Gaussian Process Modeling with Genotype × Environment Kernels for Wheat Performance Prediction

Lea FRIEDLI, Tim STEINERT, Nathalie WUYTS, Fabian GUIGNARD, Lilia Levy HÄNER, Juan Manuel HERRERA, and David GINSBOURGER

Optimizing wheat variety selection for high performance in different environmental conditions is critical for reliable food production and stable incomes for growers. We employ a statistical machine learning framework utilizing Gaussian process (GP) models to capture the effects of genetic and environmental factors on wheat yield and protein content. The GP approach is closely related to kernel linear mixed-effect and kernel regression models, which are commonly used for genotype × environment predictions. A notable advantage of the GP formulation is that it provides probabilistic predictions in closed form. Building on this, there is extensive literature for optimal decision-making and data acquisition strategies, opening the door to variety recommendation and experimental design strategies. By means of a wheat test case and using a novel dataset collected in Switzerland, we demonstrate that the GP approach delivers comparable or superior prediction results in comparison with state-of-the-art methods, while providing improved uncertainty quantification and supporting better informed decision-making.

Key Words: Covariance kernels; Genotype-environment predictions; Leakage; Multi-environment variety trials; Uncertainty quantification.

1. INTRODUCTION

Wheat is a staple crop essential for global food security, and its growth is shaped by an interplay of genetic, environmental, and management factors. Understanding how different wheat varieties respond to varying environmental conditions is crucial, especially in the face of climate change. When the goal is to identify broadly adapted varieties or those suited to specific environments, the patterns of genotype × environment (G × E; Kang and

L.Friedli (✉) Engineering Risk Analysis Group, Technical University of Munich, Munich, Germany (E-mail: *lea.friedli@hotmail.com*).
T. Steinert, D. Ginsbourger Institute of Mathematical Statistics and Actuarial Science, University of Bern, Bern, Switzerland.
N. Wuyts, L. Levy. Häner, Juan Manuel Herrera D. Pellet Agroscope, Plant Production Systems, Nyon, Switzerland.
F. Guignard Federal Institute of Metrology METAS, Wabern, Switzerland.

Journal of Agricultural, Biological, and Environmental Statistics
https://doi.org/10.1007/s13253-026-00749-2

Published online: 23 August 2026

Gorman 1989; Cooper and DeLacy 1994) interactions must be considered. Since not all combinations can be tested in the field, statistical and machine learning models serve as efficient surrogates to predict the performance at new combinations (e.g., Freeman 1973; van Dijk et al. 2021).

As a nonparametric, probabilistic, and interpretable statistical machine learning alternative, we explore Gaussian process (GP) modeling, leveraging its ability to model G × E interactions through kernels defined on the G × E product space. GP models originate from the geostatistical interpolation method known as kriging (Krige 1951), but have broadened their application far beyond geostatistics (Rasmussen and Williams 2006). GP modeling, aiming (in its most common form) to learn a function $f : \mathcal{X} \to \mathbb{R}$, can be applied to a wide range of input spaces $\mathcal{X}$, handling the diverse inputs through tailored covariance kernels. Apart from its versatility, GP modeling is popular for enabling uncertainty quantification, handling small training datasets, allowing the incorporation of prior knowledge through the covariance kernel, and providing interpretable results. In recent years, GPs have been increasingly applied across a variety of domains, including chemistry (e.g., Griffiths et al. 2023), engineering (e.g., Su et al. 2017), and environmental modeling (e.g., Ray 2021). In the considered GP approaches for G × E modeling, there are three major degrees of freedom illustrated in Fig. 1: the choice of the kernel for i) the genetic effects, ii) the environmental effects, and iii) the strategy used to construct a joint kernel over the product space of genotypes and environments. Note: Here, 'GP' refers to Gaussian process, not genomic-enabled prediction, which is also commonly abbreviated as 'GP' in agricultural literature.

GP models offer promising potential for decision-making and data acquisition strategies. Once a model is fitted, it provides probabilistic predictions at any candidate input (from the considered input space), enabling the search for inputs that yield optimal responses. In the context of G × E, this means GP models can predict how different genotypes will perform across a range of environmental conditions, for instance enabling to compare and select varieties for specific environments. In this realm, GP models have been used for global optimization for several decades and the resulting field now referred to as Bayesian optimization (BO) has become very active from engineering to machine learning and beyond (Močkus 1974; Jones et al. 1998; Garnett 2023). Here, a clear advantage of GP modeling is its capability to handle associated uncertainties, for instance by separating controllable (e.g., variety selection) from non-controllable variables (e.g., temperature), aiming to optimize performance over the controllable ones under average or worst-case conditions (Williams et al. 2000; Janusevskis and Le Riche 2013; Ginsbourger et al. 2014). Apart from variety selection, GP models may be leveraged to design novel agronomical experiments, allowing to work out acquisition functions dedicated to targeted experimental design, e.g., focusing on specific ranges of the response (Chevalier et al. 2014). In the context of G × E, this allows researchers to strategically explore risk-aware combinations for diverse or changing climates.

We evaluate GP methods using a novel dataset targeting wheat yield and grain protein content, collected between 1990 and 2023 in multi-environment trials (METs) across the Swiss wheat production zone (Levy Häner et al. 2025). We rely on the concept of 'environment', which is referring to a combination of trial location and harvest year, and described only by meteorological data without considering the location or the year. Doing so enables notably extrapolating to future years (and new locations) by relying on projections from cli-

mate models. Also for the genotype, we only consider covariates, given by single-nucleotide polymorphisms (SNPs; e.g., Rafalski 2002). Whereas this strategy has the drawback that the considered covariates may not fully describe differences across locations, years, and varieties (including climate trends, soil conditions, management shifts, or genetic progress), it leads to a very general prediction model, applicable for any combination of meteorological conditions and genetic markers. This resonates with the presented test case as Switzerland has a very heterogeneous and complex landscape, such that geographically close locations are not necessarily similar.

We consider two prediction scenarios: new environment and new genotype. Throughout the manuscript, we use the term 'variety' to refer to the categorical factor that can be selected in an experiment and the term 'genotype' to refer to the underlying genetic information. Predicting the performance of genotypes in new environments not only enables predictions in future meteorological conditions but also enables tailored recommendations for farmers under specific local conditions. Predicting new varieties is closely linked to variety evaluation in plant breeding and official value for cultivation and use (VCU; Strebel et al. 2025; Ramakers et al. 2026). Varieties are usually only tested during two to three years, which means that not all environments, e.g., meteorological conditions, can be tested. For example, environments with drought stress at flowering may be missing. GP models can then be used to predict the performance of new varieties in these unobserved stress environments. This generalizes to the handling of incomplete METs arising from sowing issues, management errors, unexpected damage, or cost constraints. Especially in incomplete METs, information from one or more trials conducted in the target environment or including the genotype of interest may already be available. We consider this scenario by means of controlled leakage, allowing one observation into the training set to mimic minimal prior knowledge.

With this manuscript, our goal is to introduce the GP approach for G × E prediction, and to bridge the gap between the GP and the state-of-the art LMM and regression methods. We also provide a straightforward proof-of-concept case study using BO to demonstrate the potential of the GP approach for decision-making and targeted data acquisition. In doing so, we hope that this work paves the way for promising future applications of GP modeling. The manuscript is organized as follows: Sect. 2 presents a literature review of G × E prediction models. Section 3 reviews GP modeling and links to the state-of-the art methods. Section 4 outlines the dataset and implementation, while Sect. 5 reports the results. Then, Sect. 6 is about the proof-of-concept case study using BO. Finally, Sect. 7 concludes with a discussion and summary.

2. LITERATURE REVIEW G × E PREDICTION

Linear mixed-effect models are a widely used approach in G × E prediction (LMMs; e.g., Piepho et al. 1998; Meuwissen et al. 2001; Bernardo et al. 2002; VanRaden 2007; Buntaran et al. 2021). A basic single-environment model analyzes varieties within each environment individually (Bernardo et al. 2002). When genetic markers are available, genetic effects are typically modeled using regression (Meuwissen et al. 2001). However, the regression coefficients cannot be estimated when the number of markers greatly exceeds the number

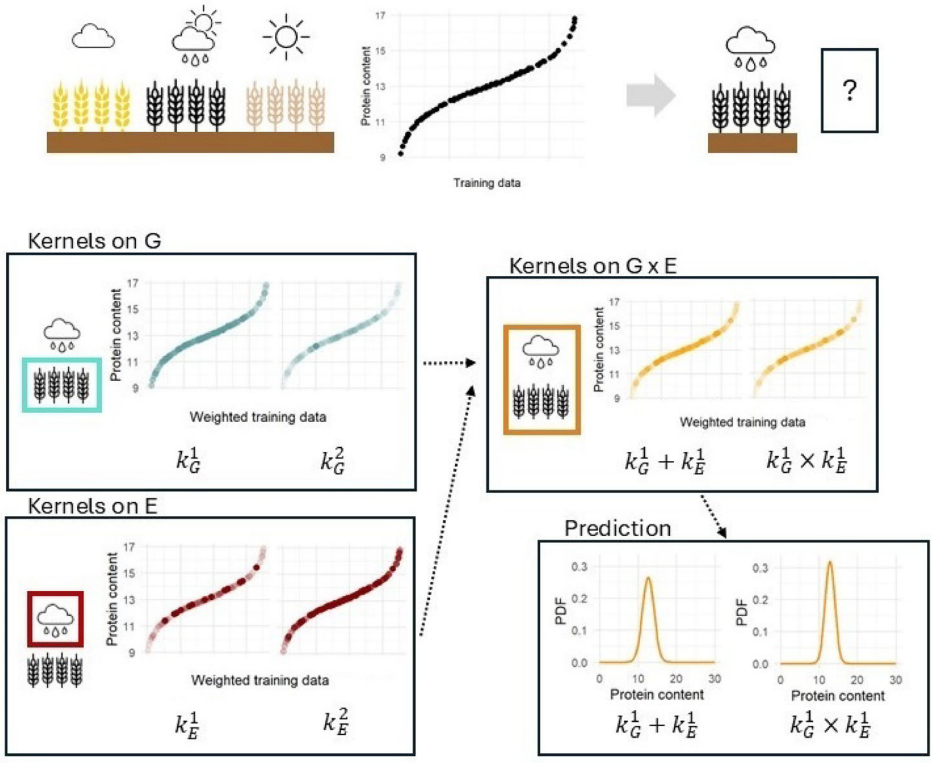

Figure 1. Illustration of GP modeling for G × E prediction: Given training data from specific G × E configurations, the goal is to predict outcomes (here the wheat grain protein content) for a new scenario. Kernels on *G* and *E* are used to quantify similarity between the training data and the target scenario, effectively weighting the training points (indicated by transparency of the colorful dots). The kernels are then combined to construct a joint kernel over the product space, and predictions are made based on the weighted training data according to this combined kernel. In GP modeling, the predictive distributions are Gaussian, as illustrated by the bell-shaped probability density functions depicted in prediction.

of observations (Crossa et al. 2017). Two solutions to this are ridge regression (Hoerl and Kennard 1970) and the least absolute shrinkage and selection operator (LASSO; Tibshirani 1996).

