Crossling Outlook

Forecasting heading dates and site performance
from a breeding program's own records

Jakob Richert

Crossthread · crossling.app

Technical white paper · November 2026

Abstract

Crossling's Outlook forecasts two things from a breeding program's own records: the day each plot will head, and how each line will do at each site in that site's weather. Both forecasts come from one engine, a REML mixed model that borrows strength from relatives through pedigree and DNA. Every program's models are refitted every night on that program's data alone, in about a second per trait.

The two models feed each other. The heading model learns how much warmth each line needs to head, slowed by short days when the data say so. Its predicted heading dates then set the crop-stage windows that the site model measures weather in: rain before heading, heat around flowering, heat in grain fill.

On Crossling's simulated demonstration wheat program (22 site-seasons, 1,468 plots, 182 lines), each site-season was left out in turn and predicted from the others. Heading forecasts were off by 1.0 day on average, with 97% of plots within 3 days; the same model counting days from sowing was off by 4.1 days. Site-average grain yield came within 223 kg/ha, against 515 kg/ha for the same model without weather. The heading model also recovered the photoperiod sensitivity built into the simulator (0.1 per hour) from five candidates.

These numbers come from a simulator whose true rules we wrote, run on real NASA POWER weather. They show the method finds structure that is known to be there. They are not a measure of accuracy on real fields; testing on real program data is the next step.

1 The problem

Breeders plan around two questions the weather keeps changing: when will these plots reach the stage I need to score, and which lines will do best where they'll be grown.

Timing. Heading has to be scored within a few days at every site, and sites can be hundreds of kilometres apart. Most programs plan by calendar, or by "it headed around 5 July last year". Counting days from sowing ignores that a cool year slows development: a line that heads in 52 days at a warm site can take 58 at a cool one, while needing about the same warmth in both.

Performance across sites. A line's mean over trials, or its BLUP, assumes it ranks the same everywhere. Lines respond to heat and drought differently, so rankings move between sites and years (genotype-by-environment interaction).

Two families of tools exist, and neither fits a breeding program's daily work well.

- **Crop growth models** such as APSIM and DSSAT simulate development from weather. Each cultivar needs calibrated coefficients, which isn't practical for hundreds of new lines a year.
- **Genomic prediction** handles new lines through their relatives. Usually the weather isn't in the model, or it enters as a site label that can't be carried to next season.

The Outlook sits between them. It is statistical and learned from the program's own scores, like genomic prediction. Like a crop model, it reads daily weather and the crop's stage. It is light enough to refit every night.

2 What it reads

The Outlook needs nothing a program doesn't already keep in its field book, plus daily weather, which Crossling fetches itself (Table 1). Models are fitted for each program on that program's records only. Relatives are found only within the program's own pedigree and DNA. Nothing is pooled across programs.

Table 1: What the Outlook reads, and where it comes from.

Input	Where it comes from	Used by
Plot scores (days to heading, yield, height, disease and so on)	The program's own trials	Both models
Sowing date and site coordinates	Trial and location records	Both models
Pedigree	Germplasm records (coefficient of parentage)	Relatives
DNA markers, where a program has them	Genomics lab results	Relatives
Daily maximum and minimum temperature, rain	The program's own weather station if it has one; otherwise NASA POWER daily data	Both models
The past 15 years' weather at each site	NASA POWER, fetched nightly	Playing out the rest of a season
The coming days' forecast	MET Norway Locationforecast	Heading, during the season

Figure 1 shows how the parts connect. Every input feeds both models. The one link between the models is the heading day, which tells the site model when each crop stage fell that season.

3 Model 1: when each plot will head

A plot heads once it has had enough warmth, not after a fixed number of days. So each line is described by its *requirement*: how many development units it needs from sowing to heading.

