# Assessing the Impact of Global Warming and Environmental Factors on the Spectacled Eider Using Various Machine Learning Models.

By Yan Sharma and Noah Abbas

Staten Island Technical High School, 10/2024

# Table Of Contents:

## Abstract

The purpose of this paper is to analyze how various temporal and climatic variables affect the migration patterns of the spectacled eider (*Somateria fischeri*). The eider is a waterfowl that migrates between Northeastern Siberia and Coastal Alaska. In recent years, the eider's population size has had a consistent decrease, which global warming has been theorized to play a role in mitigating harvest sizes and nesting locations. To address the gap in the knowledge, data was gathered from 1993 to 2011 from the Department of Interior (DOI) Satellite database and was analyzed using various Python scripts. Exploratory data analysis (EDA) was conducted to clean, visualize, and explore correlations within the data. Data was graphed to analyze migratory patterns, average durations, and differences in durations between sexes. Additionally, a polynomial regression model was created to analyze the correlation between average global temperature and maximum latitude reached. A weak correlation was determined, with a root mean square error value of 0.153. Afterwards, a Long Short Term Memory (LSTM) model was used to predict the average maximum latitude of a spectacled eider based on average global temperature. The model used TensorFlow and Keras and reformatted the data into twelve scaled time steps to capture seasonal data. The model's predictions of the average latitude were off by 2.6 degrees, suggesting that the model can capture general trends with low precision.

## Introduction:

Throughout the 21st century, a consistent rise in temperature of .2°C per decade globally has led to a sharp increase of global sea levels and the distribution of various species' migratory patterns (Hansen et al., 2006). The overall shift in both climate and habitat has led to the mass extermination of various species, with approximately 50% of all animal species and 18% of all land-based species put at risk of extinction by 2100 (IPCC, 2023; Myers & Knoll, 2001). One such species affected by global warming is the spectacled eider (Somateria fischeri), a sea duck native to the Arctic coasts of Alaska and Siberia. From 1967 to 1992, the eider had an estimated 87% population decline in total population density (Stehn et al., 1993). While the reasoning for the decline is not fully understood, global warming has been theorized to indirectly contribute to it through mitigating harvest sizes and nesting locations (Brambilla, 2008; Both, 2010).

In recent years, machine learning has been used to analyze both structured and unstructured data to find correlations between complex variables, enabling the analysis of intricate situations (Amini & Rahmani, 2023). While machine learning has been used in a wide range of fields, its ability to conduct predictive analytics on large population datasets has given it use in environmental science (Lopez-Martínez et al., 2018). This application has allowed researchers to model the impacts of global warming by assessing the potential effects of climate change on wildlife populations.

The main gap in the existing research for the eider is the lack of specific analysis on how environmental factors may be affecting the population of the eider. Rapid climate change has historically contributed to the distribution of avian species' breeding cycle, such as the abundance of food supplies and average poleward migration (Hansen et al., 2006). Despite this, little research has been conducted on how these variables may affect the population of the Spectacled Eider, and to what extent. We hypothesize that we can develop various machine learning models to fill the gap in the knowledge to evaluate the impact of climate change on the eider’s migration patterns.

Exploratory Data Analysis:

The data gathered from the DOI was first visualized in Python using the Plotly library to map the migration patterns of individual eiders. To ensure there was a large enough dataset to draw valuable conclusions from, we filtered the dataset to include only eiders with over three hundred recorded data points. Each eider's migration was displayed on a graph, highlighting key locations such as Northern Alaska and Coastal Russia. The graph also included key details like the start and end dates, species ID (a unique identifier for each eider), and the total number of recorded data points.

To eliminate inconsistencies and unreliable data, we removed any entries in the CSV file which were missing data and excluded any eiders tracked for fewer than 80 days. After filtering, the data showed that eiders were tracked for an average of 292 days (median: 185.5 days), with an interquartile range of 94.25 to 537.25 days. To explore the differences between various groups of eiders, the data was separated and analyzed between males and females. While the average tracking duration was slightly different for males (288 days) and females (294.8 days), the analysis showed no statistically significant difference between the two, as shown by a t-statistic of -0.9178 and a p-value of 0.36.