The genetic effect can also be modeled using random effects with a centered multivariate Gaussian distribution, where the covariance matrix is modeled with a genomic relationship matrix derived from a kernel applied to the marker data. A specific choice of linear kernel yields the genomic best linear unbiased prediction (GBLUP) model (VanRaden 2007). With an appropriate specification of the genomic relationship matrix, GBLUP and ridge regression are equivalent in terms of predictive performance. In practice, linear kernels frequently encounter limitations in their ability to capture nonlinear phenomena. Consequently, there is a growing preference for alternative kernel-based methods, which have been shown to be effective in modeling complex nonlinear relationships (Montesinos-López et al. 2022). Nonlinear kernels, such as the Gaussian kernel, have shown improved performance in several studies (Cuevas et al. 2016, 2017; Bandeira e Sousa et al. 2017). More recently, kernels

informed by deep learning techniques have also been explored and delivered improved biological accuracy and processing speed, as did the Gaussian Kernel (Cuevas et al. 2019; Costa-Neto et al. 2021). In this realm, we would also like to mention reproducing kernel Hilbert space (RKHS) regression, also called kernel ridge regression (Gianola et al. 2006; Montesinos-López et al. 2022), which is the generalization of ridge regression including nonlinear kernels. For instance, Crossa et al. (2010) obtained promising results using RKHS regression for genetic selection.

Multienvironment trials (METs) allow to borrow information across environments. Thereby, variety performance varies across different environments (Cooper and DeLacy 1994). Explicitly modeling each interaction potentially results in a large number of parameters (Burgueño et al. 2012). One method to address this involves defining mega-environments by grouping environments based on patterns of environmental stress factors (e.g., Chapman et al. 2000). Another approach treats the interactions as low genetic correlation between environments, and models performances in different environments as different correlated traits (Falconer and Mackay 1996; Piepho et al. 1998; Burgueño et al. 2012). However, since these methods rely on observed covariance between environments, they cannot predict variety performance in new, untested environments (Heslot et al. 2014; Malosetti et al. 2016).

To enable prediction in new environments, environmental covariates responsible for the interactions can be incorporated (Jarquín et al. 2014; Heslot et al. 2014). In factorial regression models (Van Eeuwijk and Elgersma 1993; Piepho et al. 1998), the response is regressed on environmental covariates, where the regression coefficients are variety-specific. While this approach provides a flexible framework for modeling variety-by-environment interactions, its applicability may be limited when dealing with high-dimensional environmental data (Jarquín et al. 2014; Raffo et al. 2022). A simple solution would be to introduce variable selection, but this could result in important volumes of information being discarded (Bernardo et al. 2002). Another approach is to integrate crop growth models, which link genetic markers to physiological traits, which then lead to interactions when combined with environmental data to simulate trait performance (Chenu et al. 2008; Messina et al. 2009; Heslot et al. 2014; Technow et al. 2015; Cooper et al. 2016; Washburn et al. 2020).

Jarquín et al. (2014) discuss a family of random effects models that rely on environmental covariance to model dependencies among the random environmental effects. This approach extends the GBLUP model and can be interpreted as reaction norm model (Woltereck 1909; Gregorius and Namkoong 1986; Falconer 1990). Thereby, Jarquín et al. (2014) assume that the interaction effect follows a normal distribution, with the covariance structure given by the element-wise (Hadamard) product of the two separate covariance matrices. However, Jarquín et al. (2014)'s reaction norm model has faced criticism from Cuevas et al. (2017) for a restrictive structure imposed on the interaction term. This has been addressed for instance by Cuevas et al. (2016; 2017), Bandeira E Sousa et al. (2017), and Costa-Neto et al. (2021), who demonstrated that incorporating Gaussian and Deep kernels significantly enhances predictive performance compared to models employing standard linear kernels.

Finally, we refer to the recent reviews by Van Dijk et al. (2021) and Crossa et al. (2025) for insights into the growing importance of machine learning in plant science and breeding. This topic is not covered in the present discussion. Furthermore, we would like to draw attention

to the recent open benchmark of G × E models by Washburn et al. (2025), demonstrating that no single model or strategy is far superior to all others.

3. GAUSSIAN PROCESS REGRESSION WITH COMBINATIONS OF GENOTYPE AND ENVIRONMENT KERNELS

In G × E prediction, we are typically interested in a performance trait such as yield or protein content. We observe (possibly noisy) responses at some G × E combinations and aim to predict the trait at new ones. In this work, we assume additional information for each combination, given by the covariate tuple $\boldsymbol{x}=(\boldsymbol{x}_{\mathrm{G}}, \boldsymbol{x}_{\mathrm{E}})$, with the environmental variables $\boldsymbol{x}_{\mathrm{E}} \in \mathcal{X}_{\mathrm{E}}$ and the genetic vectors $\boldsymbol{x}_{\mathrm{G}} \in \mathcal{X}_{\mathrm{G}}$.

3.1. GAUSSIAN PROCESS BASICS

Gaussian processes (GPs) are a flexible, probabilistic approach to modeling unknown functions. A GP, denoted by $\boldsymbol{\xi} = (\xi(\boldsymbol{x}))_{\boldsymbol{x}\in\mathcal{X}}$, is a collection of real-valued random variables defined over the same probability space. The defining property of a GP is that, for any finite set of points $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\} \subset \mathcal{X}$ with $n \geq 1$, $(\xi(\boldsymbol{x}_1), \ldots, \xi(\boldsymbol{x}_n))$ is multivariate Gaussian distributed. A GP is characterized (in distribution) by a mean function $\mu : \mathcal{X} \to \mathbb{R}$, $\mu(\boldsymbol{x}) := \mathbb{E}[\xi(\boldsymbol{x})]$, and a covariance function (kernel) $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, $k(\boldsymbol{x}, \boldsymbol{x}') := \mathrm{Cov}(\xi(\boldsymbol{x}), \xi(\boldsymbol{x}'))$.

In the considered GP modeling context, the target function $f : \mathcal{X} \to \mathbb{R}$ is treated as a sample from a GP $\boldsymbol{\xi}$, which, in a Bayesian context, can be seen as a prior model. We assume that we observe noisy measurements of f at n input points $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$, resulting in observations of the form $Z_i = f(\boldsymbol{x}_i) + \varepsilon_i$, where the noise terms are i.i.d. with $\varepsilon_i \sim \mathcal{N}(0, \tau^2)$. By conditioning the GP on the observed data $\mathcal{A}_n = \{\boldsymbol{Z}_n = \boldsymbol{z}_n\}$, we obtain a posterior GP characterized by a predictive mean function $m_n(\boldsymbol{x})$ and a posterior covariance function $k_n(\boldsymbol{x}, \boldsymbol{x}')$. We consider the case of ordinary kriging, assuming a constant but unknown mean μ. Then, the best linear unbiased estimator (BLUE) for the mean is,

$$\widehat{\mu} = \frac{\mathbf{1}_n^T \left(\boldsymbol{K}_n + \tau^2 \boldsymbol{I}_n\right)^{-1} \boldsymbol{z}_n}{\mathbf{1}_n^T \left(\boldsymbol{K}_n + \tau^2 \boldsymbol{I}_n\right)^{-1} \mathbf{1}_n}, \tag{1}$$

and one obtains (Roustant et al. 2012),

$$m_n(\boldsymbol{x}) = \widehat{\mu} + \boldsymbol{k}_n(\boldsymbol{x})^T \left(\boldsymbol{K}_n + \tau^2 \boldsymbol{I}_n\right)^{-1} (\boldsymbol{z}_n - \widehat{\mu}\mathbf{1}_n), \tag{2}$$

$$\begin{aligned} k_n(\boldsymbol{x}, \boldsymbol{x}') =& k(\boldsymbol{x}, \boldsymbol{x}') - \boldsymbol{k}_n(\boldsymbol{x})^T \left(\boldsymbol{K}_n + \tau^2 \boldsymbol{I}_n\right)^{-1} \boldsymbol{k}_n(\boldsymbol{x}') \\ &+ \frac{\left(1 - \mathbf{1}_n^T \left(\boldsymbol{K}_n + \tau^2 \boldsymbol{I}_n\right)^{-1} \boldsymbol{k}_n(\boldsymbol{x})\right)\left(1 - \mathbf{1}_n^T \left(\boldsymbol{K}_n + \tau^2 \boldsymbol{I}_n\right)^{-1} \boldsymbol{k}_n(\boldsymbol{x}')\right)}{\mathbf{1}_n \left(\boldsymbol{K}_n + \tau^2 \boldsymbol{I}_n\right)^{-1} \mathbf{1}_n}, \end{aligned} \tag{3}$$

where $\boldsymbol{k}_n(\boldsymbol{x}) = [k(\boldsymbol{x}, \boldsymbol{x}_i)]_{1\leq i\leq n}$ and $\boldsymbol{K}_n = [k(\boldsymbol{x}_i, \boldsymbol{x}_j)]_{1\leq i,j\leq n}$. When considering a Bayesian framework with an improper uniform prior for the mean μ and fixed kernel hyper-

parameters, the expression for the posterior mean coincides with the classical conditional expectation for Gaussians (see, e.g., Roustant et al. 2012).

