Enhancing Ecological Forecasting with LSTM Models: The Impact of Partition-based Data Shuffling on Predictive Accuracy

Zhengnan Li, Juan Francisco Leonhardt Chavez, Shubham Pant, David Li
Katz School of Science and Health
Yeshiva University
New York, NY
Emails: zli6@mail.yu.edu, jleonhar@mail.yu.edu, spant1@mail.yu.edu, david.li@yu.edu

Abstract—**This study explores the impact of a novel partition-based shuffling technique on the predictive accuracy of Long Short-Term Memory (LSTM) models in ecological forecasting. By utilizing datasets on mosquito population dynamics and environmental variables, we demonstrate that partition-based shuffling—which randomizes inter-year sequences while preserving intra-year temporal patterns—significantly enhances model performance. Experimental results show that models trained with this method achieve a notable reduction in root mean square error (RMSE) compared to those trained on non-shuffled data, underscoring its ability to balance temporal integrity with generalization. The study offers a robust methodological framework that improves the applicability of LSTM models in ecological and epidemiological research by addressing biases caused by sequential data dependencies. The findings advocate for the adoption of partition-based shuffling as a crucial preprocessing step to enhance the reliability and robustness of time-series forecasting models in similar domains.**

Index Terms—**Partition-based Data Shuffling, LSTM Models, Time-Series Forecasting, Predictive Accuracy.**

I. Introduction

Accurate forecasting of ecological phenomena, such as biological populations and environmental dynamics, is critical for research and real-world applications, including public health planning and resource management. In recent years, advancements in machine learning have enabled the development of powerful models capable of capturing complex patterns in time-series data. Among these, Long Short-Term Memory (LSTM) networks have emerged as a robust solution for modeling sequential data, offering significant improvements in predictive performance over traditional statistical approaches. [1]

Despite these advancements, the effectiveness of LSTM models heavily depends on the quality of data preprocessing. The way data is prepared and presented to these models can introduce biases or hinder generalization, ultimately limiting their utility. One such preprocessing technique—data shuffling—plays a pivotal role in mitigating sequence-induced biases by randomizing the order of training data points. While widely applied in machine learning, the specific impact of shuffling on LSTM models in ecological forecasting has received limited attention. [5]

This study addresses this gap by introducing a partition-based shuffling methodology tailored to the unique characteristics of ecological time-series data. Unlike conventional shuffling methods that risk disrupting critical temporal structures, partition-based shuffling randomizes inter-year sequences while preserving intra-year continuity. This approach ensures that seasonal and other short-term patterns remain intact, enabling the model to learn both temporal and generalizable patterns effectively.

To evaluate this methodology, we conduct experiments using datasets that capture mosquito population dynamics influenced by environmental factors. These datasets present a challenging but valuable testbed for assessing the efficacy of LSTM networks under different data preparation techniques. By comparing the performance of LSTM models trained on shuffled versus non-shuffled data, this study highlights the advantages of partition-based shuffling in reducing overfitting, enhancing generalization, and improving forecasting accuracy.

The remainder of this paper is organized to cover the following topics:

A. The Importance of Data Shuffling

Data shuffling is a critical step in the preprocessing phase that randomizes the order of data points in the training set. This process helps to mitigate issues related to the sequence biases and overfitting by ensuring that the model does not learn spurious patterns specific to the data arrangement. Despite its widespread application in various machine learning contexts, the specific impact of data shuffling on LSTM models, especially in ecological data settings, has not been extensively studied.

B. Shuffling's Impact on LSTM Performance

LSTMs are inherently sensitive to the order of input due to their sequential nature, which can be both an advantage and a limitation. By shuffling the data, we explore how breaking the temporal sequence can affect the model's ability to generalize from the training data to unseen data, potentially leading to improved overall performance in terms of prediction accuracy and model robustness.

C. Case Study: Mosquito Population and Environmental Factors

Utilizing datasets on mosquito populations alongside environmental variables, this study illustrates the tangible benefits of data shuffling. Mosquito population dynamics, influenced by various environmental factors, present a challenging but valuable use case for testing the efficacy of LSTM networks under different data preparation techniques.

