

Process-oriented models of autumn leaf phenology: ways to sound calibration and implications of uncertain projections

Michael Meier, Christof Bigler

<https://doi.org/10.5194/gmd-16-1-2023-supplement>

Correspondence to: Michael Meier (michael.meier@cefe.cnrs.fr)

S1 Effects of individual factors – Generalized additive models

We fitted generalized additive models (GAM; Wood, 2017) to estimate the effects of the individual phenology models, optimization algorithms, sampling procedures, sample size, and climate projection scenarios on the root mean square error (RMSE) in site- and species-specific calibration (Eqs. S1 and S2, respectively) and on the 100-year shift at site level (Eq. S1):

$$\mathbf{Y}_{i,j} = \mathbf{X}_{i,j}\boldsymbol{\beta} + s_i(\mathbf{Z}_i) + s_j(\mathbf{Z}_j) + \boldsymbol{\epsilon}_{i,j} \quad (\text{Eq. S1})$$

$$\mathbf{Y}_j = \mathbf{X}_j\boldsymbol{\beta} + s_j(\mathbf{Z}_j) + \boldsymbol{\epsilon}_j \quad (\text{Eq. S2})$$

$\mathbf{Y}_{i,j}$ and \mathbf{Y}_j are the n -dimensional vectors of the response variable $\log(\text{RMSE})$ respective to the site- and species-specific calibration. $\mathbf{X}_{i,j}$ and \mathbf{X}_j are the $n \times p$ matrices of the intercept (i.e. 1) and the $p - 1$ explanatory variables, which are factors for phenology model, optimization algorithm, and sampling procedure and continuous log-transformed sample size proxies (N:q regarding the site-specific or \bar{N} :q and s:S regarding species-specific calibration). $\boldsymbol{\beta}$ is the corresponding p -dimensional vector of the fixed-effects, \mathbf{Z}_i and \mathbf{Z}_j are the $n \times p$ matrices of the sites and species, respectively, s_i and s_j are the corresponding smooth functions that mimic random intercepts, and $\boldsymbol{\epsilon}_{i,j}$ and $\boldsymbol{\epsilon}_j$ are the n -dimensional vectors of the errors [with $\boldsymbol{\epsilon}_{i,j} \sim N(0, \sigma^2 \mathbf{I})$ and $\boldsymbol{\epsilon}_j \sim N(0, \sigma^2 \mathbf{I})$] for the i^{th} site and the j^{th} species. The number of observations (n) and p depend on the calibration mode, while n further depends on whether NA values were replaced and included or not. Regarding the alternative evaluation, $\mathbf{X}_{i,j}$ and \mathbf{X}_j are the $n \times p$ matrices of the intercept (i.e. 1) and the $p - 1$ explanatory variables, which are factors for phenology model, optimization algorithm, and sampling procedure and continuous log-transformed sample size proxies (N:q regarding the site-specific or \bar{N} :q and s:S regarding species-specific calibration) together with the interaction terms between optimization algorithms and sample size proxies.

S1.1 Linear mixed-effect formulation of the generalized additive models

We estimated the influence on RMSE and Δ_{100} with GAMs such that they mimicked linear mixed effects models (LMMs, Pinheiro and Bates, 2000) with random intercepts of the form

$$\mathbf{Y}_{ij} = \mathbf{X}_{ij}\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i + \mathbf{Z}_j\mathbf{b}_j + \boldsymbol{\epsilon}_{ij} \quad (\text{Eq. S3})$$

$$\mathbf{Y}_j = \mathbf{X}_j\boldsymbol{\beta} + \mathbf{Z}_j\mathbf{b}_j + \boldsymbol{\epsilon}_j \quad (\text{Eq. S4})$$

Here, \mathbf{Y}_{ij} and \mathbf{Y}_j are the n -dimensional vectors of $\ln(\text{RMSE})$ or Δ_{100} respective to the site- and species-specific calibration (Eq. S3 and S4) or projections (Eq. S3). \mathbf{X}_{ij} and \mathbf{X}_j are the $n \times p$ matrices of the intercept (i.e. 1) and the $p - 1$ explaining variables and $\boldsymbol{\beta}$ is the corresponding p -dimensional vector of the fixed-effects. \mathbf{Z}_i and \mathbf{Z}_j are the $1 \times n_i$ and $1 \times n_j$ matrices and \mathbf{b}_i and \mathbf{b}_j are the corresponding n_i - and n_j -dimensional vectors of the crossed random effects of the sites and species, respectively. $\boldsymbol{\epsilon}_{ij}$ and $\boldsymbol{\epsilon}_j$ are the n -dimensional vectors of the errors [with $\boldsymbol{\epsilon}_{ij} \sim N(0, \sigma^2\mathbf{I})$ and $\boldsymbol{\epsilon}_j \sim N(0, \sigma^2\mathbf{I})$] for the i^{th} site and the j^{th} species.

