Tabular vs. Temporal: Modeling the Impact of CO_2 Emissions on Global Forest Loss using Regression and Sequence Learning

Keonvin Park [1]

[1] *Interdisciplinary Program in Artificial Intelligence, Seoul National University, Seoul, Republic of Korea*. Correspondence to: Keonvin Park kbpark16@snu.ac.kr.

1. Introduction

Forests are vital to global climate regulation and carbon storage, yet deforestation continues at a rate exceeding 10 million hectares per year [1]. Greenhouse gas (GHG) emissions not only result from but also contribute to deforestation, forming a self-reinforcing feedback loop.

The Global Forest Watch (GFW) platform [1] provides high-resolution annual data on tree cover loss and carbon emissions. Prior studies have leveraged GFW to classify deforestation drivers [2], quantify degradation [3], and analyze policy impacts [4]. Recent work also includes machine learning models such as random forests [5] and graph neural networks [6].

Despite this, limited attention has been paid to systematically comparing tabular and sequence-based models for forecasting forest loss, especially given the short and structured nature of historical forest datasets.

In this study, we investigate the relationship between CO_2 emissions and tree cover loss in the top 10 deforestation-prone countries using GFW data. We benchmark several tabular and sequential models to assess their suitability for this forecasting task.

2. Methods

We compare two modeling paradigms for forecasting annual forest loss from CO_2 emissions: tabular and sequence models.

Tabular models treat each country-year (x, y) independently, using current-year CO_2 emissions as input. We implement:

- Linear Regression
- Random Forest [7]
- GBDT [8]
- XGBoost [9]

Sequence models use rolling input sequences $\{x_{t-k}, ..., x_{t-1}\}$ to predict y_t, capturing temporal patterns. We use:

- RNN [10]
- LSTM [11]

All models are trained on country-level annual data from 2001–2021 without strict train/test splits, emphasizing comparative trends over generalization.

3. Data

We use annual forest loss (ha) and CO_2-equivalent emissions (Mg) data from the **Global Forest Watch (GFW)** platform[1]. Our analysis focuses on the top 10 countries by cumulative tree loss from 2001–2024.

Tabular input uses single-year features for static models. **Sequence input** uses fixed-length windows of prior years for RNN/LSTM.

Table 1 summarizes the variability in annual tree cover loss across selected countries.

Table 1: Summary statistics of tree cover loss (ha), 2001–2024

Country (ISO)	Mean	Std
AUS	384,097	517,218
BOL	407,445	352,818
BRA	3,054,878	844,980
CAN	2,610,311	1,473,990
CHN	531,947	140,433
COD	878,014	418,622
IDN	1,331,796	461,416
MYS	396,398	124,188
RUS	3,701,295	1,454,337
USA	2,060,953	359,360

4. Results

4.1 Correlation Analysis

CO_2 emissions and forest loss exhibit strong positive correlations in most countries, with R^2 values often exceeding 0.9 (e.g., CAN, COD, BOL), as shown in Table 2. The U.S. shows a weaker relationship ($R^2 = 0.31$), indicating other potential influencing factors.

4.2 Regression Fit Example

Figure 1 illustrates regression fits for RUS and BRA, two countries with high cumulative forest loss. Both show high alignment between emissions and loss trends.

4.3 Model Performance Comparison

Table 3 and 4 present RMSE scores for tabular and sequence models. Tree-based tabular models, especially XGBoost, consistently outperform RNN and LSTM across all countries. Temporal models underperform, likely due to short sequence length (24 years).

[1] `https://www.globalforestwatch.org/`

Table 2: Per-country linear regression results: CO_2 emissions as predictor of tree cover loss

Country (ISO)	Correlation	R^2	P-value
AUS	0.9827	0.9658	0
BOL	0.9889	0.9779	0
BRA	0.9391	0.8819	0
CAN	0.9931	0.9863	0
CHN	0.9694	0.9398	0
COD	0.9986	0.9972	0
IDN	0.9730	0.9468	0
MYS	0.9513	0.9050	0
RUS	0.9383	0.8804	0
USA	0.5556	0.3087	0.0048

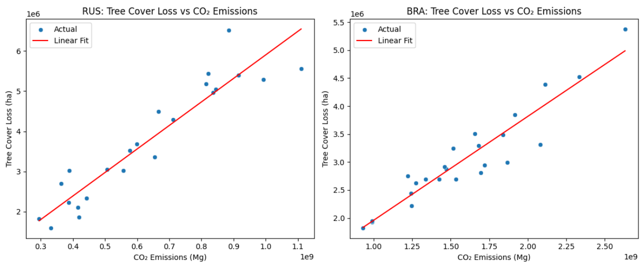

Fig. 1: Linear regression fit for tree cover loss vs. CO_2 emissions in Russia and Brazil

