see Elements 5 and 6 below for discussion).
4. *Accurate forecasts can be made without considering genetic interactions, trait correlations or trait plasticity*. Currently, the complex dynamics of within- and among-locus interactions, pleiotropy and plasticity are not explicitly incorporated into genomic offset models. Fitness may be affected by non-additive allelic interactions within loci (dominance) and among loci (epistasis) (Conner & Hartl, 2004). Pleiotropic effects of a gene on multiple traits may facilitate or constrain adaptation and therefore complicate forecasts (Hoffmann et al., 2021). Plasticity is the environmentally induced production of different phenotypes from a given genotype (DeWitt & Scheiner, 2004), which suggests that trait data in addition to environmental and genetic data will be needed to produce accurate forecasts. These dynamics are not explicitly incorporated into current genomic forecasting models, but the elements of experimental design

discussed below will also be germane to more complex models that include these dynamics.

3 | ELEMENTS OF EXPERIMENTAL DESIGN AND REPORTING

How robust and accurate genomic forecasts models are—even when implicit assumptions are met—is an open question for many study systems. The most rigorous way to test model forecasts is with controlled experiments and compare a forecast against ground-truth data from the experiment. There are many types of experiments that one could design, including common garden experiments, growth chambers or microcosms, experimental evolution or within-generation selection experiments. Here, the focus is largely on common garden experiments because they will be the most feasible for many study systems (de Villemereuil et al., 2016; Sork et al., 2013), but many of the principles will apply to other kinds of experiments.

3.1 | Element 1: The type of evaluation to do

The appropriate experimental design used to evaluate a model depends on the management context or application (Augusiak et al., 2014). The question remains as to how well a genomic offset measurement estimates some kind of fitness offset for different applications. Unfortunately, there are many different ways that fitness offsets can be calculated, and depending on the pattern of local adaptation in the metapopulation, different types of fitness offsets may not be correlated with each other (Lotterhos, 2024). For this reason, the experimental design that should be used to evaluate a forecast of climate change vulnerability is not the same design that should be used to evaluate a forecast for a restoration project.

In a *current-future evaluation*, one aims to understand whether the forecast can predict population vulnerability in a future climate. The appropriate experimental design would be to raise the focal genotype in current and future environments and use an estimate of fitness (or fitness offset from the current to the future environment) in each treatment as a ground-truth metric (Figure 1c, *x*-axis) (Lotterhos, 2024). The relevant forecast to evaluate would estimate the focal genotype's fitness in the different environmental treatments (Figure 1c, *y*-axis). For a genomic offset, genotypes with larger offsets to a new environment are predicted to have lower fitness in that environment, so a strong negative correlation indicates high predictive performance of the forecast.

In a *restoration* or *common garden evaluation*, one aims to understand whether the forecast can predict the most fit genotype(s) at a restoration site, so that the best genotype(s) can be chosen for that environment and give the project the highest chances of success. The appropriate experimental design would be to raise multiple genotypes in a common garden at the restoration site and use an estimate of fitness (or fitness offset) of each genotype in that common garden as a ground-truth metric (Figure 1d, *x*-axis) (Lotterhos, 2024). The relevant forecast to evaluate would estimate the (relative) fitnesses of different genotypes in the restoration environment (Figure 1d, *y*-axis).

To date, all evaluations of genomic offsets have been common garden evaluations (Fitzpatrick et al., 2021; Gain et al., 2023; Lind et al., 2024; Rhoné et al., 2020), so it is still unclear how well genomic offsets predict population vulnerability to future climate change. Evaluation experiments that have multiple genotypes in many current and future environmental treatments can simultaneously conduct *current-future* and *common garden evaluations* (for an extensive discussion on this topic see Lotterhos, 2024).

3.2 | Element 2: Independence between and sample sizes of training and ground-truth data

Current genomic forecasting models require data to build the model that makes the forecast. This model-building step uses 'training data' to calibrate the model. After the model is developed, predictions are made for a set of 'ground-truth data' or 'test data', which are then used to evaluate the predictive performance of the model. If the same or overlapping data are used for both training and ground-truthing, this can lead to inflated model performance and poor generalizability (i.e., overfitting) because one is essentially asking how good the data is at predicting itself (Goodall, 1972; Grimm et al., 2015; Molnar, 2021). Therefore, such circularity should be avoided by using independent data for both model training and ground-truthing.

An example of independent training data and ground-truth data is illustrated in Figure 2 for a hypothetical coastal marine species that is locally adapted to temperature and salinity. In this case, seascape genomic data with individual SNP genotypes and population-level environmental variables (temperature and salinity) are used to train a forecasting model (Figure 2, *Step 1*, orange colour). In the independent ground-truthing phase, different individuals are collected from across the metapopulation and bred to create offspring for the test (Figure 2, *Step 2*, black colour). In the example shown in Figure 2, some test individuals come from populations that are used in training the forecasting model and some come from populations that were not used in training, which can give insight into the domain of applicability of the predictive model.

Offspring are then raised in 'test' common garden environments at high or low salinity (Figure 2, *Step 3*, green colour). Ideally, there would be multiple common garden sites, although this might not be feasible in all systems for logistical (e.g., broadcast-spawning marine invertebrates that require a hatchery for rearing) or permitting reasons (e.g., permits do not allow transfer of foreign genetic material to the local site). The genomic offset is calculated from the forecasting model using the test environment and, depending on the specific forecasting method being used, the genotypes of the test individuals (Figure 2, *green arrow from Step 3 to Step 4*) (see

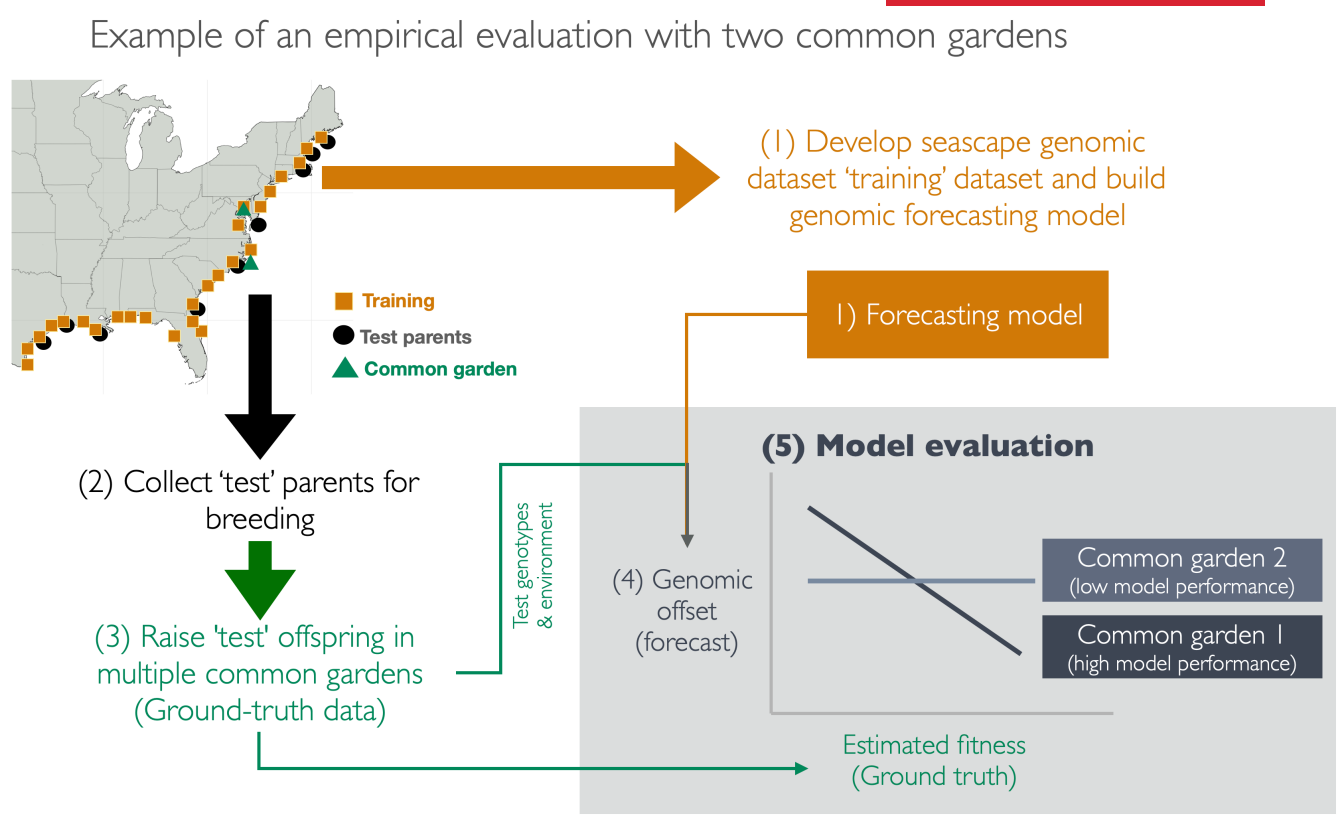

FIGURE 2 Overview of an empirical evaluation with two common gardens. An overview of the model evaluation process with independent training and test phases, using the hypothetical example of a widespread coastal marine invertebrate. *Step 1*: Train a forecasting model on existing data, in this example from seascape genomic data with salinity, pH, dissolved oxygen and temperature (orange squares and text). *Step 2*: Collect test parents that will be used to create test offspring (black circles). *Step 3*: Raise test offspring in the experimental common gardens where the model forecast will be tested (green text). In this example, one common garden is at a low salinity site, while the other is at a high salinity site (green triangles). *Step 4*: The common garden environment and test genotypes are input into the model to make a forecast. *Step 5 (grey box)*: The forecast is evaluated against an estimate of fitness in the common garden. The predictive power of the forecasting model may vary among common gardens (as shown here), which would inform the domain of applicability for the model.

