

# Analyzing the Relationship Between Cosmic Rays and Total Cloud Cover with LSTM Networks

## Kozmik Işınlar ve Toplam Bulut Örtüsü Arasındaki İlişkinin LSTM Ağlarıyla Analizi

Ahmet POLATOĞLU 

Atatürk University, Department of Astronomy  
and Space Science, Faculty of Sciences,  
Erzurum, Turkey



### Abstract

Understanding the interactions between cosmic phenomena and terrestrial weather patterns, particularly the relationship between cosmic rays (CRs) and cloud cover, has been a significant scientific endeavor. CRs, high-energy particles originating from supernovae, can ionize air molecules upon entering Earth's atmosphere, potentially influencing cloud formation. Cloud cover plays a vital role in Earth's climate system by regulating energy balance through reflecting solar radiation and trapping infrared radiation. This study aims to analyze the relationship between CRs and Total Cloud Cover (TCC) globally using Long Short-Term Memory (LSTM) networks, a type of recurrent neural network suited for time series data. We used data from the University of Oulu's Cosmic Ray Station and the Copernicus Climate Change Service's ECMWF European Reanalysis V5 (ERA5). A correlation matrix was constructed to identify relationships between CRs and TCC across various regions, including the Antarctic, Arctic, Europe, and globally. The results indicated generally weak and negative correlations between CR and TCC, with weak positive correlations in the Southern Hemisphere and globally. Negative correlations were more pronounced in the Antarctic and Arctic regions, suggesting region-specific climate mechanisms. The LSTM model's predictions of CR values did not closely follow actual values, indicating a significant gap in capturing dynamic changes and peaks, and suggesting the need for more data, additional features, or further tuning. The training process showed rapid initial learning but overfitting after several epochs. The final model's performance, measured by test mean squared error (MSE), suggested inadequate generalization. These findings highlight the complexity of modeling the CR-TCC relationship using machine learning. Future research should focus on enhancing data quality, incorporating detailed cloud metrics, and exploring advanced models to better understand CRs influence on cloud formation and climate. This study contributes to the debate on CR role in climate systems, providing insights for improved climate models and predictions.

**Keywords:** Cosmic Ray (CR), Total Cloud Cover (TCC), Long Short-Term Memory (LSTM) Networks, Climate Modeling.

### Öz

Dünya dışı radyasyon ile yeryüzündeki hava olayları arasındaki etkileşimleri, özellikle kozmik ışınlar (CR) ile bulut örtüsü arasındaki ilişkiyi anlamak, son yıllarda önemli bir bilimsel alan oluşturmuştur. Süpernova gibi kaynaklardan oluşan yüksek enerjili kozmik parçacıklar, Dünya atmosferine girdiklerinde hava moleküllerini iyonize edebilmekte ve bu da potansiyel olarak bulut oluşumunu etkileyebilmektedir. Bulut örtüsü, güneş radyasyonunu yansıtarak ve kızılötesi radyasyonu tutarak enerji dengesini düzenlediği için Dünya'nın iklimi üzerinde hayati bir rol oynar. Bu çalışma, CR ile Toplam Bulut Örtüsü (TCC) arasındaki ilişkiyi küresel ölçekte analiz etmeyi amaçlamaktadır. Bu amaçla, zaman serisi verileri için uygun bir tekrarlayan sinir ağı türü olan Uzun Kısa Dönemli Bellek (LSTM) ağları kullanılmıştır. CR verileri, Oulu Üniversitesi Kozmik Işın İstasyonu'ndan ve bulut verileri de Copernicus Climate Change Service's ECMWF European Reanalysis V5 (ERA5)'in Maine Üniversitesi climate reanalyzer sayfası üzerinden alınmıştır. CR'ler ve TCC arasındaki ilişkileri belirlemek için Antarktika, Arktik, Avrupa ve küresel dahil olmak üzere çeşitli bölgelerde bir korelasyon matrisi oluşturulmuştur. Sonuçlar, CR ve TCC arasında genellikle zayıf negatif korelasyonlar olduğunu, Güney Yarımküre ve küresel ölçekte zayıf pozitif korelasyonlar olduğunu göstermiştir. Antarktika ve Arktik bölgelerinde negatif korelasyonlar daha belirgindir ve bu da bölgeye özgü iklim mekanizmalarını işaret eder. LSTM modelinin CR değerlerinin tahminleri, dinamik değişiklikleri ve zirveleri yakalamada önemli bir boşluk olduğunu ve daha fazla veri, ek özellikler veya daha fazla parametre gereksinimi olduğunu göstererek, gerçek değerleri yakından takip etmemiştir. Eğitim süreci hızlı başlangıç öğrenmesi göstermiş ancak birkaç dönemden sonra aşırı öğrenme (overfitting) ortaya çıkmıştır. Nihai modelin performansı, test ortalama kare hatası (MSE) ile ölçüldüğünde, yetersiz genelleme yapıldığını göstermiştir. Bu bulgular, makine öğrenimi kullanarak CR-TCC ilişkisini modellemenin karmaşıklığını vurgulamaktadır. Gelecek araştırmalar, veri kalitesini artırmaya, ayrıntılı bulut metriklerini eklemeye ve CR'lerin bulut oluşumu ve iklim üzerindeki etkisini daha iyi anlamak için ileri modelleri keşfetmeye odaklanmalıdır. Bu çalışma, CR'nin iklim sistemlerindeki rolü hakkındaki tartışmalara katkıda bulunarak, daha iyi iklim modelleri ve tahminleri için öngörüler sunmaktadır.

