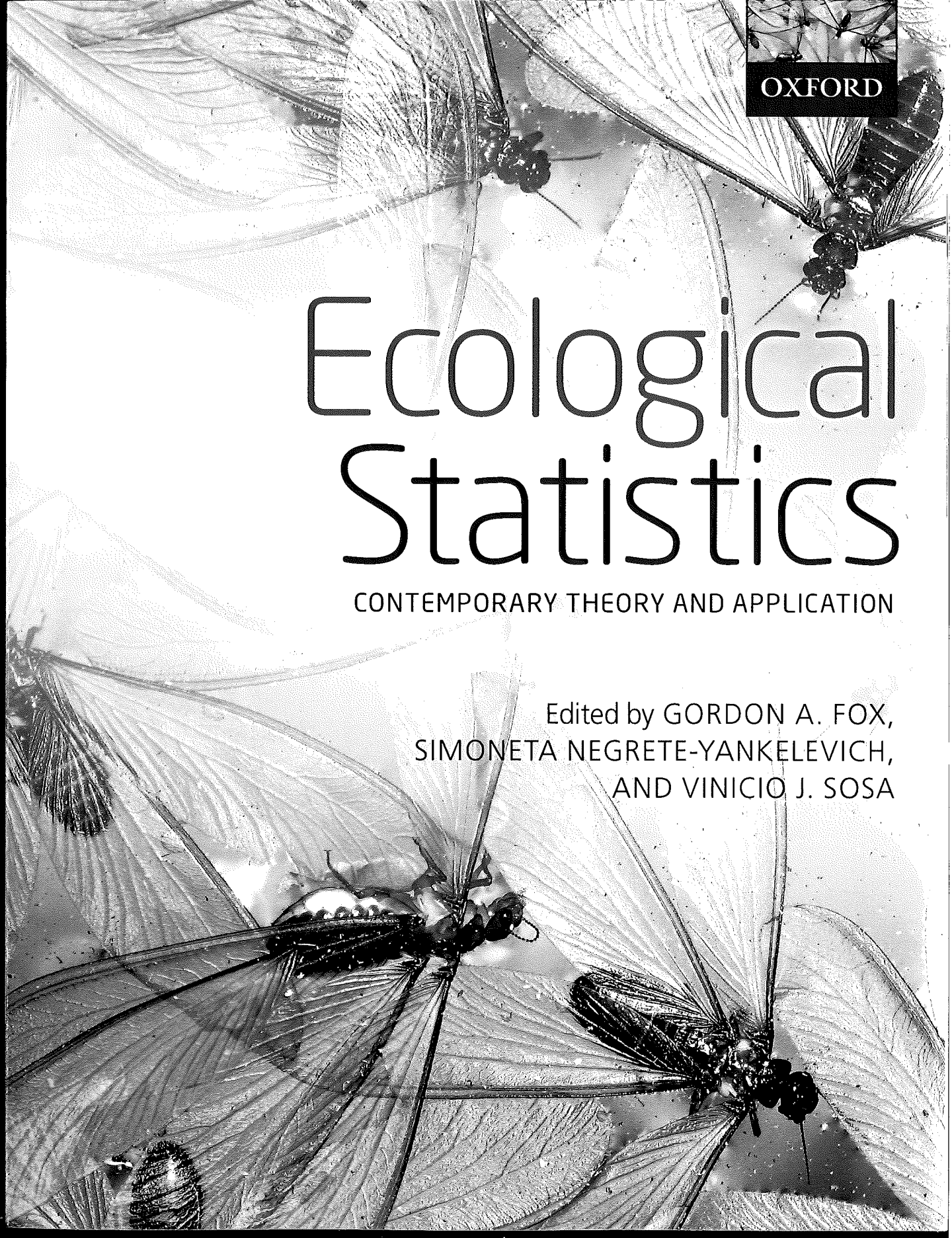

OXFORD

Ecological Statistics

CONTEMPORARY THEORY AND APPLICATION

Edited by GORDON A. FOX,
SIMONETA NEGRETE-YANKELEVICH,
AND VINICIO J. SOSA

CHAPTER 13

Linear and generalized linear mixed models

Benjamin M. Bolker

13.1 Introduction to generalized linear mixed models

Generalized linear mixed models (GLMMs) are a powerful class of statistical models that combine the characteristics of generalized linear models (GLMs: chapter 6) and *mixed models* (models with both fixed and random predictor variables). They handle a wide range of types of response variables, and a wide range of scenarios where observations have been sampled in some kind of groups rather than completely independently. While they can't do everything—an expert might sometimes choose custom-built models for greater flexibility (Bolker et al. 2013)—GLMMs are fast, powerful, can be extended to handle additional complexities such as zero-inflated responses, and can often be fitted with off-the-shelf software. The only real downsides of GLMMs are due to their generality: (1) some standard recipes for model testing and inference do not apply, and (2) it's easy to build plausible models that are too complex for your data to support. GLMMs are still part of the statistical frontier, and even experts don't know all of the answers about how to use them, but this chapter will try to provide practical solutions to allow you to use GLMMs with your data.

GLMs allow modeling of many kinds of response variables, particularly those with binomial and Poisson distributions; you should definitely feel comfortable with GLMs before attempting the methods described in this chapter. In contrast, you may be unfamiliar with the mixed models, and with the central distinction between *fixed effects* (the typical way to compare differences between treatments or the effects of continuous predictor variables) and *random effects* (roughly speaking, experimental or observational blocks within which you have several observations). Models with normally distributed responses that incorporate some kind of random effects are called *linear mixed models* (LMMs); they are a special, slightly easier case of GLMMs. This chapter will review the basic idea of experimental blocks (for a reminder see Gotelli and Ellison (2004) or Quinn and Keough (2002)). If you are already well-versed in classic ANOVA approaches to blocked experimental designs, you may actually have to unlearn some things, as modern approaches to random effects are quite different from the classical approaches taught in most statistics courses.

As well as using different conceptual definitions of random effects (section 13.3.1), modern mixed models are more flexible than classic ANOVAs, allowing, for example, non-Normal responses, unbalanced experimental designs, and more complex grouping

Ecological Statistics: Contemporary Theory and Application. First Edition. Edited by Gordon A. Fox, Simoneta Negrete-Yankelevich, and Vinicio J. Sosa.
Published in 2015 by Oxford University Press.

structures. Equally important is a new philosophy: modern approaches use a model-building rather than a hypothesis-testing approach (chapter 3). You can still test hypotheses, but instead of a list of F statistics and p-values the primary outputs of the analysis are quantitative parameter estimates describing (1) how the response variable changes as a function of the fixed predictor variables, and (2) the variability among the levels of the random effects.

Random effects such as variation among experimental blocks are often neglected in model-based analyses because they are relatively difficult to incorporate in custom-built statistical models. While one can use software such as WinBUGS, AD Model Builder, or SAS PROC NLMIXED to incorporate such components in a general model (Bolker et al. 2013), generalized linear mixed models are general enough to encompass the most common statistical problems in ecology, yet can be fitted with off-the-shelf software.

Section 13.2 (Running examples) introduces several case studies from the literature, and from my own work, for which the data are freely accessible. Section 13.3 (Concepts) gets philosophical, exploring different definitions of random effects; related concepts like pooling, shrinkage, and nested vs. crossed experimental designs; the statistical issues of overdispersion and variable correlation within groups; and the extended definitions of likelihood required for mixed models (see chapter 3 for the basic definition). Section 13.4 (Setting up a GLMM) is practical but short; once you understand the ins and outs of random effects, and the concepts of GLMs from chapter 6, writing the code to define a GLMM is actually quite straightforward. Sections 13.5 (Estimation) and 13.6 (Inference) go into nitty-gritty detail about the choices you have when fitting a GLMM and translating the results back from statistical to scientific answers.

13.2 Running examples

- *Tundra carbon dynamics*: Belshe et al. (2013) did a meta-analysis of previous studies of carbon uptake and release in tundra ecosystems. They asked how the CO_2 flux, or net ecosystem exchange at measured experimental sites, was changing over time. The residual variation of the primary response variable (`GS.NEE`, net carbon flux during the growing season) was assumed to be Normal, so the model is a linear mixed model. Sites were treated as random effects, meaning that site was a *grouping variable* (a categorical predictor across which effects are assumed to vary randomly), with the baseline CO_2 flux (i.e., the intercept term of the model) varying across sites. Time (`Year`) was the primary fixed effect, although the paper also considered the effects of mean annual temperature and precipitation, as well as the additional response variables of winter and total annual carbon flux.
- *Coral symbiont defense*: McKeon et al. (2012) ran a field experiment with coral (*Pocillopora* spp.) inhabited by invertebrate symbionts (crabs [*Trapezia* spp.] and shrimp [*Alpheus* spp.]) and exposed to predation by sea stars (*Culcita* spp.). They asked whether combinations of symbionts from different species were more, less, or equally effective in defending corals from predators, compared to expectations based on the symbionts' independent protective effects. The design is a randomized complete block design with a small amount of replication: 2 replications per treatment per block; 4 treatments (no symbionts, crabs only, shrimp only, both symbionts), with each of these units of 8 repeated in 10 blocks. The response (`predation`) is binomial with a single trial per unit (also called Bernoulli or binary, see book appendix); treatment (`ttt`), a categorical

variable, is the only fixed-effect input variable; block is the only grouping variable, with intercepts (i.e., baseline predation probability) varying among blocks.