Element 6 for more discussion on quantifying the environment). The predictive performance of the forecasting model is estimated as the correlation between the forecast (*y*-axis) and a ground-truth estimate of fitness measured in each common garden (*x*-axis) (Figure 2, *Step 5*) (see Element 4 for more discussion on estimating predictive performance). The variation in predictive performance among many different common gardens gives insight into how generalizable the forecasting model will be. In the hypothetical example in Figure 2 (*Step 5*), the model has high predictive performance in Common Garden 1 (as evidenced by the strong negative correlation between predicted offset and fitness), but low predictive performance in Common Garden 2, which suggests that the model is not generalizable.

The feasibility of creating offspring from parents for an evaluation experiment will vary among study systems. It can be advantageous to split test offspring from each family among common garden test sites and measure fitness proxies on both test parents and test offspring (see Element 3 for more discussion). If such a design is not feasible and test individuals are relocated directly from their collection site to an experimental common garden, the experiment may be confounded by direct carry-over effects from the organism's environment to the common garden. On the other hand, this may be exactly how one wants to conduct a test if such relocations are what would be practised for management, as in corals. Regardless of how the test is conducted, many of the following recommended design elements arise from this basic principle of independence between training and ground-truthing.

3.2.1 | How large do sample sizes need to be?

It is well established in the machine learning literature that the predictive capacity of a model is limited by the data on which it is based (Bergstrom & West, 2021; Molnar, 2021). Thus, models that are trained on data that is a limited subset of the situations for which they will be applied often lead to inaccurate results when extrapolated to new situations. For example, if a forecasting model uses information about allele frequency correlations with environments, then a comprehensive training data set would include a large number of subpopulations across the environmental range of the species, with enough samples from each site to accurately estimate allele frequency in that environment. Understanding how a forecasting model

is built is fundamental to planning robust sample sizes for training and testing. The simulation studies (Láruson et al., 2022; Lind & Lotterhos, 2024) and empirical validations (Fitzpatrick et al., 2021; Lind et al., 2024) that find high predictive performance of genomic offsets suggest that much larger sample sizes are needed for training (hundreds of individuals from dozens of populations) than for testing (dozens of individuals from tens of populations). This is because a comprehensive set of data is needed for model training, whereas a smaller dataset spanning a range of scenarios can be sufficient to calculate predictive performance. At this time, the effect of sample size on the performance of forecasts has not been sufficiently explored in the literature and this is an important area for future work.

3.3 | Element 3: The ground-truth metric(s)

To evaluate the predictive performance of a model, the forecast must be compared with a ground-truth data set that estimates fitness (Figure 2, *Step 5*, *x*-axis). There could be multiple ways to estimate fitness based on traits, fitness components or organism performance. While fitness components include viability (survival) and fecundity, these can be difficult to measure, and growth rate or size are often used as proxies for fitness instead. For the design shown in Figure 2, ground-truth measures could be calculated from (i) the fitness proxy of the parental genotypes measured as the viability of their offspring, or (ii) the fitness proxies of the offspring genotypes measured as their survival, growth or fecundity. Both of these measures have strengths and drawbacks.

3.3.1 | Fitness proxies of parents

The fitness of parental genotypes can be estimated as the viability of their descendants, which can be estimated at a handful of genetic markers with a parentage or lineage analysis. The benefit of using a parental fitness proxy as a ground-truth is that genotypes of parents can be obtained before or immediately after offspring are produced. Thus the parental fitness metric will reflect any early-stage mortality that occurs due to the treatment before offspring are large enough to count, genotype or phenotype accurately. Depending on the study species, it may be hard to control for the exact same number of offspring from each parent, and in this case the proportion of offspring from each parent should be quantified at the beginning of the experiment and used as a baseline.

3.3.2 | Fitness proxies of offspring

Measuring fitness proxies directly on offspring themselves, such as their growth and survival, also has strengths and drawbacks. When families can be tracked in the design, best linear unbiased predictors or estimates (BLUPs or BLUEs) may be used to estimate a breeding value for a fitness proxy such as height and used as a ground-truth in the evaluation (e.g., Fitzpatrick et al., 2021; Lachmuth et al., 2023; Lind et al., 2024). The advantage of using offspring fitness as a ground-truth is that a forecast based on each individual's genotype can be tested within that individual's lifetime. Drawbacks arise from the timing of genotyping relative to when mortality arises from selection, because one has to wait until the offspring are large enough to collect tissue without causing mortality. If offspring are put into the common garden test environments before they are genotyped (e.g., as seeds), this has the drawback of potentially excluding offspring samples that die due to selection (e.g., never sprout or die shortly after sprouting) from the ground-truth data, and their absence could bias the overall evaluation.

3.3.3 | Direct transplantation

Another alternative would be to transplant genotypes directly from their home site to the common garden test environment, which might be necessary for some organisms that are difficult to breed. These approaches have the potential drawbacks of (i) being biased by direct carry-over effects from one environment to another, and/or (ii) missing selection that would have happened during the early part of the life cycle in the test environment. Despite these drawbacks, studies that choose these alternatives for practical or restoration reasons can still give relevant insights.

3.3.4 | Should population size be used as a ground-truth metric?

While measuring fitness proxies on both test parents and their offspring in a controlled experiment meets the highest level of rigour for a ground-truth measure, some studies have interpreted relationships between genomic offsets and population size as evidence that forecasts are accurate (Bay et al., 2018; Ruegg et al., 2018). A computational study found that genetic offset values were correlated with population size in purely neutral simulations due to genetic drift affecting the offset values, which highlights issues with using population size as a ground-truth metric (Láruson et al., 2022).

3.4 | Element 4: The evaluation metric(s)

An important question in the evaluation (Figure 2, *Step 5*) is how to quantify the predictive performance of the forecasting model. The *evaluation metric* is a statistic that summarizes the model performance from the *evaluation model*—a statistical model of the relationship between the forecast and the ground-truth data. In the case that the forecast and the ground-truth data are both numerical variables in different units, the model performance can be quantified as the strength of the relationship between these two variables, which can be measured as a correlation or as R^2 in a general/generalized linear model (Lotterhos et al., 2022). Note that other evaluation metrics

such as mean error and its derivatives are only appropriate when the forecast and the ground truth are in the same units (Lotterhos et al., 2022). To date, both correlation (Láruson et al., 2022; Lind et al., 2024; Rhoné et al., 2020) and R^2 (Fitzpatrick et al., 2021; Gain et al., 2023) have been used as evaluation metrics to quantify the performance of a forecast. Although this section focuses on evaluation metrics, the process of evaluation should also consider logical consistency in the forecasting model structure (e.g., underlying theories and assumptions), as major flaws in the forecasting model structure could mislead a decision even when evaluation metrics suggest that predictive performance is high (Augusiak et al., 2014).

3.4.1 | Correlation as an evaluation metric

When correlation is used as an evaluation metric, it measures the strength and significance of the association between ground-truth fitness and the forecast. When the assumption of linearity is met, Pearson's correlation should be used, while if the relationship is monotonic but non-linear then a rank correlation should be used (Whitlock & Schluter, 2009). The benefit of correlation measures is that they make few assumptions and can be easily compared among studies.

3.4.2 | R^2 as an evaluation metric

When R^2 from some type of general/generalized linear model or linear mixed model is used as an evaluation metric, different philosophical approaches to constructing linear models makes it difficult to compare among studies. The question for evaluating forecasts is: Should the ground-truth be the response variable (e.g., the forecast predicts the ground-truth fitness) or explanatory variable (e.g., the ground-truth fitness predicts the forecast)? Logical arguments can be constructed for both approaches.

Arguments for constructing an evaluation model with ground-truth fitness as an explanatory variable and the model forecast as a response variable come from method evaluations in data science (Lotterhos et al., 2022). Evaluations are constructed in this way because it is the amount of error in the forecast (response variable) that gives information about how good the forecast is. This type of model construction is shown in Figures 1c,d and 2 (*Steps 4–5*) with the ground-truth on the *x*-axis and the forecast on the *y*-axis. Most types of linear models assume that the explanatory variable is known without error and models error in the response variable. However, ground-truth estimates of fitness are imperfect and contain error, which violates the assumptions of ordinary least squares regression. When error exists in an explanatory variable, ordinary least squares regression will underestimate the true regression slope (McArdle, 1988). In this case, one can construct an evaluation model with reduced major axis regression, which will yield a more accurate slope (McArdle, 1988).

