Proceedings of the Canadian Society for Mechanical Engineering International Congress
32nd Annual Conference of the Computational Fluid Dynamics Society of Canada
Canadian Society of Rheology Symposium
CSME-CFDSC-CSR 2025
May 25–28, 2025, Montréal, Québec, Canada

Using Random Forest Feature Importance to Rank Environmental Factors Affecting PV Degradation Rates

Nahid Alladini[1], Rolf Wuthrich[2], Lucas Hof[3],
[1]Département de génie logiciel et des technologies de l'information, ÉTS, Montréal, Canada
[2]Department of Mechanical and Industrial Engineering, Concordia, Montréal, Canada
[3]Département de génie mécanique, ÉTS, Montréal, Canada
lucas.hof@etsmtl.ca

Abstract—This study uses machine learning techniques to explore the degradation of solar photovoltaic (PV) panels under varying environmental conditions. Using data from PV plants in Portugal alongside detailed weather datasets, the research aims to rank the influence of environmental factors on PV degradation. A Random Forest regression model was employed, demonstrating significant improvements in predictive performance through data augmentation and hyperparameter optimization. The analysis highlights wind speed, temperature, humidity, and cloud cover as critical degradation factors, providing actionable information to optimize solar panel installation and maintenance strategies. Despite promising findings, the study acknowledges limitations, including the size of the dataset, geographic scope, and computational constraints of the research process. This work contributes to the field by offering a systematic approach to understanding and mitigating environmental stressors on the performance of photovoltaic panels.

PV panels; Photovoltaic (PV) Degradation; Solar Panel Performance; Environmental Factors; Machine Learning; Random Forest Regression; Feature Importance; Renewable Energy Optimization.

I. Introduction

Solar photovoltaic (PV) panels are essential in meeting global sustainable energy demands [1]. However, their efficiency and longevity are significantly influenced by environmental factors such as changes in temperature, humidity levels, ultraviolet (UV) radiation, and pollution [2]. These conditions contribute to the degradation of PV panels over time, leading to variability in performance. There have been many studies on the effect of some environmental factors on the degradation of solar panels. Still, they all focus on prediction without ranking environmental factors according to their impact. We cannot control these environmental factors or always have the best environment for solar panels; therefore, understanding and systematically classifying these factors by their impact can significantly help us improve panel designs, select optimal installation sites, and implement maintenance strategies tailored to specific environments. To address these challenges, this study adopts a data-driven approach to better understand the factors contributing to solar panel degradation.

Machine learning (ML) has emerged as a powerful tool in assessing and predicting PV degradation. These models offer significant computational capabilities, enabling the analysis of large datasets with diverse features in minimal timeframes. ML is particularly suitable for evaluating the performance and degradation of PV modules across manufacturing, field deployment, and laboratory settings [3]. Hence, various ML models, including Decision Tree (DT), Support Vector Machine (SVM), Random Forest (RF), KNN, Neural Networks, and Naive Bayes (NB), have been employed in previous research for defect detection, classification, and PV estimate degradation rates analysis [3].

Supervised learning methods such as linear regression, decision trees, gradient boosting models (LightGBM, XGBoost), artificial neural networks (ANNs), and support vector machines (SVMs) have been widely applied to analyze PV performance. For instance, Suanpang and Jamjuntr [4] employed models such as the Light Gradient Boosting Machine (LGBM) and K-Nearest Neighbors (KNN) to forecast solar energy performance under urban stressors, while Pronichev and Shishkov [2] applied the Statistical Clear Sky Fitting Algorithm and Prophet algorithm to evaluate degradation. Similarly, Dhingra, Tyagi, and Tomar [5] identified the Extra Tree Regressor as particularly effective for accurately forecasting PV degradation. These studies highlight the predictive potential of machine learning but do not prioritize environmental factors systematically.

High-stress environmental conditions have also been extensively studied. Hassane et al. [6] and Ihaddadene, Tabet, Guerira, Ihaddadene N., and Bekhouche [7] identified dust, humidity, and extreme temperatures as critical to degradation, while Sameera, Tariq, and Rihan [8] emphasized cloud cover, dust, and haze. Despite these findings, comparative analyses of environmental factors to identify the most significant drivers of degradation are still scarce. Existing research extensively investigates how environmental factors influence PV degradation but lacks a systematic classification of these factors by their impact.

