Noname manuscript No.
(will be inserted by the editor)

Scaling for Dynamical Systems in Biology

Glenn Ledder

the date of receipt and acceptance should be inserted later

Abstract Asymptotic methods can greatly simplify the analysis of all but the simplest mathematical models and should therefore be commonplace in such biological areas as ecology and epidemiology. One essential difficulty that limits their use is that they can only be applied to a suitably-scaled dimensionless version of the original dimensional model. Many books discuss nondimensionalization, but with little attention given to the problem of choosing the right scales and dimensionless parameters. In this paper, we illustrate the value of using asymptotics on a properly-scaled dimensionless model, develop a set of guidelines that can be used to make good scaling choices, and offer advice for teaching these topics in differential equations or mathematical biology courses.

Keywords asymptotics, scaling, epidemiology, modeling

Mathematics Subject Classification (2000) 97M60

1 Introduction

Asymptotic methods are techniques for obtaining approximate solutions to mathematical problems that contain small dimensionless parameters. Their use is ubiquitous in the physical sciences, but relatively rare in biology (but see Banasiak and Lachowicz (2014) and Huang et al (2015)). This is unfortunate, because many biological settings in areas such as ecology and epidemiology have processes with widely disparate time scales, allowing for significant use of asymptotics.

Mathematical biologists can acquire the necessary background in asymptotics from a number of excellent books (for example: Bush 1992; Fowler 1997; Hinch 1991; Holmes 2013; Howison 2005; Logan 2013; Murdock 1991). However, knowledge of asymptotic methods is not sufficient to allow their use in biological modeling, because models do not start out in a suitable dimensionless form for exploiting the presence of small parameters. Recasting models in some dimensionless form is a simple matter, but finding a "suitable" form is not.

Nondimensionalization is merely the systematic replacement of dimensional quantities with dimensionless quantities, without regard for the magnitudes of the new quantities. For example, if E is the population of exposed individuals in an epidemiology model and N is the (fixed) total population, then $x = E/N$ is a dimensionless version of E using N as a reference population value. Any reference value can be used, as having the correct dimensions is the only requirement.

Nondimensionalization eliminates unnecessary parameters (in the above example, N will become a common factor of those equations where it occurs), and that alone confers some benefit (Section 2.1). However, asymptotics requires that a dimensionless model be properly scaled, which means that any significant difference in magnitudes of terms in an equation is explicitly indicated by the presence of small or large parameters.

G. Ledder
Department of Mathematics, University of Nebraska-Lincoln, Lincoln, NE 68588-0130,
E-mail: gledder@unl.edu

For example, suppose ϵ is a small parameter in the dimensionless equation

$$\epsilon \frac{dx}{dt} = x.$$

The smallness of ϵ means that either x is also small or dx/dt is large; in either case, the equation is not properly scaled. In contrast, the dimensionless equation

$$\epsilon \frac{dx}{dt} = x - 1$$

is properly scaled if x is not changing rapidly, which happens when x differs from 1 by only a small amount.

Scaling is nondimensionalization with reference values that are representative of the values the variables will actually have. For example, if the fraction of the population that is exposed is never more than 1%, then the quantity E/N will always be small, which means that the "smallness" of the corresponding term will appear in the term itself rather than the coefficient of the term. If instead we choose a smaller reference value for E, then the smallness of the term will appear explicitly as a small dimensionless parameter. Small parameters multiplying terms that are properly scaled make possible the use of asymptotic arguments to simplify the analysis of a model or to obtain a simpler model that approximates the original one (Section 2.2).

Few mathematical biology texts incorporate nondimensionalization, much less with scaling. Thieme (2003) makes sporadic use of nondimensionalization, but without justifying the choices of reference quantities. Banasiak and Lachowicz (2014) scale most of their models in preparation for asymptotic analysis, but again without discussion of the choices they make.

Outside of mathematical biology, nondimensionalization and scaling are standard topics in texts on mathematical modeling and asymptotics, beginning with the work of the pioneering applied mathematician Lee Segel (Segel 1972; Lin and Segel 1988). Segel was careful in his writing to draw a distinction between nondimensionalization and scaling; these writings offer some guidance to the modern practitioner, but the examples are relatively simple physical science problems with few parameters and correspondingly few choices for scales. In contrast, most of the more recent modeling texts use the words "nondimensionalization" and "scaling" almost interchangeably, largely reducing the choice of reference quantities to the trivial problem of finding parameter combinations having the right dimensions, which they do by dimensional analysis alone. Given a problem with multiple scaling choices, mere nondimensionalization is as likely to lead to poor choices as good ones. The relatively simple example of Section 3 and the student example in Section 5 show the need for nuance in scaling choices.

The only reliable approach to finding proper scales is to use mathematical arguments based on physical/biological intuition. Individual practitioners of mathematical modeling are very good at doing this on an ad hoc basis, but people new to the topic would benefit greatly from some guidelines.

Even after a good set of scales has been chosen for a model, it is still necessary to find good choices for the dimensionless parameters that will appear in the final version of the model. For physical science models, choosing dimensionless parameters is facilitated by the availability of standard parameters such as the Reynolds number and the Mach number. However, for population modeling there are few obvious choices for dimensionless parameters, the exception being the basic reproductive number, which is ubiquitous even among mathematical biologists who do not use scaling. Very little, if anything, has been said in print to guide the choice of dimensionless parameters.

1.1 Plan for the paper

The primary purposes of this paper are to develop a set of guidelines that generally leads to good choices for scales and dimensionless parameters in biological models and to offer suggestions for how scaling and asymptotics can be taught in courses at various levels. This is difficult to do without first seeing the benefits conferred by good choices of scales and dimensionless parameters. We therefore begin in Section 2 with an example of a well-scaled model that illustrates some of these benefits.

In Section 3 we show in some detail how the scaled model of Section 2 was derived, with some time spent examining the effect of less optimal scaling choices. In general, appropriate scaling choices are based either on scales that appear directly in a problem, such as an initial value for a decreasing quantity or a theoretical

maximum for an increasing quantity, or on dominant balance arguments, in which scale relations are derived by comparing two terms thought to be the most important ones in an equation.

Section 4 extends the discussion to the common case where critical processes occur on widely disparate time scales. In these cases, we need a separate scaling for each phase, and sometimes these require different scales for dynamic variables as well as time. In problems with more than one relevant time scale, small parameters that identify time scale ratios can be particularly useful.

We conclude in Section 5 with a discussion of how to incorporate scaling and asymptotics into courses in differential equations or mathematical biology.

2 Benefits of Scaling

As an example, consider a mathematical model for onchocerciasis (on-ko-ser-KĪ-a-sis), a vector-borne disease that is endemic in parts of tropical Africa. The disease is caused by the parasitic worm *Onchocerca volvulus*, which has a complicated life cycle split between a black fly host for the early larval stages and a human host for the later larval stages, the adults, and the microfilaria that develop into larvae after ingestion by the flies (Basáñez and Boussinesq 1999). The simplest reasonable model for onchocerciasis requires susceptible (S), exposed (E), and infective (I) classes among humans and uninfected (U) and infected (V) classes among the black fly vector. A separate exposed class is needed for the human host because it takes about a year for the larvae that infect humans to develop into the adults that make humans infective to flies. Adult worms live for about 10 years, so it takes about that long for infective humans to become uninfective, and then only if they are not reinfected while their adult worms age. The flies live for only about a month, so the disease has the unusual feature that the incubation period in humans is significantly longer than the lifespan of the vector. With no migration, no disease-induced mortality, and constant human and fly populations of N and F, respectively, we have the SEIS-UV model depicted in Figure 1. See Ledder et al (2017) for description and analysis of an onchocerciasis model that includes a treatment protocol to fight the disease.

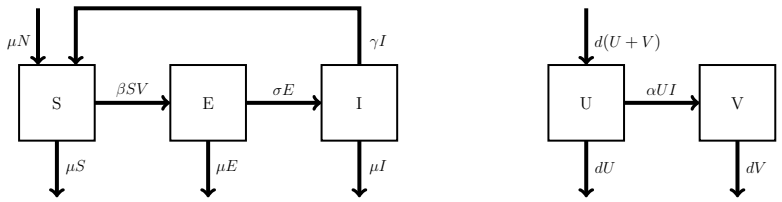

Fig. 1 A schematic representation of a constant population SEIS-UV model. Boxes indicate the populations in each class. Arrows indicate flows into, out of, and between the different classes, with labels giving the flow rates in people/time (left) and flies/time (right). The human population is assumed to be constant, so $N \equiv S + E + I$.

