INTRODUCTION

Periodically recurring, often seasonal, biological events (phenology) are influenced by environmental factors and interactions between organisms (Lieth, 1974). Such phenomena are of particular interest because anthropogenic influences, such as climate change, might alter important ecological processes that are intimately correlated (Forrest & Miller-Rushing, 2010). A biologically meaningful description of phenological events is essential to understanding their temporal dynamics, and offers an opportunity to assess the significance of its potential drivers (Rafferty, Caradonna, Burkle, Iler, & Bronstein, 2013).

Certain phenological events are recorded as a binary change from one recognizable state into another, either for a whole organism, as when a winter migrant has arrived for the breeding season (Gordo, 2007) or for individual parts, as when individual leaf or flower buds on a plant burst (Cole & Sheldon, 2017).

While it is clear that variability in individual events is expressed at the population level as a time distribution, phenological observations are often restricted to recording only extreme events, such as the date of the first flower to bloom (Fitter & Fitter, 2002) or the first migrant of the season to arrive (Gordo & Sanz, 2006). This approach ignores the population-level dynamics which contains a wealth of information about the ecological processes driving the events.

APPLICATION

NLSTIMEDIST: An R package for the biologically meaningful quantification of unimodal phenology distributions

Nicola C. Steer | Paul M. Ramsay | Miguel Franco

School of Biological and Marine Sciences, Plymouth University, Plymouth, UK

Correspondence
Nicola C. Steer
Email: nicola.steer@plymouth.ac.uk

Funding information
School of Biological and Marine Sciences, University of Plymouth, Grant/Award Number: QR Scholarship

Handling Editor: Samantha Price

Abstract

1. Phenological investigation can provide valuable insights into the ecological effects of climate change. Appropriate modelling of the time distribution of phenological events is key to determining the nature of any changes, as well as the driving mechanisms behind those changes.

2. Here, we present the NLSTIMEDIST R package, a distribution function and modelling framework that describes the temporal dynamics of unimodal phenological events. The distribution function is derived from first principles and generates three biologically interpretable parameters.

3. Using seed germination at different temperatures as an example, we show how the influence of environmental factors on a phenological process can be determined from the quantitative model parameters.

4. The value of this model is its ability to represent various unimodal temporal processes statistically. The three intuitively meaningful parameters of the model can make useful comparisons between different time periods, geographical locations or species' populations, in turn allowing exploration of possible causes.

KEYWORDS
biologically interpretable parameters, climate change, model fitting, nonlinear regression, phenological processes, phenology, temporal dynamics, time distribution
information regarding, for example, the duration of the phenomenon, its temporal skew and its shape. Other scalar values may be conveniently chosen thresholds (Zhang et al., 2003), such as the 50% of completion commonly used in the investigation of canopy phenology (Richardson, Bailey, Denny, Martin, & O’Keefe, 2006), and varying percentage values are a key feature of the BBCH scale used to identify phenological developmental stages in plants (Meier, 2001). All of these approaches result in a single date, which is intended to capture useful information about the phenological process.

When single dates are used to describe the timing of a phenological event, they are often compared across years or linked to changes in an environmental condition, such as temperature, using regression (Sparks & Tryjanowski, 2010). As useful as these scalars may be to summarize key features and changes in phenology, they inevitably miss potentially important information about the shape of the overall time course (CaraDonna, Iler, & Inouye, 2014; Carter, Saenz, & Rudolf, 2018; Clark & Thompson, 2011).

A more thorough assessment should aim to model the entire phenological time distribution (CaraDonna et al., 2014; Carter et al., 2018). This is frequently accomplished using classic growth functions, such as the logistic and Richards (Richardson et al., 2006; Sun & Freligh, 2011; Yin, Goudriaan, Lantinga, Vos, & Spiertz, 2003; Zhang et al., 2003), as their sigmoid shape resembles the time course of a phenological event. The logistic model is symmetrical around its point of inflection which is always halfway along the asymptotes (Birch, 1999), but there is no theoretical basis for a phenological event to be symmetrical around its mid-point. The Richards (or generalized logistic) model is more flexible due to an additional shape parameter but its parameters cannot be interpreted in a meaningful way (Birch, 1999; Damgaard & Weiner, 2008; Richards, 1959; Zeide, 1993).