- *Gopher tortoise shells*: Ozgul et al. (2009) analyzed the numbers of gopher tortoise shells found at different sites to estimate whether shells were more common (implying a higher mortality rate) at sites with higher prevalence of a mycoplasmal pathogen (`prev`). The response is the count of fresh shells (`shells`), for which we will consider Poisson and negative binomial distributions (book appendix); seroprevalence of mycoplasma (`prev`: i.e., the fraction of tortoises carrying antibodies against the disease) is a continuous, fixed predictor variable. We initially considered year and site as crossed grouping variables (section 13.3.1) with variation in baseline shell counts (intercepts) among them; we also included the logarithm of the site area (`Area`) as an *offset* term to account for variation in site area, effectively modeling shell density rather than shell numbers.
- *Red grouse ticks*: Elston et al. (2001) used data on numbers of ticks sampled from the heads of red grouse chicks in Scotland to explore patterns of aggregation. Ticks have potentially large fitness and demographic consequences on red grouse individuals and populations, but Elston et al.'s goal was just to decompose patterns of variation into different scales (within-brood, within-site, by altitude and year). The response is the tick count (`TICKS`, again Poisson or negative binomial); altitude (`HEIGHT`, treated as continuous) and year (`YEAR`, treated as categorical) are fixed predictor variables. Individual within brood (`INDEX`) and brood within location are nested random-effect grouping variables, with the baseline expected number of ticks (intercept) varying among groups.

All of these case studies include some kind of grouping (sites in the tundra carbon example; experimental blocks for the sea star example; areas and years for the gopher tortoise example; and individuals within broods within sites for the tick example), requiring mixed models. The first has Normal responses, requiring a LMM, while the latter three have non-Normal response variables, requiring GLMMs.

13.3 Concepts

13.3.1 *Model definition*

The complete specification of a GLMM includes the distribution of the response variable; the link function; the definition of categorical and continuous fixed-effect predictors; and the definition of the random effects, which specify how some model parameters vary randomly across groups. Here we focus on random effects, the only one of these components that is not already familiar from chapter 6.

Random effects

The traditional view of random effects is as a way to do correct statistical tests when some observations are correlated. When samples are collected in groups (within sites in the tundra example above, or within experimental blocks of any kind), we violate the assumption of independent observations that is part of most statistical models. There will be some variation within groups (σ^2_{within}) and some among groups (σ^2_{among}); the total variance is $\sigma^2_{\text{total}} = \sigma^2_{\text{within}} + \sigma^2_{\text{among}}$; and therefore the correlation between any two observations in the same group is $\rho = \sqrt{\sigma^2_{\text{among}}/\sigma^2_{\text{total}}}$ (observations that come from *different* groups are uncorrelated). Sometimes one can solve this problem easily by taking group averages.

For example, if we are testing for differences between deciduous and evergreen trees, where every member of a species has the same leaf habit, we could simply calculate species' average responses, throwing away the variation within species, and do a *t*-test between the deciduous and evergreen species means. If the data are balanced (i.e., if we sample the same number of trees for each species), this procedure is exactly equivalent to testing the fixed effect in a classical mixed model ANOVA with a fixed effect of leaf habit and a random effect of species. This approach correctly incorporates the facts that (1) repeated sampling within species reduces the uncertainty associated with within-group variance, but (2) we have fewer *independent* data points than observations—in this case, as many as we have groups (species) in our study.

These basic ideas underlie all classical mixed-model ANOVA analyses, although the formulas get more complex when treatments vary within grouping variables, or when different fixed effects vary at the levels of different grouping variables (e.g., randomized-block and split-plot designs). For simple nested designs, simpler approaches like the averaging procedure described above are usually best (Murtaugh 2007). However, mixed-model ANOVA is still extremely useful for a wide range of more complicated designs, and as discussed below, traditional mixed-model ANOVA itself falls short for cases such as unbalanced designs or non-Normal data.

We can also think of random effects as a way to combine information from different levels within a grouping variable. Consider the tundra ecosystem example, where we want to estimate linear trends (slopes) across time for many sites. If we had only a few years sampled from a few sites, we might have to *pool* the data, ignoring the differences in trend among sites. Pooling assumes that σ^2_{among} (the variance in slopes among sites) is effectively zero, so that the individual observations are uncorrelated ($\rho = 0$).

On the other hand, if we had many years sampled from each site, and especially if we had a small number of sites, we might want to estimate the slope for each site individually, or in other words to estimate a fixed effect of time for each site. Treating the grouping factor (site) as a fixed effect assumes that information about one site gives us no information about the slope at any other site; this is equivalent, for the purposes of parameter estimation, to treating σ^2_{among} as infinite. Treating site as a random effect compromises between the extremes of pooling and estimating separate (fixed) estimates; we acknowledge, and try to quantify, the variability in slope among sites. Because the trends are assumed to come from a population (of slopes) with a well-defined mean, the predicted slopes in CO_2 flux for each site are a weighted average between the trend for that site and the overall mean trend across all sites; the smaller and noisier the sample for a particular site, the more its slope is compressed toward the population mean (figure 13.1).

For technical reasons, these values (the deviation of each site's value from the population average) are called *conditional modes*, rather than *estimates*. The conditional modes are also sometimes called *random effects*, but this could also refer to the grouping variables (the sites themselves, in the tundra example). Confusingly, both the conditional modes and the estimates of the among-site variances can be considered parameters of the random effects part of the model. For example, if we had independently estimated the trend at one site (i.e., as a fixed effect) as –5 grams C/m^2/year, with an estimated variance of 1, while the mean rate of all the sites was –8 g C/m^2/year with an among-site variance of 3, then our predicted value for that site would be $(\mu_{\text{site}}/\sigma^2_{\text{within}} + \mu_{\text{overall}}/\sigma^2_{\text{among}})/(1/\sigma^2_{\text{within}} + 1/\sigma^2_{\text{among}}) = (-5/1 + -8/3)/(1/1 + 1/3) = -5.75$ g C/m^2/year. Because $\sigma^2_{\text{within}} < \sigma^2_{\text{among}}$—the trend estimate for the site is relatively precise compared to the variance among sites—the random-effects prediction is closer to the site-specific value than to the overall mean. (Stop and plug in a few different values of among-site variance to convince yourself that

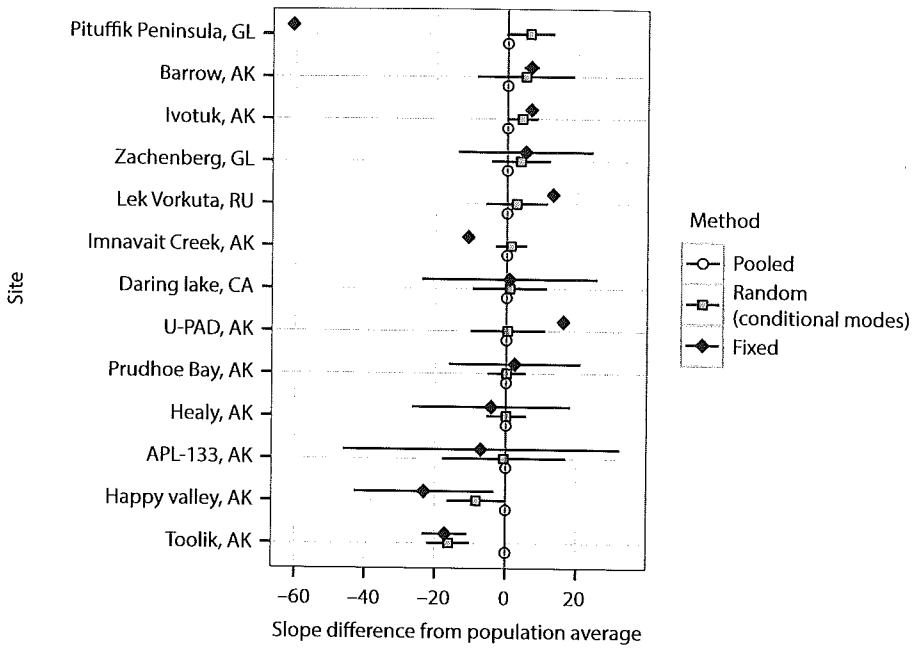

Fig. 13.1 Estimated differences in slope (annual change in growing season NEE) among sites, with 95% confidence intervals. The conditional modes are (mostly) intermediate between the fixed estimates and the pooled estimate of zero (the two exceptions, Pituffik Peninsula and Imnaivit Creek, have compensating differences in their intercept estimates); sites with only one year's data, for which a fixed-effect slope cannot be estimated, are not shown. The confidence intervals are generally much narrower for the conditional modes than for the fixed-effect estimates (the four fixed-effect estimates with error bars not shown have 95% CIs that extend beyond the limits of the plot).

this formula agrees with verbal description above of how variance-weighted averaging works when σ^2_{among} is either very small or very large relative to σ^2_{within}.)

Random effects are especially useful when we have (1) lots of levels (e.g., many species or blocks), (2) relatively little data on each level (although we need multiple samples from most of the levels), and (3) uneven sampling across levels (box 13.1).

Frequentists and Bayesians define random effects somewhat differently, which affects the way they use them. Frequentists define random effects as categorical variables whose levels are chosen *at random from a larger population*, e.g., species chosen at random from a list of endemic species. Bayesians define random effects as sets of variables whose parameters are drawn from a distribution. The frequentist definition is philosophically coherent, and you will encounter researchers (including reviewers and supervisors) who insist on it, but it can be practically problematic. For example, it implies that you can't use species as random effect when you have observed *all* of the species at your field site—since the list of species is not a sample from a larger population—or use year as a random effect, since researchers rarely run an experiment in randomly sampled years—they usually use either a series of consecutive years, or the haphazard set of years when they could get into the field. This problem applies to both the gopher tortoise and tick examples, each of which use data from consecutive years.