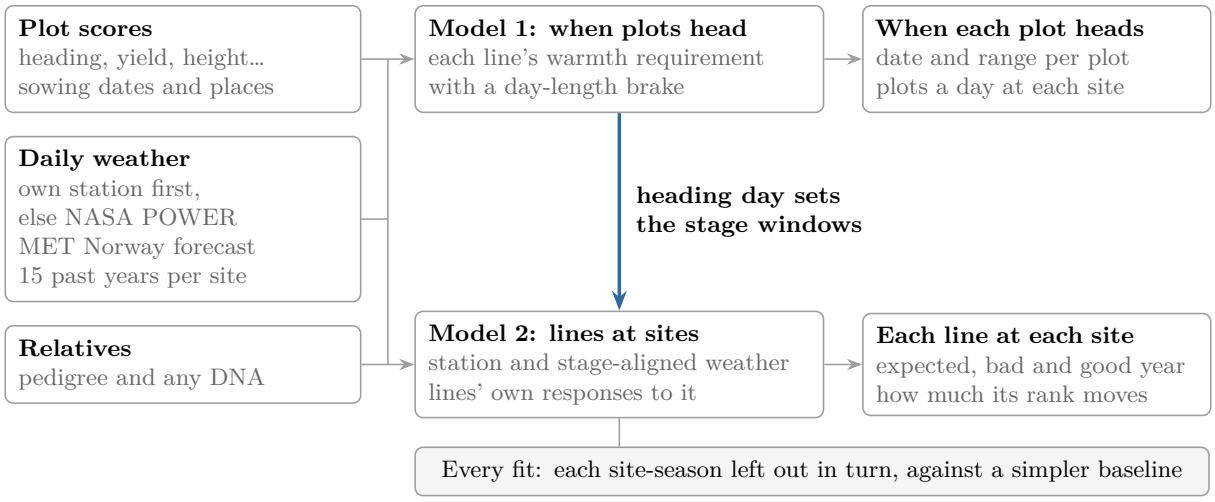


Figure 1: How the Outlook's parts connect. The heading model is the site model's clock.

Development units. A day's units are its growing degree days, slowed when the days are shorter than a long-day crop likes:

$$u_t = \max\!\left(0,\ \frac{T_{\max,t} + T_{\min,t}}{2} - T_b\right) \cdot \exp\!\big(-s \cdot \max(0,\ 20 - L_t)\big). \tag{1}$$

Here T_b is the crop's base temperature (0 °C for wheat, 5 °C for canola). L_t is the day length in hours, from the start to the end of civil twilight (the sun 6° below the horizon), from the CBM model of Forsythe et al. (1995). For short-day crops such as soybean and maize, the brake applies to days longer than 12.5 hours instead. With $s = 0$ this is plain thermal time (McMaster and Wilhelm, 1997). Treating day length as a brake on the rate of development follows Ritchie and NeSmith (1991).

Requirements as the trait. Every days-to-heading score is turned into the units that plot had from sowing to the day it headed. A mixed model then splits those requirements:

$$R_{ijk} = \mu + g_i + a_i + e_j + \varepsilon_{ijk}. \tag{2}$$

The terms are, in order: the average requirement, line i's share explained by its relatives (pedigree and DNA), line i's own remainder, site-season j, and plot error. The line terms carry over to a new season. A line with few scores leans on its relatives.

Choosing the day-length brake. The model is fitted with $s = 0$, 0.05, 0.1, 0.2 and 0.3, and each is tested by leaving every site-season out in turn. Plain thermal time stays unless a brake cuts the error by at least 3%. Then the setting with the lowest error wins.

Forecasting a season. For a site in the ground, Crossling adds up the units so far from the observed weather. Next come the forecast days from MET Norway. Then the rest of the season is played out once with each past year's weather at that place, giving up to 15 possible running totals.

Each plot's requirement is taken at ten points across its uncertainty (the 5th to 95th percentile). For every one of those points and every past weather, the model finds the day the requirement is met. Pooling them gives the middle date and a 10–90% range per plot, and a curve of plots expected to head each day at the site.