An alternative approach builds on an existing body of work on niche overlap (Castro-Arellano, Lacher, Willig, & Rangel, 2010; Fleming & Partridge, 1984; Pleasants, 1980; Totland, 1993), and allows species interactions to be compared as measures of temporal overlap (e.g., Carter et al., 2018). These approaches take account of whole phenological distributions through time, accommodating multimodal or skewed responses. Temporal overlap is an outcome of interactions between distributions rather than a direct consideration of their shapes, but is a sensible approach where the comparison focuses on time alone and where there are multimodal, complex probability distributions.

For unimodal phenology distributions, a model that describes the entire phenological time distribution well, is sufficiently flexible to accommodate asymmetrical distributions, and generates biologically interpretable shape parameters would be more useful. In particular, the model should be derived from basic principles applicable to a wide spectrum of biological time distributions. Importantly, goodness-of-fit alone should not be used to justify model selection; it is always preferable to choose a model that has biologically meaningful parameters (Paine et al., 2012).

Here, we present a model for describing the temporal dynamics of unimodal phenological events. It has been derived from first principles and generates biologically meaningful parameters that can be compared and used to assess potential driving mechanisms.

## 2 | THE MODEL

A phenological process of events (y) unfolding over time (x) at a constant rate (r) would follow an exponential distribution (Franco, 2018). Phenological processes, however, do not occur at a constant rate (Sparks & Tryjanowski, 2010) and individual events are more likely to be distributed according to a probabilistic process described by the inverse logit governed by an additional parameter, c (Franco, 2018). Finally, phenological processes do not occur instantly, but happen sometime after exposure to a specific set of conditions (Wu et al., 2015), which requires a third parameter, the time-lag, t. By incorporating the lagged form of the inverse logit function into the exponential distribution, a suitable biological time distribution can be derived (Franco, 2018). This cumulative distribution function (cdf) has the form:

\[ y = \left(1 - \frac{r}{1 + e^{-ct}}\right)^x \]

The derivative of this function quantifies the probability density function (pdf):

\[ \frac{dy}{dx} = \left(1 - \frac{r}{1 + e^{-ct}}\right)^x \times \left(1 - \frac{r}{1 + e^{-ct}}\right) \times \ln \left(1 - \frac{r}{1 + e^{-ct}}\right) \]

which describes the population-level rate at which the phenomenon occurs. Each of the function’s parameters has clear meaning and units: r quantifies the maximum proportional rate at which the process occurs (it is dimensionless); c is the rate at which r converges on its maximum value (units: time^{-1}); and t is an overall measure of the process’ time-lag (units: time) (Figure 1). Parameter t can also be thought of as a weighted measure of the process’ duration – weighted in relation to the values of r and c, that is. It correlates with, but is not equivalent to, any of the distribution’s various measures of central tendency.

### 2.1 | THE R PACKAGE

**nlstimedist** is an R package that provides a convenient way to fit the time course of a unimodal phenological time distribution employing nonlinear regression. **nlstimedist** combines functions for data preparation, model fitting and data visualization into one complete package, allowing efficient, accurate and meaningful analysis.

The model is fitted to data using the `timedist()` function. The function requires data in the form of the proportion of cumulative number of events through time, together with column identifiers (allowing the analysis of multi-column data) and starting values for r, c and t. If data are in their raw form of counts versus time, they can...
be cleaned and converted to proportions (range: 0–1) for model use, using the built-in tidy function `tdData()`.