We compared our GAM formulation with the above LMM formulation and found the coefficient estimates and corresponding p -values to be sufficiently similar to each other. For the comparison, we fitted the GAMs with R-function `mgcv::bam` (Wood, 2011, 2017) and the LMMs with the R-function `lme4::lmer` (Bates et al., 2015) on a random sample of 1000 calibration runs of our dataset. Thus, we compared coefficient estimates, p -values, and 99% coefficient intervals (results not shown).

S1.2 Back-transformation of effects on log-transformed response variable

The effect on model performance was analyzed with generalized additive models, which fitted factor and log-transformed continuous explanatory variables to the log-transformed root mean square error. The resulting coefficient estimates may be back-transformed for easy interpretation as follows.

Here, the log-transformed \hat{y} is the response variable of a linear equation with intercept $\hat{\alpha}$ and slope $\hat{\beta}_i$ of the explanatory factor variable x_i :

$$\ln(\hat{y}_i) = \hat{\alpha} + \hat{\beta}_i x_i \quad (\text{Eq. S5})$$

Solving for \hat{y} leads to

$$\hat{y}_i = e^{\hat{\alpha}} \times e^{\hat{\beta}_i x_i} \quad (\text{Eq. S6})$$

To derive the relative increase q of \hat{y} when $x_0 = 0$ and $x_l = 1$, we write

$$\frac{\hat{y}_0 (1 + q)}{\hat{y}_0} = \frac{e^{\hat{\alpha}} \times e^{\hat{\beta}_i}}{e^{\hat{\alpha}}} \quad (\text{Eq. S7})$$

which leads to

$$q = e^{\hat{\beta}_i} - 1 \quad (\text{Eq. S8})$$

Here, the log-transformed \hat{y} is the response variable of a linear equation with intercept $\hat{\alpha}$ and slope $\hat{\beta}_i$ of the log-transformed continuous explanatory variable x_i :

$$\ln(\hat{y}_i) = \hat{\alpha} + \hat{\beta}_i \ln(x_i) \quad (\text{Eq. S9})$$

Solving for \hat{y} leads to

$$\hat{y}_i = e^{\hat{\alpha}} \times e^{\ln(x_i)\hat{\beta}_i} = e^{\hat{\alpha}} \times x_i^{\hat{\beta}_i} \quad (\text{Eq. S10})$$

To derive the relative increase q of \hat{y} when x_i observes a relative increase of p , we write

$$\frac{\hat{y}_i(1+q)}{\hat{y}_i} = \frac{e^{\hat{\alpha}} \times [x_i(1+p)]^{\hat{\beta}_i}}{e^{\hat{\alpha}} \times x_i^{\hat{\beta}_i}} \quad (\text{Eq. S11})$$

which leads to

$$q = (1+p)^{\hat{\beta}_i} - 1 \quad (\text{Eq. S12})$$

and may be approximated for $p = 0.01$ by

$$q \approx p \times \hat{\beta}_i \quad (\text{Eq. S13})$$

S2 Influence of aggregated factors – ANOVA

To estimate the relative influence (i.e. the explained variance) of aggregated factors (I_i) of phenology models, optimization algorithms, sampling procedures, or climate projections as well as of size proxies on either model performance or model projections, we derived a type-III ANOVA (Yates, 1934; Herr, 1986; Chandler and Scott, 2011) from the corresponding GAM output and thus divided the sum of squares (SSQ) of each factor family i by the total sum of squares (Eq. S14):

$$I_i = \frac{SSQ_i}{\sum_{i=1}^n SSQ_i} \quad (\text{Eq. S14})$$

SSQ_i and corresponding p -values were calculated with the R-functions `stats::aov` and `stats::drop1` (R Core Team, 2022).

References

- Bates, D., Machler, M., Bolker, B. M., and Walker, S. C.: Fitting Linear Mixed-Effects Models Using lme4, *J. Stat. Softw.*, 67, 1-48, 2015.
- Chandler, R. E. and Scott, E. M.: Statistical methods for trend detection and analysis in the environmental sciences, *Statistics in practice*, Wiley, Chichester, 368, 2011.
- Herr, D. G.: On the History of ANOVA in Unbalanced, Factorial Designs: The First 30 Years, *The American Statistician*, 40, 265-270, 10.2307/2684597, 1986.
- Pinheiro, J. C. and Bates, D. M.: Mixed-effects models in S and S-PLUS, *Statistics and computing*, Springer, New York, 528, 2000.
- R Core Team: R: A language and environment for statistical computing. R Foundation for Statistical Computing, 2022.
- Wood, S. N.: Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73, 3-36, <https://doi.org/10.1111/j.1467-9868.2010.00749.x>, 2011.
- Wood, S. N.: Generalized Additive Models: An Introduction with R, 2nd edition, Chapman and Hall/CRC, New York, <https://doi.org/10.1201/9781315370279>, 2017.
- Yates, F.: The Analysis of Multiple Classifications with Unequal Numbers in the Different Classes, *J. Am. Stat. Assoc.*, 29, 51-66, 10.2307/2278459, 1934.