In summary, while machine learning shows great promise in predicting PV degradation, there remains a significant research gap in ranking environmental factors. Addressing this gap could support more precise location-specific optimizations and strategic planning for PV installations.

The principal objective of this study is to leverage machine learning techniques, specifically the RF regression model, and environmental datasets to classify and rank ecological factors based on their influence on PV degradation. The findings can provide actionable insights for enhancing the efficiency and lifespan of photovoltaic panels in various climatic conditions.

II. Methodology

This study aims to evaluate the performance and degradation of PV panels from a dataset supplied by the non-profit organization Coopérnico [9]. The dataset includes the performance of nine PV panels in six Portuguese cities over four years (2019–2022), which correlates with weather data from the corresponding weather stations. It involves several stages: data collection, preparation, model selection, and performance evaluation.

A schematic chart representing the workflow is presented in Fig. 1 to provide an overview of the methodological approach. This chart illustrates the sequential stages of data collection, preprocessing, augmentation, model training, feature analysis, and evaluation, ensuring a structured and reproducible framework. The first stage will combine the three primary datasets: PV production data, metadata (contextual data about the PV plants), and weather data for the exact locations and periods. All these datasets must be combined into one structured format, aligning timestamps, serial numbers, and location data. The data undergoes cleaning based on the timestamp transformation, duplication removal, and disregarding of NaN or inconsistent data entries. Specifically, rows with 0 energy produced will be eliminated to prevent working with misleading outcomes in subsequent analysis.

Following data preparation, a modeling approach will be implemented to analyze PV performance and degradation. A Random Forest (RF) regression model is selected due to its ability to handle complex, nonlinear relationships and provide feature importance scores, thereby identifying the key environmental factors contributing to PV degradation [10], [11]. To systematically rank environmental factors affecting PV degradation, we utilized the feature importance scores provided by the RF model. This method was chosen due to its ability to handle non-linear relationships and measure the contribution of each variable to prediction accuracy [10]. Additionally, a simple linear regression model will be used to estimate the degradation rate of each panel.

As Pronichev and Shishkov [2] state, “Degradation is the loss of performance over time, caused by various factors such as environmental conditions, climate, and technical equipment specifications.”. To evaluate the performance and degradation of PV panels, a simple approach is to calculate the efficiency of each panel at a given time t as the ratio of the amount of energy generated per unit of installed capacity (Specific Energy) to the total capacity of the solar energy system (Installed Power), as shown in Equation (1):

$$\eta_t = \frac{Specific\ Energy\ \left(\frac{kWh}{kWp}\right)}{Installed\ Power\ (kWp)} \tag{1}$$

Here η_t represents the panel’s efficiency at a given time step t, which is a fundamental measure of solar panel efficiency. It normalizes the energy generated (“kWh”) by the installed capacity (“kWp”), providing a practical and standardized metric for evaluating performance under varying environmental conditions, as supported by Hassane et al. [6] and Breiman [10], using normalized efficiency metrics allows for reliable comparisons across systems and better insights into the impact of environmental factors. This approach is particularly useful in identifying degradation trends caused by dust, humidity, and extreme temperatures [6].

A linear regression model will be applied to the efficiency values over time to estimate the degradation rate. The degradation rate β is defined as the slope of the best-fit regression line

$$\eta_t = \beta \cdot t + c \tag{2}$$

where t is the numeric representation of time, and c is the intercept. This slope β quantifies the rate at which efficiency declines over time, representing the degradation rate of each panel. Equation (2) models the relationship between efficiency

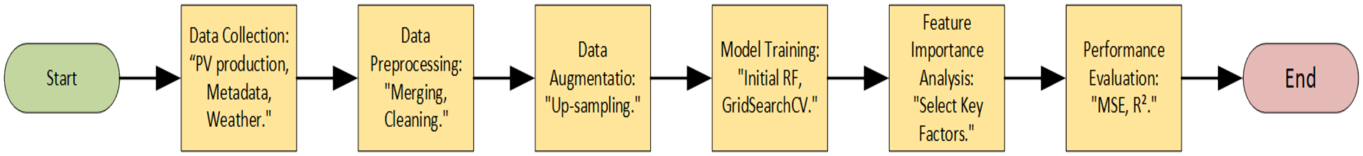
Start
Data Collection: “PV production, Metadata, Weather."
Data Preprocessing: "Merging, Cleaning."
Data Augmentatio: "Up-sampling."
Model Training: "Initial RF, GridSearchCV."
Feature Importance Analysis: "Select Key Factors."
Performance Evaluation: "MSE, R^2."
End

Figure 1. Schematic Chart of Methodology

(η_t) and time (t). This linear model approximates degradation rates over time, consistent with standard practices in PV degradation research. These equations underpin the study's methodology for quantifying and ranking environmental factors affecting PV panel performance.