Additionally, a scatter plot graphing deployment dates versus deployment duration was created, highlighting a significant shift in average deployment length before and after 2008. The data covers the period from 1993 to 2011, though there is a noticeable gap in the late 1990s and early 2000s due to missing data within the CSV file. After 2008, the average deployment duration for eiders rose sharply to 348.1 days, compared to just 129.5 days in earlier years. Additionally, the range of deployment durations widened, from 281 days to 911 days. Further analysis will be needed to investigate potential causes for this increase.

Basic Polynomial Regression Model:

A polynomial regression algorithm was used to analyze the correlation between the maximum latitude transmitted by an individual and the average global temperature when they reached that location. Global temperature was defined as the average temperature of the air two meters above the ground above both land and sea. The monthly average temperature was gathered from the Our World dataset. From there, basic data cleaning was performed to ensure there was sufficient data to construct the model. The data was also filtered to only include eiders which had at least 6 months of data to ensure robust data was used.

The polynomial model was constructed using various sklearn tools, such as normalizing the latitudes and temperature arrays via the StandardScaler. From there, the data was used to train the model via an 80-20 test-train split. To determine the reliability of the model, the mean squared error and root mean squared error (R squared) values were calculated, and a second-degree polynomial was used to capture nonlinear features while preventing overfitting.

The model had a mean squared error of 0.604, and a R squared value of 0.153. As the R squared value is extremely low, it suggests that temperature explains only a small fraction of the variation in latitude, which thus may imply that there is a weak correlation between average global temperature and latitude. Further studies may consider tweaking elements of the model to improve its accuracy, such as adjusting the degree, varying the test-validation split, and analyzing specific time periods.

LSTM Model:

After testing the polynomial regression model, we attempted to use a Long Short Term Model to predict the average maximum latitude of a spectacled eider during a month based on the average global temperature during that month. For the training data for the model, we got the range of months from the first location transmission to the last (May 1993 to August 2012). For the months within this range with no transmissions, the rows were filled with nan values. The maximum latitude achieved in that month by every transmitting individual was then averaged. For all months within the range, the Our World in Data temperature dataset was used to get the average monthly global surface temperature. We decided to use the global temperature as opposed to the temperature in Canada, the USA, or Russia because of the higher correlation we achieved with that set of temperatures with the polynomial regression model.

The LSTM model construction for predicting duck migration patterns involves several key steps. Initially, the latitude features are preprocessed using sklearn's MinMaxScaler, which normalizes the data to a 0-1 range. This scaling is crucial for ensuring all input features are on a comparable scale, potentially improving the model's performance and convergence speed. Following this, the data is transformed into sequences suitable for LSTM processing. The sequences are 12 time steps long to represent the 12 months of the year. This approach allows the LSTM to learn from patterns in the duck migration data over a full year, potentially capturing seasonal trends.

The LSTM model itself is constructed using TensorFlow and Keras, with an architecture consisting of a Masking layer to handle potential padding in input sequences, an LSTM layer with 50 units to process the sequential data, a Dropout layer with a rate of 0.2 to prevent overfitting, and a Dense output layer with a single unit for temperature prediction. The model is compiled using the Adam optimizer and means squared error as the loss function.

For the validation section of the dataset, the LSTM model achieved a mean squared error (MSE) of 0.0504, a root mean squared error (RMSE) of 0.2245, and a mean absolute error (MAE) of 0.1919. Since the latitude values were scaled from 0 to 1 using the MinMaxScaler, these error metrics are based on scaled data. The error values indicate that the model's predictions deviate moderately from the actual values. Specifically, the RMSE value of 0.2245 suggests that, on average, the model's predictions are off by 22% of the normalized range, indicating some level of inaccuracy.