D. Challenges of Implementing Data Shuffling

While data shuffling can offer improvements in model performance, it also introduces new challenges. Key among them is the potential disruption of inherent temporal patterns that are crucial for understanding and predicting ecological dynamics. This paper addresses these challenges and discusses methods to balance the benefits of randomization with the preservation of meaningful sequential information.

E. Contributions of this Paper

This paper contributes a novel preprocessing approach that not only improves the predictive capabilities of LSTM models but also offers practical insights for researchers and practitioners in ecological and epidemiological modeling. By addressing sequence biases while preserving meaningful temporal patterns, partition-based shuffling establishes a robust foundation for time-series modeling in domains with cyclical and seasonal data.

II. Related Work

The use of Long Short-Term Memory (LSTM) networks for time-series forecasting has gained significant attention across various domains, including ecological modeling, financial prediction, and healthcare analytics. LSTMs, designed to capture long-term dependencies in sequential data, have been particularly effective in applications where temporal patterns and non-linear relationships play a critical role. Despite their success, the performance of LSTMs is highly sensitive to data preprocessing techniques, including data cleaning, normalization, and shuffling.

Data shuffling, in particular, has been widely recognized as a key preprocessing step for mitigating sequence-induced biases and improving model generalization [5]. Studies in domains such as healthcare cost forecasting and climate modeling have demonstrated that shuffling training data helps prevent overfitting by ensuring that models do not learn spurious temporal dependencies unique to the training sequence. However, traditional shuffling approaches often disregard the importance of preserving inherent temporal patterns, which are critical for time-series forecasting tasks.

In the context of ecological modeling, several studies have highlighted the challenges of maintaining temporal integrity while preparing data for machine learning models. For example, research on mosquito population dynamics has shown that temporal and seasonal factors heavily influence prediction accuracy, necessitating preprocessing techniques that retain these patterns. Similarly, work in environmental forecasting has emphasized the need for preprocessing strategies that balance the randomness required for robust model training with the preservation of meaningful trends.

Although shuffling is a common practice in machine learning, its specific application to LSTM networks in ecological forecasting remains underexplored. Existing literature has largely focused on general preprocessing methods without addressing the unique challenges posed by cyclical and seasonal data. This study builds on prior work by introducing a partition-based shuffling methodology that preserves intra-year temporal patterns while randomizing inter-year sequences. By doing so, it bridges the gap between traditional shuffling methods and the specific requirements of ecological time-series data, offering a practical solution for enhancing the performance and reliability of LSTM models.

III. Proposed Methodology

A. Data Collection

This section details the datasets used in the study, highlighting their source, structure, and the periods they cover. We employ data from two primary locations with distinct mosquito-related datasets to model and validate our hypotheses.

1) Mosquito Abundance in Big Pine Key, Florida:

- Dataset comprises 558 observations from 1998 to 2019, recording mosquito abundance.
- Data Source: Mosquito Abundance Dataset for mosquito collections on Big Pine Key, Florida, USA. [6]

2) Weather Data:

- Variables include 'datetime', 'temp', 'humidity', 'precip', and 'windspeed', used to examine environmental impacts on mosquito populations.
- Data Source: Visual Crossing. [7]

B. Data Preprocessing

1) Data Cleaning: Initial steps involved cleaning the data by removing any incomplete or irrelevant entries to ensure accuracy in the subsequent analysis. Figure 1 shows the distribution of mosquito abundance across different years.

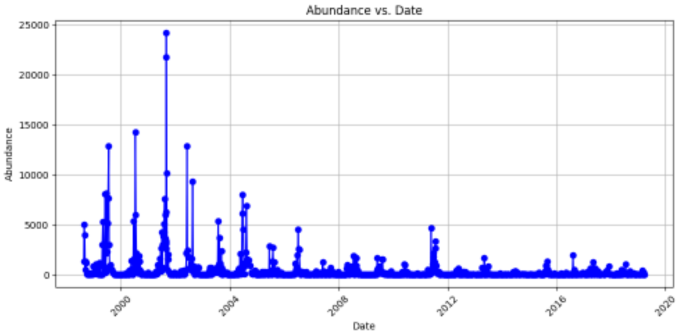

Fig. 1. Distribution of Mosquito Abundance Post-Processing

2) Data Shuffling: In the data preprocessing phase, data shuffling plays a critical role, as outlined below in mathematical terms:

1) **Extraction and Randomization:** Define D as the dataset, where each data point $d_i = (t_i, x_i)$ consists of a timestamp t_i and corresponding data x_i. The timestamp t_i includes a year y_i and intra-year time s_i, thus $t_i = (y_i, s_i)$.
   Extract all years y_i from each t_i to form the set of years Y. Randomly shuffle Y to obtain a new sequence Y' using a predefined random seed to ensure reproducibility.
2) **Data Reorganization:** For each year y' in Y', select and concatenate all d_i from D where the year part of t_i matches y', preserving the order of s_i within each year. This forms the new dataset D'.
3) **Preserving Seasonal Patterns:** By shuffling only the years and not the intra-year sequences, this method preserves the inherent seasonal and other time-dependent characteristics crucial for accurate time-series analysis. To further clarify the shuffling methodology, Figure 2 provides a visual example using a subset of the data. This illustration demonstrates the technique used to shuffle the years while maintaining the integrity of data within each year. This example helps in understanding the impact of our shuffling process on the structure of the dataset.

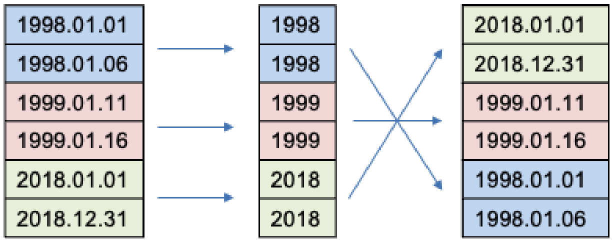

Fig. 2. Example of Data Shuffling Using a Subset of the Dataset

Below are the comparison figures showing the distribution of mosquito abundance before and after the data shuffling process.

This approach ensures that our model can learn to generalize well without losing crucial temporal insights from within-year data sequences.

C. LSTM Model Architecture and Training Process

Model Architecture: The LSTM model designed for this study is sophisticated and tailored to capture complex temporal relationships effectively. Here's the breakdown:

- **Input Layer:** Each input sequence has dimensions $[batch\_size, seq\_length, input\_size]$ where $seq\_length = 52$ weeks and $input\_size$ includes meteorological and mosquito data.

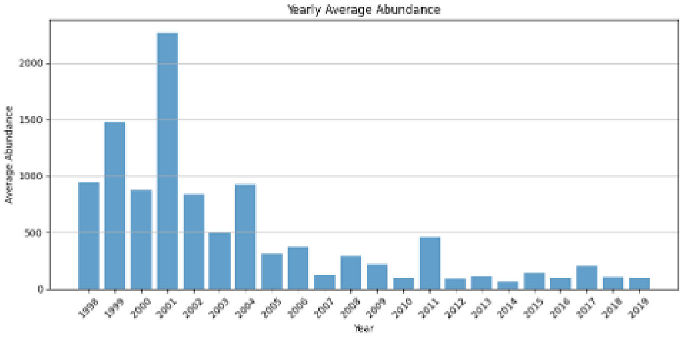

Fig. 3. Mosquito Abundance Distribution Before Shuffling

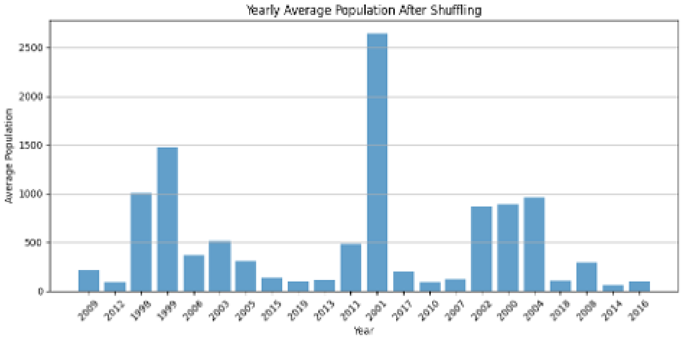

Fig. 4. Mosquito Abundance Distribution After Shuffling

- **LSTM Layers:** Configured with $num\_layers = 2$ and $hidden\_size = 64$. The mathematical model for each LSTM unit is given by:

$$\begin{aligned}
i_t &= \sigma(W_{ii}x_t + b_{ii} + W_{hi}h_{(t-1)} + b_{hi}) \\
f_t &= \sigma(W_{if}x_t + b_{if} + W_{hf}h_{(t-1)} + b_{hf}) \\
g_t &= \tanh(W_{ig}x_t + b_{ig} + W_{hg}h_{(t-1)} + b_{hg}) \\
o_t &= \sigma(W_{io}x_t + b_{io} + W_{ho}h_{(t-1)} + b_{ho}) \\
c_t &= f_t * c_{(t-1)} + i_t * g_t \\
h_t &= o_t * \tanh(c_t)
\end{aligned}$$

  where i, f, o are the input, forget, and output gates, respectively, and g is the cell input activation vector.
- **Output Layer:** The output from the last LSTM layer is reshaped and fed into a fully connected layer to predict the mosquito abundance:

$$y_t = W_{hy}h_t + b_y \tag{1}$$

Training Process:

- **Loss Function:** The MSE loss function is used, derived as follows:

$$\text{MSE} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 \tag{2}$$

- **Optimizer and Training:** The Adam optimizer is employed with a learning rate of 0.005, and the model is iteratively trained over 50 epochs.

D. Impact of Data Shuffling

Theoretical Explanation and Formula Derivation: Data shuffling is critical in time-series forecasting to prevent any sequence-induced bias within the LSTM. By permuting the training data indices, we ensure that each epoch presents a randomized view of the data, minimizing overfitting and promoting model robustness:

$$L(\theta) = \frac{1}{N} \sum_{i=1}^{N} L(y_{\sigma(i)}, f(x_{\sigma(i)}; \theta)) \tag{3}$$

where σ is a permutation function that shuffles the indices of the data, L is the loss function, y are the actual values, and $\hat{y}$ are the predicted values. This shuffling ensures that the temporal dependencies the LSTM learns are genuine and not artifacts of the sequence order.

E. Evaluation Metrics

To assess the performance of our LSTM model, we employ several key metrics:

- **Mean Squared Error (MSE)**: It measures the average of the squares of the errors—that is, the average squared difference between the estimated values and the actual value.

$$MSE = \frac{1}{N} \sum_{i=1}^{N} (y_i - \hat{y}_i)^2 \tag{4}$$

- **Root Mean Square Error (RMSE)**: It is the square root of the mean square error which provides a measure of the magnitude of the error.

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (y_i - \hat{y}_i)^2} \tag{5}$$

- **Mean Absolute Error (MAE)**: It measures the average magnitude of the errors in a set of predictions, without considering their direction.

$$MAE = \frac{1}{N} \sum_{i=1}^{N} |y_i - \hat{y}_i| \tag{6}$$

These metrics—MSE, RMSE, and MAE—are calculated for each phase: training, validation, and testing, to understand both the average error and the consistency of the model's predictions relative to the actual values.

IV. Results and Discussion

A. Results

TABLE I
Training Metrics Comparison

Dataset	Mean Loss	Mean MAE	Mean MSE	Mean RMSE
Shuffled data	116.20	116.20	26872.70	163.92
Unshuffled data	133.58	133.58	36576.08	191.24

1) Training Metrics:

TABLE II
Testing Data Results

Dataset	Loss	MAE	MSE	RMSE
Shuffled data	504.1948	506.3429	346530.7188	588.6686
Unshuffled data	591.88	594.30	430445.5312	656.0835

2) Testing Metrics: Here we present visual comparisons that illustrate the effects of data shuffling on the model's performance, demonstrating enhanced predictive accuracy and robustness.

