Streamflow forecasting under non-homogeneous and discontinuous data conditions using LSTM: Application to Nam Pan basin

Truong Van Anh[1*], Huynh Thi Lan Huong[1], Hoang Thi Nguyet Minh[1], Tran Duy Kieu[1]

Abstract: Streamflow forecasting in small mountainous basins faces significant challenges due to non-homogeneous and discontinuous observational data caused by sensor failures, irregular sampling, and missing records. This study presents a Long Short-Term Memory (LSTM) deep learning approach to handle discontinuous time series for water level forecasting in the Nam Pan River basin, northern Vietnam. The methodology integrates multiple data preprocessing techniques including linear and PCHIP interpolation, outlier removal, and Z-score normalization to address data irregularities. Input features comprise rainfall observations from multiple stations, lagged water levels (6–24 hours), and cyclical time encoding. The LSTM model achieves Nash–Sutcliffe Efficiency of 0.86–0.91 for 6-hour forecasts and 0.70–0.78 for 24-hour forecasts, with R^2 values of 0.88–0.94 and forecast assurance of 82–90%. Results demonstrate the model's robustness in handling imperfect data, confirming its applicability for operational flash flood early warning systems in data-limited mountainous catchments.

Keywords: LSTM, Nam Pan river basin, flow forecasting, small mountainous catchment.

1. Introduction

Streamflow forecasting plays a critical role in water resource management, disaster prevention, and reservoir operation, particularly in small mountainous basins where flash floods frequently occur due to steep terrain, seasonally concentrated rainfall, and short water concentration times. In such contexts, traditional hydrological models, which typically require spatially continuous and physically consistent datasets (e.g., topography, land use, infiltration parameters, and dense rainfall–runoff monitoring networks), are often constrained by sparse measurement infrastructure and incomplete datasets.

The rapid development of artificial intelligence, especially deep learning, over the past two decades has introduced more effective approaches for modeling complex nonlinear rainfall–runoff processes. Among these, the Long Short-Term Memory (LSTM) neural network has demonstrated significant advantages due to its ability to learn and retain long-term temporal dependencies, enabling reliable modeling of hydrological series that are temporally discontinuous, non-homogeneous, or fragmented-conditions under which traditional ANN, ARIMA, or NARX models commonly fail (Shen, 2018; Kratzert et al., 2019).

Recent studies have shown that LSTM and its enhanced variants such as CNN–LSTM and Attention-LSTM can forecast hourly or daily streamflow with high accuracy even under data-scarce or temporally unstable conditions (Hu et al., 2021; Bai et al., 2023). These capacities are particularly relevant to the Nam Pan basin, a small mountainous catchment within the Da River system, which exhibits strong seasonal rainfall variability, complex topography, and high flash-flood susceptibility. However, the basin's hydrometeorological observation network remains limited, leading to rainfall and runoff data that are spatially uneven and temporally discontinuous due to measurement interruptions or incomplete historical records. Such characteristics introduce substantial challenges for traditional physically based hydrological models and emphasize the need for data-driven forecasting approaches capable of handling fragmented and non-homogeneous datasets.

Therefore, developing an LSTM-based streamflow forecasting model for the Nam Pan basin under conditions of limited spatial data and non-homogeneous time series is not only scientifically meaningful in evaluating the capability of deep learning for poorly gauged basins but also practically important for improving flash flood early-warning systems in mountainous regions of Vietnam, where extreme hydro-meteorological events are increasingly intensified by climate change.

2. Study area

The Nam Pan River basin is located in Son La Province and is one of the tributary catchments draining into the Da River. This basin has a relatively small area and is characterized by mountainous terrain with steep slopes and highly dissected topography. The steep longitudinal gradient results in rapid hydrological responses, short concentration times, and flood waves that may rise and fall abruptly within just a few hours. These characteristics create favorable conditions for the occurrence of flash floods, particularly during intense short-duration rainfall events.

