**Rainfall Weather Forecasting**

**1. Problem Definition**

The project aimed to leverage machine learning (ML) to predict whether it will rain tomorrow and quantify the expected rainfall, utilising over a decade of daily weather observations across various Australian locales. The predictive model sought to assist in planning and decision-making by accurately forecasting rain.

**2. Data analysis**

The endeavour of weather forecasting, particularly the prediction of rainfall and its intensity, stands at the confluence of data science's precision and the capricious nature of atmospheric phenomena (Bochenek & Ustrnul, 2022). This complex domain necessitates a meticulous analysis of historical weather data, employing an extensive dataset derived from over a decade of daily weather observations across various Australian locales. This dataset is not a mere aggregation of numbers; it represents a detailed narrative of meteorological conditions, comprising variables such as temperature, rainfall, humidity, wind speed, and atmospheric pressure. These variables, integral to the fabric of weather storytelling, offer insights into the climatic rhythm (Latif et al., 2023).

The analysis of this dataset uncovers its dual nature: a rich source of meteorological insight and a minefield of analytical challenges. Primary among these challenges are missing values, particularly in critical variables like "Evaporation" and "Sunshine," which pose significant risks to the integrity and reliability of predictive models (Nguyen et al., 2023). Outliers further complicate the landscape, requiring a nuanced approach to distinguish between anomalies that represent genuine extreme weather events and those that are mere errors (Hael & Yuan, 2020).).

Addressing these challenges necessitates a comprehensive and methodical approach to data preprocessing. This includes employing robust imputation methods for missing data and sophisticated algorithms for outlier detection and treatment. Furthermore, the significance of feature selection and engineering cannot be understated, as the relevance of each variable to rainfall prediction must be meticulously evaluated to optimise model performance (Liu et al., 2010).

The initial examination of the dataset highlights its depth and the granular view it offers of daily weather conditions across different Australian cities. While beneficial, this granularity introduces hurdles such as the aforementioned missing values, outliers, and the inherent variability of weather phenomena. Such issues are not merely obstacles but opportunities to refine analytical techniques and enhance model accuracy (Bauer et al., 2015).

In the project aimed at predicting rainfall, the Exploratory Data Analysis (EDA) phase was pivotal, leveraging a suite of sophisticated tools to unravel the complex interplay of meteorological factors (McKinney, 2012). Among these tools, Matplotlib and Seaborn stood out for their visualisation capabilities, while Pandas streamlined statistical analysis and initial data assessment (Paczkowski, 2022).

Matplotlib, a foundational Python library for creating a wide array of visualisations, was indispensable for examining the distributions and relationships within the dataset. Its versatility allowed for detailed customisations, making it possible to tailor visualisations precisely to the project's analytical needs (Bloice & Holzinger, 2016). This flexibility was crucial for uncovering patterns in variables such as temperature, pressure, and rainfall, which are integral to predicting weather events.

Building upon Matplotlib, Seaborn offered advanced visualisation techniques and aesthetically pleasing default settings that enhanced the interpretability of statistical graphics (Sial et al., 2021). Its functionalities were particularly beneficial for correlation analysis. Seaborn's pair plots also proved invaluable for exploring bivariate relationships, deepening the understanding of how individual factors might influence rainfall.

Pandas, a library renowned for its data manipulation capabilities, was at the core of the statistical analysis. It enabled the computation of descriptive statistics, offering insights into weather variables' central tendencies and dispersion (McKinney, 2012).

Together, these tools formed the analytical backbone of the EDA phase, enabling a nuanced exploration of the dataset. The insights gleaned from this phase informed the predictive modelling strategy and underscored the importance of a thorough preliminary analysis in tackling complex data-driven projects. Through the judicious application of visualisation and statistical analysis tools, the project adeptly navigated the challenges inherent in weather data, setting a solid foundation for accurate rainfall prediction (Meem et al., 2023).

**3. EDA Concluding Remarks**

The EDA revealed nuanced interactions between various meteorological factors influencing rainfall. Seasonal patterns emerged as a significant predictor of rainfall, with variations in precipitation levels aligning with specific times of the year. This observation underscores the importance of considering temporal dynamics in rainfall prediction models, echoing the findings of studies that highlight the critical role of seasonality in weather patterns (Peel et al., 2007).

Geographic variability further complicated the prediction landscape, underscoring the need to integrate spatial data into the models. The variance in rainfall patterns across different Australian regions highlighted the influence of local geographical features and climatic zones on precipitation, aligning with research that emphasises regional climatic influences on weather outcomes (Pagano et al., 2016).

