Disaggregating Daily Precipitation Data 1990 to 2022 into Half-Hourly Intervals Using LSTM Models

Harrison Oates[1,2], Nayan Arora[2], Hong Gic Oh[2] and Trevor Lee[2]

[1]*School of Computing, The Australian National University, Canberra ACT 2601, Australia*
[2]*Exemplary Energy, Fadden ACT 2904, Australia*
harrison@harrisonoates.com exemplary.energy@exemplary.com.au

Introduction

Building modelling software is increasingly used to optimize design parameters for efficiency and to predict building performance under various conditions (de Wilde, 2023). A critical component of these systems is the availability of meaningful weather and climate data. The World Meteorological Organization (WMO) recommends using at least thirty years of historical data to define climate norms and extremes, as shorter periods may not produce reliable statistics due, for example, to annual and shorter-term precipitation variances (WMO, 2023).

In the Australian context, the Bureau of Meteorology (BoM) has only measured precipitation at half-hourly intervals since the progressive installation of automatic Tipping Bucket Rain Gauges from the late 1990s. Prior to this, precipitation data was primarily collected through daily manual readings by post office staff or volunteers at 9:00AM clock time. However, for reliable built environment modelling, hourly or sub-hourly data is essential. This discrepancy highlights a clear need for algorithms capable of producing fine-scale temporal data based on daily readings and hourly measurements of other weather elements.

The process of generating finer temporal resolution data (e.g., half-hourly) from coarser scale measurements (e.g., daily) is known as disaggregation. In the context of precipitation, this technique is crucial for bridging the gap between available historical data and the requirements of modern building performance simulations, hygrothermal models, and hydrological models.

In this abstract, we present a machine learning approach for disaggregating daily precipitation data into half-hourly intervals, specifically tailored for the Australian climate data collection context[1]. The basis of our model is long short-term memory (LSTM), a type of recurrent neural network that can capture temporal correlations in sequential data over an arbitrary timeframe (Gers et al., 2000). We demonstrate the effectiveness of our approach in four Australian capital cities using BoM data.

Data Selection and Preparation

Using data sourced from BoM, we present results for four stations, each representing a different climate zone as per the Australian Building Codes Board (2024). These are presented in Table 1.

Table **1 Climate zones of investigated locations**

Location	Climate Zone	Precipitation Half-hourly Record Start
Brisbane	Climate Zone 2	2000-03
Sydney	Climate Zone 5	1998-12
Melbourne	Climate Zone 6	1997-10
Canberra	Climate Zone 7	2000-04

For each weather station, we utilised 33 years of hourly non-precipitation data in TMY2 format from 1990 – 2022, which we linearly interpolated to half-hourly intervals. While Brisbane's data required

[1]EPW and ACDB formats use different timestamp conventions: EPW represents the hour before the timestamp, while ACDB represents the hour centred on the timestamp. Half-hourly data generation is necessary to accommodate this difference.

no temporal adjustment, the other locations' daily precipitation measurements needed to be aligned to account for daylight saving time. These locations undergo biannual one-hour shifts, requiring temporal alignment between the daily precipitation readings and other meteorological elements, as the half-hourly precipitation data is consistently recorded in standard time. We then perform an inner join of this interpolated data with the station's complete half-hourly precipitation data. This results in a time series dating from the half-hourly precipitation record start date. We test the model on all data from 2020 to 2022, use 2018 and 2019 as the validation set to protect against model overfitting and train on the remaining data. This results in a train-validation-test split of roughly 75%-10%-15%. The training dataset was shuffled to allow the model to learn from a more representative sample in each batch.

We selected dew point temperature (DPT), dry bulb temperature (DBT), atmospheric pressure, and relative humidity (RH) as the model's input features based on Pearson correlation to precipitation. These were standardised to have a mean of zero and a standard deviation of one before being passed to the model[2].