3.2. KERNEL COMBINATION APPROACHES ON PRODUCT SPACES

First assume that valid kernels k_G and k_E are given on $\mathcal{X}_G$ and $\mathcal{X}_E$, respectively. We now discuss ways to combine them into valid kernels on $\mathcal{X}_G \times \mathcal{X}_E$. We are thus interested in combining kernels k_G and k_E into valid kernels k_{GE} on $\mathcal{X}_G \times \mathcal{X}_E$, taking their inputs in $(\mathcal{X}_G \times \mathcal{X}_E)^2 = (\mathcal{X}_G \times \mathcal{X}_E) \times (\mathcal{X}_G \times \mathcal{X}_E)$. A first important remark is that k_G and k_E can directly be used as kernels on the product space, in the sense that kernels (informally) defined by $k^G_{GE}((\boldsymbol{x}_G, \boldsymbol{x}_E), (\boldsymbol{x}'_G, \boldsymbol{x}'_E)) = k_G(\boldsymbol{x}_G, \boldsymbol{x}'_G)$ and $k^E_{GE}((\boldsymbol{x}_G, \boldsymbol{x}_E), (\boldsymbol{x}'_G, \boldsymbol{x}'_E)) = k_E(\boldsymbol{x}_E, \boldsymbol{x}'_E)$ are both valid on the product space. Either option would amount to ignore part of the variables, which is expected to be detrimental in prediction but would not alter the non-negative definiteness of resulting kernel matrices, thereby delivering valid kernels on $\mathcal{X}_G \times \mathcal{X}_E$. The combination approaches considered in this work essentially revolve around (blockwise) tensor sums and products. First, the tensor sum,

$$k^+_{GE}((\boldsymbol{x}_G, \boldsymbol{x}_E), (\boldsymbol{x}'_G, \boldsymbol{x}'_E) = k_G(\boldsymbol{x}_G, \boldsymbol{x}'_G) + k_E(\boldsymbol{x}_E, \boldsymbol{x}'_E),$$

defines a valid kernel $\mathcal{X}_G \times \mathcal{X}_E$, following the same mechanism as for usual additive kernels. This operation is immediately extended to non-negatively weighted sums. Beyond this,

$$k^\times_{GE}((\boldsymbol{x}_G, \boldsymbol{x}_E), (\boldsymbol{x}'_G, \boldsymbol{x}'_E) = k_G(\boldsymbol{x}_G, \boldsymbol{x}'_G) \times k_E(\boldsymbol{x}_E, \boldsymbol{x}'_E),$$

is also known to provide valid kernels on $\mathcal{X}_G \times \mathcal{X}_E$, in virtue of the stability of the cone of symmetric non-negative matrices by matrix product. Using again summation and non-negative linear combinations, we arrive at a class of kernels synthesizing and extending the latter approaches, whereby $\alpha, \beta, \gamma \geq 0$:

$$k^{\sim}_{GE}((\boldsymbol{x}_G, \boldsymbol{x}_E), (\boldsymbol{x}'_G, \boldsymbol{x}'_E) = \alpha k_G(\boldsymbol{x}_G, \boldsymbol{x}'_G) + \beta k_E(\boldsymbol{x}_E, \boldsymbol{x}'_E) + \gamma k_G(\boldsymbol{x}_G, \boldsymbol{x}'_G) \times k_E(\boldsymbol{x}_E, \boldsymbol{x}'_E). \tag{4}$$

A second remark is that taking a different (k_G, k_E) pair (e.g., with different hyperparameter values) in the sum and product parts of Equation (4) still leads to valid kernels on $\mathcal{X}_G \times \mathcal{X}_E$.

3.3. KERNELS ON G AND E

Commonly used families of covariance kernels for Euclidean input spaces include the isotropic Matérn kernels (Stein 2012). In this context, isotropy implies that the kernel values for pairs of locations depend solely on the Euclidean distance between them. For multidimensional inputs, anisotropic versions of these kernels are frequently employed. Here, we

consider the isotropic exponential and Gaussian kernel for $\boldsymbol{x}, \boldsymbol{x}' \in \mathbb{R}^{\nu}$,

$$k_{\text{EXP}}(\boldsymbol{x}, \boldsymbol{x}') = \exp\left(-\frac{||\boldsymbol{x} - \boldsymbol{x}'||}{\theta}\right), \quad k_{\text{GAU}}(\boldsymbol{x}, \boldsymbol{x}') = \exp\left(-\frac{||\boldsymbol{x} - \boldsymbol{x}'||^2}{\theta^2}\right),$$

with $||\boldsymbol{x} - \boldsymbol{x}'|| = \sqrt{(\boldsymbol{x} - \boldsymbol{x}')^\top (\boldsymbol{x} - \boldsymbol{x}')}$ and the correlation length $\theta > 0$.

On the genetic covariates G, we consider an Exponential–Hamming kernel ($k_{\text{G:EXP-HAM}}$). The Hamming distance for string sequences $\boldsymbol{x}_{\boldsymbol{G}}, \boldsymbol{x}_{\boldsymbol{G}}{}' \in \mathcal{X}_{\mathcal{G}}$ is given by,

$$d_{\text{H}}(\boldsymbol{x}_{\boldsymbol{G}}, \boldsymbol{x}'_{\boldsymbol{G}}) = \frac{1}{\nu_{\text{G}}} \sum_{i=1}^{\nu_{\text{G}}} \mathbb{1}(\boldsymbol{x}_{\boldsymbol{G}}[i] \neq \boldsymbol{x}'_{\boldsymbol{G}}[i]),$$

and can be used in the exponential kernel k_{EXP} instead of the Euclidean distance. The Hamming distance, which counts the number of differing entries between two vectors, can be viewed as the L_∞ distance. Hutter et al. (2014) introduced a generalized (weighted) version of this kernel and demonstrated its positive definiteness. While this result applies to the exponential-Hamming kernel, it does not hold for the Gaussian–Hamming kernel (following a foundational theorem by Schoenberg, see Berg et al. 1984). As a second kernel on G, we consider the traditional Gaussian-GBLUP kernel ($k_{\text{G:GAU-GBLUP}}$). In this approach, the SNP data are numerically encoded based on the major allele frequency. A Gaussian kernel is then applied to the resulting Euclidean distance matrix, following the methodology described by Cuevas et al. (2016) and Costa-Neto et al. (2021). On the environmental covariates E, we test both an exponential-Euclidean ($k_{\text{E:EXP-EUCL}}$) and a Gaussian–Euclidean kernel ($k_{\text{E:GAU-EUCL}}$), where we directly apply the exponential and Gaussian kernel to the Euclidean distances of the environmental vectors.

We have also tested alternative kernels for non-Euclidean spaces. However, as they did not significantly improved or even deteriorated the results, the results are not shown here.

3.4. LINK TO RKHS REGRESSION AND KERNEL LMMS

Given the same setup as in Sect. 3.1 with n observation pairs, we consider the regression model $Z_i \sim \mathcal{N}\left(f(\boldsymbol{x}_i), \tau^2\right)$. To infer the function f, it is necessary to specify a set or space of candidate functions. In reproducing kernel Hilbert spaces (RKHS) regression (Gianola et al. 2006; Montesinos-López et al. 2022), the space of candidate functions $\mathcal{H}_k$ is given by a RKHS defined by a kernel k. By the representer theorem (Wahba 1990), the resulting solution admits a linear representation with kernel functions centered at the training points, $f(\boldsymbol{x}_i) = m + \boldsymbol{k}_n(\boldsymbol{x}_i)^T \boldsymbol{\beta}$, with $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_n)^T$ a vector of coefficients. To estimate $\boldsymbol{\beta}$ and m employing frequentist methods, RKHS regression relies on the negative log-likelihood while controlling the complexity of the function by adding the squared norm in $\mathcal{H}_k$, $||f||^2_{\mathcal{H}_k} = \boldsymbol{\beta}^T \boldsymbol{K}_n \boldsymbol{\beta}$. Using the Gaussian assumption of the noise for the log-likelihood, this leads to a minimization of,

$$(\boldsymbol{Z}_n - \mathbf{1}_n \mu - \boldsymbol{K}_n \boldsymbol{\beta})^T (\boldsymbol{Z}_n - \mathbf{1}_n \mu - \boldsymbol{K}_n \boldsymbol{\beta}) + \lambda \boldsymbol{\beta}^T \boldsymbol{K}_n \boldsymbol{\beta}, \tag{5}$$

employing the regularization parameter λ. Plugin in the BLUE $\widehat{\mu}$ (see Sect. 3.1), and minimizing over $\boldsymbol{\beta}$, we obtain $\widehat{\boldsymbol{\beta}} = (\boldsymbol{K}_n + \lambda \boldsymbol{I_n})^{-1} \left(\boldsymbol{Z}_n - \hat{\mu}\mathbf{1}_n\right)$. This solution to the beta coefficients is known as kernel ridge regression (Hastie et al. 2009; Shawe-Taylor and Cristianini 2004). For predicting in an input $\boldsymbol{x}$, we can then use,

$$\widehat{z} = \hat{\mu} + \boldsymbol{k}_n(\boldsymbol{x})^T \widehat{\boldsymbol{\beta}} = \hat{\mu} + \boldsymbol{k}_n(\boldsymbol{x})^T \left(\boldsymbol{K}_n + \lambda \boldsymbol{I_n}\right)^{-1} \left(\boldsymbol{z}_n - \widehat{\mu}\mathbf{1}_n\right),$$

which corresponds to the posterior mean of the GP in Equation (2) if $\lambda = \tau^2$ (Rasmussen and Williams 2006). The regularization approach also incorporates smoothness assumptions concerning f, but is typically not used to obtain probabilistic predictions as in GP modeling. However, Bayesian perspectives on RKHS regression have been proposed, e.g., with independent priors on $\boldsymbol{\varepsilon} \sim \mathcal{N}\left(0, \tau^2 \boldsymbol{I_n}\right)$ and $\boldsymbol{\beta} \sim \mathcal{N}\left(0, \boldsymbol{K}_n^{-1}\right)$ (de Los Campos et al. 2010). Under such assumptions, the joint distribution of $\boldsymbol{K}_n\boldsymbol{\beta}$ and $\boldsymbol{z}_n$ is Gaussian, the conditional distribution is Gaussian as well, and we recover the simple kriging equations (for fixed μ). When also employing an improper uniform prior for the mean m, we recover the ordinary kriging equations (Eq. 2). Connections between RKHS regression and Gaussian processes were made as early as in Kimeldorf (1970). In a full Bayesian approach for RKHS regression, also priors for the unknown hyperparameters are used.

