Predicting London's Precipitation: A Spatio-Temporal Neural Network Approach

Huma Zafar[1], Stelios Kapetanakis[1], Giacomo Nalli[2], Khuong An Nguyen[3]

[1] Distributed Labs, 30 Churchill Pl, London E14 5RE, UK
[2] Middlesex University, The Burroughs, London NW4 4BT, UK
[3] Royal Holloway University of London, Egham Hill, Egham TW20 0EX, UK
[1] {huma, stelios}@dstr.co.uk, [2] giacomo.nalli@mdx.ac.uk,
[3]Khuong.Nguyen@rhul.ac.uk

Abstract. This study presents a data-driven approach to forecasting total precipitation in London using an Artificial Neural Network (ANN) within a spatio-temporal framework. Leveraging ERA5 data from 2010 to 2025, the methodology includes automated NetCDF extraction, feature engineering with lagged precipitation and cyclic time encodings, and dimensionality reduction via a trained Autoencoder. The ANN, designed in a GenCast-style architecture, was trained using the Adam optimiser over 50 epochs and achieved strong performance. SHAP analysis highlighted the importance of lag features and seasonal time variables, enhancing interpretability and supporting the model's application in urban flood risk management and climate resilience.

Keywords: Total precipitation forecasting, Artificial Neural Networks (ANN), Spatio-temporal modelling

1 Introduction

Total precipitation significantly impacts urban infrastructure, water management, and socio-economic stability. Extreme or irregular rainfall can lead to flash floods, overwhelmed drainage systems, and disruptions in transport and agriculture (Yao et al., 2020).

Accurate forecasting of total precipitation is essential for urban mobility and strategic planning across agriculture and industry (Aguasca-Colomo et al., 2019). Traditional statistical models, which rely on linear associations with weather variables, often fall short due to the complexity of rainfall dynamics (AbuSaleh et al., 2024).

Nevertheless, precipitation forecasting remains challenging due to its spatial and temporal variability, especially in urban areas like London (dao Santos, 2020). ANN effectiveness still depends heavily on data quality, model design, and the specific forecasting objective (AbuSaleh et al., 2024). Luk et al. 2000 paved the way for advancements in short-term rainfall forecasting within urban environments by proposing the use of ANNs. Their work was further complemented by Valverde Ramírez et al., 2005 proposed an ANN-based method for Sao Paulo, Brazil. Their approach involved establishing a non-linear relationship between regional ETA model output data and surface rainfall data. Moreover, Bilgili and Sahin (2010) successfully employed ANN architecture to model monthly rainfall and temperature in the context of Turkey. Suparta and Samah (2020) built three different ANN models for rainfall forecasting, evaluating their respective performance and potential for further development. This suggests ANFIS could be a valuable tool for flood risk management.

This work focuses on London, a densely populated urban place, which faces significant challenges with total precipitation. This is due to a unique combination of its population density, aging drainage infrastructure, and increasing climate volatility. Erratic or extreme rainfall events have far-reaching consequences, leading to flash floods, overwhelmed drainage systems, and disruptions to both transportation and agriculture (Hanlon et al., 2021). Precipitation patterns are also linked to increased air pollution episodes, adding another layer of environmental concern (Czarnecka et al., 2011). Ultimately, accurate forecasting of total precipitation is vital for London's urban planning, climate resilience, and early warning systems.

2 Background

2.1 Dataset Description

The data for this study is sourced from the Copernicus Climate Data Store (CDS), a leading provider of climate data. The data is in NetCDF format and has been preprocessed using xarray, an open-source library that helps managing numerical arrays, allowing for more concise and less error-prone user experience whilst efficiently handling multi-dimensional arrays. The temporal coverage spans from 2010 to 2025, focusing on total precipitation (tp) recorded at 00:00 UTC. To ensure relevance to the study area, the spatial resolution is clipped specifically to the London boroughs, using a coordinate system defined by Latitude, Longitude, sand Time.