Box 13.1 WHEN TO TREAT A PREDICTOR VARIABLE AS A RANDOM EFFECT

You may want to treat a predictor variable as a random effect if you:

- don't want to test hypotheses about differences between responses at particular levels of the grouping variable;
- do want to quantify the variability among levels of the grouping variable;
- do want to make predictions about unobserved levels of the grouping variable;
- do want to combine information across levels of the grouping variable;
- have variation in information per level (number of samples or noisiness);
- have levels that are randomly sampled from/representative of a larger population;
- have a categorical predictor that is a nuisance variable (i.e., it is not of direct interest, but should be controlled for).

Cf. Crawley (2002); Gelman (2005)

If you have sampled fewer than five levels of the grouping variable, you should strongly consider treating it as a fixed effect even if one or more of the criteria above apply.

Random effects can also be described as predictor variables where you are interested in making inferences about the distribution of values (i.e., the variance among the values of the response at different levels) rather than in testing the differences of values between particular levels. Choosing a random effect trades the ability to test hypotheses about differences among particular levels (low vs. high nitrogen, 2001 vs. 2002 vs. 2003) for the ability to (1) quantify the variance among levels (variability among sites, among species, etc.) and (2) generalize to levels that were not measured in your experiment. If you treat species as a fixed effect, you can't say anything about an unmeasured species; if you use it as a random effect, then you can guess that an unmeasured species will have a value equal to the population mean estimated from the species you did measure. Of course, as with all statistical generalization, your levels (e.g., years) must be chosen in some way that, if not random, is at least *representative* of the population you want to generalize to.

People sometimes say that random effects are "factors that you aren't interested in." This is not always true. While it is often the case in ecological experiments (where variation among sites is usually just a nuisance), it is sometimes of great interest, for example in evolutionary studies where the variation among genotypes is the raw material for natural selection, or in demographic studies where among-year variation lowers long-term growth rates. In some cases fixed effects are also used to control for uninteresting variation, e.g., using mass as a covariate to control for effects of body size.

You will also hear that "you can't say anything about the (predicted) value of a conditional mode." This is not true either—you can't formally test a null hypothesis that the value is equal to zero, or that the values of two different levels are equal, but it is still perfectly sensible to look at the predicted value, and even to compute a standard error of the predicted value (e.g., see the error bars around the conditional modes in figure 13.1). Particularly in management contexts, researchers may care very much about *which* sites are particularly good or bad relative to the population average, and how much better or worse they are than the average. Even though it's difficult to compute formal inferential summaries such as p-values, you can still make common-sense statements about the conditional modes and their uncertainties.

The Bayesian framework has a simpler definition of random effects. Under a Bayesian approach, a fixed effect is one where we estimate each parameter (e.g., the mean for each species within a genus) independently (with independently specified priors), while for a random effect the parameters for each level are modeled as being drawn from a distribution (usually Normal); in standard statistical notation, species_mean ~ Normal(genus_mean, $\sigma^2_{species}$).

I said above that random effects are most useful when the grouping variable has many measured levels. Conversely, random effects are generally ineffective when the grouping variable has too few levels. You usually can't use random effects when the grouping variable has fewer than five levels, and random effects variance estimates are unstable with fewer than eight levels, because you are trying to estimate a variance from a very small sample. In the classic ANOVA approach, where all of the variance estimates are derived from simple sums-of-squares calculations, random-effects calculations work as long as you have at least two samples (although their power will be very low, and sometimes you can get negative variance estimates). In the modern mixed-modeling approach, you tend to get warnings and errors from the software instead, or estimates of zero variance, but in any case the results will be unreliable (section 13.5 offers a few tricks for handling this case). Both the gopher tortoise and grouse tick examples have year as a categorical variable that would ideally be treated as random, but we treat it as fixed because there are only three years sampled: treating years as a random effect would most likely estimate the among-year variance as zero.

Simple vs. complex random effects

The most common type of random effect quantifies the variability in the baseline values of the response variable among levels of a categorical grouping variable (e.g., baseline numbers of ticks in different locations). Although technically location is the *grouping variable* in this case, and the thing that varies among levels is the intercept term of a statistical model, we would often call this simply a random effect of location. This is a random intercept model, which is also a scalar random effect (i.e., there is only one value per level of the grouping variable). In R it would be specified within a modeling formula as `~group` or `~(1):group` (`MCMCglmm` library), `~1|group` (`nlme` or `glmmADMB` libraries), or `(1|group)` (`lme4` or `glmmADMB` libraries) (the `1` specifies an intercept effect; it is implicit in the first example).

More generally, we might have observed the effects of a treatment or covariate within each level, and want to know how these effects (described by either a categorical or a continuous predictor) vary across levels; this is the case for slopes (i.e., the effect of time) in the tundra example. Since the intercept as well as all of the parameters describing the treatment would vary across levels, this would be called a non-scalar or a vector random effect. This could be specified as `~1+x|group` in `nlme` or `glmmADMB`, `(1+x|group)` in `lme4` or `glmmADMB`), or `~us(1+x):group` in `MCMCglmm`. In many cases the `1` is optional—`(x|group)` would also work—but I include it here for concreteness. The `us` in the third specification refers to an unstructured variance–covariance matrix: `MCMCglmm` offers several other options (see the *Course Notes* vignette that comes with the library).

For example, the coral symbiont data follow a randomized block design, with replicates of all treatments within each block. So we could in principle use the random effects model `(1+ttt|block)` (equivalent to `(ttt|block)` because the intercept is implicitly included) to ask how the effects of symbionts varied among different blocks, with four random parameters per block (intercept and three treatment parameters), where the intercept parameter describes the variation among control treatments across blocks and the

treatment parameters describe the variation in the effects of symbionts (crab vs. control, shrimp vs. control, and crab + shrimp vs. control) among blocks. However, this is another case where the ideal and the practical differ; in practice this approach is not feasible because we have too little information—there are only two binary samples per treatment per block—so we would likely proceed with an intercept-only (scalar) random effect of blocks.

Non-scalar effects represent *interactions* between the random effect of block and the fixed effect (symbionts), and are themselves random—we assume, for example, that the difference in predation rate between corals with and without symbionts is drawn from a distribution of (differences in) predation rates. The interaction between a random effect and a continuous predictor is also random; the tundra carbon example includes a site × year interaction which describes the variation in temporal trends among sites. This type of interaction is the only case in which it makes sense to consider a random effect of a continuous variable; a continuous variable (year in this example) cannot itself be a *grouping* variable, but can vary across grouping variables (sites). One should in general consider the random × fixed effect interactions whenever it is feasible, i.e., for *all* treatments that are applied within levels of a random effect; doing otherwise assumes a priori that there is no variation among groups in the treatment effect, which is rarely warranted biologically (Schielzeth and Forstmeier 2009; Barr et al. 2013). It is often impossible or logistically infeasible to apply treatments within groups: in the gopher tortoise example the prevalence of disease is fundamentally a site-level variable, and can't vary within sites. Or, as in the coral symbiont example, we may have so little statistical power to quantify the among-group variation that our models don't work, or that we estimate the variation as exactly zero. In these cases we have to accept that there probably is a real interaction that we are ignoring, and temper our conclusions accordingly.

Nesting and crossing

What about the interaction between two random effects? Here we have to specify whether the two effects are *nested* or *crossed.* If at least one of the levels of each effect is represented in multiple levels of the other effect, then the random effects are crossed; otherwise, one is nested in the other. In the gopher tortoise example, each site is measured in multiple years, and multiple sites are measured in each year, so site and year are crossed (although as pointed out above we don't actually have data for enough years to treat them as random); this would be specified as `(1|site)+(1|year)`. On the other hand, in the tick example each chick occurs in only one brood, and each brood occurs in only one site: the model specification is `(1|SITE/BROOD/INDEX)`, read as "chick (`INDEX`) nested within brood nested within site," or equivalently `(1|SITE)+(1|SITE:BROOD)+(1|SITE:BROOD:INDEX)`. If the broods and chicks are uniquely labeled, so that the software can detect the nesting, `(1|SITE)+(1|BROOD)+(1|INDEX)` will also work (do *not* use `(1|SITE)+(1|SITE/BROOD)+(1|SITE/BROOD/INDEX)`; it will lead to redundant terms in the model). Another way of thinking about the problem is that, in the gopher tortoise example, there is variation among sites that applies across years, variation among years that applies across all sites, and variation among site-by-year combinations. In the tick example, there is variation among broods and variation among chicks within broods, but there is no sensible way to define variation among chicks *across* broods. In this sense a nested model is a special case of crossed random effects that sets one of the variance terms to zero.

Crossed random effects are more challenging computationally than nested effects (they are largely outside the scope of classical ANOVAs), and so this distinction is often ignored in older textbooks. Most of the software that can handle both crossed and nested random

effects can automatically detect when a nested model is appropriate, provided that the levels of the nested factor are uniquely labeled. That is, the software can only tell individuals are nested if they are labeled as A1, A2, . . . , A10, B1, B2, . . . B10, . . . If individuals are instead identified only as 1, 2, . . . 10 in each of species A, B, and C, the software can't tell that individual #1 of species A is not related to individual #1 of species B. In this case you can specify nesting explicitly, but it is safer to label the nested individuals uniquely.

You should usually treat interactions between two or more fixed effects as crossed, because the levels of fixed effects are generalizable across levels of other fixed effects ("high nitrogen" means the same thing whether we are in a low- or high-phosphorus treatment). Random effects can be nested in fixed effects, but fixed effects would only be nested in random effects if we really wanted (for example) to estimate different effects of nitrogen in each plot.