The computed degradation rates for each panel will be mapped back to the primary dataset for further analysis and comparison.

To establish a baseline, the initial RF model will be trained without hyperparameter tuning. Model performance will be evaluated using Mean Squared Error (MSE) and the coefficient of determination (R^2). If the model exhibits low predictive accuracy, Grid Search with Cross-Validation (GridSearchCV) will be employed for hyperparameter tuning. Data augmentation techniques will also be applied to compensate for the dataset's limited size, improving model robustness. Environmental factors such as wind speed, temperature, humidity, and cloud cover will be analyzed to determine their relative impact on PV panel degradation.

The last phase of the framework implements an optimized RF model, features importance analysis, and predictive versatility of the model. This method will help identify the significant environmental drivers behind PV panel decay and provide guidance in potentially optimizing solar panel placement practices.

III. Implementation and Results

The methodology was carried out systematically, starting with data integration and preparation. This successful merging resulted in synchronized PV production records, metadata, and weather variables. Timestamps were restructured to split the date and time during preprocessing, and duplicate entries, as well as rows where the produced energy was zero, were dropped to keep the consistency of data. This cleaned dataset is saved for later analyses.

Without hyperparameter tuning, the first RF regression model used 100 decision trees. The model performance was evaluated with MSE and R^2, which gave an MSE of 2.36×10^{-12} and an R^2 of 0.016. The low R^2 showed that the model explained very little of the variance in the target variable.

To improve model performance, GridSearchCV was employed to tune the hyperparameters, considering the computational resources available. The training and optimization of the RF model were performed on a 64-bit Windows operating system with an Intel(R) Core(TM) i7-10510U CPU running at 1.80 GHz (2.30 GHz max) and 16 GB of RAM (15.8 GB usable).

The optimized RF model was trained with 200 decision trees, each having a maximum depth of 20 and a minimum split of 2. The model achieved an MSE of 2.32×10^{-12} and an R^2 score of 0.035. Although R^2 improved slightly compared to the initial model, it remained low, indicating that the model could explain only 3.52 percent of the variance.

One significant challenge in model training was the limited variability in the dataset, aside from seasonal trends. Fig. 2 illustrates the seasonal variations in solar panel performance over the four-year study period, showing noticeable produced energy drops during winter. This figure, created using the cleaned dataset, plots the produced energy for all PV panels. These drops align with known seasonal patterns, such as reduced solar irradiance, lower temperatures, and increased cloud cover during winter, all contributing to diminished panel efficiency.

This visualization underscores the dataset's inherent periodicity and highlights the importance of accounting for seasonal trends in predictive modeling. The insights from this analysis provided a critical baseline for understanding the influence of external factors on panel degradation. They guided the selection of key variables for the Random Forest model.

A heatmap of the correlations between environmental variables was generated to investigate further the underlying factors affecting PV panel degradation, as shown in Fig. 3. This visualization is based on Pearson correlation coefficients, which measure the linear relationship between two variables. The correlation values range from -1 to 1, where:

- 1 indicates a perfect positive correlation,
- -1 indicates a perfect negative correlation, and
- 0 indicates no linear correlation.

The correlation matrix was calculated using the pandas.DataFrame.corr() function, which computes the Pearson correlation coefficients for the selected environmental variables in the augmented dataset. These variables include shortwave radiation, wind speed, apparent temperature, relative humidity, air temperature, dew point, and cloud cover. The heatmap was then plotted using the Seaborn library, with annotations to display the exact correlation values for clarity.

The heatmap reveals strong correlations between certain environmental variables:

- Short-wave radiation is strongly correlated with temperature.
- Apparent temperature exhibits a strong correlation with temperature.
- Dew point also demonstrates a strong correlation with temperature.