2.2 Precipitation Patterns in London (2010-2025)

Analysis of total precipitation data over London for selected years (2010, 2015, 2020, 2025) reveals distinct spatio-temporal patterns. The provided precipitation maps, with London boroughs outlined in red, illustrate the distribution of precipitation in millimeters (mm) at 00:00 UTC for these years.

2010 (Figure 1) In 2010, the total precipitation across London at 00:00 UTC varied, with some areas experiencing higher levels (darker shades, up to 22 mm) and others relatively lower (lighter shades, around 18 mm). The western and south-eastern parts of London appear to have received more precipitation compared to the central and eastern regions.

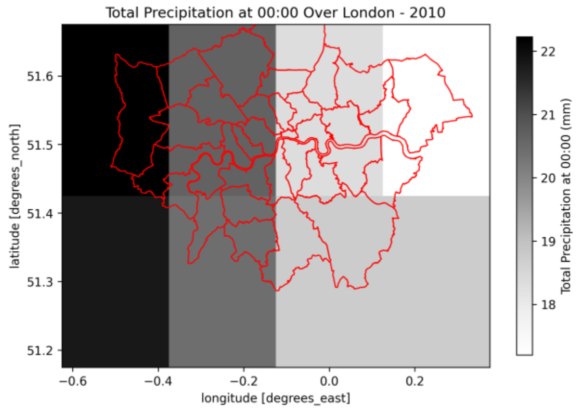


Fig 1: 2010, Total Precipitation across London

2015 (Figure 2) By 2015, there was a noticeable increase in total precipitation values compared to 2010, with some regions reaching up to 30 mm. The overall precipitation levels seem higher across the city, indicating a potentially wetter period or a specific event captured at this time. The spatial variability remains, but the intensity has increased.

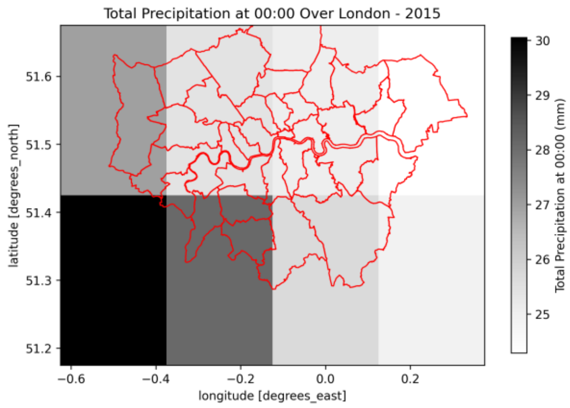


Fig 2: 2015 - Total Precipitation across London

2020 (Figure 3) The precipitation map for 2020 shows a shift in the distribution and intensity. While some areas still exhibit higher precipitation (up to 25 mm), there are also significant areas with lower values. Notably, the southern central area appears to have experienced particularly low precipitation at this specific time, highlighted by the darkest shades indicating values around 22 mm. This suggests localized variations in rainfall.

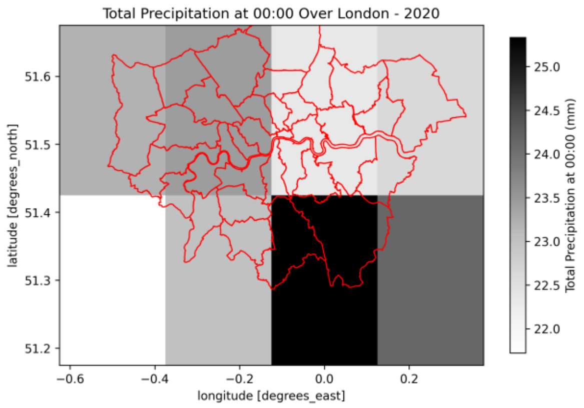


Fig 3: 2020 - Total Precipitation across London

2025 (Figure 4) The forecast for 2025 indicates a significant decrease in total precipitation compared to previous years, with values ranging from approximately 9 mm to 10.2 mm. This suggests a much drier outlook at 00:00 UTC for London. The shift towards lower precipitation levels across the entire city is quite pronounced, potentially implying a drier climate trend or a specific dry period.