Overdispersion and observation-level random effects

Linear mixed models assume the observations to be normally distributed conditional on the fixed-effect parameters and the conditional modes. Thus, they need to estimate the residual variance at the level of observations. If there is only one observation for each level of a grouping variable, the variance of the corresponding random effect will be confounded with the residual variance—we say that the variance of the observation-level random effect is *unidentifiable*. For example, if we decided to treat year as categorical variable in the tundra ecosystem analysis, and included a random effect of the site × year interaction, we would have exactly one observation for each site-by-year combination, and this random effect variance would be confounded with the residual variance. Many libraries (e.g., `nlme`) will fail to detect this problem, and will give arbitrary answers for the residual variance and the confounded random-effect variance. The same situation applies for any GLMM where the scale parameter determining the variance is estimated rather than fixed, such as Gamma GLMMs or quasi-likelihood models.

Most GLMMs, in contrast, assume distributions such as the binomial or Poisson where the scale parameter determining the residual variance is fixed to 1—that is, if we know the mean then we assume we also know the variance (equal to the mean for Poisson distributions, or to $Np(1-p)$ for binomial distributions, see book appendix). However, as discussed in chapters 3, 6, and 12, we frequently observe *overdispersion*—residual variances higher than would be predicted from the model, due to missing predictors or among-individual heterogeneity. Overdispersion does not occur in LMMs or in GLMMs with an estimated scale parameter, because the scale or residual variance parameter adjusts the model to match the residual variance. Overdispersion occurs, but is not identifiable, with binary/Bernoulli responses, unless the data are grouped so that there are multiple observations with the same sets of predictor variables (e.g., in the coral predation data there are two replicates in each site/treatment combination). If so, the data can be collapsed to a binomial response, in this case by computing the number of predation events (out of a maximum of 2) for each site/treatment combination, and then overdispersion will be identifiable.

You can allow for overdispersion in GLMMs in some of the same ways as in regular GLMs—use quasi-likelihood estimation to inflate the size of the confidence intervals appropriately, or use an overdispersed distribution such as a negative binomial. These options may not be available in your GLMM software: at present, none of the libraries discussed here offers quasi-likelihood estimation, and only `glmmADMB` has a well-tested negative binomial option.

A GLMM-specific solution to overdispersion is to add *observation-level* random effects, i.e., to add a new grouping variable with a separate level for every observation in the data set. This seems like magic—how can we estimate a separate parameter for every observation in the data set?—but it is just a way to add more variance to the data distribution. For Poisson distributions, the resulting *lognormal-Poisson* distribution is similar to a negative binomial distribution (sometimes called a *Gamma-Poisson* distribution because it represents a Poisson-distributed variable with underlying Gamma-distributed heterogeneity). Most GLMM packages allow observation-level random effects: for technical reasons, `MCMCglmm` *always* adds an observation-level random effect to the model, so you can *only* fit overdispersed models. Another advantage of using observation-level random effects is that this variability is directly comparable to the among-group variation in the model; Elston et al. (2001), the source of the grouse tick data, exploit this principle (see also Agresti 2002, section 13.5).

Correlation within groups (R-side effects)

As described above, grouping structure induces a correlation $\rho = \sqrt{\sigma^2_{\text{among}}/\sigma^2_{\text{total}}}$ between every pair of observations within a group. Observations can also be differentially correlated within groups; that is, an observation can be strongly correlated with some of the observations in its group, but more weakly correlated with other observations in its group. These effects are sometimes called *R-side effects* because they enter the model in terms of correlations of residuals (in contrast with correlations due to group membership, which are called *G-side effects*). The key feature of R-side effects is that the correlation between pairs of observations within a group typically decreases with increasing distance between observations. As well as physical distance in space or time, pairs of observations can be separated by their amount of genetic relatedness (distance along the branches of a pedigree or phylogeny). To include R-side effects in a model, one typically needs to specify both the distance between any two observations (or some sort of coordinates—observation time, spatial location, or position in a phylogeny—from which distance can be computed), as well as a model for the rate at which correlation decreases with distance. While incorporating R-side effects in *linear* mixed models is relatively straightforward—Belshe et al. (2013) included temporal autocorrelation in their model, and chapter 10 gives other examples—putting them into GLMMs is, alas, rather challenging at present.

Fixed effects and families

For a complete model, you need to specify the fixed effects part of your model, and the family (distribution and link function) as well as the random effects. These are both specified in the usual way as for standard (non-mixed, fixed-effect-only) GLMs (chapter 6).

Depending on the package you are using, the fixed effects may be specified separately or in the same formula as the random effects; typically the fixed-effect formula is also where you specify the response variable (the model has only one response variable, which is shared by both the fixed and the random effects). In the tundra ecosystem example, time (`year`) is the only fixed effect. In the coral symbiont example, the fixed effect is the categorical treatment variable `ttt` (control/shrimp/crabs/both). In the gopher tortoise example we have the effects of both disease prevalence and, because we didn't have enough levels to treat it as random, year (treated as a categorical variable); we also have an offset term that specifies that the number of shells is proportional to the site area (i.e., we add a log(area) term to the predicted log number of shells). Finally, the grouse tick example uses fixed effects of `YEAR` and `HEIGHT`.

13.3.2 *Conditional, marginal, and restricted likelihood*

Once you have defined your GLMM, specifying (1) the conditional distribution of the response variable (`family`) and link function (chapter 6); (2) the categorical and continuous predictors and their interactions (chapter 6); and (3) the random effects and their pattern of crossing and nesting (table 13.1), you are ready to try to fit the model. Chapter 3 describes the process of maximum likelihood estimation, which we extend here to allow for random effects.

Conditional likelihood

If we somehow knew the values of the conditional modes of the random effects for each level (e.g., the predation rates for each block), we could use standard numerical procedures to find the maximum likelihood estimates for the fixed-effect parameters, and all of the associated things we might like to know: confidence intervals, AIC values, and p-values for hypothesis tests against null hypotheses that parameters or combinations of parameters were equal to zero. The likelihood we obtain this way is called a conditional likelihood, because it depends (is conditioned on) a particular set of values of the conditional modes. If x is an observation, β is a vector of one or more fixed effects parameters, and b is a vector of the conditional modes of a random effect, then the conditional likelihood for x would be expressed as $L(x|\beta, b)$. If b were a regular fixed effect parameter, then we could go ahead and find the values of β and b that jointly gave the maximum likelihood, but that would ignore the fact that the conditional modes are random variables that are drawn from a distribution. In order to account for this extra variability, we need to define the marginal likelihood.

Marginal likelihood

The marginal likelihood is the modified form of the likelihood that allows for the randomness of the conditional modes. It compromises between the goodness of fit of the conditional modes to their overall distribution and the goodness of fit of the data within grouping variable levels. For example, a large number of attacks on a coral defended by both crabs and shrimp, which would be typically expected to be well protected, could be explained either by saying that the coral was an unlucky individual within its (perfectly typical) block or by saying that the coral was not unlucky but that the block was unusual, i.e., subject to higher-than-average attack rates. Because the block effect is treated as a random variable, in order to get the likelihood we have to average the likelihood over *all possible values* of the block effect, weighted by their probabilities of being drawn from the Normal distribution of blocks. The result is called the marginal likelihood, and we can generally treat it the same way as an ordinary likelihood. In mathematical terms, this average is expressed as an integral. If we take the definitions of x (observation), b (conditional mode), and β (fixed effect parameter) given above, and abbreviate the among-group variance introduced above (σ^2_{among}) as σ^2, then the likelihood of a given value of b is $L(b\,|\,\sigma^2)$ (the b values are defined as having a mean of zero) and the marginal likelihood of x is the integral of the conditional likelihood weighted by the likelihood of b:

$$L(x\,|\,\beta, \sigma^2) = \int L(x\,|\,b, \beta) \cdot L(b\,|\,\sigma^2)\, db.$$

Figure 13.2 shows the conditional likelihood $L(x\,|\,b, \beta)$ as a dashed line; the likelihood of the conditional mode $L(b\,|\,\sigma^2)$ as a dotted line; and the marginal likelihood as the gray area under the product curve. The marginal likelihood is a function of β and σ^2, which are the parameters we want to estimate. In a more complex model, σ^2 would be replaced

Table 13.1 Model specifications in R syntax for the examples

	nlme/glmmADMB	lme4/glmmADMB	MCMCglmm

tundra CO_2	`fixed = NEE ~ year,`
	`random = ~ year | Site`
	(`family` not specified for LMMs)
		`formula = NEE ~ year +`
		`(year | Site)`
		(`family` not specified for LMMs)
			`fixed = NEE ~ year,`
			`random = ~ us (year) : Site,`
			`family = "gaussian"`

coral symbiont	`fixed = pred ~ ttt,`
	`random = ~ 1 | block,`
	`family = "binomial"`
		`formula = pred ~ ttt + (1 | block),`
		`family = "binomial"`
			`fixed = pred ~ ttt,`
			`random = ~ block,`
			`family = "categorical"`

gopher tortoise	`fixed = shells ~ factor (year) +`
	`prev + offset (log (Area)),`
	`random = ~ 1 | Site,`
	`family = "poisson"`
		`formula = shells ~ factor (year) +`
		`prev + offset (log (Area)) +`
		`(1 | Site),`
		`family = "poisson"`
			`fixed = shells ~`
			`factor (year) + prev +`
			`offset (log (Area)),`
			`random = ~ Site,`
			`family = "poisson"`

grouse ticks	`fixed = ticks ~ 1 +`
	`factor (year) + height,`
	`random = ~ (1 | location / brood /`
	`index),`
	`family = "poisson"`
		`formula= ticks ~ 1 +`
		`factor (year) + height +`
		`(1 | location / brood / index),`
		`family = "poisson"`
			`fixed = ticks ~ 1 +`
			`factor (year) +`
			`height,`
			`random = ~ location + brood +`
			`index,`
			`family = "poisson"`

