

When can we infer mechanism from parasite aggregation? A constraint-based approach to disease ecology

Mark Q. Wilber^{1,*}

Pieter T. J. Johnson²

Cheryl J. Briggs¹

October 5, 2016

1. University of California, Santa Barbara, Santa Barbara, CA, 93106;

2. University of Colorado, Boulder, CO 80309;

* Corresponding author; e-mail: mark.wilber@lifesci.ucsb.edu

Keywords: macroparasites, feasible sets, maximum entropy, trematodes, amphibians, negative binomial, geometric, host-heterogeneity, parasite-induced host mortality

Abstract

Few hosts have many parasites while many hosts have few parasites - this axiom of macroparasite aggregation is so pervasive it is considered a general law in disease ecology, with important implications for the dynamics of host-parasite systems. Because of these dynamical implications, a significant amount of work has explored both the various mechanisms leading to parasite aggregation patterns and how to infer mechanism from these patterns. However, as many disease mechanisms can produce similar aggregation patterns, it is not clear whether aggregation itself provides any additional information about mechanism. Here we apply a “constraint-based” approach developed in macroecology that allows us to explore whether parasite aggregation contains any additional information beyond what is provided by mean parasite load. We tested two constraint-based null models, both of which were constrained on the total number of parasites P and hosts H found in a sample, using data from 842 observed amphibian host-trematode parasite distributions. We found that constraint-based models captured $\sim 85\%$ of the observed variation in host-parasite distributions, suggesting that the constraints P and H contain much of the information about the shape of the host-parasite distribution. However, we also found that extending the constraint-based null models can identify the potential role of known aggregating mechanisms (such as host-heterogeneity) and disaggregating mechanisms (such as parasite-induced host mortality) in constraining host-parasite distributions. Thus, by providing robust null models, constraint-based approaches can help guide investigations aimed at detecting biological processes that directly affect parasite aggregation above and beyond those that indirectly affect aggregation through P and H .

Introduction

Disease ecology has traditionally emphasized mechanistic descriptions of infection patterns (Anderson and May 1978, Duerr et al. 2003, Poulin 2007). One particular pattern observed in macroparasites, such as parasitic helminths and arthropods that do not directly reproduce within their host (Anderson and May 1979), is that many hosts in a population tend to have few para-

27 sites and a few hosts tend to have many. In statistical parlance this means that parasites tend to
be aggregated within their hosts. This pattern is so ubiquitous in parasites that it has been called
one of the few general laws in disease ecology (Poulin 2007).

30 Canonical models of host-macroparasite dynamics have illustrated that a balance between
parasite pathogenicity and parasite aggregation plays an important role in the ability of a para-
site to regulate a host population (Anderson and May 1978, Tompkins et al. 2002). In general, the
33 stability of a host-parasite system and the regulation of a host population by parasites requires
some level of parasite aggregation and that parasite pathogenicity is not too high (Anderson
and May 1978). Because of the importance of parasite aggregation, much empirical and the-
36 oretical work has sought to understand both the mechanisms that can lead to aggregation in
host-macroparasite systems (Anderson and Gordon 1982, Wilson et al. 2002, Raffel et al. 2011,
Gourbière et al. 2015), and how to infer the dominant mechanisms structuring a host-parasite
39 system from observed aggregation patterns (Duerr et al. 2003, Grear and Hudson 2011, Wilber
et al. 2016).

Traditionally, studies of macroparasite aggregation have relied on a process-based approach
42 where various aggregating and disaggregating mechanisms are sequentially incorporated into
unaggregated null models until observed levels of aggregation are obtained (Anderson and Gor-
don 1982, Isham 1995, Chan and Isham 1998, Pugliese et al. 1998, Rosà and Pugliese 2002, Rosà
45 et al. 2003, Grear and Hudson 2011, Fowler and Hollingsworth 2016). While the process-based
approach has usefully illuminated various aggregating and disaggregating mechanisms in host-
parasite systems (summarized in Wilson et al. 2002), it suffers from the “many-to-one problem”
48 inherent in much of ecology (Frank 2014): there are many process-based models that can result in
similar levels of parasite aggregation making it difficult to identify the specific processes leading
to aggregation from patterns alone. When lab or field experiments are not a viable option to
51 identify mechanism in host-parasite systems, it would be useful to have some criteria to identify
when observed patterns of parasite aggregation may provide some information about the mech-
anisms influencing a host-parasite system or when most of the information is provided in the
54 mean parasite abundance.

Recently developed constraint-based models used in macroecology provide such a criteria. These models are different from the process-based approach in that they attempt to predict the most-likely form of a population- or community-level distribution using only a known set of statistical constraints (Harte 2011, Locey and White 2013, Newman et al. 2014, Xiao et al. 2015b). The constraint-based approach does not propose that biological mechanisms are not acting in a system; it contends that many different combinations of these mechanisms lead to similar patterns of aggregation with predictable statistical properties (Frank 2009, McGill and Nekola 2010, Frank 2014). This is important because these models can then be used as robust null models (i.e. models that do not trivially fail) to identify when a given observed distribution contains biological information beyond that given by the constraints used to predict the distribution (Locey and White 2013, Harte and Newman 2014). Similarly, null model approaches have been used in community ecology to identify when signals of an ecological process can be discerned from observed changes in a community metric (e.g. changes in β diversity) when a change in this metric can also be a concomitant result of changes in another metric (e.g. changes in α diversity; Chase and Myers 2011, Chase et al. 2011). These constraint-based and null model approaches have had much success in understanding patterns in free-living populations and communities (Leibold and Mikkelsen 2002, White et al. 2012, Ulrich and Gotelli 2013, Harte et al. 2015) and we argue that they can also be useful in addressing mechanistic questions about parasite aggregation in disease ecology.

For example, any observed host-parasite distribution is constrained by the total number of parasites P and the total number of hosts H in the sample. Given this, there are only a finite number of shapes that this sampled host-parasite distribution can take (i.e. the feasible set of the host-parasite distribution, Locey and White 2013). If the shape of this observed host-parasite distribution is similar to the most likely distribution within this feasible set, then making inferences about the biological mechanisms leading to the shape of this distribution is difficult as the observed distribution is simply the most-likely distribution of all possible distributions (Haegeman and Loreau 2009). In other words, many different combinations of host and/or parasite-related processes will lead to the same host-parasite distribution, and this is the distribution that is

predicted by the constraint-based model.

84 This has important implications for disease ecology where one often wants to understand something about the mechanisms affecting a host-parasite system from the level of aggregation observed (Anderson and Gordon 1982). Having some robust criteria for when a sampled host-
87 parasite distribution shows “unusual” aggregation can help identify host-parasite systems where particular aggregating or disaggregating mechanisms are disproportionately constraining the distribution beyond the inherent (but biologically important) constraints imposed by P and H .
90 We define “unusual” aggregation as a level of aggregation that is significantly different than the level of aggregation predicted by the most likely distribution in the feasible set (e.g. Locey and White 2013, Harte and Newman 2014).

93 The traditional null model for the distribution of parasites across hosts follows a Poisson distribution, which is typically derived from a death-immigration process (Anderson and Gordon 1982). The constraint-based approach for aggregation is asking a different question than
96 tests of the traditional Poisson null model. Rejecting the Poisson null model indicates that this simple model is not capturing the important aggregating or disaggregating mechanisms in a host-parasite system, however failing to reject the Poisson null model is not proof that a system
99 is following a simple death-immigration process. In contrast, the constraint-based null models make no assumptions about the particular processes leading to its predicted level of aggregation and simply predict the most likely level of aggregation for a system with P parasites and H
102 hosts. Unlike the classic Poisson null hypothesis, failing to reject the constraint-based model tells us something important: our empirical pattern of aggregation does not contain any information about process beyond that already contained in P and H (Harte and Newman 2014). Because
105 this approach robustly identifies “unusual” aggregation, it can help us better understand when the effects of processes, such as parasite-induced mortality or host-heterogeneity, can be reliably inferred from observed host-parasite distributions.

108 This study has two goals. First, we use a dataset consisting of 22 unique amphibian host-trematode parasite pairings with over 8000 amphibians sampled at 205 sites over 5 years to test whether constraint-based models used in free-living systems also provide robust null models

for host-parasite distributions. Second, we explore how, upon failing to describe host-parasite distributions, these constraint-based models can be extended to account for known aggregating mechanisms (such as host-heterogeneity) and disaggregating mechanisms (such as parasite-induced host mortality) in host-parasite systems. We find that the shape of host-macroparasite distributions are generally well-predicted by the constraint-based approach. These results show that to reliably infer something about biological mechanism directly affecting patterns of parasite aggregation, one must first account for the strong constraints imposed on aggregation by P and H .

Methods

The methods section is organized as follows. The first section gives an overview of two constraint-based null models that have been recently used in the macroecological literature (Haegeman and Etienne 2010, Locey and White 2013, Xiao et al. 2015a). The second section describes how we generated predicted host-parasite distributions from these two constraint-based null models. The third section describes how we compared the constraint-based null models to data. Finally, the fourth section describes how we extended these constraint-based null models to account for known aggregating and disaggregating mechanisms in host-parasite systems. Table 1 contains a list of terms and definitions used to define the constraint-based models.

Defining the weighted feasible sets for constraint-based null models of parasite aggregation

The constraint-based null models that we consider have two constraints inherent in any sampled host-parasite distribution: the total number of parasites sampled P and the total number of hosts sampled H . Given these constraints, both models proceed by enumerating the feasible set of all possible macrostates of P parasites and H hosts (Locey and White 2013). We define a macrostate as one possible unordered host-parasite distribution resulting from distributing P parasites among H hosts (Table 1). For example, the feasible set of possible macrostates that we can observe given $P = 3$ parasites and $H = 3$ hosts is $F = \{\{3, 0, 0\}, \{2, 1, 0\}, \{1, 1, 1\}\}$. The

macrostate $\{3, 0, 0\}$ specifies that one host has three parasites and two hosts have zero parasites.

After specifying all of the possible macrostates in a feasible set constrained by P and H , each macrostate is then assigned a weight. For example, some macrostates may be combinatorially more likely to occur than others and thus will have a larger weight in the feasible set. Determining how to weight each macrostate depends on whether hosts and/or parasites are considered labeled or unlabeled (Table 1; Haegeman and Etienne 2010).