The distribution of rain events, particularly the preponderance of days without rainfall, highlighted the challenges in modelling rainfall occurrences. The skewed distribution towards zero rainfall days necessitated consideration of models capable of handling imbalanced data, reinforcing the complexity of accurately forecasting rain events.

Informed by the EDA, several preliminary decisions were made to refine the predictive models further. The incorporation of temporal and spatial features into the dataset was a direct response to the observed seasonal and geographic variability in rainfall patterns. This approach was designed to capture the temporal trends and regional differences that significantly impact rainfall, thereby enhancing the model's predictive accuracy.

The strong correlations observed between certain weather features and rainfall informed the feature selection process for the predictive models. Temperature and humidity, given their significant relationships with rainfall, were prioritised as key predictors. This decision was grounded in the understanding that these variables play a crucial role in the formation of precipitation, as evidenced by their statistical association with rainfall events in the EDA (Yin et al., 2024).

The skewed distribution of rainfall events led to the adoption of data transformation techniques, such as log transformation, to normalise the target variable's distribution (Feng et al., 2014). This step aimed to mitigate the impact of outliers and address the imbalance in the dataset, thereby improving the models' ability to predict rainfall accurately.

**4. Pre-processing Pipeline**

Ensuring the integrity of meteorological data is crucial in the field of rainfall prediction, where the accuracy of forecasts heavily depends on the quality of input data. Missing data, an omnipresent challenge in meteorological datasets, necessitates a strategic approach to data cleaning and imputation to maintain this integrity. Sterne et al. (2009) highlight the potential and pitfalls of multiple imputations for missing data, underscoring the importance of a careful approach to dealing with data gaps in epidemiological and clinical research, a principle that is equally applicable in the context of meteorological data analysis.

To mitigate the impact of missing data, an evaluative process is implemented to ascertain the proportion of missing information for each variable, alongside establishing a criterion for exclusion. Variables where the percentage of missing data surpasses a set threshold, such as 30%, are typically flagged for potential removal from the dataset. However, exceptions are made for variables with direct meteorological significance to precipitation events, such as Humidity and Pressure readings, due to their indispensable role in forecasting (Wilks, 2011). Conversely, due to their high levels of missing data, which compromise forecast reliability, variables like Evaporation and Sunshine are excluded from the analysis. This decision reflects the guiding principle that the quality of weather forecasting is fundamentally linked to the integrity and reliability of the input data, affirming the necessity for accurate and comprehensive data in making reliable predictions.

Various imputation techniques can be employed for essential variables with manageable levels of missing data. Median imputation might be favoured for continuous variables with a skewed distribution, as it preserves the central tendency without being influenced by outliers, a key consideration in variables like Rainfall, where extreme events are rare but significant (Austin et al., 2021). Decision-making in this phase would be guided by statistical rationale and meteorological insights, ensuring that imputation strategies do not inadvertently distort the data's natural variability.

The creation of new features is a deliberate process aimed at enriching the dataset with predictive insights. In this context, the integration of temporal features such as the month or the season reflects an understanding of the annual rainfall cycle, informed by historical climate patterns exhibiting precipitation seasonality (Cao et al., 2021). Similarly, engineering features to capture daily weather fluctuations, such as the difference between the day's highest and lowest temperatures, taps into the thermodynamic principles influencing weather systems, acknowledging that significant temperature variations can precede weather fronts bringing rainfall (Sillmann et al., 2017). The strategic approach adopted for cleaning and imputation, feature engineering, and the application of encoding and scaling techniques was informed by both the unique characteristics of the dataset and established meteorological principles.

The cleaning process commenced with a thorough assessment of missing data across the dataset. Recognising the detrimental impact that missing data could have on the accuracy of rainfall predictions, variables with a significant portion of missing values exceeding a pragmatically set threshold were judiciously evaluated for removal. This decision-making process was guided by the principle that preserving data integrity was paramount, aligning with the observations made by Ferrari and Ozaki (2014). However, median imputation was employed for indispensable variables like Pressure9am and Pressure3pm, which bear direct meteorological relevance to precipitation events. This choice was influenced by the skewed nature of environmental data, where outliers—such as extreme weather events—could potentially distort the mean. Thus, median imputation served to maintain the central tendency of these critical indicators, a strategy supported by the findings of Austin et al. (2021).

