Evaluating the Impact of Weather on Salmonella Prevalence and Serovar Diversity

Connor Norris

**Authors**

* Connor Norris

**Author affiliations**

1. College of Public Health, University of Georgia, Athens, GA, USA.

Corresponding author: cgn18694@uga.edu

# 1. Introduction

*Salmonella enterica* is a very important cause of foodborne illness around the globe. In the United States alone, nontyphoidal *Salmonella* serotypes are estimated to cause 1.3 million illnesses and 12,500 hospitalizations annually (1). Many *Salmonella* outbreaks are commonly associated with contaminated meat and poultry. However, 44% of domestic *Salmonella* illnesses are attributed to produce consumption. It is possible that contaminated surface water used in irrigation may contribute to disseminating these foodborne pathogens (2,3). As such, having a better understanding the dynamics of *Salmonella enterica* in these water sources is important for continued improvement of public health management efforts against this pathogen.

There are over 2,600 *Salmonella* serovars, as characterized by their somatic and flagellar antigens (4,5). These serovars can widely differ in phenotypes, and some serovars are more important to public health than others (6–9). With serovars in mind, this analysis sought to establish which individual serovars are most prevalent across different creek systems and how weather patterns can affect the prevalence or relative abundance of these serovars.

# 2. Methods

## 2.1 Data Collection

Over a period of two years (November 2021 - October 2023), monthly water samples were taken from four different creek systems (labeled A-D) in North Georgia. Up to six sites were sampled at each creek, with the sites being numbered by their distance from the creek systems’ headwaters (i.e. higher numbered sites were further down steam than lower numbered sites). From each of these water samples, *Salmonella* was selectively enriched, and DNA was extracted. This DNA was used as input for CRISPR-SeroSeq, a novel deep serotyping technique based on the PCR amplification of the unique CRISPR arrays of *Salmonella* serovars (10). Weather information for each creek system was obtained from the geographically nearest weather station operated by the UGA weather network. Creek systems C and D had the same nearest weather station.

## 2.2 Initial Data Processing

Relevant data for this analysis came from two sources. The first source of data was the results of CRISPR-SeroSeq for each sample were *Salmonella* was detected. Included was the month of the study, the season, the system and site, and individual columns for each serovar that was detected at any point of sampling containing the proportion of that serovar in each sample. Instances where a serovar was not detected in a sample were coded as NA. The second source was an excel file with spreadsheets containing extracted data from each of the included weather stations. For each station, the maximum and minimum air temperature in degrees Fahrenheit, the average relative humidity as a percentage, the average wind speed in miles per hour, the maximum wind speed in miles per hour, the total solar radiation in megajoules per square meter, and the total rainfall in inches was recorded for each day of sampling. Three day averages including the days before and after sampling and the average for each month were also included in separate spreadsheets. Other information included in the data taken just from the day of sampling included the day of the year (ranging 1 to 365) and the year itself (ranging 2021 to 2023).

Cleaning the CRIPSR-SeroSeq data began with recoding all NA values to 0. There were 41 serovars represented in the CRISPR-SeroSeq data. Including all of these serovars in modelling would have likely lead to overly complex models, especially for the serovars that were rarely found in the two years of sampling. To only capture the most relevant serovars in further analysis, included serovars in further analysis were present at least five times in each system. Other serovars were included if they returned a statistically significant Spearman’s rank correlation coefficient, as evaluated by a t-test. Of the remaining serovars, additional binary variables indicating whether or not each given serovar was found in that sample were created as factors. The month of the study was also converted into the month and year formatted as a Date object.

Cleaning the weather data from each station involved converting the day of the year to the actual date of sampling, then reformatting it to just include the month and year in accordance with the CRISPR-SeroSeq data. Since the data from each weather station was in separate data frames, the CRISPR-SeroSeq data was split into separate data frames according to each system. Systems C and D were kept together as they shared a common weather station. For each system, the weather data and the CRISPR-SeroSeq data were combined using a full join, using the date expressed as the month and year as the joining variable. These three data sets were then combined along the rows to form a final processed dataset.

## 2.3 Exploratory Data Analysis

Further exploratory analysis of the processed data included assessing the distributions of each of the numeric variables (each weather variable and the proportions of each of the included serovars) using histograms and boxplots. Correlations of each of these numeric variables were also calculated to assess the potential for multicolinearity in later modeling. Scatterplots were also created for each pair of numeric variables to visually assess this potential. Categorical variables (the season, the system, and the combined year and month) were evaluating using bar graphs.

## 2.4 Statistical Analysis

Three model types were fit in this analysis: logistic regression, beta regression, and regression-based random forest. For the logistic regression models, the binary prevalence of each serovar present at least five times in each system were individually modeled as the response variable. The binary prevalence of all other included serovars and all weather variables were included as predictors. Predictors that appeared to have a relationship when conducting exploratory analysis were also included as interaction terms. Each logistic regression model was evaluated fit within the tidymodels framework. Prior to modeling, the data was split into a training set comprised of 80% of the processed data and a testing set comprised of the other 20%. The models were trained using 5-fold cross-validation of the training data using a random seed of 1234. The accuracy and ROC-AUC were evaluated for these models and compared to the same metrics in similarly fit null models. These models were then fit to the training data and the testing data, and the performance of both were compared using the accuracy and ROC-AUC.