The `timedist()` function returns an object which contains all of the fitted model information. This includes the equation used to fit the estimated time distribution, estimated values for \( r, c \) and \( t \), the model fit’s residual sum of squares, and the number of iterations to convergence. The object can be examined with all of the generic `nls` functions, such as `summary()`, and can also be used by packages such as ‘nlstools’ (Baty et al., 2015). Functions and packages such as these can be used to assess how well the model fits the data and the reliability of the parameter estimates. The statistical moments and percentiles of the fitted distribution can be obtained from the model object. The `nlstimedist` package also has two built-in functions for plotting the estimated time distribution as either a cumulative distribution function (cdf) `tdCdfPlot(object, …)` or a probability density function (pdf) `tdPdfPlot(object, …)`.

`nlstimedist` is based on the framework provided by nlsLM from the `minpack.lm` package (Elzhov, Mullen, Spiess, & Bolker, 2016). nlsLM is a modification of the standard nls function that uses the Levenberg–Marquardt algorithm (Marquardt, 1963) for model fitting (Elzhov et al., 2016). This fitting procedure was chosen because it is considered robust (Lourakis, 2005). Because the method of nonlinear regression fitting uses an iterative optimization procedure to converge on the least squares solution, fairly accurate starting values need to be chosen (Ruckstuhl, 2010). `nlstimedist` is not a self-starting model, therefore, guidelines are provided to assist with the selection of appropriate starting values for the three parameters (see package vignette).

Fitting to the underlying cumulative distribution function (as opposed to the more usual practice of fitting a probability density function to binned data) allows datasets with few observations to be analysed. The temporal resolution of the data must be sensible and representative of the whole phenology under investigation. This model cannot be applied to complex, multimodal phenologies.

As shown in Figure 1, each parameter has a unique effect on three different aspects of the distribution’s shape. In summary, \( r \) is a scaled rate of completion (without units), \( c \) is a measure of its temporal concentration (units: time\(^{-1}\)) and \( t \) is an overall measure of temporal delay (units: time). In combination, these parameters provide insight into potential drivers and mechanisms associated with the phenological process, such as rates of development, climate change, competition between species, genetic diversity, resource availability and environmental heterogeneity. Exploring the relationships between model parameters and statistical moments with biological and environmental variables might offer additional understanding of possible determinants.
3 | APPLICATION OF THE MODEL

The model can be applied across a wide range of phenological studies, including aspects of reproduction and development (e.g., pollination, gestation, egg laying, egg hatching, germination, life stages), seasonal population dynamics (of leaves, flowers, whole organisms, etc.), species interactions (trophic mismatch, predator–prey dynamics, competition, pest outbreaks), migration and dispersal (in relation to cues and invasion dynamics), and mortality in response to environmental challenge (climate change, ecotoxicology). The model has also been fitted successfully to the distribution of reproductive value of perennial plants as a means of quantifying the duration (by parameter \( t \)) and the speed (parameter \( c \)) of life (Mbeau-Ache & Franco, 2013).

As a worked example, we present data from a controlled seed germination experiment for *Puya raimondii*, a giant rosette plant from the Andes. The experiment tested the effect of temperature on germination along a temperature gradient ranging from 8.4°C to 23.7°C. We use this example to illustrate how the new function is able to quantify accurately the changing temporal dynamics of a phenological process. We also show how quantification of the model’s parameters can be used to determine the influence that an environmental factor, in this case temperature, has on seed germination.