Given these relationships, wind speed, relative humidity, temperature, and cloud cover were selected as the primary environmental factors in the analysis. This decision minimizes redundancy while ensuring the selected variables accurately capture key influences on solar panel degradation.

Since dataset size was identified as a limiting factor, data augmentation was used to expand the dataset. The original dataset was up-sampled sixfold, increasing the number of rows from 157,755 to 946,554. The RF model was then retrained using this augmented dataset. Optimized with 300 decision trees and an unrestricted maximum depth, the final model achieved an MSE of 4.98×10^{-14} and an R^2 of 0.981. This significant improvement demonstrates the impact of hyperparameter tuning and dataset size on predictive accuracy.

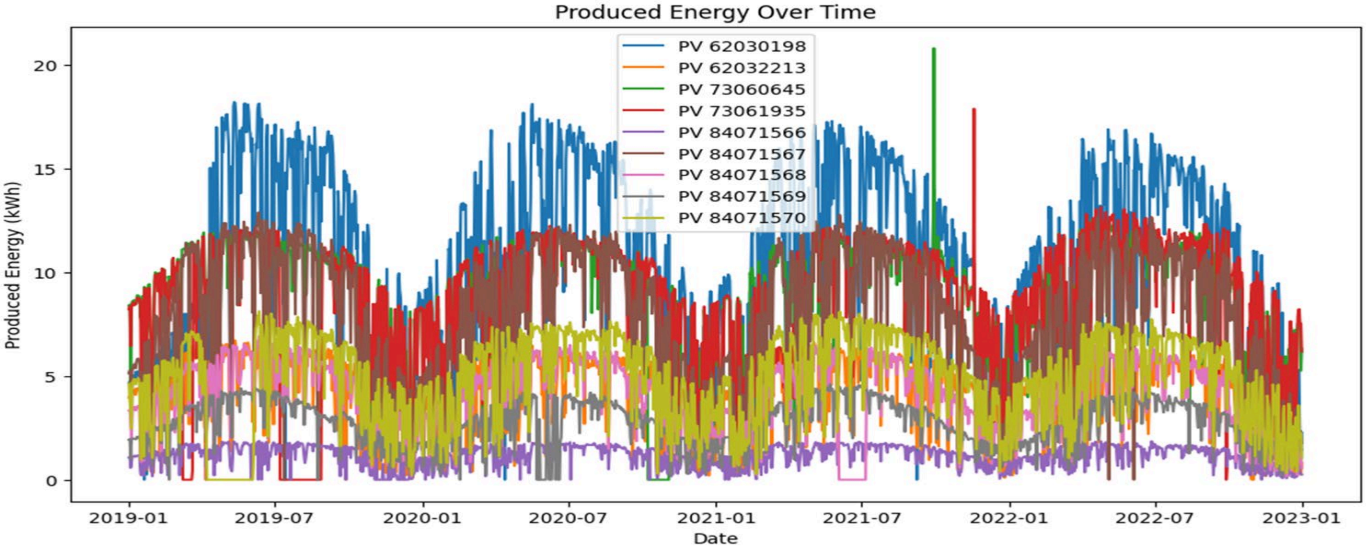

Figure 2. Produced energy by solar panels over time

Finally, the feature importance analysis was plotted, as shown in Fig. 4. It highlights wind speed as the most influential factor affecting PV panel degradation, followed by temperature, relative humidity, and cloud cover. These findings suggest that mechanical stress due to wind speed and temperature fluctuations plays a crucial role in efficiency degradation.

The validated RF model accurately predicted solar panel degradation trends, demonstrating its potential as a decision-making tool for optimizing maintenance schedules and deployment strategies under varying climate conditions. The findings emphasize the importance of expanding dataset size and refining feature selection to enhance model performance. Additionally, this study reinforces the significance of environmental risk assessment in predicting PV panel degradation, contributing to a data-driven approach for solar energy management.

