Bayesian Analysis of Golden Eagle Counts

Youssef Abdulghani YOUSSEFABDULGHANI@MONTANA.EDU
Department of Physics
Montana State University
Bozeman, MT 59715, USA

Michael Fawcett SWMONTANAJJ@GMAIL.COM
Department of Mathematical Sciences
Montana State University
Bozeman, MT 59715, USA

Mads Munro MADISON.MUNRO@STUDENT.MONTANA.EDU
Gianforte School of Computing
Montana State University
Bozeman, MT 59715, USA

Benjamin DeVries BENJAMIN.DEVRIES1@STUDENT.MONTANA.EDU
Department of Mathematical Sciences
Montana State University
Bozeman, MT 59715, USA

1. Introduction

In our project we are going be using Bayesian Time Series Analysis to analyze Golden Eagle count data that is collected every year in the Bridger Mountains during the fall.

Bayesian time series analysis, is exactly what it sounds like. The Bayesian approach to time series. These Bayesian methods differ from frequentist types of models in all the same fundamental ways that one might observe in other statistical courses. Namely that, Bayesian methods use a tool called "prior" to help inform newly acquired data when fitting a model. Using Bayes's theorem from probability theory we calculate the likelihood of the observed data given our prior parameter distributional knowledge. The resulting output for parameter estimate is called a posterior distribution.

Bayesian time series approaches have a number of advantages over the methods that we have used throughout our course. The first shortcoming of the models we have learned about in class, is that they assume a continuous response variable. For discrete variables, like count data, and their subsequent distributions, we have nothing in our current toolkit to help model these situations. Bayesian Time series gives us a method for this type of data and is explored in this paper.

Bayesian time series analysis enhances our capacity to handle data sets that may exhibit limitations from a frequentist standpoint. Specifically, the use of a prior, can help data limitations such as: sparse data, small data sets, or noisy data sets. Moreover, the prior information does not have to come from an empirical source, as demonstrated in our study. It can come from a number of sources. Having the ability to include prior expertise, or distributional assumptions into the final posterior models enables modeling that is more effective and accurate in situations where frequentist models fall short.

In this project, we examine Golden Eagle count data. By using Bayesian methods we effectively handle the discrete distribution of our response variable. We also utilize a model that is dynamic. In the sense that it can be updated with every new year for which the count data is collected.

The posterior distribution we produce also has the advantage of potentially being used as a starting prior for modeling data when the HawkCount project expands to new areas that have similar features as the Bridgers.

2. Methods

Our main goal in this analysis is to generate a forecast for the 2024 Golden Eagle counts. To do this, we began by splitting the data into training (years; 2019-2022), testing (2023), and validation (2024). Our plan was to first fit the training set and forecast for the testing period, performing variable selection and updating priors to obtain the best possible fit. Subsequently, we refit the model with data from 2019 to 2023 and generate a forecast for 2024. In the following section, we provide a detailed description of the model building process.

For our analysis of the Golden Eagle counts, we chose to implement a fairly simple state-space model using the Bayesian Structural Time Series (bsts) R package [5]. Using this package helped address concerns such as gaps in data collection, but severely limited our control of the priors for parameters. Our final model has three main components, a first-order autoregression, a seasonal component, and regression with an intercept and the hourly observed Sharp-Shined Hawk rate. We chose to use the period of time when the data was collected, as using years worth of daily data would require estimating many seasonal parameters, making the model computationally intensive. This leaves a large artifact on the first day of the season, as the autoregressive component associates the first and last day. Ideally, we would have implemented a short buffer period between migrations. Computational intensity also led us to use dummy seasonal variables as opposed to harmonics. Our seasonal component has 22 dummy variables, each corresponding to three days of the season. The data show a fairly abrupt transition every year around the beginning of October, so we wanted small bins to adequately capture the transition. During this initial fitting period, we also performed variable selection using the Poisson Zellner prior.