Feature engineering was a deliberate and thoughtful process, aiming to imbue the dataset with new dimensions of predictive power. The creation of features like TemperatureRange and HumidityChange was directly inspired by insights garnered during the EDA phase. These features, capturing the variability in daily temperature and humidity, were predicated on the thermodynamic principles that govern weather systems. The rationale behind TemperatureRange—the difference between the day's highest and lowest temperatures—stemmed from the observation that significant temperature variations often precede weather fronts that could lead to precipitation, echoing the dynamics discussed by Sillmann et al. (2017). Similarly, HumidityChange was conceptualised based on the principle that fluctuations in humidity levels, especially when coupled with appropriate atmospheric conditions, are indicative of impending rainfall events, as elucidated by Ahrens (2015).

The structural diversity of the dataset necessitated the careful encoding of categorical variables and scaling of numerical features. Wind direction, a categorical variable with substantial predictive value, was encoded using one-hot encoding to transform it into a machine-readable numerical format. This method was specifically chosen to preserve the variable's nominal nature without implying any ordinal relationship, addressing the complexity of capturing wind direction's cyclical aspect, as Kosaraju et al. (2023) noted. Scaling, particularly the application of StandardScaler to features like WindSpeed9am and Humidity, was critical in normalising the data, ensuring uniform contribution across variables. This standardisation is indispensable for algorithms that depend on distance measures or gradient descent optimisation, fostering a more equitable learning process and enhancing model performance, as highlighted by Ozsahin et al. (2022).

**5. Building Machine Learning Models**

At the core of the project was a strategic selection of ML models, carefully chosen to align with the dual objectives of predicting the occurrence of rain (classification) and quantifying potential rainfall amounts (regression). For the classification task, the models explored included Logistic Regression, Random Forest Classifier, and XGBoost Classifier, each selected for their unique data analysis and pattern recognition strengths. Logistic Regression was noted for its simplicity and interpretability, the Random Forest Classifier for its ensemble learning approach to combat overfitting, and the XGBoost Classifier for its efficiency and superior performance driven by gradient boosting (Hosmer et al., 2013; Liu et al., 2012; Chen & Guestrin, 2016).

In a similar vein, for the regression task aimed at estimating rainfall amounts, the project leaned on the proven capabilities of the Random Forest Regressor and XGBoost Regressor. Their selection was justified by their ability to adeptly handle the nonlinear relationships often present in meteorological data, showcasing the project's commitment to applying models that accurately capture weather patterns' complexity (Breiman, 2001; Chen & Guestrin, 2016).

Hyperparameter tuning, particularly through RandomizedSearchCV, played a crucial role in optimising the performance of ML models in a rainfall prediction project (Probst et al., 2019). By exploring the parameter space in a randomised manner, the project effectively pinpointed the settings that significantly improved model performance, focusing on models like the XGBoost Classifier. This method struck a delicate balance, enhancing the models' sensitivity to nuanced weather indicators while preventing overfitting, thus maintaining their generalisability to new data. The use of RandomizedSearchCV not only elevated model accuracy and efficiency but also conserved computational resources and provided insightful revelations about model behaviour and weather data dynamics (Bergstra et al., 2011). Specifically, adjusting the XGBoost Classifier's parameters, such as the learning rate and max depth, markedly boosted its predictive accuracy for rainfall events. This improvement highlighted the critical importance of hyperparameter tuning in refining models for intricate tasks like weather forecasting (Chen & Guestrin, 2016).

The implementation of cross-validation provided a robust framework for assessing the models' generalisability to unseen data (James et al., 2013). This methodological approach was crucial in verifying the models' ability to provide reliable predictions across varying weather conditions, affirming the project's dedication to developing robust predictive tools (Raschka, 2018).

The evaluation phase extended beyond mere accuracy measures, adopting a multifaceted approach to assess model performance. For the classification task, precision, recall, and the F1 score offered insights into the models' effectiveness in accurately predicting rainfall occurrences (Powers, 2020). These metrics illuminated the careful considerations involved in model selection, guiding the project towards models that offered a balanced approach to prediction.

The regression models were similarly scrutinised, with Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and Mean Absolute Error (MAE) serving as critical lenses for evaluating the accuracy of rainfall amount predictions (Chai & Draxler, 2014). This nuanced evaluation process highlighted the XGBoost Classifier and Regressor for their exceptional performance, underscoring their potential as reliable forecasting tools (Chen & Guestrin, 2016).