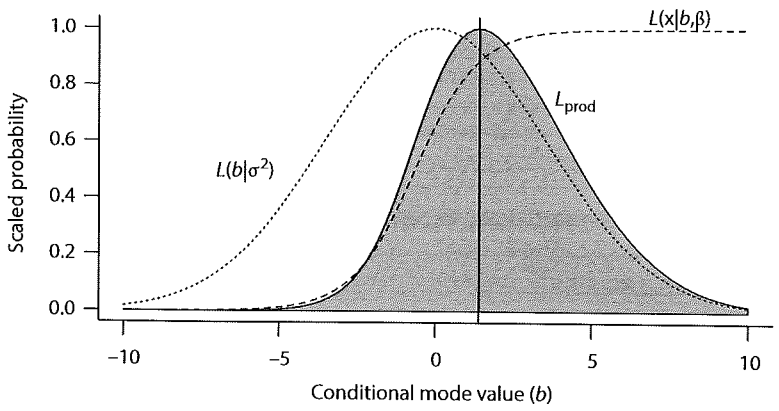

Fig. 13.2 Conditional and marginal likelihoods. For block 5, "shrimp" treatment, replicate 2. The Normal curve (dotted line) shows the likelihood of the conditional mode b; the logistic curve (dashed line) shows the conditional likelihood of the observation x given b; the solid line shows their product, and the gray area under the curve represents the marginal likelihood. (All likelihoods are scaled to a maximum of 1.0 for ease of presentation.) If the focal observation were the only one in the block, the conditional mode would be estimated at the peak of L_{prod}, $\hat{b}_5 = 1.4$. The contribution of the other 7 observations in the block makes the overall estimate of the conditional mode $\hat{b}_5 = -0.43$.

by a vector of parameters, representing the variances of all of the random effects and the covariances among them.

Restricted likelihood

Many of the useful properties of maximum likelihood estimates, such as efficiency and lack of bias, only hold *asymptotically*—that is, when the data set is large. In particular, maximum likelihood estimates of variances are biased downward because they ignore uncertainty in the sample means. You may remember that the usual formula for estimating sample variance is $\sum(x-\bar{x})^2/(n-1)$, rather than $\sum(x-\bar{x})^2/n$ (the latter is the maximum likelihood estimate), for exactly this reason: dividing by a smaller number ($n-1$ rather than n) increases the estimate just enough to account for the uncertainty in $\bar{x}$. Restricted maximum likelihood (REML) generalizes this idea to allow for less biased estimates of the variances in mixed models. Technically, it is based on finding some way to combine the observations that factors out the fixed effects. For example, in a pairwise t-test the average difference between the two observations in a pair is equal to the difference between treatments, which is the fixed effect. Since we are usually interested in the difference between the treatments, we compute the difference between treatments in each pair. If instead we took the average of each pair, we would cancel out the fixed effect, and could then compute an unbiased estimate of the variance among the pairs. A broader way of thinking about REML is that it applies to any statistical method where we integrate over the fixed effects when estimating the variances. When using REML, you *cannot* compare the restricted likelihoods of two models with different sets of fixed effects, because they are likelihoods of completely different models for the variance. While REML in principle applies to GLMMs as well as LMMs, they are more easily defined and more accessible in

software for LMMs than for GLMMs (Bellio and Brazzale 2011; Millar 2011). It's generally good to use REML, if it is available, when you are interested in the magnitude of the random effects variances, but *never* when you are comparing models with different fixed effects via hypothesis tests or information-theoretic criteria such as AIC.

13.4 Setting up a GLMM: practical considerations

13.4.1 *Response distribution*

The conditional distribution of the response variable, which we often abbreviate to "the response distribution" or "the distribution of the data," is the expected distribution of each observed response around its predicted mean, given the values of all of the fixed and random effects for that observation. That is, when we collect a data set of (for example) counts, we don't expect the overall (marginal) distribution of the data to be Poisson; we expect each point to be drawn from a Poisson distribution with its own mean that depends on the predictors for that point (chapter 6). In the gopher tortoise example, the distribution of number of shells S in a given site s (with infection seroprevalence $P(s)$) and year y is $S_{sy} \sim \text{Poisson}(\beta_0 + \beta_y + \beta_P P(s) + b_s)$, where β_0 is the baseline (year-0, 0-prevalence, average site) expectation; β_y is the difference between year y and the baseline; β_P is the effect of an additional percentage of seroprevalence; and b_s gives the difference between site s and the overall average.

If the conditional distribution is Gaussian, or can sensibly be transformed to be Gaussian (e.g., by log transformation), as in the tundra ecosystem example, then we have a *linear* mixed model, and several aspects of the modeling process are simpler (we can more easily define R-side effects and restricted maximum likelihood; statistical tests are also easier: see section 13.6). As with GLMs (chapter 6), binomial (including binary or Bernoulli, i.e., 0/1 responses) and Poisson responses comprise the vast majority of GLMMs. The Gamma distribution is the other common distribution handled by GL(M)Ms; it is useful for continuous, skewed distributions, but treating such data as lognormal (i.e., log-transforming and then using a linear mixed model) is easier and usually gives very similar results.

In addition to these standard distributions, there are other useful distributions that do not technically fall within the scope of GLMMs, but can sometimes be handled using simple extensions. These include the negative binomial distribution for overdispersed count data; zero-inflated distributions for count data with excess zeros (chapter 12); the beta distribution for proportional data that are not proportions out of a known total count; and the Tweedie distribution for continuous data with a spike at zero (`glmmADMB` handles the first three cases). Ordinal responses (i.e., categorical responses that have more than two ordered categories) and multinomial responses (categorical responses with more than two categories, but without ordering) can be handled by extensions of binomial GLMMs, implemented in the `clmm` function in the `ordinal` library. These extensions are often useful, but using them will generally make it harder to analyze your model (you are more likely to run into computational difficulties, which will manifest themselves as warnings and errors from software), and restrict your choice of software more than if you stick to the simpler (Normal, binomial, Poisson) distributions.

As is typical in ecological applications, the examples for this chapter all use either Normal (tundra ecosystem), binary (coral symbiont), or Poisson (gopher tortoise, grouse tick) conditional distributions (table 13.1). The family is specified almost exactly as in

standard GLMs, with a few quirks. For linear mixed models, you should use the `lme` (in the `nlme` library) or `lmer` (in the `lme4` library), or `family="gaussian"` in `MCMCglmm` or `glmmADMB`. `MCMCglmm` and `glmmADMB` require the family argument to be given as a quoted string (e.g., `family="poisson"`), in contrast to `lme4`, which allows more flexibility (e.g., `family="poisson"` or `poisson()`). `MCMCglmm` has different names from the standard R conventions for binary/logit (`family="categorical"`) and binomial (`family="multinomial2"`) models.

13.4.2 *Link function*

As with GLMs, we also have to choose a link function to describe the shape of the response curve as a function of continuous predictor variables. The rules for picking a link function are the same as for GLMs: when in doubt, use the default (canonical) link for the response distribution you have chosen. We will follow this rule in the examples, using the default logit link for the coral symbiont (binary) example and a log link for the gopher tortoise and grouse tick (Poisson) examples (table 13.1), although we did also consider a log link for the coral symbiont example. In `lme4` links are specified along with the family as for standard GLMs in R, e.g., `family=binomial(link="logit")` or `binomial(link="log")`; in `glmmADMB` they are specified as a separate string (`link="logit"`); and `MCMCglmm` uses alternative family names where alternate links are available (e.g., `family="ordinal"` for a binary/probit link model).

13.4.3 *Number and type of random effects*

As discussed in section 13.3.1, it is not always easy to decide which variables to treat as random effects. The more random effects a model includes, the more likely you are to run into computational problems. It is also more likely that the fit will be *singular*: some random effects variances will be estimated as exactly zero, or some pairs of random effects will be estimated as perfectly correlated. While this does not necessarily invalidate a particular model, it may break model-fitting software in either an obvious way (errors) or a non-obvious way (the model is more likely to get stuck and give an incorrect result, without warning you). Model complexities interact: for example, some of the software available to fit models with non-standard distributions can only handle a single random effect. In general you should avoid: (1) fitting random effects to categorical variables with fewer than five levels, and, unless you have very large data sets and a fast computer, (2) fitting more than two or three random effects in a single model or (3) fitting vector-valued random effects (i.e., among-group variation of responses to categorical variables) for categorical predictors with more than two or three levels.

13.5 Estimation

Once the model is set up, you need to estimate the parameters—the fixed-effect parameters that describe overall changes in the response, the conditional modes of the random effects that describe the predicted differences of each level of the grouping variable from the population average, and the variances of, and covariances among, the random effects. This isn't always easy; there are a variety of methods, with trade-offs in speed and availability.

13.5.1 *Avoiding mixed models*

Sometimes fitting a mixed model is difficult: for example, if you have too few levels of your random effect, or repeated measurements within just a few blocks. In this case fitting a mixed model doesn't have many advantages, and you may be able to take a shortcut instead.