The dataset used in this example is available on the Dryad Digital Repository. The file can be read directly into R using the following command.

```r
> Puya <- read.csv("PuyaGermination.csv", header = TRUE, sep = ",")
```

To obtain the estimated parameter values \( (r, c, t) \) at each temperature, the model was fitted to each column in the ‘Puya Germination’ dataset separately using the `timedist()` function. Starting values for parameter estimates are dependent on the length of the time course under investigation and as such, starting values were adjusted for each model fit.

```r
> Puya1.1 <- timedist(Puya, x = "x", y = "T8.4", r = 0.04, c = 0.5, t = 40)
```

### TABLE 1

| Temp. (°C) | N  | \( y_{\text{max}} \) (%) | \( r \) | \( c \) | \( t \) | \( R^2 \) | Skew | Kurtosis | Entropy |
|------------|----|--------------------------|-----|-----|-----|-----|-----|-------|-------|
| 8.4        | 148| 74.0                      | 0.073(0.005)*** | 0.447(0.025)*** | 37.368(0.345)*** | 0.497 | 35.661 | 5.892 | 4.156 |
| 9.3        | 156| 78.0                      | 0.053(0.003)*** | 0.653(0.032)*** | 29.592(0.158)*** | 0.448 | 29.461 | 6.334 | 4.848 |
| 12.5       | 161| 80.5                      | 0.012(0.002)*** | 0.806(0.002)*** | 22.018(0.230)*** | 0.448 | 16.021 | 3.925 | 4.621 |
| 13.8       | 164| 82.0                      | 0.129(0.004)*** | 1.485(0.043)*** | 16.993(0.071)*** | 0.417 | 15.970 | 4.218 | 5.188 |
| 14.7       | 167| 79.5                      | 0.176(0.005)*** | 1.418(0.048)*** | 14.020(0.053)*** | 0.417 | 14.452 | 3.918 | 4.774 |
| 16.7       | 157| 76.5                      | 0.339(0.008)*** | 1.418(0.048)*** | 11.193(0.073)*** | 0.417 | 14.777 | 3.918 | 4.774 |
| 19.5       | 159| 79.5                      | 0.076(0.005)*** | 0.801(0.025)*** | 15.970(0.073)*** | 0.417 | 16.061 | 4.431 | 3.893 |
| 20.0       | 155| 77.5                      | 0.055(0.003)*** | 0.507(0.025)*** | 13.970(0.071)*** | 0.417 | 15.580 | 3.982 | 4.774 |
| 21.7       | 146| 75.0                      | 0.076(0.005)*** | 0.283(0.025)*** | 12.193(0.073)*** | 0.417 | 14.542 | 3.571 | 5.188 |
| 22.4       | 144| 72.0                      | 0.076(0.005)*** | 0.283(0.025)*** | 11.970(0.073)*** | 0.417 | 14.154 | 3.418 | 5.144 |
| 23.7       | 88 | 44.0                      | 0.052(0.005)*** | 0.203(0.015)*** | 10.970(0.073)*** | 0.417 | 14.542 | 3.571 | 5.188 |

Fitting accuracy was verified using a range of functions. The reliability of the parameter estimates was obtained for each fit using the generic `summary()` function for `nls` objects. Standard errors of
parameter estimates were very small, and model fit was highly significant in all cases ($p < .001$; Table 1).

\begin{verbatim}
> summary(Puya1.1, correlation = TRUE, symbolic.cor = FALSE)

Formula: T8.4 ~ 1 - (1 - (r/(1 + exp(-c * (x - t))))^x

Parameters:

  Estimate  Std. Error  t value   Pr(>|t|)
  r  0.073472  0.005016     14.65     4.51e-11***
  c  0.447144  0.024582     18.19     1.40e-12***
  t 37.367538  0.344737    108.39      < 2e-16***

---

Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.02214 on 17 degrees of freedom

Correlation of Parameter Estimates:

     r    c
  r 1.00 -0.64
  c -0.64 1.00

Nonlinear regression has no direct $R^2$. However, a pseudo $R^2$

\[ R^2 = 1 - \frac{\sum(y - \hat{y})^2}{\sum(y - \bar{y})^2} \]

calculated as $1 - \frac{\sum(y - \hat{y})^2}{\sum(y - \bar{y})^2}$, which defines a similar quantity for nonlinear regression and is able to describe the proportion of variance explained by the model (Cameron & Windmeijer, 1997; Kvålseth, 1985). Extracting this quantity from each model object provided another measure of how well the model fitted the data. $R^2$

was over 0.99 for all temperature treatments (Table 1), although we recommend caution in the interpretation of this statistic, as it provides an over-optimistic measure of fit (Spiess & Neumeyer, 2010).