One option is to specify that both hosts and parasites are unlabeled such that all possible macrostates are equally likely to occur. This is equivalent to integer partitions used in combinatorics (Bóna 2006, Xiao et al. 2015a), so we call this model the “partition model”. The partition model is process-independent and makes no assumptions about any potential mechanisms leading to a given macrostate (Locey and White 2013). Therefore, no macrostate is more likely to occur than any other macrostate (Xiao et al. 2015a). Assuming that each macrostate is equally likely is not equivalent to assuming that any single host is equally likely to have a parasite abundance from zero to P . The probability of a single host having a parasite abundance of $x = 0, \dots, P$ is $p(x|P, H) = \sum_{m \in F} p(x|m, P, H)p(m|P, H)$ where m is a macrostate in the feasible set F . Using our example from above, the partition model assigns each of the three macrostates in the feasible set an equal probability of $1 / 3$. The probability of observing a single host with $x = 0, 1, 2$, or 3 parasites is $p(0) = 3/9$, $p(1) = 4/9$, $p(2) = 1/9$, and $p(3) = 1/9$.

A second option for weighting macrostates is to again assume that parasites are unlabeled, but now assume that hosts are labeled. This is equivalent to integer compositions used in combinatorics (Bóna 2006, Xiao et al. 2015a), so we call this model the “composition model”. Using the composition model, particular macrostates are more likely to occur because they are associated with a larger number of possible configurations. For example, given labeled hosts and unlabeled parasites, the macrostate $\{3, 0, 0\}$ could be realized from three different configurations: $(3, 0, 0)$, $(0, 3, 0)$, and $(0, 0, 3)$. Enumerating all the configurations for the other macrostates in our example feasible set, we see that the macrostate $\{2, 1, 0\}$ can occur six ways and $\{1, 1, 1\}$ can occur one way. Therefore, the macrostate $\{3, 0, 0\}$ has a weight of $3 / 10$, $\{2, 1, 0\}$ has a weight of $6 / 10$, and $\{1, 1, 1\}$ has a weight of $1 / 10$. The probability of observing a single host with

165 $x = 0, 1, 2$, or 3 parasites is $P(0) = 4/10$, $P(1) = 3/10$, $P(2) = 2/10$, and $P(3) = 1/10$.

More generally, for any macrostate m in a feasible set with P unlabeled parasites and H unlabeled hosts there are a total of $b_m = \frac{H!}{\prod_{i \in A} h_i!}$ configurations with unlabeled parasites and
168 labeled hosts (Brualdi 2010). A is a set containing the unique parasite abundances found in macrostate m , i is a particular member of that set, and h_i is the number of hosts in macrostate m that have a parasite abundance i . Note that $\sum_{i \in A} h_i = H$. The total number of possible
171 configurations of all macrostates using the composition model is given by $D = \frac{(H + P - 1)!}{P!(H - 1)!}$ (Harte 2011). Taken together, the weight on any particular macrostate m using the composition model is $\frac{b_m}{D}$.

174 In summary, both the composition and partition models place our observed host-parasite distribution in the context of all possible observable host-parasite distributions. In particular, this allows us to ask an important question in parasite ecology: does a host-parasite distribution
177 contain any information about biological mechanism beyond that already contained in P and H ? Because there is no general consensus on which approach is preferable (Haegeman and Etienne 2010, Xiao et al. 2015a), we consider both the partition model and the composition model in this
180 study. Figure 1 gives a visual comparison of these two models.

We could have also considered two other approaches: labeled hosts and labeled parasites or unlabeled hosts and labeled parasites. We chose not to consider these approaches because
183 assuming labeled parasites is neither consistent with the pattern that we are interested in (i.e. the host-parasite distribution) nor how host-parasite systems are sampled. Assuming labeled parasites tracks the location of each individual parasite in the host population, whereas we are
186 interested in the population-level distribution of parasites across hosts (see Xiao et al. 2015a, for the equivalent argument in free-living individuals). Moreover, specifying labeled parasites assumes that the system could be sampled by randomly choosing a parasite and assigning it
189 a unique label and a label corresponding to the host in which it was found (assuming labeled hosts). This process would then be repeated until some number of P parasites were sampled. The total number of hosts H would then be given by the number of unique host labels on our
192 P sampled parasites. This is not how host-parasite systems are sampled. Instead, H hosts are

randomly sampled and the P parasites within these hosts are counted. This is more consistent with unlabeled parasites.

195 Despite these issues, the case with labeled hosts and labeled parasites is noteworthy because it results in a Poisson distribution of parasites across hosts (see Appendix S1). However, the model resulting in the Poisson distribution has exactly the same number of assumptions as the partition and composition models, so there is no *a priori* reason to favor one model over the
198 other. The only way to discriminate between the approaches is to compare them to empirical host-parasite distributions (Haegeman and Loreau 2009, Haegeman and Etienne 2010, Xiao et al. 2015a), against which the Poisson almost universally fails (Shaw and Dobson 1995, Shaw et al. 1998, Wilson et al. 2002). Appendix S6: Figure 1 illustrates the completely unsurprising result that the Poisson distribution also does not capture the level of parasite aggregation in the data
201 we present here.
204

Moving from weighted feasible sets to constraint-based null model predictions

The proceeding section described how we enumerated and weighted the feasible sets for the
207 partition and composition models. This section describes how we generate the predicted host-parasite distributions from these two models.

Given a weighted feasible set of macrostates from either the partition or composition model,
210 the central tendency of this feasible set provides a prediction for the most likely host-parasite distribution given the constraints P and H (Locey and White 2013). We define the central tendency of a weighted feasible set as the vector of marginal medians of this feasible set (Appendix
213 S2). For most realistic values of P and H it is computationally intractable to enumerate all possible macrostates in the feasible set to compute this central tendency. To address this problem, we used the algorithms provided by Locey and McGlinn (2013) to randomly draw macrostates
216 from all possible macrostates in a feasible set defined by P and H . We then computed the central tendency of this sample as an estimate of the central tendency of the full feasible set (Figure 1; Locey and White 2013). To generate the predicted host-parasite distribution for the partition
219 model, we drew 1000 random macrostates from a feasible set defined by P and H and used the

central tendency of this sample as our predicted host-parasite distribution (Figure 1).

While we could have used the same approach to compute the predicted host-parasite distribution for the composition model by weighting each randomly drawn macrostate m by b_m/D , we instead used the analytical result from maximum entropy theory that the probability $p(x|P, H)$ of a single host having x parasites under the composition model is (Haegeman and Etienne 2010)

$$p(x|P, H) = \frac{\binom{P-x+H-2}{P-x}}{\binom{P+H-1}{P}} \quad (1)$$

The predicted rank abundance distribution of equation 1 is equivalent the central tendency of the weighted feasible set for the composition approach (Appendix S2). Moreover, equation 1 shows us that the composition approach is equivalent to assuming that host-parasite distributions follow a finite negative binomial distribution with $k = 1$ (Zillio and He 2010). Note that we are not arbitrarily setting $k = 1$ – this is a direct result from maximizing entropy with respect to the constraints P unlabeled parasites and H labeled hosts. Moreover, the finite nature of this distribution is a direct result of the constraint P , which can lead to better descriptions of aggregation in finite populations (Zillio and He 2010). However, P is similar to the number of trials in a binomial distribution and cannot be tuned to improved the fit of the model.

In summary, we used both a sampling based approach and an analytical formula to generate predicted host-parasite distributions from the weighted feasible sets of our two constraint-based null models. In the context of more commonly used distributions in disease ecology, these two constraint-based null models have one less parameter than a negative binomial model, which is a very flexible distribution that often fits host-parasite distributions very well (Shaw et al. 1998). We stress that the goal of this study is not to ask whether these distributions do better or worse than a negative binomial in predicting a host-parasite distribution, but whether host-parasite distributions tend to contain information beyond what is given by P and H .

Comparing constraint-based models to empirical data

Description of empirical data

To test whether empirical host-parasite distributions contained information beyond that given by the constraints P and H , we used an extensive dataset of all macroparasites found in 8099 amphibian hosts across 205 ponds (sites) in the East Bay region of California (Alameda, Contra Costa and Santa Clara counties) from 2009-2014 (Johnson et al. 2013). This included ponds from publicly accessible parks, open space preserves, municipal watershed districts, and private ranches. In this field study, we sampled recently metamorphosed amphibians, as these provide a reliable and standardized indicator of infections acquired during aquatic development from the associated pond. In a given survey event, we randomly collected at least 10 of each host species as they approached metamorphosis using the methods described in Johnson et al. (2016). To measure parasite abundance, we performed a systematic examination of all major tissues and organs in the sampled hosts for parasites (Hartson et al. 2011). The sampled amphibians consisted of *Pseudacris regilla* (Pacific chorus frog, $n = 4431$), *Anaxyrus boreas* (Western toad, $n = 1309$), *Lithobates catesbeianus* (American bullfrog, $n = 410$), *Taricha torosa* (California newt, $n = 1568$), and *Taricha granulosa* (Rough-skinned newt, $n = 381$).

We focused the following analyses on the five most common macroparasites in the system in terms of both prevalence and abundance. These were the larval trematodes *Ribeiroia ondatrae* (RION), *Echinostoma* sp. (ECSP), *Alaria* sp. (ALAR), *Cephalogonimus* sp. (CEPH), and *Manodistomum* sp. (MANO). All of these trematodes have complex life cycles in which their first intermediate hosts are pulmonate snails, their second intermediate host can be amphibians, snails or fishes, and their definitive hosts are water-associated vertebrates (reptiles, amphibians, birds, or mammals) (Johnson and McKenzie 2008).

Comparing models to data

We determined whether the empirical distributions of parasites across hosts deviated from the predictions of our two constraint-based models for each combination of host species and parasite species at each site during each year. We included a year-by-site-by-host-by-parasite distribution only if it had at least 10 parasites and 10 hosts. Given this criterion, we were able to compare the constraint-based models to 842 host-parasite distributions. As expected, 837 of these

distributions were aggregated with a $\ln(\text{variance to mean ratio})$ greater than zero (Appendix S6: Figure 1). For each of these distributions, we extracted the total number of individuals of a given amphibian species (H) and parasites of a given trematode species (P) and calculated the corresponding rank abundance distribution (RAD) for the constraint-based models as the central tendency of the weighted feasible set (see *Moving from weighted feasible sets to constraint-based null model predictions*). The RAD gives the predicted parasite abundances from a given distribution for H hosts and assigns a rank of 1 to the host with highest abundance and a rank of H to the host with the lowest abundance (Harte 2011, White et al. 2012).