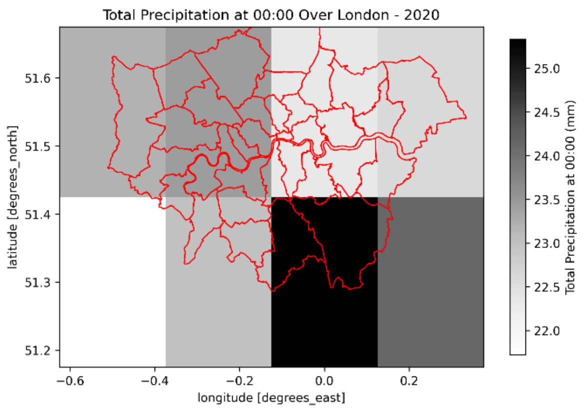


Fig 4: 2025 - Total Precipitation across London

2.3 Monthly Mean Precipitation Analysis

A deeper insight into London's precipitation patterns can be gained by examining the monthly mean precipitation for the selected years (2010, 2015, 2020, 2025), as illustrated in the histogram (Figure 5).

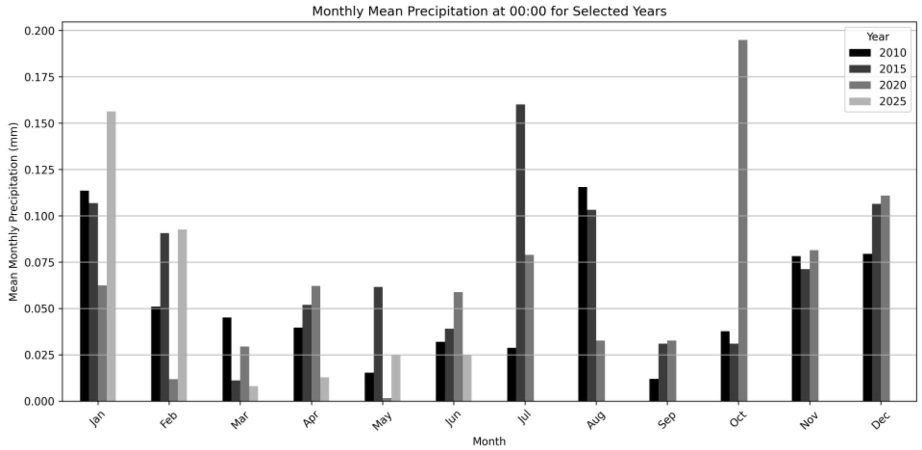


Fig 5: Monthly Mean Precipitation at 00:00 for 2010 -2025

The histogram, "Monthly Mean Precipitation at 00:00 for Selected Years," provides a comparative view of average monthly rainfall across these four years.

First, there is considerable variability in monthly precipitation across the years. For example, January recorded relatively high rainfall in 2010 and 2025, while showing much lower values in 2015 and 2020. February displays similar fluctuations, with a notable peak in 2015. Second, seasonal peaks do not occur consistently in the same months each year, reflecting the non-stationary nature of precipitation patterns. July experienced a significant peak in 2010, whereas October showed an unusually high spike in 2020, indicating a particularly wet autumn. Third, some months appear consistently drier across all years; September, for instance, generally receives lower average precipitation, with May and June also tending to show reduced rainfall. Hence, the data reveals a high degree of year-to-year variability, with extreme differences such as the wet October of 2020 compared to much drier conditions in the same month during 2015 and 2025. Likewise, while July 2010 was markedly wet, July in both 2020 and 2025 was significantly drier. These patterns underscore the complexity of long-term precipitation forecasting and the importance of models capable of capturing such extremes.

Based on the analysis of precipitation maps and monthly mean histograms highlights the complex and dynamic nature of rainfall in London. High spatial and temporal variability, along with shifting climate patterns, makes accurate total precipitation forecasting both essential and challenging. Therefore, there is a clear need for advanced AI models that can learn from and adapt to these details, thereby supporting effective urban planning, climate resilience, and early warning systems.