**Anahtar Kelimeler:** Kozmik Işın (CR), Toplam Bulut Örtüsü (TCC), Uzun Kısa-Dönemli Bellek (LSTM) Ağları, İklim Modelleme.

Corresponding Author/ Sorumlu Yazar:

A. POLATOĞLU

E-mail:

[ahmet.polatoglu@atauni.edu.tr](mailto:ahmet.polatoglu@atauni.edu.tr)

Received/ Geliş Tarihi 26.04.2024

Accepted/Kabul Tarihi 31.05.2024

Publication Date/ 27.06.2024

Yayın Tarihi

Cite this article

Polatoglu, A. (2024). *Analyzing the Relationship Between Cosmic Rays and Total Cloud Cover with LSTM Networks*. *Journal of Anatolian Physics and Astronomy*, 3(1), 19-26.



Content of this journal is licensed under a Creative Commons Attribution-Noncommercial 4.0 International License.

## Introduction

Understanding the intricate interactions between cosmic phenomena and terrestrial weather patterns has long been a topic of scientific inquiry. One such interaction that has garnered significant attention is the relationship between cosmic ray (CR) and cloud cover. CRs are high-energy particles thought to originate from outside our solar system, primarily from supernovae (Polatoğlu et al., 2023). When these particles enter Earth's atmosphere, they can ionize air molecules, potentially influencing cloud formation processes (Carslaw et al., 2002). Cloud cover plays a critical role in Earth's climate system by regulating the planet's energy balance through the reflection of solar radiation and the trapping of infrared radiation (IPCC, 2021). As such, understanding the factors influencing cloud cover variability is crucial for accurate climate modeling and prediction.

CRs reaching Earth's troposphere are influenced by solar activity. Studies from about 30 years ago suggested a connection between CR and low-cloud cover, showing variations in cloud cover between solar maximum and minimum periods. This link is significant as clouds impact Earth's radiative balance and climate. Recent research also indicates correlations between CR and cloud cover on shorter timescales. However, there's debate over these connections due to methodological differences among studies. Despite this uncertainty, the potential impact of CR on climate has sparked interest in understanding how CR might affect clouds and historical climate change. The ion-aerosol clear-sky hypothesis has emerged as a prominent proposed mechanism for the CR-cloud-climate connection. Previous studies have suggested a potential link between CRs and cloud cover, with variations in CR flux potentially affecting cloud nucleation processes (Kirkby et al., 2011; Svensmark et al., 2007).