The argument for constructing an evaluation model with ground-truth fitness as a response variable and the model forecast as an explanatory variable is that a number of other explanatory variables could be put into the model to determine how much variance in estimated fitness is explained by different factors. For a genomic offset evaluation constructed in this way, the slope informs how much fitness decline is expected for a given amount of genomic offset. However, when there are multiple explanatory variables, the overall model R^2 is not a measure of predictive performance of the forecast or genomic offset: R^2 is based on unexplained variation after accounting for all explanatory variables (Whitlock & Schluter, 2009) and therefore is not a measure of the performance of the genomic forecast specifically. Moreover, when models are constructed in this manner, investigators should take care to show that they avoided overfitting (e.g., inflating R^2 by including many explanatory variables) by performing model selection to determine the most parsimonious model that explains the data (Whitlock & Schluter, 2009).

Thus, the interpretation of the evaluation model depends on its formulation. When the model contains only two variables, a model with *forecast~ground-truth* will have the same R^2 or correlation as a model with *ground-truth~forecast*, and the R^2 or correlation value will give information about the predictive performance. However, the R^2 from a model with *ground-truth~forecast+other variables* does not give information about the predictive performance of the forecast specifically, because other variables could be explaining most of the variation. However, this latter type of model could still be useful to understand how various factors are related to the ground-truth fitness proxy via their slopes.

Thus, the overall R^2 as an evaluation metric should be interpreted carefully based on the formulation of the evaluation model. Regardless of how the evaluation model is formulated, it will be informative to compare the shape and magnitude of the forecast/ground-truth relationship across multiple studies, treatments and/or common garden environments (e.g., Figure 2, *Step 5*), because that gives information about how generalizable the forecast is.

3.5 | Element 5: Relationships among training samples and test samples

To keep the training and test phases independent, a necessary design element is that the set of individuals used in training is different from the set of individuals used in ground-truthing. A central question in determining the domain of applicability is whether model performance degrades as specific test samples (individuals or populations) become more genetically distant from those used in training (e.g., extrapolation of the forecasting model to samples that are 'novel' to that model). Design considerations that should be clearly reported include the number of samples used in training, the number of test samples and populations, whether test samples come from populations used in training, and the degree of independence between test and training samples. Individuals within a species will never truly be independent due to shared evolutionary history. Sources of non-independence in

the data arise among genes within genomes, among individuals within a population due to relatedness, and among populations due to migration and/or spatially autocorrelated selection. The degree of (non)independence among test and training samples can be estimated as values of relatedness, F_{ST}, or other measures of genetic distance.

Take for instance a test that was conducted on a mix of genotypes, some that are genetically distant ('novel') compared with the training genotypes (Figure 3a). If the fitness of novel genotypes are not predicted well by the model, then removing them from the evaluation would increase the predictive performance of the model (Figure 3a). This concept could be tested more formally by using a jackknife in the evaluation to understand the sensitivity of the evaluation to specific genotypes. A jackknife removes one genotype at a time and re-evaluates the relationship between the forecast and ground-truth data. One would predict that the decline in predictive performance when a test sample is removed scales directly with the genetic distance between that test population (or genotype) and all the training genotypes, and the rate of this decline would give insight into the domain of applicability of a particular forecasting model. This type of analysis remains to be performed.

A major design consideration is the set of genetic markers that will be used to genotype all the individuals. For some forecasting models, the same set of genetic markers must be used across all training and test individuals because the model uses this information to make a prediction (Figure 4a). Unless the genetic basis of adaptation to the multivariate environment is accurately known, researchers should make the case that their marker set provides high enough coverage of the genome so as to have markers in linkage disequilibrium with adaptive loci that determine local adaptation (Lowry et al., 2017). Without such justification, it is difficult to determine whether a poorly performing forecasting model is due to insufficient coverage of the genome or some other reason.

There is currently no consensus in the literature as to the best marker set to use for building a forecasting model. Although it is common practice to narrow the loci set down to outliers in GEAs, both simulation and empirical studies that have explored the sensitivity of forecasts to the set of loci used have found that a random set of loci has similar performance to GEA outliers (Fitzpatrick et al., 2021; Lachmuth et al., 2023; Láruson et al., 2022; Lind et al., 2024). This may occur because evolution in multivariate environments can lead to non-monotonic patterns in allele frequencies across environmental gradients that would be missed by GEA methods (Lotterhos, 2023), and offset models may be driven instead by genome-wide patterns of isolation by environment. Thus, it is important for investigators to report how sensitive a forecast is to the set of loci used (see Element 8 for more discussion).

3.6 | Element 6: The training and test environments: Quantification, novelty, uncertainty and variability

In order to obtain a genomic forecast for the test individuals, most models require the test environment as input (Figure 2, *Step 3*). Thus, the multivariate environment measured for the test must be measured in the same way as the multivariate environment was measured across populations or individuals used in training the model (Figure 4b). The test environment should only include data collected during the period over which individuals are reared/grown, in case there are any extreme climatic events that affect mortality. The

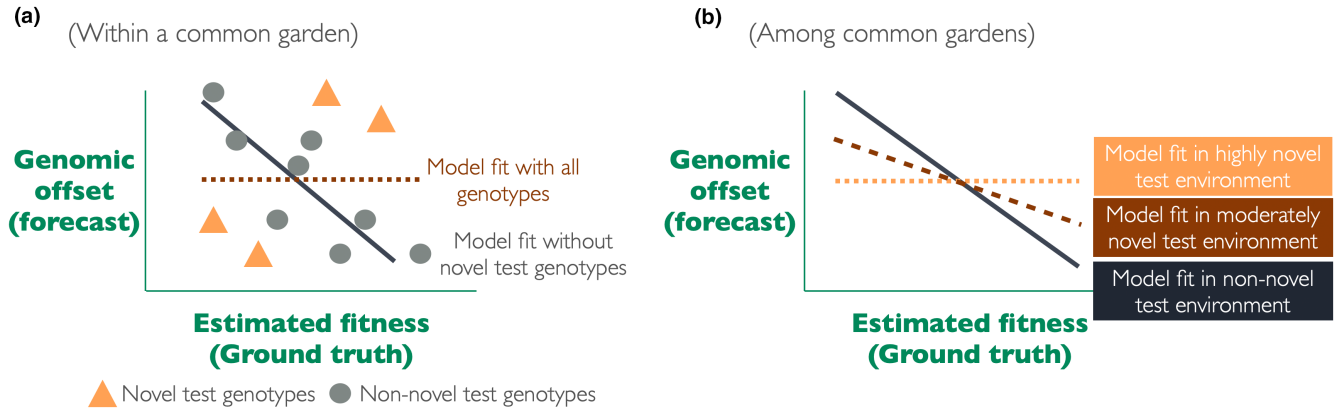

FIGURE 3 Change in model performance with extrapolation. Hypothetical examples of how the accuracy of the forecasting model might change as the model is extrapolated to new situations compared to the training data, such as novel genotypes or novel environments. (a) The relationship between the estimate of fitness (ground truth) and a forecast within a single test common garden may be sensitive to the novelty of the test genotypes relative to the training genotypes. Here, a novel test genotype is one that is genetically distant from the training genotypes. The performance of the model, measured as the correlation between ground truth and forecast, may improve when novel test genotypes are removed from the evaluation. A jackknife could be used to understand how particular test genotypes influence the model performance. (b) The relationship between the ground truth and a forecast among test common gardens may be sensitive to the novelty of the test environment (where the ground-truth data were collected) relative to the training environments. The degree of climate novelty of a particular test environment can be quantified as described in the main text. Model accuracy may decrease as test environments become more novel compared with the training environments, as shown here.

model's predictive performance may decline in test environments that are outside the range of environments from which the populations used in training were collected. Therefore, investigators should quantify how novel the test environment is compared with the environments used in training the forecasting model, which will help define the domain of applicability of the model.

A *novel climate* is defined as a climate observation without a recent analogue in a local or global area, and the degree of novelty is typically measured as the number of standard deviations of a climate observation from a historical baseline (Mahony et al., 2017; Williams et al., 2007). Novel climates for a species can also be estimated by the niche margin index, with decreasing negative values representing climatic distance outside the niche (Broennimann et al., 2021; DeSaix et al., 2022). We can extend these definitions of climate novelty to model ground-truthing, where the test/ground-truth environment does not have an analogue in the training environment. With the domain of applicability of a forecasting model in mind, there are various kinds of situations that would be interesting to examine, for example, ground-truth environments that are: (i) typical of the environments used in training (Figure 5, *circle*), (ii) extreme but within the

FIGURE 4 Best practices for data collection. Examples of best practices for data collection (top) and potentially not useful data (bottom) for (a) single nucleotide polymorphisms (SNPs) (left column) and (b) environmental data (right column). The 'training' data are used to parameterize the forecasting model and the 'test' data are used as a ground-truth for the evaluation of the forecast. Investigators can easily end up with potentially not useful data if the landscape genomic data used for model training used one type of sequencing (e.g., RADseq) and available environmental data, while the experiment for model testing used a different type of sequencing (e.g., RNAseq) and did not quantify the multivariate test environment in the exact same way.