To determine whether an observed host-parasite distribution deviated from the central tendency of a constraint-based model, we plotted the observed RAD (obs_i) versus the predicted RAD (pred_i). We then calculated the R^2 value based on a fit to the 1:1 line using the equation (White et al. 2012, Xiao et al. 2015b)

$$R^2 = 1 - \frac{\sum_i (\ln(\text{obs}_i + 1) - \ln(\text{pred}_i + 1))^2}{\sum_i (\ln(\text{obs}_i + 1) - \ln(\text{obs}_i + 1))^2}$$

where i is the rank ($i = 1, \dots, H$) of each observed or predicted host in a distribution. R^2 to the 1:1 line describes how much variance in the observed data is described by the model prediction. If the model describes a large portion of the variation in the observed data then the R^2 value will be larger, with unity being a perfect prediction. If the model is a poor fit, the R^2 value will be much less than unity and possibly negative if the 1:1 line was a worse fit than assuming that each host had a parasite abundance equal to the mean of the observed distribution (White et al. 2012). We calculated R^2 values for each distribution independently as well as for all distributions combined. We also explored a number of alternative measures of goodness-of-fit that gave consistent results (Appendix S3).

Extending the constraint-based null models to account for aggregating and disaggregating mechanisms

When an observed host-parasite distribution deviated from the central tendency of a constraint-based null model, this provided evidence that additional constraints/mechanisms beyond just P

and H were disproportionately affecting the system (Harte and Newman 2014). We developed two ways to extend the constraint-based null models to detect whether classic aggregating and disaggregating mechanisms may be affecting host-parasite distributions beyond P and H .

Accounting for disaggregating mechanisms

Disaggregating mechanisms such as parasite-induced host mortality can play an important role in structuring empirically observed host-parasite distributions (Anderson and Gordon 1982). The parasite *Ribeiroia ondatrae* is known to have a strong, intensity-dependent effect on the survival of some amphibian hosts where increased parasite intensity leads to increased limb-malformations and decreased survival (Johnson 1999). This means that hosts with large parasite burdens are removed from the system, making the parasite distribution more uniform. Therefore, *Ribeiroia*-induced host mortality may interact with P and H to further constrain the shape of host-*Ribeiroia* distributions.

We included *Ribeiroia*-induced host mortality as an additional constraint on the partition and composition null models. To do this, we used laboratory-derived survival curves that describe how *Ribeiroia* intensity affects amphibian host survival probability (Johnson et al. 2012). We focused on the amphibian species *Pseudacris regilla* because *Ribeiroia*-induced mortality and malformations in this species have been documented in the field and in the lab (Johnson 1999, Johnson and McKenzie 2008) and there were a large number of *P. regilla*-*Ribeiroia* distributions in the dataset on which to test the extended models ($n = 133$). The intensity-dependent survival curve specified the probability of an amphibian host surviving from larva to recent metamorph with some observed parasite intensity. We assumed that this curve followed a logistic function and estimated the parameters of this function from independent laboratory data (Figure 2A; Appendix S4; Johnson 1999).

We then used this result to further constrain the partition and composition model predictions by assigning each macrostate a likelihood using the estimated survival function (Figure 2A). For each constraint-based model, the macrostates were then weighted by this likelihood such that macrostates with small likelihoods (e.g. ones that contained hosts with high parasite loads) were

less likely to be observed than macrostates with large likelihoods. Using this weighting scheme, we sampled from models that were constrained on P , H and *Ribeiroia*-induced host mortality using a Metropolis-Hastings algorithm (Figure 2B; see Appendix S4 for a full description of the algorithm used).

Once we obtained estimates of the mortality-constrained partition and composition models, we compared the resulting predictions to the observed *P. regilla*-*Ribeiroia* distributions using the methods described in *Comparing models to data*. In addition, we also calculated an approximate AICc for the constraint-based model with and without an additional mortality constraint. We compared these models using ΔAICc where we considered an absolute value of $\Delta\text{AICc} > 2$ as evidence that one model was better than the other (Burnham and Anderson 2002). The AICc values were approximate because there was no analytically defined likelihood for the mortality-constrained models. Therefore, we approximated the likelihood by drawing a large number of samples (e.g. 500 samples) from the mortality-constrained feasible set and computing the likelihood of a single host having x parasites using the equation $p(x|P, H) = \sum_{m \in \hat{F}} p(x|m, P, H)p(m|P, H)$, where \hat{F} is the sampled feasible set. As we did not perform any additional model fitting to derive the mortality-constrained model, it was not statistically inevitable that the central tendencies of the mortality-constrained models would provide a better representation of the data. Therefore, an improvement in agreement between model and data, reflected in an increased R^2 or decreased AICc for the mortality models relative to models without mortality, is strong evidence that *Ribeiroia*-induced mortality is constraining the distribution beyond P and H .

Accounting for aggregating mechanisms

Host heterogeneity, whether it be in susceptibility, parasite encounter rates, behavior or other factors, is an important mechanism leading to aggregation in host-parasite systems (Cornell 2010, Raffel et al. 2011). We accounted for this aggregating mechanism by extending the constraint-based models to include empirically observed levels of host heterogeneity. In particular, we explored discrete host heterogeneity where we assumed that overaggregation relative to the pre-

dicted model was a result of mixing discrete groups of hosts (Grafen and Woolhouse 1993, Wilson
et al. 2002). This approach is different than the standard practice of fitting a negative binomial
distribution to overaggregated host-parasite distributions. If the goal of an analysis is to obtain
the best possible fit to an observed host-parasite distribution then it is well-known that fitting
a negative binomial distribution provides an excellent model of overaggregated host-parasite
distributions (Shaw et al. 1998, Calabrese et al. 2011). However, if the goal of an analysis is to
determine whether a host-parasite distribution contains any information beyond what is con-
tained in P and H , fitting a negative binomial model does not provide immediate insight into
what constitutes unusual aggregation or the potential host attributes leading to this overaggrega-
tion (but see Alonso and Pascual 2006, Fowler and Hollingsworth 2016, for various mechanistic
interpretations of the negative binomial k parameter). Extending a constraint-based model to in-
clude discrete host-heterogeneity, as is done here, can help generate more specific hypotheses as
to the relative importance of different levels of host-heterogeneity in structuring a host-parasite
distribution.

To incorporate discrete host-heterogeneity, we used 5 observed host attributes by which we
could bin hosts into groups of heterogeneity. The first attribute was host body size (i.e. snout-vent
length), which is a well-known attribute affecting parasite exposure and aggregation (Grutter
and Poulin 1998, Poulin 2013). The other 4 host attributes were the parasite abundances of the
larval trematodes, excluding the focal trematode, infecting an individual host (see Fig. 3 for an
example). Coinfection can potentially increase aggregation by increasing heterogeneity in host
susceptibility to the focal parasite (Cattadori et al. 2008), but can also decrease aggregation by
increasing intra-host parasite negative density dependence (Pacala and Dobson 1988). Here we
consider coinfection as a mechanism leading to increased aggregation.

Using these 5 host attributes, we used regression trees in which the response variable was
the focal parasite abundance and the predictor variables were the 5 host attributes described
above (Fig. 3). Separate regression trees were run for each of the 842 host-parasite distributions.
For a given host-parasite distribution, we found the best regression tree with 2-5 of groups of
host heterogeneity and calculated the relative importance of each predictor variable based on

381 how much they reduced the sum of squared error compared to the other predictor variables
(Fig. 3). We restricted each group to have at least 2 hosts. Within each of these j groups we
determined the total number parasites P_j and the total number of hosts H_j . The regression tree
384 approach explores how various predictor variables affect mean parasite load (James et al. 2013),
which is consistent with the constraint-based assumption that much of the information about the
host-parasite distribution is contained in P and H .

387 To generate a constraint-based model RAD from the results of the regression tree, the RADs
for each group j were computed with P_j and H_j and the predicted RAD was given by the concate-
nation for these j vectors (Fig. 3). This predicted mixture RAD could then be analyzed using the
390 various methods described above. We also computed approximate AICc values for each hetero-
geneity model that we applied to an observed distribution. As described in the previous section,
we did this by drawing 500 macrostates from the heterogeneity model to generate an estimate
393 for the probability of a single host having x parasites under either the partition or composition
assumptions. Finally, we employed a randomization test to ensure that any increase in R^2 after
including host heterogeneity was due to the host attributes considered, rather than just the act of
396 grouping itself (described in Appendix S5).

In summary, while P and H alone may sometimes not sufficiently constrain an observed host-
parasite distribution, this approach is testing whether allowing P and H to vary as a function of
399 host heterogeneity can account for deviations from the constraint-based null models. All analyses
were performed in Python (version 2.7.11) and the code to replicate the analysis can be found at
https://github.com/mqwilber/feasible_parasites.

402 Results

Do host-parasite distributions contain information beyond that contained in P and H ?

405 Overall, the partition model and composition model described 86% and 85% of the variation
in all of the observed host-parasite distributions combined, respectively (Fig. 4A, B). For any

particular host-parasite distribution, the median R^2 for the partition and composition models was
408 0.78 and 0.76, respectively (Fig. 4A, B). Examining the models with regard to host-by-parasite
combinations, the median R^2 for the constraint-based models tended to be close to 80% for the
various host-by-parasite combinations (Appendix S6: Fig. 2), with some notable exceptions for
411 the host *Lithobates catesbeianus* and the parasites *Alaria* sp. and *Cephalogonimus* sp. (Appendix S6:
Fig. 2-5).

The partition model tended to describe more variation in host-parasite distributions than the
414 composition model (Fig. 4, Appendix S6: Fig. 6). The composition model is equivalent to a finite
negative binomial model with $k = 1$ and many of the observed host-parasite distributions in
this study had maximum-likelihood estimated k parameters (\hat{k}) less than one (Appendix S6: Fig.
417 6). While $\hat{k} \neq 1$ is not necessarily incompatible with the composition model due to estimation
error in the negative binomial k parameter (Lloyd-Smith 2007), the partition model did predict
more aggregated distributions than the composition model (Fig. 1). This led to the partition
420 model accounting for a larger amount of the variance in host-parasite distributions with $k < 1$
(Appendix S6: Fig. 6).