The exact nature and significance of this relationship remain contentious, with conflicting results and a need for more robust statistical analyses and observational data (Erylkin & Wolfendale, 2010; Laken et al., 2012). Machine learning (ML) techniques offer a promising avenue for exploring these complex relationships. With their ability to handle large datasets and uncover non-linear patterns, ML methods can provide new insights into the interactions between CRs and cloud cover that traditional statistical approaches might miss (Goodfellow et al., 2016). By leveraging advanced ML algorithms, we can analyze vast amounts of meteorological and cosmic ray data to identify potential correlations and causal mechanisms with greater precision.

In this study, we aim to analyze the relationship between CRs and Total Cloud Cover (TCC) globally using state-of-the-art deep learning techniques. We utilized a comprehensive dataset comprising CR flux measurements, cloud cover observations, and various meteorological parameters collected over several years. Our approach involves the application of supervised and unsupervised ML algorithms to identify patterns and test hypotheses regarding the influence of CRs on cloud formation. We begin by providing an overview of the data sources and preprocessing steps. Next, we detail the machine learning methodologies employed, including feature selection, model training, and validation. The results section presents our findings, highlighting significant correlations and the potential implications for climate science. Finally, we discuss the limitations of our study and suggest directions for future research. Through this research, we hope to contribute to the ongoing debate about the role of cosmic rays in cloud formation and, by extension, their influence on Earth's climate system. Our findings could enhance our understanding of the factors driving cloud cover variability and improve the accuracy of climate models, with broader implications for global climate predictions.

## Material and Methods

### Data Collection

TCC data are taken from Climate reanalysis website (<https://climatereanalyzer.org/>). Climate reanalysis describes the use of physically grounded numerical models to mimic Earth's climate throughout time, with frequent input from real-world observations (such as satellites, weather stations, radiosondes, and ocean buoys). In situations where direct observations are unavailable, reanalysis models are a crucial resource for comprehending climatic variability and change. For the most popular reanalysis products, common meteorological variables are accessible from this page. The NCAR Climate Data Guide is a good resource for users to learn more about reanalysis, including its methodology, advantages, disadvantages, and product comparisons. Advancing Reanalysis has further information about reanalysis. Copernicus C3S provided the ECMWF European Reanalysis V5 (ERA5) (0.25°x0.25°) download. Here, the data files are regridded using bilinear interpolation to

0.5°x0.5° in order to minimize access time and server burden. Gridded data products use interpolation techniques to fill in data gaps by overlaying point-based or spatially discontinuous measurements of Earth's climate (such as temperature, precipitation, wind, and sea surface temperature) onto time-registered grids. Similar to reanalysis, gridded datasets are helpful for studying climate in regions without direct observations, but where relevant information can be obtained by combining nearby input data. Numerous frequently used gridded datasets are accessible through Climate Reanalyzer. The Climate Change Institute at the University of Maine in the United States provided Climate Reanalyzer to compile this data (<https://climatechange.umaine.edu>).

As seen in Table 1, CR data was selected globally. However, Total Cloud Cover (TCC) data was selected based on different regions of the world. Cloud parameters are abbreviated as T1, T2, ..., T7. The cosmic ray data utilized in this research was sourced from the Cosmic Ray Station, a component of the University of Oulu's Sodankylä Geophysical Observatory (<https://cosmicrays oulu.fi/>).

**Table 1.** CR and TCC Parameters

Code	Parameters/Region	Coordinate
CR	Cosmic Ray (count/min) - World	Globally
T1	TCC (%) - Antarctica	90° S – 66.5° S, 0° E – 360° E
T2	TCC (%) - Arctic	66.5° N – 90° N, 0° E – 360° E
T3	TCC (%) - Europe	35° N – 60° N, 12° W – 30° E
T4	TCC (%) - Northern Hemisphere	0° – 90° N, 0° E – 360° E
T5	TCC (%) - Southern Hemisphere	90° S – 0° , 0° E° – 360° E
T6	TCC (%) - Atlantic Ocean	50° S – 60° N, 70° W – 10° E
T7	TCC (%) - World	90° S – 90° N, 0° E – 360° E