For the beta regression and random forest models, the proportions of each serovar found at least five times in each system were individually modeled as the response, with the proportion of the other serovars, the weather variables, and any relevant interaction terms included as predictors. Beta regression was chosen over a multivariate linear regression because predictions for beta regression models are bounded from zero to one, making them better equipped to model proportions. Additional data processing steps occurred prior to fitting these two model types. Beta regression models cannot model true zero or true one, so instances of zero or one in the serovar data were recoded to 0.0001 and 0.9999, respectively. In addition, all predictors involved in interaction terms were standardized to have a mean of zero and standard deviation of one. The interaction terms were then recalculated and included in the training data. The modeling workflow for the beta regression models followed a similar workflow to the logistic regression models. However, as beta regression is not included in the tidymodels framework, the cross-validation fitting and model evaluation metric calculation were done manually within R. The random forest models were fit within the tidymodels framework using the same variables as the beta regression models, but without the additional data processing steps. These random forests were set to have 1000 trees and were fit using the ranger engine. For the beta regression models and the random forest models, root-mean-square error (RMSE), and R2 were used as model performance metrics.

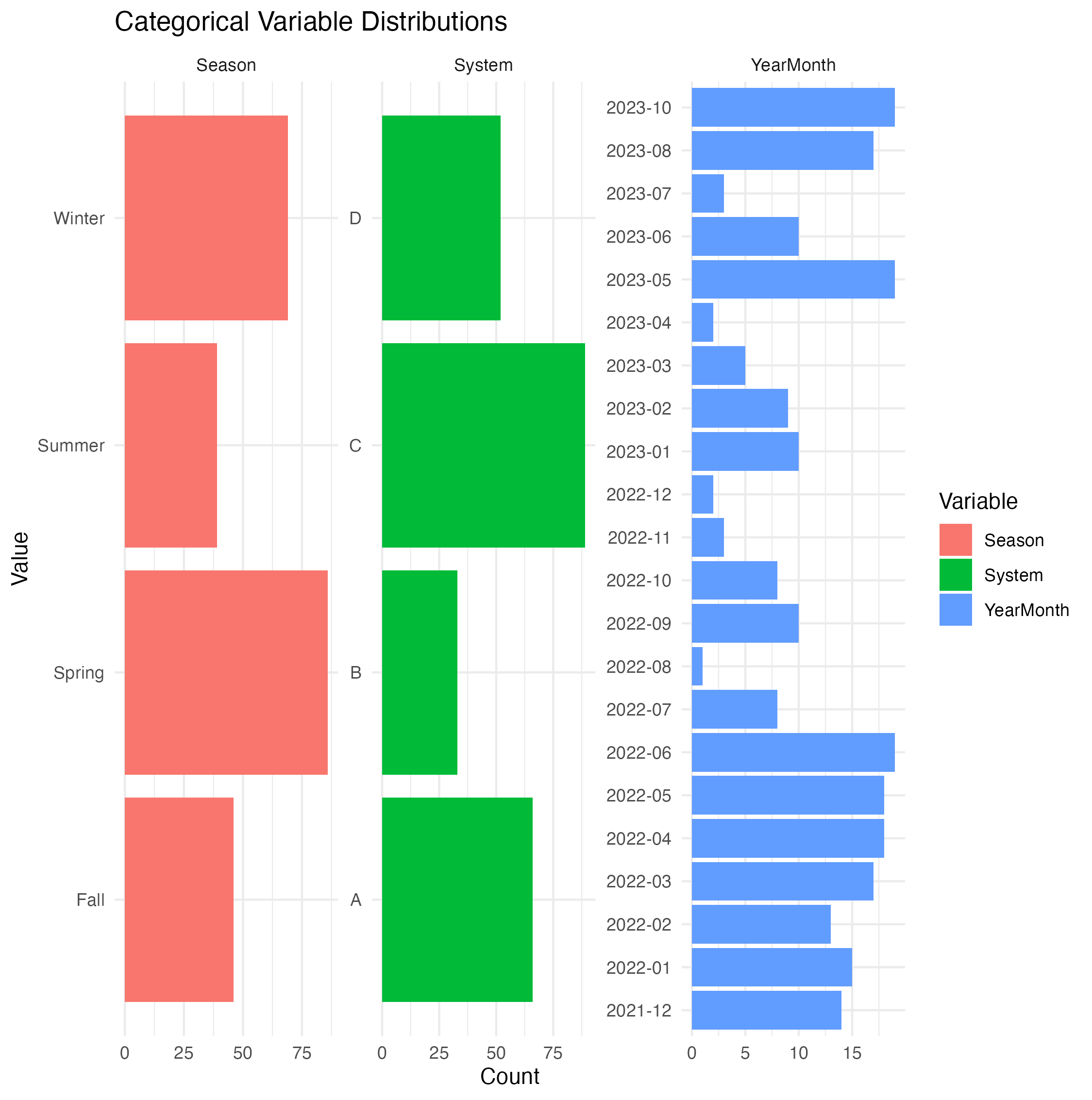
# 3. Results

## 3.1 Data Processing

The CRISPR-SeroSeq data contained 240 observations from the two years of sampling. From these observations, four serovars were represented at least five times in each creek system: Give I, Muenchen I, Rubislaw, and Typhimurium. Additional correlation analysis showed that serovar Infantis was correlated with Give I (R2 = 0.276, p = 0.005) and Muenchen I (R2 = -0.250, p = 0.039) and that serovars Aqua and Inverness were correlated with Muenchen I (R2 = 0.276, p = 0.005). Serovar Give I was also correlated with serovar Rubislaw (R2 = -0.322, p < 0.001). The serovars included in further analysis were serovars Give I, Muenchen I, Rubislaw, Typhimurium, Infantis, and Aqua/Inverness. Serovars Aqua and Inverness are separate serovars, but they cannot be distinguished using CRISPR-SeroSeq. As such, they are considered together in this data.

## 3.2 Exploratory Analysis

Figure 1 shows the frequency of *Salmonella* positive samples among the categorical variables in the data.



The most positive samples were found in the spring, in creek system C, and in June of 2022.

Figure 2 shows the distribution of the numeric variables expressed as boxplots.

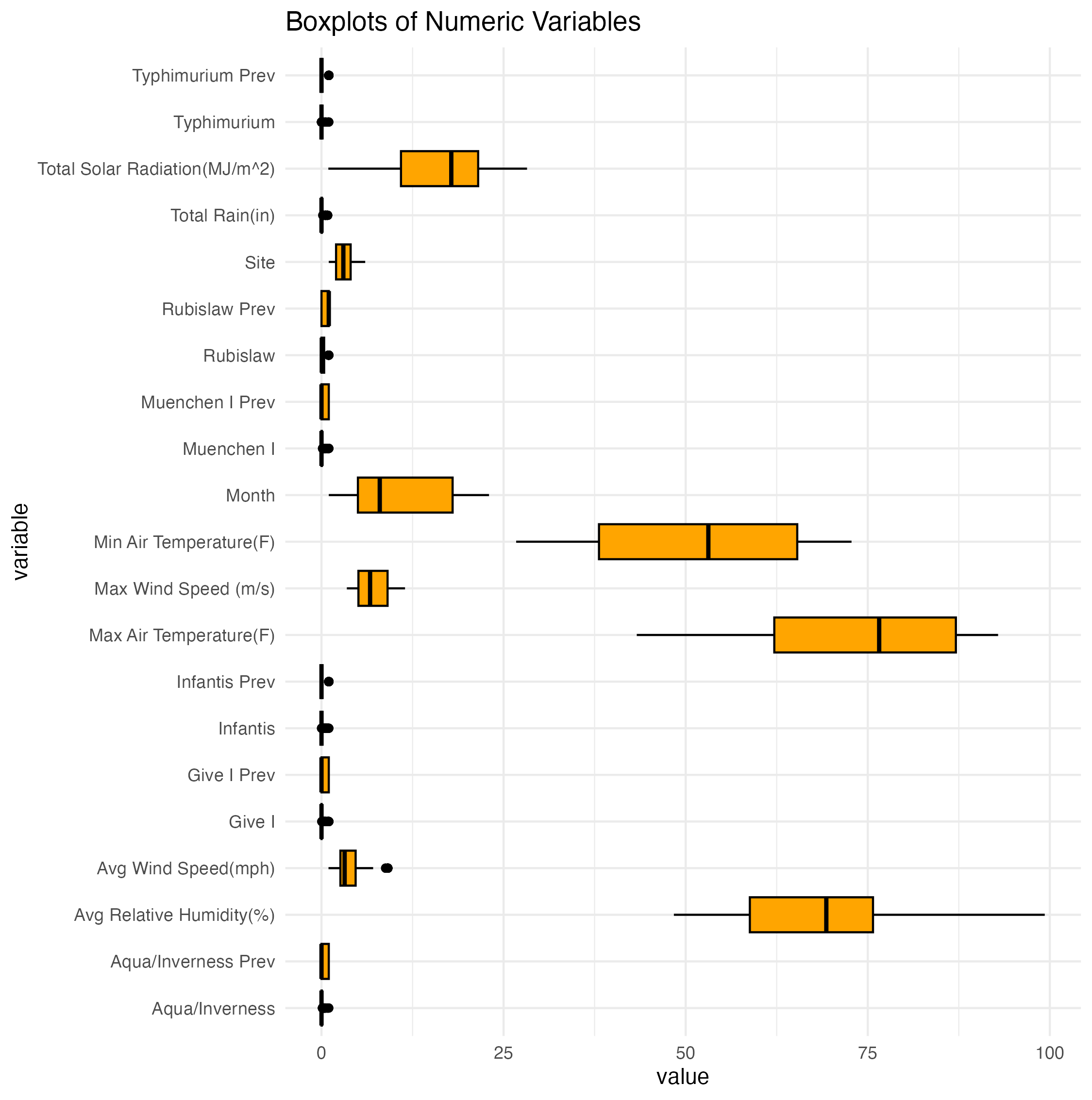
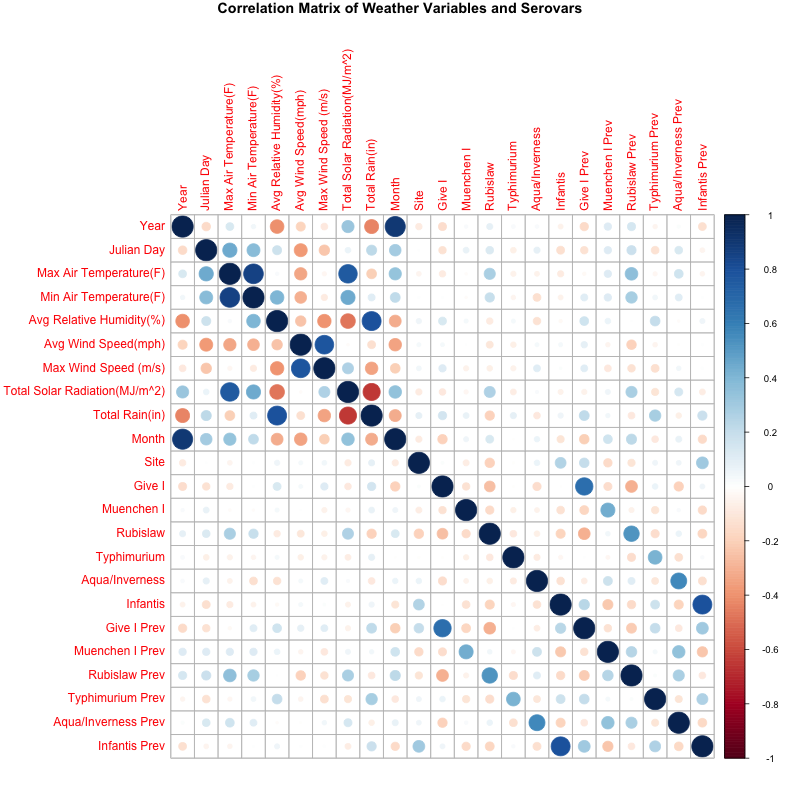