The Poisson Zellner prior is a type of spike and slab prior. Spike and slab priors are hierarchical, conditioning on a Bernoulli random variable. For the Poisson Zellner, the spike comes from a normal that degenerates to a point mass at zero. The slab is a non-degenerate normal. Coefficients the model deems unimportant are shrunk to zero making variable selection fairly simple. With the few covariates we have, manual variable selection techniques may be more appropriate, as spike and slab type priors often struggle with multicollinearity. Furthermore, we must note that the Poisson Zellner prior uses the design matrix, making our analysis not truly Bayesian. This affects our credibility intervals as the "prior" does not reflect our previous beliefs about the effect of various predictors on the log of the hourly Golden Eagle rate. Comparing our predictions to the observed data, we may be able to argue that our model is useful; but we have made significant sacrifices in explainability. Ideally, we would have changed the prior after selecting coefficients, but the bsts package does not allow this. While there are many downsides to this prior, it certainly helped expedite variable selection.

3. Data

Our chosen data for Bayesian analysis comes from three places: HawkCount, MSU Optical Remote Sensor Lab's Grafana Weather Dashboard, and the Sacajawea Peak SNOTEL Weather Site. HawkCount is a database containing current and historic raptor bird observation data across the US. We specifically looked at historic data from August 2019 to November 2024 for raptor bird counts in the Bridger Mountain range. Raptor birds of interest to us are the Golden Eagle, the Sharp-Shinned Hawk, and Cooper's Hawk as well as the number of hours these birds were observed by bird watchers in the area.

While HawkCount included weather data with raptor bird counts and observed hours, the database contained no API to pull historic weather data. Thus, we turned to weather data provided by the ORSL at MSU. The ORSL's database, provided by Grafana, contained weather information such as mean temperature, precipitation, absolute air pressure, and wind speed for Bozeman. However, Bozeman weather is often different than weather up in the Bridger Mountains, thus we pulled most of the weather data from Sacajawea Peak's SNOTEL weather station. The SNOTEL station contained historic weather information for mean, minimum, and maximum temperature, snow depth, and precipitation, but not wind speed or air pressure. Therefore, some data transformations were necessary.

As we were pulling from multiple data sources, we needed to process and transform the data before analysis. First, we extended the raptor bird data to include dates before August 27th and after October 31st for each year of interest (2019-2023). To remove missing values for bird counts and observed hours, we zero-imputed those values. We chose zero-imputation because 1) negative counts and observed hours are impossible and 2) zero make sense as a value because any zeroes present in the original data indicate that no one was present to observe birds and/or no birds were seen that day.

$$P_{Adj} = P_{origin} \times \left(1 - \frac{\text{TempLapse} \times \text{Elevation}_{\text{diff}}}{Temp_{origin}}\right)^{\frac{9.81 \times 0.0289644}{8.3144598 \times \text{TempLapse}}} \tag{1}$$

The raptor bird data was not the only data that needed transformations. As we wanted to use weather data that closer aligned with where the raptor birds were located, we used the weather data provided by the SNOTEL station over the ORSL's database. However, we wanted to include air pressure and wind speed. We included wind speed as is in the aggregated weather dataset, but we adjusted the air pressure to match Bridger's air pressure using the Barometric formula (EQ. 1). After aggregating the two weather datasets together, we filled in the missing weather values by imputing each value with their weekly mean. We aggregated the raptor bird and weather datasets together after transformations were complete (Table 1 contains the variables and their description).

For this analysis, our response of interest are Golden Eagle counts in the Bridgers. We selected this bird over the other raptor birds of interest due to their higher observation frequency across 2019-2023 (Fig. 1). While the two hawk breeds appear early in the season (August-September), they become less frequent in October unlike the Golden Eagle. Furthermore, the peak number of Golden Eagles observed is higher than both the Sharp-Shinned and Cooper's hawks.

We also investigated the correlation between each numeric variable (Fig. 2). Our response variable had high positive correlation between Sharp-Shinned Hawk counts and observed hours, and our response also had relatively high negative correlation between snow depth. Other variables with high correlation include all three temperature measurements with each other, snow depth and each temperature variable, wind speed and year, and snow depth and day.

