European Journal of Agronomy 159 (2024) 127227

Contents lists available at ScienceDirect

European Journal of Agronomy

journal homepage: www.elsevier.com/locate/eja

A Bayesian approach to analyzing long-term agricultural experiments

J.W.G. Addy [a,*,1], C. MacLaren [b,c], R. Lang [d]

[a] Intelligent Data Ecosystems, Rothamsted Research, Harpenden AL5 2JQ, United Kingdom
[b] Department of Crop Production Ecology, Swedish University of Agricultural Sciences, Uppsala 75007, Sweden
[c] International Maize & Wheat Improvement Centre (CIMMYT), Southern Africa Regional Office (SARO), P.O. Box MP163, Harare, Zimbabwe
[d] Department of Ecology, Swedish University of Agricultural Sciences, Uppsala 75007, Sweden

ARTICLE INFO

Keywords:
Bayesian Regression
Bayesian Multiple Regression
Hierarchical Models
Long-Term Experiments
Random Effects
Linear Models

ABSTRACT

Effective and flexible statistical analyses are key to getting the most out of long-term experiments (LTEs). Here, we aim to introduce Bayesian analysis to the wider LTE community and show how the modelling process differs from traditional statistical analyses. Bayesian methods have become increasingly popular due to more flexibility in model development with better access to statistical software and sampling algorithms. Using Bayes' Theorem, model coefficients are estimated by incorporating any prior knowledge we may have on model terms. Including prior knowledge in this way requires a different estimating procedure for a fitted model. Bayesian model coefficients are usually sampled from thousands of samples from one or more runs of a Markov Chain. We present the use of Bayesian analyses through three examples. Example 1 illustrates a single regression with and without factors using the Broadbalk Long-Term Experiment, showing how the estimated model changes with more uncertainty in our prior knowledge of model coefficients. Example 2 demonstrates the use of multiple regression, predicting grain yield from factor variables and seasonal weather variables. Example 3 shows an estimation of soil carbon changes under crop rotation and fertilization treatments with a hierarchical time series model using a Swedish soil fertility experiment.

1. Introduction

Long-term field experiments (LTEs) are important research resources in agriculture and ecology, enabling the study of slow processes and the robust distinction of treatment effects from background variability (Eckl and Piepho, 2015; Grosse et al., 2020; Onofri et al., 2016; Rasmussen et al., 1998; Richter and Kroschewski, 2006; Storkey et al., 2016; Payne, 2018). Given this potential of LTEs to make unique contributions to science, it is important that researchers are equipped with effective and flexible analysis methods to make the most of LTE datasets. The use of Bayesian methods has become increasingly popular due to advancements in statistical software and sampling algorithms. Bayesian methods make use of Bayes' Theorem to include prior information about model terms in the estimation of model parameters. Traditional methods of analysis which do not include prior information are called Frequentist methods. The use of prior information offers a level of model flexibility or parameter regulation that is not present in non-Bayesian (Frequentist) methods. An example of using Bayesian modelling to achieve robust results from a complex analysis of LTE data includes the direct modelling of a mean-variance function in forecasting future hay production from the Park Grass LTE (Addy et al., 2022).

However, many researchers working with LTEs may still be unaware of Bayesian methods and may not be aware of the flexibility these tools offer, nor know how to make use of them. There are various other teaching documents which explain the motivation and use of Bayesian statistics in other academic disciplines (Gelman et al., 2013, 2020; McElreath, 2018; van de Schoot et al., 2021). However, none so far explain the use of Bayesian methods for LTEs. To address this gap, this paper has two aims: (1) to introduce researchers to key Bayesian concepts involved in modelling LTE datasets, and (2) to provide examples of simple Bayesian analyses to enable researchers to see how Bayesian methods can be applied to a range of commonly encountered analysis scenarios with LTE datasets.

To understand Bayesian statistics, we first must understand traditional Frequentist statistical methods. In traditional statistical modelling we wish to find a series of model parameters which provide the best fit to our data. The classical Frequentist approach maximises the likelihood function to obtain the best-fitting model parameters. The likelihood

* Corresponding author.
E-mail address: John.Addy@bioss.ac.uk (J.W.G. Addy).
[1] Current address: Biomathematics and Statistics Scotland, Invergowrie, Dundee, DD2 5DA, UK

https://doi.org/10.1016/j.eja.2024.127227

Available online 21 June 2024

function can be thought of as the probability of our observed data given the parameters of the model, and by maximising the likelihood function we obtain the most probable model coefficients given the data we have measured. A Bayesian approach would combine any prior information we have about the model parameters with the model likelihood to regulate the model parameters. Prior information can take many forms. For example, when we know the mean of some measured data must lie within a given range, or if we know the relationship between two measured responses has a strong positive association. We can translate our prior information into prior distributions and combine this with the model likelihood (best fitting model to our data) and obtain a now regulated version of the likelihood function called the posterior distribution. The reason we wish to regulate our model parameters is to prevent a mis-specified model, dominated by incorrectly parameterised model terms. This often occurs when using traditional statistical methods to create models with a high number of estimable parameters (Tibshirani, 1996). The use of priors in Bayesian statistics is sometimes criticised because the choice of the prior distribution can be subjective and therefore the final model may not be as objective as a Frequentist analysis. However, for more complex models with many parameters, the