When these errors are unscaled back to the original latitude range, the model's predictions of the average latitude are off at approximately 2.6 degrees. This suggests that the model can capture some trends in the data but still struggles to provide precise predictions. For a bird species like the Spectacled Eider, which migrates over large distances and whose movement is influenced by multiple environmental factors (besides temperature), an average error of 2.6 degrees of latitude represents a significant gap in predictive accuracy. It implies that the model failed to fully capture the complexity of the migration patterns concerning temperature changes alone.

The training metrics provide further insights into the model's performance. The LSTM model recorded a training MSE of 0.0793, a RMSE of 0.2815, and an MAE of 0.2243, all of which are higher than the validation values. Typically, when the training error exceeds the validation error, it suggests the model is underfitting the training data. In this case, the model cannot fully capture the underlying patterns in the training data, which could be due to the model's simplicity or insufficient training epochs. The fact that the validation loss is lower than the training loss indicates that the model generalizes well to new data but has not yet learned enough from the training data.

The higher training error points to limitations in the dataset such as gaps in data or an inadequate number of months (only 80), which makes it harder for the model to fully understand migration patterns. The current errors suggest that the model is not capturing the full complexity of the migration patterns based on temperature alone. There are several factors that could explain why the model's predictions are inaccurate. For one, the model was attempting to predict average latitude based solely on average temperature. While temperature is a key factor in bird migration, other environmental variables—such as wind patterns, food availability, sea ice coverage, and photoperiod—also play crucial roles in determining migration routes and timing. The absence of these variables may have limited the model's ability to make accurate predictions. Additionally, while LSTM models are well-suited for capturing temporal dependencies, the architecture used here might not have been complex enough to fully model the relationships between temperature and migration patterns. The model employed only a single LSTM layer with fifty units, which might not have been sufficient to capture the seasonal trends or subtle interactions in the data.

To address these sources of error in future research, several improvements could be made to enhance the model's accuracy. First, incorporating additional environmental variables such as wind patterns, food availability, sea ice coverage, and photo period could provide a more comprehensive dataset for the model to learn from. These features would capture more of the complex factors that influence migration, helping the model predict latitude more accurately. Additionally, the LSTM architecture could be made more sophisticated by using stacked LSTM layers or increasing the number of units per layer, allowing the model to capture more nuanced temporal dependencies and seasonal trends. Alternative architectures, such as GRUs or even

Transformer models, could also be tested for their ability to handle long-term dependencies and non-linear relationships more effectively.

However, a large portion of the inaccuracy is likely due simply to the limitations of the data. The large gaps in the data where no Eiders transmit for months or even years limited the model's ability to deduce the temporal patterns in the data. LSTMs are designed to capture sequential dependencies over time, but when there are missing months, the model loses critical temporal information, making it harder to learn the migration patterns accurately. Additionally, with only 79 months of data, the model was working with a relatively small dataset. LSTM models, especially, require a significant amount of data to learn temporal patterns effectively. A small dataset limits the model's ability to distinguish between genuine trends and noise in the data. Since LSTMs are designed to capture long-term dependencies, insufficient data means it may fail to recognize the full scope of migration patterns or seasonal changes.

To account for these weaknesses in the dataset, there are multiple things future researchers can do. One solution is to impute missing values using statistical techniques like linear interpolation, moving averages, or more advanced methods like K-Nearest Neighbors (KNN) or time series-specific imputation (e.g., Seasonal Decomposition of Time Series - STL). This would allow the model to have continuous data without gaps, making it easier to learn from the complete sequences. Researchers could also consider transfer learning, using a pre-trained model on a similar task (e.g., migration patterns of another bird species), and fine-tuning it on the Spectacled Eider data could enhance model performance. This approach allows the model to start with learned representations and then adapt to the limited data available.