Let us now turn to mixed-effects modeling and consider the LMM specified by

$\boldsymbol{Z}_n = \mathbf{1}_n\, m + \boldsymbol{M}\, \boldsymbol{u} + \boldsymbol{\varepsilon}$, using the random effect $\boldsymbol{u} \sim \mathcal{N}\,(0, \boldsymbol{K}_v)$ with $\boldsymbol{K}_v$ denoting the covariance matrix of all the possible G × E input combinations and $\boldsymbol{M} \in \mathbb{R}^{n\times}$ a design matrix encoding the combinations present in $\boldsymbol{Z}_n$. Furthermore, we assume a fixed intercept m and noise $\boldsymbol{\varepsilon} \sim \mathcal{N}\left(0, \tau^2 \boldsymbol{I_n}\right)$. When LMMs are fitted using a frequentist approach, estimates typically rely on Henderson's mixed model equations (Henderson 1950), which are based on a maximization of the joint density of $\boldsymbol{z}_n$ and $\boldsymbol{u}$ with respect to m and $\boldsymbol{u}$ (e.g., Meuwissen et al. 2001; Gianola et al. 2006; Buntaran et al. 2021). After estimating the noise variance and kernel parameters with ML or REML, this leads to the empirical BLUE for m and BLUP for $\boldsymbol{u}$ (e.g., Buntaran et al. 2021). To make predictions for new inputs, the matrix M can simply be adapted to the combinations of interest, with the estimates then plugged in. This leads to the same point predictions as with the posterior mean in GP modeling (ordinary kriging, Eqs. 2). In the same way, an LMM approach with fixed effects on covariates could be linked to a universal kriging approach (Matheron 1969; Handcock and Stein 1993). However, in contrast to BLUEs and BLUPs, GP models provide full predictive distributions. LMMs can also be implemented in a Bayesian framework, where all unknown quantities are treated as random variables with assigned prior distributions (Lindley and Smith 1972; Sorensen and Gianola 2007; Montesinos-López et al. 2016; Costa-Neto et al. 2021; Liu et al. 2025). Doing this, the Bayesian LMM methods typically rely on posterior sampling at some stage.

4. DATASET AND IMPLEMENTATION

4.1. WHEAT DATASET

The wheat dataset (Levy Häner et al. 2025) was collected in multienvironment trials conducted by Agroscope (the Swiss Federal Centre for Agricultural Research) and partners

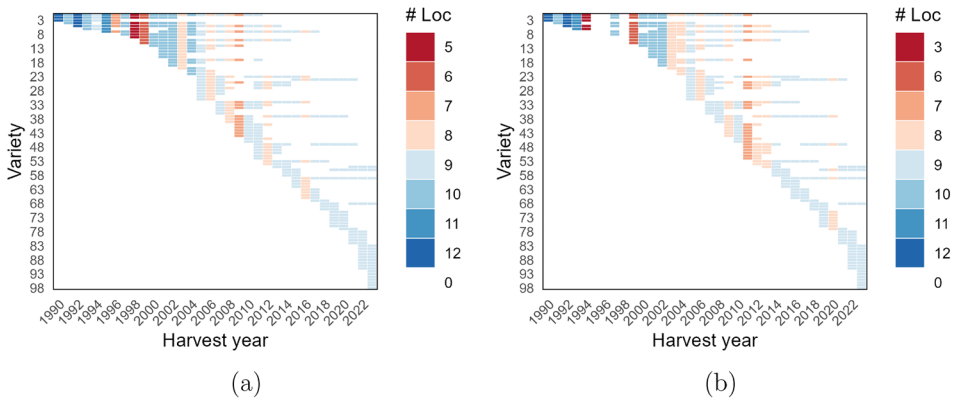

Figure 2. Number of locations tested for the different varieties in the different harvest years: **a** yield and **b** protein content.

as part of the official Swiss winter wheat variety trials for cultivation and use (Strebel et al. 2025). These trials were carried out across the Swiss wheat production zone (from western to eastern Switzerland, at an altitude between 390 and 640 MASL). The data include the performance traits yield (15% moisture content, dt per ha), used here as an average of three replicates per trial location and year, and grain protein content (%), measured on a mixture of the three replicates per trial location and year. The trial design was a lattice when known. Data are available for sets of 3 to 25 varieties grown at 5–12 locations over the period 1990–2023 (Fig. 2). The total number of varieties is 98 and was limited by the availability of genetic marker data. To use the same dataset for predicting both yield and protein content, we included only records with measurements for both traits.

The term 'environment' is referring to a combination of trial location and harvest year; we have 263 environments available. Each environment is associated with meteorological data that are averaged over six periods covering the Swiss winter wheat growing season: winter (October-February, sowing and vernalization), March (tillering), April (stem elongation), May (heading-flowering), June (flowering and seed filling), and July (harvest). Daily minimum, maximum and mean temperatures, total and average daily precipitation, and monthly dry days were extracted for each location from the MeteoSwiss spatial climate analysis products (1 km grid; MeteoSwiss 2024b). Daily solar radiation was obtained from the nearest automated meteorological station (MeteoSwiss 2024a).

Genetic information consists of 12106 SNPs (single-nucleotide polymorphisms) marker data. The data were characterized by a SNP call frequency above 80%, a frequency of homozygous calls above 90%, and a sample call rate above 90%. The SNPs are in IUPAC notation with 'G,' 'A,' 'T,' and 'C' representing guanine, adenosine, thymine, and cytosine bases, and 'K,' 'M,' 'R,' and 'Y' representing combinations of these bases as specified in Cornish-Bowden (1985). For the GBLUP kernel, bi-allelic marker data are employed with homogeneous 'G = GG,' 'A =AA,' 'T=TT,' 'C=CC' and heterogeneous 'K=GT,' 'M=AC,' 'R=AG,' 'Y=CT' (Cuevas et al. 2016; Costa-Neto et al. 2021).

4.2. MODEL EVALUATION

For evaluation, data are split into training $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ and test set $\boldsymbol{x}_{n+1}, \ldots, \boldsymbol{x}_m$ and we perform cross-validation on 80 % of the data using 30 splits. We evaluate predictive performance in two scenarios: new environment and new genotype. Group-based splitting, by environment or genotype, prevents information from leaking into the training. Here, leakage refers to situations where observations from the same group appear in both training and test set, allowing the model to exploit within-group dependence. We also consider controlled leakage, allowing one observation in the training set to mimic minimal prior knowledge from each target genotype or environment. For a more detailed discussion of data leakage and its implications in group-based cross-validation, we refer to Guignard et al. (2024). While traditional leave-one-out approaches exclude and predict individual environments or genotypes (e.g., Tadese et al. 2024), our approach predicts multiple targets while keeping the number of training data roughly consistent.

To assess the predictive accuracy of the GP, we first use the mean squared error (MSE). This metric assesses the mean $m_n(\boldsymbol{x})$ as a predictor, ignoring full probabilistic performance. To capture the probabilistic aspect, the Gaussian predictive distribution must be compared to a single observed value. This can be achieved through the use of a scoring rule (Gneiting and Raftery 2007). Scoring rules assess performance by assigning lower scores to 'better' probabilistic predictions, ideally capturing both calibration (agreement with observed outcomes) and sharpness (concentration of the predictive distribution). We consider the Continuous Ranked Probability Score (CRPS; Sanders 1963; Murphy 1973), defined as $\mathrm{CRPS}(F, y) = \int_{-\infty}^{\infty} [F(u) - \mathbb{1}\{y \leq u\}]^2 \, \mathrm{d}u$, where $F(\cdot)$ is the cumulative distribution function of the predictive distribution (Gneiting and Raftery 2007). For a deterministic point predictor, where the predictive distribution reduces to a Dirac delta, the CRPS simplifies to the absolute difference between the predicted and the true value. For evaluation and comparison purposes, we compute the median CRPS for the test set. As a second scoring rule, we consider the log-score (logS; Good 1952), $\mathrm{logS}(f, y) = -\log p(y)$, where $p(\cdot)$ is the probability density function of the predictive distribution. Again, we consider the median score over the test set. The GP acts as a predictor for the noiseless function f. To evaluate it using noisy data, we rely on the observational GP model, in which the predicted variance is extended by τ^2.

4.3. IMPLEMENTATION AND COMPARISON METHODS

We implemented the GP model manually and the R code used in this study can be found in a GitHub (WheatGP 2025). We denote the GPs with the different kernel combinations and choices by:

- GP^{G} for $k_{\mathrm{GE}}^{\mathrm{G}}$, GP^{E} for $k_{\mathrm{GE}}^{\mathrm{E}}$,
- GP^{+} for k_{GE}^{+},
- $\mathrm{GP}^{\times}$ for $k_{\mathrm{GE}}^{\times}$,

- GP$^{\sim}$ for $k_{\mathrm{GE}}^{\sim}$,
- GP$_1$: $k_{\mathrm{E:GAU-EUCL}}$, $k_{\mathrm{G:GAU-GBLUP}}$,
- GP$_2$: $k_{\mathrm{E:GAU-EUCL}}$, $k_{\mathrm{G:EXP-HAM}}$,
- GP$_3$: $k_{\mathrm{E:EXP-EUCL}}$, $k_{\mathrm{G:GAU-GBLUP}}$,
- GP$_4$: $k_{\mathrm{E:EXP-EUCL}}$, $k_{\mathrm{G:EXP-HAM}}$.