Accounting for disaggregating and aggregating mechanisms

423 *Disaggregating mechanisms: Parasite-induced host mortality*

Including independently-estimated *Ribeiroia*-induced parasite mortality into the constraint-based
models improved the overall fit of the models to *Pseudacris regilla*-*Ribeiroia* distributions. This
426 was seen in three different metrics. First, there was a significant increase in the overall R^2 when
the mortality constraint was included (bootstrapped 95% confidence interval for the difference
in overall R^2 between the mortality constraint-based model and the null constraint-based model
429 from 1000 re-samples: feasible set model, [0.019, 0.033]; maximum entropy model, [0.018, 0.035];
neither interval includes 0; Fig. 5A-D). This improvement in fit can be visualized by observing
the tightening of the points to the 1:1 line when *Ribeiroia*-induced mortality was included in
432 the model (Fig. 5A-D). Second, the median R^2 for individual *Pseudacris*-*Ribeiroia* distributions
increased and the variance around the individual R^2 values decreased (Fig. 5B,D). Third, for

individual distributions that were better fit under either the mortality or no mortality models
435 based on the absolute value of $\Delta\text{AICc} > 2$, a significant or marginally significant proportion were
better under the mortality model (partition model: Binomial test, $N = 26$, better under mortality
model = 21, $p = 0.002$; composition model: Binomial test, $N = 31$, better under mortality model
438 = 21, $p = 0.07$; Fig. 5B,D).

Aggregating mechanisms: Host-heterogeneity

There were 124 unique host-parasite distributions that had $R^2 < 0.5$ and an observed variance
441 to mean ratio greater than the variance to mean ratio of one of the constraint-based models. We
considered these distributions to be overaggregated with respect to the constraint-based models.
Of these 124 distributions, 48 were *Echinostoma* (12% of all *Echinostoma* distributions) 29 were
444 *Alaria* (35% of all *Alaria* distributions), 17 were *Cephalogonimus* (24% of all *Cephalogonimus* distri-
butions), 17 were *Manodistomum* (33% of all *Manodistomum* distributions), and 13 were *Ribeiroia*
(5% of all *Ribeiroia* distributions).

447 Considering only these overaggregated distributions, we used the regression tree analysis
described above to test whether further constraining P and H based on known host attributes
improved the fit of the constraint-based models to the observed host-parasite distributions. The
450 heterogeneity models built from the regression tree analysis improved the fit of the constraint-
based models to the empirical data beyond what would be expected by the inevitable increase in
fit by simply grouping hosts (overall R^2 greater than the 95% interval from randomly permuting
453 hosts into groups; Fig. 6). The improvement in fit can be visualized in Figure 6 by noting how
the data points compress to the 1:1 line as more groups of heterogeneity are included. Note
that this increase in model fit was not achieved by minimizing or maximizing any criteria about
456 how well the heterogeneity model fit the observed host-parasite distribution. Finally, for three or
more groups of host-heterogeneity, including just host body-size heterogeneity as an additional
constraint yielded better models in terms of both higher R^2 values and larger AICc weights than
459 including just heterogeneity in coinfection as an additional constraint (Fig. 7A-D).

Discussion

The shape of a sampled host-parasite distribution is necessarily constrained by the total number of hosts H and the total number of parasites P found in that sample (Haegeman and Etienne 2010, Locey and White 2013). While there are indisputably biological mechanisms leading to the shape of this distribution (Wilson et al. 2002), inferring anything about these mechanisms may be difficult without first accounting for the constraints imposed by P and H . Here we use an extensive dataset of 22 host-parasite combinations and 842 empirical host-parasite distributions to show that aggregated host-parasite distributions tend to be consistent with the most likely distribution given P and H . This suggests that when trying to make inference about biological mechanism directly affecting patterns of parasite aggregation one must account for the mechanisms indirectly affecting aggregation through changes to P and H .

This finding has three important implications for disease ecology. First, there is a rich history in parasitology of using the shape of host-parasite distributions in combination with dynamic models and statistical techniques to infer which mechanisms may be affecting a given host-parasite system (Crofton 1971, Anderson and Gordon 1982, Grear and Hudson 2011). While these approaches are in no way inappropriate, our results show that the instances in which the shape of a host-parasite distribution contains more information beyond what is contained in P and H may be more rare than previously thought. This result is consistent with other findings showing that log mean parasite load describes up to 88% of the variation in the log variance of parasite load, leaving only 13% of the variation to be described by biological mechanisms acting on something other than the mean (Shaw and Dobson 1995, Poulin 2013). Our results take this a step further by using constraint-based models to explicitly predict the entire host-parasite distribution given P and H . We find that, similar to Poulin (2013), much of the variability in the entire host-parasite distribution (not just the variance) is well predicted by mean parasite load and how many hosts are present in the sample. As a next step, explicitly considering whether specific attributes of observed host-parasite distributions systematically deviate from constraint-based predictions, such as the number of predicted uninfected hosts, could shed additional light on when host-parasite distributions contain much information beyond P and H .

The second implication is that the success of constraint-based models in disease ecology will
489 allow them to be adopted as robust null models against which empirical host-parasite distribu-
tions can be compared. Constraint-based models are being increasingly used as robust null mod-
els in community ecology to determine when ecological mechanism may be disproportionately
492 affecting the shape of population- and community-level distributions (Ulrich and Gotelli 2013,
Newman et al. 2014, Xiao et al. 2015b). In disease ecology, by using a constraint-based model
to predict parasite aggregation given P and H , we can determine when a host-parasite system
495 is showing unusual levels of aggregation to help direct modeling and experimental efforts. For
example, future studies could explore whether factors such as the complexity of parasite life
cycles (Lester and McVinish 2016), self-reinfection processes (Gear and Hudson 2011), or the
498 composition of the host and parasite community in which a distribution is observed (Krasnov
et al. 2006) lead to consistent deviations from constraint-based predictions.

Third, the general success of constraint-based models in describing host-parasite distributions
501 has important implications for understanding the dynamics of host-macroparasite systems. Most
macroparasite models explicitly model the state variables H and P (Anderson and May 1978,
Dobson and Hudson 1992) and examine, in addition to other biological factors, how either fixed
504 (Anderson and May 1978) or dynamic aggregation (Kretzschmar and Alder 1993, Rosà et al. 2003)
influences host and parasite dynamics. Constraint-based models in turn predict that aggregation
is largely determined by exactly these state variables. Therefore, a constraint-based approach
507 to parasite ecology can be directly linked back to a more familiar mechanistic framework by
examining the implications of constraint-based, aggregation predictions on dynamics of the total
number of hosts and parasites in a system. Linking constraint-based models for describing
510 aggregation to dynamic equations for the state variables of a system has often been alluded to
in macroecology (Supp et al. 2012, White et al. 2012), but has been difficult to implement (Harte
2011). The rich empirical and theoretical understanding of biological factors affecting the total
513 number of hosts and the total number of parasites in a system (Kretzschmar and Alder 1993,
Hudson et al. 1992, Dobson and Hudson 1992) makes disease ecology an ideal field in which to
make this connection.

516 In addition to providing robust null models and a unique opportunity to link dynamic, mechanistic models with a constraint-based approach, the constraint-based models can also be extended beyond null models to test the importance of potential aggregating and disaggregating mechanisms affecting host-parasite distributions. In this study, we extended the constraint-based models to include independently estimated relationships between parasite intensity and amphibian survival and found that accounting for the well-described negative effect of *Ribeiroia* on *P. regilla* (Johnson 1999, Johnson et al. 2012) improved the fit of the constraint-based model to empirical host-parasite distributions. While this improvement in model fit was not drastic as *P* and *H* already accounted for 87% of the variation in the distributions, it was achieved using a survival curve estimated from an independent dataset (Johnson 1999), providing strong evidence that parasite-induced mortality is influencing *P. regilla*-*Ribeiroia* distributions beyond just changes to *P* and *H*.

528 Moreover, we also found that extending constraint-based models to include heterogeneity in host body size and coinfection with other trematode parasites accounted for much of the over-aggregation in observed distributions that were not well described by the constraint-based null models. In particular, we found evidence that host body size was generally a more important constraint on the host-parasite distribution than a host's level of coinfection with other trematodes. This result is consistent with previous studies which have shown the importance of host age/body size heterogeneity for increasing parasite aggregation due to changes in host immunity and/or exposure to parasites with host age/body size (Pugliese et al. 1998, Poulin 2013). Moreover, while previous work has shown that coinfection can act as a type of host heterogeneity and increase parasite aggregation (Cattadori et al. 2008), this same work has also shown that host characteristics such as age/body size, sex, and breeding status can often be more important factors affecting parasite aggregation and host-parasite dynamics than coinfection. While we have considered host-heterogeneity and parasite-induced mortality separately in this study, there is no reason that the constraint-based approach cannot be extended to included multiple mechanistic constraints. However, this must be done judiciously as imposing too many constraints could lead to trivial agreements between the model and data (Haegeman and Loreau 2009).

In conclusion, constraint-based models provide a powerful framework for understanding when we can reliably infer mechanism from parasite aggregation. However, we are not advocating that the constraint-based approach should replace the process-based approach that has been so successful in disease ecology. Rather, the constraint-based approach is another tool in the disease ecologist's belt that can highlight when observed parasite aggregation is telling us something novel about the mechanisms acting in our system and when we should acknowledge the statistical inevitability that sometimes host-parasite distributions simply look how they must look given P and H .

Acknowledgments

We would like to thank the numerous members of the Johnson Lab at University of Colorado, Boulder who collected and processed the thousands of samples that comprise this dataset, as well as the many land managers who generously provided access to study sites, including East Bay Regional Parks, East Bay Municipal Utility District, Santa Clara County Parks, Hopland Research and Extension Center, Blue Oak Ranch Reserve, California State Parks, The Nature Conservancy, Open Space Authority and Mid-peninsula Open Space. We would also like to thank Bill Murdoch, Roger Nisbet, and three anonymous reviewers for helpful comments on this manuscript. The National Institutes of Health (USA) Grant 1R01GM109499 from the Ecology of Infectious Disease program, the National Science Foundation (USA) (DEB-0841758, DEB-1149308), the National Geographic Society, and the David and Lucile Packard Foundation provided support for this work. M.W. was supported by a National Science Foundation, USA, Graduate Research Fellowship (Grant No. DGE 1144085) and the University of California Regents (USA).

References

- Alonso, D. and M. Pascual, 2006. Comment on "A keystone mutualism drives pattern in a power function". *Science* **313**:1739; author reply 1739.
- Anderson, R. M. and D. M. Gordon, 1982. Processes influencing the distribution of parasite numbers within host populations with special emphasis on parasite-induced host mortalities. *Parasitology* **85**:373–398.