The corresponding system of equations is

$$V' = \alpha(F - V)I - dV, \tag{1}$$

$$E' = \beta SV - (\sigma + \mu)E, \tag{2}$$

$$I' = \sigma E - (\gamma + \mu)I, \tag{3}$$

$$S + E + I = N, \tag{4}$$

where all parameters are positive; d, σ, μ, and γ have dimension 1/time; α has dimension 1/(people-time); and β has dimension 1/(flies-time). We have not included the differential equations for U or S because the constant population assumption allows us to replace U in the V equation with $F - V$ and to replace one of the three human differential equations with (4).

The standard analysis of the model (1)–(4) involves finding non-trivial equilibrium solutions from the algebraic system

$$\alpha(F - V)I = dV, \quad \beta SV = (\sigma + \mu)E, \quad \sigma E = (\gamma + \mu)I, \quad S + E + I = N, \tag{5}$$

followed by the determination of stability of those equilibria along with the disease-free equilibrium solution $E = I = V = 0$, through finding eigenvalues or applying the Routh-Hurwitz conditions in 3 dimensions (see Ledder (2013), for example). Numerical simulations require estimates for the system parameters. Those that can be connected to expected duration of an individual species or status (d, μ, σ, γ) can be estimated with reasonable accuracy, while some of the other parameters (α, β, F) are much harder to estimate.

As an alternative to the analysis and numerical simulation of the model (1)–(4), we can instead analyze and simulate a dimensionless version of the model in which the new variables s, x, i, v correspond to the original variables S, E, I, V:

$$\delta v' = a(1-v)i - v, \tag{6}$$

$$\eta x' = bsv - x, \tag{7}$$

$$i' = x - i, \tag{8}$$

$$s + i + \eta(1+\epsilon)x = 1, \tag{9}$$

with equilibria determined by

$$a(1-v)i = v, \quad bsv = x = i, \quad 1 = s + i + \eta(1+\epsilon)x, \tag{10}$$

where

$$\delta = \frac{\gamma+\mu}{d}, \quad \epsilon = \frac{\mu}{\sigma}, \quad \eta = \frac{\gamma+\mu}{\sigma+\mu}, \quad a = \frac{\alpha N}{d}, \quad b = \frac{\beta F}{(\gamma+\mu)}\,\frac{\sigma}{(\sigma+\mu)}. \tag{11}$$

The definitions of the dimensionless variables and these particular choices for dimensionless parameters will be explained in Section 3.

2.1 Algebraic benefits

All of the terms in the original equations appear in the dimensionless ones, so the difference seems at first glance to be merely cosmetic. A deeper look shows that there is at least a minimal improvement in that the original 8 parameters have been reduced to 5, and a more significant improvement in that the system of 4 algebraic equations for the equilibria is immediately reduced to a system of just 3 equations. With careful reorganization, the equilibria can be determined from a single string of calculations:

$$1 + ai = aiv^{-1} = axv^{-1} = abs = ab - ab(1+\eta+\epsilon\eta)i,$$

with result

$$x = i = \frac{1-(ab)^{-1}}{1+(1+\epsilon)\eta+b^{-1}}, \quad \text{provided} \quad R_0 = ab > 1. \tag{12}$$

The corresponding result for the original system can of course be obtained through the same set of calculations, but with the significant cosmetic complications of the presence of a parameter in every term; thus, it is much harder in practice to obtain the dimensional result in a relatively simple form without working through the dimensionless form and translating back. Of course the equations can be solved using a computer algebra system, but it is unlikely the results will be put in the most convenient form, especially if, as in this case, we have not previously calculated the basic reproductive number R_0. At minimum, therefore, any nondimensionalization makes the algebraic part of the analysis easier.

Nondimensionalization is often advantageous in another way if producing meaningful simulations is part of the study goals. In this example, the parameters η, δ, and ϵ are known with reasonable accuracy because they are combinations of the four life cycle parameters listed above. In contrast, a and b are dimensionless counterparts for α and β, which makes them hard to calculate from their definitions. However, the equilibrium values of i and v represent the respective fractions of the human and fly populations that are infective; if these values can be estimated, then the equilibrium formulas can be used to recover indirect estimates of a and b. So nondimensionalization can make simulations more accurate by grouping parameters that are difficult to measure into combinations that are easier to determine than their individual components.

2.2 Analytical benefits

The algebraic differences already noted are minor compared to the enormous analytical differences between a model that has been properly scaled and one that has merely been nondimensionalized or is still in dimensional form. Variables in a properly-scaled dimensionless model have values that are neither very large nor very small, independent of the units that are used in the original model. This assumption, along with the more problematic assumption that the derivatives of the variables are also of unit order of magnitude, means that the relative importance of terms in an equation can be determined simply by examining the magnitudes of the dimensionless parameters. Given the definitions of the parameters, we know that both δ and ϵ are very small, roughly $\delta = 1/120$ and $\epsilon = 0.02$. Given that the model itself has uncertainty in the identification of relevant biological processes, the quantitative assumptions about those processes, and the values of the parameters, it is certainly reasonable to consider simplifying the model by setting these parameters to zero. Setting small parameters to zero under appropriate circumstances can simplify the analysis, and sometimes the model itself. The results of the simplified analysis can then be checked against numerical simulations; if the simplified results are not a good approximation of the full model, then the scaling choices should be rejected, while the scaling choices will have been supported if the simplified results are good. In either case, something has been gained from the asymptotic analysis of a scaled model.

Neglecting terms with small parameters reduces the scaled system (6)–(9) to

$$\eta x' = bsv - x, \tag{13}$$

$$i' = x - i, \tag{14}$$

$$s + i + \eta x = 1, \quad v = \frac{ai}{1 + ai}. \tag{15}$$

We could have replaced the original differential equation for the vector (6) with its quasi-steady approximation (15b) without having scaled the problem; however, that would have been based on intuition alone, whereas the quasi-steady approximation at this stage is based on mathematical reasoning.

The approximate system is only 2-dimensional rather than 3-dimensional, which makes the analysis much easier. In this case, the Jacobian is

$$J = \begin{bmatrix} -\eta^{-1} - bv & \eta^{-1} b \left(-v + \frac{as}{(1+ai)^2} \right) \\ 1 & -1 \end{bmatrix}.$$

Stability requires that the trace of J be negative, which is immediately true, and that the determinant be positive, where

$$\det J = (\eta^{-1} + 1)bv + \eta^{-1} \left[1 - \frac{abs}{(1 + ai)^2} \right].$$

For the disease-free equilibrium, the determinant reduces to $\eta^{-1}(1 - ab) = \eta^{-1}(1 - R_0)$, so we have stability when $R_0 < 1$. For the endemic disease equilibrium, $abs = 1 + ai$, so the factor in the square brackets is necessarily positive and hence the equilibrium is stable.

It is not that difficult to apply the 3-dimensional Routh-Hurwitz conditions to the original Jacobian, but the 2-dimensional case is certainly preferable, especially when working with students who have little experience with differential equations or linear algebra. The corresponding reduction from a 4- or 5-component system in a research problem to just 3 components is huge.

As noted earlier, we should use a numerical simulation to check for possible negative impacts of neglecting small terms in the dimensionless model. Figure 2 compares the results of the full model (6)–(9) and the simplified model (13)–(15) for a scenario in which a small population of human infectives $i_0 = 0.01$ is introduced into a system in which there are no infected humans or flies. The other parameters are

$$d = 12, \quad \sigma = 1, \quad \gamma = 0.08, \quad \mu = 0.02, \quad a = 1, \quad b = 4.$$

The first four values are from reasonable approximations of the relevant expected times, while the parameters a and b were chosen because they yield an equilibrium solution in which about 60% of the human population and 3/8 of the fly population are infective, these values being consistent with the data for the hardest hit areas in Africa.