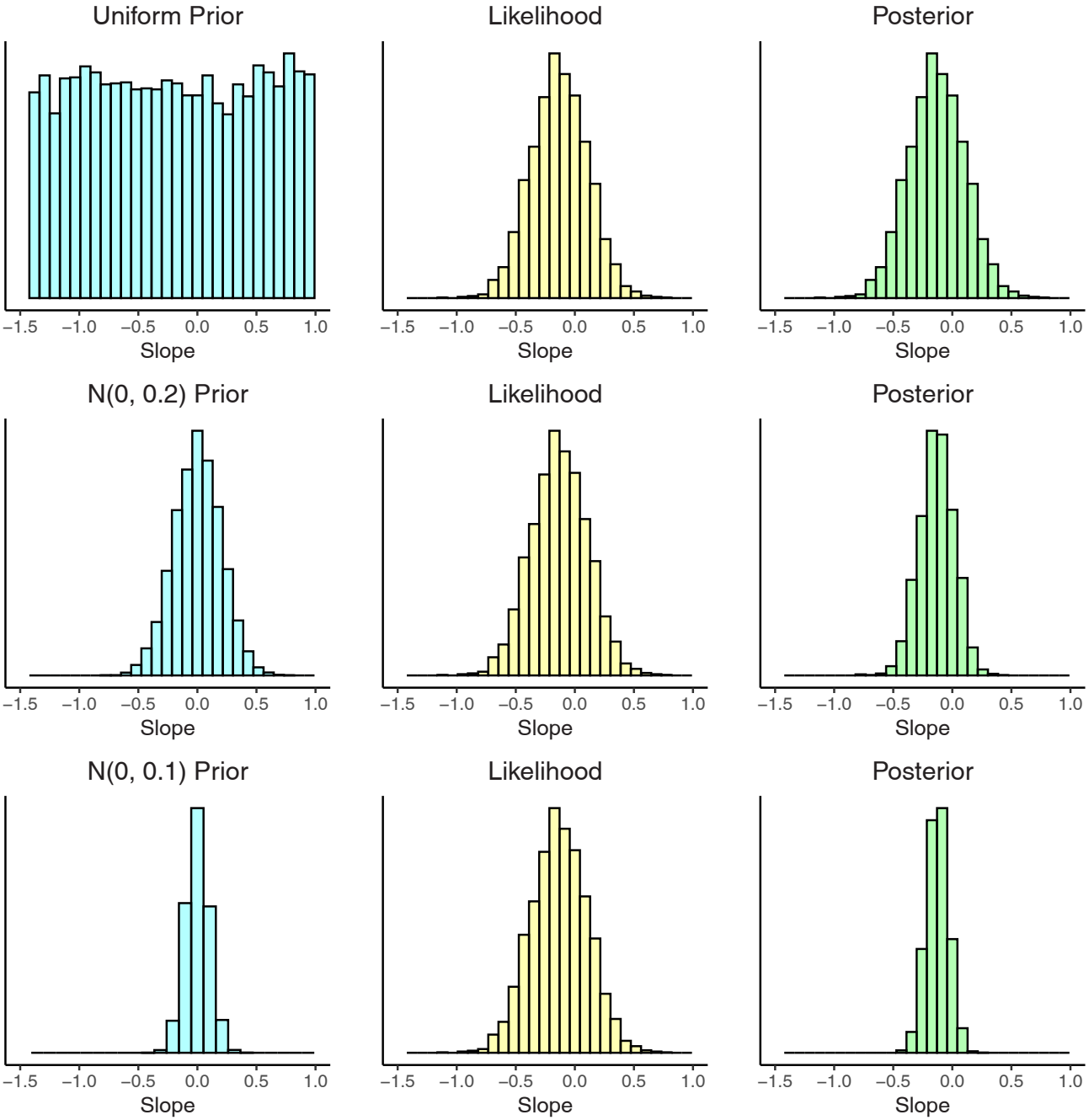

Fig. 1. The influence of different priors on the posterior of slope parameter for the simple regression given in Example 1. The left column shows different prior distributions for the slope parameter in Example 1a. The middle column shows the distribution of the slope parameter from the maximum likelihood. The right column shows the effect of different prior distributions on the posterior estimate of the slope parameter in Example 1a. The top row shows the effect of a Uniform prior on the model posterior. The second row shows the influence of a slightly informative Normal distribution prior with mean 0 and standard deviation 0.2 on the slope parameter in Example 1a. The bottom row shows the influence of a highly informative Normal distribution prior with mean 0 and standard deviation 0.1 on the slope parameter in Example 1a.

benefits of avoiding misspecification can outweigh the risk of reduced objectivity.

Including prior knowledge requires a different estimating procedure from traditional statistical models. Bayesian model coefficients and posterior distributions are usually sampled from thousands of runs of a plausible model via a Markov Chain, and most available statistical software includes very efficient Markov Chains (Neal, 2011). The process of Bayesian model fitting in this way is called the Bayesian workflow (Gabry et al., 2019), and rather than obtaining the most plausible model parameters for each model coefficient (these are called point estimates), we instead obtain thousands of model iterations which provide the estimated posterior distribution of model parameters (Fig. 1). From these iterations we can obtain standard deviations and 95 % credible intervals (this term is explained later) for each model parameter.

Here we first introduce readers to key concepts in Bayesian models via a discussion of the use of model likelihoods, priors, and posteriors, using a simple regression analysis predicting wheat yield (t ha^{-1}) from annual mean temperature (°C) (Perryman and Scott, 2020) using data from the Rothamsted Broadbalk wheat experiment (1968–2016) (Glendining and Poulton, 2023). We then follow with more complex analysis examples, firstly showing how to fit a Bayesian varying intercept model with factors such as wheat variety, and a multiple regression model where annual temperature in the Broadbalk example dataset is replaced by seasonal summaries of temperature (Addy, 2023). Mean annual and seasonal temperatures were used to predict Broadbalk's yields as an example of using Bayesian single and multiple regression, as there is a known relationship between yield and temperature (Addy et al., 2020), but please note that the analysis has been simplified for providing constructive examples (we do not for example explore concepts such as autoregression that may be relevant in a comprehensive analysis of these data - such as Macholdt et al. (2020)). We give a brief introduction and general overview of the methods involved in Bayesian model selection and comparison, such as shrinkage and model averaging. A final example is given on how the Bayesian workflow deals with random effects structures using data from the Swedish long-term soil fertility experiments. This paper is intended as an introduction to the use of Bayesian statistics for long-term experiments. When applying these methods, considerations on the design of the experiments used should be considered in the application of these methods. All examples have accompanying R code in the Supplementary Materials with data held in their respective repository.

2. Likelihood, priors, posteriors and credible interval

When modelling a straight line for non-Bayesian methods, we have a linear model with intercept β_0 and slope β_1,

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

Where y_i and x_i are the response and the explanatory variable for each observation i, and ε_i is the residual for each observation. The residual for each observation can be calculated as $\varepsilon_i = y_i - \beta_0 - \beta_1 x_i$. Given the residuals are Normally distributed, y is proportional to a Normal distribution with mean estimates equal to $\beta_0 + \beta_1 x_i$ and a common variance estimate σ^2 across all observations. This can be written using mathematical notation as

$$Y_i \sim Normal(\beta_0 + \beta_1 x_i, \sigma^2).$$

Since the residuals are proportional to a Normal distribution, we can obtain an estimate of how probable each observation of y_i is given the estimates of the intercept (β_0), slope (β_1), and common variance (σ^2). This can be represented in mathematical notation as $p(y_i|\beta_0, \beta_1, \sigma^2)$, which could be read as how probable each observation is given the model parameters we have estimated. The likelihood over all datapoints is the product of all $p(y_i|\beta_0, \beta_1, \sigma^2)$ over all observations, given as $p(y|\beta_0, \beta_1, \sigma^2)$. To estimate the β_0, β_1, and σ^2 parameters in a Frequentist way, we wish to maximise the $p(y|\beta_0, \beta_1, \sigma^2)$ function. This is described as Maximum Likelihood estimation, where we use convenient mathematical properties to estimate the standard error of the intercept and slope.

Bayesian methods consider how prior information on β_0, β_1, and σ^2 can penalise the model likelihood through the introduction of Bayes' rule. Rather than having the likelihood of all observations given model parameters $p(y|\beta_0, \beta_1, \sigma^2)$, we can obtain a distribution of model terms given all observations,

$$p(\beta_0, \beta_1, \sigma^2|y) = p(y|\beta_0, \beta_1, \sigma^2) \times p(\beta_0) \times p(\beta_1) \times p(\sigma^2).$$

Where, $p(\beta_0)$, $p(\beta_1)$ and $p(\sigma^2)$ are distributions about our prior knowledge of the intercept, slope, and common variance parameters, and $p(\beta_0, \beta_1, \sigma^2|y)$ is given as the model posterior. Although Bayesian methods allow for the use of prior knowledge, there is potential for conflicting information from our prior and likelihood. The reason for this could be that either our data are flawed and therefore the likelihood is not representative of what we believe should happen, or our prior knowledge is incorrect. It is generally good practice to define priors before looking at the data.

Changing our prior knowledge penalises the model likelihood for each parameter. In Fig. 1 we can see the posterior distribution of the slope parameter narrowing the more confident we are with our prior. If we are confident in our prior knowledge, the choice of prior distribution will have more weight on the posterior. However, it is possible to use a uniform prior that assumes no prior knowledge. Here the model posterior becomes similar to the model likelihood (Fig. 1), and this choice of prior is called a non-informative prior, and the model will produce a similar analysis to a Frequentist analysis using the same data and model terms. Other non-informative priors include the Jeffreys' prior (Jeffreys, 1961) which is a prior on model terms derived from the Maximum Likelihood estimate. In Bayesian inference we still would like to obtain a credible region for each model parameter, or more specifically, a confidence region for estimated model terms. In Frequentist statistics this is known as a Confidence Interval. A Confidence Interval is defined as a critical region which will contain the parameter we are estimating. The Bayesian equivalent is the Credible Interval, defined as an interval posterior distribution containing our parameters of interest. The main difference here is that Confidence Intervals are estimated using a point estimate and an assumed distribution, whereas Credible Intervals of a posterior distribution are estimated through thousands of samples of a Markov Chain Monte Carlo (MCMC) procedure. We discuss MCMC sampling towards the end of the manuscript.