Anderson, R. M. and R. M. May, 1978. Regulation and stability of host-parasite interactions: I. Regulatory processes. *Journal of Animal Ecology* **47**:219–247.

573 Anderson, R. M. and R. M. May, 1979. Population biology of infectious diseases: Part I. *Nature* **280**:361 – 367.

Bóna, M., 2006. A Walk Through Combinatorics: An Introduction to Enumeration and Graph
576 Theory. World Scientific Publishing Co., Toh Tuck Link, Singapore, second edition.

Brualdi, R., 2010. Introductory Combinatorics. Pearson Education, Inc, Upper Saddle River, New Jersey, fifth edition.

579 Burnham, K. P. and D. R. Anderson, 2002. Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach. Springer, New York.

Calabrese, J. M., J. L. Brunner, and R. S. Ostfeld, 2011. Partitioning the aggregation of parasites on
582 hosts into intrinsic and extrinsic components via an extended Poisson-gamma mixture model. *PloS one* **6**:e29215.

Cattadori, I. M., B. Boag, and P. J. Hudson, 2008. Parasite co-infection and interaction as drivers
585 of host heterogeneity. *International Journal for Parasitology* **38**:371–380.

Chan, M. S. and V. S. Isham, 1998. A stochastic model of schistosomiasis immuno-epidemiology. *Mathematical Biosciences* **151**:179–198.

588 Chase, J. M., N. J. B. Kraft, K. G. Smith, M. Vellend, and B. D. Inouye, 2011. Using null models to disentangle variation in community dissimilarity from variation in α -diversity. *Ecosphere* **2**:1–11.

591 Chase, J. M. and J. A. Myers, 2011. Disentangling the importance of ecological niches from stochastic processes across scales. *Philosophical Transactions of the Royal Society B: Biological Sciences* **366**:2351–2363.

594 Cornell, S. J., 2010. Modelling stochastic transmission processes in helminth infections. In *Modelling Parasite Transmission and Control*, chapter 5, pages 66–78.

Crofton, H. D., 1971. A quantitative approach to parasitism. *Parasitology* **62**:179–193.

597 Dobson, A. P. and P. J. Hudson, 1992. Regulation and stability of a free-living host-parasite system: *Trichostrongylus tenuis* in red grouse. II. Population models. *Journal of Animal Ecology* **61**:487–498.

600 Duerr, H. P., K. Dietz, and M. Eichner, 2003. On the interpretation of age–intensity profiles and dispersion patterns in parasitological surveys. *Parasitology* **126**:87–101.

Fowler, A. C. and T. D. Hollingsworth, 2016. The dynamics of *Ascaris lumbricoides* infections.
603 *Bulletin of Mathematical Biology* **78**:815–833.

Frank, S. A., 2009. The common patterns of nature. *Journal of Evolutionary Biology* **22**:1563–1585.

Frank, S. A., 2014. Generative models versus underlying symmetries to explain biological pattern.
606 *Journal of Evolutionary Biology* **27**:1172–1178.

Gourbière, S., S. Morand, and D. Waxman, 2015. Fundamental factors determining the nature of parasite aggregation in hosts. *Plos One* **10**:1–17.

609 Grafen, A. and M. E. J. Woolhouse, 1993. Does the negative binomial distribution add up? *Parasitology Today* **9**:475–477.

612 Gear, D. A. and P. Hudson, 2011. The dynamics of macroparasite host-self-infection: a study of the patterns and processes of pinworm (*Oxyuridae*) aggregation. *Parasitology* **138**:619–27.

615 Grutter, A. and R. Poulin, 1998. Intraspecific and interspecific relationships between host size and the abundance of parasitic larval gnathiid isopods on coral reef fishes. *Marine Ecology Progress Series* **164**:263–271.

Haegeman, B. and R. S. Etienne, 2010. Entropy maximization and the spatial distribution of species. *The American Naturalist* **175**:E74–90.

618 Haegeman, B. and M. Loreau, 2009. Trivial and non-trivial applications of entropy maximization in ecology: A reply to Shipley. *Oikos* **118**:1270–1278.

621 Harte, J., 2011. Maximum Entropy and Ecology: A Theory of Abundance, Distribution, and Energetics. Oxford University Press, Oxford, United Kingdom.

Harte, J. and E. A. Newman, 2014. Maximum information entropy: a foundation for ecological theory. *Trends in Ecology & Evolution* **29**:384–389.

624 Harte, J., A. Rominger, and W. Zhang, 2015. Integrating macroecological metrics and community taxonomic structure. *Ecology Letters* **18**:1068–1077.

627 Hartson, R. B., S. A. Orlofske, V. E. Melin, R. T. Dillon, and P. T. J. Johnson, 2011. Land use and wetland spatial position jointly determine amphibian parasite communities. *EcoHealth* **8**:485–500.

630 Hudson, P. J., D. Newborn, and A. P. Dobson, 1992. Regulation and stability of a free-living host-parasite system: *Trichostrongylus tenuis* in red grouse. 1. Monitoring and parasite reduction experiments. *Journal of Animal Ecology* **61**:477–486.

633 Isham, V., 1995. Stochastic models of host-macroparasite interaction. *The Annals of Applied Probability* **5**:720–740.

James, G., D. Witten, T. Hastie, and R. Tibshirani, 2013. Introduction to Statistical Learning with Applications in R. Springer, New York, USA.

636 Johnson, P. T., 1999. The effect of trematode infection on amphibian limb development and survivorship. *Science* **284**:802–804.

639 Johnson, P. T. J. and V. J. McKenzie, 2008. Effects of Environmental Change on Helminth Infections in Amphibians: Exploring the Emergence of *Ribeiroia* and *Echinostoma* Infections in North America. In *The Biology of Echinostomes*, chapter 11, pages 249–280.

642 Johnson, P. T. J., D. L. Preston, J. T. Hoverman, and K. L. D. Richgels, 2013. Biodiversity decreases disease through predictable changes in host community competence. *Nature* **494**:230–233.

Johnson, P. T. J., J. R. Rohr, J. T. Hoverman, E. Kellermanns, J. Bowerman, and K. B. Lunde, 2012. Living fast and dying of infection: Host life history drives interspecific variation in infection and disease risk. *Ecology Letters* **15**:235–242.

Johnson, P. T. J., C. L. Wood, M. B. Joseph, D. L. Preston, S. E. Haas, and Y. P. Springer, 2016. Habitat heterogeneity drives the host-diversity-begets-parasite-diversity relationship: evidence from experimental and field studies. *Ecology Letters* **19**:752–761.

Krasnov, B. R., M. Stanko, D. Miklisova, and S. Morand, 2006. Host specificity, parasite community size and the relation between abundance and its variance. *Evolutionary Ecology* **20**:75–91.

Kretzschmar, M. and F. R. Alder, 1993. Aggregated distributions in models for patchy populations. *Theoretical Population Biology* **43**:1–30.

Leibold, M. A. and G. M. Mikkelsen, 2002. Coherence, species turnover, and boundary clumping: elements of meta-community structure. *Oikos* **97**:237–250.

Lester, R. J. G. and R. McVinish, 2016. Does moving up a food chain increase aggregation in parasites? *Journal of the Royal Society, Interface* **13**:1–11.

Lloyd-Smith, J. O., 2007. Maximum likelihood estimation of the negative binomial dispersion parameter for highly overdispersed data, with applications to infectious diseases. *PLoS ONE* **2**:1–8.

Locey, K. J. and D. J. McGlinn, 2013. Efficient algorithms for sampling feasible sets of macroecological patterns. *PeerJ* pages 1–23.

Locey, K. J. and E. P. White, 2013. How species richness and total abundance constrain the distribution of abundance. *Ecology Letters* **16**:1177–85.

McGill, B. J. and J. C. Nekola, 2010. Mechanisms in macroecology: AWOL or purloined letter? Towards a pragmatic view of mechanism. *Oikos* **119**:591–603.

Newman, E. N., M. E. Harte, N. Lowell, M. Wilber, and J. Harte, 2014. Empirical tests of within- and across species energetics in a diverse plant community. *Ecology* **95**:2815–2825.

Pacala, S. W. and A. P. Dobson, 1988. The relation between the number of parasites/host and host age: population dynamic causes and maximum likelihood estimation. *Parasitology* **96**:197–210.

Poulin, R., 2007. Are there general laws in parasite ecology? *Parasitology* **134**:763–76.

Poulin, R., 2013. Explaining variability in parasite aggregation levels among host samples. *Parasitology* **140**:541–6.

Pugliese, A., R. Rosà, and M. L. Damaggio, 1998. Analysis of model for macroparasitic infection with variable aggregation and clumped infections. *Journal of Mathematical Biology* **36**:419–47.

Raffel, T. R., J. O. Lloyd-Smith, S. K. Sessions, P. J. Hudson, and J. R. Rohr, 2011. Does the early frog catch the worm? Disentangling potential drivers of a parasite age–intensity relationship in tadpoles. *Oecologia* **165**:1031–1042.

Rosà, R. and A. Pugliese, 2002. Aggregation, stability, and oscillations in different models for host-macroparasite interactions. *Theoretical Population Biology* **61**:319–34.