\begin{verbatim}
> Puya1.1$m$rss()
[1] 0.99693
\end{verbatim}

The statistical moments and the percentiles of the distribution can also be extracted from each model object. These facilitate comparison of different temperature treatments throughout time (Table 1).

\begin{verbatim}
> Puya1.1$m$getMoments()

  mean  variance     sd     skew     kurtosis  entropy
  1  35.66111 34.71189  5.891679  4.156171  36.20471
   4.096621
\end{verbatim}

\textbf{FIGURE 2} Cumulative distribution functions (left) and corresponding probability density functions (right) for 

\textit{Puya raimondii} germination occurring along a temperature gradient ranging from 8.4°C to 23.7°C. Probability density functions describe the population-level rate of germination and the area under each curve is equal to the maximum percentage of germination.

\textbf{FIGURE 3} The relationship between temperature and the values of parameter estimates produced from each model fit (a) parameter $r$, (b) parameter $c$ and (c) parameter $t$. All three quadratic relationships were significant (a) $R^2 = 0.915, p < .000$, (b) $R^2 = 0.672, p = .007$, (c) $R^2 = 0.945, p < .000$. Error bars represent the standard errors of parameter estimates.
Plotting the model fits as both cumulative distribution functions and probability density functions provide a useful summary of how germination is affected across a range of temperatures (Figure 2). These plots provide an informative visual summary of the maximum per capita rate of germination, temporal spread and time delay of seed germination at each temperature.

A key feature of the model is the production of numerically meaningful parameter values. These parameters, when plotted against biological or environmental variables, allow potential driving mechanisms to be tested. In this example, temperature affected all three parameter estimates in a curvilinear fashion (Figure 3). Parameters $r$, $c$ and $t$ displayed significant quadratic relationships with temperature, helping to identify the temperature at which germination was fastest, more concentrated and least delayed after sowing. This optimal temperature was remarkably similar for all three parameters: $r = 15.6^\circ C$, $c = 15.5^\circ C$ and $t = 15.9^\circ C$.

Although the quadratic relationship with temperature was significant and each parameter predicted similar optimal temperatures, there is no reason to expect either a similar optimum for all three parameters or a symmetrical response on either side of the optima. The analysis of other phenological processes may yield different statistical relationships. Temperature was used in this example to illustrate the effect that an environmental factor has on the time course of seed germination. However, the same principles would apply to other environmental conditions that vary on a continuous scale.

### 4 | CONCLUSIONS

The NLSIMDIST package was built to facilitate the application of Franco's (2018) distribution function to phenological data. The model adequately describes a unimodal phenological process of events that are usually recorded as completions, that is, on a binary scale. It is conceptually simple and is able to capture the essence of a phenological process because its three parameters quantify properties of the distribution with known units: a maximum net per capita rate (dimensionless), a rate at which this maximum rate is achieved (units: time$^{-1}$) and an overall measure of the process' time-lag (units: time).

Both biological and environmental variables have been shown to affect the individual parameters in a predictable way (Franco, 2018; Mbeau-Ache & Franco, 2013; and examples provided here). The flexibility of the model in representing various continuous distributions, the interpretability of its parameters and its ability to estimate the underlying statistical distribution of an often highly asymmetrical temporal process make it a useful tool in the analysis of unimodal phenological phenomena.

### AUTHORS' CONTRIBUTIONS

M.F. conceived the idea and designed the distribution function; P.M.R. collected the data; N.C.S. analysed the data, interpreted the results and led the writing of the manuscript. All authors contributed critically to the drafts and gave final approval for publication.