The last step in our exploratory analysis was to fit a regression model to our data and assess for autocorrelation. Since out response variable is a count of Golden Eagles, we fit a full Poisson regression model (EQ. 2). Furthermore, we used the observed hours variable as on offset value for our response.

$$\log(\lambda \times ObsHrs) = \beta_0 + \beta_1(Sharp\_Shinned\_Hawk) + ... + \beta_x(\textbf{variable in raptor dataset}),$$
$$\lambda = Golden\_Eagle\ \textbf{count rate} \tag{2}$$

Once the model was fit, we assessed the model residuals to determine if the model exhibited autocorrelation. We plotted the partial ACF of our model residuals (Fig. 3) and found that the model was an AR[9] autoregressive process. Thus, time series is appropriate for this dataset.

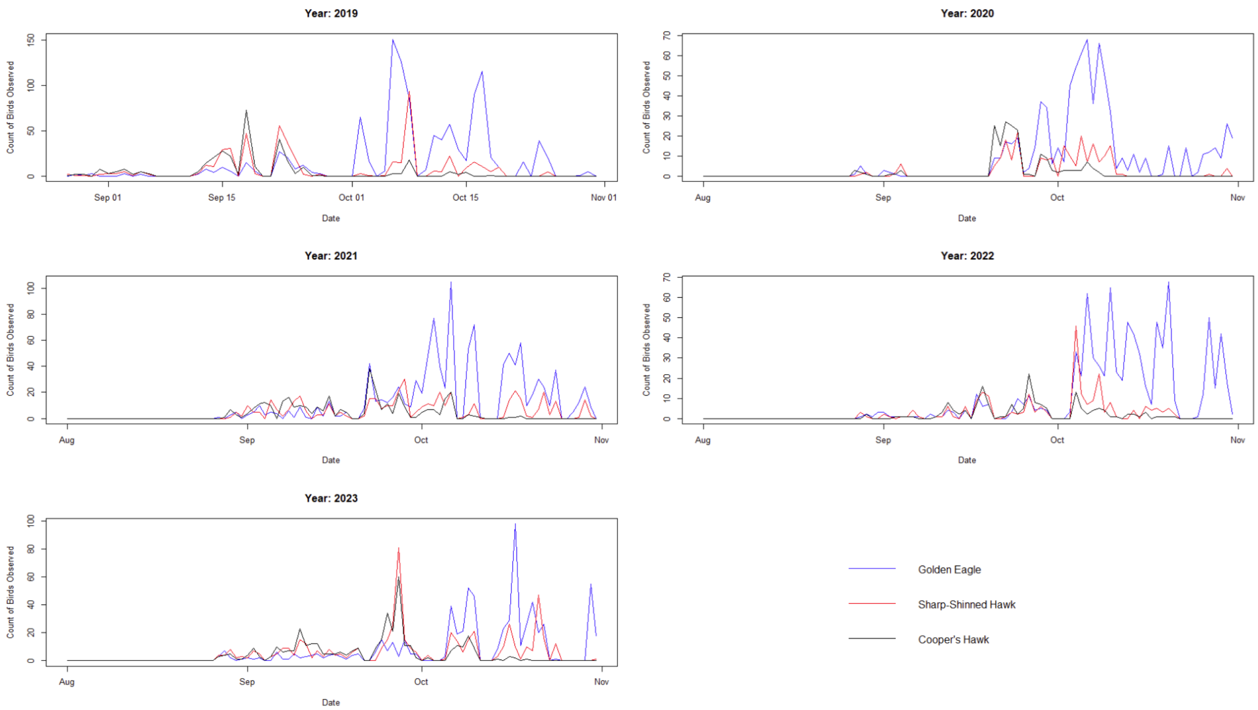

Figure 1: Raptor Bird counts in the Bridger Mountain range across five years

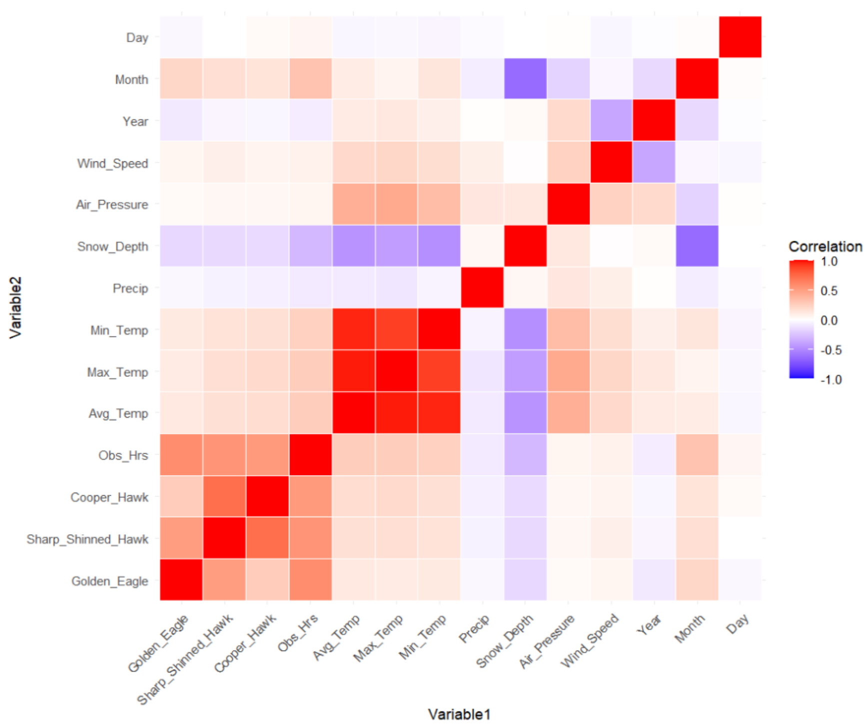

Figure 2: Correlation Heatmap between numeric variables

Table 1: Descriptions of variables in the raptor bird dataset.

Variable	Description
Golden_Eagle	Counts of Golden Eagles observed in the Bridger Mountains in a day.
Sharp_Shinned_Hawk	Counts of Sharp-Shinned Hawks observed in the Bridger Mountains in a day.
Cooper_Hawk	Counts of Cooper's Hawks observed in the Bridger Mountains in a day.