Example 1a

We want to model the relationship between annual mean temperature (°C) and grain yield (t ha^{-1}) from the Broadbalk long-term experiment at Rothamsted Research (Fig. 2) as a linear model. The model is $yield = \beta_0 + \beta_1 \times Annual\ Temperature$, with β_0 the intercept and β_1 the slope parameter, the residuals of this *yield* model are assumed to be Normally distributed with constant variance. Here we use only a subset of yield data from the Broadbalk experiment, using data from a single treatment (Continuous Wheat section Section 1 and the 192 kg N ha^{-1} treatment) between 1968 and 2018 (Glendining and Poulton, 2023) to explore how annual mean temperature affects grain yields within this treatment. Note that Broadbalk is an unreplicated experiment (the design dates back to 1843), so there is only one yield value per treatment. Varieties have changed over this period, so each variety contains a subset of years. There were six varieties sown over this period, but we will include the effect of these varieties in the next example and here focus only on the overall effect of temperature on mean grain yield (for the sake of a simple example, we do not include other potentially relevant climate covariates). We observe how the choice of prior can influence the posterior distribution of the slope parameter in Fig. 1 and Table 1. In this example we use a Uniform(-1.5, 1.5) prior as a

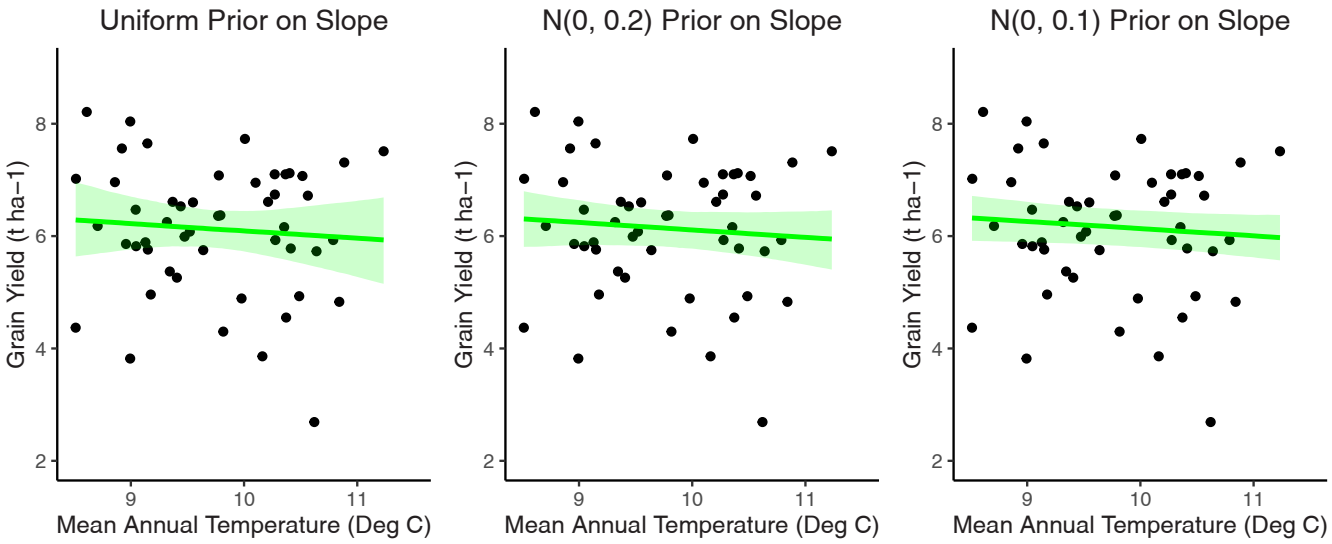

Fig. 2. The influence of different priors on the fitted model posterior of the simple regression model given in Example 1a. The solid line is the mean estimate with 95 % credible intervals.

Table 1
Estimated model coefficients over various priors on the slope parameter for the annual temperature grain yield model given in Example 1a.

Prior setting	Coefficient	Mean	Standard Deviation	2.50% CI	97.50% CI
Uniform	Intercept	7.41	2.33	2.95	11.94
	Slope	-0.13	0.24	-0.60	0.33
Normal(0, 0.2)	Intercept	7.43	1.46	4.67	10.32
	Slope	-0.13	0.15	-0.43	0.15
Normal(0, 0.1)	Intercept	7.39	0.93	5.56	9.20
	Slope	-0.13	0.09	-0.31	0.05

non-Informative prior (similar to a Frequentist approach), a Normal distribution prior with a mean of 0 and a standard deviation of 0.2 (N(0, 0.2)) as an informative prior, and a Normal distribution prior with a mean of 0 and a standard deviation of 0.1 (N(0, 0.1)) as a very informative prior. The more confident we are in our prior knowledge the more the likelihood function is regulated, and the resulting posterior distribution of the slope parameter is narrower (Fig. 1 & Table 1). In Fig. 2 we observe how the 95 % credible interval of the mean estimates from the straight-line relationship narrows the more confident we are in our prior. We discuss MCMC sampling towards the end of the manuscript.

3. Factors

The use of categorical variables (factors) does not change when using Bayesian regression, although, we do need to specify priors for each level of the factor. In Example 1b, an intercept and a prior must be specified for each variety's intercept term. We discussed previously how the distribution of the model residuals follows a Normal distribution. For a model with multiple factors we still assume all the residuals follow a Normal distribution, but it is more efficient to present the model in matrix notation, with mean estimates equal $X\boldsymbol{\beta}$, X is the model design matrix and $\boldsymbol{\beta}$ is a vector of model parameters. The model notation for the data y becomes $Y \sim Normal(X\boldsymbol{\beta}, \sigma^2)$.

Example 1b

From the Broadbalk data in Example 1a, there were six wheat varieties used on Broadbalk from 1968 to 2018. In this example we only consider varying intercept models. Yield is modelled as a linear model with individual intercepts β_0 for each variety and a common slope parameter β_1. The residuals from the linear model are assumed to follow a Normal distribution with constant variance. We can see from Fig. 3 that there are different intercept values being estimated for each variety of wheat sown on Broadbalk, with the Apollo variety being fitted as the reference factor level within the model as default. Where, in Fig. 3 the intercept term refers to the estimated Apollo intercept and the effects of other varieties are added to this term. A reference factor level is a standard procedure used in non-Bayesian regression. We can see the effect of annual mean temperature on grain yield in Fig. 4, with Cappelle estimated to have the lowest intercept value. As you can see we now have six variety parameters to estimate than the previous model in Example 1a. Models with lots of parameters can be difficult to find or think of useful priors to regulate the model. The good news is that in many statistical software packages they automatically choose priors to regulate your model for you. For this analysis we selected weakly-informative priors for each level of the factor which is a default prior specification from RStanarm (Goodrich et al., 2023). Weakly-informative priors provide moderate regulation of model parameters and prevent the domination of the prior within a model (Gelman et al., 2008). Remember, Bayesian modelling is all about regulating the model and caution should be given when selecting a prior distribution that is too narrow, because this can lead to a prior becoming non-representative of the data, resulting in the final model being mis-specified. However, it is advised that weakly informative priors are more beneficial than fitting a non-informative or informative prior as this can help sample model parameters without priors dominating parameter estimation (Simpson et al., 2017).

4. Multiple regression and comparison

Performing multiple regression in a Bayesian workflow is similar to performing a Frequentist multiple regression, particularly when using a software package such as RStanarm that is designed to have similar syntax to common Frequentist regression software. However, there are some key differences. Firstly, it is often of interest to include predictors that are providing useful information when predicting our response y. In a non-Bayesian context, this is particularly important, because in Frequentist statistics there is a finite number of degrees of freedom based on the total number of observations and each parameter we fit has a cost. Some redundant predictors in Frequentist model selection are removed using forwards or backwards selection via the AIC (Akaike, 1973) or BIC

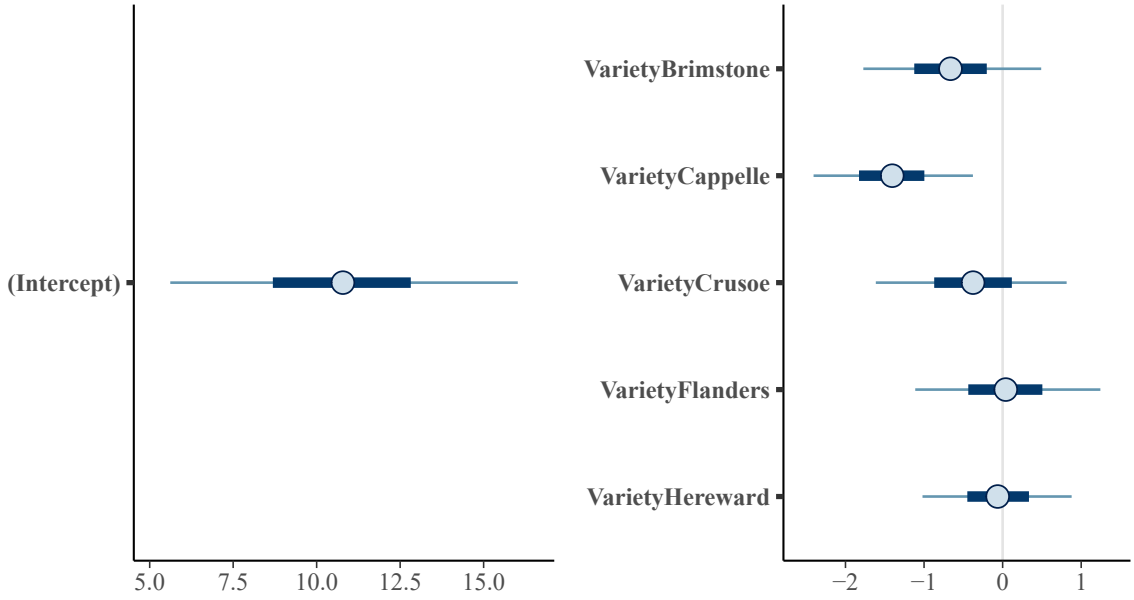

Fig. 3. The posterior estimates for the variety terms from Example 1b with 50 % (thick blue line) and 95 % (thin blue line) credible intervals. The (Intercept) term refers to the default Apollo variety intercept term. Variety Brimstone term is the added intercept term for the Brimstone variety, this term needs to be added to (Intercept). This is true for all Variety terms.

