Appendices

Appendix A: The relationship between maximum entropy and feasible set predictions

The different predictions given by the maximum entropy and feasible set constraint-based models can be shown to be a result of the prior weights on the possible configurations for each approach. For example, take a system with $H = 3$ hosts and $P = 3$ parasites. Assume that the only constraints on the system are that total number of parasites in any configuration is 3 and the mean number of parasite per host in any configuration is $3/3 = 1$. In the terminology of Haegeman and Etienne (2010), these are hard constraints on the system.

We begin with the maximum entropy approach following the steps and terminology of Haegeman and Etienne (2010). We first specify that hosts are labeled such that we can distinguish between host 1, host 2, and host 3, but parasites are unlabeled such that we can not differentiate individual parasites within a host. Given this and the hard constraints specified above, we can enumerate all the possible configurations of this system (Table 1). There are a total of 10 possible configurations given these constraints, all with an equal probability of occurring. For the ordered configurations, $(3, 0, 0)$ has a 3 / 10 probability of occurring, $(2, 1, 0)$ has a 6 / 10 probability of occurring, and $(1, 1, 1)$ has a 1 / 10 probability of occurring. The most common ordered configuration gives the maximum entropy solution: in this case it is $(2, 1, 0)$.

The feasible set approach proceeds similarly, but it begins by assuming that both hosts and parasites are unlabeled. For the problem defined above, we can enumerate all possible configurations of the system based on the feasible set assumption (Table 2). There are three possible configurations and all configurations have an equal weight of 1 / 3. In this case, because all ordered configurations are equally probable we could get the predicted configuration by finding the center of the feasible set. This could be done by finding the mean, median or mode of the given feasible set. While there are advantages and disadvantages to each of these methods of

centering, we prefer the median over the mode as it is not as sensitive to sample size as the mode and eliminates the subjectivity of which value to chose if there are multiple modes in a finite sample. We prefer the median over the mean as it is less sensitive to skew in the distribution. Using the median, the predicted center of the feasible set is $(2, 1, 0)$.

While the maximum entropy approach and the feasible set approach give the same predictions in this case, that is not generally true. However, one can see that the maximum entropy predictions could be recovered with a feasible set approach by weighting each feasible configuration with the probabilities predicted by the maximum entropy approach. In other words, $(3, 0, 0)$ is given a weight of 3 / 10, $(2, 1, 0)$ is given a weight of 6 / 10, and $(1, 1, 1)$ is given a weight of 1 / 10. The median of this weighted feasible set gives the maximum entropy prediction.

This conversion of the feasible set predictions into the maximum entropy predictions can be generalized to any system with P parasites and H hosts as follows. First, draw a system configuration from an equally weighted feasible set with P parasites and H hosts. Second, calculate the maximum entropy weight of this configuration. This can by done by first noting that the total number of configurations with P unlabeled parasites and H labeled hosts is given by (Harte, 2011)

$$D = \frac{(H + P - 1)!}{P!(H - 1)!} \tag{1}$$

For a given feasible set configuration, the total number of ways this configuration can be realized given unlabeled parasites and labeled hosts is

$$b = \frac{H!}{\Pi_{i \in A} h_i!} \tag{2}$$

where A is a set containing the unique parasite loads found in a particular feasible set configuration, i is a particular member of that set, and h_i is the number of hosts in the feasible configuration that have parasite load i. Note that $\sum_{i \in A} h_i = H$. The maximum entropy weight of that configuration is then given by $p = \frac{b}{D}$. Third, reject the given feasible set configuration

with probability $1 - p$. Repeat this procedure until the desired number of configurations have been drawn. The resulting set of configurations will represent a random sample from all possible system configurations under the aforementioned maximum entropy assumption. Adjusting the weighting probabilities allows one to sample any desired weighted feasible set, such as the parasite-induced mortality feasible sets discussed in Appendix B. Moreover, one can also see that the binomial model, which is the finite equivalent to the commonly used Poisson distribution in disease ecology, can be obtained by weighting each configuration by the corresponding multinomial coefficient (Appendix B; Haegeman and Etienne, 2010).