- 681 Rosà, R., A. Pugliese, A. Villani, and A. Rizzoli, 2003. Individual-based vs. deterministic models for macroparasites: host cycles and extinction. *Theoretical Population Biology* **63**:295–307.
- Shaw, D. J. and A. P. Dobson, 1995. Patterns of macroparasite abundance and aggregation in wildlife populations: a quantitative review. *Parasitology* **111**:111–133.
- 684 Shaw, D. J., B. T. Grenfell, and A. P. Dobson, 1998. Patterns of macroparasite aggregation in wildlife host populations. *Parasitology* **117**:597–610.
- Supp, S. R., X. Xiao, K. M. Ernest, and E. P. White, 2012. An experimental test of the response of macroecological patterns to altered species interactions. *Ecology* **93**:2505–2511.
- 687 Tompkins, D. M., A. P. Dobson, P. Arneberg, M. Begon, I. M. Cattadori, J. V. Greenman, J. A. P. Heesterbeek, P. J. Hudson, D. Newborn, A. Pugliese, A. P. Rizzoli, R. Rosa, F. Rosso, and K. Wilson, 2002. Parasites and host population dynamics. In P. J. Hudson, A. Rizzoli, B. T. Grenfell, H. Heesterbeek, and A. P. Dobson, editors, *The Ecology of Wildlife Diseases*, chapter 3, pages 45–62. Oxford University Press, Oxford.
- 690 Ulrich, W. and N. J. Gotelli, 2013. Pattern detection in null model analysis. *Oikos* **122**:2–18.
- White, E. P., K. M. Thibault, and X. Xiao, 2012. Characterizing species abundance distributions across taxa and ecosystems using a simple maximum entropy model. *Ecology* **93**:1772–8.
- 696 Wilber, M. Q., S. B. Weinstein, and C. J. Briggs, 2016. Detecting and quantifying parasite-induced host mortality from intensity data: Method comparisons and limitations. *International Journal for Parasitology* **46**:59–66.
- 699 Wilson, K., O. N. Bjoernstad, A. P. Dobson, S. Merler, G. Poglayen, A. F. Read, and A. Skorpington, 2002. Heterogeneities in macroparasite infections: patterns and processes. In P. J. Hudson, A. Rizzoli, B. Grenfell, H. Heesterbeek, and A. Dobson, editors, *The Ecology of Wildlife Diseases*, chapter 2, pages 6–44. Oxford University Press, Oxford.
- 702 Xiao, X., K. J. Locey, and E. P. White, 2015a. A process-independent explanation for the general form of Taylor's law. *The American Naturalist* **186**:E51–E60.
- 705 Xiao, X., D. J. McGlinn, and E. P. White, 2015b. A strong test of the Maximum Entropy Theory of Ecology. *The American Naturalist* **185**:E70–80.
- Zillio, T. and F. He, 2010. Modeling spatial aggregation of finite populations. *Ecology* **91**:3698–3706.
- 708

Table 1: Definitions of terms used to describe the constraint-based null models.

Term	Definition
Labeled	Hosts or parasites are distinguishable
Unlabeled	Hosts or parasites are indistinguishable
Macrostate	Unordered vector of unlabeled parasite abundances. e.g. Given $P = 3$ and $H = 2$, the vector $\{3, 0\}$ is a macrostate
Configuration	Ordered vector of unlabeled parasite abundances. e.g. Given $P = 3$ and $H = 2$, the macrostate $\{3, 0\}$ has two configurations: $(3, 0)$ and $(0, 3)$
Feasible set	All possible macrostates given P and H . e.g. The feasible set given $P = 3$ and $H = 2$ is $\{\{3, 0\}, \{2, 1\}\}$
Weighted feasible set	Feasible set in which macrostates have particular weights.
Partition model	Weights each macrostate in the feasible set by assuming unlabeled hosts and parasites. All macrostates have equal weights.
Composition model	Weights each macrostate in the feasible set by assuming labeled hosts and unlabeled parasites. Analogously, each macrostate can be realized by multiple configurations.

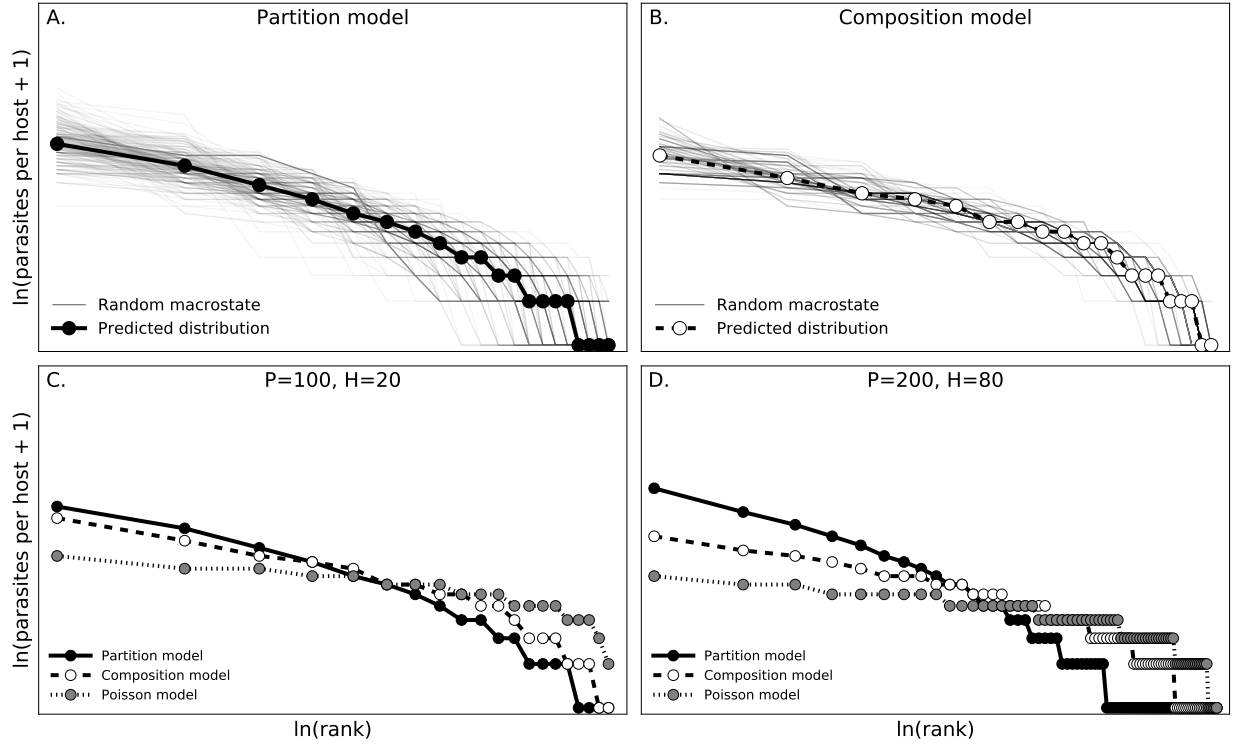


Figure 1: **A.** and **B.** show how the predictions from the partition model and composition model were generated. Random macrostates were drawn from the weighted feasible set (light gray lines) and the predicted distribution was computed as the central tendency of these randomly drawn weighted macrostates (thick lines with dots). Each dot represents a host in the predicted distribution with a given parasite abundance and rank. Hosts with low ranks (low $\ln(\text{ranks})$) have a larger number of parasites than hosts with high ranks (high $\ln(\text{ranks})$). **C.** and **D.** compare the partition and composition models for two different values of P and H . The more familiar Poisson model is also included for reference. The partition and composition models predict more aggregated host-parasite distributions than the Poisson model and the partition approach tends to produce more aggregated distributions than the composition model. The degree that the predictions from the partition and composition models differ depends on the values of P and H .

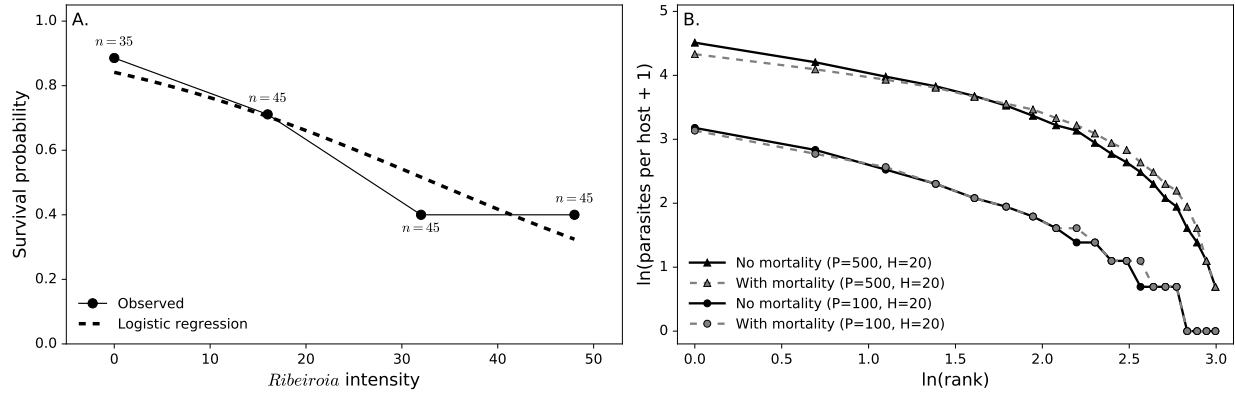


Figure 2: **A.** The black line and dots give the parasite-induced host mortality data from Johnson (1999) where *Pseudacris regilla* hosts were infected with varying *Ribeiroia* intensities. Frogs were exposed to *Ribeiroia* cercariae as tadpoles and the experiment was stopped after tadpoles metamorphosed. The dashed line gives the mean predictions of the logistic regression model fit to the data. This logistic regression was then imposed as a mortality constraint on the partition model and the composition model. **B.** An example of the effect of including the laboratory-estimated survival curve on the predictions of partition models with $P = 500$ parasites and $H = 20$ hosts and $P = 100$ and $H = 20$. The symbols indicate a given host in a predicted distribution with a particular parasite abundance and rank. Hosts with low ranks (low $\ln(\text{ranks})$) have a larger number of parasites than hosts with high ranks (high $\ln(\text{ranks})$). Depending on the values of P and H , the mortality constraint could noticeably reduce aggregation (triangles) or have little effect on aggregation (circles).