(a)

Best practice

	SNP 1	SNP 2	SNP 3	SNP 4	SNP 5
Training-1	A	G	T	A	G
Training-2	C	G	T	A	C
Test-1	T	A	T	A	C
Test-2	C	A	C	T	G

(b)

	Env 1	Env 2	Env 3	Env 4	Env 5
Training-1	19	12.1	1	6.5	33
Training-2	20	15.4	2	6.8	21
Test-1	21	9.2	3	5.3	24
Test-2	21	9.2	3	5.3	24

Potentially not useful

	SNP 1	SNP 2	SNP 3	SNP 4	SNP 5
Training-1				A	G
Training-2				A	C
Test-1	T	A	T		
Test-2	C	A	C		

	Env 1	Env 2	Env 3	Env 4	Env 5
Training-1	19	12.1	1		33
Training-2	20	15.4	2		21
Test-1	21		3	5.3	24
Test-2	21		3	5.3	24

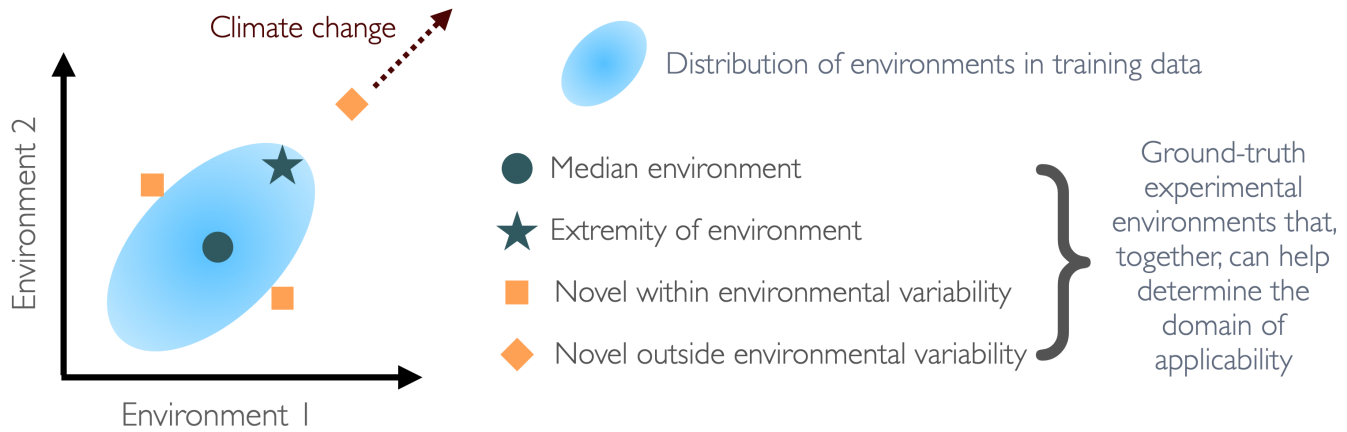

FIGURE 5 Selection of experimental common gardens that can help evaluate the domain of applicability of a forecasting model. The blue ellipse on the graph shows the distribution of values for the two environmental variables in the training data, and the dark red arrow shows the direction of climate change. Testing the model in treatments that are (i) within the environmental envelope of the training data (circle and star), (ii) novel to the training data but within the historical variability of each variable (squares) and (iii) novel to the training data and outside the historical variability of multiple variables (diamond) can inform situations in which the model yields (or does not yield) accurate results. Note that the most informative treatments for understanding the domain of applicability to future climate change are not necessarily factorial combinations of individual environmental values.

range of environmental combinations used in training (Figure 5, *star*), (iii) a novel combination of environmental variables, but within the historical variability of the training environments (Figure 5, *squares*), and/or (iv) a novel combination of environmental variables and outside the historical variability of the training environments (Figure 5, *diamond*). The degree of climate novelty (between a test environment and all the training environments) for any situation can be statistically estimated as the degree of novelty or niche margin index (e.g., DeSaix et al., 2022; Lotterhos et al., 2021; Mahony et al., 2017).

With enough common gardens, one could then test how model performance degrades as the novelty of the test environment relative to the training environment increases (Figure 3b), which has particular relevance for assisted gene flow (inside current species range) and assisted migration (outside current species range). Analysis of thousands of simulated data sets spanning a range of demographies has demonstrated that the predictive performance of genomic offsets declines with climate novelty, and it declines more rapidly for strongly locally adapted species (Lind & Lotterhos, 2024). Similar results have been observed in genomic selection models, where the predictive performance declines in ground-truth environments were outside the range of training data (Rogers & Holland, 2022). Note that the set of ground-truth environments that may be the most relevant for determining the domain of applicability is not necessarily factorial combinations of individual environmental values (Figure 5).

An additional design consideration is which environmental variables to use in training and ground-truthing a forecast. Ideally, all abiotic and biotic environmental variables that affect fitness are included in the model and the grain size is relevant to biology of the system (Dauphin et al., 2023). These goals are logistically difficult, especially given the multitude of diseases and competitive interactions that affect natural populations. Some methods, such as redundancy analysis or multiple regression, assume that the environmental predictor variables used in training are uncorrelated (Legendre & Legendre, 1998), and so some environmental variables may need to be excluded for statistical reasons. However, not all environmental measures will be important in determining fitness and could mis-parameterize offset models if all possible environmental variables are included in training (Lind & Lotterhos, 2024). In addition, there is considerable variability among different climate projections (IPCC, 2023), which is an additional source of uncertainty in forecasts. Therefore, investigators should report the sensitivity of the forecast to the set of environmental variables chosen for model-building, as well as the range of values for each variable (see Element 8 for more discussion).

Another open question with regard to building forecasting models is how to incorporate environmental variability. For example, a forecasting model could be trained on some combination of the mean temperature (which might reflect long-term thermal tolerance), the maximum temperature (which might reflect selective events due to heat stress), minimum temperature (which might reflect selective events due to cold stress) or features of temperature variability such as standard deviation and predictability (e.g., temporal autocorrelation) that might reflect the degree of phenotypic plasticity (Bitter et al., 2021). More research is currently needed to understand the best ways to incorporate environmental variability into genomic forecasting models and assess model sensitivity to the way that environmental variability is quantified.

3.7 | Element 7: Model training

For some types of forecasting models, users may have to choose hyperparameters for the algorithm. *Hyperparameters* are values that are given before the model is trained and specify the details of the algorithm (e.g., tree depth within random forests or the number of neurons in the *i*'th hidden layer of neural nets). Other methods do not require the user to choose hyperparameters, because the parameters are estimated by minimizing residual error (e.g., general linear models and redundancy analysis). This section is relevant to building models that require the user to explore a potentially large hyperparameter space available for use in model training. In this case, training data and ground-truth data are still independent datasets, but training data are further subdivided into data used for model parameterization and *holdout* data used to assess model fit for hyperparameter tuning.

Cross-validation is the process of evaluating model fit on different subsets of a single data set, and this process can be leveraged for hyperparameter tuning. In cross-validation, the training data is split into multiple subsets or *folds*, and in each fold a different subset of the training data is used as holdout data on which the model fit is assessed (Figure 6). Cross-validation is a specific technique used to summarize model generalizability, and should not be confused with model validation performed on independent ground-truth data, because the latter is the process of determining whether a model is 'good enough' for application in the real world (Lotterhos et al., 2018, 2022). The result of cross-validation is a distribution of model accuracies when different subsets of the training data are used for training and assessing model fit. A model that generalizes well to new situations will have both high mean predictive performance with minimal variation across folds.

One critical question in cross-validation is how to divide the training data into folds, as there are multiple ways to do this. The 'leave-one-out' strategy leaves one sample out in each fold, and has as many folds as there are observations in the dataset. This is computationally intensive for large datasets, so it is rarely performed. It is more common for the data to be divided into *k* folds (where the choice of *k* is informed by the sampling design), and this can be performed in an unstratified or stratified manner. An *unstratified k-fold cross-validation* randomly chooses samples from the entire dataset to place into folds. If there are unequal sample sizes among genetic clusters, the unstratified design will result in some populations being over- or under-represented in a particular fold, and as a result it may be difficult to interpret what drives the model accuracy in each fold. A *stratified k-fold cross-validation* populates the holdout data so that each stratum (i.e., some shared characteristic among samples, such as population ID or geographic region) is roughly equally represented

