Ecological Archives **E089-082-A2**

Lindsay A. Turnbull, Cloé Paul-Victor, Bernhard Schmid, and Drew W. Purves. 2008. Growth rates, seed size, and physiology: Do small-seeded species really grow faster? ***Ecology*** **89:1352–1363.**

Appendix B. PARAMETER ESTIMATION FOR THE DAILY GROWTH MODEL.

The likelihood analysis used measurements of above- and belowground biomass of individual plants, growing in different temperature and nutrient regimes, to estimate the parameters in the daily growth model (see main text and Eqs. 1 - 4). The first step of a likelihood analysis is to define the log-likelihood of the data **X**, given the model, and parameter values $\boldsymbol{\theta}$, $\ell(\mathbf{X}|\boldsymbol{\theta})$ (hereafter referred to as the likelihood only). Given the definition of $\ell(\mathbf{X}|\boldsymbol{\theta})$, the analysis proceeds by finding the parameters $\boldsymbol{\theta}$ that maximize $\ell(\mathbf{X}|\boldsymbol{\theta})$ (the so-called maximum-likelihood estimates, or MLEs), and the confidence intervals on each parameter. In special cases involving simple models and simple error structures (e.g., linear regression), the MLEs and confidence intervals can be found analytically. For complex, nonlinear models, such as those employed here, this is not possible, and computational methods are needed to explore different parameter sets. Nonetheless, the underlying approach in these more complex cases is formally identical to simpler, more familiar statistical methods, including (for example) linear regression and ANOVA. For an introduction to likelihood analysis in ecology see Hilborn and Mangel (1997).

The sections below explain the definition of the likelihood, including the error structure employed in the statistical model, how model predictions were generated given a particular parameter set, and how the data were manipulated to produce a set of observations, against which the model predictions could be compared. The final section details the Monte Carlo Markov Chain (MCMC) computational methods employed to find the MLEs and confidence intervals on parameters.

DEFINITION OF THE LIKELIHOOD

The analysis begins by defining $\ell(\mathbf{X}|\boldsymbol{\theta})$, the likelihood of the observed data **X**, given a vector of model parameters $\boldsymbol{\theta}$. In this case, the vector **X** consisted of a total of 2556 destructive measurements of above- and belowground biomass, and 7107 nondestructive measurements of above- and belowground biomass, taken from 1724 individuals of the nine different species, growing in different temperature and nutrient regimes, and taken at various times after the beginning of the experiment (see the main text).

The vector $\boldsymbol{\theta}$ consisted of 13 parameters for each of the nine species *j*. Nine of these parameters (G, M_{ref}, γ_0, γ_{mass}, γ_{nutr}, T_{opt}, σ_t, α, μ) governed the behavior of the daily growth model itself, including the size-independent growth coefficient , allocation above- and belowground, and frost tolerance (see Eqs. 5 - 9 in the main text). The other four parameters ($\rho_{abv,dest}, \rho_{blw,dest}, \rho_{abv,non-dest}, \rho_{blw,non-dest}$) set the magnitude of individual-to-individual variation in biomass, which was required to define the error distributions (see below). The model was fit several different times, with the parameters in $\boldsymbol{\theta}$ either being global (shared between all species), or species-specific. This meant that the number of parameters to be estimated also varied from 9 (all parameters global) to 117 (all parameters species-specific).

Conditional on a particular vector $\boldsymbol{\theta}$, the daily growth model gives a prediction corresponding to any observation *k*, where *k* is an observation of above- or belowground biomass, taken $\hat{d}(k)$ days after the start of the experiment, for an individual of species *j*, growing in nutrient regime *nutr*, and temperature regime *temp*. Comparing all available observations *k* with model predictions, and assuming log-normal error distributions, yields the likelihood $\ell(\mathbf{X}|\boldsymbol{\theta})$. The key assumption here is that each observation *k* is a random draw from a log-normal probability distribution with a mode equal to the prediction from the daily growth model:

$$\ell(\mathbf{X}\,|\,\boldsymbol{\theta}) = \sum_{k\in S(abv,dest)} \ln\left[normal\left(\ln\{M_k^{obs}\},\ln\{M_k^{mod}\},\rho_{abv,dest}\right)\right] + \sum_{k\in S(blw,dest)} \ln\left[normal\left(\ln\{M_k^{obs}\},\ln\{M_k^{mod}\},\rho_{blw,dest}\right)\right] + \sum_{k\in S(abv,non-dest)} \ln\left[normal\left(\ln\{M_k^{obs}\},\ln\{M_k^{mod}\},\rho_{abv,non-dest}\right)\right] + \sum_{k\in S(blw,dest)} \ln\left[normal\left(\ln\{M_k^{obs}\},\ln\{M_k^{mod}\},\rho_{blw,dest}\right)\right] \quad \text{(B2.1)}$$