Dataset Background

The dataset used in this study was sourced from the Copernicus Climate Data Store (CDS), a leading repository for climate information. Provided in NetCDF format, the data was preprocessed using the xarray library to efficiently handle its multi-dimensional structure. Covering the period from 2010 to 2025, it focuses on total precipitation (denoted as ‘tp’) recorded at 00:00 UTC. The spatial resolution was clipped to the boundaries of the London boroughs, using a coordinate system defined by latitude, longitude, and time to ensure geographic relevance.

The primary variable, 'tp', was structured along the dimensions of ‘valid_time’, ‘latitude’, and ‘longitude’, reflecting its spatio-temporal nature. Metadata revealed that the dataset originated from the European Centre for Medium-Range Weather Forecasts (ECMWF), and conformed to CF-1.7 climate data conventions. The history attribute indicated prior conversion from GRIB to NetCDF. In total, the dataset comprises 45,272 precipitation records, forming a robust foundation for analysis.

3 Methodology

3.1 Data Loading and Exploration

The NetCDF file was accessed programmatically from a compressed archive. Using xarray, the dataset was loaded and examined to understand its variables, coordinates, and attributes. Key dimensions included valid_time for temporal indexing, and latitude and longitude for spatial mapping. Additional metadata coordinates such as number and expver provided contextual details about data versions and origins.

3.2 Feature Engineering

To prepare the dataset for machine learning models, several features were engineered to capture complex temporal dynamics:

- **Lag Features:** Time-lagged precipitation values were generated (e.g., tp_lag_1, tp_lag_3, tp_lag_6, tp_lag_12) to allow the model to learn from past rainfall events.
- **Cyclical Time Encoding:** Sine and cosine transformations were applied to cyclic time variables such as month, day, and day of year (e.g., month_sin, dayofweek_cos) to retain their periodic nature while avoiding linear assumptions.
- **Autoencoder-Derived Features:** An Autoencoder trained on scaled precipitation data produced low-dimensional representations (3D and 7D), capturing latent, non-linear patterns. These were visualised using scatter plots and pairwise comparisons, offering deeper insight into precipitation variability.