Learning in season. Once five plots have been scored at a site, the site-season's own pace is estimated from them. The rest of that site's forecasts shift to match.

The same machinery answers two more questions. *Look back* asks what the forecast would have said on a past date, using only what was known then, beside what actually happened. *Plan a season* takes a location, a sowing date and a list of lines, and gives expected heading over the past years' weather.

4 Model 2: how each line will do at each site

A dry June hurts at Lethbridge because the crop is flowering then. The same June at Edmonton may fall weeks before flowering. So the site model measures weather by the crop's own clock, not the calendar.

Stage windows. For each site-season, Crossling takes the heading day h. That is the mean scored heading if at least three plots were scored, otherwise Model 1's prediction, otherwise 55 days. The windows are set from it, in days from sowing (Figure 2). The early window always runs to at least day 10, and frost is only counted from day 10, once the crop is up.

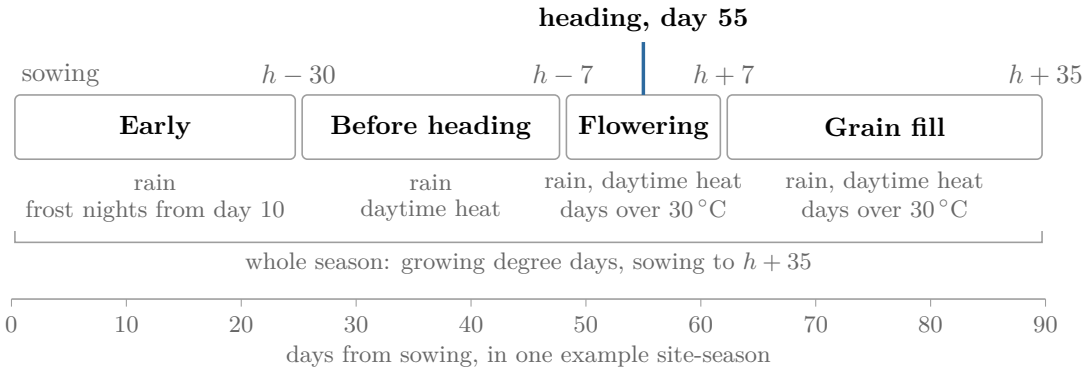


Figure 2: Weather is measured in windows set by each site-season's heading day. The windows move with heading, so a window means the same crop stage at every site.

Eleven covariates summarise each site-season: rain in each of the first four windows; mean daytime high before heading, around flowering and in grain fill; days over 30 °C around flowering and in grain fill; frost nights in the early window, counted from day 10; and growing degree days over the season. Each is standardised across the program's site-seasons.

The model is a reaction norm in the sense of Jarquín et al. (2014):

$$y_{ij} = \mu + s_{l(j)} + \mathbf{w}_j^\top \boldsymbol{\beta} + e_j + g_i + a_i + \sum_{f=1}^{10} \sum_{k=1}^{11} z_{if}\, w_{jk}\, \gamma_{fk} + \varepsilon_{ij}. \tag{3}$$

y_{ij} is line i's mean at site-season j. The station term s_l is what station l is like every year: soil, altitude, how trials are run there. The weather $\mathbf{w}_j$ then shifts the mean through $\boldsymbol{\beta}$, for how this year differs from that station's usual. The site-season e_j holds what neither explains. The line terms g_i and a_i are as in Model 1.

The last term is the line-by-weather response. Each line's first ten relatives features z_{if} are multiplied by each covariate, so lines can differ in their slope on heat, rain or frost. A new line takes its relatives' slopes.

We tried letting lines respond to the main patterns of the weather (principal components), as Jarquín et al. (2014) do with a kernel. In synthetic tests with a known heat tolerance, that version tied the line response to rain, and its estimates correlated 0.25 with the true tolerances. With each covariate on its own, the response was picked up on heat around flowering and correlated 0.87. The Outlook uses each covariate directly.