### DATA AVAILABILITY STATEMENT

The package is available on CRAN [https://cran.r-project.org/package=nlsimdist](https://cran.r-project.org/package=nlsimdist) and the data and R script used in this study are available on the Dryad Digital Repository [https://doi.org/10.5061/dryad.f01pr47](https://doi.org/10.5061/dryad.f01pr47) (Steer, Ramsay, & Franco, 2019).

### CITATION OF NLSIMDIST

Studies using NLSIMDIST should cite this article.

### ORCID

Nicola C. Steer [https://orcid.org/0000-0002-6446-8601](https://orcid.org/0000-0002-6446-8601)

Paul M. Ramsay [https://orcid.org/0000-0001-7631-6480](https://orcid.org/0000-0001-7631-6480)

Miguel Franco [https://orcid.org/0000-0002-7249-4981](https://orcid.org/0000-0002-7249-4981)

### REFERENCES

Baty, F., Ritz, C., Charles, S., Brutsche, M., Flandrois, J.-P., & Delignette-Muller, M.-L. (2015). A toolbox for nonlinear regression in R: The package nlsstools. *Journal of Statistical Software, 66*, 1–21. [https://doi.org/10.18637/jss.v066.i05](https://doi.org/10.18637/jss.v066.i05)

Birch, C. P. D. (1999). A new generalized logistic sigmoid growth equation compared with the Richards growth equation. *Annals of Botany, 83*, 713–723. [https://doi.org/10.1006/anbo.1999.0877](https://doi.org/10.1006/anbo.1999.0877)

Cameron, C. A., & Windmeijer, F. A. G. (1997). An R-squared measure of goodness of fit for some common nonlinear regression models. *Journal of Econometrics, 77*(2), 329–342. [https://doi.org/10.1016/S0304-4076(96)01818-0](https://doi.org/10.1016/S0304-4076(96)01818-0)

CaraDonna, P. J., Iler, A. M., & Inouye, D. W. (2014). Shifts in flowering phenology reshape a subalpine plant community. *Proceedings of the National Academy of Sciences, 111*, 4916–4921. [https://doi.org/10.1073/pnas.1323073111](https://doi.org/10.1073/pnas.1323073111)

Carter, S. K., Saenz, D., & Rudolf, V. H. W. (2018). Shifts in phenological distributions reshape interaction potential in natural communities. *Ecology Letters, 21*, 1143–1151. [https://doi.org/10.1111/ele.13081](https://doi.org/10.1111/ele.13081)

Castro-Arellano, I., Lacher, T. E. Jr, Willig, M. R., & Rangel, T. F. (2010). Assessment of assemblage-wide temporal niche segregation using null models. *Methods in Ecology and Evolution, 1*, 311–318. [https://doi.org/10.1111/j.2041-210X.2010.00031.x](https://doi.org/10.1111/j.2041-210X.2010.00031.x)

Clark, R. M., & Thompson, R. (2011). Sampling bias in the determination of first and last occurrences. *Plant Ecology & Diversity, 4*, 201–211. [https://doi.org/10.1016/j.pld.2011.04.007](https://doi.org/10.1016/j.pld.2011.04.007)

Cole, C. B., & Sheldon, B. C. (2017). The shifting phenological landscape: Within- and between-species variation in leaf emergence in a mixed-deciduous woodland. *Ecology and Evolution, 7*, 1135–1147. [https://doi.org/10.1002/ece3.2718](https://doi.org/10.1002/ece3.2718)

Damgaard, C., & Weiner, J. (1998). Modeling the growth of individuals in crowded plant populations. *Journal of Plant Ecology, 1*, 111–116. [https://doi.org/10.1093/jpe/tn008](https://doi.org/10.1093/jpe/tn008)

Ezhov, T. V., Mullen, K. M., Spiess, A.-N., & Bolker, B. (2016). minpack. Im: r interface to the Levenberg-Marquardt nonlinear least-squares algorithm found in Minpack, Plus Support for Bounds.