Obs_Hrs	Number of hours raptor birds were watched by at least one investigator.
Avg_Temp	Daily average temperature (°F).
Max_Temp	Daily maximum temperature (°F).
Min_Temp	Daily minimum temperature (°F).
Precip	Daily precipitation accumulation (in).
Snow_Depth	Daily snow depth (in).
Air_Pressure	Average daily absolute air pressure (inHg).
Wind_Speed	Average daily wind speed (m/s).
Time	The date for a given observation in YYYY-MM-DD format.
Year	The year of the observation.
Month	The month of the observation.
Day	The day of the observation.

4. Results

Using the Poisson Zellner prior for variable selection, the only coefficient kept was the hourly Sharp-Shinned Hawk rate. The order of coefficients in the design matrix impacted which variable was selected. When the Cooper Hawk rate was the placed as the first coefficient, it was selected over Sharp-Shinned Hawk. We chose to use Sharp-Shinned Hawk as our lone predictor, as we had not scraped Cooper Hawk data from 2024. This illustrates further some of the issues with a Poisson Zellner prior. The coefficients removed may contain more information then a general intercept. For example, weather conditions may impact observations. However, the seasonal component of our model dominates. Furthermore, local weather might not be relevant, as some Golden Eagles have been recorded flying 314 kilometers in a day [2].