As a computational aside, the rejection algorithm described above will reject a large number of proposed configurations as P and H increase, so a more efficient alternative is needed to sample from weighted feasible sets. The Metropolis-Hastings algorithm is one scalable solution and we implement both this algorithm and the rejection algorithm for sampling from weighted feasible sets in the Python code that accompanies this manuscript (all code can be found at `https://github.com/mqwilber/feasible_parasites`).

Appendix B: Extending the constraint-based models to include parasite-induced host mortality

Ribeiroia has well-documented negative effects on amphibian survival in the lab and in the field (Johnson, 1999; Johnson et al., 2012). Therefore, we might expect that incorporating the effect of *Ribeiroia*-induced host mortality as an additional constraint on a predicted host-*Ribeiroia* distribution will improve the overall fit of a given top-down model to the observed parasite distribution.

To account for *Ribeiroia*-induced mortality, we use the data from the laboratory infection experiments described in Johnson (1999) and Johnson et al. (2012) to estimate an intensity-dependent survival curve for *Pseudacris regilla* infected with *Ribeiroia*. We use a standard logistic

survival curve given by

$$\text{logit}(p(x)) = a + bx \tag{3}$$

where logit is the logistic function, $p(x)$ is the probability of amphibian survival given a *Ribeiroia* intensity of x, b is the effect of *Ribeiroia* intensity on the log-odds of amphibian survival, and a is the "threshold" at which the host begins to experience parasite-induced mortality. Using a generalized linear model with a binomial response and a logistic link, we estimated the parameters of the *P. regilla-Ribeiroia* survival curve to be $a = 1.67$ and $b = -0.05$ (see the file `manuscript_analysis_parasite_mortality.py` for the data used to fit this GLM).

Using this estimated survival curve, we implemented the approach described in Appendix 1 to draw a weighted feasible set that accounted for the additional constraint of parasite-induced host mortality. We did this using the following Metropolis-Hastings algorithm:

1. Calculate the total number of *Ribeiroia* parasites P and *Pseudacris* hosts H in given empirical host-parasite distribution.

2. Draw an initial candidate feasible set with P and H using the algorithms provided by Locey and McGlinn (2013).

3. For the candidate feasible set, calculate the likelihood of observing this feasible set given the host-survival curve described above (Equation 3). To do this, we assumed that each host in a configuration was independent and calculated the likelihood of observing the configuration by multiplying the probabilities of observing each host with a given load. The assumption of independence is conditional on observing the configuration, not deriving the configuration where each host is inherently non-independent given that the total number of parasites in the system is fixed.

4. Propose a new feasible set configuration and calculate its likelihood from equation 3. The proposal distribution for drawing a new configuration is symmetric due to fact that the

basic feasible set model assumes that each configuration is equally likely (Locey and White, 2013).

5. Take the ratio, r, of the proposed likelihood over the candidate likelihood. If r is greater than 1, accept the proposed configuration. Otherwise, accept the proposed configuration with probability r and accept the candidate configuration with probability $1 - r$.

6. Set the accepted configuration as your new candidate configuration and repeat steps 3-5 a large number of times. Discard the first half of the iterations as warm-up/burn-in samples.

The remaining samples give the feasible set with the additional constraint of parasite-induced mortality.

To sample from a maximum entropy distribution with parasite-induced host mortality we used the same procedure described above, but in addition to assigning each proposed configuration a likelihood based on the survival function, we also assign each proposed configuration a likelihood based on equation 2. This amounts to multiplying the two likelihoods. As above, the likelihood ratio of the proposed configuration and the candidate configuration determines whether to accept or reject the proposed configuration.