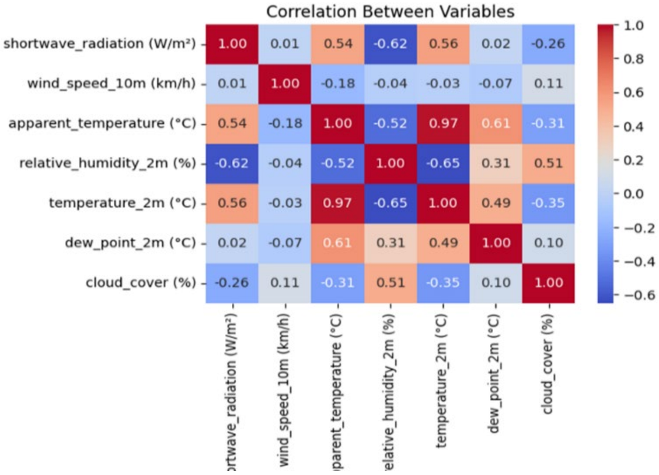

Figure 3. Heat map showing correlations among environmental factors affecting solar panel degradation.

IV. Discussion

This study provides valuable information about the degradation of PV panels under different environmental conditions. However, several caveats must be recognized, especially regarding data limitations, model assumptions, and computational restrictions.

A. Limitations

The data collected in this study were small and geographically limited to six cities in Portugal. Consequently, it did not offer climatic flexibility, potentially affecting the transferability of the results to areas with markedly different climate scenarios. No information on panel wear, maintenance history, or installation quality seems to be part of the dataset, although each is crucial for degradation. Without this contextual data, such an analysis will only tilt at environmental and extraneous factors, potentially missing other pivotal contributors to PV efficiency loss.

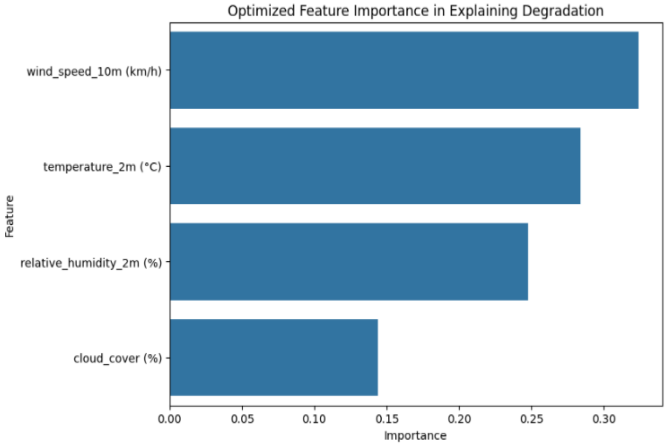

Figure 4. Feature importance ranking of environmental factors influencing PV panel degradation.

The degradation rate was estimated using a simple linear regression model, as we did not have the time to conduct a full analysis. Although this method provides a strong baseline, it does not account for nonlinear degradation effects, seasonal effects, or external factors such as dust build-up or material aging. More advanced degradation models capturing long-term efficiency trends and dynamic environmental interactions could be integrated into future studies.

To address the issue of a limited dataset size, data augmentation was done by repeating existing observations. While this increased the dataset size and helped stabilize the model, it may have raised biases by replicating old trends rather than supplementing them with new independent observations. This could constrain the model's generalization of unseen, real-world data. A more robust approach for addressing the questions posed in future studies would be to collect a more extensive and more diverse dataset across multiple geographic regions and varying climate conditions.

Another standard limitation was computational resources. The work was carried out on a personal laptop with an Intel(R) Core(TM) i7-10510U CPU and 16 GB of RAM, which limited hyperparameter tuning, processing larger datasets, and experimenting with more advanced machine learning models. To overcome these limitations and improve model performance in the future, high-performance computing (HPC) clusters or cloud-based machine learning frameworks could be employed.

B. *Key Findings and Contributions*

These constraints notwithstanding, the study effectively demonstrated the viability of applying machine learning techniques to analyze trends in PV degradation. Environmental factors included wind speed, temperature, relative humidity, and cloud cover—according to the analysis, the most relevant factors influencing the degradation of solar panels. A feature importance analysis validated wind speed and temperature as the main contributors to the performance drop, supporting the hypothesis regarding mechanical stress and the thermal impact on PV efficiency loss.

However, using data augmentation and hyperparameter tuning, a Random Forest regression model achieved an optimized R² score of 0.981, meaning that it could explain 98.13% of the variance for solar panel performance. This demonstrates the importance of both the dataset's size and the model parameters' fitting since this combination achieved a significant improvement in accuracy on unseen data.

V. Conclusion and Outlook

The present study adds to the existing knowledge on solar panel degradation by providing the first data-driven approach to ascertain and rank environmental stressors affecting degradation. This information constitutes the basis of a plan for the optimal use of solar panels, their maintenance, and their long-term performance under different climatic conditions.