where *dest* and *non-dest* signify destructive and non-destructive measurements of biomass respectively. The set $S(abv,dest)$ contain observations *k* that consist of an observation of aboveground biomass, destructively harvested; and similarly for $S(blw,dest)$, $S(abv,non-dest)$ and $S(blw,non-dest)$ (see below). M_k^{obs} is the biomass measurement associated with observation *k*, and M_k^{mod} is the model prediction for the biomass associated with *k*. M_k^{mod} is conditional on the parameter vector $\boldsymbol{\theta}$, making $\ell(\mathbf{X}\,|\,\boldsymbol{\theta})$ a function of $\boldsymbol{\theta}$. The function $normal\left(\ln\{M_k^{obs}\},\ln\{M_k^{mod}\},\sigma_{abv,dest}\right)$ is the probability density for the observation $\ln\{M_k^{obs}\}$, given a normal probability distribution with mean equal to $\ln\{M_k^{mod}\}$, and standard deviation equal to $\rho_{abv,dest}$, and similarly for the other three *normal* functions. Note that the standard deviation is specific to each of the four combinations of above- vs. belowground biomass, and destructive vs. nondestructive harvesting (i.e., $\rho_{abv,dest}$, $\rho_{blw,dest}$, $\rho_{abv,non-dest}$, $\rho_{blw,non-dest}$ are estimated as separate parameters). This is because ρ includes both true individual-to-individual variation in biomass and measurement error, both of which can be expected to differ according to species, above- vs. belowground biomass and destructive vs. nondestructive harvesting.

The statistical model underlying Eq. B2.1 can be written as

$$\ln\{M_k^{obs}\} = \ln\{M_k^{mod}(\theta)\} + \varepsilon_k \quad \text{(B2.2)}$$

where ε_k is an unexplained random deviation specific to observation *k*, which is assumed to come from a normal distribution with a mean of zero, and standard deviation equal to the ρ corresponding to *k* (e.g. $\rho_{abv,dest}$ if *k* corresponds to a destructive measurement of above-ground biomass). The analysis also assumes independence of observations. Thus, the analysis is a standard nonlinear least-squares regression, but carried out on the log-transformed values of the data, and model predictions. This implies that the most likely value of $\ln\{M_k^{obs}\}$, i.e., the mode of the log-normal distribution, is $\ln\{M_k^{mod}\}$, which in turn implies that the most likely value of M_k^{obs} is M_k^{mod}. However, the arithmetic mean of a sample taken from a log-normal distribution is not equal to the mode, which means in this case that the daily growth model (once parameterized) is not expected to reproduce the arithmetic mean of the observational data. For this reason, when comparing the model predictions to the data, we present the geometric mean of the data (e.g., see Figs. 3 and 4), which is expected to be equal to the mode of the log-normal distribution, and hence equal to the prediction from the daily growth model.

GENERATION OF MODEL PREDICTIONS

To generate the likelihood $\ell(\mathbf{X}|\boldsymbol{\theta})$, it was necessary to predict the biomass corresponding to observation *k*, M_k^{mod} (see above). These predictions were generated differently for destructive and non-destructive measurements.

Destructive measurements. For destructive measurements, the daily growth model was initialized from the earliest available observations of above- and belowground biomass for species *j* (i.e., those taken on day 14). Then, the predicted biomass for observation *k* was

generated by summing the changes in biomass up to the harvest day associated with observation *k*. Thus, where observation *k* is for plant *i* of species *j* harvested on day $\hat{d}(k)$:

$$M_{j,14}^{(abv,mod)} = \overline{M}_{j,14}^{(abv,obs)}$$

$$M_{j,14}^{(blw,mod)} = \overline{M}_{j,14}^{(blw,obs)}$$

$$M_{k}^{(abv,mod)} = M_{j,14}^{(abv,mod)} + \sum_{d=14}^{d=\hat{d}(k)} \Delta M_{k,d}^{(abv,mod)}$$