The station is situated in northern Finland, renowned for its minimal light pollution and conducive atmospheric conditions for cosmic ray observation. The data collection period covers the years from 1964 to 2023, capturing a comprehensive temporal scope for analysis. The Cosmic Ray Station employs a sophisticated network of neutron monitors, muon detectors, and other specialized instruments to capture CR interactions with Earth's atmosphere. These detectors are strategically positioned to mitigate environmental influences and ensure the fidelity of collected data. Continuous monitoring and data logging mechanisms are implemented to ensure uninterrupted data acquisition throughout the observation period. Quality control protocols are enforced to identify and rectify any anomalies or instrumental artifacts that may affect data integrity

## Deep Learning Methods

In this study, Long Short-Term Memory (LSTM), an advanced deep learning technique, is used to analyse the relationship between CR and TCC from 1964 to 2024. This method is particularly suited for time series data, allowing us to capture temporal dependencies and complex patterns within our dataset. LSTM networks are a type of recurrent neural network (RNN) architecture designed to effectively learn and remember long-term dependencies in sequential data. Traditional RNNs suffer from the vanishing gradient problem, where gradients diminish exponentially as they are backpropagated through time, making it difficult for the network to learn long-term dependencies. LSTMs address this issue through their unique cell structure, which includes three gates: the input gate, forget gate, and output gate. These gates regulate the flow of information, allowing the network to retain relevant information over long periods and discard unnecessary information. The core component of an LSTM is its cell state, a memory buffer that runs through the entire sequence, and is modified by the gates to store and manage long-term information. The input gate decides which information from the current input and the previous hidden state should be updated in the cell state. The forget gate determines which information in the cell state should be removed, ensuring that irrelevant data does not clutter the memory. The output gate controls what part of the cell state should be outputted and used for the current hidden state. This gated mechanism allows LSTMs to maintain a more constant error gradient, addressing the issues of both the vanishing and exploding gradient problems found in standard RNNs. These mechanisms allow the LSTM network to retain relevant information over long periods, making it powerful for time series prediction tasks. For our analysis, we implemented the LSTM model using the TensorFlow/Keras library (Graves & Graves, 2012; Hochreiter & Schmidhuber, 1997; Sherstinsky, 2020). The architecture includes multiple LSTM layers

followed by dense layers to output the final predictions for TCC based on CR data.

Before training the models, the CR and TCC data underwent several preprocessing steps:

- i) Normalization: Both CR and TCC data were normalized to ensure the models' efficient training and convergence.
- ii) Sequence Generation: The data were divided into sequences of appropriate length to train the RNN-based models.
- iii) Train-Test Split: The dataset was split into training and testing sets to evaluate the models' performance on unseen data.

By applying these advanced machine learning techniques, we aim to uncover potential correlations and causal relationships between CRs and TCC, contribute valuable insights to climate science research.

## Results and Discussion

The correlation matrix shows the relationships between various variables (CR, T1, T2, T3, T4, T5, T6 and T7). The correlation coefficients range from -1 to 1, where 1 indicates a perfect positive correlation, -1 indicates a perfect negative correlation and 0 indicates no correlation. In general, the correlations in the lower right part of the matrix appear to be stronger and more positive, indicating that T4, T5, T6 and T7 are more tightly correlated with each other (Figure 1). On the other hand, the correlations between CR and other variables are generally weak and negative, suggesting that CR has no apparent relationship with these variables or there may be an inverse relationship. That is, this matrix shows that there are some strong positive relationships and weak negative relationships between certain variables. This type of analysis is useful to understand which variables act together and which are independent. A positive correlation between CR and Cloud formation is expected, whereas here it is mostly negative correlations or very weak positive correlations. T5 (Southern Hemisphere) and T7 (World) TCC values were positively correlated with CR, albeit weakly. In T1 and T2, i.e. Antarctic and Arctic polar regions, negative correlations of 0.28 and 0.19 were observed. It is understood that different climate mechanisms come into play in different regions and change these connections. One thing should not be forgotten here. The positive correlation between CR and cloud is only observed between Low Cloud Cover (LTCC) in the literature (Marsh & Svensmark, 2000). However, here we have analysed directly with TCC. For this reason, we can say that there is an important relationship between the layer of clouds and CR. In Figure 2, the Time Series graph of all parameters is drawn. The TCC values worldwide and the values in T4, T5 and T6 regions are parallel. However, changes are observed in other regions. In summary, it is difficult to see a significant correlation between CR and TCC.