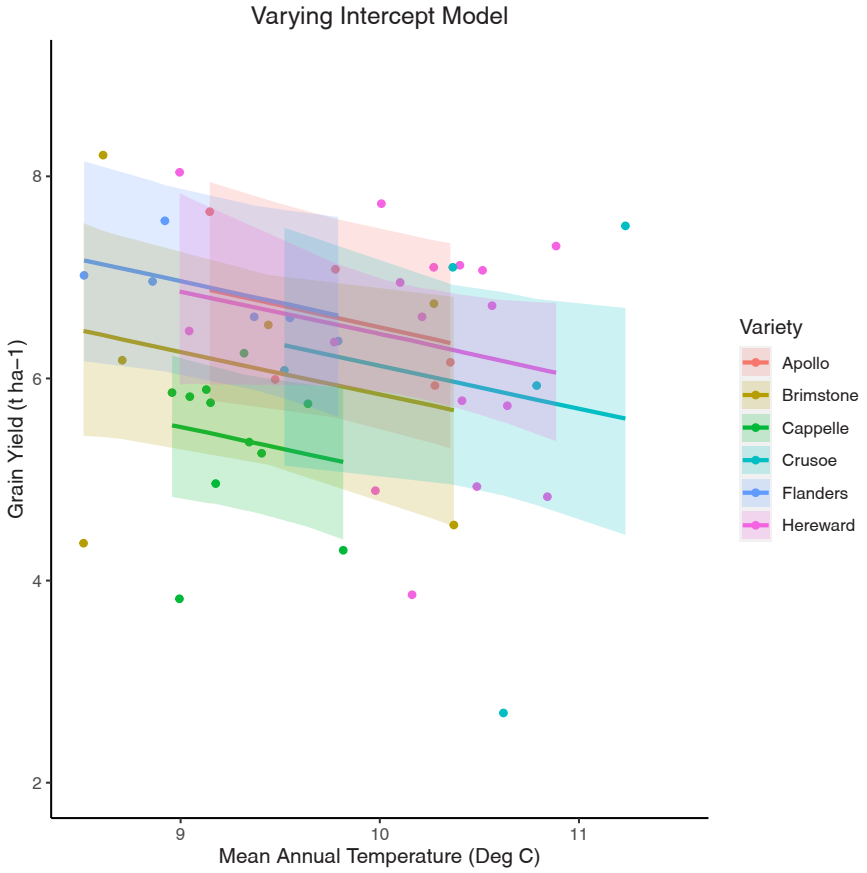

Fig. 4. The posterior estimates for the varying intercept model in Example 1b. Each Variety from Broadbalk was a different intercept but the same slope parameter. The solid lines are the mean estimates for each variety with 95 % credible intervals.

(Raftery, 1995). There are other Frequentist methods of model selection such as statistical tests, but we only highlight two here. Methods to remove redundant variables from a Bayesian Multiple Regression model include the Posterior Inclusion Probability (PIP) for each model parameter of a full model (George and McCulloch, 1993) and, forward and backwards projection using a reference model (Pavone et al., 2023).

The use of PIPs, and forward and backwards projection are outside the scope of this manuscript (for more information see Hoeting et al., 1999).

When comparing models implemented in a Bayesian workflow, to identify whether models containing different sets of variables have a better or worse predictive fit to the data, the leave-one-out Information Criterion (LOOIC) (Vehtari et al., 2017) is used to assess which model has the best posterior predictive fit over all iterations of the model. As we demonstrate in our next example, this method acts similarly to the AIC and BIC from Frequentist methods and allows for model comparison on the same response y. Other methods not explored in this manuscript include Bayesian Model Averaging (Hoeting et al., 1999) and Model Stacking (Yao et al., 2018). These methods involve averaging across a series of models based on their predictive fit rather than selecting one model with the best fit. See Gelman et al. (2020) for more examples on varying intercept and slope models, along with more on Bayesian multiple regression.

One common criticism of Bayesian statistics is the subjectivity through the choice of priors, and we have already discussed weakly-informative priors. However, in a multiple regression analysis, one benefit of Bayesian model selection is the use of priors on multiple model parameters. These priors can help regulate or shrink model terms which may not help predict the response (model parameters which are around zero). The term shrinkage regulates model parameters whose coefficients are close to zero and shrinks them closer to zero. To obtain shrinkage estimates for model parameters in Bayesian regression we set the distribution of the prior for a model parameter to be Normal and around zero, the standard deviation or spread of this prior on model parameters is estimated, narrowing the plausible parameter space for model terms. The choice of prior on the standard deviation of the model term can be specified which will determine the overall effect of shrinkage (Carvalho et al., 2010). Shrinkage is similar to LASSO regression in a Frequentist context (Tibshirani, 1996), and more advanced Bayesian shrinkage methods include the use of the regulated Horseshoe prior (Piironen and Vehtari, 2017). This is where we see the benefits of Bayesian methods. LASSO regression (non-Bayesian) regulates the model likelihood in a non-formal way, and so penalises the model likelihood via a mathematical function and has fixed penalisation to all model terms. In contrast by formalising model regulation through Bayes' Rule, we can obtain more dynamical shrinkage properties, by estimating individual shrinkage for each model term. However, shrinkage is not appropriate in our example as we have too few model predictors and in this study, we only consider weakly-informative priors for our multiple regression model.

Example 2a

So far we have only included mean annual temperature as a predictor of grain yield, but what if we want to understand within year effects of temperature? In this example, we now model grain yield as a varying intercept model with variety as a factor, along with seasonal summaries of temperature (i.e. multiple slope parameters β_1 for mean temperatures in winter, spring, summer and autumn) as continuous predictors. The Normality assumption for yield still holds in this example. We can see by comparing the LOOIC that the model including seasonal temperatures provided a better model fit than the model with annual temperature and varying intercept for variety model (Table 2). However, the standard error of both LOOIC estimates is high, which suggests the difference in predictive accuracy between models is small. The parameter estimates for each model parameter are given in Fig. 5. From all the seasonal variables mean temperature in spring had the strongest negative relationship with yield and suggested that warmer temperatures in the spring lead to a reduction in yield.

Table 2
Leave-One-Out Information Criterion (LOOIC) estimates for the Varying Intercept models of annual weather given in Example 2 and seasonal weather given in Example 2a.

Parameterised Model	LOOIC	SE
Annual Weather	164.3	14.3
Seasonal Weather	163.4	12.1

5. Model diagnostics in a Bayesian workflow

Frequentist and other non-Bayesian methods maximise the likelihood function to obtain optimum estimates of model parameters. After the optimum is found, usually algebraic derivation of the standard error is used to obtain confidence intervals of model parameters. In contrast, Bayesian methods allow prior information to regulate the model likelihood and influence the model parameters posterior distribution. The algorithm used to obtain model parameters is often a Markov Chain, or more specifically a MCMC. More advanced algorithms are used in RStanarm such as the Hamiltonian Markov Chain (HMC) (Neal, 2011). MCMC methods sample model parameters 1000 s of times over multiple chains after an initial burn-in period to obtain posterior distributions. Due to the fast computational sampling of modern computers, we can obtain 1000 s of samples or iterations relatively quickly. With more complicated models more samples and iterations are needed to obtain convergence. Convergence in the Markov Chain occurs when there is no trend in the samples, with samples distributed evenly and each chain is similar. The RStanarm package estimates 2000 samples over 4 runs as default (8000 estimates in total), but only takes the last 1000 samples from each run, this is because there is a default burn-in of 1000 samples each chain (Goodrich et al., 2023). In the Bayesian workflow we should see convergence in our MCMC iterations in order to assume we have adequate model estimates (Gabry et al., 2019). From these MCMC iterations after the burn-in period we can use the laws of large numbers and construct 95 % credible intervals based on the empirical distribution of our model terms. Once we have obtained 1000 s of posterior estimates for each parameter, we have large amounts of synthetic data to make inferences on our model. Synthetic data is data we have simulated from our model and posterior parameters given the data we have fitted, and if our model is adequate the posterior distribution of the synthetic data should follow a similar distribution to the observed data, this is called the posterior predictive distribution. There are more sophisticated posterior predictive checks you can do, but we only include the basic ones in this manuscript.