$$M_{k}^{(blw,mod)} = M_{j,14}^{(abv,mod)} + \sum_{d=14}^{d=\hat{d}(k)} \Delta M_{k,d}^{(blw,mod)} \qquad \text{(B2.3)}$$

where $\overline{M}_{j,14}^{(abv,obs)}$ is the geometric mean of all destructive measurements of aboveground biomass for species *j* taken on day 14; similarly for $\overline{M}_{j,14}^{(blw,obs)}$; and where the superscript *mod* denotes the values of biomass, and change in biomass, used in the daily growth model. The ΔM values given by the daily growth model are conditional on the parameter vector $\boldsymbol{\theta}$. Note also that the ΔM terms are specific to subscript *k*. This is because different observations *k* were subject to different combinations of temperature and nutrient regimes, which alter the predicted growth in the daily growth model. However, since there were only 10 combinations of these factors (5 nutrient regimes × 2 temperature regimes), in practice it was only necessary to implement the daily growth model 10 times for each parameter set $\boldsymbol{\theta}$.

Nondestructive measurements. Utilizing nondestructive measurements provided two principle benefits to the analysis of plant growth in this case. First, it meant that each individual plant contributed more than one pair of values for above- and belowground biomass, increasing the amount of observational information available to estimate the model parameters. Second, the nondestructive measurements provided multiple measurements of the biomass of individual plants, which could be used to measure growth over short time intervals. This provided information on how, for a given species *j* in a given combination of temperature and nutrient

regime, variation in above- and belowground biomass affected above- and belowground growth over the subsequent growth period. Since biomass is a primary driver of growth and allocation in the daily growth model (see main text), this additional information can be expected to substantially improve parameter estimation.

Therefore, for nondestructive measurements corresponding to plant *i* and measurement day $\hat{d}(k)$, the model was initialized with the nondestructive measurements of above- and belowground biomass, *specific to plant i*, taken in the measurement period proceeding day $\hat{d}(k)$, $d_{prev}(k)$ (approximately 30 days prior to $\hat{d}(k)$, because measurements were taken monthly). After this initialization, the predicted biomass for observation *k* was generated by summing the changes in biomass from $d_{prev}(k)$ up to $\hat{d}(k)$. Thus, where observation *k* is for plant *i* of species *j* measured on day $\hat{d}(k)$:

$$M_{i,d_{prev}(k)}^{(abv,mod)} = M_{i,d_{prev}(k)}^{(abv,obs)}$$

$$M_{i,d_{prev}(k)}^{(blw,mod)} = M_{i,d_{prev}(k)}^{(blw,obs)}$$

$$M_{k}^{(abv,mod)} = M_{j,d_{prev}(k)}^{(abv,mod)} + \sum_{d=d_{prev}(k)}^{d=\hat{d}(k)} \Delta M_{k,d}^{(abv,mod)}$$

$$M_{k}^{(blw,mod)} = M_{j,d_{prev}(k)}^{(blw,mod)} + \sum_{d=d_{prev}(k)}^{d=\hat{d}(k)} \Delta M_{k,d}^{(blw,mod)} \quad . \qquad \text{(B2.4)}$$

As Eqs. 2.4 show, under this scheme a particular plant *i* received an initialization specific to *i* (see the right-hand side of the first two equations in B2.4). This went on to affect the predicted growth and allocation of plant *i* (because this is affected by initial biomass; see main text), which in turn provided a prediction for the above- and belowground biomass of plant *i*, which was then compared with that observed for plant *i* . Under this scheme, the effect of individual-to-individual variation in initial biomass, for two plants of the same species, growing over the same period, and experiencing the same environmental conditions, was

explicitly incorporated into the parameter estimation. This information cannot be incorporated into an analysis utilizing destructive harvests alone.

DATA MANIPULATION

Each plant *i* provided a number of nondestructive values for above- and belowground biomass, and was then harvested, at which time it provided a pair of destructive values for above- and belowground biomass. Plants harvested at the first measurement date (day 14) provided destructive measures only. The sets $S(abv, dest)$ and $S(blw, dest)$ consisted of all available destructive harvest data, except the day-14 data, which was used only to initialize the growth model for predictions to compare with destructive harvest data (see Eqs. B2.3 above). This provided 1278 observations *k* in each of $S(abv, dest)$ and $S(blw, dest)$, a total of 2556 observations.