- For data from a nested experimental design, taking the average of each block and doing a one-way ANOVA on the results will give you exactly the same results for the fixed effects as you would get from a mixed model (Murtaugh 2007); if your data are unbalanced you can do a weighted ANOVA with weights of $1/n_i$ (where n_i is the number of observations in the ith block). If you want to allow (for example) varying slopes across blocks, you can fit a *two-stage model*, where you fit a linear regression for each block separately and then do a one-way ANOVA on the slopes. This works best with Normal data, but if you have many points per block the block averages will be approximately Normal—although you may still need to deal with heteroscedasticity, e.g., by transforming the data appropriately.
- You can try showing that random effects are ignorable by fitting a model that ignores random effects, and then using a one-way ANOVA on the residuals of the model by block to show that they do not vary significantly across blocks.
- If you need to compute the among-block variance when there are too few levels (< 5), you can fit the blocks as a fixed effect, with "sum to zero" contrasts set, and compute the mean of the squared coefficients $(\sum(\beta_i - \text{mean}(\beta))^2/(n-1))$.
- If you have paired comparisons (i.e., you are testing the difference between two fixed effect levels, such as treatment vs. control within each block) for normally distributed responses, you can replace the test of the fixed effect with a paired t-test, and estimate the among-block variance by computing the variance of ((control + treatment)/2) across blocks.

For many situations (e.g., randomized block or crossed designs, or pairwise comparisons of non-Normal data), you may not be able to use these shortcuts and will have to proceed with a mixed model.

13.5.2 *Method of moments*

The traditional way to fit a mixed ANOVA model is to compute appropriate sums of squares (e.g., the sum of squares of the deviations of the group means from the grand mean, or the deviations of observations from their individual group means) and dividing them by the appropriate degrees of freedom to obtain mean squares, which are estimates of the variances. This approach is called the method of moments because it relies on the correspondence between the sample moments (mean squares) and the theoretical parameters of the model (i.e., the random effects variances). This approach is simple, fast, always gives an answer—and is extremely limited, applying only to Normal responses (i.e., linear mixed models), in balanced or nearly balanced designs, with nested random effects only (see Gotelli and Ellison 2004 or Quinn and Keough 2002).

13.5.3 *Deterministic/frequentist algorithms*

Instead of computing sums of squares, modern estimation approaches try to find efficient and accurate ways to compute the marginal likelihood (section 13.3.2), which

can be challenging. The first class of approaches for estimating mixed models, which I call deterministic approaches (note that this is not standard terminology), are typically used in a frequentist statistical framework to find the maximum likelihood estimates and confidence intervals.

- *Penalized quasi-likelihood* (PQL, Breslow 2004) is a quick but inaccurate method for approximating the marginal likelihood. While it is fast and flexible, it has two important limitations. (1) It gives biased estimates of random-effects variances, especially with binary data or count data with low means (e.g., Poisson with mean < 5). More accurate versions of PQL exist, but are not available in R. The bias in random-effect variances may be unimportant if your questions focus on the fixed effects, but it's hard to be sure. (2) PQL computes a quantity called the "quasi-likelihood" rather than the likelihood, which means that inference with PQL is usually limited to less-accurate Wald tests (section 13.6.2).
- *Laplace approximation* is a more accurate, but slower and less flexible, procedure for approximating the marginal likelihood.
- *Gauss–Hermite quadrature* (GHQ) is a more accurate, but still slower and less flexible approach. Where Laplace approximation uses one point to integrate the marginal likelihood, GHQ uses multiple points. You can specify how many points to use; using more is slower but more accurate. The default is usually around eight; `lme4` allows up to 25, which is usually overkill. Many software packages restrict GHQ to models with a single random effect.

You should use the most accurate algorithm available that is fast enough to be practical. If possible, spot-check your results with more accurate algorithms. For example, if Laplace approximation takes a few minutes to fit your models and GHQ takes a few hours, compare Laplace and GHQ for a few cases to see if Laplace is adequate (i.e., whether the difference between the coefficient values between the two methods is small relative to their standard errors).

13.5.4 *Stochastic/Bayesian algorithms*

Another approach to GLMM parameter estimation uses *Markov chain Monte Carlo* (MCMC), a stochastic estimation algorithm. There's not nearly enough room in this chapter to give a proper explanation of MCMC; you can just think of it as a general computational recipe for sampling values from the probability distribution of model parameters. Stochastic algorithms are usually much slower than deterministic algorithms, although a single run of the algorithm provides both the coefficients and the confidence intervals, in contrast to deterministic algorithms, where computing reliable confidence intervals may take several times longer than just finding the coefficients.

Although there is at least one "black box" R library (`MCMCglmm`) that allows the user to define the fixed and random effects via the sorts of formulas shown in table 13.1, many researchers who opt for stochastic GLMM parameter estimation use the BUGS language instead (i.e., the WinBUGS package or one of its variants such as OpenBUGS or JAGS) to fit their models. BUGS is a flexible, powerful framework for fitting ecological models to data in a Bayesian context (McCarthy 2007; Kéry 2010), not just GLMMs, but it comes with its own steep learning curve.

For technical reasons, most stochastic algorithms use a Bayesian framework, usually with weak priors (chapter 1); except when you have parameters that are very uncertain, this distinction doesn't make a huge practical difference. Bayesian inference, and

stochastic algorithms in general, make it much easier to compute confidence intervals that incorporate all the relevant sources of uncertainty (section 13.6.2). If you want to use stochastic algorithms but avoid Bayesian methods, you can use a stochastic algorithm that works within a frequentist framework, such as *data cloning* (Ponciano et al. 2009; Sólymos 2010).

13.5.5 *Model diagnostics and troubleshooting*

Model checking for GLMMs overlaps a lot with the procedures for GLMs (chapter 6). You should plot appropriately scaled residuals (i.e., deviance or Pearson residuals) against the fitted values and against the input variables, looking for unexplained patterns in the mean and variance; look for outliers and/or points with large influence (leverage); and check that the distribution of the residuals is reasonably close to what you assumed. For Poisson or binomial GLMMs with $N > 1$, you should compare the sum of the squared Pearson residuals to the residual degrees of freedom (number of observations minus number of fitted parameters) to check for overdispersion (unless your data are binary, or the model already contains an observation-level random effect; appendix 13A).

The first GLMM-specific check is to see whether the model is singular: that is, whether non-zero variances (and non-perfect correlations among random effects, i.e., $|\rho| < 1$) could be estimated for all the random effects in the model. If some of the variances are zero or some correlations are ± 1, it indicates that not only was the among-group variation not significantly different from zero, the best estimate was zero. Your model is probably too complex for the data: the best way to avoid this problem in general is to try to simplify the model *in advance* to a level of complexity that you think the data can support, by leaving out random-effects terms or by converting them to fixed effects. It does take some practice to calibrate your sense of what models can be fitted. For example, in the coral symbiont example I left out the block × treatment interaction, successfully fitting a non-singular model, but in the gopher tortoise example the model with site and observation-level random effects was singular even though I had tried to be conservative by treating year as a fixed effect.

Although in principle the results from a singular model fit will be the same as if you had just left the zero-estimate terms out of the model in the first place, you should probably refit the model without them to make sure this is true (i.e., that the software hasn't run into computational problems because the model was too complicated). Another possible solution to this problem is to impose a Bayesian prior on the variances to push them away from zero, which you can do using the `blme` (Chung et al. 2013) or `MCMCglmm` libraries. Although some researchers advocate simply picking a reasonable model and sticking with it (i.e., not looking for a more parsimonious reduced model: Barr et al. 2013), you can also use information-theoretic approaches (AIC or BIC) to choose among possible candidate random-effects models (see chapter 3 and section 13.6.2), especially if you are interested in prediction rather than in testing hypotheses.

Another diagnostic specific to (G)LMMs is checking the estimates of the conditional modes. In theory these should be Normally distributed (you can check this using a *quantile–quantile* ("q–q") plot). You should only worry about extreme deviations: no-one really knows how badly a non-Normal distribution of conditional modes will compromise a (G)LMM, and fitting models with non-Normal modes is difficult. Look for extreme conditional modes and treat them as you would typically handle outliers; for example, figure out whether there is something wrong with the data for those groups, or try fitting the model with these groups excluded and see whether the results change very much.

For MCMC analyses (e.g., via `MCMCglmm`), you should use the usual diagnostics for convergence and mixing (read more about these in McCarthy (2007) and Kéry (2010)), check quantitative diagnostics such as the Gelman–Rubin statistic and effective sample size, and examine graphical diagnostics (trace and density plots) for both the fixed and random effects parameters. With small data sets, the variance–covariance parameters often mix badly, sticking close to zero much of the time and occasionally spiking near zero; the corresponding density plots typically show a spike at zero with a long tail of larger values. There are no really simple fixes for this problem, but some reasonable strategies include (1) running much longer chains; (2) adding an informative prior to push the variance away from zero; (3) taking the results with a grain of salt (appendix 13A).

As you try to troubleshoot the random-effects component of your analysis, you should keep an eye on the fixed-effect estimates and confidence intervals associated with models with different random effects structures; the fixed-effect estimates often stay pretty much the same among models with different random effects. This can be comforting if your main interest is in the fixed effects, although you should be careful since fitting multiple models also allows some scope for cherry-picking the results you like.

