# Taxonomic resolution affects host-parasite association model performance: Supplemental Material

## Variable importance across geographic scales

Variables important in the full model were also important at the scale of the United State of America and the state of Texas (Figure [S1](#fig:scale)). This means tha the same variables used for host-parasite link prediction at global scales were important at quite small spatial scale. Further, this was not strongly influenced by helminth taxonomic subset (Figure [S1](#fig:scale)).



Figure S1. Variable importance values for helminth taxonomic subset models were strongly related across geographic scale considered, with Pearson’s correlation coefficients nearing one to the global model. This suggests that the same covariates important at global scales were important at smaller geographic scales.

## Host and helminth parasite covariates data used in the analyses

Table S1. Host covariates considered in the models of host-helminth associations, obtained from the Pantheria database of life history traits. Variable descriptions below are copied over from the metadata of the Pantheria database, included here only for reference.

|  |  |  |
| --- | --- | --- |
| Variable | Units | Definition |
| Host Family | – | Host taxonomic Family |
| Host Order | – | Host taxonomic Order |
| Litter size | no. | Number of offspring born per litter per female, either counted before birth, at birth or after birth, using captive, wild, provisioned, or unspecified populations; male, female, or sex unspecified individuals; primary, secondary, or extrapolated sources; all measures of central tendency; in all localities. |
| Geographic range size | *km*2 | Total area occupied by species |
| Max latitude | UTM | Maximum latitude the species has been observed |
| Min latitude | UTM | Minimum latitude the species has been observed |
| Mid-latitudinal range | UTM | The median latitude the species has been observed |
| Max longitude | UTM | Maximum latitude the species has been observed |
| Min longitude | UTM | Minimum latitude the species has been observed |
| Mid-longitudinal range | UTM | The median latitude the species has been observed |
| Min human population density | min *n* per *km*2 | minimum human population density using the Gridded Population of the World (GPW) (CIESIN and CIAT 2005) for 1995 |
| Human population density | mean *n* per *km*2 | mean human population density using the Gridded Population of the World (GPW) (CIESIN and CIAT 2005) for 1995 |
| Human population density 5p | *n* per *km*2 | 5th percentile human population density using the Gridded Population of the World (GPW) (CIESIN and CIAT 2005) for 1995 |
| Human population density change | – | mean rate of increase in human population density using the Gridded Population of the World (GPW) (CIESIN and CIAT 2005) for 1990 and 1995 |
| Mean preciptation | mm | mean monthly precipitation |
| Mean temperature | celsius | mean monthly temperature |
| Mean AET | mean mm | mean monthly AET (Actual Evapotranspiration Rate) from 1920 to 1980 calculated using the Global Resource Information Database of UNEP |
| Mean PET | mean mm | mean monthly PET (Potential Evapotranspiration Rate) from 1920 to 1980 (mm) calculated using the Global Resource Information Database of UNEP |
| Adult body mass | g | Mass of adult (or age unspecified) live or freshly-killed specimens (excluding pregnant females) using captive, wild, provisioned, or unspecified populations; male, female, or sex unspecified individuals; primary, secondary, or extrapolated sources; all measures of central tendency; in all localities. |

(\*): more information on how extrapolation was performed in (Jones *et al*. 2009)

Table S2. Helminth parasite covariates considered in the models of host-helminth associations.

|  |  |  |
| --- | --- | --- |
| Variable | Units | Definition |
| Dominant infection site | – | first reported site of infection in definitive host |
| Secondary infection site | – | second reported site of infection in definitive host |
| Egg size (max) | *μ*m | maximum egg size in |
| Egg size (min) | *μ*m | minimum egg size |
| Female length (max) | mm | maximum length of female adults worms |
| Female length (min) | mm | minimum length of female adults worms |
| Female width (max) | *μ*m | maximum width of female adults worms |
| Female width (min) | *μ*m | minimum width of female adults worms |
| Length (max) | mm | maximum length of adult worms across male and female worms |
| Length (min) | mm | minimum length of adult worms across male and female worms |
| Male length (max) | mm | maximum length of male adults worms |
|  |  |  |
| Male length (min) | mm | minimum length of male adults worms |
|  |  |  |
| Male width (max) | *μ*m | maximum width of male adults worms |
| Male width (min) | *μ*m | minimum width of male adults worms |
| Parasite Class | – | taxonomic Class |
| Parasite Phylum | – | taxonomic Phylum |
| Width (max) | *μ*m | maximum width of adult worms across male and female worms |
| Width (min) | *μ*m | minimum width of adult worms across male and female worms |

## Models trained at smaller spatial scales

In the main text, we mainly focus on the global model, as estimating host-parasite interactions at global and local scales fundamentally differ in goal. Specifically, global models of host-parasite interactions do not consider dispersal limitation and geographic limitations of host and parasite, meaning that predicted associations are potentially impossible given the current distribution of host or parasite. Given the homogenization of host and parasite species, it seems prudent to predict these links anyways, and a focus on the global scale allowed us to leverage a larger amount of data to estimate associations. However, spatial scale is of fundamental importance, and researchers wanting to estimate host-parasite interactions at more regional scales might not need to consider taxonomic scale.