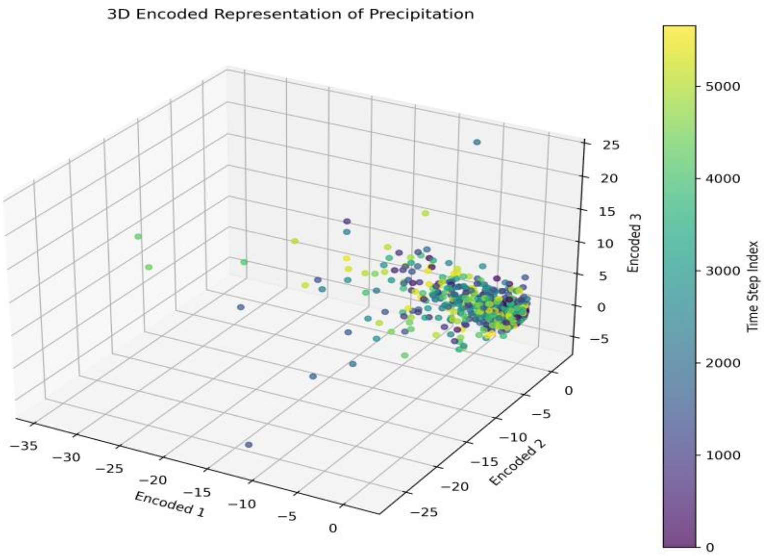


Fig 6: Encoded Representation of Precipitation

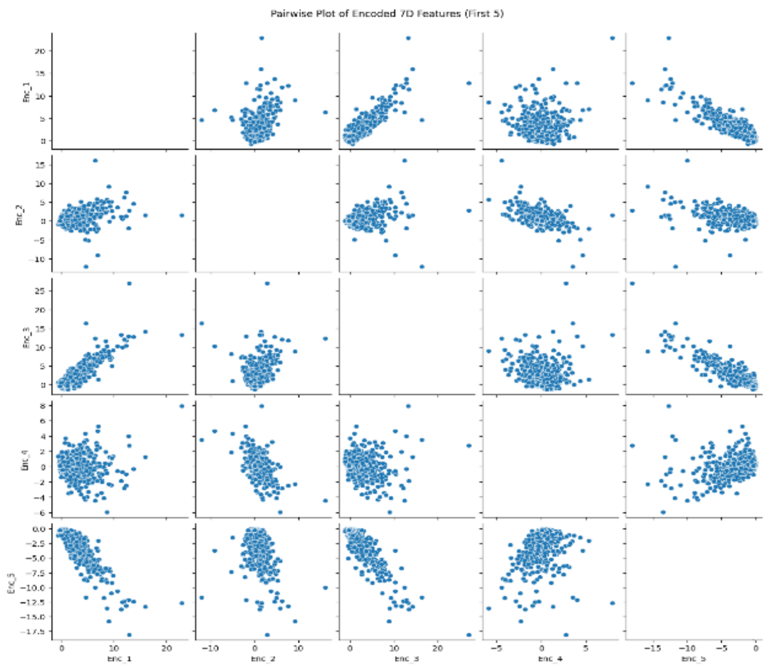


Fig 7: Pairwise plot of encoded features

All target variables (‘tp’) were normalised using StandardScaler, ensuring consistent scale and stable convergence during model training.

3.3 Data Splitting and Preprocessing

The processed dataset was split into training (60%), validation (20%), and test (20%) sets using a two-step approach via train_test_split from scikit-learn. First, 60% of the data was allocated for training, with the remaining 40% split equally for validation and testing. This stratified partitioning ensured representativeness and prevented data leakage.

Following the split, input features and target variables in each subset were further scaled using StandardScaler to maintain consistent data distributions. This step is critical for neural network models, which perform optimally when inputs are normalised across features.

3.4 Model Training and Performance Evaluation

An Artificial Neural Network (ANN), based on a structure suited for spatio-temporal data (often referred to as a *GenCast-style* model), was developed to predict total precipitation (‘tp’). This architecture was designed to leverage the engineered time series features, capturing complex temporal patterns.

The network consisted of:

- **Input layer**: Receiving all preprocessed features, with the number of neurons equal to the number of input variables.
- **Two hidden layers**: Fully connected layers using Rectified Linear Unit (ReLU) activations, which introduce non-linearity and allow the model to learn intricate precipitation dynamics that traditional linear models may miss.
- **Output layer**: A single neuron with a linear activation function, appropriate for continuous-value regression tasks.

The model was trained using the Adam optimiser, aiming to minimise the Mean Squared Error (MSE) loss function. Training occurred over 50 epochs with a batch size of 32, and incorporated a validation split to monitor generalisation performance across epochs. This validation process is crucial: while the model learns patterns in the training data, it must also avoid overfitting—performing well only on familiar data.

Post-training, model predictions and corresponding test labels were inverse-transformed to their original scales using the StandardScaler applied earlier. This enabled a direct and interpretable comparison between predicted and observed precipitation values.

Model performance was measured using three regression metrics: the R^2 score, which shows how much of the variation in the data the model can explain, with higher values meaning better predictions; the mean squared error (MSE), which calculates the average squared difference between predicted and actual values and, although it is sensitive to outliers, gives a good indication of overall accuracy; and the mean absolute error (MAE), which finds the average of the absolute differences, making it easier to understand and less affected by large errors.

4 Results

The ANN achieved an R^2 score of 0.7698, meaning it explained approximately 77% of the variance in total precipitation. While this does not represent perfect prediction, it demonstrates strong learning of underlying patterns in the data. On the other hand, both MSE and MAE were remarkably low, indicating minimal average errors between predictions and actual values. Though small errors may accumulate over longer forecasting horizons, the model's accuracy on test data remains robust.

To conclude, these results confirm that the ANN model is highly effective for short- to medium-term precipitation forecasting in complex urban environments.