Predicting a site means running the model on each past year's weather there, sown on that site's usual date. Heading in each year comes from Model 1. For every line this gives:

- the expected value, the middle over those years;
- a range, from the 10th to the 90th percentile;
- a bad year and a good year for the site, at the 10th and 90th percentiles of its mean;
- how far the line's rank moves between years, as the spread of its difference from the site mean.

5 The shared engine

Both models are one mixed model with several random blocks, written so that it stays small however many lines a program has:

$$\mathbf{y} = X\mathbf{b} + \sum_k Z_k \mathbf{u}_k + F\mathbf{u}_F + \boldsymbol{\varepsilon}, \qquad \mathbf{u}_k \sim N(0, \sigma_k^2 I), \quad \mathbf{u}_F \sim N(0, \sigma_F^2 I), \quad \boldsymbol{\varepsilon} \sim N(0, \sigma^2 I). \tag{4}$$

The Z_k are dense blocks: site-season and station indicators, relatives features, weather covariates and their products with relatives. F is the line itself. Each record has exactly one line, so the line equations are absorbed. The system that is solved is only as wide as the Z blocks, however many lines there are.

Relatives as features. The relationship matrix K comes from the pedigree's coefficients of parentage. Where at least ten of the lines have DNA, it becomes the single-step matrix H of Legarra et al. (2009) and Aguilar et al. (2010):

$$H = A + A_{\cdot 2} A_{22}^{-1} (G_w - A_{22}) A_{22}^{-1} A_{2\cdot}, \qquad G_w = 0.95\, G_s + 0.05\, A_{22}. \tag{5}$$

G_s is the genomic matrix scaled to the pedigree's diagonal. The pedigree carries the DNA's information to untested relatives. K is then written as features Z with $ZZ^\top$ close to K: its leading eigenvectors scaled by the square roots of their eigenvalues. Enough are kept to hold 95% of K's trace, at most 100 to 120. What's left goes into each line's own effect. Memory grows with the number of records, not its square, which matters on a 1 GB server.

REML. Variances are estimated by REML (Patterson and Thompson, 1971). The residual variance is profiled out, and the log variance ratios $\tau_k = \log(\sigma_k^2/\sigma^2)$ are found by L-BFGS-B, bounded between e^{-12} and e^{8}. The criterion is

$$-2\ell_R(\boldsymbol{\tau}) = (n-p)\,(\log\hat{\sigma}^2 + 1) + \log|C(\boldsymbol{\tau})| + \sum_k q_k \tau_k, \tag{6}$$

where C is the coefficient matrix of Henderson's (1984) mixed model equations (after absorbing the lines), p the number of fixed effects and q_k the width of block k.

Fast leave-one-out. The equations need only sums over records: $W^\top W$, $W^\top \mathbf{y}$, $\mathbf{y}^\top \mathbf{y}$, and each line's totals. Leaving a site-season out means subtracting its records' sums and solving again. Variance ratios stay at the full-data values, so no new likelihood search is needed.

Speed. On the demo's wheat program, the heading model with all five day-length settings and their leave-out tests takes 0.4 s on a laptop. The nightly fit of every demo program and trait takes 13 s on the production server's two cores, with the maths library limited to one thread.

6 How it is tested

Every fit is tested on site-seasons it hasn't seen, against a simpler model tested the same way. The result is shown beside every forecast.

Leave one site-season out. A site-season is a location in a year. All trials at that place in that year are hidden together, so a forecast never learns from its own neighbours. Each is predicted from the rest using its real weather, and the errors are pooled.

What is measured.

- Heading: mean absolute error in days, the share of plots within 3 days, and bias (early or late overall).
- Sites: the error in each site's mean, and how well the line ranking within each site holds (the correlation between predicted and observed line means, weighted by the number of lines).