The basin's climate follows a tropical monsoon regime, with rainfall distributed unevenly across seasons and space. The wet season typically extends from May to October and accounts for the majority of

[1]Hanoi University of Natural Resources and Environment
[*] Corresponding author

Received 21st Nov. 2025
Accepted 27th Nov. 2025
Publication date 31st Dec. 2025

the annual precipitation. Spatial rainfall distribution is also highly variable due to orographic influences and windward–leeward effects, which can produce localized heavy rainfall that generates significant flooding at downstream locations. This spatial heterogeneity represents a major challenge for hydrological modeling and complicates the interpretation of rainfall–runoff processes.

A key challenge in the Nam Pan basin is the sparse and uneven distribution of hydrometeorological monitoring stations, which is insufficient to capture the spatial variability of rainfall. Additionally, hydrometeorological datasets are often temporally discontinuous, containing missing periods due to instrument malfunction, inconsistent measurement intervals, or temporary suspension of station operations. Streamflow observations at the outlet station also suffer from temporal gaps and relatively short observation records, limiting their suitability for traditional physically based hydrological models that require continuous, high-quality data.

These characteristics0steep terrain, spatially uneven rainfall, rapid hydrological response, and limited data availability-highlight the Nam Pan basin as a representative case of small mountainous catchments in Vietnam, where the application of traditional hydrological modeling approaches faces significant constraints. In such contexts, deep learning models like LSTM, which can effectively handle non-homogeneous and discontinuous time series and learn nonlinear rainfall–runoff relationships without extensive spatial information, become a promising alternative. The selection of the Nam Pan basin for this study therefore reflects not only its scientific relevance but also the practical need to enhance flash-flood forecasting and water resource management in small mountainous basins across northern Vietnam (Figure 1).

3. Data and methodology

3.1. Data

The dataset employed in this study was collected within the Nam Pan River basin and comprises rainfall, water level. Rainfall data were initially obtained from automatic monitoring stations located at Na Cang, Co Chia, Ban Mờn, and Hat Lot for the period 31 May 2019 to 31 December 2024; however, these records were temporally discontinuous and exhibited considerable measurement uncertainties. To address these limitations, additional rainfall series with 1-hour and 6-hour intervals were drawn from national meteorological stations at Son La, Co Noi, and Yen Chau.

Water level data were acquired from the Hat Lot hydrological station, including automatic observations from 2019–2024, supplemented by manual measurements recorded during flood seasons from 2000–2017 and automatic observations from 2018–2022. These combined datasets reveal that hydrometeorological observations in the basin are fragmented in temporal domains, non-homogeneous, and subject to interruptions in monitoring—typical characteristics of small mountainous catchments in Northwest Vietnam.