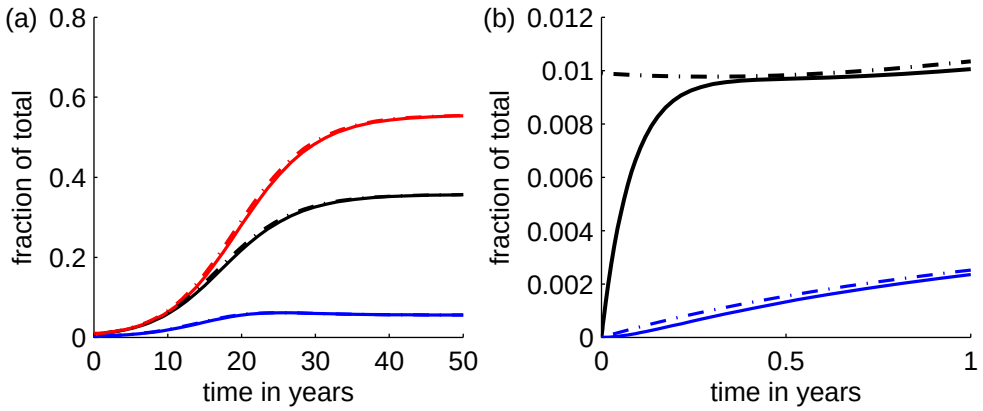

Fig. 2 Population fractions for an onchocerciasis simulation of a scenario in which a small population of infective humans is introduced into a system with no prior onchocerciasis presence, full model (solid) and simplified model (dash-dot): (a) E/N, V/F, I/N from bottom to top, (b) E/N, V/F, from bottom to top. It takes on the order of 20, 30, and 40 years, respectively, for the exposed human fraction, infective fly fraction, and infective human fraction to approach their equilibrium values. The simplified model closely matches the full model except for a very small initial transient in the infected fly fraction that resolves within about four months.

Figure 2a shows the results of a 50-year simulation. The error caused by the simplification is only just noticeable on a graph, but is far smaller than the uncertainty in the parameter values; if we were fitting the models to time series data, the Akaike Information Criterion (AIC) (see Ledder (2013), for example) would certainly favor the model with 3 parameters over the one with 5. Figure 2b highlights the only significant difference between the two models: the simplified model misses the initial transient in which the infected fly population changes rapidly from its starting value to the quasi-steady value given by (15b). This transient is resolved within about 4 months, which is not surprising given that it occurs on a time scale corresponding to the lifespan of the flies. The magnitude of the error is only 1% of the fly population at its worst, so it is clearly not a reason to reject the simplified model.

As happened here, it is often the case that initial transients in biological models are of no practical importance (see Huang et al (2015) for another example); however, it is important for anyone who uses asymptotic simplification to know how to modify the procedure to properly capture initial transients. This will be addressed in Section 4.

3 Making Good Scaling Choices

Replacing a dimensional model with a dimensionless one requires a number of choices. The process breaks down into two parts: choosing scales for the variables and choosing a set of dimensionless parameters. Considerable care must be put into these choices in order to obtain the benefits outlined in Section 2.

3.1 Choosing scales

A set of scales is judged acceptable if it leads to an asymptotic analysis that is internally consistent and also consistent with any known data. If there is more than one set of acceptable scales, there may still be one set that is preferable because of algebraic convenience. This is sometimes impossible to discover until after the analysis is finished, but at that point it is easy to redo the scaling and analysis. Nevertheless, there are some general principles that one can use to improve the chance of getting an acceptable set on the first try.

Sometimes suitable scales can be found directly using values that are inherent in the model as upper bounds. There is no doubt that the total fly population F is the correct scale for the infected fly population V. The total population N is clearly an upper bound for all of the human population classes and will be a suitable scale for any except those for which the associated population fraction remains small. For example, the infective population in a region could be as high as 60%, but that would still leave roughly 40% of the

population as susceptible. This is large enough for total population to remain a reasonable scale for the susceptible population. On the other hand, individuals transition from exposed to infective in about one year, so it is doubtful whether the exposed population is ever a significant fraction of the total. Hence, total population is probably not an appropriate scale for E.

The onchocerciasis model has four populations and time, so we need a total of five scales. It is not necessary that all of these be obtained directly. Instead, we generally need to use dominant balance arguments to obtain relations between scales. If we have a total of five scaling relations for the onchocerciasis model, inclusive of direct scales, then we have uniquely defined the five individual scales. As an example of a dominant balance argument, consider the equation (3) for the infective class. Individuals enter this class at rate σE and leave at rate $(\gamma + \mu)I$ (both in people per time). While there may be an initial period in which the population of infectives is changing rapidly, it is a reasonable assumption that the two rates will be comparable most of the time. Given this assumption, we can decide that their scales should be equal: that is, the reference values E_r and I_r should be connected by the scaling relation

$$\sigma E_r = (\gamma + \mu) I_r.$$

Given that we have already chosen $I_r = N$, this dominant balance relation serves to define the scale for E as $(\gamma + \mu)N/\sigma$.

While the onchocerciasis model has only one obvious scale for population, along with the not so obvious scale needed for E, it has a plethora of possible time scales, each taken from a rate constant or sum of rate constants:

1. $1/\mu$ is the life expectancy of a human (about 50 years in tropical Africa);
2. $1/\gamma$ is the expected time an infective individual takes to be cleared of adult worms (10–12 years);
3. $1/(\gamma + \mu)$ is the expected time an individual spends in the infective class (slightly less than $1/\gamma$)
4. $1/\sigma$ is the expected time between being infected (entering class E) and becoming infective (about 1 year);
5. $1/(\sigma + \mu)$ is the expected time an individual spends in the exposed class (slightly less than $1/\sigma$)
6. $1/d$ is the expected lifetime of a fly host (around 30 days);

Which of these we choose depends on what drives the system and what is the purpose of the analysis. Current treatment of onchocerciasis involves a medication that inhibits the production of microfilaria, thereby stopping the infection of flies by infective humans. If completely effective, it will still take about 12 years before the infection is cleared from all the infective humans. This suggests that the expected time spent as an infective $1/(\gamma + \mu)$ is the most important time scale in the scenario. If we were working from the point of view of the flies, then $1/d$ might be a better choice.

At this point, we need a brief discussion of notation. There are many style conventions in use for denoting dimensional and dimensionless counterparts of the variables. In fluid mechanics, it is common to use accent marks (stars or hats) to denote the original variables and then remove the accents for the dimensionless ones. This is not bad, but it makes the original equations hard to read. Worse yet is the opposite practice of putting the accents on the dimensionless quantities, in which case the authors find it necessary to suppress the accents in the subsequent analysis, so that the reader must recognize that the symbol "u" represents dimensional velocity in one equation and dimensionless velocity in another. I prefer a system in which dimensional and dimensionless quantities are distinguished by case. Since it is customary to use capital letters for the classes in epidemiology, I use T for the time and then use all lower case letters for dimensionless quantities. This does not always work perfectly; for example, no one with a background in mathematics can countenance using e as the dimensionless exposed population. In this case, I think of "xposed" as the dimensionless counterpart of "Exposed."

Having chosen the scales and the notational system, we now make the substitutions

$$V = Fv, \quad S = Ns, \quad I = Ni, \quad E = \frac{\gamma + \mu}{\sigma} Nx, \quad \frac{d}{dT} = (\gamma + \mu)\frac{d}{dt} \tag{16}$$

into the model (1)–(4). After removing obvious common factors, we have

$$(\gamma+\mu)\frac{dv}{dt} = \alpha N(1-v)i - dv, \tag{17}$$

$$(\gamma+\mu)\frac{(\gamma+\mu)}{\sigma}\frac{dx}{dt} = \beta F s v - (\sigma+\mu)\frac{(\gamma+\mu)}{\sigma}x, \tag{18}$$

$$\frac{di}{dt} = x - i, \tag{19}$$

$$s + \frac{\gamma+\mu}{\sigma}x + i = 1. \tag{20}$$

3.2 Choosing parameters

Our new system (17)–(20) has dimensionless variables, but it is still a dimensional model. The model formulation is not done until we have rewritten (17)–(18) so that all the terms are dimensionless; in so doing, we need to choose a set of dimensionless parameters from the combinations that emerge.