A baseline every time. The heading model is compared with the same mixed model on days from sowing. The site model is compared with the same model without the weather terms. The page reports the verdict in plain words, including when the weather isn't helping, for example: "Counting days from sowing does about as well here, so the weather isn't adding much yet."

Minimum data. Heading forecasts start at 3 site-seasons and 30 scored plots. The site model needs 5 site-seasons, each with at least 4 lines. Below that, the page says what's missing instead of guessing.

7 Results on the demonstration program

On Crossling's simulated wheat program, the heading model cut the error from 4.1 days to 1.0. The site model more than halved the error in site-average yield. For stripe rust, lodging and test weight the weather added nothing, and the page says so.

How the demo makes its scores, so it's clear what there was to find:

- **Heading:** each line has a requirement in development units, with a day-length brake of 0.1 per hour. Small site and plot noise is added. The score is the day that station's real NASA POWER weather for that year meets the requirement.
- **Other wheat traits:** a line effect, a station effect, and a hand-set effect for each year from 2019 to 2024, plus noise. The year effects were written to resemble those prairie seasons; 2021, the drought, takes 1,250 kg/ha off yield and 13 cm off height.
- **Yield** also gets a farming trend of 35 kg/ha a year, and a penalty for days over 30 °C in grain fill that differs by line.
- **Canola days to maturity** has no weather in its simulation at all.

The year effects weren't generated from weather; the site model had to find them in the real weather of those years. That makes it a fair test of the covariates. Still, the link between each year and its weather is our own reading of those seasons.

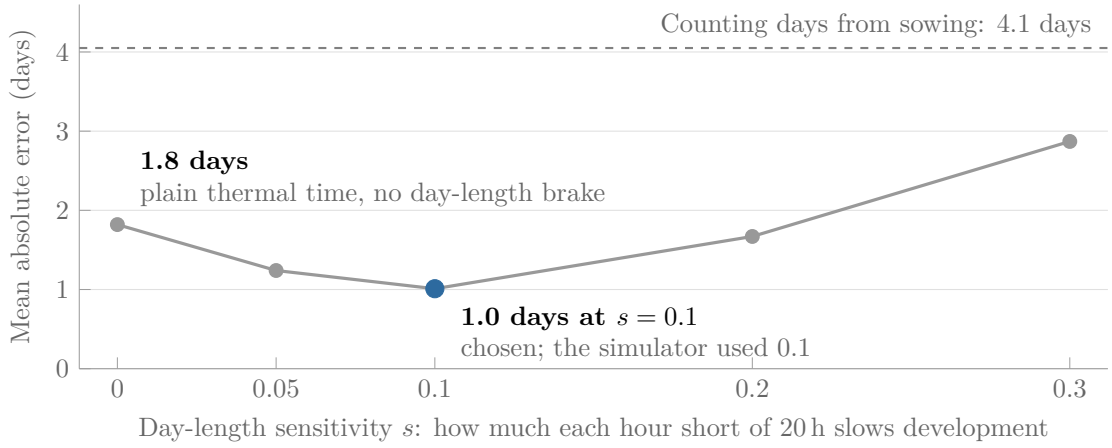


Figure 3: Heading error is lowest at the simulator's true day-length sensitivity. Mean absolute error in days, each of the 22 site-seasons left out in turn, for each candidate setting.

Plain thermal time already beat counting days (Figure 3). The day-length brake the cross-check chose took the error the rest of the way, to the value built into the simulator. Stronger settings over-correct.

Table 2: Leave-one-site-season-out results on the demo program. Errors are mean absolute errors on left-out site-seasons: per plot for heading and maturity, in the site mean for the other traits. Ranking is the within-site correlation of predicted and observed line means.