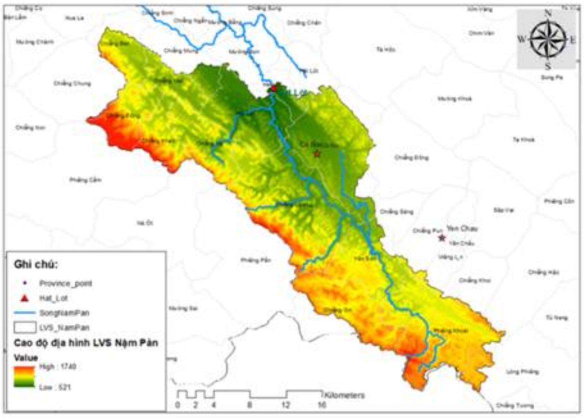

Figure 1. Nam Pan river basin

Due to the discontinuity and missing segments inherent in the observational datasets, a comprehensive preprocessing workflow was conducted to homogenize the time series prior to model development. This process involved identifying missing entries and anomalous values using functions for detecting missingness, unsorted records, and non-uniform intervals. The time series were standardized by reordering records chronologically, removing duplicates, and applying gap-filling techniques for short missing intervals. For water level data, outliers were removed, and short gaps were filled using linear or Piecewise Cubic Hermite Interpolating Polynomial (PCHIP) interpolation, while longer missing segments were excluded from the modeling dataset. Because the basin has a concentration time of approximately 10–12 hours, only short water-level gaps ($\leq$ 12 h) were interpolated, while longer gaps and all missing rainfall segments were removed to avoid producing hydrologically unrealistic values and to maintain transparency in the dataset. All variables were normalized using Z-score normalization to ensure stable convergence of the deep learning model and to minimize biases arising from differing variable scales.

3.2. Methodology

The forecasting model developed in this study is based on the Long Short-Term Memory (LSTM) neural network, selected for its strong capacity to capture nonlinear rainfall–runoff relationships and to accommodate non-homogeneous and discontinuous time series-conditions characteristic of the Nam Pan basin. In addition to conventional input variables such as rainfall series from Son La, Co Noi, and Hat Lot stations and lagged water levels at Hat Lot, the model incorporates time-dependent variables encoded using sine and cosine transformations. This cyclical encoding preserves the periodic characteristics of hydrological processes and eliminates artificial discontinuities at temporal boundaries (e.g., 23:00–00:00 or December–January), enabling the model to identify recurring patterns even when input data are incomplete or fragmented.

Furthermore, the inclusion of multiple water level lags (6 h, 12 h, 18 h, and 24 h) provides the model with information on antecedent hydrological conditions, including basin storage, soil moisture, and delayed runoff routing. These features strengthen the model's ability to capture nonlinear dependencies in a steep, fast-responding mountainous basin where hydrological states evolve rapidly over time.

The overall research workflow is summarized in Figure 3.

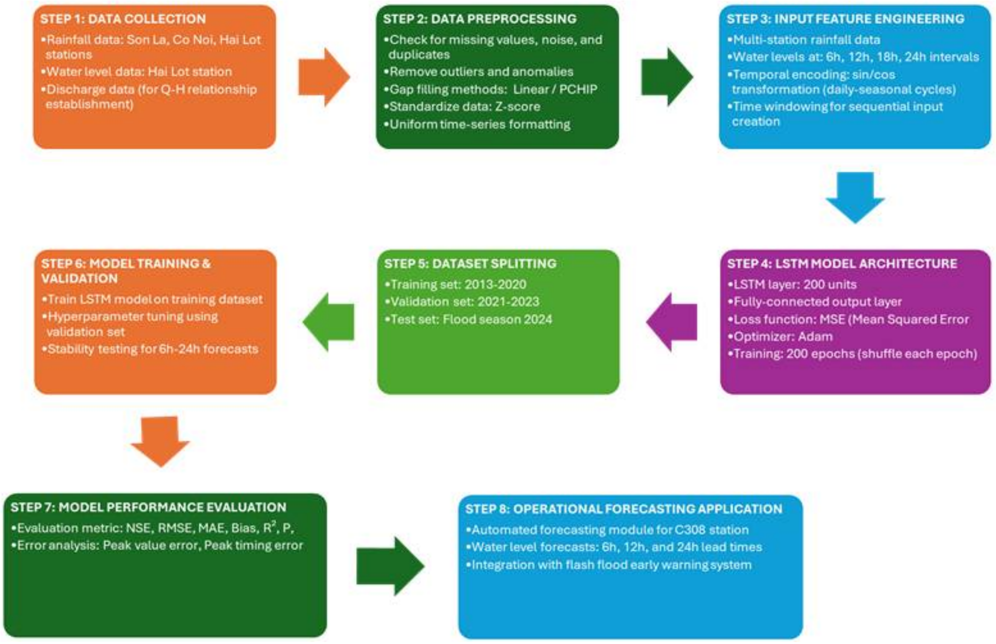
STEP 1: DATA COLLECTION
•Rainfall data: Son La, Co Noi, Hai Lot stations
•Water level data: Hai Lot station
•Discharge data (for Q-H relationship establishment)
STEP 2: DATA PREPROCESSING
•Check for missing values, noise, and duplicates
•Remove outliers and anomalies
•Gap filling methods: Linear / PCHIP
•Standardize data: Z-score
•Uniform time-series formatting
STEP 3: INPUT FEATURE ENGINEERING
•Multi-station rainfall data
•Water levels at: 6h, 12h, 18h, 24h intervals
•Temporal encoding: sin/cos transformation (daily-seasonal cycles)
•Time windowing for sequential input creation
STEP 4: LSTM MODEL ARCHITECTURE
•LSTM layer: 200 units
•Fully-connected output layer
•Loss function: MSE (Mean Squared Error
•Optimizer: Adam
•Training: 200 epochs (shuffle each epoch)
STEP 5: DATASET SPLITTING
•Training set: 2013-2020
•Validation set: 2021-2023
•Test set: Flood season 2024
STEP 6: MODEL TRAINING & VALIDATION
•Train LSTM model on training dataset
•Hyperparameter tuning using validation set
•Stability testing for 6h-24h forecasts
STEP 7: MODEL PERFORMANCE EVALUATION
•Evaluation metric: NSE, RMSE, MAE, Bias, R², P,
•Error analysis: Peak value error, Peak timing error
STEP 8: OPERATIONAL FORECASTING APPLICATION
•Automated forecasting module for C308 station
•Water level forecasts: 6h, 12h, and 24h lead times
•Integration with flash flood early warning system