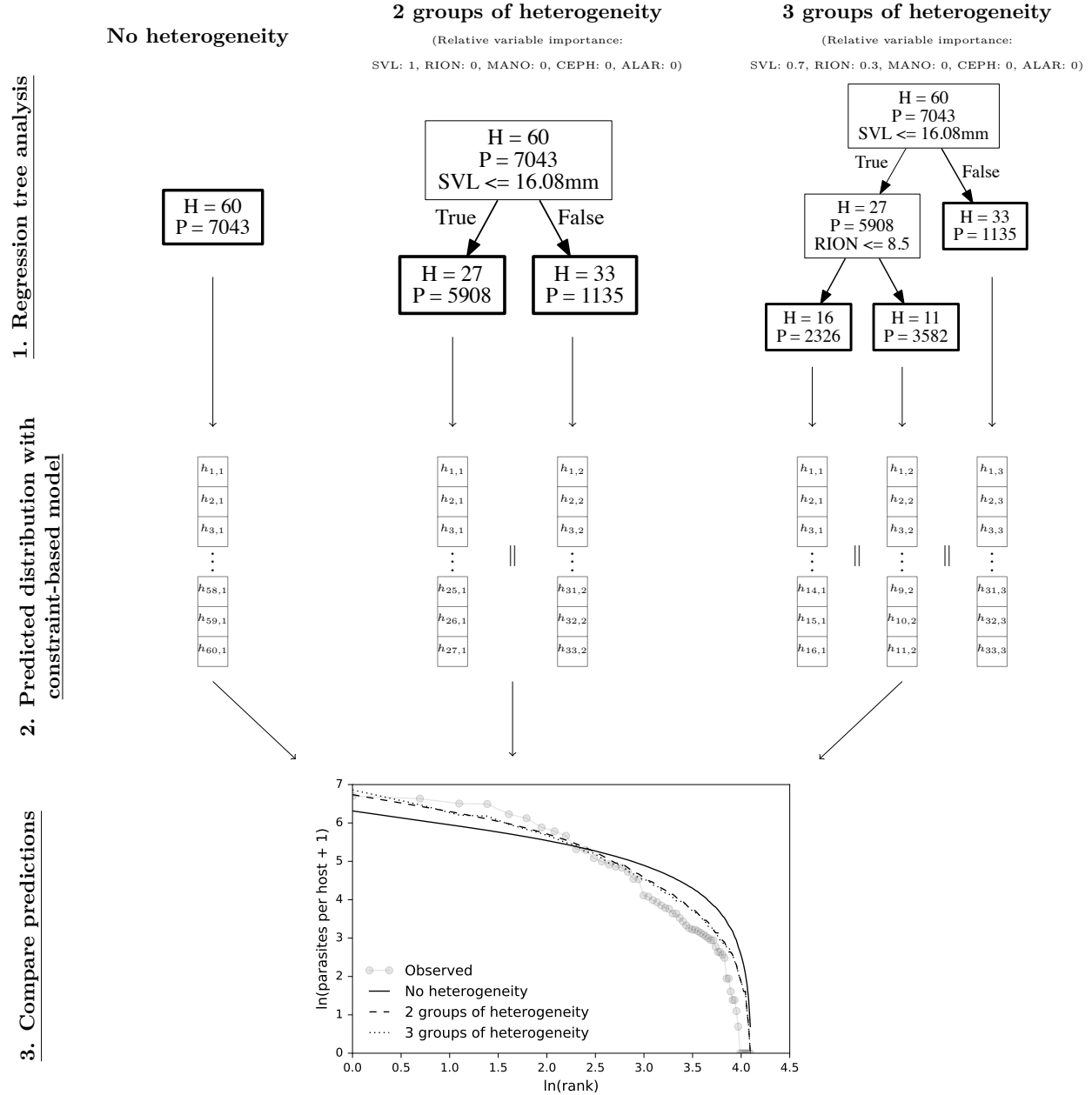


Figure 3: A diagram showing how host-heterogeneity can be incorporated into constraint-based models. **Step 1:** Consider, for example, a distribution for the parasite *Echinostoma* sp. in the host *Pseudacris regilla* with $H = 60$ hosts and $P = 7043$ parasites. When no host heterogeneity is included, the central tendency of the constraint-based model can be computed directly from H and P as described in the main text. To include groups of heterogeneity, a regression tree analysis is performed in which the response variable is *Echinostoma* abundance and the predictor variables are *P. regilla* body size (snout-vent length, SVL) and the abundance of *Ribeiroia ondatrae* (RION), *Alaria* sp. (ALAR), *Cephalogonimus* sp. (CEPH), and *Manodistomum* sp. (MANO) in a particular host. In the example above, the regression tree analysis shows that the “best” way to make two groups of heterogeneity given the predictor variables is to split the 60 *P. regilla* individuals into those with $SVL \leq 16.08$ mm and those with $SVL > 16.08$ mm. To make three groups of heterogeneity, *P. regilla* individuals with $SVL \leq 16.08$ are again split into individuals with $RION \leq 8.5$. For each of these regression trees, we can determine the relative importance of each variable in building the regression tree by how much they decrease the sum of squared error compared to the other predictors. **Step 2:** We can then compute the central tendency of the constraint-based model for each of these groups of heterogeneity (the bold boxes above) using the total number of hosts and parasites in each heterogeneity group. Each heterogeneity group has its own rank abundance distribution with $h_{i,j}$ being the i th ranked host with some number of parasites in heterogeneity group j . Concatenating (||) these rank abundance distributions together and re-ordering the resulting vector gives the predicted constraint-based model after allowing for P and H to vary with host heterogeneity. **Step 3:** These predicted distributions can then be compared to the observed host-parasite distribution.

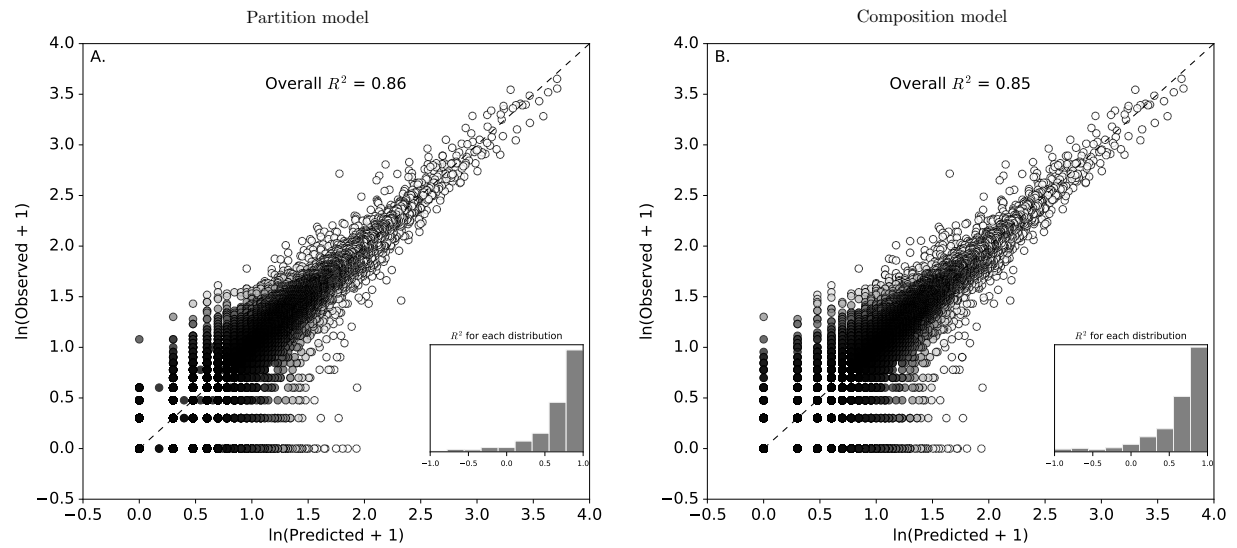


Figure 4: Plots showing the fit of the partition model **A.** and the composition model **B.** to all of the 842 observed host-parasite distributions considered in this study. The black-dashed line gives the 1:1 line and the overall R^2 gives the percent of variation that the constraint-based models explain in all of the observed data. Each point represents a single individual host from one of the 842 distributions with a given predicted and observed parasite abundance. Darker colors indicate a higher density of points in the region than lighter colors. The inset histogram shows the distribution of R^2 values calculated for each of the 842 individual host-parasite distributions.

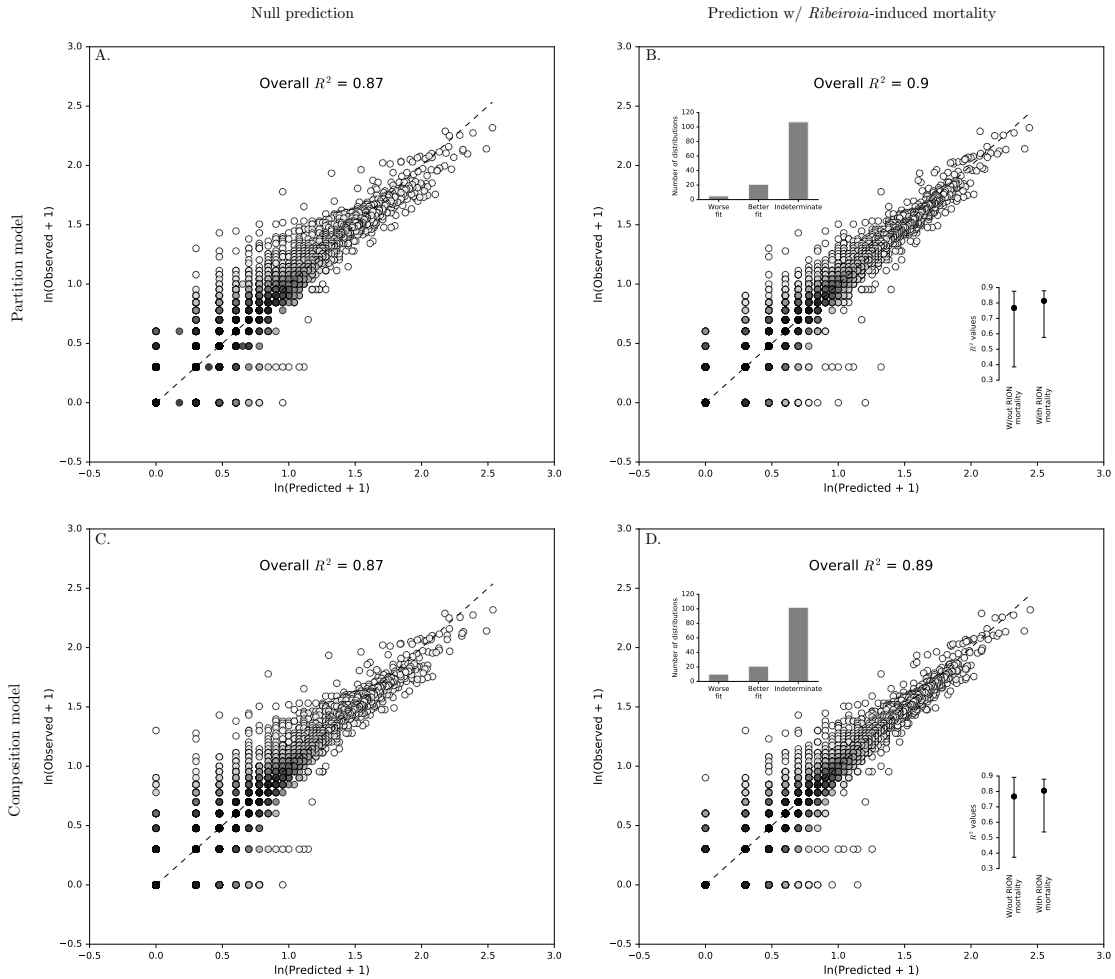


Figure 5: The effect of including empirically-estimated *Ribeiroia*-induced *Pseudacris regilla* mortality into the partition and composition models. The first column in this plot (A., C.) compares 133 observed rank abundance distributions (RAD) of *Ribeiroia*-*P. regilla* with the RADs predicted by the constraint-based models before they were constrained on parasite-induced host mortality. The second column (B., D.) compares the observed and predicted RADs after they were constrained on parasite-induced host mortality. The 1:1 line is given by the black, dashed line. Each point represents a single host with a given predicted and observed parasite abundance. Darker colors indicate a higher density of points in the region than lighter colors. The inset dot plots in B. and D. show the first through third quartiles of the R^2 values from each of the 133 distributions without and with *Ribeiroia*-induced mortality. The inset bar plots in B. and D. give the number of observed distributions that were a worse fit with the mortality constraint, a better fit with the mortality constraint, or were equally good with either (Indeterminate) based on $\Delta AICc$.