Future research can iterate towards more accurate predictive models and contribute to the widespread adoption of AI-based decision-making processes in renewable power systems by tackling dataset constraints, applying more complex modeling approaches, and utilizing even greater computing power.

Future work can overcome the mentioned limitations to enhance the reliability and generalizability of this research by:

1. Widening dataset diversity—merging the data pools from different geographical areas to attain higher generalizability.
2. Inputting panel condition data such as maintenance history, panel material characteristics, and operational performance logs.
3. Investigating more sophisticated degeneration models—for example, using nonlinear regression, time-sequence prediction models, or deep learning methods.
4. Using high-performance computing (HPC) resources so more sophisticated models can be trained on larger datasets.

Acknowledgments

I sincerely thank Professor Ali Ayub for his guidance, particularly his advice on data augmentation, which helped improve the project's outcomes. I also extend my gratitude to Mitacs Accelerate and CAPSolar Inc. for their financial support, which made this research possible.

References

[1] A. K. Tripathi *et al.*, “Advancing solar PV panel power prediction: A comparative machine learning approach in fluctuating environmental conditions,” *Case Studies in Thermal Engineering*, vol. 59, p. 104459, Jul. 2024, doi: 10.1016/j.csite.2024.104459.

[2] A. V. Pronichev and E. M. Shishkov, “Assessing and Predicting Degradation of Solar Panels Using Machine Learning Approach,” in *2023 6th International Scientific and Technical Conference on Relay Protection and Automation (RPA)*, Oct. 2023, pp. 1–16. doi: 10.1109/RPA59835.2023.10319847.

[3] Z. Ullah Khan *et al.*, “A Review of Degradation and Reliability Analysis of a Solar PV Module,” *IEEE Access*, vol. 12, pp. 185036–185056, 2024, doi: 10.1109/ACCESS.2024.3432394.

[4] P. Suanpang and P. Jamjuntr, “Machine Learning Models for Solar Power Generation Forecasting in Microgrid Application Implications for Smart Cities,” *Sustainability*, vol. 16, no. 14, Art. no. 14, Jan. 2024, doi: 10.3390/su16146087.

[5] B. Dhingra, S. Tyagi, and A. Tomar, “A Comparative Study of Machine Learning Algorithms for Photovoltaic Degradation Rate Prediction,” in *2022 Third International Conference on Intelligent Computing Instrumentation and Control Technologies (ICICICT)*, Aug. 2022, pp. 474–478. doi: 10.1109/ICICICT54557.2022.9917960.

[6] Abdelhamid Issa Hassane *et al.*, “Data Driven Analysis of the Impact of Weather Parameters on Solar Photovoltaic Panels Efficiency in a Sahel Region: Future Prospects,” *Appl. Sol. Energy*, vol. 60, no. 2, pp. 313–327, Apr. 2024, doi: 10.3103/S0003701X23601898.

[7] R. Ihaddadene, S. Tabet, B. Guerira, N. Ihaddadene, and Kh. Bekhouche, “Evaluation of the degradation of a PV panel in an arid zone; case study Biskra (Algeria),” *Solar Energy*, vol. 263, p. 111809, Oct. 2023, doi: 10.1016/j.solener.2023.111809.

[8] Sameera, M. Tariq, and M. Rihan, “Investigating the Impact of Atmospheric Factors on Solar PV Panel Performance,” in *2023 International Conference on Power, Instrumentation, Energy and Control (PIECON)*, Feb. 2023, pp. 1–6. doi: 10.1109/PIECON56912.2023.10085772.

[9] Elissaios Sarmas, Maria Kleideri, Nuno Matias, Catarina Pereira, and Ana Rita Antunes, “Photovoltaic Power Production Dataset.” Jun. 27, 2024. doi: 10.17632/dbh93b6vp8.2.

[10] L. Breiman, “Random Forests,” *Machine Learning*, vol. 45, no. 1, pp. 5–32, Oct. 2001, doi: 10.1023/A:1010933404324.

[11] A. Akselrud and C. I, “Random forest regression models in ecology: Accounting for messy biological data and producing predictions with uncertainty”, Accessed: Jan. 09, 2025. [Online]. Available: https://repository.library.noaa.gov/view/noaa/66323