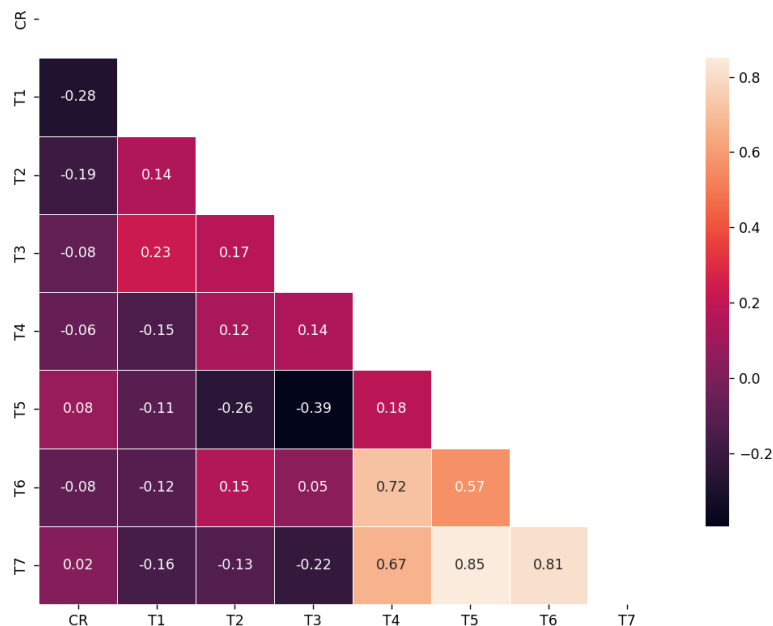
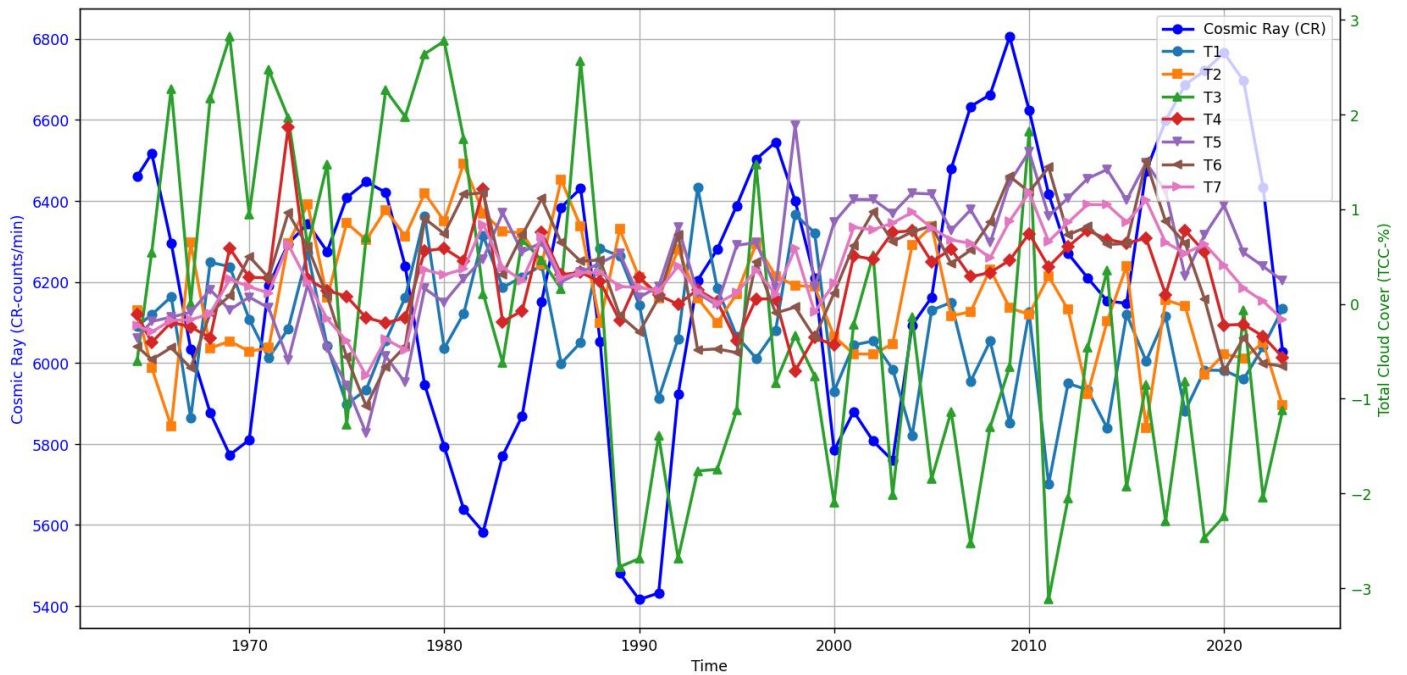