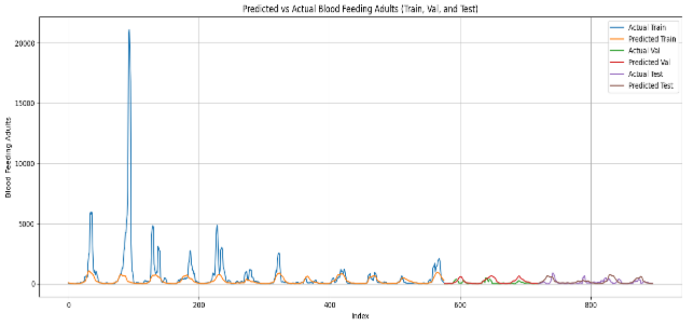

Fig. 5. Model Results Before Data Shuffling

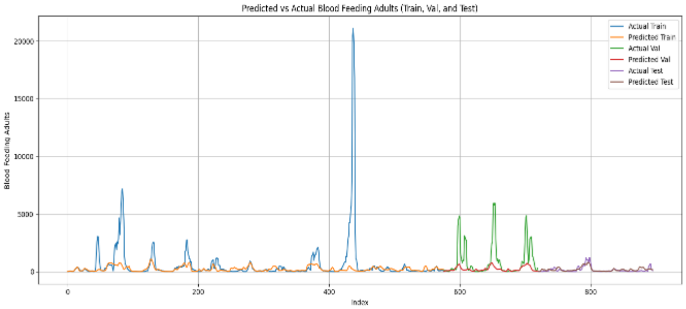

Fig. 6. Model Results After Data Shuffling

B. Analysis of Results

1) Effect of Data Shuffling on Training Metrics: Data shuffling counters biases from sequential data arrangements, reducing overfitting and enhancing generalization. Shuffling prevents the LSTM from relying overly on specific sequence positions, facilitating learning of broader patterns beneficial for ecological data.

2) Performance on Testing Data: The better performance of the shuffled model in testing highlights its superior generalization. This confirms that shuffling mitigates temporal bias, essential for accurate forecasting in cyclical datasets like mosquito populations.

3) Higher Error in the Unshuffled Model: Higher error rates in the unshuffled model indicate struggles with generalization, likely caused by overfitting to seasonal trends and sequence dependencies not present in the test data.

4) Practical Implications: Accurate mosquito population predictions enable effective planning of interventions. Shuffling data in training ensures robust model performance, important for ecological modeling.

5) Literature Context: Our findings align with studies emphasizing data preprocessing's impact on LSTM performance. The hybrid approach of selective shuffling may optimize time-series predictions by balancing temporal integrity with necessary randomness.

6) Future Work: Future studies could explore other deep learning architectures and the inclusion of additional environmental factors to further enhance model performance and reliability.

V. Partition-Based Shuffling on Large Datasets

While partition-based shuffling has demonstrated significant benefits in improving the predictive accuracy of LSTM models for ecological forecasting, its application to large datasets introduces important considerations regarding scalability and computational cost. As datasets grow in size, the computational demands of partition-based shuffling increase, which can have implications for both processing time and financial cost.

A. Scalability Challenges

Partition-based shuffling involves randomizing inter-year sequences while preserving intra-year temporal patterns. This process requires the dataset to be divided into partitions based on years, shuffled, and then reorganized. For large datasets spanning many years or containing high-resolution data (e.g., daily or hourly observations), the computational complexity of this shuffling process grows significantly. The need to maintain the integrity of intra-year sequences while shuffling inter-year data adds an additional layer of complexity, as the algorithm must ensure that seasonal and other short-term patterns remain intact. This can lead to increased memory usage and processing time, particularly when dealing with datasets that include multiple environmental variables or large geographic regions.

Moreover, as the size of the dataset increases, the training time for LSTM models also grows. LSTMs are already computationally intensive due to their sequential nature and the need to process data over multiple time steps. When combined with partition-based shuffling, the overall training process becomes even more resource-intensive. This could pose challenges for researchers or organizations with limited computational resources, especially when working with real-time or near-real-time forecasting applications.