A key implementation step is estimating the hyperparameters. Recall the kernel sum and product combinations on the product space (Sect. 3.2). To avoid over-parameterization, we constrain (α, β, γ) to sum to one. Furthermore, we consider one hyperparameter per kernel, denoted as θ_{E} for the environmental kernel and θ_{G} for the genetic kernel. For the variance, we distinguish between the total variance of the response denoted as ν, the noise variance τ^2, and the proportion of ν explained by the GP denoted by ς. Following Roustant et al. (2012)'s Appendix A2, ν can be estimated in closed form as a function of $(\theta_{\mathrm{E}}, \theta_{\mathrm{G}}, \alpha, \beta, \gamma, \varsigma)$, leading to five parameters for $k_{\mathrm{GE}}^{\sim}$, four for k_{GE}^{+} $(\gamma = 0)$, and three for k_{GE}^{G} $(\beta = \gamma = 0)$, k_{GE}^{E} $(\alpha = \gamma = 0)$, and $k_{\mathrm{GE}}^{\times}$ $(\alpha = \beta = 0)$. Hyperparameters are tuned via maximum likelihood (empirical Bayes). A coarse grid search provides first estimates to initialize the Adam optimizer (Kingma and Ba 2014) with a learning rate of 0.01 for 1000 epochs. In order to significantly decrease computation time, we use batching of size 0.5, computing the mean gradient from two randomly selected, disjoint batches, each the size of half the training dataset.

Relying on Sect. 3.4, we compare the GP with the corresponding implementations of RKHS regression and kernel LMMs. For RKHS regression, we rely on the R-package for Bayesian generalized linear regression (BGLR; Pérez and de Los Campos 2014), which performs a fully Bayesian analysis and samples the hyperparameters, effects, and predictions with Gibbs. For the kernel LMM, we employ the rrBLUP package (Endelman 2011) as suggested in López et al. (2022). We rely on the same notation with superscripts and subscripts as for the GP models, and we use the ML fit of the GP model for the hyperparameters. We have compared the computation times of the different methods and observed that most of the total runtime is consumed by the hyperparameter optimization, on which all methods rely. With regard to prediction, the GP and the kernel LMM have comparable computational times, whereas RKHS regression takes longer (see WheatGP 2025).

5. RESULTS

5.1. NEW ENVIRONMENT

In this section, we are interested in predicting for a new environment. We begin by evaluating GP$^{\sim}$ using the full kernel combination $k_{\mathrm{GE}}^{\sim}$ (Eq. 4) in conjunction with the four different kernel configurations introduced in Sect. 4.3. Table 1 summarizes the results for yield and protein content by depicting the median MSE, CRPS, and logS. First, we consider the scenario without leakage (left of the vertical bars). For both traits and with respect to all three metrics, the GPs using the exponential-Euclidean kernel for the environmental information (GP$_3^{\sim}$, GP$_4^{\sim}$) outperform the ones employing the Gaussian–Euclidean (GP$_1^{\sim}$, GP$_2^{\sim}$)

Table 1. Model evaluation results for predictions in a new environment without | with controlled leakage

Method	MSE Y	CRPS Y	logS Y	MSE P	CRPS P	logS P
$GP_1^{\sim}$	127.95 | 43.33	6.44 | 3.82	3.90 | 3.41	1.54 | 0.56	0.70 | 0.44	1.69 | 1.23
$GP_2^{\sim}$	128.15 | 41.95	6.45 | 3.77	3.90 | 3.40	1.54 | 0.57	0.70 | 0.44	1.69 | 1.23
$GP_3^{\sim}$	119.72 | 41.16	6.22 | 3.73	3.87 | 3.39	1.43 | 0.55	0.68 | 0.43	1.65 | 1.22
$GP_4^{\sim}$	119.73 | 41.13	6.22 | 3.73	3.87 | 3.39	1.43 | 0.55	0.68 | 0.43	1.65 | 1.22
GP_4^{+}	125.25 | 45.71	6.44 | 3.91	3.91 | 3.44	1.44 | 0.58	0.69 | 0.45	1.68 | 1.25
$GP_4^{\times}$	119.84 | 42.55	6.14 | 3.78	3.84 | 3.38	1.48 | 0.58	0.68 | 0.44	1.62 | 1.22
GP_4^{G}	147.22 | 147.76	6.87 | 6.90	3.94 | 3.94	1.72 | 1.73	0.75 | 0.75	1.71 | 1.71
GP_4^{E}	160.06 | 83.91	7.33 | 5.36	4.04 | 3.74	2.07 | 1.44	0.84 | 0.70	1.89 | 1.69
$LMM_4^{\sim}$	119.73 | 41.00	8.47 | 4.95	–	1.43 | 0.55	0.95 | 0.58	–
$BGLR_4^{\sim}$	123.20 | 41.21	6.15 | 3.55	3.83 | 3.31	1.43 | 0.56	0.68 | 0.42	1.61 | 1.18

Y represents yield (dt per ha) and P represents grain protein content (%). The analysis is based on 30 different train/test splits, and we show the median of the obtained metrics

kernel. Among the models $GP_3^{\sim}$ and $GP_4^{\sim}$, we observe minimal performance differences between the genetic kernels. When allowing one controlled leakage point per environment (right of the vertical bars), the scores improve drastically. In this scenario, the kernels show generally a more similar performance. However, the first kernel ($GP_1^{\sim}$) performs worse than the others in regard to yield.

Next, we compare the different kernel combination strategies within GP^{G}, GP^{E}, GP^{+}, $GP^{\times}$, and $GP^{\sim}$ (Table 1). Here, we focus on GP_4, which demonstrated convincing performance in the first comparison of the kernels within $GP^{\sim}$. We observe that without leakage for yield, $GP_4^{\times}$ and $GP_4^{\sim}$ perform similarly, while for protein content, GP_4^{+} exhibits comparable performance as the full model $GP_4^{\sim}$. However, for both traits and the case with one leakage observation, the full model $GP_4^{\sim}$ performs generally better than $GP_4^{\times}$ and GP_4^{+}. For the setting without leakage, the GP using the genetic kernel only (GP_4^{G}) performs better than the one using the environmental kernel only (GP_4^{E}). When considering a single controlled leakage observation of the target environment, the performance of the environmental kernel only (GP_4^{E}) improves and becomes clearly better than that of the genetic only. Nevertheless, the MSE is still about two times (yield) /three times (protein content) higher than the one of the full model.

To gain a deeper understanding of the inner workings of the GP approaches, Fig. 3 presents a boxplot of the estimated hyperparameters for the different kernel combinations. We only show the hyperparameters for the scenario targeting yield without leakage, the remaining scenarios can be viewed on GitHub (WheatGP 2025). We immediately observe that θ_G is estimated at values ranging from 1.5 to 3.2, while θ_E is below 1, suggesting that the ML approach aims to flatten the effect of the genetic kernels. We observe that the full model $GP_4^{\sim}$ assigns about $\alpha = 0.35$ to the genetic, $\beta = 0.45$ to the environmental and $\gamma = 0.2$ to the product kernel. The proportion of variance explained by the GP, denoted by ς, is estimated similarly across all methods as above 95 %. For protein content, the full model gives more equal weight to the environmental and genetic kernel (WheatGP 2025). These observations do not drastically change when adding one leakage observation.

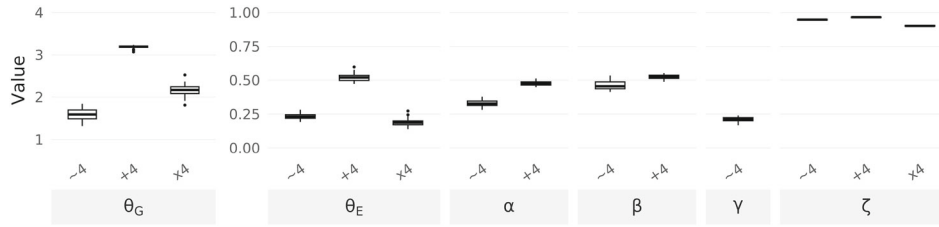

Figure 3. Estimated hyperparameters for the different GP models, fitted using the Adam optimizer. The boxplot shows the values across 30 different train/test splits for a new environment, without leakage, and for the trait yield.

Table 1 also shows the scores of $\text{LMM}_4^{\sim}$ and $\text{BGLR}_4^{\sim}$. As expected, $\text{LMM}_4^{\sim}$ and $\text{GP}_4^{\sim}$ yield very similar results in MSE for both traits and leakage scenarios. The MSE values obtained with $\text{BGLR}_4^{\sim}$ are also in the same range but slightly higher, potentially a result from the full Bayesian approach also sampling the hyperparameters. The CRPS and logS values of $\text{BGLR}_4^{\sim}$ are more aligned with the ones of $\text{GP}_4^{\sim}$. Note that the LMMs traditionally do not account for uncertainty, so the logS is missing, and the CRPS values based on the MAE are worse than those of models that explicitly model uncertainty.