The sets $S(abv, non-dest)$ and $S(blw, non-dest)$ were generated by finding all available *pairs* of nondestructive measurements of above- and belowground biomass *from the same plant i*, where the second set of measurements was taken in the measurement date immediately following the first (approximately 30 days previous to the second measurement). From this pair of observations, the second measurement corresponded to an observation *k* as described above, while the previous observation was used to initialize the daily growth model to provide a model prediction with which to compare observation *k* (see Eqs. B2.4 above).

Therefore, the observations *k* in sets $S(abv, non-dest)$ and $S(blw, non-dest)$ consisted of pairs of measurements taken either on day 35 with a previous measurement from day 14; or on day 70, with a previous measurement from day 35; or on day 98 with a previous measurement from day 70; or on day 133 with a previous measurement from day 98; or on day 161 with a previous measurement from day 133. An example of an observation *k* in set $S(abv, non-dest)$ is the aboveground biomass of plant 457, measured on day 161, where a

nondestructive measurement of the aboveground biomass of plant 457 was also taken on day 133. In total, there were 2325 paired measurements of above- and belowground biomass meeting these criteria, providing 2325 observations k in the each of the sets $S(abv, non-dest)$ and $S(blw, non-dest)$ (a total of 4650 observations).

MONTE CARLO MARKOV CHAIN SCHEME

In likelihood analysis, the vector of parameter values $\boldsymbol{\theta}$ that maximizes $\ell(\mathbf{X}|\boldsymbol{\theta})$ is taken to be the most likely vector of parameters, given the model structure and data *X*. The parameter values in this vector are referred to as the "maximum-likelihood estimates", or MLEs, and the value of the likelihood that is returned when the parameters are set to the MLEs is called the "maximum likelihood". Confidence limits on parameters can be defined very simply. Extract the set of all vectors $\boldsymbol{\theta}$ that return a value for $\ell(\mathbf{X}|\boldsymbol{\theta})$ within a certain tolerance of the maximum likelihood, where the tolerance is defined by the confidence level (e.g., for P = 0.05, the tolerance is 1.96). This set of vectors contains a range of values for a given parameter *p*, and this range is the confidence interval for *p*.

Therefore, once $\ell(\mathbf{X}|\boldsymbol{\theta})$ is defined, all that remains is to find the MLE vector, and the set of all other vectors that lies within a predetermined tolerance of the maximum likelihood. When fitting a model with few parameters, and when $\ell(\mathbf{X}|\boldsymbol{\theta})$ can be calculated very quickly, it is possible to calculate the $\ell(\mathbf{X}|\boldsymbol{\theta})$ associated with every possible vector within a pre-specified range (and with the parameters specified to a predetermined level of accuracy). However, for models with larger numbers of parameters, such as the one considered here, the number of possible vectors is so large that an approximation is needed.

In this case, we employed a Monte Carlo Markov chain (MCMC) sampling scheme, which we have found previously to be particularly efficient at finding MLEs and confidence intervals (e.g., Purves et al. 2007*a*,*b*, A. Barron, D. W. Purves, and L. O. Hedin, *unpublished*

manuscript). MCMC sampling is based around random changes in $\boldsymbol{\theta}$, which are then accepted or rejected depending on the corresponding changes in $\ell(\mathbf{X}|\boldsymbol{\theta})$ (see Chib and Greenberg 1995). Two key features of MCMC sampling make it attractive for complex, parameter-rich problems, such as that considered here. First, as might be expected, the algorithm accepts any change in $\boldsymbol{\theta}$ that increases $\ell(\mathbf{X}|\boldsymbol{\theta})$; but it also probabilistically accepts changes that decrease $\ell(\mathbf{X}|\boldsymbol{\theta})$, according to the so-called "metropolis criterion". This latter behavior is particularly important in nonlinear problems, because it allows the algorithm to escape from local maxima of $\ell(\mathbf{X}|\boldsymbol{\theta})$, and find the global maximum. Second, when configured properly, MCMC sampling has the remarkable property that, once a so-called quasi-equilibrium has been reached, each vector $\boldsymbol{\theta}$ returned by the algorithm is a random sample from the posterior probability of $\boldsymbol{\theta}$ given $\mathbf{X}$ (Chib and Greenberg 1995). Within the context of likelihood analysis, this means that (for example) 90% of the values of a given parameter *p* returned by the algorithm will be within the 90% confidence limit for *p* (given the model structure, and data $\mathbf{X}$). In practical terms, this means that, retaining a list of parameter values for *p* output by the MCMC scheme, ranking this list, and removing the tails gives a simple estimate of the confidence interval for *p*. The benefit of this technique is that the time required to obtain an estimate of the confidence intervals is dramatically reduced compared with the time required to find the set of all vectors within a fixed tolerance of the maximum likelihood. However, we have confirmed in the context of several different analyses (e.g., A. Barron, D. W. Purves, and L. O. Hedin, *unpublished manuscript*; and the current global fit in this study) that it gives the same results. This can also be proven analytically (not shown).