Table 3: RMSE (lower is better) for Tabular Models

Country	Linear	RF	GBDT	XGBoost
AUS	0.0419	0.0476	0.0028	**0.0012**
BOL	0.0307	0.0530	**0.0012**	0.0015
BRA	0.0798	0.0446	0.0045	**0.0015**
CAN	0.0229	0.0475	0.0012	**0.0011**
CHN	0.0664	0.0314	0.0046	**0.0013**
COD	0.0181	0.0155	0.0001	**0.0010**
IDN	0.0555	0.0268	0.0010	**0.0011**
MYS	0.0811	0.0368	0.0039	**0.0013**
RUS	0.0999	0.0391	0.0043	**0.0012**
USA	0.2145	0.0626	0.0073	**0.0012**

Table 4: RMSE for Sequence Models (RNN, LSTM)

Country	RNN	LSTM
AUS	0.1170	**0.1333**
BOL	**0.0928**	0.0891
BRA	**0.1400**	0.2440
CAN	0.1156	**0.0878**
CHN	0.1184	**0.1280**
COD	**0.0906**	0.1187
IDN	**0.2007**	0.1991
MYS	**0.2223**	0.1152
RUS	**0.1273**	0.2238
USA	**0.1914**	0.1736

Summary

Tabular models achieve consistently lower RMSE than sequence models. XGBoost, in particular, balances predictive accuracy and robustness across countries.

5. Conclusion

This study explored the predictive relationship between greenhouse gas emissions and annual tree cover loss using Global Forest Watch (GFW) data. Focusing on the top 10 countries with the most significant cumulative forest loss since 2001, we conducted both correlation analysis and comparative forecasting experiments using tabular and sequential modeling paradigms.

Our findings indicate a strong linear correlation between CO_2 emissions and tree cover loss in most countries, with R^2 values exceeding 0.95 in several cases (e.g., Canada, Congo, Bolivia). However, this correlation is not universal: countries such as the United States exhibited weaker associations, suggesting additional socioeconomic or ecological factors at play.

In terms of model performance:

- **Tree-based tabular models**, especially Gradient Boosted Decision Trees (GBDT) and XGBoost, consistently achieved the lowest RMSE values across countries. These models were particularly effective in capturing the non-linear relationships between emissions and forest loss in a static, year-level context.
- **Sequential models** like RNNs and LSTMs were less effective overall. Despite their theoretical advantage in modeling temporal dependencies, their performance was often inferior to simpler models. This is likely due to the limited sequence length (24 years) and small sample size per country, which restrict the benefits of deep learning-based sequence modeling.
- **XGBoost emerged as the best-performing model overall**, combining low RMSE with robust generalization across diverse national contexts.

These results underscore a counterintuitive but important insight: in environmental forecasting tasks with relatively short temporal histories and limited features, traditional tabular models can outperform more complex sequence models. This emphasizes the need to match modeling complexity to the structure and scale of available data.

Future Work

Future research could integrate additional environmental and socioeconomic variables (e.g., precipitation, land use, policy data) to improve predictive accuracy. Modeling cross-country dependencies using graph neural networks may reveal regional deforestation dynamics. Uncertainty-aware methods and causal inference frameworks could further enhance model interpretability and policy relevance.

Acknowledgments

This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) [NO.RS-2021-II211343, Artificial Intelligence Graduate School Program (Seoul National University)]

References

[1] Matthew C. Hansen, Peter V. Potapov, Rebecca Moore, Matt Hancher, Svetlana A. Turubanova, Alexandra Tyukavina, David Thau, Stephen V. Stehman, Scott J. Goetz, Thomas R. Loveland, Aruna Kommareddy, Andrey Egorov, Louise Chini, Christopher O. Justice, and John R. G. Townshend. High-resolution global maps of 21st-century forest cover change. *Science*, 342(6160):850–853, 2013.

[2] Philip G. Curtis, Christy M. Slay, Nancy L. Harris, Alexandra Tyukavina, and Matthew C. Hansen. Classifying drivers of global forest loss. *Science*, 361(6407):1108–1111, 2018.

[3] Alexandra Tyukavina, Matthew C. Hansen, Peter Potapov, Andrey Krylov, Scott J. Goetz, Evgeny Lopatin, and Svetlana Turubanova. Aboveground carbon loss in natural and managed tropical forests from 2000 to 2012. *Environmental Research Letters*, 10(7):074002, 2015.

[4] Kemen G. Austin, Aimee Schwantes, Yifan Gu, and Prasad S. Kasibhatla. What causes deforestation in indonesia? *Environmental Research Letters*, 14(2):024007, 2019.

[5] Carlos H. L. Silva Junior, Andrea C. Pessôa, Nadja S. Carvalho, José B. C. Reis, Lope T. Anderson, and Luiz E. O. C. Aragão. Random forest modeling of amazonian deforestation hotspots. *Remote Sensing of Environment*, 240:111692, 2020.

[6] Lian Zhu, Yujing Wang, Tianqi Li, and Xiangyu Chen. Spatio-temporal graph neural networks for global deforestation forecasting. *International Journal of Remote Sensing*, 43(17):6243–6262, 2022.

[7] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[8] Jerome H. Friedman. Greedy function approximation: A gradient boosting machine. *Annals of statistics*, pages 1189–1232, 2001.

[9] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794, 2016.

[10] John J Hopfield. Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the national academy of sciences*, 79(8):2554–2558, 1982.

[11] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.