We think that our potential predictors and their relationship with the response give a prime example of correlation versus causation. We ultimately kept the Sharp-

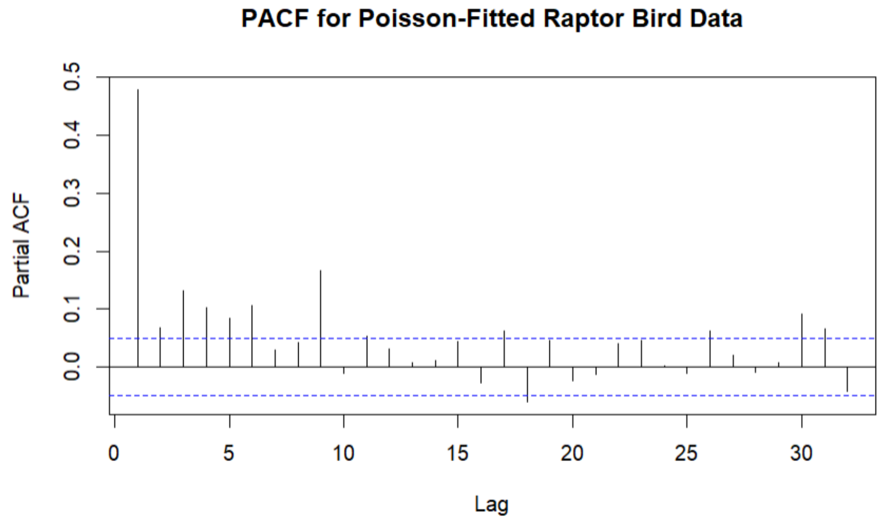

Figure 3: Partial ACF for Golden Eagle Poisson Model

Shinned Hawk rate in the model to help us learn about incorporating regression into a Bayesian time series analysis framework.

The final theoretical model is presented below.

$$\log(\lambda_t \cdot H_t) = \mu_t + \tau_t + \boldsymbol{\beta}^T \boldsymbol{x}_t$$
$$\mu_t = \phi\mu_{t-1} + \epsilon_{1_{t-1}}$$
$$\tau_t = \sum_{i=1}^{22} \gamma_{i,t-1} I_{season=i} + \epsilon_{2,t-1}$$
$$\mu_1 \sim N(0, \sigma_{\mu_1})$$
$$\gamma_{i,1} \sim N(0, \sigma_{\gamma_1})$$
$$\epsilon_{1,t} \sim N(0, \sigma_\mu)$$
$$\epsilon_{2,t} \sim N(0, \sigma_\gamma)$$
$$\sigma_y \sim \text{InvGamma}(a,b)$$
$$\sigma_\mu \sim \text{InvGamma}(a, b)$$
$$\boldsymbol{\beta}|\boldsymbol{\alpha} \sim MVN(\boldsymbol{b}, \boldsymbol{V}^{-1})$$
$$\alpha_j \sim \text{Bernouilli}(0, p_j)$$
$$\boldsymbol{V}^{-1} = \kappa((1-\delta)\boldsymbol{x}^T\boldsymbol{w}\boldsymbol{x}/n + \delta\text{diag}(\boldsymbol{x}^T\boldsymbol{w}\boldsymbol{x}))$$

For the final model, $j = 1, 2$ is the index of the regression coefficient, $n = 330$ is the sample size, $\kappa = 0.01$ is a shrinkage parameter, $b_0 = 0$ corresponds to the intercept,

$b_1 \approx 0.014$ is the coefficient for the Sharp-Shined Hawk rates, $p_1 = 0$, $p_2 = 1$ are the inclusion probabilities for coefficients, $a \approx 0.0131$, $b = 0.01$, $\sigma_{\mu_1} = \sigma_{\gamma_1} = 0.1$. λ is a quantity of interest, denoting the hourly rate Golden Eagles are observed. Finally, H is the number of survey hours, accounting for exposure.

Using this model, we generated forecasts for 2024. Figure 4 shows the observed Golden Eagle counts, mean prediction, and 95% predictions intervals. Our prediction intervals are quite wide, and tend to overestimate the Golden Eagle count. 65/66 observations are within the prediction interval, indicating we are slightly above the nominal rate, but the primary issue is the width of the intervals. Implementing a similar model in JAGS or Stan would have given us more control, potentially allowing us to shrink these intervals. Even though we believe we are overestimating the hourly Golden Eagle rate, we present Figure 5, showing a 95% prediction interval for the hourly Golden Eagle rate.