4.1 SHAP Analysis of the ANN Model

To improve the interpretability of the Artificial Neural Network (ANN) and understand how features affect its predictions, SHAP (SHapley Additive exPlanations) analysis was performed on test data. SHAP values show each feature's contribution to predictions, indicating their importance and whether their impact is positive or negative.

The SHAP Summary Plot (Figure: SHAP Summary Plot – Feature Importance) displays the distribution of SHAP values for all features. The x-axis shows the effect on predictions, and colours indicate feature values (red = high, blue = low). The spread of points reflects each feature's overall importance.

Top 5 Influential Features (ranked by mean absolute SHAP values):

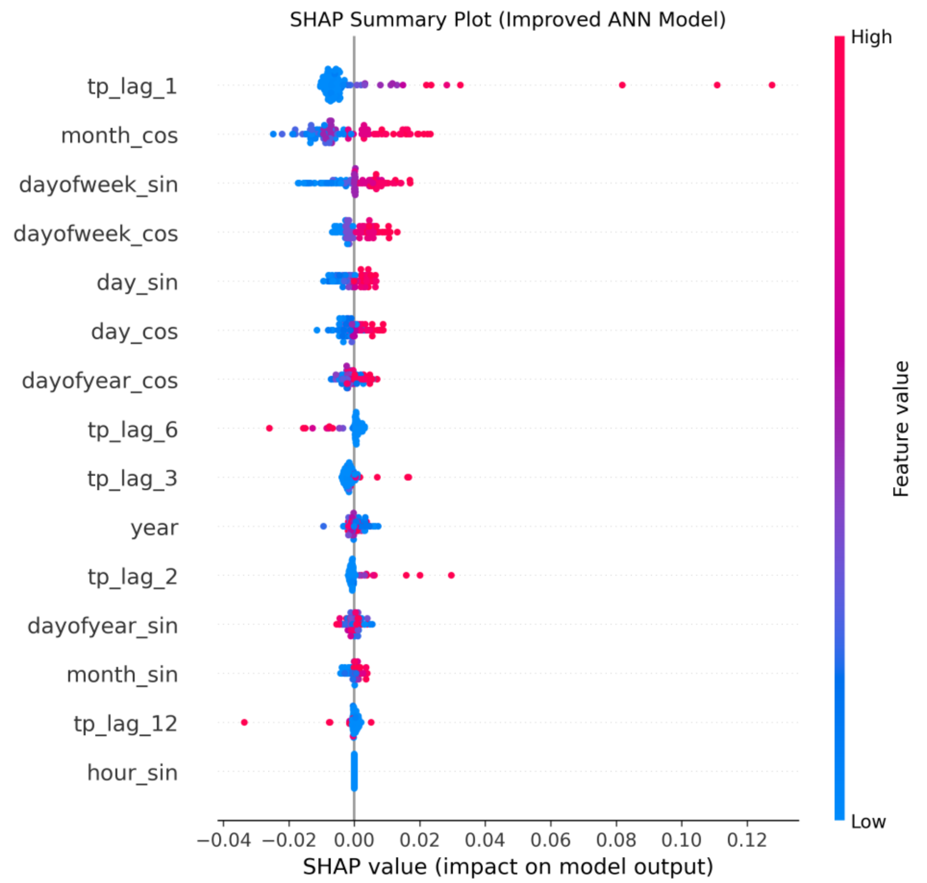
1. tp_lag_1 (0.0104): Previous precipitation was the strongest predictor, showing a positive link with current precipitation, especially seasonally.

2. month_cos (0.0100): Captures seasonal cycles, showing significant variation in precipitation by time of year.

3. dayofweek_sin (0.0064): Suggests a mild weekly precipitation pattern, possibly linked to weather or human activity.

4. dayofweek_cos (0.0041): Complements the weekly pattern with variations across years.

5. day_sin (0.0037): Reflects finer monthly or yearly cycles influencing precipitation.



Although the ANN is inherently a black-box model, SHAP offers a robust framework for revealing its internal decision processes. In contrast, the analysis confirms that the model relies on features with strong meteorological grounding—particularly lagged precipitation and cyclic time components. Even so, certain features exert weaker influence, yet still contribute to nuanced predictions.