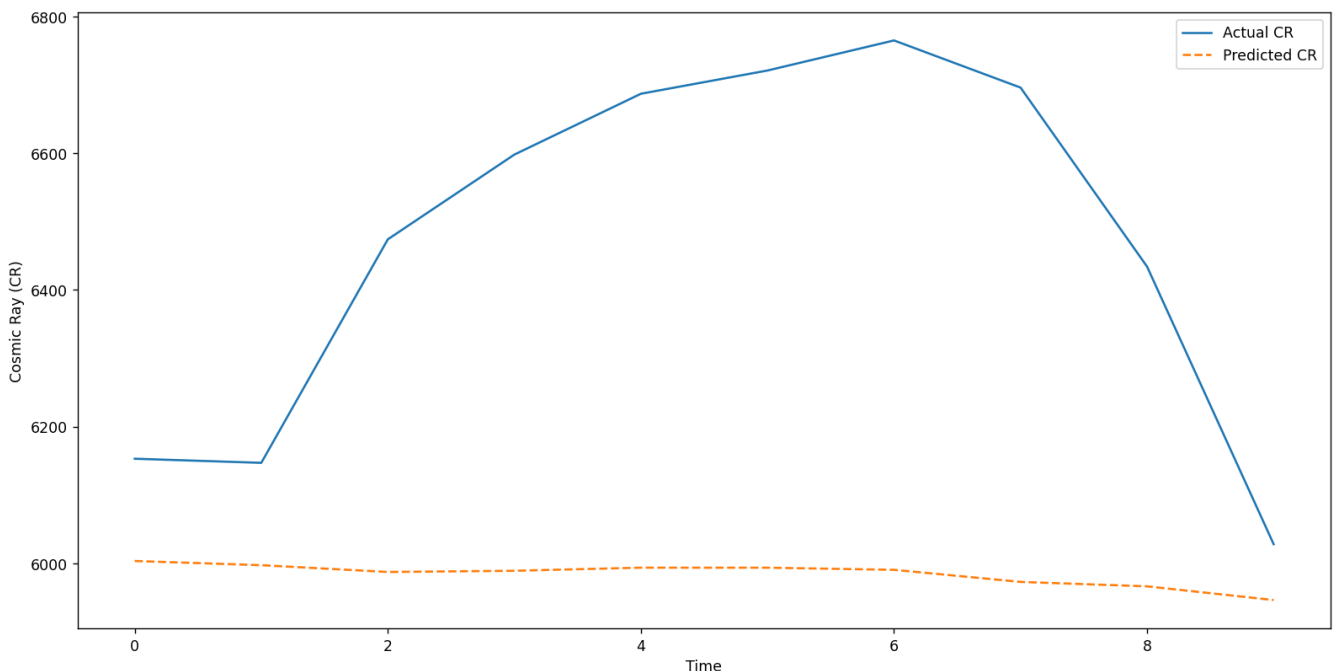