How we complete the nondimensionalization after replacing the dimensional variables with dimensionless ones makes a big difference in the ease of subsequent analysis and our ability to obtain biological insights from the model. In physical systems, there are standard predefined dimensionless parameters, such as the Reynolds number, that can safely be used without careful thought. Since most, if not all, written guides to scaling are done with physical systems in mind, there is generally no attention given to the choice of parameters. In biology, there are no standard dimensionless parameters that apply across the variations in dynamic models, so we must take as much care to choose parameters as we do to choose scales.

The naive approach to completing the nondimensionalization of (17)–(18) is to divide through by the coefficients of the derivatives, giving

$$\frac{dv}{dt} = \frac{\alpha N}{\gamma+\mu}(1-v)i - \frac{d}{\gamma+\mu}v,$$
$$\frac{dx}{dt} = \frac{\sigma\beta F}{(\gamma+\mu)^2}sv - \frac{\sigma+\mu}{\gamma+\mu}x.$$

We would then have 5 dimensionless parameters:

$$\frac{\alpha N}{\gamma+\mu}, \quad \frac{d}{\gamma+\mu} \quad \frac{\sigma\beta F}{(\gamma+\mu)^2}, \quad \frac{\sigma+\mu}{\gamma+\mu}, \quad \frac{\gamma+\mu}{\sigma}, \tag{21}$$

with the last of these coming from (20).

To see the flaws in this set of parameters, consider the advantages conferred by the parameter choices given in (11):

1. Only 4 of the 5 parameters appear in the equations for the equilibria.
2. Only 2 of the 5 parameters appear in the formula for the basic reproductive number.
3. The presence of two small parameters in convenient locations allows the full model of 3 differential equations and 5 parameters to be approximated by a simpler model with only 2 differential equations and 3 parameters.

In contrast, the choices given in (21) lead to equilibrium equations with 5 parameters, a basic reproductive number formula with 4 parameters, and no way to obtain a good approximate system with fewer than the original 3 differential equations with 5 parameters.

While there are no definitive rules for choosing dimensionless parameters, there are some guiding principles that almost always lead to good choices:

1. Dimensionless parameters that appear only on the left side of a differential equation are advantageous, as equilibrium solutions are independent of these parameters.
2. When possible, it is best to choose ratios of time scales for processes that have a clear biological connection.
3. Small parameters allow for a model to be simplified by an asymptotic approximation. (Large parameters are equally useful, but it is common to define these instead as reciprocals of small parameters.)

4. It is best to avoid pairs of parameters whose values differ only slightly (or whose product is close to 1).

There are three ways to complete the nondimensionalization of (17), depending on which of the three coefficients we divide the equation by. If we pick the term on the left, then we get two dimensionless parameters on the right side of the equation, violating principle 1. If we pick the first term on the right, then neither of the parameters we create will be ratios of time scales, violating principle 2. Dividing by the coefficient of the final term results in the parameters δ and a, satisfying both of these principles, with δ the ratio between expected lifespan of the fly host and expected duration of infectivity for a human host.

Although we did not use principle 3 deliberately, the parameter δ that arose from principles 1 and 2 is small, as recommended by principle 3. This was more than a happy accident. Each component in a population model has its own natural time scale, and in most systems there are components whose time scales are very different. Here there is a large difference between expected life of a fly and expected infectivity duration of a human. In infectious disease models, the expected duration of infectivity is usually much shorter than the expected life of the host (See Section 4). In ecology, the expected life of a prey animal is often much shorter than that of a predator, at least in cases where the predator is much larger than the prey. Usually when principles 1 and 2 are followed, the parameters that appear on the left sides of the equations are ratios of expected lifespans and/or times in category, so at least one will be large or small if any two components of the system have very different expected durations.

The best way to complete the nondimensionalization of (18) is also to divide by the last coefficient, leading to the time scale parameter η and the dimensionless infectivity parameter b. The choice for the parameter in (20) is less clear. It would be more algebraically convenient to use a single parameter $(\gamma + \mu)\sigma$ for the coefficient of x, and this would clearly be best in a more general setting. However, this would obscure the fact that (in this case, but perhaps not for a different SEIS-UV disease) this parameter and η are almost identical in magnitude. Following principle 4, we benefit slightly in the long run by using the small parameter $\epsilon = \mu/\sigma$ as the additional parameter, since it is eliminated in the simplification from (6)–(10) to (13)–(15). Alternatively, we could have defined $\eta = (\gamma + \mu)/\sigma$ and $\epsilon = \mu/(\sigma + \mu)$, in which case the parameter in (9) would have been η and the parameter in (7) would have been $\eta(1 - \epsilon)$. We would still have got the same simplified model (13)–(15). My preference for the choices made here is simply that $\sigma + \mu$ seems more directly relatable to $\gamma + \mu$ than is σ, thus making η as I have defined it slightly more biologically clear than the alternative choice would have been.

4 Models With More Than One Principal Time Scale

The onchocerciasis model has five time scales, from which there are arguably three principal ones: the expected lifespan of the flies and the expected times spent in the exposed and infective human classes. Only the last of these turned out to be important. However, models that have demographic effects as well as epidemiological effects often have more than one time scale that needs to be considered. As an example, consider the common case of an infectious disease in which recovery confers immunity, such as smallpox or measles. Given a focus on long-term effects, we make the simplifying assumption of a short incubation period, so that newly infected individuals move directly from the susceptible class to the infective class (SIR rather than SEIR). The model must include processes for natural birth and death as well as death from the disease.

There are several different ways to conceptualize an SIR scenario with demographics. The version depicted in Figure 3 is not the most common one, but the few differences from the standard formulation (Brauer 2008; Hethcote 2000) can be justified by the small but significant advantages they confer. First, it is more common to choose separate rate constants for recovery and disease-induced mortality; instead, the model of Figure 3 uses a single rate constant γ (1/time) for disease duration, as in the onchocerciasis model, along with an additional dimensionless parameter m that indicates the fraction of individuals who die. This is merely a notational difference, but it has the benefit of incorporating a well-scaled biologically-significant dimensionless parameter in the initial formulation. Second, the birth rate $B(N)$ is not given explicitly; instead, the model assumes that the total population is governed by the logistic growth equation in the absence of disease-induced mortality. Conservation of total population ($N = S + I + R$) must then be used to find the correct expression for the birth rate. The standard model can be recovered from our model by a specific choice of parameters, but the more general formulation allows us to investigate the biologically interesting question of what transient and permanent effects disease-induced mortality has on the overall demographics of a population.

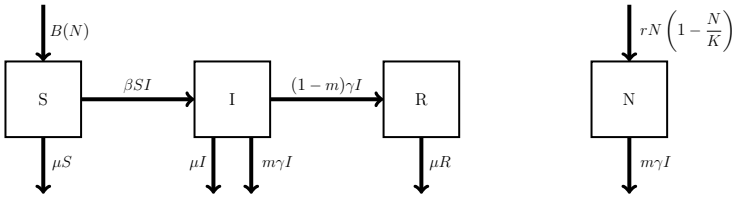

Fig. 3 A schematic representation of an SIR model with variable population and disease-induced mortality. The total population satisfies the logistic growth equation in the absence of disease-induced mortality. The birth rate $B(N)$ must be calculated from the population balance equation $N = S + I + R$.

Using T for dimensional time, we have the system

$$\frac{dS}{dT} = B(N) - \beta SI - \mu S, \tag{22}$$

$$\frac{dI}{dT} = \beta SI - (\gamma + \mu)I, \tag{23}$$

$$\frac{dR}{dT} = (1 - m)\gamma I - \mu R, \tag{24}$$

$$\frac{dN}{dT} = rN\left(1 - \frac{N}{K}\right) - m\gamma I, \tag{25}$$

where all parameters are positive; r, γ, and μ have dimension 1/time; K has dimension "people"; β has dimension 1/(people-time); and m is dimensionless.

The identity $N = S + I + R$ then establishes the correct expression for $B(N)$, leading to the susceptible equation

$$\frac{dS}{dT} = rN\left(1 - \frac{N}{K}\right) - \beta SI + \mu(N - S). \tag{26}$$

The final dimensional model consists of the equations (26), (23), and (25).