**6. Concluding Remarks**

The culmination of this project in ML to predict rainfall stands as a testament to the prowess of the XGBClassifier and XGBRegressor models, showcasing their exceptional performance in forecasting rain occurrence and quantifying rainfall amounts, respectively (Chen & Guestrin, 2016). These results validate the models' robustness and capability to navigate the complex, nonlinear intricacies of meteorological data and exemplify the project's triumph in harnessing advanced data science methodologies for a crucial environmental science domain.

The XGBClassifier, distinguished by its adeptness in handling diverse datasets and its strategic minimisation of overfitting, has confirmed its superiority in accurately classifying weather events. Achieving an impressive accuracy of 87.91% on the test dataset—after meticulous hyperparameter tuning focusing on 'max\_depth' and 'n\_estimators'—it stands out for its precision, significantly surpassing baseline models in predicting rain occurrence, thereby reducing false positives and enhancing forecast precision (Probst et al., 2019).

Similarly, the XGBRegressor has manifested its precision in rainfall quantification, achieving a Mean Absolute Error (MAE) of just 2.3 mm on rainfall amount predictions, post adjustments in 'learning rate' and 'subsample' parameters. This accuracy is pivotal for agriculture and urban planning sectors, where precise rainfall data can critically influence decision-making processes (Friedman, 2001).

The integration of these models into existing weather forecasting frameworks presents a promising pathway for improving the reliability and accuracy of weather predictions—critical for disaster preparedness, agricultural planning, and water resource management (Karpatne et al., 2018). This project not only underlines the necessity for ongoing model enhancement and adaptation through iterative training and leveraging advancements in ML but also signals the immense potential for future research and application in meteorological forecasting (Reichstein et al., 2019).

The XGBClassifier's potential for integration into agricultural advisory services could significantly benefit farmers by informing decisions on planting and irrigation, thereby optimising resource use and crop yields (Friedman, 2001). Concurrently, the XGBRegressor's accuracy in estimating rainfall amounts could revolutionise flood prediction systems, offering more precise and timely rainfall forecasts to bolster emergency response and mitigation efforts (Chen & Guestrin, 2016).

**7. Recommendations**

Implementing ML models in the field of meteorology involves a multifaceted approach, ensuring that these advanced analytical tools remain effective and relevant over time. A critical element in this process is the periodic retraining of models to adapt to the ever-changing weather patterns and climate data dynamics. As Aljohani (2023) highlighted, incorporating the latest observations into models through routine retraining schedules is vital for maintaining accuracy and enhancing predictive capabilities. This practice allows models to capture new trends and anomalies in meteorological data, improving their performance.

Another essential aspect is the continuous monitoring of model performance against real-time weather events. McGovern et al. (2017) emphasise the importance of evaluating models to promptly gain insights into their accuracy and reliability. The use of drift detection techniques, as discussed by Yeshchenko, (2019), plays a significant role in identifying when models begin to deviate from expected performance standards. This ongoing evaluation ensures that models remain robust and generalisable to unseen data, enabling practitioners to make necessary adjustments or retrain models to maintain their predictive reliability.

Moreover, the integration of user feedback into the model development and refinement process is crucial. Fundel et al. (2019) advocate for engaging end-users, such as meteorologists, emergency response teams, and the public, to provide practical insights into the models' utility and accuracy. Establishing channels for collecting and analysing feedback allows for the iterative improvement of models, aligning their development with user needs and enhancing their applicability in real-world scenarios. Together, these recommendations form a comprehensive strategy for implementing ML models in the dynamic and complex field of meteorology.Top of Form

**References**

Aljohani, A. (2023). Predictive analytics and machine learning for real-time supply chain risk mitigation and agility. *Sustainability*, *15*(20), 15088. https://doi.org/10.3390/su152015088

Ahrens, C. D. (2015). *Meteorology today: an introduction to weather, climate, and the environment*. Cengage Learning Canada Inc.

Austin, P. C., White, I. R., Lee, D. S., & van Buuren, S. (2021). Missing Data in Clinical Research: A Tutorial on Multiple Imputation. *The Canadian journal of cardiology*, *37*(9), 1322–1331. https://doi.org/10.1016/j.cjca.2020.11.010

Bauer, P., Thorpe, A., & Brunet, G. (2015). The quiet revolution of numerical weather prediction. *Nature*, *525*(7567), 47-55. https://doi.org/10.1038/nature14956

Bergstra, J., Bardenet, R., Bengio, Y., & Kégl, B. (2011). Algorithms for hyper-parameter optimization. *Advances in neural information processing systems*, *24*.