Figure 1. Correlations of CR and TCC Parameters



**Figure 2.** Time Series of CR and TCC

The predicted CR values estimated using the LSTM network do not closely follow the actual CR values (Figure 3). The predictions are quite different from the trends and changes of the actual values and follow an almost constant line. While the actual CR values show a significant seasonality or time-dependent variation, the model's predictions fail to capture these variations. This indicates that the model cannot accurately learn or represent dynamic changes and peaks. To improve the performance of the model, more data, more complex models or retraining of the model may be required. It may also be considered to add new featurettes so that the model can better capture changes. The fact that the predictions in the graph are not close to the actual CR values indicates that the current state of the model is inadequate and significant improvements are required.



**Figure 3.** Actual and Predicted CR Values



Training Loss and Validation Loss decrease rapidly in the first few epochs. The term loss is used as a metric to measure the performance of the model on the training set. At the end of each epoch, this is the average loss value of the model on the training set. The term Val Loss is used as a metric to measure the performance of the model on the validation set. At the end of each epoch, this is the average loss value of the model on the validation set. This indicates that the model initially learnt the data well. As can be seen in Table 2, after Epoch 7, Val Loss reaches a minimum level and then starts to increase. This may indicate that the model is overfitting the validation set. The model fits the training set very well, but fits the validation set less well. Test Loss, Test MSE: 0.1719 indicates the performance of the model on the test set. This value is higher than the model's lowest loss on the validation set. This may indicate that the model does not generalise well to new and unseen data. The fluctuation in the validation loss during the model's training process indicates that the model is not learning consistently and is starting to overfit. This situation can usually be improved by early stopping or using more data in the training process. Also, hyperparameter adjustments can be made to improve the performance of the model. For example, parameters such as the size of the LSTM layers, learning rate, number of epochs or batch size can be adjusted. More data can be used to improve the performance of the model.

**Table 2.** Training Process Values

Epoch	Loss	Val Loss
0	0.3101	0.5429
1	0.2207	0.4011
2	0.1512	0.2807
3	0.1017	0.1827
4	0.0729	0.1096
5	0.0655	0.0644
6	0.0761	0.0444
7	0.0912	0.0390
8	0.0979	0.0409
9	0.0947	0.0483
10	0.0858	0.0610
11	0.0761	0.0782
12	0.0685	0.0983
13	0.0641	0.1194
14	0.0627	0.1395
15	0.0634	0.1570
16	0.0653	0.1708
17	0.0673	0.1803
18	0.0690	0.1854
19	0.0699	0.1863

### Conclusion

Understanding the intricate interactions between cosmic phenomena and terrestrial weather patterns has long been a topic of scientific inquiry. One such interaction that has garnered significant attention is the relationship between CRs and cloud cover. CRs are high-energy particles thought to originate primarily from supernovae outside our solar system. When these particles enter Earth's atmosphere, they can ionize air molecules, potentially influencing cloud formation processes. Cloud cover plays a critical role in Earth's climate system by regulating the planet's energy balance through the reflection of solar radiation and the trapping of infrared radiation. As such, understanding the factors influencing cloud cover variability is crucial for accurate climate modeling and prediction. This study aims to analyze the relationship between CRs and TCC globally using state-of-the-art machine learning techniques, specifically LSTM networks. We utilized a comprehensive dataset comprising CR flux measurements, cloud cover observations, and various meteorological parameters collected over several decades. The cosmic ray data was sourced from the Cosmic Ray Station at the University of Oulu's Sodankylä Geophysical Observatory, while TCC data was obtained from the Copernicus Climate Change Service's ECMWF European Reanalysis V5 (ERA5).