We could also sample from a binomial distribution with a parasite-induced mortality constraint by changing our proposal distribution to a multinomial distribution where the probability of any one of the H hosts encountering a parasite is $1/H$ (Haegeman and Etienne, 2010). The multinomial distribution from which we propose a new configuration X is then given by

$$X \sim \text{Multinomial}(P, p_1 = \frac{1}{H}, p_2 = \frac{1}{H}, \ldots, p_H = \frac{1}{H}) \tag{4}$$

We can draw proposal configurations from this multinomial model and, as described above, assign them a likelihood based on both 1) their likelihood given by the estimated survival function and 2) their probability under the multinomial model. Then we accept or reject our proposed configuration based on the ratio of the likelihoods for the proposed and candidate configuration

times the probability ratio of the candidate and proposed configurations under the multinomial model. This is the additional weighting on the acceptance ratio imposed by the Metropolis-Hastings algorithm. In summary, this is a long-winded way of saying that if the survival function likelihood is 1 and there is no effect of parasite mortality, the algorithm will sample from a multinomial distribution whose predicted, ordered rank abundance distribution is equal to the predicted RAD from a binomial distribution with P parasites and H hosts. These algorithms are implemented and tested in the accompanying code.

We applied the algorithms and *P. regilla-Ribeiroia* survival curve to all 133 *P. regilla-Ribeiroia* distributions in the dataset. For each constraint-based model with parasite mortality, we ran the Metropolis-Hastings algorithm for 2000 iterations, discarding the first 1000 iterations as warm-up/burn-in samples. We ran this analysis multiple times from different random starting points to ensure the chains were converging to the same stationary distribution. In general, visual inspection of the trace plots of the mortality-constrained feasible set chains showed consistent convergence, good mixing, and generally had acceptance rates above 50%. This high acceptance rate was expected as these chains were designed to have an acceptance rate of 1 (i.e. sampling from the unconstrained distribution) if parasite-induced mortality was not important. The chains of the constrained maximum entropy model had an average acceptance rate of 0.34 and the majority of the chains showed good mixing. However, 9 of the 133 constrained maximum entropy chains had acceptance rates of less than 10% and showed high autocorrelation between samples. Both excluding the distributions resulting from these chains from the analysis and running the chains for longer had no effect on the conclusions we drew about parasite-induced mortality improving the fit of the maximum entropy model. Moreover, these chains were not problematic in the constrained feasible set model for which we also concluded that parasite-induced host mortality improved the fit of the top-down model. Because these under-sampled chains did not affect our inference or conclusions, we included the distributions resulting from these chains in the analysis presented in the main paper.

Appendix C: Randomization test for heterogeneity

To test whether the host-parasite distribution predicted by the regression tree was better than a host-parasite distribution generated by a randomly grouping hosts, we randomly permuted hosts with their corresponding parasite intensities into groups with the same number of H_j as predicted by the regression tree analysis. We then calculated the total number of parasites in each group j and used the procedure described in Figure 1 of the main text to generate the predicted RAD for these randomly permuted mixture distributions. More specifically, this is equivalent to finding the rank abundance distribution for the mixture model $g(x) = \sum_{j=1}^{G} H_i / Hf(x, \frac{P_{j,rand}}{H_j})$ where $P_{j,rand}$ indicates that the total number of parasites in each group j varies with each random permutation. We repeated this randomization 200 times for every empirical host-parasite distribution for both the feasible set and maximum entropy models. This generated 400 permuted host-parasite distributions (200 using the feasible set model and 200 using the maximum entropy model) for every observed host-parasite distribution. We could then use these permuted samples to determine whether the host groupings produced by the regression tree analysis improved the fit of the constraint-based models to the empirical distributions more than we would expect by randomly permuting hosts into groups. If the mixture model from the regression tree provided a significantly better fit to the host-parasite distribution then randomly grouping hosts, we expected its R^2 value to be significantly higher than the upper bound of the 95% quantile of the R^2s from the randomly generated groupings.

References

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

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

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

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.

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.

Table 1: All of the possible configurations of $P = 3$ unlabeled parasites among $H = 3$ labeled hosts. This corresponds to the maximum entropy assumption.

#	host 1	host 2	host 3
1.	3	0	0
2.	0	3	0
3.	0	0	3
4.	2	1	0
5.	2	0	1
6.	1	2	0
7.	1	0	2
8.	0	2	1
9.	0	1	2
10.	1	1	1

Table 2: All of the possible configurations of $P = 3$ unlabeled parasites among $H = 3$ unlabeled hosts. This corresponds to the feasible set assumption.