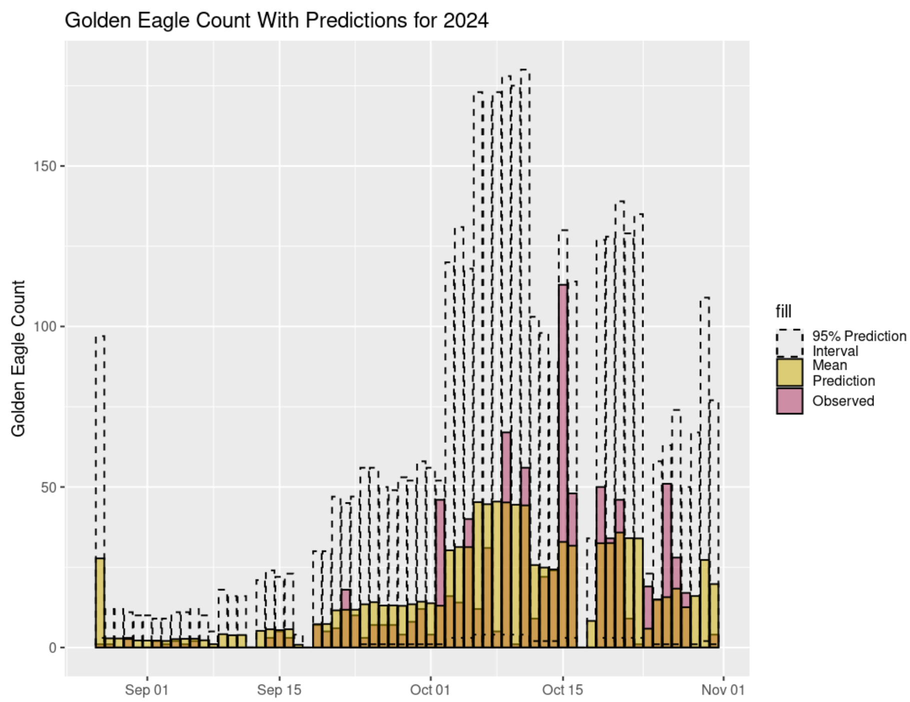

Figure 4: 95% Prediction Interval for number of Golden Eagles Observed Along With Observed Counts