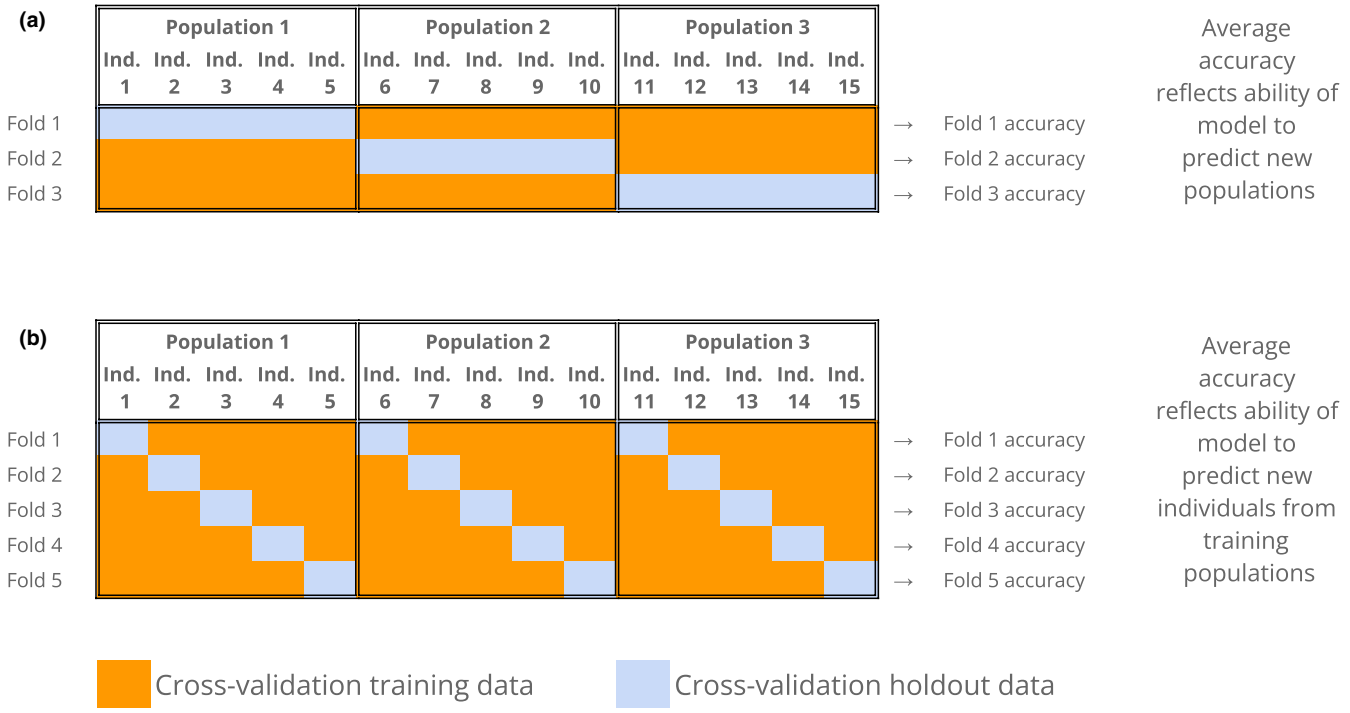

FIGURE 6 Types of cross-validation (CV). CV involves partitioning the data into folds and within each fold using a different subset of data for training the model (orange boxes) and assessing model accuracy (grey boxes). In this example, each row represents a fold of the training data. Populations are based on genetic groupings rather than sampling locations. Note that the number of populations and individuals shown here are much fewer than would be necessary for building a robust model. (a) *Leave-one-population-out cross-validation* splits the data into the number of populations and retains one population per split for estimating model accuracy. In this case, the mean accuracy for all splits reflects the ability of the model to predict new populations. (b) *Leave-one-individual-per-population-out cross-validation* splits the data into the number of individuals per population and retains one individual-per-population for estimating model accuracy. In this case, the mean accuracy for all folds reflects the ability of the model to predict new individuals from the same populations.

in both training and holdout sets. This strategy ensures that the generalizability reflects the hyperparameters being evaluated and is not affected by differences between the training and holdout sets.

Most sampling designs sample individuals within locations, and this can inform a *k*-fold strategy. Two examples of *k*-fold strategies are 'leave-one-individual-per-population-out' and 'leave-one-population-out' cross-validation (Rellstab et al., 2021). 'Leave-one-individual-per-population-out' is a stratified *k*-fold strategy in which the average accuracy in cross-validation reflects the ability of the model to generalize to new individuals from training populations (Figure 6b). 'Leave-one-population-out' has the advantage of informing the average accuracy in cross-validation reflects the ability of the model to generalize a new population (Figure 6a), but note that this 'leave-out' strategy is not stratified (because each population is not equally represented in the holdout data).

Real datasets may not be as straightforward to stratify as Figure 6, so investigators should describe how they chose to partition the data into folds based on genetic, phenotypic, geographic or climatic groups (or other logical strata in the datasets). The cross-validation should be interpreted based on the stratification used in folding because different holdout strategies test different ideas about model generalizability. Which folding strategy to choose may depend on the research goals in the study system, but in many cases it can be worthwhile to compare multiple strategies. For example, genomic prediction studies that predict traits from genotypes and environments often compare different 'leave out' strategies in which specific genotypes and/or environmental variables are excluded from the training data (Millet et al., 2019; Rogers & Holland, 2022). For instance, a study in maize (*Zea mays*) found that predictive performance depended more on the environmental similarity between the holdout data and training data than the genetic similarity between the two sets (Rogers & Holland, 2022).

Finally, it is important to understand what cross-validation is doing in each fold. For models that automatically estimate some parameters from the data (e.g., the slope and intercept in a simple linear model), each fold in cross-validation may be testing a slightly different model (because different subsets of the data lead to different parameter estimates in the model). Having a family of models creates a philosophical problem for evaluation, which seeks to evaluate a single model by comparing it to ground-truth. To develop a single model while still keeping the training and testing data independent of each other, one can split the training data into folds and explore model accuracy over a large parameter space (Figure 7a). One could combine multiple 'leave out' strategies in cross-validation to determine the effects of different kinds of folding on model accuracy (Figure 7a). Once an optimal hyperparameter set is determined, the training data can be used to build the full model (Figure 7b). Finally, the full model can be evaluated with the independent ground-truth data (Figure 7c).

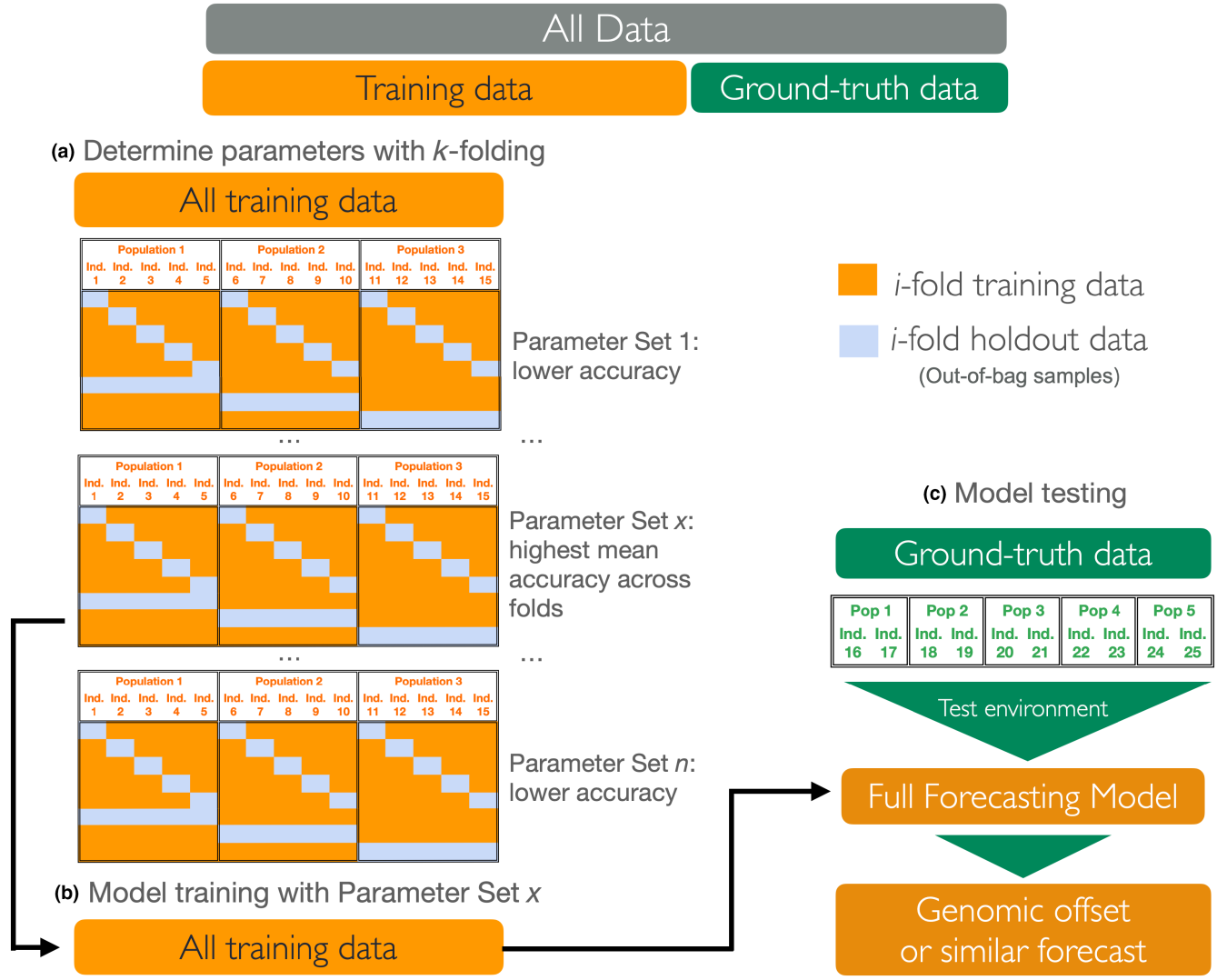

FIGURE 7 Cross-validation (CV) for hyperparameter tuning as an independent phase from model ground-truthing. (a) Some forecasting models may have hyperparameters that must be fine-tuned by the user. In this example, the data are split into 8 folds (each row within the dataset is a fold). Within each fold, the CV-training data is used to train the model given the parameter set, while the CV-test data are used to estimate model accuracy for that parameter set. The average accuracy for each parameter set is calculated as the mean over all the folds. This process is repeated for different parameter sets, until the parameter set with the highest average accuracy is determined (centre). (b) With the most accurate parameter set, all the training data can be run with that parameter set to build the full forecasting model (see also Figure 2, Step 1). (c) Test genotypes and environments, which were not used in model training, are then input into the full model to obtain a forecast in the test environment. Depending on the type of model, genotypes of the test individuals may not have to be input into the model to make a forecast. Note that the number of populations and individuals shown here are much fewer than would be necessary for building and evaluating a robust model.