#	rank 1	rank 2	rank 3
1.	3	0	0
2.	2	1	0
3.	1	1	1

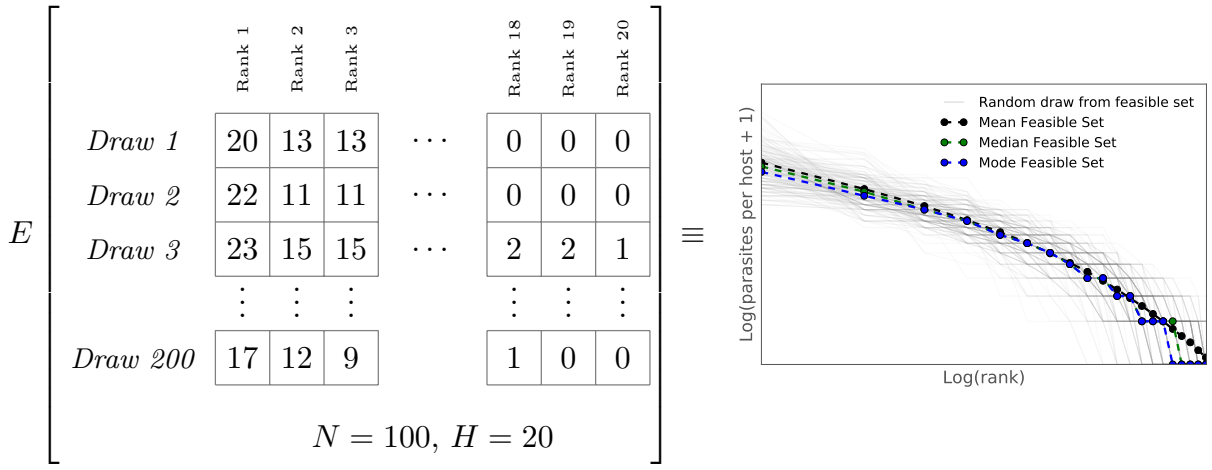

Figure 1: Given a host-parasite system $P = 100$ parasites and $H = 20$ hosts, the feasible set can be approximated by drawing some number of random configurations from the full feasible set (200 in this example) and ranking the hosts in each drawn configuration where the host with the most parasites has a rank of 1 and the host with the fewest individuals has a rank of H. The plot shows the graphical representation of this procedure where each gray line is a sampled configuration from the feasible set and the dashed lines are different measures of the center of the sampled feasible set.