Bloice, M. D., & Holzinger, A. (2016). A tutorial on machine learning and data science tools with python. *Machine Learning for Health Informatics: State-of-the-Art and Future Challenges*, 435-480. https://doi.org/10.1007/978-3-319-50478-0\_22

Bochenek, B., & Ustrnul, Z. (2022). Machine learning in weather prediction and climate analyses—applications and perspectives. *Atmosphere*, *13*(2), 180. <https://doi.org/10.3390/atmos13020180>

Breiman, L. (2001). Random forests. *Machine learning*, *45*, 5-32.

Cao, F., Gao, T., Dan, L., Zhao, F., Gong, X., Zhang, X., ... & Zhan, J. (2021). Contributions of natural climate variability on the trends of seasonal precipitation extremes over China. *International Journal of Climatology*, *41*(11), 5226-5242.

Chai, T., & Draxler, R. R. (2014). Root mean square error (RMSE) or mean absolute error (MAE)?–Arguments against avoiding RMSE in the literature. *Geoscientific model development*, *7*(3), 1247-1250. https://doi.org/10.5194/gmd-7-1247-2014

Chen, T., & Guestrin, C. (2016, August). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* (pp. 785-794).

Feng, C., Wang, H., Lu, N., Chen, T., He, H., Lu, Y., & Tu, X. M. (2014). Log-transformation and its implications for data analysis. *Shanghai archives of psychiatry*, *26*(2), 105–109. https://doi.org/10.3969/j.issn.1002-0829.2014.02.009

Ferrari, G. T., & Ozaki, V. (2014). Missing data imputation of climate datasets: Implications to modeling extreme drought events. *Revista Brasileira de Meteorologia*, *29*, 21-28.

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of statistics*, 1189-1232.

Fundel, V. J., Fleischhut, N., Herzog, S. M., Göber, M., & Hagedorn, R. (2019). Promoting the use of probabilistic weather forecasts through a dialogue between scientists, developers and end‐users. *Quarterly Journal of the Royal Meteorological Society*, *145*, 210-231.

Hael, M. A., & Yuan, Y. (2020). Identifying extreme rainfall events using functional outliers detection methods. *Journal of Data Analysis and Information Processing*, *8*(04), 282. https://doi.org/10.4236/jdaip.2020.84016

Hosmer Jr, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied logistic regression*. John Wiley & Sons.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning* (Vol. 112, p. 18). New York: springer.

Karpatne, A., Ebert-Uphoff, I., Ravela, S., Babaie, H. A., & Kumar, V. (2018). Machine learning for the geosciences: Challenges and opportunities. *IEEE Transactions on Knowledge and Data Engineering*, *31*(8), 1544-1554.

Kosaraju, N., Sankepally, S. R., & Mallikharjuna Rao, K. (2023). Categorical data: Need, encoding, selection of encoding method and its emergence in machine learning models—a practical review study on heart disease prediction dataset using pearson correlation. In *Proceedings of International Conference on Data Science and Applications: ICDSA 2022, Volume 1* (pp. 369-382). Singapore: Springer Nature Singapore. https://doi.org/10.1007/978-981-19-6631-6\_26

Latif, S. D., Hazrin, N. A. B., Koo, C. H., Ng, J. L., Chaplot, B., Huang, Y. F., ... & Ahmed, A. N. (2023). Assessing rainfall prediction models: Exploring the advantages of machine learning and remote sensing approaches. *Alexandria Engineering Journal*, *82*, 16-25. <https://doi.org/10.1016/j.aej.2023.09.060>

Liu, Y., Wang, Y., & Zhang, J. (2012). New machine learning algorithm: Random forest. In *Information Computing and Applications: Third International Conference, ICICA 2012, Chengde, China, September 14-16, 2012. Proceedings 3* (pp. 246-252). Springer Berlin Heidelberg.

Liu, Y., Li, Z., Xiong, H., Gao, X., & Wu, J. (2010). Understanding of internal clustering validation measures. In *2010 IEEE international conference on data mining* (pp. 911-916). IEEE. https://doi.org/10.1109/ICDM.2010.35.