The Model

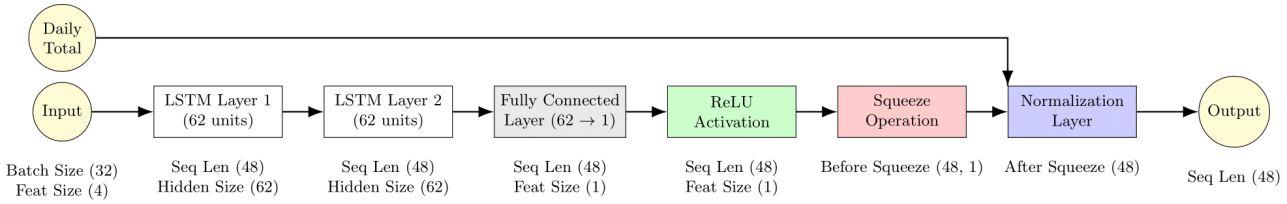

Figure 1: Model Architecture

We have designed a novel deep learning model, based on the LSTM architecture, specifically for precipitation disaggregation in the Australian climate context. Figure 1 illustrates the model's architecture. Each sequence of 48 half-hourly time steps is processed through two successive LSTM layers. Each LSTM layer contains 62 LSTM units, which capture temporal correlations in the data. The output of these layers then passes through a fully connected layer, reducing the feature dimension from 62 to 1 for each time step. A ReLU (rectified linear unit) activation follows, which ensures that all predicted precipitation values are non-negative, as rainfall cannot be negative. This is essential to prevent unrealistic outputs from the model. We then adjust the dimensions of our data before passing it to a normalisation layer. The normalisation layer scales the estimated half-hourly precipitation values to ensure they sum to the given daily total. The final output is a single tensor (an ordered set of values) representing the day's half-hourly precipitation estimates.

The model was implemented and trained using PyTorch. We defined our loss function for a predicted tensor p and target tensor q as the sum of the mean squared error between p and q, the Kullback-Liebler divergence between the softmax outputs of p and q, and the difference in variance:

$$\ell(p, q) = \mathrm{MSE}(p, q) + KL(\sigma(p), \sigma(q)) + |V(p) - V(q)|$$

The first two terms aim to measure overall prediction accuracy and assess differences in the relative distribution of rainfall within the day, while the last term aims to preserve the statistical characteristics of the target tensor in the prediction, particularly the magnitude of extreme values.

To train the model, we employed the Adam optimizer with an initial learning rate of 1e-3 and a batch size of 32. Learning rate scheduling was implemented with a reduction on plateau strategy to fine-tune the model's performance. After an initial run of 140 epochs (training cycles), the epoch

[2] Correlation with cloud cover, while strong in reality, has been overlooked because it is rarely available in the early years other than as a derivative of solar irradiation data which leaves all night hours cloud cover data as unreliable linear interpolations between pre-dusk and post-dawn values.

with the lowest validation loss was selected and trained with a learning rate of 5e-6 for a further 50 epochs. This helped to further reduce the validation loss rate.

The model was trained on a single NVIDIA 4070 Ti Super GPU and took around 2.8 seconds per epoch. The entire training and testing pipeline takes just over 9 minutes to complete per location. The model weights corresponding to the lowest validation loss usually occur relatively early in the training run, before epoch 70, so there is scope to significantly reduce the number of epochs the model is trained to further cut runtime without sacrificing model quality.