5.2. NEW GENOTYPE

In this section, we consider prediction for a new genotype. Table 2 reports the same assessment metrics as in the previous section. We observe that performance in the no-leakage case is substantially better than in the setting targeting a new environment and the difference between leakage and no leakage is less pronounced. The scores for $\text{GP}^{\sim}$ using the four kernels show that the difference between the kernels is minimal. We compare again the different kernel combination approaches, focusing on GP_4. We observe only small difference between GP^{+}, $\text{GP}^{\times}$, and $\text{GP}^{\sim}$. However, as in the scenario involving a new environment, we observe that $\text{GP}_4^{\times}$ and $\text{GP}_4^{\sim}$ perform better than GP_4^{+} in terms of yield without leakage, with $\text{GP}_4^{\times}$ performing best in the presence of leakage. GP^{G} performs very poorly in this setting for both traits, likely due to the lack of training data of the same genotype. For yield, GP^{E} performs remarkably well. As before and as expected, $\text{LMM}_4^{\sim}$ and $\text{GP}_4^{\sim}$ produce similar results in MSE. Also here, $\text{BGLR}_4^{\sim}$ is worse in MSE but comparable in CRPS and logS.

6. PROOF OF CONCEPT: BAYESIAN OPTIMIZATION

In this section, we demonstrate with a proof of concept how the GP model could potentially be leveraged for decision-making and data acquisition. GP models have been used to optimize functions based on queries and sequential model updates (Kushner 1964) in the framework now known as Bayesian optimization (BO). BO typically relies on an acquisition function to decide where to evaluate f next. A very common acquisition function is the expected improvement (EI),

$$\text{EI}_n(\mathbf{x}) = (m_n(\mathbf{x}) - f_n^{\star})\Phi\left(u_n(\mathbf{x})\right) + \sqrt{k_n(\mathbf{x},\mathbf{x})}\,\phi(u_n(\mathbf{x}))\,,$$

Table 2. Model evaluation results for predictions for a new genotype without | with one controlled leakage point

Method	MSE Y	CRPS Y	logS Y	MSE P	CRPS P	logS P
$GP_1^{\sim}$	51.32 | 39.01	4.07 | 3.61	3.45 | 3.35	0.65 | 0.48	0.47 | 0.41	1.31 | 1.16
$GP_2^{\sim}$	52.35 | 39.25	4.09 | 3.62	3.46 | 3.35	0.65 | 0.48	0.47 | 0.41	1.31 | 1.16
$GP_3^{\sim}$	52.43 | 39.10	4.11 | 3.61	3.46 | 3.35	0.66 | 0.49	0.47 | 0.41	1.31 | 1.16
$GP_4^{\sim}$	51.50 | 38.97	4.10 | 3.62	3.46 | 3.35	0.66 | 0.49	0.47 | 0.41	1.31 | 1.16
GP_4^{+}	53.55 | 39.93	4.19 | 3.65	3.48 | 3.37	0.67 | 0.51	0.48 | 0.41	1.33 | 1.18
$GP_4^{\times}$	50.53 | 41.45	4.00 | 3.68	3.41 | 3.35	0.67 | 0.50	0.47 | 0.41	1.29 | 1.18
GP_4^{G}	162.32 | 160.36	7.34 | 7.27	4.04 | 4.02	2.10 | 1.97	0.84 | 0.82	1.88 | 1.84
GP_4^{E}	53.80 | 53.51	4.17 | 4.15	3.47 | 3.47	0.96 | 0.96	0.57 | 0.56	1.45 | 1.45
$LMM_4^{\sim}$	51.55 | 38.83	5.59 | 4.81	–	0.66 | 0.49	0.65 | 0.55	–
$BGLR_4^{\sim}$	54.60 | 39.15	4.09 | 3.48	3.45 | 3.31	0.66 | 0.48	0.45 | 0.39	1.21 | 1.11

Y represents yield (dt per ha) and P represents grain protein content (%). The analysis is based on 30 different train/test splits, and we show the median of the obtained metrics

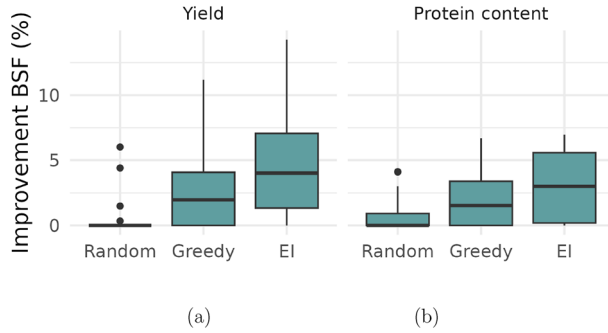

Figure 4. Improvement of BSF values (in %) when starting with 10% of the data from a small subset of environments and incrementally adding 50 training points selected using the different criteria for **a** yield and **b** protein content. The boxplots summarize the results of 30 runs.

where $u_n(\mathbf{x}) = (m_n(\mathbf{x}) - f_n^\star)/\sqrt{k_n(\mathbf{x},\mathbf{x})}$ with the best value observed so far $f_n^\star$, and ϕ denote the Gaussian cumulative and probability density function, respectively. In each step, BO samples the point $\mathbf{x}$ that maximizes the $\mathrm{EI}_n(\mathbf{x})$ and then updates the GP model with the new data. When observations are noisy, $f_n^\star$ is latent and can be estimated by the maximum of the current posterior mean $f_n^\star := \max_{1\le i\le n} m_n(\mathbf{x}_i)$ (see also Picheny et al. 2013, for more options in the noisy case).

We now test this on our wheat dataset. The changing set of varieties over the years (Fig. 2) makes chronological evaluation difficult, but the GP model does not use site or year identifiers and it relies solely on meteorological variables, so temporal order is not considered. We examine a test case in which we start with 10% of the data from a small subset of environments (meteorological scenarios). The remaining data serve as the candidate set from which we sequentially select 50 additional combinations via EI maximization. To ensure verifiability, we restrict the analysis to observed data only. This could mirror a decision-making problem in which we must select promising variety × environment combinations across different trial sites with distinct meteorology and regional varieties. We add candidates one by one via EI maximization and compare the resulting designs

against a random baseline. To demonstrate the benefits of the GP's closed-form predictive distribution, we also evaluate a greedy posterior mean policy that consistently selects the combination with the highest predicted mean (which could also be implemented using an LMM). We are interested in the best-so-far (BSF) values, i.e., the highest trait values observed up to each step. Given noisy observations, we quantify the BSF values using the posterior mean of a GP fitted to the full dataset for the final evaluation. Figure 4 shows the percentage improvements in BSF values after adding 50 points, relative to the initial BSF values, for (a) yield and (b) protein content. We find that adding points at random only provides modest gains, particularly for yield. On the other hand, EI consistently outperforms the greedy strategy, achieving roughly twice the median improvement.

7. DISCUSSION AND CONCLUSIONS

Optimizing multienvironment trials and recommending varieties based on the local environment are two very common themes in agricultural research. This paper introduces a GP modeling approach for G $\times$ E prediction, which is closely connected to kernel-based LMM and kernel regression methods. However, a key advantage of the GP formulation is that it provides closed-form probabilistic predictions, opening the door to optimal decision-making and data acquisition strategies. We hope this work provides a useful proof of concept showing that the GP framework offers a promising foundation for future applications.

There remains room to refine the kernel choices used in our GP models. In our experiments, exponential outperformed Gaussian kernel approaches. In contrast, more specialized kernels for sequences or time series (such as the global alignment and spectrum kernels) did not improve performance in our setting (results not shown). Other kernel families remain promising for further study. In general, the genotype kernels performed poorly when predicting the yield for new genotypes. As shown in Table 2, the environmental-only model GP^E was not substantially improved by adding genetic information in $GP^{\sim}$. However, when considering leakage observations from the target genotypes, we did observe an improvement. This indicates, that for yield, the considered genetic kernels effectively only 'work' for the same genotype, meaning that their measure of genetic similarity is not very effective in-between genotypes. On the other hand, for protein content, adding genetic information in $GP^{\sim}$ clearly improved the performance, in both leakage and no-leakage scenarios.

The environmental kernels performed better overall, and adding environmental information improved performance for both traits across all testing scenarios (Tables 1 and 2). However, including one leakage observation from the same environment substantially improved prediction accuracy, highlighting the importance of site- and year-specific attributes such as management and soil characteristics for wheat performance. This effect of local information is also evident in the stronger prediction results for new genotypes (Table 2), compared to prediction for new environments (Table 1). To mitigate this effect, the inclusion of additional covariates capturing soil characteristics and management practices, as suggested by Buntaranet al. (2021) and Canella Vieira et al. (2022), may prove beneficial. Such information could also improve predictions for new environments without leakage, including future climate scenarios for which no prior observations are available.

In Sect. 6, we present a simple proof of concept illustrating how GP models can support decision-making and guide data acquisition. We show that the EI criterion (Močkus 1974; Jones et al. 1998) identifies more promising input combinations than a greedy strategy (Fig. 4). The sequential one-by-one acquisition could readily be replaced by batch selection (e.g., Marmin et al. 2015), and many alternative criteria or multiobjective strategies are available (e.g., Chevalier et al. 2014; Cole et al. 2023; Friedli et al. 2025). Finally, we draw attention to the potential of GP models to jointly model yield and protein content, enabling local exploration of their relationship.

ACKNOWLEDGEMENTS

This work was supported by Agroscope, swiss granum, the Swiss Federal Office for Agriculture [project 'Wheat Advisor', grant no. 19.03], the Schweizerischer Getreideproduzentenverband, Prometerre, Fresh Food & Beverage Group and Timac Agro Swiss. GP calculations were performed on UBELIX (https://www.id.unibe.ch/hpc), the HPC cluster at the University of Bern. Tim Steinert and David Ginsbourger acknowledge the support of the Digitization Commission (DigiK) of the University of Bern via the project "Perception in Statistics, Econometrics and Probability". David Ginsbourger would like to thank the Isaac Newton Institute for Mathematical Sciences, Cambridge, for support and hospitality during the program Representing, calibrating & leveraging prediction uncertainty from statistics to machine learning, where work on this paper was undertaken that was partially supported by EPSRC grant EP/Z000580/1 and by a grant from the Simons Foundation. Lea Friedli acknowledges financial support from the Swiss National Science Foundation under grant number 225353. Finally, we thank two anonymous reviewers and the editorial team for constructive reviews that helped to improve the clarity of the paper.