4.1 Identifying scales and parameters

As is typical in biological systems, each dynamic variable has its own time scale; here we have $1/r$, which represents the time scale of population growth, for S and N, $1/(\gamma + \mu)$ for I, and $1/\mu$, which represents expected lifespan, for R. We can also think of $1/(\beta K)$ as an additional time scale representing the infection process. The associated rate constants fall into two broad groups: γ and βK are fast rates associated with the disease process, while r and μ are slow rates associated with demographic changes. Any parameters formed as ratios of short times to long times will be suitable small parameters for asymptotics. It is generally easier to do the analysis if parameters are defined so that most are not small in this sense. Based on an understanding of the biological dynamics, we can identify $1/(\gamma + \mu)$, representing the expected time spent as infective, as the most important short time scale and $1/r$, representing the time scale of demographic change, as the most important long time scale. We will consider these two choices for time scale separately, but with a common set of parameters:

$$\epsilon = \frac{r}{\gamma + \mu} \ll 1, \quad b = \frac{\beta K}{\gamma + \mu}, \quad \phi = \frac{\mu}{r}, \tag{27}$$

where the notation "$\ll 1$" is used to indicate a parameter that can be assumed to be asymptotically small. There is another good reason for choosing b as one of the parameters. One infective produces new infectives in a susceptible population at the mean rate βK over a time interval of $1/(\gamma+\mu)$; thus, b is the basic reproductive number. Finally, ϕ has a simple biological interpretation as the ratio of the population growth time scale $1/r$ to the life expectancy time scale $1/\mu$.

4.2 The disease duration scaling

Since we will have two versions of the scaled model, we follow the standard convention in asymptotics of using τ for the faster time scale and t for the slower one. Assuming K is the appropriate population scale for each class, the short-term scaling is then defined by the substitutions

$$S = Ks, \quad I = Ki, \quad N = Kn, \quad \frac{d}{dT} = (\gamma + \mu)\frac{d}{d\tau}, \tag{28}$$

resulting in the system

$$\begin{aligned}
(\gamma + \mu)\frac{ds}{d\tau} &= rn(1 - n) - \beta Ksi + \mu(n - s),\\
(\gamma + \mu)\frac{di}{d\tau} &= \beta Ksi - (\gamma + \mu)i,\\
(\gamma + \mu)\frac{dn}{d\tau} &= rn(1 - n) - m\gamma i.
\end{aligned}$$

With this short-time scaling, we cannot get small parameters on the left sides of any of the differential equations, so there is nothing better than to divide through by $\gamma + \mu$ in each equation, yielding

$$\frac{ds}{d\tau} = \epsilon n(1 - n) - bsi + \epsilon\phi(n - s), \tag{29}$$

$$\frac{di}{d\tau} = bsi - i, \tag{30}$$

$$\frac{dn}{d\tau} = \epsilon n(1 - n) - (1 - \epsilon\phi)mi. \tag{31}$$

Setting $\epsilon = 0$ in this system does not eliminate any differential equations, but it does decouple the total population equation from the others, leaving the standard SIR model without demographics:

$$\frac{ds}{d\tau} = -bsi, \tag{32}$$

$$\frac{di}{d\tau} = bsi - i. \tag{33}$$

This system is usually derived from a model in which the demographic terms have been omitted from the start, but we will get much more information through having derived it from a more general system that also yields a long-time approximation.

4.3 The demographic scaling

Neglecting the order ϵ terms in the scaled model changes the long-term behavior of the system, so equilibrium analysis of the full problem requires us to rewrite it on a demographic time scale. The obvious way to do this is to keep the same scales for the populations while using the dimensionless time defined by $\frac{d}{dT} = r\frac{d}{dt}$. This yields the equations

$$\begin{aligned}
r\frac{ds}{d\tau} &= rn(1 - n) - \beta Ksi + \mu(n - s),\\
r\frac{di}{d\tau} &= \beta Ksi - (\gamma + \mu)i,\\
r\frac{dn}{d\tau} &= rn(1 - n) - m\gamma i.
\end{aligned}$$

There are several ways to work out the correct dimensionless problem from here. Given that we want to compare results from the two scalings, one simple starting point is to divide all three equations by $\gamma + \mu$, just

as we did in working on the short time scale. The difference is that we will now have factors of ϵ in front of each derivative:

$$\epsilon\frac{ds}{dt} = \epsilon n(1-n) - bsi + \epsilon\phi(n-s), \tag{34}$$

$$\epsilon\frac{di}{dt} = bsi - i, \tag{35}$$

$$\epsilon\frac{dn}{dt} = \epsilon n(1-n) - (1-\epsilon\phi)mi. \tag{36}$$

Whenever assessing the correctness of a proposed scaling, it is important to make sure that removing terms with small parameters yields a meaningful approximation. Equation (35) simplifies to a quasi-steady approximation, which is fine. Equations (34) and (36) are problematic, as each has only one term that is not marked by a small parameter. This is always unacceptable. In (36), for example, the $\epsilon = 0$ approximation is $i = 0$, which contradicts the whole idea of approximation: the most important term in an equation cannot be 0! Equation (34) has a similar problem.

The contradiction inherent in the $\epsilon = 0$ approximation is proof that our scaling failed. At least two terms in each equation have to matter, and this is only possible if we drop the assumption that the size of each term can be determined merely by looking at the small parameters. Although it is not obvious, the inescapable conclusion is that i should be of order ϵ. In deriving the long-term scaled model, we need to use a different scaling for I:

$$I = \epsilon K y, \tag{37}$$

where y, rather than i, is order 1. This results in the system

$$\frac{ds}{dt} = n(1-n) - bsy + \phi(n-s), \tag{38}$$

$$\epsilon\frac{dy}{dt} = bsy - y, \tag{39}$$

$$\frac{dn}{dt} = n(1-n) - (1-\epsilon\phi)my. \tag{40}$$

This system makes good biological sense. There is a small parameter in front of the equation for the infectives but not for the other two; this is because I changes on the τ time scale while S and N change on the demographic scale.

4.4 Model analysis

Our SIR example has two key time scales, one for the initial transient and one for the approach to demographic equilibrium. Problems that feature limit cycles rather than stable equilibria can have more than two distinct phases.[1]

The short time system:(32)–(33) Linearized stability analysis does not work for this system because every point $(s, 0)$ is an equilibrium; however, we can write the problem as

$$\frac{di}{ds} = \frac{1-bs}{bs}, \quad (s,i)(t_0) = (1-i_0, i_0), \tag{41}$$

which defines solution curves in the si-plane with initial populations of i_0 infectives and $1-i_0$ susceptibles. Integration of the differential equation from initial conditions to terminal conditions ($s = s_\infty$, $i = 0$) yields the implicit relation

$$s_\infty - b^{-1}\ln s_\infty = 1 - b^{-1}\ln(1-i_0)$$

for the final susceptible population s_∞ in terms of the basic reproductive number b. Given $i_0 \ll 1$, we can approximate this result as

$$s_\infty - b^{-1}\ln s_\infty = 1 + b^{-1}i_0; \tag{42}$$

[1] See Ledder (2007) for an example of a limit cycle that requires separate scalings for 5 different phases.

s_∞, of course, is the fraction of the population that do not get the disease.