3.8 | Element 8: Forecast sensitivity and uncertainty

Forecasts have many sources of uncertainty. *Sensitivity analysis* quantifies how differences in the model outputs can be allocated to differences in the model inputs. The inputs of a genomic forecasting model that may affect the output (the forecast) include the environmental data, the set of SNPs used for training, the set of individuals used for training, the climate period used for training and the climate projections. For instance, various types of environmental data have different spatial resolutions and low-resolution data can lead to high uncertainty regarding the environment at the study site (Dauphin et al., 2023). Authors can use their knowledge about the data and study system to design their own sensitivity analysis by subsetting their data in thoughtful ways, re-running the analysis and seeing how it affects the outputs. To date, only a few studies with ground-truth data have conducted some kind of sensitivity analysis. In general, forecasts have been found to be insensitive to the set of SNPs used in training (e.g., Fitzpatrick et al., 2021; Lachmuth et al., 2023; Lind et al., 2024), suggesting that the models are driven by genome-wide patterns of isolation by environment rather than specific adaptive alleles. In another study, Lachmuth et al. (2023)

explored how forecasts in red spruce (*Picea rubens*) to future climate risks were influenced by altering the genomic marker set, the set of climate variables used in training, and the climate change scenario (SSP2-RCP4.5 vs. SSP5-RCP8.5). They found that forecasts were by far the most sensitive to climate change scenario, which introduced significant uncertainty in parts of the species range that could not easily be reduced. However, they did not evaluate sensitivity to the set of populations used in model training, and it was not clear whether the model was extrapolated to novel climates. Conducting extensive sensitivity analyses as part of the routine evaluation of forecasting models, mapping geographic variation in the degree of uncertainty and showing geographic regions of extrapolation (as in DeSaix et al., 2022, see Element 5 for discussion) can aid in our understanding of the precision of forecasts.

3.9 | Element 9: Data ethics and reproducibility

Genomic forecasting is a rapidly advancing field, with new methods constantly being developed. Careful and ethical curation of data, code and forecasts will drive rapid advances in the field of genomic forecasting. The FAIR Guiding Principles for data stewardship state that archived data should be findable, accessible, interoperable and reusable (Wilkinson et al., 2016). In addition, the CARE Principles for Indigenous Data Governance provides guidelines for engagement with Indigenous Peoples rights and interests (Carroll et al., 2021, 2022). Since recent reviews have revealed that a large proportion of archived genomic data sets lack spatiotemporal metadata, growing awareness of the community on best practices for data curation is urgently needed (Crandall et al., 2023; Toczydlowski et al., 2021). The Genomics Observatories Metadatabase (GEOME) (Deck et al., 2017; Riginos et al., 2020) provides a user-friendly portal for uploading spatiotemporal FAIR metadata and linking it to genomic data stored in the International Nucleotide Sequence Database Collaboration (Cochrane et al., 2016). Crandall et al. (2023) provides a thorough overview of common genomic metadata gaps and guidelines for avoiding them, and Leigh et al. (2024) provides a set of recommendations for genomic data archival.

In addition to genomic and spatiotemporal metadata, what else should be archived? The basic elements of the ODMAP (Overview, Data, Model, Assessment and Prediction) protocol, a standard protocol for reporting species distribution models, can serve as a guideline for archiving data, code, and outputs (Zurell et al., 2020). Even when code is properly archived, it can be difficult to reproduce as hardware and software changes over time. Thus, key outputs of the code, such as the evaluation and forecasts (the 'AP' in 'ODMAP'), should be carefully archived in addition to the data. When the forecasts by a specific method are made publicly available, this facilitates comparison of new methods on the same data set. In addition, studies should follow best practices for crafting clean code (Filazzola & Lortie, 2022) and for archiving data and code (Jenkins et al., 2023).

Studies that follow the best practices for the curation of data, code and outputs will enable methods comparisons, which in turn will drive rapid advances in the field. In a best case scenario, such studies will become standardized benchmark data sets and used to compare methods on a common ground. Standardized benchmark data sets are commonly used in computer science to drive advancements in algorithm development, such as in the image recognition of handwritten digits (LeCun et al., 2021), and benchmarks are urgently needed in the field of genomic forecasting.

4 | SUMMARY AND CHECKLIST

In summary, it is becoming more accessible to molecular ecologists to build forecasting models, but evaluating their limitations in a particular study system requires careful planning and experimentation. If investigators conduct such experiments without clearly reporting basic design elements, it may be hard to understand why models work in some cases and not others. It is important for reviewers to recognize that field experiments have many logistical hurdles, and for this reason it is unlikely that any single study will meet all these design elements. Nevertheless, investigators should strive to design informative experiments and explain the limitations of their design. The following checklist will be helpful in planning and reporting:

- Element 1: The type of evaluation
  - Use a current-future evaluation design to test predictions of a population's response to climate change.
  - Use a common garden evaluation design to test predictions of the performance of multiple genotypes at the restoration site.
- Element 2: Independence between training and ground-truth data
  - To avoid circularity, samples used to train the model are different from samples used to ultimately ground-truth the model.
  - Show that sample sizes are sufficient for model training and ground-truthing.
- Element 3: The ground-truth metric(s)
  - Report how each ground-truth metric was measured or calculated.
  - Report when test individuals that will form the basis for the ground-truth data are genotyped relative to when the ground-truthing started.
  - Consider the strengths and drawbacks of different ground-truth metrics in the interpretation.
- Element 4: The evaluation metric(s)
  - Report the correlation between the ground-truth metric and the forecast for each common garden, because this can be easily compared across studies.
  - Show that the evaluation metric accurately captures predictive performance of the model.
  - Avoid inflating model performance via overfitting.
- Element 5: Relationships among training samples and test samples
  - Report the number of training and test samples.
  - Show the marker set has sufficient coverage of the genome.

- If necessary, genotype the same markers in training samples and test samples.
- Report the degree of (non)independence among test and training samples as relatedness and/or FST.
- Use jackknife to explore how model performance changes when particular test genotypes are removed from the evaluation.

- Element 6: The training and test environments: quantification, novelty, uncertainty and variability
  - Measure the multivariate environment in the same way in the training and ground-truth/test datasets.
  - Show that the chosen environmental variables are relevant to fitness.
  - Quantify the common garden test environment during the period of the test.
  - Report the degree of climate novelty between each test environment and all the training environments.
- Element 7: Model training
  - Maintain independent training and ground-truth datasets by partitioning training data into folds for hyperparameter tuning.
  - Report how training data were divided into folds for cross-validation.
  - Interpret cross-validation as a summary of model fit, informed by the type of folding.
- Element 8: Forecast sensitivity and uncertainty
  - Report the sensitivity of the forecasts to the set of populations used in training, genomic marker set, the set of climate variables used in training and/or the climate change scenario, as applicable.
- Element 9: Data Ethics and Reproducibility
  - Deposit genomic data in the International Nucleotide Sequence Database and link it to spatiotemporal metadata with GEOME.
  - Follow FAIR and CARE Guiding Principles for data stewardship.
  - Follow the ODMAP (Overview, Data, Model, Assessment and Prediction) protocol for archiving the forecasting model and the evaluation.
  - Follow best practices for crafting and archiving code.

AUTHOR CONTRIBUTIONS

Katie E. Lotterhos conceived the ideas and wrote the manuscript.

ACKNOWLEDGEMENTS

The author wishes to thank Stephen Keller, Brandon Lind, Madeline Eppley and two anonymous reviewers for helpful comments on the manuscript. This project was supported by funds from the National Science Foundation (2043905).

CONFLICT OF INTEREST STATEMENT

The author has no conflict of interest to declare.

PEER REVIEW

The peer review history for this article is available at https://www.webofscience.com/api/gateway/wos/peer-review/10.1111/2041-210X.14379.

DATA AVAILABILITY STATEMENT

There is no data or code associated with this article. This study did not involve fieldwork.

ORCID

Katie E. Lotterhos https://orcid.org/0000-0001-7529-2771

REFERENCES

Augusiak, J., Van den Brink, P. J., & Grimm, V. (2014). Merging validation and evaluation of ecological models to "evaludation": A review of terminology and a practical approach. *Ecological Modelling*, *280*, 117–128.

Barrett, R. D. H., & Hoekstra, H. E. (2011). Molecular spandrels: Tests of adaptation at the genetic level. *Nature Reviews Genetics*, *12*(11), 767–780.

Bay, R. A., Harrigan, R. J., Underwood, V. L., Gibbs, H. L., Smith, T. B., & Ruegg, K. (2018). Genomic signals of selection predict climate-driven population declines in a migratory bird. *Science*, *359*(6371), 83–86.

Bergstrom, C. T., & West, J. D. (2021). *Calling bullshit: The art of skepticism in a data-driven world*. Random House Publishing Group.