Funding Open Access funding enabled and organized by Projekt DEAL.

Data Availability The anonymised wheat dataset analysed during the current study has recently been published on the Zenodo repository (Levy Häner et al. 2025). The code can be found on a GitHub repository (WheatGp, 2025)

Declarations
Conflict of interest The authors report no conflict of interest relevant to this article.

[Received August 2025. Revised July 2026. Accepted July 2026.]

REFERENCES

Bandeira e Sousa M, Cuevas J, de Oliveira Couto EG et al. (2017). Genomic-enabled prediction in maize using kernel models with genotype×environment interaction. G3: Genes Genomes Genet 7(6):1995–2014. https://doi.org/10.1534/g3.117.042341

Berg C, Christensen JPR, Ressel P (1984) Harmonic analysis on semigroups: theory of positive definite and related functions. vol 100. Springer. https://doi.org/10.1007/978-1-4612-1128-0

Bernardo R et al (2002) Breeding for quantitative traits in plants, vol 1. Stemma press Woodbury, MN

Buntaran H, Forkman J, Piepho H-P (2021) Projecting results of zoned multi-environment trials to new locations using environmental covariates with random coefficient models: accuracy and precision. Theor Appl Genet 134:1513–1530. https://doi.org/10.1007/s00122-021-03786-2

Burgueño J, de los Campos, G, Weigel K, Crossa J (2012) Genomic prediction of breeding values when modeling genotype×environment interaction using pedigree and dense molecular markers. Crop Sci 52(2):707–719. https://doi.org/10.2135/cropsci2011.06.0299

Canella Vieira C, Persa R, Chen P, Jarquin D (2022) Incorporation of soil-derived covariates in progeny testing and line selection to enhance genomic prediction accuracy in soybean breeding. Front Genet 13:905824. https://doi.org/10.3389/fgene.2022.905824

Chapman SC, Cooper M, Hammer GL, Butler D (2000). Genotype by environment interactions affecting grain sorghum. ii. frequencies of different seasonal patterns of drought stress are related to location effects on hybrid yields. Aust J Agric Res 51(2):209–222. https://doi.org/10.1071/AR99021

Chenu K, Chapman SC, Hammer GL, Mclean G, Salah HBH, Tardieu F (2008) Short-term responses of leaf growth rate to water deficit scale up to whole-plant and crop levels: an integrated modelling approach in maize. Plant Cell Environ 31(3):378–391. https://doi.org/10.1111/j.1365-3040.2007.01772.x

Chevalier C, Bect J, Ginsbourger D et al (2014) Fast parallel kriging-based stepwise uncertainty reduction with application to the identification of an excursion set. Technometrics 56(4):455–465. https://doi.org/10.1080/00401706.2013.860918

Cole DA, Gramacy RB, Warner JE, Bomarito GF, Leser PE, Leser WP (2023) Entropy-based adaptive design for contour finding and estimating reliability. J Qual Technol 55(1):43–60

Cooper M, DeLacy I (1994) Relationships among analytical methods used to study genotypic variation and genotype-by-environment interaction in plant breeding multi-environment experiments. Theor Appl Genet 88(5):561–572. https://doi.org/10.1007/BF01240919

Cooper M, Technow F, Messina C, Gho C, Totir LR (2016) Use of crop growth models with whole-genome prediction: application to a maize multienvironment trial. Crop Sci 56(5):2141–2156. https://doi.org/10.2135/cropsci2015.08.0512

Cornish-Bowden A (1985) Nomenclature for incompletely specified bases in nucleic acid sequences: rcommendations 1984. Nucleic Acids Res 13(9):3021–3030. https://doi.org/10.1093/nar/13.9.3021

Costa-Neto G, Fritsche-Neto R, Crossa J (2021) Nonlinear kernels, dominance, and envirotyping data increase the accuracy of genome-based prediction in multi-environment trials. Heredity 126(1):92–106. https://doi.org/10.1038/s41437-020-00353-1

Crossa J, Campos G, d. l., Pérez P, Gianola D, Burgueno J, Araus JL, Makumbi D, Singh RP, Dreisigacker S, Yan J, et al (2010) Prediction of genetic values of quantitative traits in plant breeding using pedigree and molecular markers. Genetics 186(2):713–724. https://doi.org/10.1534/genetics.110.118521

Crossa J, Montesinos-Lopez OA, Costa-Neto G, Vitale P, Martini JW, Runcie D, Fritsche-Neto R, Montesinos-Lopez A, Pérez-Rodríguez P, Gerard G et al (2025) Machine learning algorithms translate big data into predictive breeding accuracy. Trends Plant Sci 30(2):167–184. https://doi.org/10.1016/j.tplants.2024.09.011

Crossa J, Pérez-Rodríguez P, Cuevas J, Montesinos-López O, Jarquín D, De Los Campos G, Burgueño J, González-Camacho JM, Pérez-Elizalde S, Beyene Y et al (2017) Genomic selection in plant breeding: methods, models, and perspectives. Trends Plant Sci 22(11):961–975. https://doi.org/10.1016/j.tplants.2017.08.011

Cuevas J, Crossa J, Soberanis V et al. (2016). Genomic prediction of genotype× environment interaction kernel regression models. Plant Genome 9(3). https://doi.org/10.3835/plantgenome2016.03.0024

Cuevas J, Crossa J, Montesinos-López OA et al. (2017). Bayesian genomic prediction with genotype×environment interaction kernel models. G3: Genes Genomes Genet 7(1):41–53. https://doi.org/10.1534/g3.116.035584

Cuevas J, Montesinos-López O, Juliana P et al. (2019). Deep kernel for genomic and near infrared predictions in multi-environment breeding trials. G3: Genes Genomes Genet 9(9):2913–2924. https://doi.org/10.1534/g3.119.400493

de Los Campos G, Gianola D, Rosa GJ et al (2010) Semi-parametric genomic-enabled prediction of genetic values using reproducing kernel Hilbert spaces methods. Genet Res 92(4):295–308. https://doi.org/10.1017/S0016672310000285

Endelman JB (2011). Ridge regression and other kernels for genomic selection with R package rrBLUP. The Plant Genome 4(3). https://doi.org/10.3835/plantgenome2011.08.0024

Falconer D (1990) Selection in different environments: effects on environmental sensitivity (reaction norm) and on mean performance. Genet Res 56(1):57–70. https://doi.org/10.1017/S0016672300028883

Falconer DS, Mackay TFC (1996) Introduction to quantitative genetics. 4 edn. Pearson Prentice Hall, Harlow

Freeman G (1973) Statistical methods for the analysis of genotype-environment interactions. Heredity 31(3):339–354. https://doi.org/10.1038/hdy.1973.90

Friedli L, Gautier A, Broccard A, Ginsbourger D (2025) CRPS-based targeted sequential design with application in chemical space. Preprint at https://arxiv.org/abs/2503.11250

Garnett R (2023) Bayesian optimization. Cambridge University Press. https://doi.org/10.1017/9781108348973

Gianola D, Fernando RL, Stella A (2006) Genomic-assisted prediction of genetic value with semiparametric procedures. Genetics 173(3):1761–1776. https://doi.org/10.1534/genetics.105.049510

Ginsbourger D, Baccou J, Chevalier C et al (2014) Bayesian adaptive reconstruction of profile optima and optimizers. SIAM/ASA J Uncertain Quantif 2(1):490–510. https://doi.org/10.1137/130949555

Gneiting T, Raftery AE (2007) Strictly proper scoring rules, prediction, and estimation. J Am Stat Assoc 102(477):359–378. https://doi.org/10.1198/016214506000001437

Good IJ (1952) Rational decisions. J R Stat Soc Ser B 14(1):107–114. https://www.jstor.org/stable/2984087

Gregorius H-R, Namkoong G (1986) Joint analysis of genotypic and environmental effects. Theor Appl Genet 72(3):413–422. https://doi.org/10.1007/BF00288581

Griffiths R-R, Klarner L, Moss H et al (2023) Gauche: a library for gaussian processes in chemistry. Adv Neural Inf Process Syst 36:76923–76946

Guignard F, Ginsbourger D, Levy Häner L, Herrera JM (2024) Some combinatorics of data leakage induced by clusters. Stoch Environ Res Risk Assess. pp 1–14. https://doi.org/10.1007/s00477-024-02715-1

Handcock MS, Stein ML (1993) A Bayesian analysis of kriging. Technometrics 35(4):403–410. https://doi.org/10.1080/00401706.1993.10485354

Hastie T, Tibshirani R, Friedman J et al (2009). The elements of statistical learning data mining inference and prediction. https://doi.org/10.1007/978-0-387-84858-7

Henderson CR (1950) Estimation of genetic parameters. Ann Math Stat 21:309–310

Heslot N, Akdemir D, Sorrells ME, Jannink J-L (2014) Integrating environmental covariates and crop modeling into the genomic selection framework to predict genotype by environment interactions. Theor Appl Genet 127:463–480. https://doi.org/10.1007/s00122-013-2231-5

Hoerl AE, Kennard RW (1970) Ridge regression: biased estimation for nonorthogonal problems. Technometrics 12(1):55–67. https://doi.org/10.1080/00401706.1970.10488634

Hutter F, Xu L, Hoos HH et al (2014) Algorithm runtime prediction: methods & evaluation. Artif Intell 206:79–111. https://doi.org/10.1016/j.artint.2013.10.003

Janusevskis J, Le Riche R (2013) Simultaneous kriging-based estimation and optimization of mean response. J Glob Optim 55(2):313–336. https://doi.org/10.1007/s10898-011-9836-5

Jarquín D, Crossa J, Lacaze X et al (2014) A reaction norm model for genomic selection using high-dimensional genomic and environmental data. Theor Appl Genet 127:595–607. https://doi.org/10.1007/s00122-013-2243-1

Jones DR, Schonlau M, Welch WJ (1998) Efficient global optimization of expensive black-box functions. J Glob Optim 13(4):455–492. https://doi.org/10.1023/A:1008306431147