Example 2b

Consider the varying intercept model for seasonal summarised variables in Example 2a. We now have 8000 runs of the model, from which we can simulate $Y \sim Normal(X\boldsymbol{\beta}, \sigma^2)$ over a random selection of our 8000 iterations. We can use a smooth histogram of the posterior predictive distribution to check if it follows the distribution of the data (Fig. 6), or we can check if the model average across our synthetic data follows a straight-line relationship with our observed Broadbalk grain yield data (Fig. 6). The Markov Chain for the model terms estimated in Example 2a shows convergence over 4 chains of 1000 iterations (Fig. 7), so model terms are adequately estimated.

6. Hierarchical models and random effects

Hierarchical or multilevel models allow for the modelling of data measured at different levels taking into account the complex dependency structures in the grouping units (Bürkner, 2017; Gelman and Pardoe, 2006). The use of hierarchical structure is very common in field experiments, for example, when the same experiment is conducted on multiple sites, or an experiment is conducted at the same site with samples measured repeatedly over time. Therefore, the effect of treatments is nested within multiple levels of experimental units, either across sites or through time. This type of experiment is often analysed using linear or non-linear mixed-effect models. Mixed-effect models can include additional variance terms within the model while estimating model parameters for each level of the experiment. For example, in the same linear model as in Example 1a, $y_i = \beta_0 + \beta_1 x_i$ is estimated across

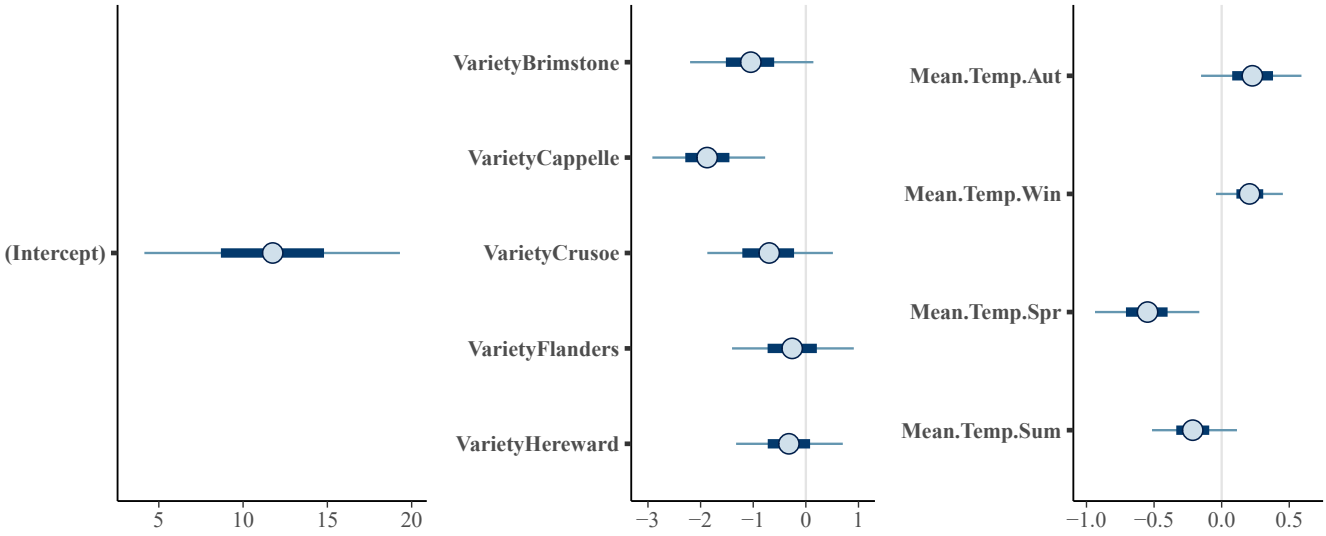

Fig. 5. The posterior estimates of seasonal temperature model coefficients from the varying intercept model given in Example 2a with 50 % (thick blue line) and 95 % (thin blue line) credible intervals.

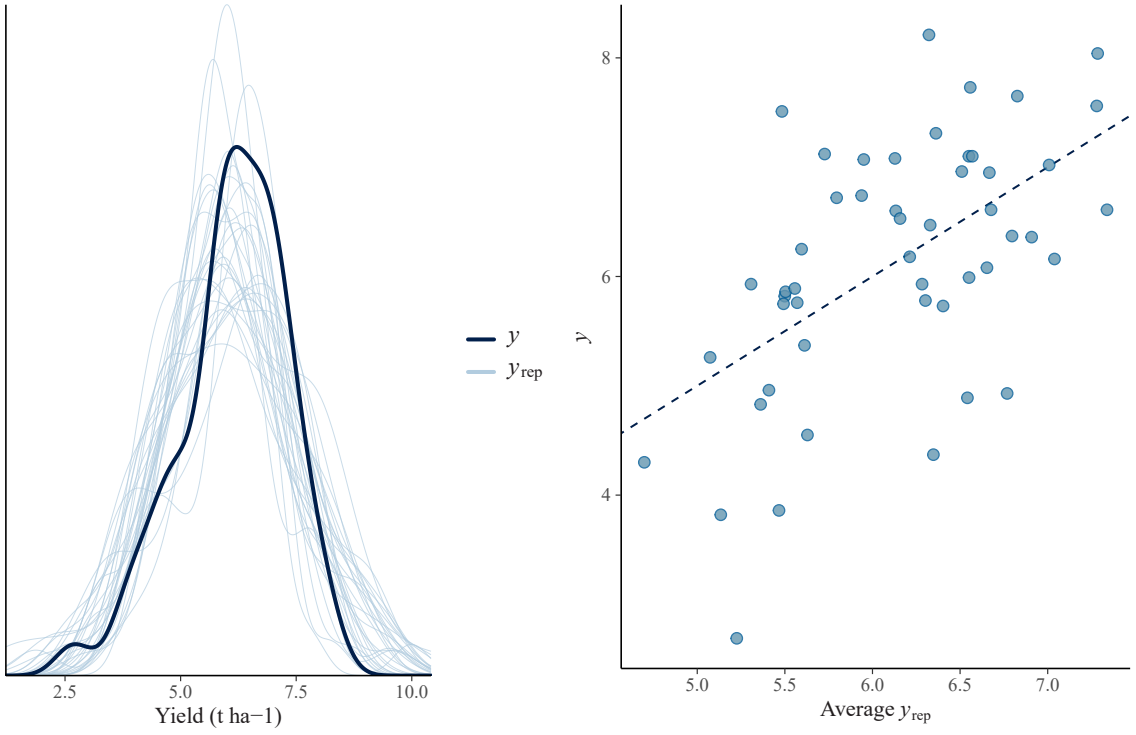

Fig. 6. The posterior predictive checks of the varying intercept seasonal temperature model in Example 2a. Left is the smooth histogram with 30 random samples of synthetic data sampled from the Markov Chain from the fitted model. The dark blue line is the smoothed histogram of the data in the example. Right is the scatter-average (mean) plot of the observed data on the y-axis and the average posterior estimates from the fitted model for each observation.