Trait	Tested on	Outlook error	Baseline	Baseline error	Ranking, Outlook / baseline
Days to heading, wheat	22 site-seasons, 1,468 plots	1.0 days (97% within 3 days)	Counting days from sowing	4.1 days (52% within 3 days)	–
Grain yield	22 site-seasons, 46 trials	223 kg/ha	No weather	515 kg/ha	0.77 / 0.75
Plant height	20 site-seasons, 37 trials	4.0 cm	No weather	5.1 cm	0.82 / 0.82
Grain protein	19 site-seasons, 32 trials	0.3 points	No weather	0.5 points	0.79 / 0.79
Stripe rust	22 site-seasons, 49 trials	0.8 score points	No weather	0.7 score points	0.74 / 0.74
Lodging	5 site-seasons, 9 trials	0.6 score points	No weather	0.6 score points	0.74 / 0.74
Test weight	5 site-seasons, 7 trials	0.5 kg/hL	No weather	0.5 kg/hL	0.82 / 0.82
Days to maturity, canola	4 site-seasons, 171 plots	4.5 days	Counting days from sowing	4.4 days	–

What else the results show.

- **Ranking within a site barely moves with weather**, for example 0.77 against 0.75 for yield. The line-specific heat penalty is small next to the differences between lines. Even so, the response term singled out days over 30 °C in grain fill as the weather lines differ most in, which is the one the simulator built in.
- **The negative control behaves.** Canola maturity has no weather signal in the simulation, and the model finds no gain over counting days. The page tells the user counting days does as well.
- **Weather can hurt slightly** when there's nothing in it to find, as for stripe rust (0.8 against 0.7). The verdict on the page reports that too.

8 What is new here

Each building block is established: thermal time, the photoperiod brake, REML, single-step relationships, reaction norms. What's new is how they are put together and run, inside a field book, for every program every night.

1. **The requirement is the trait.** Heading scores are turned into development-unit requirements before the genetic model sees them. A line's effect is then a property of the line, not of the site and year it happened to be scored in, so it carries to next season's weather. Genomic

prediction of heading often models days to heading directly.

2. **The program's data choose the day-length response.** Day-length sensitivity is picked by leave-out testing on the program's own seasons, with a rule that the simpler model wins unless the brake earns its place.
3. **One model is the other's clock.** The heading model sets the stage windows for the site model, so "heat around flowering" means the same crop stage at an early site and a late one. Environmental covariates in genomic prediction are often summarised by calendar month or by fixed periods from sowing.
4. **A reaction norm small enough to run nightly.** The genotype-by-weather kernel of Jarquín et al. (2014) is written as low-rank features, with the line equations absorbed. Every program's models refit in seconds on a 2-core, 1 GB server.
5. **Ensembles of real past seasons, not one average.** The rest of a season is played out with each of the past 15 years' weather at that place, so a forecast carries the spread of real seasons, not a smoothed normal.
6. **Every forecast comes with its own test.** The leave-out error and a plain-words verdict against a simpler baseline are shown beside each forecast, including when the weather isn't helping.

How it relates to other approaches. Crop growth models integrated with genomic prediction (Technow et al., 2015) capture more physiology. They need a crop model and much heavier computation. Environmental covariates and crop-model outputs have been added to genomic selection before (Heslot et al., 2014; Millet et al., 2019). The Outlook takes a lighter statistical route, built for in-season decisions on one program's data.

9 Limitations and next steps

The biggest gap is real data: everything above was measured on a simulator. The next step is a backtest on real programs' historical trials.

Known limitations.

- **The leave-out tests are slightly optimistic.** Variance ratios are estimated once on all the data and held fixed in each fold. The day-length setting is chosen by the same test it's scored on.
- **Weather covariates overlap.** With around 20 site-seasons and 11 covariates, the model's split of an effect between, say, heat before heading and heat around flowering is unstable. The page ranks drivers by size; their individual numbers shouldn't be over-read.
- **Spring crops only, for now.** There's no vernalisation, so winter wheat and winter canola heading isn't modelled properly yet. Base temperatures are fixed per crop, and there's no upper temperature cutoff.
- **No soil water.** Rain is counted in windows, but there's no water balance and no soil type.
- **Fixed window widths.** The windows (30 and 7 days before heading, 7 and 35 after) suit small-grain cereals. Other crops may need their own.