There are many particular choices to be made in the implementation of MCMC sampling, but providing that certain criteria are met, these choices only affect the efficiency and not the output of the analysis. The scheme used and outlined here is the result of experimentation with a variety of problems in likelihood and Bayesian analysis (see, for

example, Purves et al. 2007*a*; Purves et al. 2007*b*; A. Barron, L. O. Hedin and D. W. Purves, *unpublished manuscript*). For further details, or C code to implement the scheme, please contact D.W. Purves. To generate the proposed parameter vector $\boldsymbol{\theta}'$ from the current vector $\boldsymbol{\theta}$, the scheme (1) chooses, at random, the number, *n*, of parameters to be altered between 1 and *N*, where *N* is the total of parameters in $\boldsymbol{\theta}$; (2) selects, at random, which of the *N* parameters are to be altered; (3) increments the current value of *p* by Δp for each parameter *p* chosen, (i.e. $p \to p + \Delta p$) where Δp is drawn from a Gaussian distribution with mean zero and standard deviation σ_P (if *p* is logically constrained to be positive, then the transformation $p \to p.\exp[\Delta p]$ is used, which corresponds to a Gaussian jump in the logarithm of *p*). The change from $\boldsymbol{\theta}$ to $\boldsymbol{\theta}'$ is accepted according to the metropolis criterion: i.e., with certainty if $\ell(\mathbf{X}|\boldsymbol{\theta}') > \ell(\mathbf{X}|\boldsymbol{\theta})$, and with probability $\exp[\ell(\mathbf{X}|\boldsymbol{\theta}') - \ell(\mathbf{X}|\boldsymbol{\theta})]$ if $\ell(\mathbf{X}|\boldsymbol{\theta}') < \ell(\mathbf{X}|\boldsymbol{\theta})$. The algorithm begins with σ_p set to half the allowable range in *p* (where the allowable range is much larger than the realistic range). The values of σ_p are adjusted automatically, to achieve a target acceptance ratio of 0.25 (i.e., the aim is to accept one quarter of the proposed changes, which is often considered optimal). Importantly, the scheme for adjusting the σ_p values (which is too complicated to be given here) allows the σ_p values to become specific to *p* (in practice, this means that σ_p becomes smaller for parameters that are better constrained by the data). This improves the efficiency of the sampling of vectors from the posterior.

When fitting this model, we allowed a "burn in" of 250000 iterations for the MCMC algorithm to reach quasi-equilibrium, after which an additional 250000 were used to provide samples of the parameters from the posterior distribution. These numbers of iterations were many times greater than required for convergence in this case. The MLE vector was taken to be the single vector $\boldsymbol{\theta}$ that returned the highest value of the likelihood $\ell(\mathbf{X}|\boldsymbol{\theta})$ from the total sample of 500000 vectors. 67% and 95% confidence intervals for parameter *p* were generated

by cutting the tails from a rarified sample from the 250000 post-burn-in vectors, as described above.

LITERATURE CITED

Chib, S., and E. Greenberg. 1995. Understanding the Metropolis-Hasting algorithm. The American Statistician 49:327–335.

Hilborn, R., and M. Mangel. 1997. The ecological detective: confronting models with data. Princeton University Press, Princeton, New Jersey, USA.

Purves, D. W., J. W. Lichstein, and S. W. Pacala. 2007*a*. Crown plasticity and competition for canopy space: a new spatially implicit model parameterized for 250 North American tree species. PLoS-ONE 2(9): e870. [doi:10.1371/journal.pone.0000870]

Purves, D. W., M. A. Zavala, K. Ogle, F. Prieto, and J. Ray-Benayas. 2007*b*. Environmental heterogeneity, bird-mediated directed dispersal, and oak woodland dynamics in Mediterranean Spain. Ecological Monographs **77**:77–97.