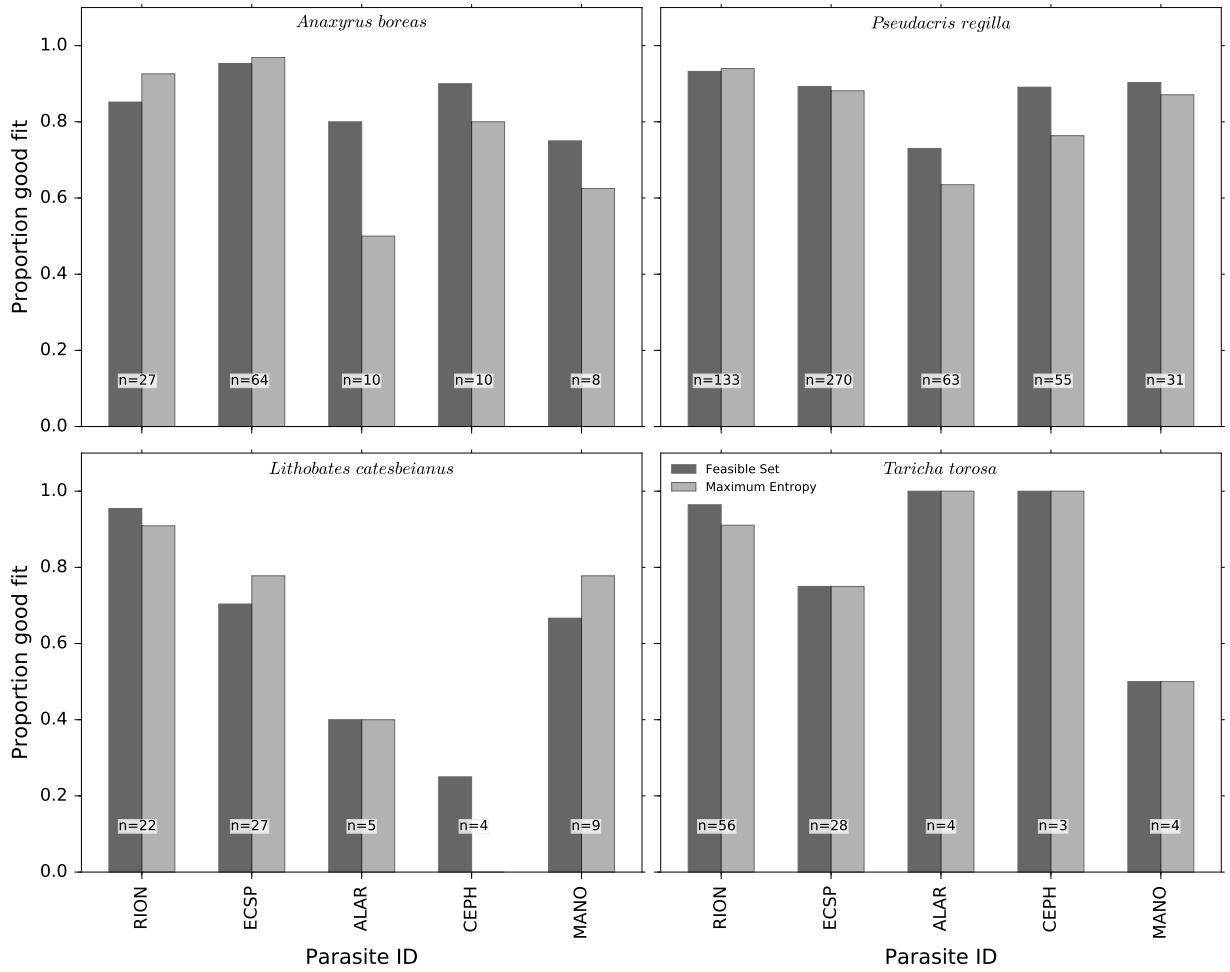

Figure 2: The proportion of predicted host-parasite distributions for either top-down (feasible set and maximum entropy) or bottom-up (Poisson) models that show a good fit to empirical host-parasite distributions. A predicted distribution was considered a good fit when the *p*-value from and Anderson-Darling test comparing the predicted and observed distribution was greater than 0.1. The number of distributions compared for any given host-parasite combination are also displayed on the figure. The x-axis gives the 5 trematode parasites examined in this analysis: *Ribeiroia ondatrae* (RION), *Echinostoma* sp. (ECSP), *Alaria* sp. (ALAR), *Cephalogonimus* sp. (CEPH), and *Manodistomum* sp. (MANO). *Taricha granulosa* is not shown in this plot as it was never infected with ALAR, CEPH or MANO.