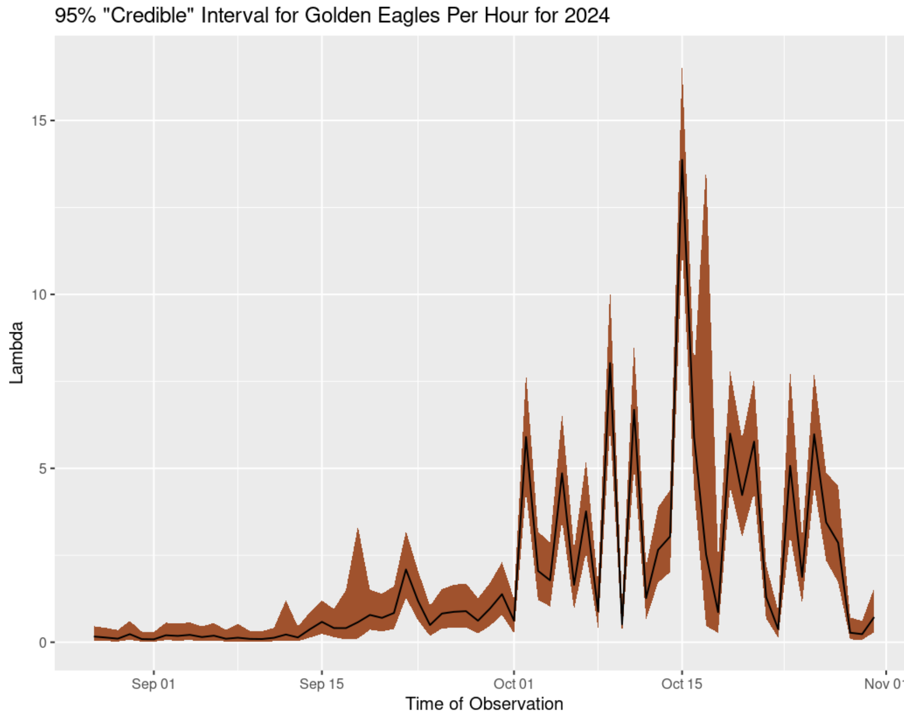

Figure 5: 95% Prediction Interval with Median Hourly Rate of Golden Eagles

5. Discussion

The primary question addressed in this project was whether Bayesian time series analysis could accurately model and forecast Golden Eagle counts in the Bridger Mountains. Our analysis revealed several insights and limitations. First, this was an exercise on the chaotic nature of real-world datasets. Handling missing data as well as having data only in specific season of the year was something that we did not encounter in textbooks example datasets. Secondly, the Bayesian model we considered demonstrated its utility in handling count data, a discrete response variable, and its ability to incorporate dynamic updates with additional data. The inclusion of the Sharp-Shinned Hawk rate as the sole predictor highlighted the correlation between this variable and the Golden Eagle counts. As already discussed, the Poisson Zellner prior is effective but it also introduced various biases and limitations.

Furthermore, the seasonal component was the dominant part of the model, suggesting that local weather conditions might be less critical for Golden Eagle migration patterns, given their ability to traverse large distances. This aligns with ecological understanding but limits the explanatory power of the covariates. The wide prediction intervals for the 2024 forecast reflect the modeling limitations and highlight potential overestimation of the hourly Golden Eagle rate.

Despite these challenges, the project provided valuable experience in applying regression analysis using a Bayesian analysis framework. It also underscored the importance of balancing computational efficiency with model precision, particularly when working with real-world ecological data.

Future work could focus on refining the model by leveraging more sophisticated priors or alternative software platforms to improve accuracy and interpretability. Moreover, incorporating additional ecological or meteorological predictors, along with a more robust treatment of multicollinearity, might enhance model performance. As a final note, the project demonstrated the potential of Bayesian methods for ecological forecasting, paving the way for more nuanced analyses in similar contexts.

References

[1] Rob Hyndman, George Athanasopoulos, Christoph Bergmeir, Gabriel Caceres, Leanne Chhay, Mitchell O'Hara-Wild, Fotios Petropoulos, Slava Razbash, Earo Wang, and Farah Yasmeen. *forecast: Forecasting functions for time series and linear models*, 2024. URL `https://pkg.robjhyndman.com/forecast/`. R package version 8.23.0.

[2] None Listed. SSE renewables. URL `https://www.sserenewables.com/media/1ldehntt/how-far-do-eagles-fly-in-a-day.pdf`.

[3] José Pinheiro, Douglas Bates, and R Core Team. *nlme: Linear and Nonlinear Mixed Effects Models*, 2024. URL `https://CRAN.R-project.org/package=nlme`. R package version 3.1-166.

[4] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2024. URL `https://www.R-project.org/`.

[5] Steven L. Scott. *bsts: Bayesian Structural Time Series*, 2023. URL `https://CRAN.R-project.org/package=bsts`. R package version 0.9.8.

[6] W. N. Venables and B. D. Ripley. *Modern Applied Statistics with S*. Springer, New York, fourth edition, 2002. URL `https://www.stats.ox.ac.uk/pub/MASS4/`. ISBN 0-387-95457-0.

[7] Hadley Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2016. ISBN 978-3-319-24277-4. URL `https://ggplot2.tidyverse.org`.

[8] Hadley Wickham, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, Alex Hayes, Lionel Henry, Jim Hester, Max Kuhn, Thomas Lin Pedersen, Evan Miller, Stephan Milton Bache, Kirill Müller, Jeroen Ooms, David Robinson, Dana Paige Seidel, Vitalie Spinu, Kohske Takahashi, Davis Vaughan, Claus Wilke, Kara

Woo, and Hiroaki Yutani. Welcome to the tidyverse. *Journal of Open Source Software*, 4(43):1686, 2019. doi: 10.21105/joss.01686.