Bitter, M. C., Wong, J. M., Dam, H. G., Donelan, S. C., Kenkel, C. D., Komoroske, L. M., Nickols, K. J., Rivest, E. B., Salinas, S., Burgess, S. C., & Lotterhos, K. E. (2021). Fluctuating selection and global change: A synthesis and review on disentangling the roles of climate amplitude, predictability and novelty. *Proceedings of the Royal Society B: Biological Sciences*, *288*(1957), 20210727.

Blanquart, F., Kaltz, O., Nuismer, S. L., & Gandon, S. (2013). A practical guide to measuring local adaptation. *Ecology Letters*, *16*(9), 1195–1205.

Brauer, C. J., Sandoval-Castillo, J., Gates, K., Hammer, M. P., Unmack, P. J., Bernatchez, L., & Beheregaray, L. B. (2023). Natural hybridization reduces vulnerability to climate change. *Nature Climate Change*, *13*(3), 282–289.

Broennimann, O., Petitpierre, B., Chevalier, M., González-Suárez, M., Jeschke, J. M., Rolland, J., Gray, S. M., Bacher, S., & Guisan, A. (2021). Distance to native climatic niche margins explains establishment success of alien mammals. *Nature Communications*, *12*(1), 2353.

Burford Reiskind, M. O., Moody, M. L., Bolnick, D. I., Hanifin, C. T., & Farrior, C. E. (2021). Nothing in evolution makes sense except in the light of biology. *Bioscience*, *71*, 370–382. https://doi.org/10.1093/biosci/biaa170

Capblancq, T., & Forester, B. R. (2021). Redundancy analysis: A Swiss Army Knife for landscape genomics. *Methods in Ecology and Evolution*, *12*(12), 2298–2309.

Carroll, S. R., Herczog, E., Hudson, M., Russell, K., & Stall, S. (2021). Operationalizing the CARE and FAIR Principles for Indigenous data futures. *Scientific Data*, *8*(1), 108.

Carroll, S. R., Plevel, R., Jennings, L. L., Garba, I., Sterling, R., Cordova-Marks, F. M., Hiratsuka, V., Hudson, M., & Garrison, N. A. (2022). Extending the CARE Principles from tribal research policies to benefit sharing in genomic research. *Frontiers in Genetics*, *13*, 1052620.

Chen, Y., Jiang, Z., Fan, P., Ericson, P. G. P., Song, G., Luo, X., Lei, F., & Qu, Y. (2022). The combination of genomic offset and niche modelling provides insights into climate change-driven vulnerability. *Nature Communications*, *13*(1), 4821.

Cochrane, G., Karsch-Mizrachi, I., & Takagi, T. (2016). International nucleotide sequence database collaboration. *Nucleic Acids Research*, *44*(D1), D48–D50.

Conner, J. K., & Hartl, D. L. (2004). *A primer of ecological genetics*. Sunderland, MA, USA: Sinauer Associates.

Crandall, E. D., Toczydlowski, R. H., Liggins, L., Holmes, A. E., Ghoojaei, M., Gaither, M. R., Wham, B. E., Pritt, A. L., Noble, C., Anderson,

T. J., Barton, R. L., Berg, J. T., Beskid, S. G., Delgado, A., Farrell, E., Himmelsbach, N., Queeno, S. R., Trinh, T., Weyand, C., … Toonen, R. J. (2023). Importance of timely metadata curation to the global surveillance of genetic diversity. *Conservation Biology*, *37*, e14061.

Dauphin, B., Rellstab, C., Wüest, R. O., Karger, D. N., Holderegger, R., Gugerli, F., & Manel, S. (2023). Re-thinking the environment in landscape genomics. *Trends in Ecology & Evolution*, *38*(3), 261–274.

de Villemereuil, P., Gaggiotti, O. E., Mouterde, M., & Till-Bottraud, I. (2016). Common garden experiments in the genomic era: New perspectives and opportunities. *Heredity*, *116*(3), 249–254.

Deck, J., Gaither, M. R., Ewing, R., Bird, C. E., Davies, N., Meyer, C., Riginos, C., Toonen, R. J., & Crandall, E. D. (2017). The Genomic Observatories Metadatabase (GeOMe): A new repository for field and sampling event metadata associated with genetic samples. *PLoS Biology*, *15*(8), e2002925.

DeSaix, M. G., George, T. L., Seglund, A. E., Spellman, G. M., Zavaleta, E. S., & Ruegg, K. C. (2022). Forecasting climate change response in an alpine specialist songbird reveals the importance of considering novel climate. *Diversity and Distributions*, *28*(10), 2239–2254.

DeWitt, T. J., & Scheiner, S. M. (2004). *Phenotypic plasticity: Functional and conceptual approaches*. Oxford University Press.

Exposito-Alonso, M., 500 Genomes Field Experiment Team, Burbano, H. A., Bossdorf, O., Nielsen, R., & Weigel, D. (2019). Natural selection on the Arabidopsis thaliana genome in present and future climates. *Nature*, *573*(7772), 126–129.

Filazzola, A., & Lortie, C. J. (2022). A call for clean code to effectively communicate science. *Methods in Ecology and Evolution*, *13*(10), 2119–2128.

Fitzpatrick, M. C., Chhatre, V. E., Soolanayakanahally, R. Y., & Keller, S. R. (2021). Experimental support for genomic prediction of climate maladaptation using the machine learning approach Gradient Forests. *Molecular Ecology Resources*, *21*(8), 2749–2765.

Fitzpatrick, M. C., & Keller, S. R. (2015). Ecological genomics meets community-level modelling of biodiversity: Mapping the genomic landscape of current and future environmental adaptation. *Ecology Letters*, *18*(1), 1–16.

Fitzpatrick, M. C., Keller, S. R., & Lotterhos, K. E. (2018). Comment on "genomic signals of selection predict climate-driven population declines in a migratory bird". *Science*, *361*, eaat7279.

Foden, W. B., Young, B. E., Akçakaya, H. R., Garcia, R. A., Hoffmann, A. A., Stein, B. A., Thomas, C. D., Wheatley, C. J., Bickford, D., Carr, J. A., Hole, D. G., Martin, T. G., Pacifici, M., Pearce-Higgins, J. W., Platts, P. J., Visconti, P., Watson, J. E. M., & Huntley, B. (2019). Climate change vulnerability assessment of species. *Wiley Interdisciplinary Reviews: Climate Change*, *10*(1), e551.

Gain, C., & François, O. (2021). LEA 3: Factor models in population genetics and ecological genomics with R. *Molecular Ecology Resources*, *21*(8), 2738–2748.

Gain, C., Rhonê, B., Cubry, P., Salazar, I., Forbes, F., Vigouroux, Y., Jay, F., & François, O. (2023). A quantitative theory for genomic offset statistics. *Molecular Biology and Evolution*, *40*(6), msad140.

Goodall, D. W. (1972). Building and testing ecosystem models. In J. N. J. Jeffers (Ed.), *Mathematical models in ecology* (pp. 173–194). Oxford: Blackwell.

Grimm, D. G., Azencott, C.-A., Aicheler, F., Gieraths, U., MacArthur, D. G., Samocha, K. E., Cooper, D. N., Stenson, P. D., Daly, M. J., Smoller, J. W., Duncan, L. E., & Borgwardt, K. M. (2015). The evaluation of tools used to predict the impact of missense variants is hindered by two types of circularity. *Human Mutation*, *36*(5), 513–523.

Hoffmann, A. A., Weeks, A. R., & Sgrò, C. M. (2021). Opportunities and challenges in assessing climate change vulnerability through genomics. *Cell*, *184*(6), 1420–1425.

Ingvarsson, P. K., & Bernhardsson, C. (2020). Genome-wide signatures of environmental adaptation in European aspen (*Populus tremula*) under current and future climate conditions. *Evolutionary Applications*, *13*(1), 132–142.

IPCC. (2023). *Synthesis report of the IPCC sixth assessment report (AR6)*. Intergovernmental Panel on Climate Change.

Jenkins, G. B., Beckerman, A. P., Bellard, C., Benítez-López, A., Ellison, A. M., Foote, C. G., Hufton, A. L., Lashley, M. A., Lortie, C. J., Ma, Z., Moore, A. J., Narum, S. R., Nilsson, J., O'Boyle, B., Provete, D. B., Razgour, O., Rieseberg, L., Riginos, C., Santini, L., … Peres-Neto, P. R. (2023). Reproducibility in ecology and evolution: Minimum standards for data and code. *Ecology and Evolution*, *13*(5), e9961.

Jordan, R., Hoffmann, A. A., Dillon, S. K., & Prober, S. M. (2017). Evidence of genomic adaptation to climate in *Eucalyptus microcarpa*: Implications for adaptive potential to projected climate change. *Molecular Ecology*, *26*(21), 6002–6020.

Lachmuth, S., Capblancq, T., Keller, S. R., & Fitzpatrick, M. C. (2023). Assessing uncertainty in genomic offset forecasts from landscape genomic models (and implications for restoration and assisted migration). *Frontiers in Ecology and Evolution*, *11*, 1155783. https://doi.org/10.3389/fevo.2023.1155783

Láruson, Á. J., Fitzpatrick, M. C., Keller, S. R., Haller, B. C., & Lotterhos, K. E. (2022). Seeing the forest for the trees: Assessing genetic offset predictions from gradient forest. *Evolutionary Applications*, *15*(3), 403–416.