- **Gridded weather can miss local conditions.** NASA POWER is a coarse grid, which is why a program's own station is used first when it has one.
- **Big programs are capped.** Past 1,500 lines, only the lines with the most records get relatives; the rest keep their own effect only.

Next steps.

1. Backtest on real programs' past trials, scored with the page's own "look back" mode.
2. Re-estimate variance ratios inside each fold from time to time, to measure how optimistic the fixed-ratio test is.
3. Add vernalisation for winter crops, and a simple soil-water balance.
4. Learn the stage windows' widths per crop from the data, as the day-length setting is.

Data sources

- NASA POWER (Prediction Of Worldwide Energy Resources), daily point data: `https://power.larc.nasa.gov/`; daily API documentation: `https://power.larc.nasa.gov/docs/services/api/temporal/daily/`.
- MET Norway Locationforecast 2.0: `https://api.met.no/weatherapi/locationforecast/2.0/documentation`. Data from MET Norway under NLOD 2.0 and CC BY 4.0 (`https://api.met.no/doc/License`).

References

Aguilar I, Misztal I, Johnson DL, Legarra A, Tsuruta S, Lawlor TJ (2010). Hot topic: A unified approach to utilize phenotypic, full pedigree, and genomic information for genetic evaluation of Holstein final score. *Journal of Dairy Science* 93(2):743–752. doi:10.3168/jds.2009-2730

Forsythe WC, Rykiel EJ, Stahl RS, Wu H, Schoolfield RM (1995). A model comparison for daylength as a function of latitude and day of year. *Ecological Modelling* 80(1):87–95. doi:10.1016/0304-3800(94)00034-F

Henderson CR (1984). *Applications of Linear Models in Animal Breeding*. University of Guelph, Guelph, Ontario. ISBN 0-88955-030-1.

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. *Theoretical and Applied Genetics* 127(2):463–480. doi:10.1007/s00122-013-2231-5

Jarquín D, Crossa J, Lacaze X, Du Cheyron P, Daucourt J, Lorgeou J, Piraux F, Guerreiro L, Pérez P, Calus M, Burgueño J, de los Campos G (2014). A reaction norm model for genomic selection using high-dimensional genomic and environmental data. *Theoretical and Applied Genetics* 127(3):595–607. doi:10.1007/s00122-013-2243-1

Legarra A, Aguilar I, Misztal I (2009). A relationship matrix including full pedigree and genomic information. *Journal of Dairy Science* 92(9):4656–4663. doi:10.3168/jds.2009-2061

McMaster GS, Wilhelm WW (1997). Growing degree-days: one equation, two interpretations. *Agricultural and Forest Meteorology* 87(4):291–300. doi:10.1016/S0168-1923(97)00027-0

Millet EJ, Kruijer W, Coupel-Ledru A, Alvarez Prado S, Cabrera-Bosquet L, Lacube S, Charcosset A, Welcker C, van Eeuwijk F, Tardieu F (2019). Genomic prediction of maize yield across European environmental conditions. *Nature Genetics* 51(6):952–956. doi:10.1038/s41588-019-0414-y

Patterson HD, Thompson R (1971). Recovery of inter-block information when block sizes are unequal. *Biometrika* 58(3):545–554. doi:10.1093/biomet/58.3.545

Ritchie JT, NeSmith DS (1991). Temperature and crop development. In Hanks J, Ritchie JT (eds), *Modeling Plant and Soil Systems*, Agronomy Monograph 31, pp. 5–29. ASA, CSSA, SSSA, Madison, WI. doi:10.2134/agronmonogr31.c2

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. doi:10.1371/journal.pone.0130855