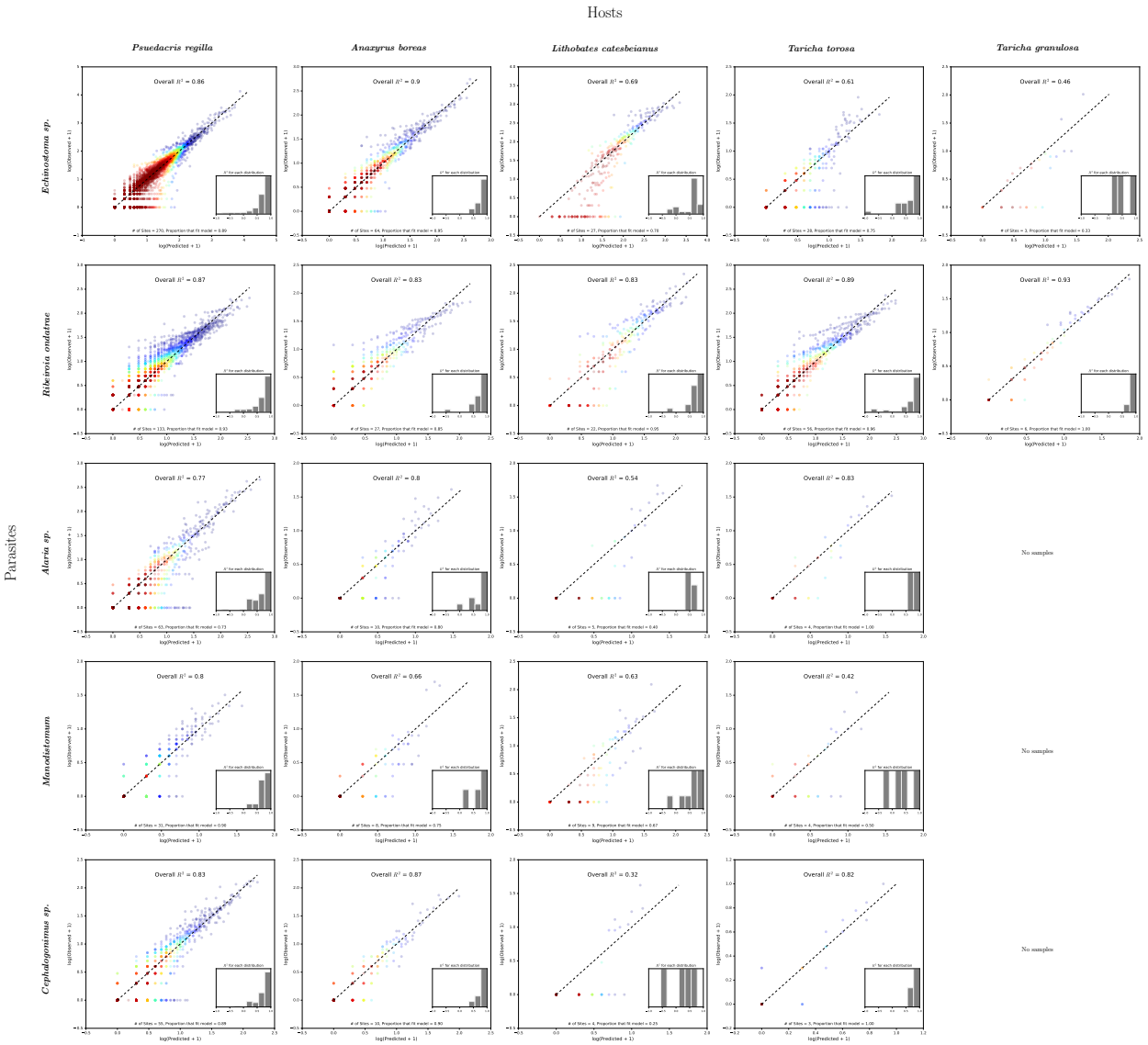

Figure 3: Comparison of observed and predicted host-parasite distributions from the feasible set model (top-down) for each host-parasite combination. In each subplot, the x-axis gives the model predicted parasite intensity and the y-axis given the observed parasite intensity. Each point gives a particular host's predicted and observed parasite intensity. The color of the points represent the density of points in that region. "Hotter" colors mean there are more points in that region while "cooler" colors mean there are less points in that region. The dashed black line gives the 1:1 line, along which we would expect the points to fall if the model was a perfect fit. The overall R^2 gives the measure of how well all the data fit the 1:1 line. The text at the bottom of the plot gives the total number of distributions that were tested and the proportion of them that fit the model (i.e. the number of distributions with a p-value from the Anderson Darling test that is greater than 0.1). Finally, the histogram in the lower right hand corner of each plot gives the histogram of the R^2 values for each of the specified host-parasite distributions.