13.5.6 *Examples*

Some technical issues that arose as I fitted and diagnosed models for the examples above were (appendix 13A for more details):

- *Tundra CO_2 flux*: Overall the fits were well-behaved, but one site (Toolik) differed from the others; its observations were poorly fitted by the full model (they had large residuals with high variance) and its conditional mode for the slope was an outlier. Including the Toolik data made the model harder to fit, and generated autocorrelation in the residuals that could not be completely accounted for. However, the primary estimate of the population-level rate of increasing CO_2 flux remained qualitatively similar whether we included the Toolik data or not.
- *Coral symbionts*: One observation in the data set was poorly predicted—it was a coral that escaped predation although it had a high expected predation risk (it was in the no-symbiont treatment in a frequently attacked block). Refitting the model without this observation led to nearly complete separation (chapter 6), making the estimates even more extreme. In the end we retained this data point, since including it seemed to give conservative estimates. Other aspects of the model looked OK—the distribution of conditional modes was sensible, and using GHQ instead of the Laplace approximation changed the estimates only slightly.
- *Gopher tortoise shells*: Poisson sampling accounted for nearly all the variation in the data—the estimated variances both among observations and among sites were very close to zero. Thus, the conditional modes were also all near zero. In other words, we would have obtained similar results from a simple Poisson GLM. The residuals looked reasonable, with similar variation in each site. The `MCMCglmm` fit, which included both among-site and among-observation variation, showed unstable estimates of the random-effects variance, as described in section 13.5.5. We couldn't simplify the model, but instead used a stronger prior on the among-site variance to stabilize it. This didn't change the estimated effect of disease prevalence, but did increase its uncertainty. In Ozgul et al. (2009) we fitted the full model with WinBUGS; if I ran the analysis again today I would either fit a simple Poisson model or use `MCMCglmm` or `blme` to fit the full model with stabilizing priors.

- *Grouse ticks*: The residuals and estimated conditional modes all looked reasonable. We didn't test for overdispersion since the model includes observation-level random effects. Deterministic algorithms (`lme4` and `glmmADMB`) gave positive estimates for all of the variances, but `MCMCglmm` disagreed; unless we added a prior, it estimated the among-location variance as nearly zero, suggesting that the separation of variation into among-brood vs. among-location components is unstable.

13.6 Inference

13.6.1 *Approximations for inference*

Estimates of parameters are useless without confidence intervals, or hypothesis tests (p-values), or information criteria such as AIC, that say how much we really know. Inference for GLMMs inherits several assumptions from GLMs and linear mixed models that do not apply exactly for GLMMs, and which (as with the estimation methods for GLMMs) require trade-offs between accuracy, computation time, and convenience or availability in software. As with estimation (section 13.5), you should generally use the slowest but most accurate method that is practical, double-checking your results with a slower and more accurate method if possible. Inference for GLMMs involves three separate types of approximation, which we will discuss in general before discussing specific methods for inference in section 13.6.2.

Shape: the fastest but least accurate approaches to GLMM inference (Wald intervals and tests) make strong assumptions about the shape of the likelihood curve, or surface, that are exactly true for linear models (ANOVA/regression), but only approximately true for GLMs, LMMs, and GLMMs. These approximations are more problematic for smaller data sets, or for data with high sampling variance (binary data or Poisson or binomial data with small observed counts or numbers of successes/failures).

Finite-size effects: When the data set is not very large (e.g., < 40 observations, or < 40 levels for the smallest random-effect grouping variable) we have to make further assumptions about the shapes of distributions of summaries such as the likelihood ratio or F statistic. In the classical ANOVA or regression framework, these assumptions are taken care of by specifying the "denominator degrees of freedom," that is, specifying the effective number of independent observations. For GLMs, for better or worse, people usually ignore these issues completely.

- For LMMs that don't fit into the classical ANOVA framework (i.e., with unbalanced designs, crossed random effects, or R-side effects), the degrees of freedom for the t distribution (for testing individual parameters), or the denominator degrees of freedom for the F distribution (for testing effects), are hard to compute and are at best approximate. If your experimental/observational design *is* nested and balanced, you can use a software package that computes the denominator degrees of freedom for you or you can look the experimental design up in a standard textbook (e.g., Gotelli and Ellison 2004 or Quinn and Keough 2002). If not, then you will need to use the approximation methods implemented in the `lmerTest` and `pbkrtest` libraries (Kenward and Roger 1997; Halekoh and Højsgaard 2013), or use a resampling-based approach (section 13.6.2).
- GLMMs involve a different finite-size approximation (the distribution of the likelihood ratio test statistic is approximate rather than exact). Stroup (2014) states that the Kenward–Roger approximation procedure developed for LMMs works reasonably well for GLMMs, but neither it nor Bartlett corrections (another approximation method

described in McCullagh and Nelder 1989) are implemented for GLMMs in R; you will need to use stochastic sampling methods (section 13.6.2) if you are concerned about finite-size inference for GLMMs.

Boundary effects: statistical tests for linear models, including GLMMs, typically assume that estimated parameters could be either above or below their null value (e.g., slopes and intercepts can be either positive or negative). This is not true for the random effect variances in a (G)LMM—they must be positive—which causes problems with standard hypothesis tests and confidence interval calculations (Pinheiro and Bates 2000). In the simplest case of testing whether a single random-effect variance is zero, the p-value derived from standard theory is twice as large as it should be, leading to a conservative test (you're more likely to conclude that you can't reject the null hypothesis). To test the null hypothesis that the sole random-effect variance in a model is equal to zero you can just divide the p-value by 2. If you want to test hypotheses about random effects in a model with more than one random effect you will need to simulate the null hypothesis (section 13.6.2).

13.6.2 *Methods of inference*

Wald tests

The standard errors and p-values that R prints out when you summarize a statistical model (*Wald* standard errors and tests) are subject to artifacts in GLM or GLMM modeling. They're especially bad for binomial data where some categories in the data have responses that are mostly (or all) successes or failures (*complete separation*: the related inference problems are called the *Hauck–Donner effect*: Venables and Ripley 2002). The typical symptom of these problems is large parameter estimates (e.g., absolute value>10) in conjunction with huge standard errors and very large ($p \approx 1$) p-values: sometimes, but not always, you will also get warnings from the software. More generally, Wald statistics are less accurate than the other methods described below. However, they are quick to compute, can be useful for a rapid assessment of parameter uncertainty, and are reasonably accurate for large data sets. If you can guess the appropriate residual degrees of freedom, then you may try to use appropriate t statistics rather than Z statistics for the p-values and confidence interval widths in order to account for finite sample sizes, but this is a crude approximation in the case of GLMMs.

Likelihood ratio tests

Likelihood ratio tests and profile confidence intervals are an improvement over Wald statistics, but come at a computational cost that may be significant for large data sets. You can use the likelihood ratio test to compare nested models (via the `anova` command in R, or by computing the p-value yourself based on the χ^2 distribution); this provides a significance test for the factors that differ between the two models. The corresponding confidence intervals for a parameter are called profile confidence intervals. Profile confidence intervals are computationally challenging—they may take dozens of times as long to compute as the original model fit. Furthermore, because profile likelihood calculations have to evaluate the likelihood for extreme parameter values, they are much more subject to computational problems than the original model fit.

Finally, although likelihood-based comparisons are more reliable than Wald statistics, they still assume infinite denominator degrees of freedom. If your effective sample size is large enough (e.g., the smallest number of levels of any grouping variable in your model

is > 40), you don't need to worry. Otherwise you may need to use a stochastic resampling method such as parametric bootstrapping or Markov chain Monte Carlo for accurate inference.

Bootstrapping

Bootstrapping means resampling data with replacement to derive new pseudo-data sets, from which you can estimate confidence intervals (chapter 1). Parametric bootstrapping (PB) instead simulates pseudo-data from the fitted model (or from *reduced* models that omit a parameter you are interested in making inferences about). You can then refit your model to these pseudo-data sets to get reliable *p*-values or confidence intervals.

PB is very slow (taking hundreds or thousands of times as long as fitting the original model), and it does make assumptions—that the model structure is appropriate, and that the estimated parameters are close to the true parameters—but it is the most accurate way we know to compute *p*-values and confidence intervals for GLMMs.

Specialized forms of PB are faster. For example, the `RLRsim` library in R (Scheipl et al. 2008) does a kind of PB to compute *p*-values for random-effect terms in LMMs, orders of magnitude faster than standard PB.

You can also use non-parametric bootstrapping—resampling the original data values—but you must respect the grouping structure of the data. For example, for a model with a single grouping variable you could do two-stage bootstrapping (Field and Welsh 2007), first sampling with replacement from the levels of the grouping variable, then sampling with replacement from the observations within each sampled group. For more complex models (with crossed random effects, or R-side effects), appropriate resampling may be difficult.

MCMC

The results of an MCMC fit (section 13.5.4) give estimates and confidence intervals on parameters; you can also get *p*-values from MCMC, although it is unusual (since most MCMC is based in a Bayesian framework). MCMC is very powerful—it automatically allows for finite size effects, and incorporates the uncertainty in all the components of the model, which is otherwise difficult. It's so powerful, in fact, that some frequentist tools (such as AD Model Builder) use a variant of MCMC to compute confidence intervals. This pseudo-Bayesian approach is convenient, but may have problems when the information in the data is weak. For small, noisy data sets the distribution of the variance parameters is often composed of a spike at zero along with a second component with a mode away from zero. In this case, many MCMC algorithms can get stuck sampling either the spike or the non-zero component, and thus give poor results.

Information-theoretic approaches

Many ecological researchers use information-theoretic approaches to select models and generate parameter importance weights or weighted multimodel averages of parameters and predictions (chapter 3; Burnham and Anderson 2002). In principle, AIC (and other indices like BIC) do apply to mixed models, but several of the theoretical difficulties discussed in section 13.6.1 affect information criteria (Greven and Kneib 2010; Müller et al. 2013).

- AIC comparisons among models with different variance parameters have the same problem as null-hypothesis tests of variances (section 13.6.1)—they tend to understate the importance of variance terms.