We found that model performance – measured in terms of discrimination and accuracy – did not differ between full model and helminth taxonomic submodels when models were built at the scale of the United States of America (Table S3) or the state of Texas (Table S4). Despite this lack of difference between models at these scales, the same host and helminth parasite covariates important at the global scale were important at the scale of the United States of America (Figure S2) and the state of Texas (Figure S3)

Table S3. Considering host-helminth interactions in the United States of America, model performance – quantified using AUC, accuracy, and TSS – did not differ between full model and helminth taxonomic subsets, based on Welch’s two-sample t-tests comparing model performance across the 50 trained boosted regression models.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Performance | Helminth taxa | *t* | df | *p* value |
| AUC | Acanthcephalans | 0.01 | 64 | 0.99 |
|  | Platyhelminthes | -0.03 | 52 | 0.74 |
|  | Nematodes | -0.87 | 52 | 0.39 |
|  |  |  |  |  |
| Accuracy | Acanthcephalans | 0.91 | 95 | 0.36 |
|  | Platyhelminthes | -0.03 | 80 | 0.97 |
|  | Nematodes | 0.26 | 85 | 0.79 |
|  |  |  |  |  |
| TSS | Acanthcephalans | 3.14 | 97 | **0.002** |
|  | Platyhelminthes | 1.83 | 98 | 0.07 |
|  | Nematodes | 0.12 | 96 | 0.91 |

Table S4. Considering host-helminth interactions in Texas, model performance – quantified using AUC, accuracy, and TSS – did not differ between full model and helminth taxonomic subsets, based on Welch’s two-sample t-tests comparing model performance across the 50 trained boosted regression models.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Performance | Helminth taxa | *t* | df | *p* value |
| AUC | Acanthcephalans | 0.14 | 42 | 0.89 |
|  | Platyhelminthes | -0.17 | 42 | 0.86 |
|  | Nematodes | -0.15 | 26 | 0.88 |
|  |  |  |  |  |
| Accuracy | Acanthcephalans | -0.58 | 80 | 0.57 |
|  | Platyhelminthes | -1.39 | 87 | 0.17 |
|  | Nematodes | 2.24 | 98 | **0.03** |
|  |  |  |  |  |
| TSS | Acanthcephalans | 0.22 | 83 | 0.83 |
|  | Platyhelminthes | 0.25 | 98 | 0.80 |
|  | Nematodes | -0.004 | 97 | 0.99 |



Figure S2. Variable importance for each model trained on host-helminth interactions in the United States of America – with helminth taxonomic group denoted by point color – tended to be conserved, with host family and the site of infection as dominant predictors across models (panel *a*; host variables are italicized, helminth parasite covariates are bolded; only the top 10 predictor variables are shown here). The rank order of mean variable importance tended to be positively correlated among models as well (panel *b*). Finally, while important variables tended to be the same across models, the relative importance of helminth parasite covariates (darker colors in the pie charts in panel *c*) compared to host covariates (lighter shaded regions) did show variation.



Figure S3. Variable importance for each model trained on host-helminth interactions in the US state of Texas – with helminth taxonomic group denoted by point color – tended to be conserved, with host family and the site of infection as dominant predictors across models (panel a; host variables are italicized, helminth parasite covariates are bolded; only the top 10 predictor variables are shown here). The rank order of mean variable importance tended to be positively correlated among models as well (panel b). Finally, while important variables tended to be the same across models, the relative importance of helminth parasite covariates (darker colors in the pie charts in panel c) compared to host covariates (lighter shaded regions) did show variation.

## Partial dependence plots

Our focus was largely on the variation in predictive performance of models trained on different taxonomic subsets of helminth parasites. But the directionality of the variable effects in each model are also important, as predictive covariates could have entirely different effects on host-parasite associations depending on parasite subset.

We do not find evidence of this, but instead that variable effects tend to be conserved across models (Figure S4). It is difficult to distinguish between the effects of different host families or infection sites due to the sheer number of feature levels, but the continuous variable corresponding to the maximum latitude where a host species is found was a consistent indicator of host-parasite associations, and the functional form of this relationship was conserved across different models.



Figure S4. Partial dependence plots of a representative GBM model considering all helminth groups (full model), or helminth taxonomic subsets (i.e., Acanthocephalans, Nematodes, and Platyhelminthes). Each row corresponds to a different helminth taxonomic subset, while each column corresponds to the top variables in the given boosted regression tree (GBM) model (1st column is the most important covariate, 2nd column is the second most important, and 3rd column is the third most important).