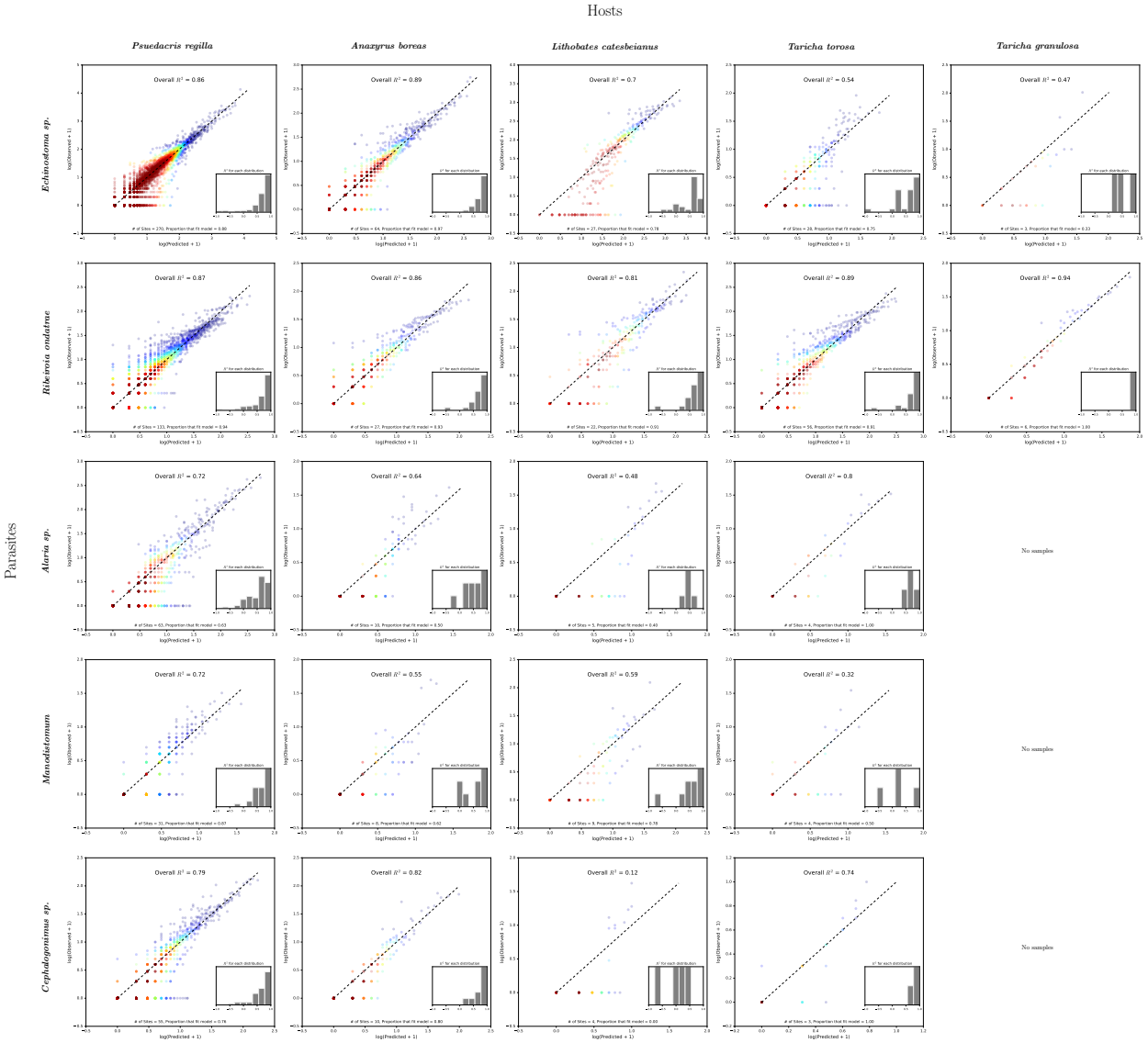

Figure 4: Comparison of observed and predicted host-parasite distributions from the maximum entropy model (top-down) for each host-parasite combination. In each subplot, the x-axis gives the model predicted parasite intensity and the y-axis given the observed parasite intensity. Each point gives a particular host's predicted and observed parasite intensity. The color of the points represent the density of points in that region. "Hotter" colors mean there are more points in that region while "cooler" colors mean there are less points in that region. The dashed black line gives the 1:1 line, along which we would expect the points to fall if the model was a perfect fit. The overall R^2 gives the measure of how well all the data fit the 1:1 line. The text at the bottom of the plot gives the total number of distributions that were tested and the proportion of them that fit the model (i.e. the number of distributions with a p-value from the Anderson Darling test that is greater than 0.1). Finally, the histogram in the lower right hand corner of each plot gives the histogram of the R^2 values for each of the specified host-parasite distributions.