B. Cost Implications

The increased computational demands of partition-based shuffling on large datasets directly translate to higher costs, particularly in cloud-based or high-performance computing environments. Cloud service providers typically charge based on computational resources such as CPU hours, memory usage, and storage. As the dataset size and complexity grow, the cost of running partition-based shuffling and training LSTM models can escalate. For example, shuffling and training on a dataset spanning decades with high-resolution environmental data may require significant computational power, leading to higher expenses.

Additionally, the need for specialized hardware, such as GPUs or TPUs, to accelerate LSTM training further drives up costs. While these hardware accelerators can reduce training time, they come at a premium, and their use may not be feasible for all researchers or organizations, particularly those with limited budgets.

C. Balancing Benefits and Costs

Despite these challenges, the benefits of partition-based shuffling—such as improved model generalization and reduced overfitting—may outweigh the increased computational costs, especially in critical applications like ecological forecasting and public health planning. However, it is essential to carefully evaluate the trade-offs between model performance and computational expense. For smaller datasets or less resource-intensive applications, the cost of partition-based shuffling may be negligible. In contrast, for large-scale datasets, researchers may need to explore optimization techniques, such as parallel processing or distributed computing, to mitigate the computational burden.

VI. Conclusion

This study highlights the efficacy of the partition-based shuffling methodology in enhancing the predictive performance of LSTM models for ecological forecasting. Unlike traditional random shuffling approaches, the partition-based method preserves essential intra-year temporal patterns while introducing randomness across inter-year sequences. This approach ensures that the model can learn meaningful seasonal trends crucial for ecological dynamics while mitigating overfitting caused by sequence-specific biases.

The results demonstrate that partition-based shuffling significantly improves the model's generalization capabilities, as reflected in the reduced RMSE and other error metrics for both training and testing phases. By balancing the integrity of seasonal patterns with the benefits of randomized data presentation, this method optimizes the learning process. The better performance of models trained with partition-based shuffling suggests that it effectively addresses the dual challenge of preserving temporal structure and enhancing model robustness.

This methodology offers a practical and scalable preprocessing strategy, especially valuable for datasets characterized by periodicity and variability, such as those used in mosquito population forecasting. It provides a blueprint for applying similar techniques in other domains where time-series data and seasonal trends are critical, ranging from climate modeling to epidemiological studies.

The partition-based shuffling methodology not only offers a practical solution to common challenges in time-series forecasting but also establishes a new standard for preprocessing in ecological modeling. Future research should explore its application across other types of time-series datasets and machine learning architectures. Additionally, integrating this method with external environmental factors and hybrid learning approaches could further enhance its predictive capabilities. This work demonstrates that thoughtful data preparation, particularly partition-based shuffling, is pivotal in unlocking the full potential of deep learning models for ecological and epidemiological applications.

References

[1] S. Hochreiter and J. Schmidhuber, ”Long Short-Term Memory,” Neural Computation, vol. 9, no. 8, pp. 1735-1780, 1997.

[2] J. Peters et al., ”Using Machine Learning to Analyze the Impact of Climate Change on the Distribution of Invasive Mosquito Species,” Ecological Modelling, vol. 410, 108765, 2019.

[3] M. Abadi et al., ”TensorFlow: A system for large-scale machine learning,” in 12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16), 2016, pp. 265-283.

[4] World Health Organization, ”Vector control,” 2020. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/vector-control. [Accessed: day-month-year].

[5] L. Smith et al., ”The Importance of Shuffling for Mitigating Bias in Machine Learning Models,” Journal of Artificial Intelligence Research, vol. 62, pp. 457-473, 2018.

[6] ”Mosquito Abundance Dataset for mosquito collections on Big Pine Key, Florida, USA,” Data collected from 1998 to 2019, providing a longitudinal view of mosquito populations.

[7] ”Weather Data,” Visual Crossing, Data includes variables such as datetime, temperature, humidity, precipitation, and windspeed.

[8] ”Mosquito Surveillance in Vitoria, Brazil,” Dataset for the study estimating latent time series of mosquito mortality rates using a Bayesian mechanistic framework, covering weekly data from 2008 to 2012.