LeCun, Y., Cortes, C., & Burges, C. J. C. (2021). *The mnist database of handwritten digits*. http://yann.lecun.com/exdb/mnist/

Legendre, P., & Legendre, L. (1998). *Numerical ecology*. Elsevier.

Leigh, D. M., Vandergast, A. G., Hunter, M. E., Crandall, E. D., Funk, W. C., Garroway, C. J., Hoban, S., Oyler-McCance, S. J., Rellstab, C., Segelbacher, G., Schmidt, C., Vázquez-Domínguez, E., & Paz-Vinas, I. (2024). Best practices for genetic and genomic data archiving. *Nature Ecology & Evolution*. https://doi.org/10.1038/s41559-024-02423-7

Lind, B. M., Candido-Ribeiro, R., Singh, P., Lu, M., Vidakovic, D. O., Booker, T. R., Whitlock, M., Isabel, N., Yeaman, S., & Aitken, S. N. (2024). How useful is genomic data for predicting maladaptation to future climate? *Global Change Biology*, *30*, e17227.

Lind, B. M., & Lotterhos, K. E. (2024). The limits of predicting maladaptation to future environments with genomic data. *bioRxiv*. https://doi.org/10.1101/2024.01.30.577973

Lotterhos, K. E. (2023). The paradox of adaptive trait clines with nonclinal patterns in the underlying genes. *Proceedings of the National Academy of Sciences of the United States of America*, *120*(12), e2220313120.

Lotterhos, K. E. (2024). Interpretation issues with "genomic vulnerability" arise from conceptual issues in local adaptation and maladaptation. *Evolution Letters*, *8*, 331–339. https://doi.org/10.1093/evlett/qrae004

Lotterhos, K. E., Fitzpatrick, M. C., & Blackmon, H. (2022). Simulation tests of methods in evolution, ecology, and systematics: Pitfalls, progress, and principles. *Annual Review of Ecology, Evolution, and Systematics*, *53*, 113–136. https://doi.org/10.1146/annurev-ecolsys-102320-093722

Lotterhos, K. E., Láruson, Á. J., & Jiang, L.-Q. (2021). Novel and disappearing climates in the global surface ocean from 1800 to 2100. *Scientific Reports*, *11*(1), 15535.

Lotterhos, K. E., Moore, J. H., & Stapleton, A. E. (2018). Analysis validation has been neglected in the age of reproducibility. *PLoS Biology*, *16*(12), e3000070.

Lowry, D. B., Hoban, S., Kelley, J. L., Lotterhos, K. E., Reed, L. K., Antolin, M. F., & Storfer, A. (2017). Breaking RAD: An evaluation of the utility of restriction site-associated DNA sequencing for genome scans of adaptation. *Molecular Ecology Resources*, *17*(2), 142–152.

Mahony, C. R., Cannon, A. J., Wang, T., & Aitken, S. N. (2017). A closer look at novel climates: New methods and insights at continental to landscape scales. *Global Change Biology*, *23*, 3934–3955. https://doi.org/10.1111/gcb.13645

McArdle, B. H. (1988). The structural relationship: Regression in biology. *Canadian Journal of Zoology*, *66*(11), 2329–2339.

Millet, E. J., Kruijer, W., Coupel-Ledru, A., Alvarez Prado, S., Cabrera-Bosquet, L., Lacube, S., Charcosset, A., Welcker, C., van Eeuwijk, F., & Tardieu, F. (2019). Genomic prediction of maize yield across European environmental conditions. *Nature Genetics*, *51*(6), 952–956.

Molnar, C. (2021). *Interpretable machine learning: A guide for making black box models explainable*. https://christophm.github.io/interpretable-ml-book/

Rellstab, C., Dauphin, B., & Exposito-Alonso, M. (2021). Prospects and limitations of genomic offset in conservation management. *Evolutionary Applications*, *14*(5), 1202–1212.

Rellstab, C., Gugerli, F., Eckert, A. J., Hancock, A. M., & Holderegger, R. (2015). A practical guide to environmental association analysis in landscape genomics. *Molecular Ecology*, *24*(17), 4348–4370.

Rellstab, C., Zoller, S., Walthert, L., Lesur, I., Pluess, A. R., Graf, R., Bodénès, C., Sperisen, C., Kremer, A., & Gugerli, F. (2016). Signatures of local adaptation in candidate genes of oaks (*Quercus* spp.) with respect to present and future climatic conditions. *Molecular Ecology*, *25*(23), 5907–5924.

Rhoné, B., Defrance, D., Berthouly-Salazar, C., Mariac, C., Cubry, P., Couderc, M., Dequincey, A., Assoumanne, A., Kane, N. A., Sultan, B., Barnaud, A., & Vigouroux, Y. (2020). Pearl millet genomic vulnerability to climate change in West Africa highlights the need for regional collaboration. *Nature Communications*, *11*(1), 5274.

Riginos, C., Crandall, E. D., Liggins, L., Gaither, M. R., Ewing, R. B., Meyer, C., Andrews, K. R., Euclide, P. T., Titus, B. M., Therkildsen, N. O., Salces-Castellano, A., Stewart, L. C., Toonen, R. J., & Deck, J. (2020). Building a global genomics observatory: Using GEOME (the Genomic Observatories Metadatabase) to expedite and improve deposition and retrieval of genetic data and metadata for biodiversity research. *Molecular Ecology Resources*, *20*(6), 1458–1469.

Rochat, E., & Joost, S. (2021). Spatial areas of genotype probability: Predicting the spatial distribution of adaptive genetic variants under future climatic conditions. *Diversity and Distributions*, *27*, 1076–1090.

Rogers, A. R., & Holland, J. B. (2022). Environment-specific genomic prediction ability in maize using environmental covariates depends on environmental similarity to training data. *G3*, *12*(2). https://doi.org/10.1093/g3journal/jkab440

Ruegg, K., Bay, R. A., Anderson, E. C., Saracco, J. F., Harrigan, R. J., Whitfield, M., Paxton, E. H., & Smith, T. B. (2018). Ecological genomics predicts climate vulnerability in an endangered southwestern songbird. *Ecology Letters*, *21*(7), 1085–1096.

Sang, Y., Long, Z., Dan, X., Feng, J., Shi, T., Jia, C., Zhang, X., Lai, Q., Yang, G., Zhang, H., Xu, X., Liu, H., Jiang, Y., Ingvarsson, P. K., Liu, J., Mao, K., & Wang, J. (2022). Genomic insights into local adaptation and future climate-induced vulnerability of a keystone forest tree in East Asia. *Nature Communications*, *13*(1), 6541.

Sork, V. L., Aitken, S. N., Dyer, R. J., Eckert, A. J., Legendre, P., & Neale, D. B. (2013). Putting the landscape into the genomics of trees: Approaches for understanding local adaptation and population responses to changing climate. *Tree Genetics & Genomes*, *9*(4), 901–911.

Toczydlowski, R. H., Liggins, L., Gaither, M. R., Anderson, T. J., Barton, R. L., Berg, J. T., Beskid, S. G., Davis, B., Delgado, A., Farrell, E., Ghoojaei, M., Himmelsbach, N., Holmes, A. E., Queeno, S. R., Trinh, T., Weyand, C. A., Bradburd, G. S., Riginos, C., Toonen, R. J., & Crandall, E. D. (2021). Poor data stewardship will hinder global genetic diversity surveillance. *Proceedings of the National Academy of Sciences of the United States of America*, *118*(34), e2107934118. https://doi.org/10.1073/pnas.2107934118

Waldvogel, A.-M., Feldmeyer, B., Rolshausen, G., Exposito-Alonso, M., Rellstab, C., Kofler, R., Mock, T., Schmid, K., Schmitt, I., Bataillon, T., Savolainen, O., Bergland, A., Flatt, T., Guillaume, F., & Pfenninger, M. (2020). Evolutionary genomics can improve prediction of species' responses to climate change. *Evolution Letters*, *4*(1), 4–18.

Whitlock, M. C., & Schluter, D. (2009). *The analysis of biological data* (1st ed.). Roberts and Company Publishers.

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., … Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, *3*, 160018.

Williams, J. W., Jackson, S. T., & Kutzbach, J. E. (2007). Projected distributions of novel and disappearing climates by 2100 AD. *Proceedings of the National Academy of Sciences of the United States of America*, *104*(14), 5738–5742.

Zurell, D., Franklin, J., König, C., Bouchet, P. J., Dormann, C. F., Elith, J., Fandos, G., Feng, X., Guillera-Arroita, G., Guisan, A., Lahoz-Monfort, J. J., Leitão, P. J., Park, D. S., Peterson, A. T., Rapacciuolo, G., Schmatz, D. R., Schröder, B., Serra-Diaz, J. M., Thuiller, W., … Merow, C. (2020). A standard protocol for reporting species distribution models. *Ecography*, *43*(9), 1261–1277.

How to cite this article: Lotterhos, K. E. (2024). Principles in experimental design for evaluating genomic forecasts. *Methods in Ecology and Evolution*, *15*, 1466–1482. https://doi.org/10.1111/2041-210X.14379