three sites, where the experiment is replicated at three locations. We are now estimating three intercepts (β_0) and three slopes (β_0) across all sites. However, we are less interested in the effect of each β_0 and β_1 parameters at each site than the average effect across all sites and the variability associated with the average effect across all sites. Including a random term allows intercept (β_0) and slope (β_1) parameters to vary at each experimental level. Estimates of the variability of the parameters are characterized by a variance-covariance matrix. Variance estimates of β_0 and β_1 across three sites are given as $\sigma^2_{\beta_0}$ and $\sigma^2_{\beta_1}$, these statistics inform how uniform the linear model was across all three sites. In a Bayesian analysis, prior distributions can be assumed on variance terms $\sigma^2_{\beta_0}$ and $\sigma^2_{\beta_1}$ (Gelman, 2006). The covariance and correlation estimates between β_0 and β_1 informs us how similar these estimates are across three sites. For more sophisticated models, there can be more complicated multi-level structures such as temporal correlations in repeated measurements, which can be estimated using different covariance structures. Widely used R packages such as lme4 (Bates et al., 2015) and nlme (Pinheiro and Bates, 2023) have been developed to fit Frequentist

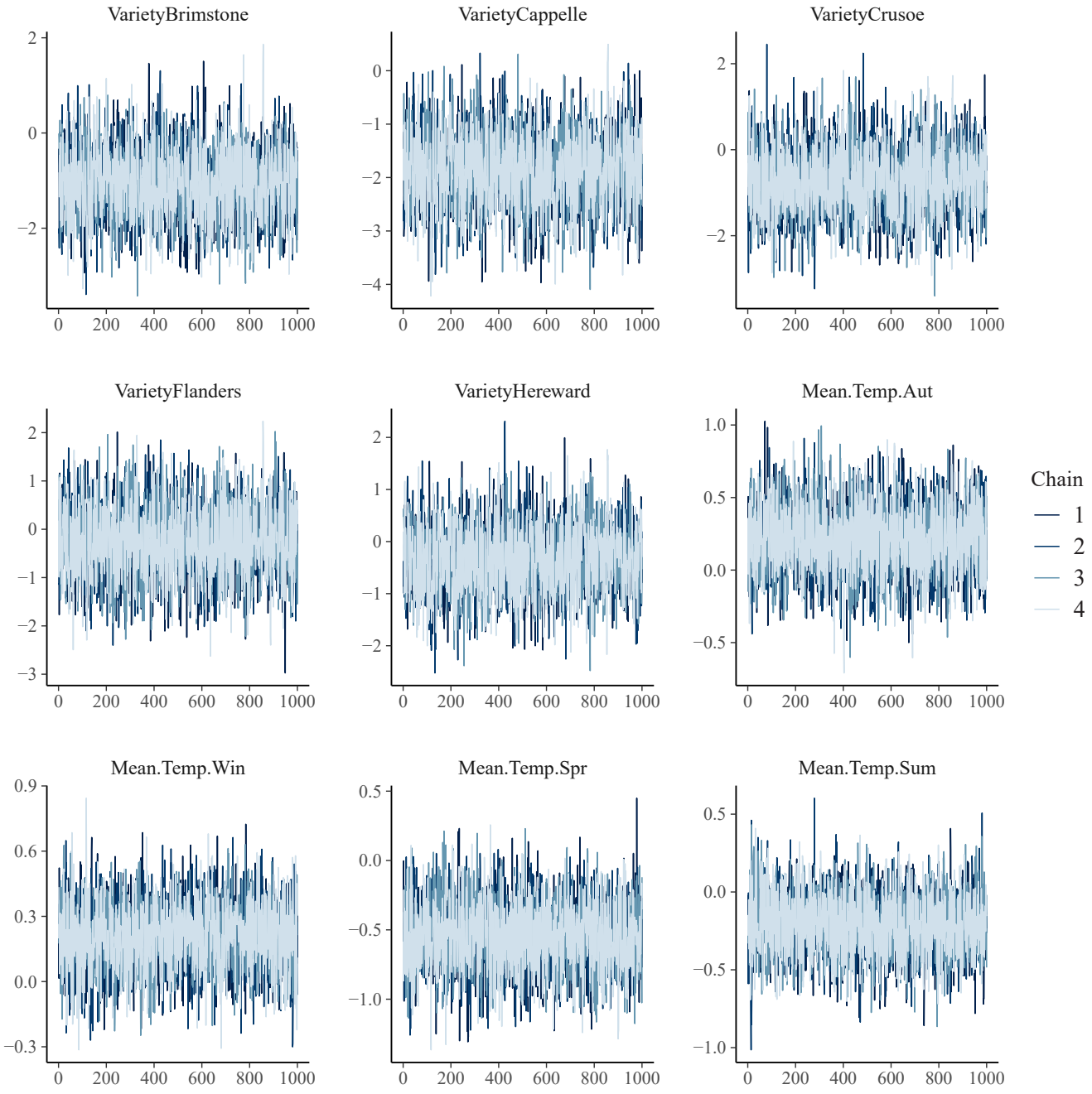

Fig. 7. The four sampled Markov Chains for estimated parameters of the varying intercept seasonal weather model across four chains. Chains are sequentially overlaid (1, 2, 3, 4) for each sub-figure to illustrate good mixture of chains.

multilevel mixed-effect models. Despite the flexibilities of multilevel mixed-effect models, the conventional Frequentist approach is more likely to encounter convergence issues when the variance components of random effects are close to zero (on the boundary of their corresponding parameter space – we cannot have negative estimates of variance), resulting in dropping levels of random effects (Bates et al., 2014). A Bayesian approach is an alternative to fit complex multilevel mixed-effect models using priors on structural variance terms.

A Bayesian approach typically defines a prior for the variance-covariance matrix of the random effects that incorporate prior knowledge about parameters or derive probability statements for interested parameters (Gelman et al., 2013). The R package brms (Bürkner, 2017) applies extended formula syntax that is similar to lme4 (Bates et al., 2014), allowing for complex mixed-effect models using Bayesian methods. In brms, a wide range of distributions and link functions are available, and users can use the default implementation of priors of model coefficients or define them explicitly. The most common prior distributions are Normal and Student's t for fixed effect regression coefficients. Multilevel models are supported by setting up multiple grouping factors, which specify the parameters of random effects (variance components), including random intercepts, slopes and correlations at group-level. The temporal or/and spatial autocorrelations can be modelled with available functions in brms, such as compound symmetry (COSY), autoregressive (AR), spatial conditional autoregressive (CAR), etc. A typical workflow of using brms includes defining the distribution of response variable (e.g. Normal, Poisson or Binomial),

specifying prior distribution and model parameters, adjusting the sampling behaviour of Stan through control argument, and analysing results including estimations from posterior samples.

Example 3

The Swedish long-term soil fertility trial in central Sweden was established in the 1960s to investigate crop rotation and fertilization effects on soil fertility and crop yield. The experiment is replicated at three sites in central Sweden with a split-split-block design. Two crop rotations representing farming systems with and without livestock (Rotation I and II) are randomized to two blocks (main plot factor) at each site, four levels of phosphorus and potassium fertilization (sub-plot factor) were nested under two types of crop rotation and randomized in columns of each block, and four levels of nitrogen fertilization (sub-subplot factor) were nested under each level of phosphorus and potassium fertilization and randomized in columns of phosphorus and potassium fertilization (Carlgren and Mattsson, 2001; Ivarsson and Bjarnason, 1988). Topsoil (0–20 cm) samples from two plots of the same treatment at each site were mixed and analysed as composite samples after each rotation cycle (every six years). Therefore, we treated Site as Block in setting up random effect levels. In this example, we analysed the crop rotation effect on total carbon content in the topsoil (%) using a multilevel regression model with brms package (Bürkner, 2017). The model has a linear term for year and crossed with rotation, and fertilization treatments were not included in fixed effects here but used in setting up group-level effects (random effects) and temporal autocorrelation (covariance structure) (Fig. 8). We set up the multilevel model structure in brms similarly to the structure in nlme, i.e. a Normal distribution for the response variable and compound symmetry for the covariance structure. In the compound symmetry cosy(time|group) we specified year as time and crossed Block and Sub-Block with the lowest level of fertilization treatments as autocorrelated subjects (Supplementary Material code Example 3). Standard deviation estimates of model terms at the Block and Sub-Block were estimated as 0.2231 (SE: 0.3544) and 0.0570 (SE: 0.0855) (Table 3), which suggests there was much higher variation in Total Carbon (%) across Block than there was within Sub-Block. There was also strong temporal autocorrelation across all years. Despite large variations across sites (first level of group effect), the negative regression slopes at the population level suggested soil