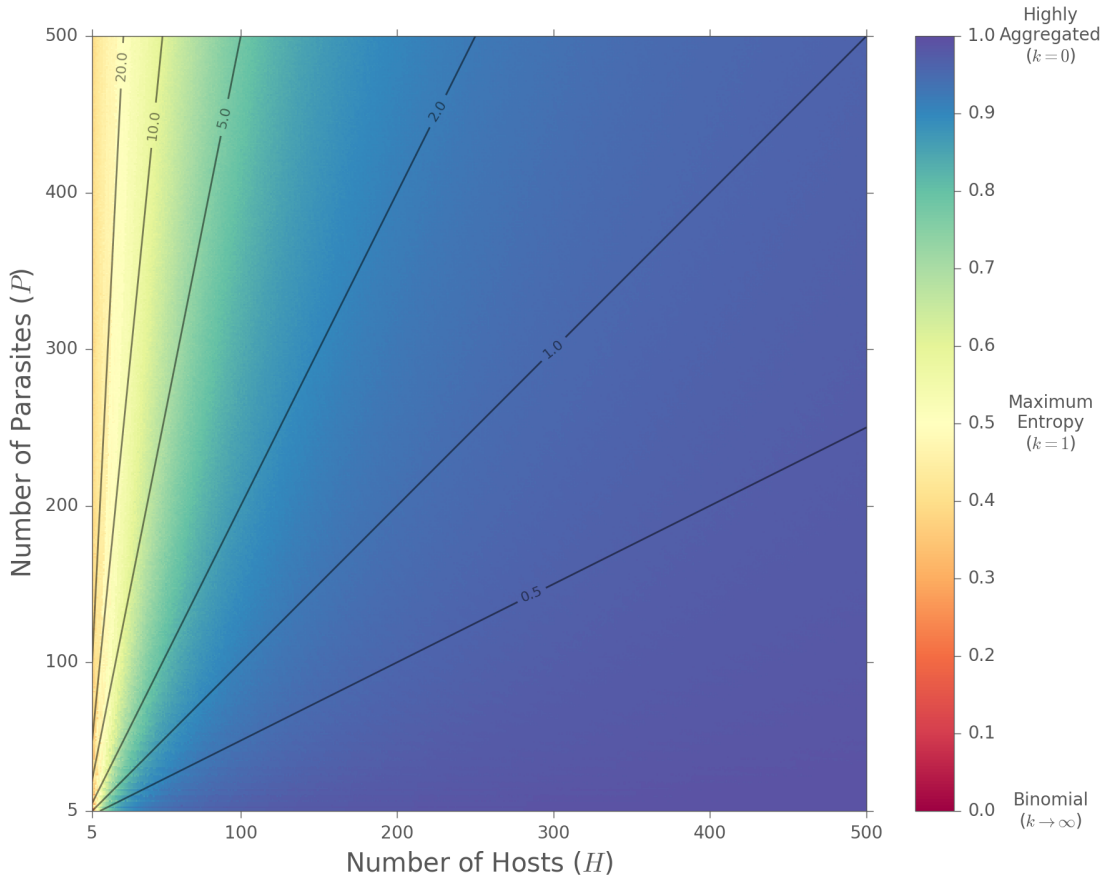

Figure 5: This plot shows the feasible set predictions of parasite aggregation over values of H (hosts) and P (parasites) from 5 - 500. For each combination of H and P, 1000 random samples were taken from the feasible set using the algorithms provided by Locey and McGlinn (2013) and the median of the sampled feasible set was computed as the feasible set prediction. For each feasible set prediction, the maximum likelihood estimate of the k parameter of a negative binomial was computed. The above plot displays the results in terms of the transformation $1/(1+k)$, where a value of 1 corresponds to a highly aggregated distribution $k \to 0$, 0.5 corresponds to $k = 1$, and 0 corresponds to a binomial distribution ($k \to \infty$). The contour lines given on the plot show the mean number of parasites in a host for the various combinations of P and H. The feasible set approach predicts substantial aggregation across nearly all combinations of P and H (k typically less than 2). Moreover it predicts that k increases with increasing mean parasites per host, but that the nature of this increase depends on P and H.