Kang M, Gorman D (1989) Genotype × environment interaction in maize. J Agron 81(4):662–664. https://doi.org/10.2134/agronj1989.00021962008100040020x

Kimeldorf G, Wahba G (1970) A correspondence between Bayesian estimation on stochastic processes and smoothing by splines. Ann Math Stat 41(2):495–502. https://www.jstor.org/stable/2239347

Kingma DP, Ba J (2014) Adam: a method for stochastic optimization. CoRR. arXiv:abs/1412.6980

Krige DG (1951) A statistical approach to some basic mine valuation problems on the witwatersrand. J S Afr Inst Min Metall 52(6):119–139

Kushner HJ (1964) A new method of locating the maximum point of an arbitrary multipeak curve in the presence of noise. J Basic Eng 86:97–106. https://api.semanticscholar.org/CorpusID:62599010

Levy Haner, L. and Strebel, S. and Visse-Mansiaux, M. and Wuyts, N. and Herrera, J. M. and Pellet, D. (2025) Winter wheat performance data for varieties grown under low input conditions in Switzerland over the period 1990–2023, together with trial site weather data and variety genetic marker data (1.0.0) [Data set]. https://doi.org/10.5281/zenodo.16420898

Lindley DV, Smith AF (1972) Bayes estimates for the linear model. J R Stat Soc Ser B Stat Methodol 34(1):1–18. https://www.jstor.org/stable/2985048

Liu J, Gock A, Ramm K et al (2025) Incorporating gene expression and environment for genomic prediction in wheat. Front Plant Sci 16:1506434. https://doi.org/10.3389/fpls.2025.1506434

Malosetti M, Bustos-Korts D, Boer MP, van Eeuwijk FA (2016) Predicting responses in multiple environments: issues in relation to genotype× environment interactions. Crop Sci 56(5):2210–2222. https://doi.org/10.2135/cropsci2015.05.0311

Marmin S, Chevalier C, Ginsbourger D (2015) Differentiating the multipoint expected improvement for optimal batch design. In: International workshop on machine learning, optimization and big data Springer, pp 37–48. https://doi.org/10.1007/978-3-319-27926-8_4

Matheron G (1969) *Le krigeage universel (Universal kriging)*, volume 1 of *Cahiers du Centre de Morphologie Mathématique*. Fontainebleau: École des Mines de Paris.https://doi.org/10.1016/j.geoderma.2003.08.018

Messina C, Hammer G, Dong Z, Podlich D, Cooper M (2009) Modelling crop improvement in a g×e×m framework via gene–trait–phenotype relationships. In: Crop physiology: applications for genetic improvement and agronomy. Academic Press, pp 235–265. https://doi.org/10.1016/B978-0-12-374431-9.00010-4

MeteoSwiss (2024a). Automatic measurement network. https://www.meteoswiss.admin.ch/weather/measurement-systems/land-based-stations/automatic-measurement-network.html

MeteoSwiss (2024b). Spatial climate analyses. https://www.meteoswiss.admin.ch/climate/the-climate-of-switzerland/spatial-climate-analyses.html

Meuwissen TH, Hayes BJ, Goddard M (2001) Prediction of total genetic value using genome-wide dense marker maps. Genetics 157(4):1819–1829. https://doi.org/10.1093/genetics/157.4.1819

Močkus J (1974) On Bayesian methods for seeking the extremum. In: IFIP Technical Conference on Optimization Techniques. Springer, pp 400–404. https://doi.org/10.1007/3-540-07165-2_55

Montesinos López OA, Montesinos López A, Crossa J (2022) Reproducing kernel Hilbert spaces regression and classification methods. In: Multivariate statistical machine learning methods for genomic prediction. Springer, pp 251–336.https://doi.org/10.1007/978-3-030-89010-0_8

Montesinos-López OA, Montesinos-López A, Crossa J et al. (2016) A genomic Bayesian multi-trait and multi-environment model. G3: Genes Genomes Genet 6(9):2725–2744.https://doi.org/10.1534/g3.116.032359

Montesinos-López OA, Montesinos-López A, Kismiantini n, Roman-Gallardo A, Gardner K, Lillemo M, Fritsche-Neto R, Crossa J (2022) Partial least squares enhances genomic prediction of new environments. Front Genet 13:920689. https://doi.org/10.3389/fgene.2022.920689

Murphy AH (1973) A new vector partition of the probability score. J Appl Meteorol Climatol 12(4):595–600. https://doi.org/10.1175/1520-0450(1973)012<0595:ANVPOT>2.0.CO;2

Pérez P, de Los Campos G (2014) Genome-wide regression and prediction with the BGLR statistical package. Genetics 198(2):483–495. https://doi.org/10.1534/genetics.114.164442

Picheny V, Wagner T, Ginsbourger D (2013) A benchmark of kriging-based infill criteria for noisy optimization. Struct Multidiscip Optim 48(3):607–626. https://doi.org/10.1007/s00158-013-0919-4

Piepho H-P, Denis J-B, van Eeuwijk FA (1998) Predicting cultivar differences using covariates. J Agric Biol Environ Stat, pp 151–162. https://doi.org/10.2307/1400648

Rafalski A (2002) Applications of single nucleotide polymorphisms in crop genetics. Curr Opin Plant Biol 5(2):94–100. https://doi.org/10.1016/S1369-5266(02)00240-6

Raffo MA, Sarup P, Andersen JR, Orabi J, Jahoor A, Jensen J (2022) Integrating a growth degree-days based reaction norm methodology and multi-trait modeling for genomic prediction in wheat. Front Plant Sci 13:939448. https://doi.org/10.3389/fpls.2022.939448

Ramakers JJ, Malik WA, Welcker C et al (2026) Evaluation and optimization of wheat and maize national variety evaluation systems in Europe. Field Crop Res 341:110420. https://doi.org/10.1016/j.fcr.2026.110420

Rasmussen CE, Williams CKI (2006) Gaussian processes for machine learning. The MIT Press. https://doi.org/10.7551/mitpress/3206.001.0001

Ray A (2021) Bayesian inversion using nested trans-dimensional Gaussian processes. Geophys J Int 226(1):302–326. https://doi.org/10.1093/gji/ggab114

Roustant O, Ginsbourger D, Deville Y (2012) DiceKriging, DiceOptim: Two R packages for the analysis of computer experiments by kriging-based metamodeling and optimization. J Stat Softw 51(1):1–55. https://doi.org/10.18637/jss.v051.i01

Sanders F (1963) On subjective probability forecasting. J Appl Meteorol Climatol 2(2):191–201. https://doi.org/10.1175/1520-0450(1963)002<0191:OSPF>2.0.CO;2

Shawe-Taylor J, Cristianini N (2004) Kernel methods for pattern analysis. Cambridge University Press. https://doi.org/10.1017/CBO9780511809682

Sorensen D, Gianola D (2007) Likelihood, Bayesian, and MCMC methods in quantitative genetics. Springer Science & Business Media. https://doi.org/10.1007/b98952

Stein ML (2012) Interpolation of spatial data: some theory for kriging. Springer Science & Business Media. https://doi.org/10.1007/978-1-4612-1494-6

Strebel S, Levy Häner L, Imhoff Y et al. (2025) Liste der empfohlenen Getreidesorten für die Ernte 2026. Agroscope Transfer, 591. https://link.ira.agroscope.ch/de-CH/publication/59585

Su G, Peng L, Hu L (2017) A Gaussian process-based dynamic surrogate model for complex engineering structural reliability analysis. Struct Saf 68:97–109. https://doi.org/10.1016/j.strusafe.2017.06.003

Tadese D, Piepho H-P, Hartung J (2024) Accuracy of prediction from multi-environment trials for new locations using pedigree information and environmental covariates: the case of sorghum (*Sorghum bicolor* (L.) Moench) breeding. Theor Appl Genet 137(8):181. https://doi.org/10.1007/s00122-024-04684-z

Technow F, Messina CD, Totir LR, Cooper M (2015) Integrating crop growth models with whole genome prediction through approximate Bayesian computation. PLoS ONE 10(6):e0130855. https://doi.org/10.1371/journal.pone.0130855

Tibshirani R (1996) Regression shrinkage and selection via the lasso. J R Stat Soc Ser B Stat Methodol 58(1):267–288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x

van Dijk ADJ, Kootstra G, Kruijer W, de Ridder D (2021) Machine learning in plant science and plant breeding. Iscience 24(1). https://doi.org/10.1016/j.isci.2020.101890

Van Eeuwijk F, Elgersma A (1993) Incorporating environmental information in an analysis of genotype by environment interaction for seed yield in perennial ryegrass. Heredity 70(5):447–457. https://doi.org/10.1038/hdy.1993.66

VanRaden P (2007) Genomic measures of relationship and inbreeding. Interbull Bull 37:33–33

Wahba G (1990) Spline models for observational data. SIAM DOI 10(1137/1):9781611970128

Washburn JD, Burch MB, Franco JAV (2020) Predictive breeding for maize: Making use of molecular phenotypes, machine learning, and physiological crop models. Crop Sci 60(2):622–638. https://doi.org/10.1002/csc2.20052

Washburn JD, Varela JI, Xavier A, et al. (2025) Global genotype by environment prediction competition reveals that diverse modeling strategies can deliver satisfactory maize yield estimates. Genetics 229(2):iyae195. https://doi.org/10.1093/genetics/iyae195

WheatGP (2025). GitHub. https://github.com/Timst98/Wheat-GP

Williams BJ, Santner TJ, Notz WI (2000) Sequential design of computer experiments to minimize integrated response functions. Stat Sin 10(4):1133–1152. http://www.jstor.org/stable/24306770

Woltereck R (1909) Weitere experimentelle Untersuchungen über Artveränderung, speziell über das Wesen quantitativer Artunterschiede bei Daphniden. Verhandlungen der Deutschen Zoologischen Gesellschaft 19:110–173. https://doi.org/10.1007/BF01876686