Two other quantities of interest are the maximum infective population, which is determined from (41) with terminal condition $(s, i) = (b^{-1}, i_{\max})$, and the final total population, which is determined by noting that $n - m(s + i)$ is conserved, with value $1 - m$ in the limit $\epsilon \to 0$. Thus,

$$i_{\max} = 1 - b^{-1} + b^{-1} \ln(b^{-1}) + b^{-1} i_0, \qquad n_\infty = 1 - m + m s_\infty. \tag{43}$$

The long time system:(38)–(40) We cannot use a quasi-steady approximation for the y equation because setting $\epsilon = 0$ does not reduce it to an algebraic equation that can be solved for y. Nevertheless, we can set $\epsilon = 0$ in the n equation to obtain the simplified endemic disease equilibrium equations $bs = 1$ and

$$my = n(1 - n), \qquad y = n(1 - n) + \phi(n - b^{-1}), \tag{44}$$

and the Jacobian

$$J = \begin{bmatrix} -\phi - by & -bs & -(2n - 1 - \phi) \\ \epsilon^{-1} by & \epsilon^{-1}(bs - 1) & 0 \\ 0 & -m & -(2n - 1) \end{bmatrix}.$$

For the disease-free equilibrium ($s = 1$, $y = 0$, $n = 1$), we can immediately see that the eigenvalues are -1, $-\phi$, and $\epsilon^{-1}(b - 1)$; hence, we have stability when the basic reproductive number b is less than 1. It is best to postpone solution of (44); however, we can simplify the Jacobian with $bs = 1$, $\phi + by = bn(1 + \phi - n)$, yielding

$$J = \begin{bmatrix} -bn(1 + \phi - n) & -1 & -(2n - 1 - \phi) \\ \epsilon^{-1} by & 0 & 0 \\ 0 & -m & -(2n - 1) \end{bmatrix}. \tag{45}$$

Stability is most easily determined by the Routh-Hurwitz criteria (Ledder 2013)

$$c_1 > 0, \quad c_3 > 0, \quad c_1 c_2 > c_3,$$

where

$$c_1 = bn(1 + \phi - n) + (2n - 1) > bs(1 + \phi - n) + (2n - 1) = \phi + n > 0, \tag{46}$$

$$c_2 = \epsilon^{-1} by + bn(1 + \phi - n)(2n - 1) \sim \epsilon^{-1} by, \tag{47}$$

$$c_3 = \epsilon^{-1} by[m\phi + (1 - m)(2n - 1)], \tag{48}$$

and we have used the asymptotic assumption of small ϵ to simplify c_2. The requirement $c_1 c_2 > c_3$ follows from

$$m\phi + (1 - m)(2n - 1) < m\phi + (1 - m)n < \phi + n < c_1,$$

leaving only the requirement $c_3 > 0$. Rather than calculating n to demonstrate that this last requirement is satisfied, we can eliminate y from (44) and rearrange to get

$$m\phi = (1 - m)(1 - n) \cdot \frac{n}{n - b^{-1}} > (1 - m)(1 - n);$$

hence, $c_3 > \epsilon^{-1}(1 - m) byn > 0$. Finally, the stable equilibrium population, from (44), is

$$n = q + \sqrt{q^2 - 2qb^{-1} + b^{-1}}, \qquad q \equiv \frac{1 - m - \phi m}{2(1 - m)}. \tag{49}$$

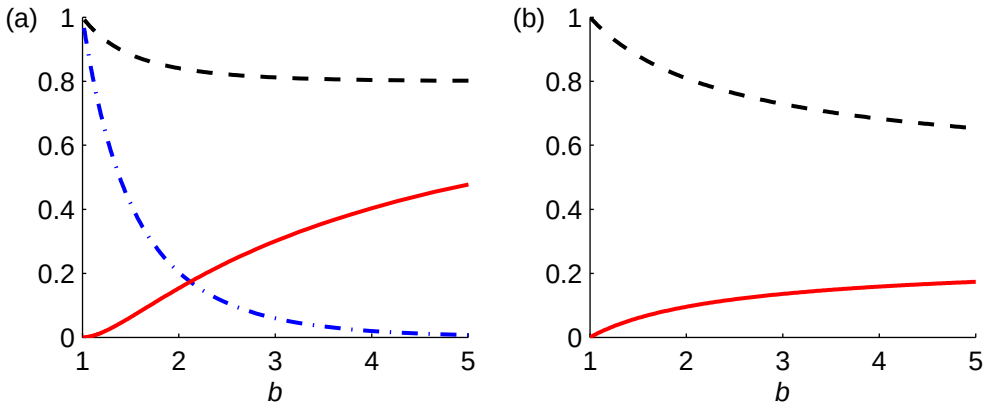

Fig. 4 Analytical results for the SIR model in terms of the basic reproductive number b with $\phi = 2$, $m = 0.2$, and $\epsilon = 0.001$: (a) The short-time model: uninfected population s_∞ (dash-dot), maximum simultaneous infective population $i_{\max}$ (solid), and final surviving population n_∞ (dashed); (b) The endemic equilibrium for the long term model: percentage infective $100I/N$ (solid), and population relative to carrying capacity N/K (dashed). Given an endemic disease ($b > 1$), the effect of the basic reproductive number is more pronounced at the short time-scale, with the fraction simultaneously infected rising to half of the population and the fraction that remains uninfected decreasing rapidly.

The combined system Some important biological implications of the model are apparent merely from doing the two scalings correctly. Given an initial population that is almost entirely susceptible and a basic reproductive number somewhat larger than 1, the number of infectives in the short term is a significant fraction of the total population. In the long term, the number of infectives is only a small fraction of the total population, regardless of how large the basic reproductive number is (within practical limits).

The analytical results, shown in Figure 4, bear out the conclusions obtained just from scaling for a case with common parameter values of $\phi = 2$, $m = 0.2$, and $\epsilon = 0.001$, corresponding to a pre-industrial population with average life expectancy 20 years and minimum doubling time 40 years, and a disease with a duration of 2 weeks and a mortality of 20%. Given a modestly large basic reproductive number and a starting population that is almost wholly susceptible, nearly the entire population gets the disease in the short term, with m approximately the fraction who die. At peak impact there could be 40–50% of the population who are simultaneously infective. The demographic parameter ϕ does not impact these results. In the long term, the population is depressed by a larger amount because the population growth parameter r is only half of the natural death rate parameter μ, but the endemic infective population is less than 0.2% of the total.

The analytical results for the short-term and long-term models do not tell the whole story of the model behavior. Full solutions of short-term and long-term asymptotic model approximations can be matched together to get a uniformly valid solution (Bush 1992; Fowler 1997; Hinch 1991; Holmes 2013; Howison 2005; Logan 2013; Murdock 1991). That is not possible for our model because we have only partial results for the simplified models. In particular, we do not have analytical results for the time between the resolution of the short-term behavior and the approach to long-term equilibrium. The behavior of this portion of the scenario must be determined through numerical simulations. This is almost always the case for models in ecology and epidemiology.

Figure 5 shows the results of simulations using the parameters from Figure 4 along with basic reproductive numbers of 2 and 4. The analytical results for both time scales are visible in the plots. The short-term plots show the terminal behavior of the simplified short-term model, but this behavior is not sustained as the time increases beyond the short term. The long-term plots show that the approach to equilibrium is achieved gradually, with a sequence of outbreaks of diminishing amplitude. The medium-term plots show that the severity of subsequent outbreaks is far less than that of the initial one; in particular, the short-term scaling for infectives $I = Ki$ is only appropriate for the initial outbreak. In biological terms, outbreaks after the initial one occur as soon as the susceptible population is large enough for a significant increase in new infections, which is well before the susceptible population could rise to a level that would support an intense outbreak. Comparison of the results for $b = 2$ and $b = 4$ shows the influence of the basic reproductive number: larger

values cause a significant decrease in the susceptible fractions with modest changes in the other variables, along with shorter intervals between outbreaks and a faster approach to equilibrium.