Figure 3 depicts pairwise correlations between each of the numeric predictors.

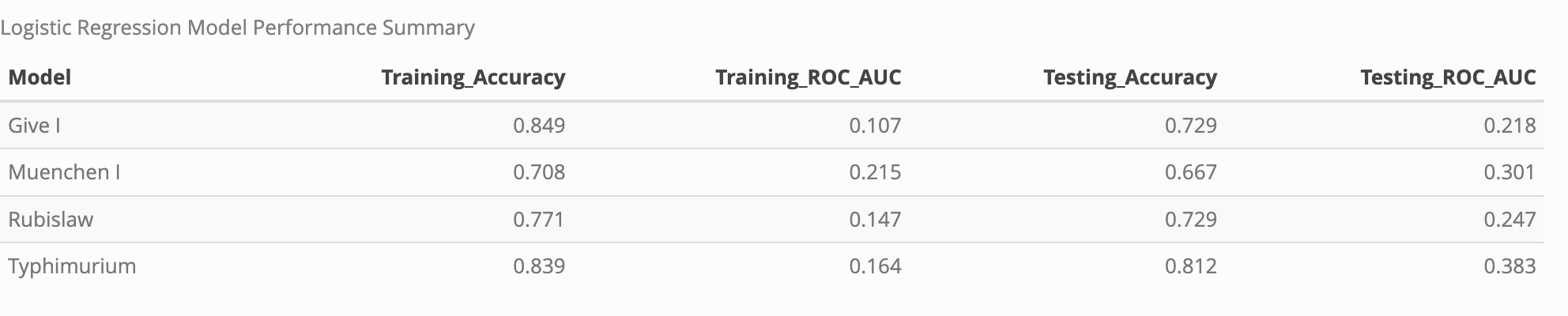


Strong positive correlations were observed between minimum and maximum air temperature and between total rainfall and average humidity, and between average wind speed and maximum windspeed. A strong negative correlation was observed between total rainfall and total solar radiation. Interaction terms for maximum and minimum temperature, total rainfall and average humidity, and total rainfall and total solar radiation were included in future modeling efforts. To handle the potential interaction between average wind speed and maximum wind speed, only the average wind speed was evaluated in modeling.