McGovern, A., Elmore, K. L., Gagne, D. J., Haupt, S. E., Karstens, C. D., Lagerquist, R., ... & Williams, J. K. (2017). Using artificial intelligence to improve real-time decision-making for high-impact weather. *Bulletin of the American Meteorological Society*, *98*(10), 2073-2090. https://doi.org/10.1175/BAMS-D-16-0123.1

McKinney, W. (2012). *Python for data analysis: Data wrangling with Pandas, NumPy, and IPython*. " O'Reilly Media, Inc.".

Meem, S. M., Hossain, M. T., Chowdhury, J. K., Miah, M. S. U., & Monir, M. F. (2023). Understanding the Dynamics of Dengue in Bangladesh: EDA, Climate Correlation, and Predictive Modeling. In *TENCON 2023-2023 IEEE Region 10 Conference (TENCON)* (pp. 1309-1314). IEEE. https://doi.org/10.1109/TENCON58879.2023.10322490

Nguyen, T., Brandstetter, J., Kapoor, A., Gupta, J. K., & Grover, A. (2023). Climax: A foundation model for weather and climate. *arXiv preprint arXiv:2301.10343*. <https://doi.org/10.48550/arXiv.2301.10343>

Ozsahin, D. U., Mustapha, M. T., Mubarak, A. S., Ameen, Z. S., & Uzun, B. (2022). Impact of feature scaling on machine learning models for the diagnosis of diabetes. In *2022 International Conference on Artificial Intelligence in Everything (AIE)* (pp. 87-94). IEEE. https://doi.org/10.1109/AIE57029.2022.00024.

Paczkowski, W. R. (2022). *Modern survey analysis: Using Python for deeper insights*. Springer Nature.

Pagano, T. C., Elliott, J. F., Anderson, B. G., & Perkins, J. K. (2016). Australian bureau of meteorology flood forecasting and warning. In *Flood forecasting* (pp. 3-40). Academic Press.

Peel, M. C., Finlayson, B. L., & McMahon, T. A. (2007). Updated world map of the Köppen-Geiger climate classification. *Hydrology and earth system sciences*, *11*(5), 1633-1644.

Powers, D. M. (2020). Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. *arXiv preprint arXiv:2010.16061*. https://doi.org/10.48550/arXiv.2010.16061

Probst, P., Wright, M. N., & Boulesteix, A. L. (2019). Hyperparameters and tuning strategies for random forest. *Wiley Interdisciplinary Reviews: data mining and knowledge discovery*, *9*(3), e1301. https://doi.org/10.1002/widm.1301

Raschka, S. (2018). Model evaluation, model selection, and algorithm selection in machine learning. *arXiv preprint arXiv:1811.12808*. https://doi.org/10.48550/arXiv.1811.12808

Reichstein, M., Camps-Valls, G., Stevens, B., Jung, M., Denzler, J., Carvalhais, N., & Prabhat, F. (2019). Deep learning and process understanding for data-driven Earth system science. *Nature*, *566*(7743), 195-204.

Sial, A. H., Rashdi, S. Y. S., & Khan, A. H. (2021). Comparative analysis of data visualization libraries Matplotlib and Seaborn in Python. *International Journal*, *10*(1), 45. <https://doi.org/10.30534/ijatcse/2021/391012021>

Sillmann, J., Thorarinsdottir, T., Keenlyside, N., Schaller, N., Alexander, L. V., Hegerl, G., ... & Zwiers, F. W. (2017). Understanding, modeling and predicting weather and climate extremes: Challenges and opportunities. *Weather and climate extremes*, *18*, 65-74.

Sterne, J. A., White, I. R., Carlin, J. B., Spratt, M., Royston, P., Kenward, M. G., ... & Carpenter, J. R. (2009). Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *Bmj*, *338*. https://doi.org/10.1136/bmj.b2393

Stull, R. B. (2015). *Practical meteorology: an algebra-based survey of atmospheric science*. University of British Columbia.

Wilks, D. S. (2011). *Statistical methods in the atmospheric sciences* (Vol. 100). Academic press.

Yeshchenko, A., Di Ciccio, C., Mendling, J., & Polyvyanyy, A. (2019). Comprehensive process drift detection with visual analytics. In *International Conference on Conceptual Modeling* (pp. 119-135). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-33223-5\_11

Yin, Y., He, J., Guo, J., Song, W., Zheng, H., & Dan, J. (2024). Enhancing precipitation estimation accuracy: An evaluation of traditional and machine learning approaches in rainfall predictions. *Journal of Atmospheric and Solar-Terrestrial Physics*, *255*, 106175. https://doi.org/10.1016/j.jastp.2024.106175