Conclusively, the SHAP analysis substantiates that the ANN model aligns closely with known physical drivers of precipitation. This added interpretability increases confidence in the model's outputs and confirms its practical value for urban forecasting, infrastructure planning, and climate risk mitigation. Specifically, the analysis shows that lagged precipitation values and cyclic temporal features (month, day of week, day) are the main factors influencing the model's predictions for total precipitation in London. Such interpretability is vital for validating the model's logic and reinforcing trust in its forecasts, as it corresponds with established meteorological knowledge.

5 Conclusion

Building on the success of this ANN-based forecasting approach, several promising avenues for future research emerge. Future research should first focus on improving computational efficiency to develop more practical and affordable models, as current advanced methods like GenCast demand high resources unsuitable for real-time use. Second, incorporating additional meteorological variables such as temperature, humidity, wind speed, and pressure could enhance forecast accuracy. Third, using higher spatial and temporal resolution data would improve localisation, crucial for urban rainfall prediction. Fourth, ensemble modelling that combines various ANN architectures and traditional methods can increase reliability by leveraging diverse strengths. Fifth, explicitly quantifying forecast uncertainty would help decision-makers better assess confidence and manage risks. Further areas include real-time deployment challenges and applying models to future climate scenarios to assess climate change impacts on London's precipitation.

References

AbuSaleh, M., Rasel, H.M. and Ray, B. (2024) A comprehensive review towards resilient rainfall forecasting models using artificial intelligence techniques, Green Technologies and Sustainability, 2, p.100-104.

Aguasca-Colomo, R., Castellanos-Nieves, D. and Méndez, M. (2019) Comparative analysis of rainfall prediction models using machine learning in Islands with complex orography: Tenerife Island, Applied Sciences, 9(22), p. 4931. Available at: http://dx.doi.org/10.3390/app9224931.

Bilgili, M. and Sahin, B. (2010) Prediction of long-term monthly temperature and rainfall in Turkey, Energy Sources, Part A: Recovery, Utilization, and Environmental Effects, 32(1), pp. 60–71. Available at: http://dx.doi.org/10.1080/15567030802467522.

Czarnecka, M. and Nidzgorska-Lencewicz, J. (2011) Impact of weather conditions on winter and summer air quality, International Agrophysics, 25(1), pp. 7–12.

Dos Santos, R.S. (2020) Estimating spatio-temporal air temperature in London (UK) using machine learning and earth observation satellite data, International Journal of Applied Earth Observation and Geoinformation, 88, p. 102066. doi:10.1016/j.jag.2020.102066.

Hanlon, H.M., Bernie, D., Carigi, G. et al. (2021) Future changes to high impact weather in the UK, Climatic Change, 166, p. 50. Available at: https://doi.org/10.1007/s10584-021-03100-5.

Luk, K.C., Ball, J.E. and Sharma, A. (2000) A study of optimal model lag and spatial inputs to artificial neural network for rainfall forecasting, Journal of Hydrology, 227(1–4), pp. 56–65. Available at: http://dx.doi.org/10.1016/S0022-1694(99)00165-1.

Suparta, W. and Samah, A.A. (2020) Rainfall prediction by using ANFIS times series technique in South Tangerang, Indonesia, Geodesy and Geodynamics, 11(6), pp. 411–417. Available at: http://dx.doi.org/10.1016/j.geog.2020.08.001.

Valverde Ramírez, M.C., de Campos Velho, H.F. and Ferreira, N.J. (2005) Artificial neural network technique for rainfall forecasting applied to the São Paulo region, Journal of Hydrology, 301(1), pp. 146–162. Available at: http://dx.doi.org/10.1016/j.jhydrol.2004.06.028.

Yao, R., Sun, P., Bian, Y., Yu, S., Sun, Z., Singh, V.P., Ge, C. and Zhang, Q. (2023) Responses of agricultural drought to meteorological drought under different climatic zones and vegetation types, Journal of Hydrology, 619, article no. 129305. doi:10.1016/j.jhydrol.2023.129305.