## 3.3 Model Performance

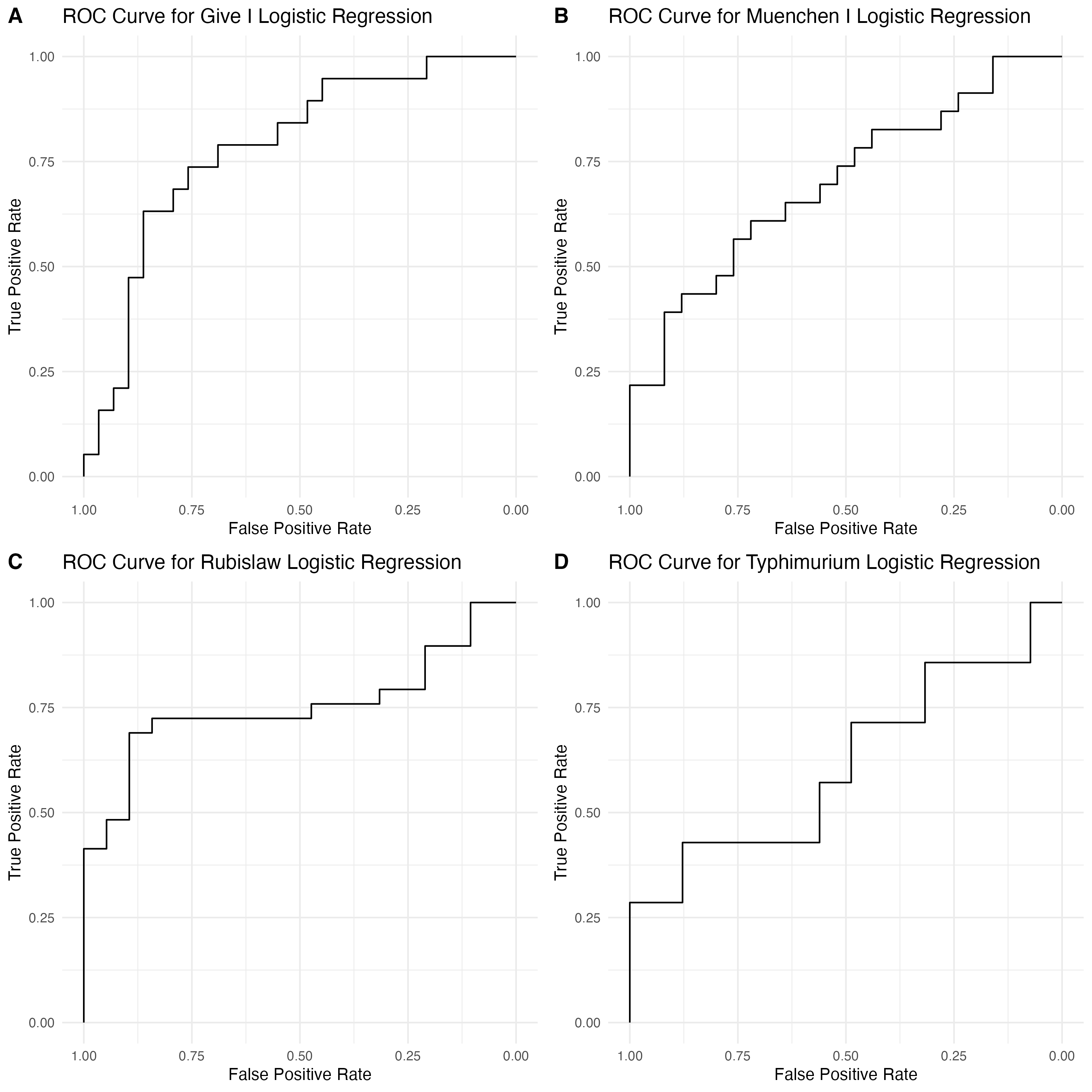
### 3.3.1 Logistic Regression

Table 1 shows the comparison of the logistic regression models fit for each serovar using the training data and the testing data, as evaluated by the accuracy and the ROC-AUC.



When fit to the training data, the models had accuracy values ranging from 0.708 to 0.849 and ROC-AUC values ranging from 0.107 to 0.215. The model performed similarly when evaluated using the testing data, having accuracy values ranging from 0.667 to 0.812 and ROC-AUC values ranging from 0.218 to 0.383.

Figure 4 shows the ROC curves for each model.