Results

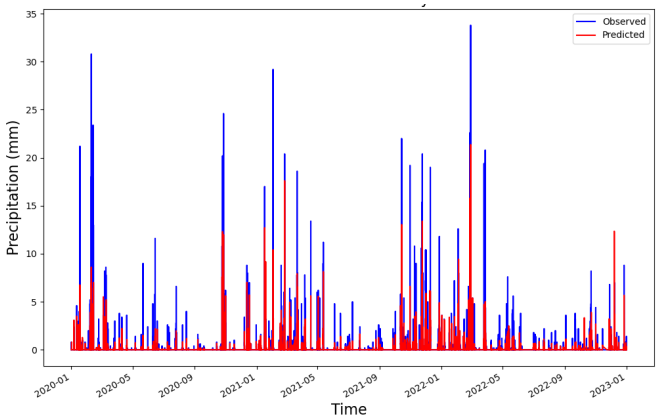

Figure 2: Half-hourly series for Brisbane

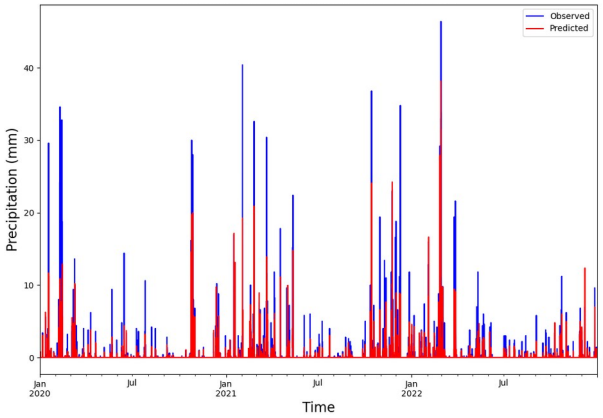

Figure 3: Hourly series for Brisbane

Table 2 Model results

Location	RMSE (mm)	Relative error in total number of precipitation half-hours (%)	Proportion of correctly detected precipitation half-hours with no error (%)
Brisbane	0.57	8.36%	61.63%
Sydney	0.51	13.15%	69.71%
Melbourne	0.23	8.02%	56.76%
Canberra	0.29	22.35%	67.08%
Mean	0.40	12.97%	63.80%

Table 3 Comparison of results with Ferrari, et al.

Model	RMSE (mm)[3]	Relative error in total number of precipitation hours (%)	Proportion of correctly detected precipitation hours (%)
LSTM (average)	0.45	12.57%	69.04%
Markov chain Monte Carlo	0.65	~7%	20%

Figure 2 illustrates the model's performance for Brisbane, comparing generated and observed half-hourly precipitation. While the model generally captures the temporal patterns of rainfall (as seen in the hourly aggregation in Figure 3), it notably underestimates the magnitude of extreme events. This limitation arises from two main factors: the inherent smoothing effect of LSTM models, which

[3]RMSE is sensitive to climate variance. Canberra RMSE values are used to enable a fair comparison.

prioritize long-term dependencies over sharp fluctuations, and the relative scarcity of extreme rainfall instances in the training data. In the Brisbane training set, only 6.13% of wet days (9AM to 9AM) featured half-hourly precipitation exceeding 10mm, with such events comprising just 1.5% of all wet half-hours.

Table 2 summarises our results on the test dataset. Across all climate zones, our model achieves an average error of 0.4 mm. Further, we evaluate its performance on two characteristics introduced by Ferrari et al. (2022) as desirable for a disaggregation model: wet half-hour frequency preservation and wet half-hour detection. The model's output displays an error of 12.97% in the number of rainfall half-hours and detects 63.80% of rainfall half-hours correctly. Table 3 compares our hourly re-aggregated series with the Markov chain Monte Carlo (MCMC) model of Ferrari et al. The results indicated that our LSTM-based approach yields lower error rates and improved detection of rainfall hours compared to MCMC. While our model shows a slightly higher error in estimating the total number of rainfall hours, this performance is still commendable, especially considering that our model primarily focuses on half-hourly, rather than hourly, disaggregation.

Conclusion

Our model demonstrates robust performance in precipitation disaggregation across various Australian climate zones. Its ability to maintain accuracy while preserving important precipitation characteristics such as total number and distribution of wet half-hours makes it valuable for generating data that can be used for building performance simulations.

The performance variations across cities suggest that local climate patterns significantly impact disaggregation accuracy. To address this, we could perform a hyperparameter search for each station to optimise the model's architecture. However, given the computational intensity of this approach, a more practical alternative might be to select an architecture per climate zone.

Several avenues for further investigation include:

- Enhancing the model's ability to capture fine-grained precipitation patterns by incorporating additional meteorological variables and refining the model architecture to better handle rainfall intermittency;
- Applying the model to more locations and conducting further performance evaluations; and
- Training the model on an entire climate zone instead of individual stations, which could potentially improve model performance due to increased data availability.

These refinements could further increase the model's accuracy under various climate contexts. The generated series can then be used to define a climate normal, ensuring that precipitation can be reliably used for modelling and simulation of built environments.

References

Australian Building Codes Board, 2024, 'Climate zone map'. Available at https://www.abcb.gov.au/resources/climate-zone-map. Accessed 5th July 2024.

De Wilde, P, 2023, 'Building performance simulation in the brave new world of artificial intelligence and digital twins: A systematic review', *Energy and Buildings, 292*, p113-171.

Ferrari, D., Mahmoodi, M., Kodagoda, C., Hameed, N.A., Lee, T., and Anderson, G., 2022, 'Disaggregation of precipitation data applicable for climate-aware planning in built enviroments'. *Australian Building Simulation 2022 Conference Proceedings, 2022*, p24-27

Gers, F.A., Schmidhuber, J., Cummins, F., 2000, 'Learning to Forget: Continual Prediction with LSTM' *Neural Computation* 12(10), p2451–24

World Meteorological Organization, 2023. 'Guide to Climatological Practices', 3rd edn.