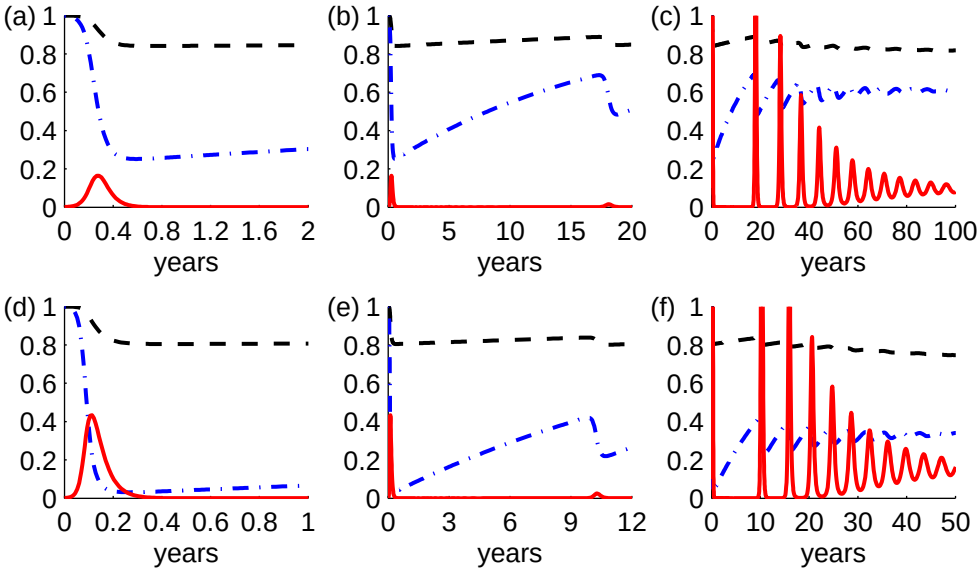

Fig. 5 Population fractions for an SIR model simulation of a scenario in which the disease is introduced into a previously uninfected population, using relatively low (top row) and high basic reproductive numbers of $b = 2$ and $b = 4$, with $\phi = 2$, $m = 0.2$, and $\epsilon = 0.001$; total population relative to carrying capacity (n – dashed), susceptible fraction (s/n – dash-dot), infective fraction (i/n in (a), (b), (d), (e) – solid), percent infected ($100i/n$ in (c), (f) – solid).

5 Teaching Asymptotics and Scaling

The rudiments of asymptotics can be taught in any differential equations or mathematical biology course. There are just three key principles for this most elementary level.

1. When the coefficients of the highest order derivatives in all equations are not small, the problem is *regular*. In such cases, neglecting terms with small parameters should yield an approximate solution that is reasonably good for all time.
2. When one or more differential equations has a small parameter in front of the highest order derivative, the problem is *singular*. In such cases, neglecting terms with small parameters can still yield an approximate solution that is good for long times; however, there will usually be significant initial transients. Numerical simulations, such as those illustrated in Figures 2 and 5, can determine whether the initial transient is important or not.
3. The first two principles assume that the magnitudes of terms can be determined from the parameters in those terms. One must always be careful to justify that neglected terms actually are small. If not, then one must reconsider the scaling choices.

The original onchocerciasis model (6)–(9) is singular, while the simplified version (13)–(15) is regular. The regular approximation misses the initial transient of the singular original, but that initial transient is not biologically relevant. The model in Huang et al (2015) has these features as well.

Despite its usefulness, scaling is generally taught only in courses devoted to a combination of mathematical modeling and asymptotics. Not only are such courses uncommon in the United States, but also students who take them do not necessarily get a solid foundation in scaling; as noted earlier, most texts for these courses offer them little guidance. The only feasible option in the absence of good coverage in the course text is to supplement with principle-based scaling such as has been offered here.

There is no need to defer scaling to advanced courses. The basic principles can be taught in any course that includes algebraic or differential equation models. There is no particular mathematical background required beyond high school algebra and the rudiments of differential calculus. However, the need for nuanced physical arguments means that students must have a deep understanding of the mathematical models they are working with (which they also need to have in order to make sense of the mathematical results). This does not necessarily mean that they have to be able to derive the models themselves, but they should be able to identify the physical significance of each term in each equation; for example, they should be able to produce the differential equations corresponding to Figure 1 and identify each term with the appropriate biological process. Of course the instructor has to be careful with the presentation, as many beginning students find any non-mechanical mathematics to be difficult. The instructor also needs to understand that scaling is nuanced and that students will often produce work that is only partially successful even though it is based on sound principles.

5.1 An Instructive Example

I gave my mathematical modeling class an exam question on an SIR epidemic model with demographics but no disease mortality (the special case of (22)–(25) with $m = 0$ and $N = K$—see (52) below). I first asked them to construct the model from a list of assumptions, calculate the basic reproductive number ($b = \beta N/(\gamma + \mu)$) from first principles, and identify at least two time scales, classifying them as long or short and providing a verbal explanation of their biological meaning. Most students got this far, correctly identifying $1/\mu$ as a long time scale and $1/(\gamma + \mu)$ as a short time scale. Because they constructed the model from a narrative, some wrote the sum as $\mu + \gamma$, which is of course correct but which puts the smaller rate first and makes it more difficult to identify the sum as a large rate. A few were able to use the basic reproductive number to identify βN as an additional short time scale, which we had not specifically done before.

The remainder of the problem asked them to scale the model using a long time scale. Given that they already knew the basic reproductive number and the only long time scale, their task essentially came down to identifying two scaling principles from which to obtain the scales for S and I and identifying the second dimensionless parameter. Generally students grab the first ideas that come to mind, but it would be better to first identify all reasonable options. In this case, there are four possible scaling relations:

1. Use N as the scale for S.
2. Use N as the scale for I.
3. Use a dominant balance argument in the I equation to obtain a scale for S $[(\gamma + \mu)/\beta]$.
4. Use a dominant balance in the S equation to obtain a scale for I $[\mu/\beta]$.

The additional parameter is some ratio of rate parameters μ, γ, and possibly one of the rates $\gamma + \mu$ and βN that combine to make b.

Choosing N as the scale for I changes the S equation to

$$\mu \frac{ds}{dt} = \mu(1 - s) - \beta N s i.$$

Dividing by μ then puts the parameter $\beta N/\mu$ in front of the last term. It might seem that we can just make this the second parameter, but this parameter is large because βN is a fast rate and μ a small one. Choosing N as the reference for I fails because it leads to an equation in which one term appears to be more important than all the others, leading to the paradoxical result that the dominant term is approximately 0. Thus, the only reasonable choice for I is the one that comes from the balance of the S equation: μ/β.

One of the students used the balance of the I equation to choose the scale for S. This is alright if $b > 1$, but it is problematic if $b < 1$ because in that case the scale for S is larger than the total population. Even if $b > 1$, this choice is less convenient than N, as the S equation becomes

$$\frac{ds}{dt} = b - s - s i.$$

This equation is misleading in that the equilibrium for s when the disease is absent is $s = b$; this seems to suggest that the population of susceptibles depends on the basic reproductive number even when the disease

is absent. Of course this statement is not true biologically; rather, it reflects the choice of a scale that assumes the disease to be present.

With the preferred reference quantity of N for S (and any reference quantity for I), the I equation becomes

$$\mu\frac{di}{dt} = \beta N si - (\gamma + \mu)i.$$

Dividing by $\gamma + \mu$ and taking $\epsilon = \mu/(\gamma + \mu)$ yields the final result

$$\epsilon\frac{di}{dt} = bsi - i.$$

This version correctly indicates that the dynamics of I are fast compared to S and that the equilibrium susceptible population is $s = 1/b$, provided $b > 1$.

One student had written the original differential equation as

$$\frac{dI}{dT} = \beta SI - \mu I - \gamma I.$$

This is correct, of course, but failing to combine the last two terms into one led the student to divide by γ rather than $\gamma + \mu$, resulting in the equation

$$\frac{\mu}{\gamma}\frac{di}{dt} = \frac{\beta N}{\gamma}si - i - \frac{\mu}{\gamma}i.$$

He then defined parameters $\epsilon = \mu/\gamma$ and $b = \beta N/\gamma$. The resulting equation,

$$\epsilon\frac{di}{dt} = bsi - i - \epsilon i,$$

meets all the necessary requirements for scaling. However, the parameter b in this case is not the basic reproductive number, which works out to be $b/(1 + \epsilon)$. It is much more difficult to understand the biological meaning of the results when the dimensionless parameters that are chosen are not the ones with the most biological significance. While the student's answer is not wrong, it is far inferior to the choice one gets by setting b to be the basic reproductive number and $\epsilon = \mu/(\gamma + \mu)$.[2]

The examples I've cited fall short of high-quality work. What can we do to help our students master scaling?