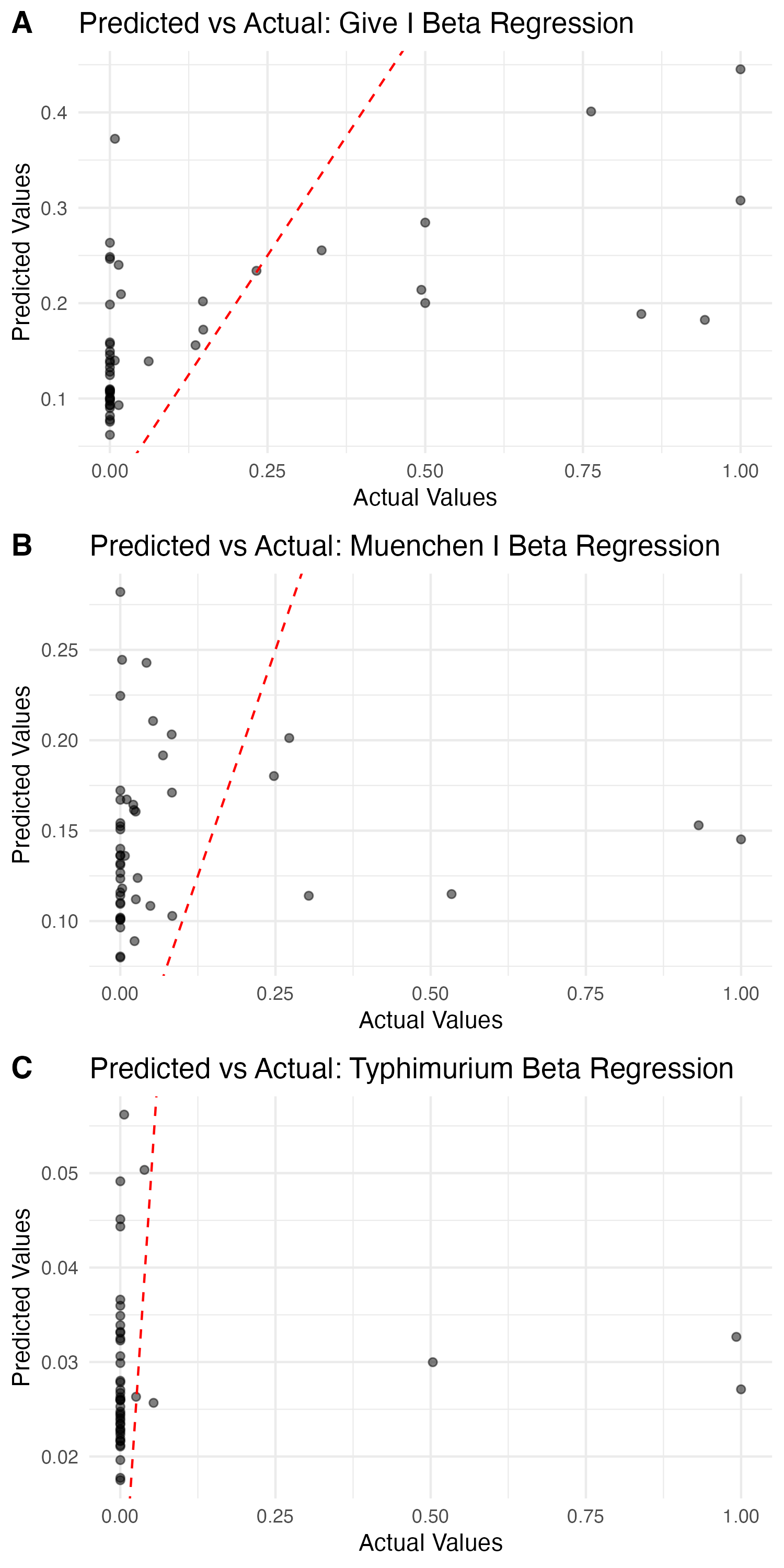
### 3.3.2 Beta Regression

Table 2 shows the RMSE and R2 for the beta regression models fit for each serovar, compared against null models.



The beta regression models had RMSE values ranging from 0.089 to 0.209 and R2 values ranging from 0.204 to 0.416. Each model performed better than the respective null model as determined by the RMSE, albeit only slightly. Modeling the serovar Rubislaw data proved unsuccessful for this model type.

Figure 5 shows scatterplots of the expected vs. predicted values for the beta regression models for each serovar. The red dashed line has intercept of zero and slope of one.



### 3.3.3 Random Forest Performance

Table 3 shows the RMSE and R2 of the random forest models for each serovar.