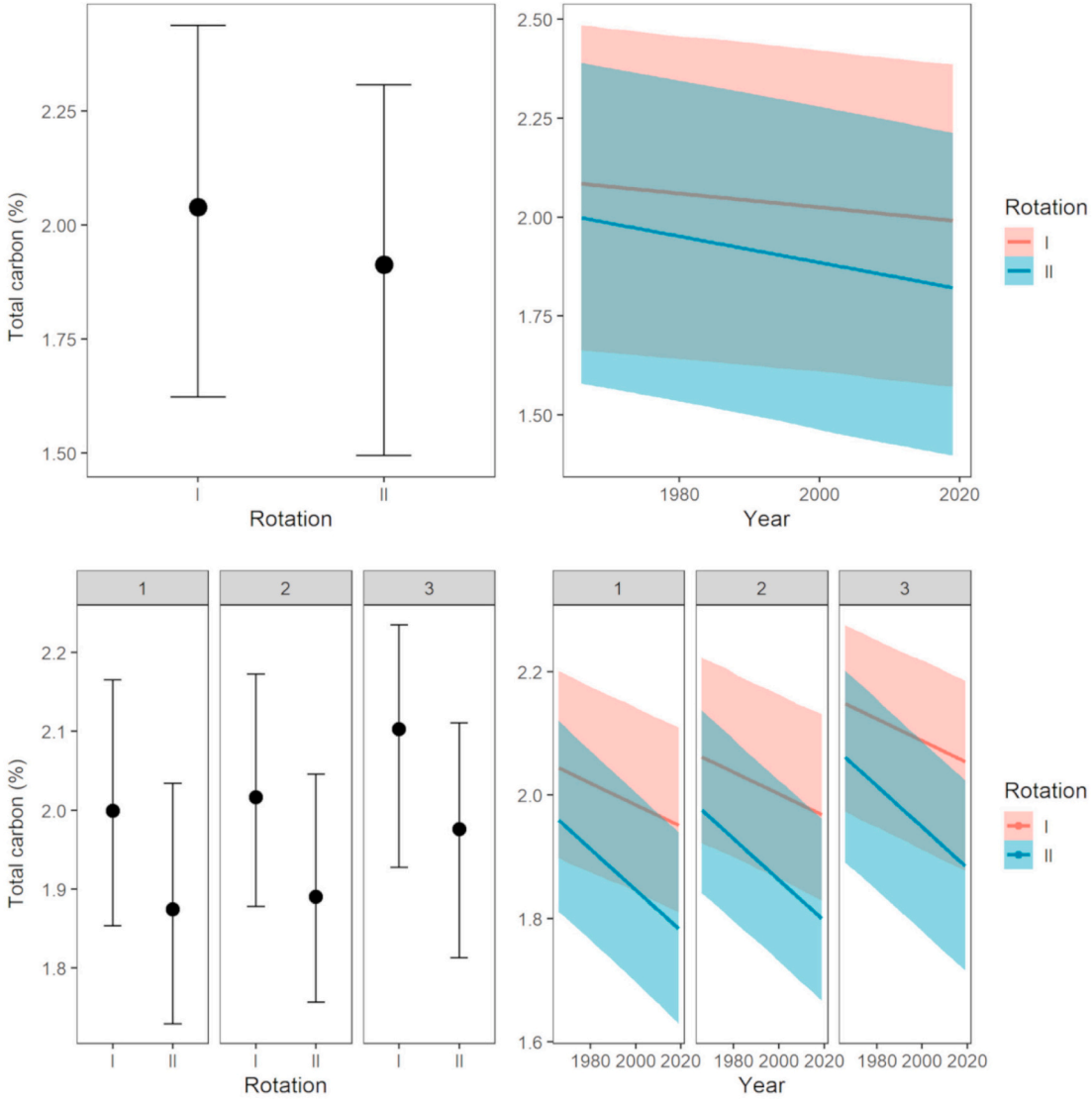

Fig. 8. Fitted hierarchical model for the crop rotation effect at the population level and the group level (Block) in Example 3. Top row is the fitted model parameters of Rotation and Year. Bottom row is the sub-Block random effects for Rotation and Year. Figures at left sides are means ± standard deviation, and at right side are regression lines with 95 % prediction credible intervals.

Table 3
Estimated model parameters at the population and group level for the multilevel mixed model given in Example 3.

	Parameters	Estimate	Est.Error	2.50 % CI	97.50 % CI
Population-level effects	Intercept	5.5499	0.8966	3.8343	7.3055
	Year	-0.0018	0.0004	-0.0026	-0.0009
	Rotation II	2.9991	1.2610	0.5121	5.4517
	Year:Rotation II	-0.0016	0.0006	-0.0028	-0.0003
Group-level effects	sd (intercept),~Block (n=3)	0.2231	0.3544	0.0110	1.1649
	sd (intercept),~Block:SubBlock (n=6)	0.0570	0.0855	0.0012	0.2544
Family specific parameter	sigma	0.2119	0.0075	0.1985	0.2282
Correlation structure	cosy	0.3561	0.0434	0.2767	0.4448

carbon content decreased in both crop rotations over time but at a slower rate in the rotation with livestock (Table 3). Fitted conditional effect (crop rotation) at the population and group levels were presented in Fig. 8.

7. Summary

This manuscript has been a short introduction to the Bayesian workflow using examples from long-term agricultural field experiments. We have presented some key ideas of the Bayesian workflow such as likelihoods, priors and posteriors, credible intervals, Markov Chain sampling, and posterior predictive checks. However, care should be taken when analysing data regardless of the statistical method used. Although we can include prior information in our model in a Bayesian workflow, the prior information used by the researcher should be appropriate under the correct circumstances. We have briefly touched on the benefits of Bayesian modelling, such as regulating models with many parameters and setting priors of variance terms in hierarchical models to aid in model convergence. Although we covered aspects of multiple regression, these are by no means the extent of Bayesian methods and we encourage readers who found this manuscript useful to explore further topics in further reading material (Gelman et al., 2013; McElreath, 2018).

Declaration of Competing Interest

The authors declare no competing interests.

Data Availability

The authors do not have permission to share data.

Acknowledgement

We thank data stewards Margaret Glendining and Sarah Perryman for access to the Rothamsted data from the Electronic Rothamsted Archive (e-RA). The Rothamsted Long-Term Experiments - National Bioscience Research Infrastructure RLTE-NBRI is supported by the Lawes Agricultural Trust and the Biotechnology and Biological Sciences Research Council (Grants BBS/E/C/00005189 (2012–2017); BBS/E/C/000J0300 (2017–2022); BB/CCG2280/1 (2023–2028)). We are thankful to the data holder and especially the Faculty of Natural Resources and Agricultural Sciences at SLU for providing resources to maintain the Swedish long-term experiments. This work is supported by the Swedish Government Research Council for Sustainable Development (FORMAS, grant no. 2022–00214). Description of soil fertility experiment (R3–9001 series) and contact for retrieving historical data are available at https://www.slu.se/en/departments/soil-environment/research/soil-nutrient-cycling/slu-field-research-plant-nutrition/, datasheets of crop yield, plant nutrients and soil properties from 2009 to 2019 are available in PDF format at https://www.slu.se/institutioner/mark-miljo/forskning/mark-naringsomsattning/langliggande-vaxtnaringsforsok/ (in Swedish).

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.eja.2024.127227.

References

Addy, J.W.G. (2023). Mean Seasonal Air Temperature and Total Seasonal Rain at Rothamsted 1968-2022. Electronic Rothamsted Archive, Rothamsted Research. https://doi.org/10.23637/rms-RothSeasonTotRainMeanTemp1968-2022.

Addy, J.W.G., Ellis, R.H., Macdonald, A.J., Semenov, M.A., Mead, A., 2020. Investigating the effects of inter-annual weather variation (1968–2016) on the functional response of cereal grain yield to applied nitrogen, using data from the Rothamsted Long-Term Experiments. Agric. For. Meteorol. *284* https://doi.org/10.1016/j.agrformet.2019.107898.

Addy, J.W.G., Ellis, R.H., Maclaren, C., Macdonald, A.J., Semenov, M.A., Mead, A., 2022. A heteroskedastic model of Park Grass spring hay yields in response to weather suggests continuing yield decline with climate change in future decades. J. R. Soc. Interface *19* (193). https://doi.org/10.1098/rsif.2022.0361.