Time series correlation:

The construction of cross-correlation functions between a time series of the Spectacled Eider's mean latitude and the average temperature for each month serves as an analysis tool to understand the relationship between environmental factors and migration patterns. Cross-correlation functions quantify the degree of association between two time series while accounting for potential time lags. This approach is essential, as bird migration is influenced not only by current temperature conditions but also by prior environmental cues that signal changes in seasonality, food availability, and habitat conditions. By employing the built-in stats.pearsonr function to compute the cross-correlation, insights can be gained into whether temperature changes precede shifts in migration patterns or if bird movements lag behind temperature fluctuations. This analysis helps determine the sensitivity of the Spectacled Eider to environmental changes driven by global warming.

Because of the gaps where there are no transmissions for some months, the time series were split up into five separate ranges. For each range, a cross-correlation function was run. The ranges were: from May to December of 1993, from May 1994 to March 1995, from June to December of 1995, from June to November of 1996, and from May 2008 to August 2012. The correlation and p values for each range and time series can be seen in Table 1.

| Date | Correlation value | P value |
|---|---|---|
| May 1993 - December 1993 | 0.97 | 0.0015 |
| May 1994 - March 1995 | 0.36 | 0.3867 |
| June 1995 - December 1995 | 0.80 | 0.0585 |
| June 1996 - November 1996 | 0.34 | 0.5696 |
| May 2008 - August 2012 | 0.61 | 0.0001 |

Table 1

## References:

Both, C. (2010). Food availability, mistiming, and climatic change. In *Effects of Climate Change on Birds* (p. 320). Oxford Biology. https://books.google.com/books?hl=en&lr=&id=diiQDwAAQBAJ&oi=fnd&pg=PA129&dq=Food+availability,+mistiming,+and+climatic+change+Christiaan+Both&ots=a0zmeO6Sd0&sig=2XnLVjLTlgbdlKxvdn19AZ3f3mU#v=onepage&q=Food%20availability%2C%20mistiming%2C%20and%20climatic%

Brambilla, M., Resano-Mayor, J., Scridel, D., Bogliani, G., Braunisch, V., Capelli, F., Cortesi, M., Horrenberger, N., Pedrini, P., Sangalli, B., Chamberlain, D. E., Rubolini, D., & Arlettaz, R. (2018, May 3). *Past and future impact of climate change on foraging habitat suitability in a high-alpine bird species: Management options to buffer against global warming effects*. Past and future impact of climate change on foraging habitat suitability in a high-alpine bird species: Management options to buffer against global warming effects. https://www.sciencedirect.com/science/article/abs/pii/S0006320717319109?casa_token=Ixw4sPYo6H0AAAAA:jnzP2IoR6QcsiLBoUIF1LfzrKAJODMsyxhMVZh2X4rSRrvfFVqkGpdopKaylLLmn8UknKaRMfpVt

Hansen, J., Sato, M., Reudy, R., Lo, K., Lea, D. W., & Medina-Elizade, M. (2006, September 26). *Global temperature change*. Global temperature change. https://www.pnas.org/doi/full/10.1073/pnas.0606291103

Intergovermental Panel on Climate Change. (2023, June 16). *Sixth Assement Report*. Sixth Assement Report. https://www.ipcc.ch/report/ar6/wg2/downloads/faqs/IPCC_AR6_WGII_Overaching_OutreachFAQ2.pdf

Lopez-Martínez, F., Schwarcz, A., Núñez-Valdez, E. R., & García-Díaz, V. (2018, November 15). *Machine learning classification analysis for a hypertensive population as a function of several risk factors*. Machine learning classification analysis for a hypertensive population as a function of several risk factors. https://doi.org/10.1016/j.eswa.2018.06.006

Myers, N., & Knoll, A. H. (2001, May 8). *The biotic crisis and the future of evolution*. The biotic crisis and the future of evolution. https://www.pnas.org/doi/full/10.1073/pnas.091092498

Stehn, R. A., Dau, C. P., Contant, B., & Butler, W. A. (1993, May 12). *Decline of Spectacled Eiders Nesting in Western Alaska*. Decline of Spectacled Eiders Nesting in Western Alaska. https://www.jstor.org/stable/40511415