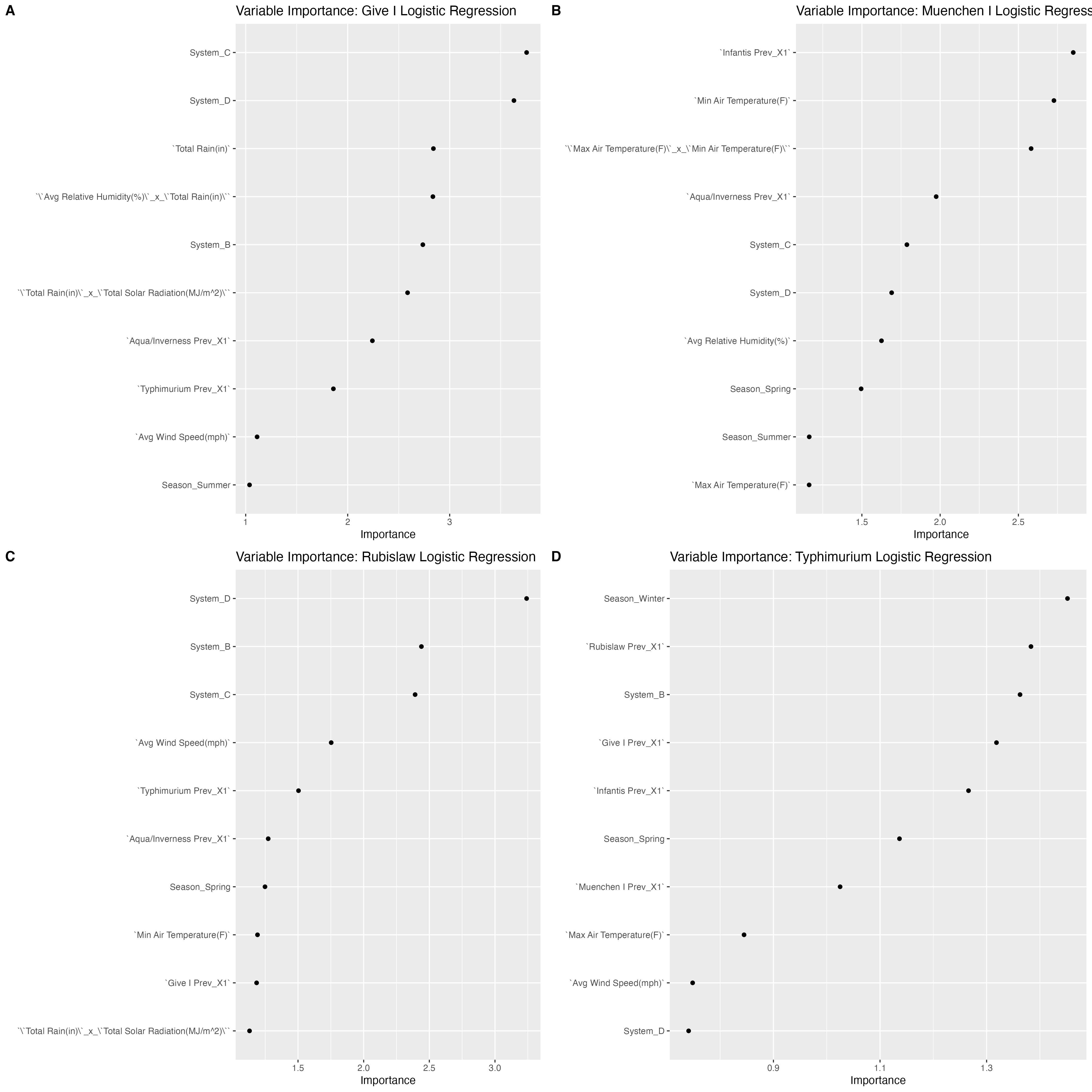


These models performed slightly worse than the beta regression models, with RMSE values ranging from 0.083 to 0.278 and R2 values ranging from 0.026 to 0.223.

## 3.4 Model Interpretability

### 3.4.1 Variable Importance

Figures 6 and 7 depict variable importance plots for each of the logistic regression, beta regression, and random forest models, respectively. Variables with a higher importance score have more impact on the model results.