Initial analyses involved constructing a correlation matrix to identify relationships between CRs and TCC across different regions, including the Antarctic, Arctic, Europe, Northern Hemisphere, Southern Hemisphere, Atlantic Ocean, and globally. The results indicated generally weak and negative correlations between CR and TCC, with a few exceptions. Notably, weak positive correlations were observed for TCC in the Southern Hemisphere and globally. Negative correlations were more pronounced in the Antarctic and Arctic regions, suggesting region-specific climate mechanisms affecting these connections. The LSTM model was employed to capture temporal dependencies and complex patterns within the dataset. Despite the sophisticated architecture of LSTM networks, the model's predictions of CR values did not closely follow the actual values, indicating a significant gap in capturing dynamic changes and peaks. This suggests that the current model configuration may require more data, additional features, or further tuning to improve accuracy. The training process revealed rapid initial learning, followed by overfitting after several epochs, as indicated by increasing validation loss. The final model's performance, measured by test MSE, suggested inadequate generalization to unseen data. These findings highlight the challenges of modeling the relationship between CRs and TCC using machine learning. The weak correlations and the model's performance indicate the complexity of these interactions and the need for more robust data and advanced modeling techniques. In these studies, it may be necessary to focus on Low-Level Cloud Cover (LCC) data to obtain more accurate results. Future research should focus on enhancing data quality, incorporating more detailed cloud cover metrics, and exploring other machine learning models to better understand the influence of cosmic rays on cloud formation and climate. This study contributes to the ongoing debate about the role of cosmic rays in climate systems, providing insights that could improve climate models and predictions:

**Peer-review:** Externally peer-reviewed.

**Author Contributions:** Concept-AP; Design-AP; Supervision-AP; Resources-AP; Data Collection and/or Processing-AP; Analysis and/or Interpretation-AP; Literature Search-AP; Writing Manuscript-AP; Critical Review-AP; Other-AP

**Conflict of Interest:** The author has no conflicts of interest to declare.

**Financial Disclosure:** The author declared that this study has received no financial support.

**Hakem Değerlendirmesi:** Dış bağımsız.

**Yazar Katkıları:** Fikir-AP; Tasarım-AP; Denetleme-AP; Kaynaklar-AP; Veri Toplanması ve/veya İşlemesi-AP; Analiz ve/veya Yorum-AP; Literatür Taraması-AP; Yazıyı Yazan-AP; Eleştirel İnceleme-AP

**Çıkar Çatışması:** Yazar, çıkar çatışması olmadığını beyan etmiştir.

**Finansal Destek:** Yazar, bu çalışma için finansal destek almadığını beyan etmiştir.

## References

- Carslaw, K.S., Harrison, R.G. & Kirkby, J. (2002). Cosmic rays, clouds, and climate. *Science*, 298(5599), 1732-1737.
- Erlykin, A.D. & Wolfendale, A.W. (2010). Cosmic rays and clouds. *Journal of Atmospheric and Solar-Terrestrial Physics*, 72(5-6), 436-438.
- Goodfellow, I., Bengio, Y. & Courville, A. (2016). *Deep Learning*. MIT Press.
- Graves, A. & Graves, A. (2012). Long short-term memory. *Supervised sequence labelling with recurrent neural networks*, 37-45.
- Hochreiter, S. & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, 9(8), 1735-1780.
- IPCC, (2021). *Climate Change 2021: The Physical Science Basis*. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change.
- Kirkby, J., Curtius, J., Almeida, J. & et al., (2011). Role of sulphuric acid, ammonia and galactic cosmic rays in atmospheric aerosol nucleation. *Nature*, 476(7361), 429-433.
- Laken, B.A., Kniveton, D.R. & Frogley, M.R. (2012). Cosmic rays linked to rapid mid-latitude cloud changes. *Atmospheric Chemistry and Physics*, 12(16), 7859-7867.
- Marsh, N. & Svensmark, H. (2000). Cosmic rays, clouds, and climate. *Space Science Reviews*, 94(1), 215-230.
- Polatoğlu, A., Yeşilyaprak, C., Kaya, M., Shameoni Niaei, M., & Er, H. (2023). New the Design and Measurements of the Portable Cosmic Ray Muons Detector (CRMD) for an Observatory. *Universal Journal of Physics and Application*, 17(4).
- Sherstinsky, A. (2020). Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network. *Physica D: Nonlinear Phenomena*, 404, 132306.

---

Svensmark, H., Bondo, T. & Svensmark, J. (2007). Cosmic ray decreases affect atmospheric aerosols and clouds. *Geophysical Research Letters*, 34(15).