- When comparing models with different random-effects terms, or when using a finite-size corrected criterion such as AICc, the proper way to compute the model complexity (number of parameters) associated with a random effect depends on whether you are trying to predict at the population level (predicting the average value of a response across all random-effects levels) or at the individual level (a *conditional* prediction, i.e., making predictions for specific levels of the random effect). For population-level prediction, you should count one parameter for each random-effects variance or covariance/correlation. For conditional prediction, the correct number of parameters is somewhere between 1 and $n-1$, where n is the number of random-effects levels: methods for computing appropriate AIC values in this case (Vaida and Blanchard 2005) are not widely implemented. Academic ecologists typically want to know about effects at the level of the whole population, which allows them to use the easier one-parameter-per-variance-parameter rule; applied ecologists might be more interested in predictions for specific groups. Bayesian MCMC has an information-theoretic metric called the *deviance information criterion* (DIC: Spiegelhalter et al. 2002), for which the so-called level of focus must be defined similarly (O'Hara 2007).
- Finite-size-corrected criteria such as AICc are poorly understood in the mixed model context. For example, for n in the denominator of the AICc correction term ($n-k-1$: chapter 3), should one count the total number of observations in a nested design, or the number of groups? For better or worse, most ecologists use AICc for model selection with GLMMs without worrying about these issues, but this may change as statisticians come to understand AICc better (Shang and Cavanaugh 2008; Peng and Lu 2012).

In general you should *pick a single approach to modeling and inference in advance*, or after brief exploration of the feasibility of different approaches, in order to avoid the ever-present temptation to pick the results you like best.

13.6.3 *Reporting the GLMM results*

Graphical summaries of statistical analyses that display the model coefficients and their uncertainty, or that overlay model predictions and their uncertainties on the original

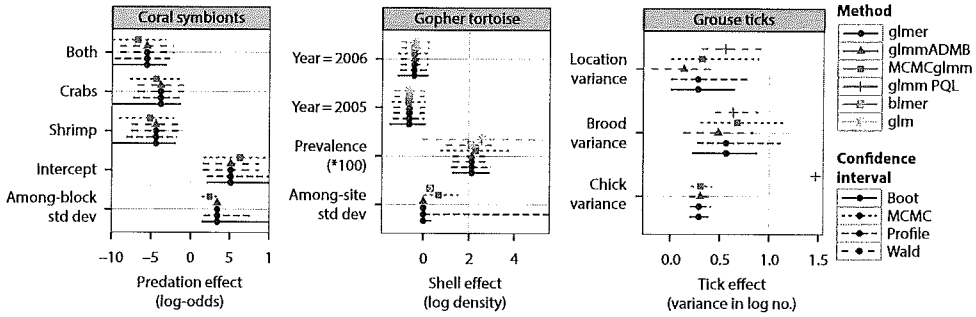

Fig. 13.3 Comparisons of all estimation/inference methods showing estimates and 95% confidence intervals for all three GLMM examples. For the most part all methods give similar results; the biggest differences are in the `MCMCglmm` estimates (which represent posterior means rather than maximum likelihood estimates) and in the estimates and confidence intervals for random-effect standard deviations or variances.

data, are important (Gelman et al. 2002). However, you also need to summarize the results in words. This summary should include the magnitudes and confidence intervals of the fixed effects; the magnitude of the among-group variation for each random effect, whether it is of primary interest or not; and possibly the confidence intervals of the among-group variation (if the random effects are included because they are part of the design, you should *not* test the null hypothesis that they are zero). If you are interested in the partitioning of variance across levels, report among-group variation as random-effect variances, or proportions of variance (see the grouse tick example below). If you are more interested in the fixed effects, report among-group variation as random-effect standard deviations, as these are directly comparable to the corresponding fixed effects. The following are sample reports for the four worked examples; appendix 13A shows the technical details of deriving these results. The results from all the combinations of estimation and inference methods in this chapter are summarized in Figure 13.3.

- *Tundra carbon*: The main effect of interest is the across-site average change in growing-season carbon flux per year; the estimated slopes are negative because the rate of carbon loss is increasing. Our conclusion from the fitted model with the year variable centered (i.e., setting `Year=0` to the overall mean of the years in the data) would be something like: "the overall rate of change of growing season NEE was –3.84 g C/m^2/season/year (t_{23} = –2.55, p = 0.018, 95% CI = {–6.86, –0.82}). We estimated a first-order autocorrelation within sites of ρ = 0.39; among-site variation in the intercept was negligible, while the among-site standard deviation in slope was 5.07 g C/m^2/season/year, with a residual standard deviation of 58.9 g C/m^2/season."
- *Coral symbionts*: For the analysis done here (logit link, one-way comparison of crab/shrimp/both to control) we could quote either the fixed-effect parameter estimates (clarifying to the reader that these are differences between treatments and the baseline control treatment, on the logit or log-odds scale), or the changes in predation probability from one group to another. Taking the first approach: "Crab and shrimp treatments had similar effects (–3.8 log-odds decrease in predation probability for crab, –4.4 for shrimp); the dual-symbiont treatment had an even larger effect (–5.5 units), but although the presence of any symbiont caused a significant drop in predation probability relative to the control (Wald p-value 0.0013; parametric bootstrap p-value < 0.003), none of the symbiont treatments differed significantly from each other (likelihood ratio test p = 0.27, parametric bootstrap test (N = 220) p = 0.23); in particular, two symbionts did not have significantly greater protective effects than one (Wald and PB p-values both $\approx$0.15). The among-block standard deviation in log-odds of predation was 3.4, nearly as large as the symbiont effect." (McKeon et al. (2012) present slightly different conclusions based on a model with a log rather than a logit link.) Alternately, one could quote the predicted predation probabilities for each group, which might be more understandable for an ecological audience.
- *Gopher tortoise*: The main point of interest here is the effect of prevalence on the (per-area) density of fresh shells. This makes reporting easy, since we can focus on the estimated effect of prevalence. Because the model is fitted on a log scale and the parameter estimate is small, it can be interpreted as a proportional effect. For example: "A 1% increase in seroprevalence was associated with an approximately 2.1% increase (log effect estimate = 0.021) in the density of fresh shells (95% CI = {0.013, 0.031} by parametric bootstrap [PB]). Both of the years subsequent to 2004 had lower shell densities (log-difference = –0.64 (2005), –0.43 (2006)), but the differences were not statistically significant (95% PB CI: 2005 = {–1.34, 0.05}, 2006 = {–1.04, 0.18}). There was no

detectable overdispersion (Pearson squared residuals/residual df = 0.85; estimated variance of an among-observation random effect was zero). The best estimate of among-site standard deviation was zero, indicating no discernible variation among sites, with a 95% PB CI of {0, 0.38}."

- *Grouse ticks*: In this case the random-effects variation is the primary focus, and we report the among-group variance rather than standard deviation because we are interested in variance partitioning. "Approximately equal amounts of variability occurred at the among-chick, among-brood, and among-location levels (MCMCglmm, 95% credible intervals: σ^2_{chick} = 0.31 [95% CI {0.2, 0.43}, σ^2_{brood} = 0.59 {0.36, 0.93}, $\sigma^2_{\text{location}}$ = 0.57 {0.29, 1.0}]. The among-brood variance is estimated to be approximately twice the among-chick and among-location variances, but there is considerable uncertainty in the brood/chick variance ratio ($\sigma^2_{\text{brood}}/\sigma^2_{\text{chick}}$ = 2.01 {1.007, 3.37}), and estimates of the among-location variance are unstable. Year and altitude also have strong effects. In 1996, tick density increased by a factor of 3.3 relative to 1995 (1.18 {0.72, 1.6} log units); in 1997 density decreased by 38% (–0.98 {–1.49, 0.46} log units) relative to 1995. Tick density increased by approximately 2% per meter above sea level (–0.024 {–0.03, –0.017} log-units), decreasing by half for every 30 (log(2)/0.024) m of altitude."

13.7 Conclusions

I hope you are convinced by now that GLMMs are a widely useful tool for the statistical exploration of ecological data. Once you get your head around the multi-faceted concept of random effects, you can see how handy it is to have a modeling framework that naturally combines flexibility in the response distribution (GLMs) with the ability to handle data with a variety of sampling units with uneven and sometimes small sample sizes (mixed models).

GLMMs cannot do everything; especially for very small data sets, they may be overkill (Murtaugh 2007). Ecologists will nearly always have too little data to fit as sophisticated a model as they would like, but one can often find a sensible middle ground.

In this chapter I have neglected the other end of the spectrum, very large data sets. Ecologists dealing with Big Data from remote sensing, telemetry, citizen science, or genomics may have tens or hundreds of thousands of observations rather than the dozens to hundreds represented in the examples here. However, telemetry and genomic data often contain huge amounts of detail about a small number of individuals; in this case a fixed-effect or two-stage (Murtaugh 2007) model may work as well as a GLMM. The good news is that some of the computational techniques described here scale well to very large data sets, and some of the most computationally intensive analyses become unnecessary when all the grouping variables have more than 40 levels.

I have also neglected a variety of useful GLM extensions such as non-standard link functions (for fitting specific non-linear models such as the Beverton–Holt or Ricker functions); methods for handling multinomial or ordinal data; and zero-inflation. The good news is that most of these tricks are at least in principle extendable to GLMMs, but your choice of software may be more limited (Bolker et al. 2013).

Unfortunately, GLMMs do come with considerable terminological, philosophical, and technical baggage, which I have tried in this chapter to clarify as much as possible. As GLMM software, and computational power, continue to improve, many of the technical difficulties will fade, and GLMMs will continue their growth in popularity; a firm grasp of the *conceptual* basis of GLMMs will be an increasingly important part of the quantitative ecologist's toolbox (Zuur et al. 2009, 2012, 2013; Millar 2011).