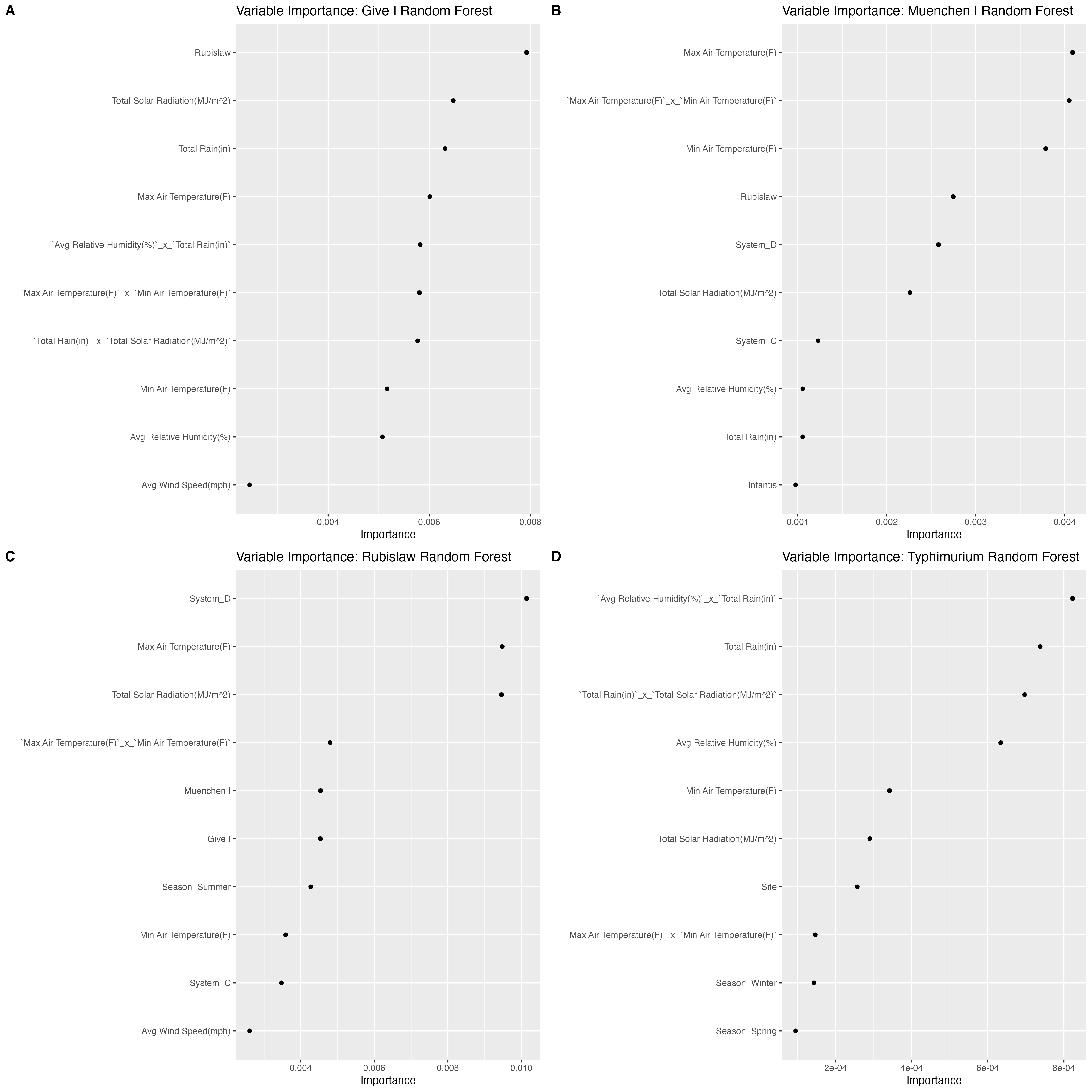


Table 4 shows the significant predictors in each of the beta regression models.



# 4. Discussion

Modeling the binary prevalence of each of the serovars of interest proved largely unsuccessful. The logistic regression models had relatively high accuracy values, and the models performed consistently between the training and the testing data. However, all of the ROC-AUC values were below 0.5, indicating that these models performed worse than what the results would be if predictions were made at random. Certain weather variables were shown to be important in these models. For serovar Give I, total rainfall, the interaction between total rainfall and the total solar radiation, the interaction between total rainfall and and the average wind speed had importance values above zero. For serovar Muenchen I, the minimum temperature, the maximum temperature, the average relative humidity, and the interaction between minimum and maximum temperature were important. For serovar Rubislaw, the average wind speed, the minimum air temperature, and the interaction between total rainfall and total solar radiation were important. For serovar Typhimurium, only the maximum air temperature was important. Given the poor performance of these models, however, this importance is difficult to contextualize.

The beta regression models performed slightly better, with the model fitting the proportion of serovar Give I in a sample performing the best. For these models, though, the weather variables were less significant. In the model for Give I, the total rainfall, the interaction between total rainfall and humidity, and the interaction between total rainfall and total solar radiation were significant predictors. In the model for serovar Typhimurium, the total rainfall was the only significant predictor. The model for serovar Muenchen I had no significant predictors at all.

The random forest models overall performed worse than the beta regression models, though the model for serovar Give I was still the best performing model of the four. A number of weather variables were important across the four models. However, with the highest importance score across the four models being just above 0.01 (System D in the serovar Rubislaw model), no real conclusions can be made about the biological significance of these predictors using these models.

An important limitation of this analysis is that only the weather conditions on each day of sampling were considered. Data that captures average weather trends over time may better capture the biological effect that weather conditions have on *Salmonella* serovar dynamics in creek systems. Overall, the total rainfall was represented the most across all models when considering instances where rainfall was considered by itself and instances where it was represented in an interaction term. This would suggest that rainfall has the overall highest importance. However, due to the mediocre to poor performance of all of these models, further analysis is required to make conclusions on weather’s impact on *Salmonella* serovar dynamics in water.

# 5. References

1. Scallan Walter EJ, Cui Z, Tierney R, et al. Foodborne illness acquired in the united states-major pathogens, 2019. *Emerg. Infect. Dis.* 2025;31(4):669–677.

2. Interagency Food Safety Analytics Collaboration. Foodborne illness source attribution estimates for *salmonella*, *escherichia coli* O157, and *listeria monocytogenes* – united states, 2022. Georgia; Washington, D.C.: U.S. Department of Health; Human Services, Centers for Disease Control; Prevention, Food; Drug Administration, U.S. Department of Agriculture’s Food Safety; Inspection Service; 2024.

3. Bell RL, Kase JA, Harrison LM, et al. The persistence of bacterial pathogens in surface water and its impact on global food safety. *Pathogens*. 2021;10(11):1391.

4. Grimont P, Weill FX. Antigenic formulae of the *salmonella* serovars: WHO collaborating centre for reference and research on *salmonella*. 9th ed. Institut Pasteur; 2007.

5. Issenhuth-Jeanjean S, Roggentin P, Mikoleit M, et al. Supplement 2008-2010 (no. 48) to the White-Kauffmann-Le minor scheme. *Res. Microbiol.* 2014;165(7):526–530.

6. Cheng RA, Eade CR, Wiedmann M. Embracing diversity: Differences in virulence mechanisms, disease severity, and host adaptations contribute to the success of nontyphoidal salmonella as a foodborne pathogen. *Front. Microbiol.* 2019;10:1368.

7. Shah DH, Paul NC, Sischo WC, et al. Population dynamics and antimicrobial resistance of the most prevalent poultry-associated salmonella serotypes. *Poult. Sci.* 2017;96(3):687–702.

8. Uzzau S, Brown DJ, Wallis T, et al. Host adapted serotypes of salmonella enterica. *Epidemiol. Infect.* 2000;125(2):229–255.

9. Gorski L, Noriega AA. Comparison of phenotype nutritional profiles and phosphate metabolism genes in four serovars of salmonella enterica from water sources. *Microorganisms*. 2023;11(8).

10. Thompson CP, Doak AN, Amirani N, et al. High-resolution identification of multiple salmonella serovars in a single sample by using CRISPR-SeroSeq. *Appl. Environ. Microbiol.* 2018;84(21).