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. N. Petrov & F. Csáki (Eds.), Proceedings to the 2nd International Symposium on Information Theory. (pp. 267–281).

Bates, D., Kliegl, R., Vasishth, S., & Baayen, R.H. (2015). Parsimonious Mixed Models. *ArXiv Preprint.* https://arxiv.org/abs/1506.04967v2.

Bates, D., Mächler, M., Bolker, B.M., Walker, S.C., 2014. Fitting linear mixed-effects models using lme4. J. Stat. Softw. *67* (1) https://doi.org/10.18637/jss.v067.i01.

Bürkner, P.C., 2017. brms: an R package for Bayesian multilevel models using stan. J. Stat. Softw. *80* (1), 28. https://doi.org/10.18637/JSS.V080.I01.

Carlgren, K., Mattsson, L., 2001. Swedish soil fertility experiments. Acta Agric. Scand., Sect. B - Soil Plant Sci. *51* (2), 49–76. https://doi.org/10.1080/090647101753483787.

Carvalho, C.M., Polson, N.G., Scott, J.G., 2010. The horseshoe estimator for sparse signals. Biometrika *97* (2), 465–480. https://doi.org/10.1093/BIOMET/ASQ017.

Eckl, T., Piepho, H., 2015. Analysis of series of cultivar trials with perennial grasses for subdivided target regions. Crop Sci. *55* (2), 597–609. https://doi.org/10.2135/cropsci2014.04.0327.

Gabry, J., Simpson, D., Vehtari, A., Betancourt, M., Gelman, A., 2019. Visualization in Bayesian Workflow. J. R. Stat. Soc. Ser. A: Stat. Soc. *182* (2), 389–402. https://doi.org/10.1111/RSSA.12378.

Gelman, A., Carlin, J.B., Stern, H.S., Dunson, D.B., Vehtari, A., Rubin, D.B., 2013. Bayesian Data Analysis. Chapman and Hall/CRC. https://doi.org/10.1201/b16018.

Gelman, A., Hill, J., Vehtari, A., 2020. Regression and other stories. Cambridge University Press.

Gelman, A., Jakulin, A., Pittau, M.G., Su, Y.S., 2008. A weakly informative default prior distribution for logistic and other regression models. Ann. Appl. Stat. *2* (4), 1360–1383. https://doi.org/10.1214/08-AOAS191.

Gelman, A., Pardoe, I., 2006. Bayesian measures of explained variance and pooling in multilevel (Hierarchical) models. Technometrics *48* (2), 241–251. https://doi.org/10.1198/004017005000000517.

George, E.I., McCulloch, R.E., 1993. Variable selection via Gibbs sampling. J. Am. Stat. Assoc. *88* (423), 881–889. https://doi.org/10.1080/01621459.1993.10476353.

Glendining, M., & Poulton, P. (2023). Broadbalk Wheat annual grain and straw yields 1968-2022. Electronic Rothamsted Archive, Rothamsted Research. https://doi.org/10.23637/rbk1-yld2667-01.

Goodrich, B., Gabry, J., Ali, I., & Brilleman, S. (2023). rstanarm: Bayesian applied regression modeling via Stan (R package version 2.26.1).

Grosse, M., Hierold, W., Ahlborn, M.C., Piepho, H.-P., Helming, K., 2020. Long-term field experiments in Germany: classification and spatial representation. SOIL *6* (2), 579–596. https://doi.org/10.5194/soil-6-579-2020.

Hoeting, J.A., Madigan, D., Raftery, A.E., Volinsky, C.T., 1999. Bayesian model averaging: a tutorial (with comments by M. Clyde, David Draper and E. I. George, and a rejoinder by the authors. Stat. Sci. *14* (4), 382–417. https://doi.org/10.1214/SS/1009212519.

Ivarsson, K., Bjarnason, S., 1988. The Long-Term Soil Fertility Experiments in Southern Sweden. Acta Agric. Scand. *38* (2), 137–143. https://doi.org/10.1080/00015128809438477.

Jeffreys, H., 1961. Theory of Probability, 3rd ed. Oxford University Press.

Macholdt, J., Piepho, H.-P., Honermeier, B., Perryman, S., Macdonald, A., Poulton, P., 2020. The effects of cropping sequence, fertilization and straw management on the yield stability of winter wheat (1986–2017) in the Broadbalk Wheat Experiment, Rothamsted, UK. J. Agric. Sci. *158* (1–2), 65–79. https://doi.org/10.1017/S0021859620000301.

McElreath, R., 2018. Statistical Rethinking. Chapman and Hall/CRC. https://doi.org/10.1201/9781315372495.

Neal, R.M. (2011). MCMC using Hamiltonian dynamics. Handbook of Markov Chain Monte Carlo.

Onofri, A., Seddaiu, G., Piepho, H.-P., 2016. Long-Term Experiments with cropping systems: Case studies on data analysis. Eur. J. Agron. *77*, 223–235. https://doi.org/10.1016/j.eja.2016.02.005.

Pavone, F., Piironen, J., Bürkner, P.C., Vehtari, A., 2023. Using reference models in variable selection. Comput. Stat. *38* (1), 349–371 https://doi.org/10.1007/S00180-022-01231-6/FIGURES/9.

Payne, R.W. (2018). The Design and Analysis of Long-term Rotation Experiments (pp. 299–317). https://doi.org/10.2134/appliedstatistics.2016.0001.c11.

Perryman, S., & Scott, T. (2020). Annual Mean Air Temperature Anomaly at Rothamsted 1878-2019. In Electronic Rothamsted Archive, Rothamsted Research. https://doi.org/10.23637/rms-RMAAtempanomaly-1.

Piironen, J., Vehtari, A., 2017. Sparsity information and regularization in the horseshoe and other shrinkage priors. Electron. J. Stat. *11* (2), 5018–5051. https://doi.org/10.1214/17-EJS1337SI.

Pinheiro, J., & Bates, D. (2023). nlme: Linear and Nonlinear Mixed Effects Models.

Raftery, A.E., 1995. Bayesian model selection in social research. Sociol. Methodol. *25*, 111. https://doi.org/10.2307/271063.

Rasmussen, P.E., Goulding, K.W.T., Brown, J.R., Grace, P.R., Janzen, H.H., Korschens, M., 1998. Long-term agroecosystem experiments: Assessing agricultural sustainability and global change. Science *282* (5390), 893–896 https://doi.org/10.1126/SCIENCE.282.5390.893/ASSET/FCFB77F9-AB25-4078-A1D7-F505D28A84DB/ASSETS/GRAPHIC/SE4386930004.JPEG.

Richter, C., Kroschewski, B., 2006. Analysis of a Long-term Experiment with Repeated-measurement Models. J. Agron. Crop Sci. *192* (1), 55–71. https://doi.org/10.1111/j.1439-037X.2006.00167.x.

Simpson, D., Rue, H., Riebler, A., Martins, T.G., Sørbye, S.H., 2017. Penalising model component complexity: a principled, practical approach to constructing priors. Stat. Sci. *32* (1), 1–28. https://doi.org/10.1214/16-STS576.

Storkey, J., Macdonald, A.J., Bell, J.R., Clark, I.M., Gregory, A.S., Hawkins, N.J., Hirsch, P.R., Todman, L.C., Whitmore, A.P., 2016. The unique contribution of Rothamsted to ecological research at large temporal scales. Adv. Ecol. Res. *55*, 3–42. https://doi.org/10.1016/BS.AECR.2016.08.002.

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 de Schoot, R., Depaoli, S., King, R., Kramer, B., Märtens, K., Tadesse, M.G., Vannucci, M., Gelman, A., Veen, D., Willemsen, J., Yau, C., 2021. Bayesian statistics and modelling. *2021 1:1* Nat. Rev. Methods Prim. *1* (1), 1–26. https://doi.org/10.1038/s43586-020-00001-2.

Vehtari, A., Gelman, A., Gabry, J., 2017. Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. Stat. Comput. *27* (5), 1413–1432 https://doi.org/10.1007/S11222-016-9696-4/FIGURES/12.

Yao, Y., Vehtari, A., Simpson, D., Gelman, A., 2018. Using stacking to average Bayesian predictive distributions (with discussion). Bayesian Anal. *13* (3), 917–1007. https://doi.org/10.1214/17-BA1091.