5.1.1 Take small steps

The first step in learning scaling is to acquire conceptual understanding and technical mastery of nondimensionalization. For example, consider the Michaelis-Menten reaction velocity equation

$$V(S) = \frac{V_{max}S}{K_m + S}, \tag{50}$$

where S is a substrate concentration in mass/volume or moles/volume, V is the rate of substrate reaction, often called the "reaction velocity," in units of concentration/time, V_{max} is a theoretical maximum reaction velocity (corresponding to the limit $S \to \infty$), and K_m is the value of substrate concentration for which the reaction velocity is half of its maximum. We can start by asking students to prepare graphs of this function with different values of the parameters. That done, we can ask them to graph V/V_{max} vs S/K for each of their examples, in order to bring home the point that all the graphs are identical when done with dimensionless variables (see Ledder (2013), Section 2.5). The alternative approach of recasting the equation in terms of dimensionless variables $v = V/V_{max}$ and $s = S/K$ results in the same graph, but much more efficiently. Similar examples can be constructed to reduce a 2-variable, 3-parameter dimensional model to a 1-parameter dimensionless model using obvious choices for the scales and whatever dimensionless parameter emerges.

A next step is to give students several possible nondimensionalizations for a simple model and ask them to determine which are correct under what circumstances and to discuss the relative advantages

[2] I awarded this student almost all the credit for the problem.

for circumstances with more than one correct nondimensionalization. The final paragraph of Section 3 is an example of such a discussion.

Once students are ready to start making their own attempts at scaling, it is important to give them graded examples, building up complicated models a little at a time. In epidemiology, we can start with the standard short-term SIR model,

$$\frac{dS}{dT} = -\beta SI, \qquad \frac{dI}{dT} = \beta SI - \gamma I, \qquad R = N - S - I \tag{51}$$

and then add constant-population demographics to get

$$\frac{dS}{dT} = \mu N - \beta SI - \mu S, \qquad \frac{dI}{dT} = \beta SI - \gamma I - \mu I, \qquad R = N - S - I \tag{52}$$

before graduating to the variable population model of Section 4.[3] Similarly, the onchocerciasis model of Sections 2 and 3 can be preceded by a standard short-term SIS model and then an SEIS model. Many other models of intermediate difficulty can be found in Chapters 5 and 7 of Ledder (2013), including a variable N SIS model for a population of laboratory mice in which newcomers are added to the population at a constant rate.

5.1.2 Identify options before making choices

A recurrent theme in my examples is my practice of identifying alternatives before making choices. This is not the way students have learned to do mathematics; they are used to working problems by trying the first thing that comes to mind and only trying a second idea if the first one doesn't work. We can encourage this new kind of thinking by preceding a question that asks for a choice of scales with a question that asks students to identify possible choices of direct scales and scaling relations derived from dominant balance arguments, as in my list of four options in the exam problem. The students should annotate their list with a statement giving a possible biological justification, such as "N is an upper bound for S" or "balancing the two terms on the right side of the I equation assumes that I is near equilibrium."

5.1.3 Cultivating a good attitude toward scaling

Many problems in mathematics have one right answer—for example, an equilibrium solution is either asymptotically stable, neutrally stable, or unstable for a given set of parameter values. Others have multiple right answers, such as how to prove a theorem. Scaling is of the latter type; as in proofs of theorems, the different correct answers may be unequal in value. A scaling is correct if it accurately captures orders of magnitude in some time and space regime in the limit as any small parameters approach zero. Correct scalings may be "better" or "worse" based on algebraic or biological reasons. Some of these reasons emerge in the analysis, so it is common to make minor changes, particularly in the choice of dimensionless parameters, as the analysis proceeds. One should consider the form of the dimensionless model to be provisional until the characterization of the model is complete. This goes against the grain for students, who are used to thinking of problems in terms of components solved in a linear order. In modeling, scaling must be done early, but it should be revisited after each subsequent step.

5.1.4 Some additional suggestions and a final word

In addition to the general suggestions in Sections 3 and 4 for how to identify scales, there are some techniques that can be used for specific problems. If typical values for all of the dimensional parameters are known, a computer simulation can be used to generate sample graphs of the solutions. These can be very helpful in checking possible reference values. For example, any simulation graph of the onchocerciasis model will show that E is always much less than N, providing evidence that a smaller scale is needed for E. Students should be encouraged to run simulations if reasonable parameter values can be given to them or found from available literature.

[3] Hethcote (2000) is a good source for both of these models; interestingly, the author nondimensionalizes the dependent variables but retains dimensional time. Brauer (2008) has only the first of these two models but features a nice explanation of the "natural decay" assumption used for the term $-\gamma I$.

If there is a clear choice for a dimensionless parameter, that choice can be very helpful in identifying good scaling choices. In Section 4, we used a biological argument to justify the preference for $1/(\gamma+\mu)$ as the long time scale rather than $1/\gamma$. Instead, we could have found the basic reproductive number first and used its form as an argument for choosing $1/(\gamma+\mu)$.

It can be helpful to note that dimensionless parameters can always be thought of as ratios of potential scales for a variable. If we parse b as $(\beta K)/(\gamma+\mu)$, we can identify $1/(\beta K)$ as a time scale we had not previously identified. But we could also parse b as $K/[(\gamma+\mu)/\beta]$, which makes it a ratio of two population scales. We can ask students if they can think of a scenario in which $(\gamma+\mu)/\beta$ would be a good population scale. The answer is "no" because b can never be very small or very large for a realistic scenario; thus, the scale $(\gamma+\mu)/\beta$ is *asymptotically* equivalent to the scale K. Clearly the latter is the better choice on algebraic grounds. But *small* parameters do suggest alternative scales that are asymptotically distinct. Given $\epsilon = r/(\gamma+\mu)$ as a small parameter, there clearly is a meaningful difference between the scaling choices of $1/r$ and $1/(\gamma+\mu)$, which we used for the slow and fast scales in the SIR model, respectively.

A particularly good example of scaling done by students appears in Flake et al (2003), work done as part of an REU project.

References

1. Basáñez, M. & Boussinesq, M. (1999). Population biology of human onchocerciasis. *Philosophical Transactions: Biological Sciences*, **1384**, p809.
2. Banasiak, J. & Lachowicz, M. (2014). *Methods of Small Parameter in Mathematical Biology*, Birkhauser.
3. Brauer, F. (2008). Compartmental models in epidemiology, in *Mathematical Epidemiology*, ed. Brauer, F., van den Dreissche, P. & Wu, J., Springer.
4. Bush, A.W. (1992). *Perturbation Methods for Engineers and Scientists*, CRC Press.
5. Flake, C., Hoang, T., & Perrigo, E. (2003). A predator-prey model with disease dynamics, *Rose-Hulman Undergraduate Math Journal*, **4**, issue 1
6. Fowler, A.C. (1997). *Mathematical Models in the Applied Sciences*, Cambridge University Press.
7. Hethcote, H.W. (2000). The mathematics of infectious diseases. *SIAM Review*, **42**, pp 599–653.
8. Hinch, E.J. (1991). *Perturbation Methods*, Cambridge University Press.
9. Holmes, M.H. (2013) *Introduction to Perturbation Methods*, Springer.
10. Howison, S. (2005). *Practical Applied Mathematics: Modelling, Analysis, Approximation*, Cambridge University Press.
11. Huang, Q., Wang, H., & Lewis, M.A. (2015). The impact of environmental toxins on predator-prey dynamics. *J Theo. Bio.*, **378**, pp 12–30.
12. Ledder, G. (2007). Forest defoliation scenarios. *Math. Biosci. Eng.*, **4**, pp 15–28.
13. Ledder, G. (2013). *Mathematics for the Life Sciences: Calculus, Modeling, Probability, and Dynamical Systems*, Springer.
14. Ledder, G., Sylvester, D., Bouchat, R. & Thiel, J. (2017). Continuous and pulsed epidemiological models for onchocerciasis with implications for eradication strategy, submitted to *Math. Biosci. Eng.*
15. Lin, C.C. & Segel, L.A. (1988). *Mathematics Applied to Deterministic Problems in the Natural Sciences*, SIAM.
16. Logan, J.D. (2013). *Methods of Applied Mathematics*, 4th edition, Wiley.
17. Murdock, J.A. (1991). *Perturbations: Theory and Methods*, Wiley.
18. Segel, L.A. (1972). Simplification and scaling. *SIAM Review*, **14**, pp 547–571.
19. Thieme, H.R. (2003). *Mathematics in Population Biology*, Princeton University Press.