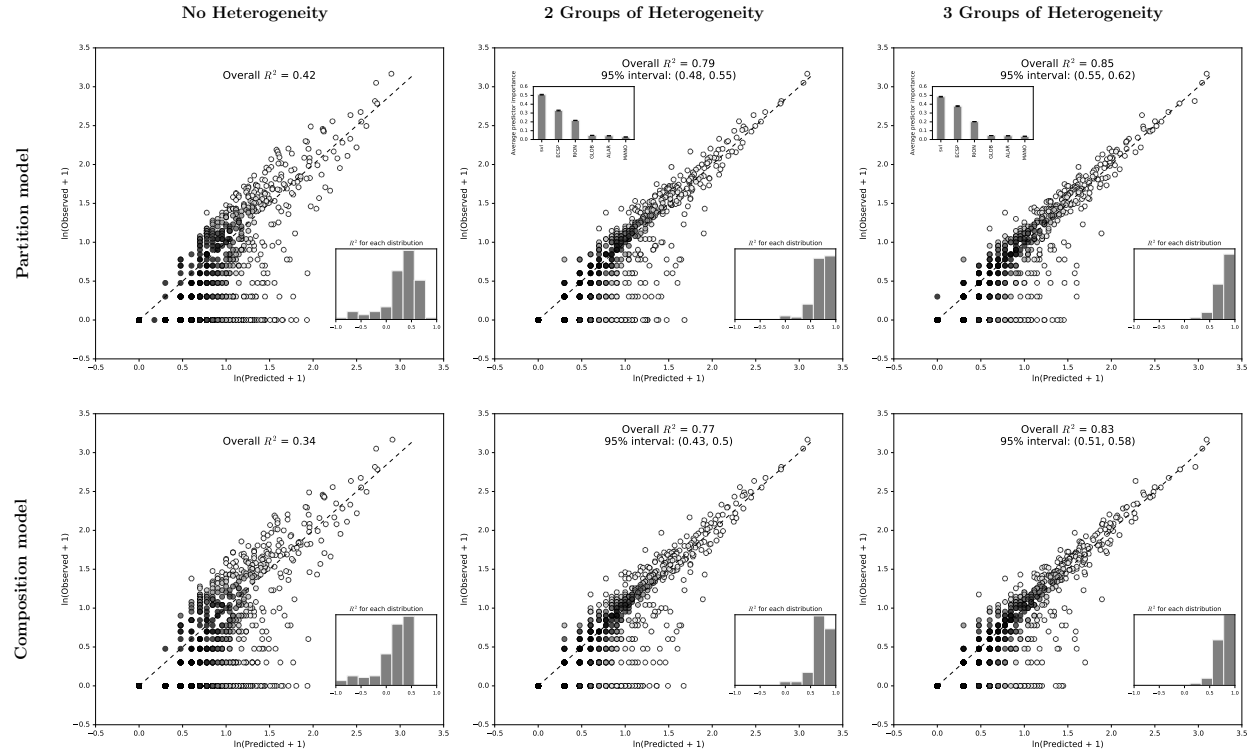


Figure 6: The effect of discrete heterogeneity on the 124 host-parasite distributions that were overaggregated relative to at least one of the constraint-based models (all hosts and parasites shown together). The first column in this plot shows the predicted rank abundance distributions (RAD) compared to the observed RADs when no host heterogeneity was included in either of the two constraint-based models. The black, dashed line gives the 1:1 line and the overall R^2 describes the amount of variation the constraint-based models described in all 124 overaggregated distributions. Each point represents a single host with a given predicted and observed parasite abundance. Darker colors indicate a higher density of points in the region than lighter colors. The histogram in the lower right hand side gives the distribution of R^2 values for each particular host-parasite distribution. The second and third columns in this plot show the effect of adding 2 and 3 groups of host heterogeneity, respectively, on the predicted host-parasite distributions based on the results from a regression tree analysis on known host attributes in the dataset. The plots in the upper left hand corner show the mean importance of a given host attribute in structuring the regression tree for all the 124 overaggregated host-parasite distributions. The predictor variables were body-size (svl), *Echinostoma* sp. (ECSP), *Ribeiroia* (RION), *Cephalogonimus* (CEPH), *Alaria* (ALAR), and *Manodistomum* (MANO). The predictor importance was the same for all models within a heterogeneity group and are therefore only displayed once for each group. Finally, the 95% interval displayed in the plot gives the 95% quantiles of overall R^2 values based on randomly permuting parasites into the groups predicted by the regression tree analysis. If the overall R^2 is greater than the interval, it shows that the increase in R^2 from the regression tree is a result of the predictors used in the regression tree analysis, rather than just grouping itself.

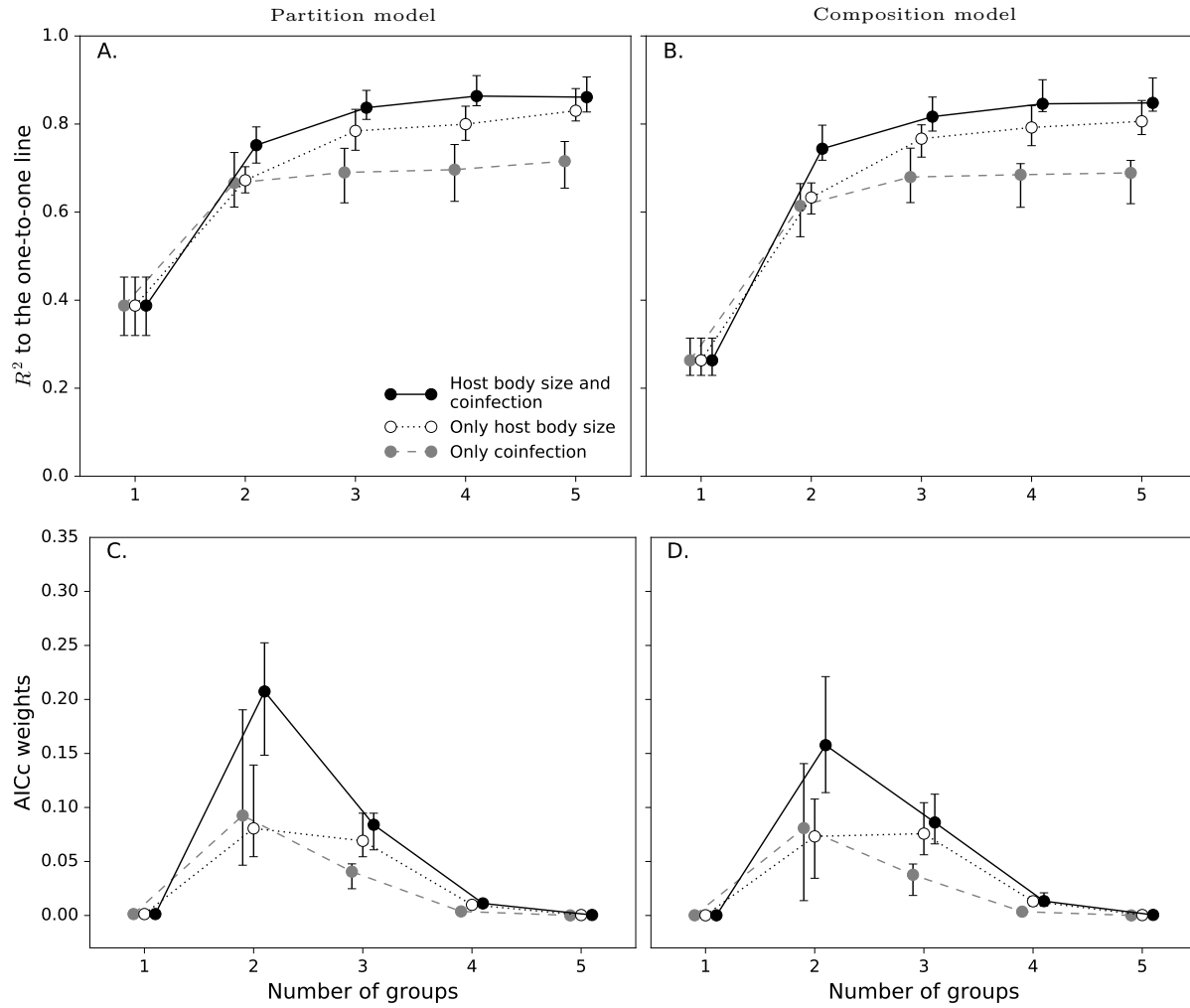


Figure 7: Plots show the effects of adding groups of heterogeneity to the constraint-based null models on two metrics: R^2 and AICc weights. Three heterogeneity models with different predictor variables were considered: only host body size, only coinfection with other trematodes, and both body size and coinfection with other trematodes. **A.** and **B.** show how the median R^2 of the 124 overaggregated distributions changes for the partition model (**A.**) and the composition model (**B.**) when heterogeneity in body size and and/or coinfection with other trematodes was considered. The points represent the median R^2 and the error bars give the approximate 95% confidence interval around the median. **C.** and **D.** show the median AICc weights for all 124 distributions for a given group size and heterogeneity model. For example, for the partition model every observed host-parasite distribution had 13 candidate models: the no-heterogeneity model (1 group) and three heterogeneity models times four groupings (2 groups, 3 groups, 4 groups, 5 groups). The AICc weights were calculated for these 13 models for a single distribution and then the median AICc weights for a given model was computed for all 124 overaggregated distributions. The error bars give the approximate 95% confidence intervals around these medians. These AICc weights are only comparing within a constraint-based model and are not comparing the partition model (**C.**) to the composition model (**D.**).