Figure 2. Methodological framework for LSTM-based water level forecasting at Hat Lot Station. The process consists of five main stages: data collection and preprocessing of rainfall and water level data from multiple stations with quality control procedures; feature engineering incorporating temporal lag features (6h, 12h, 18h, 24h) and cyclical time encoding; dataset creation defining input matrix X (rainfall, lagged water levels, time features) and target variable Y (water level at time t) with train-validation-test split; LSTM model architecture with 200 units, fully connected layer, and Adam optimization; and comprehensive model evaluation using NSE, RMSE, MAE, and R² metrics, followed by 6-hour, 12-hour, and 24-hour ahead forecasts.

4. Result and Discussion

4.1. Training Performance

The LSTM model was trained and validated using hydrometeorological data in flood season from 2013 to 2020. The LSTM model demonstrates a strong agreement between observed and simulated water levels, particularly during the rising limb of major flood events (Figure 3).

The evaluation results demonstrate that the model effectively captures the temporal dynamics of rising and falling flood hydrographs, reproducing both the timing and magnitude of major peaks. During the calibration period, the LSTM model achieved Nash–Sutcliffe Efficiency (NSE) values ranging from 0.86 to 0.91 for 6-hour lead times, with RMSE values between 0.22 m and 0.31 m and coefficients of determination (R^2) between 0.89 and 0.94. These values indicate strong agreement between simulated and observed water levels, particularly during flood peaks when rapid changes occur over short durations.

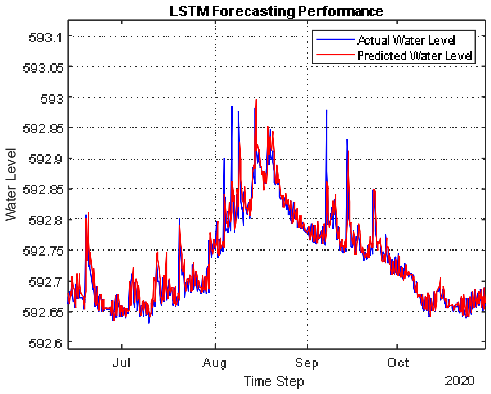

Figure 3. Trainning results showing observed and LSTM-simulated water levels during the 2019-2023 period: zoom to 2020

Following the testing of forecasting performance demonstrate that the LSTM model provides high predictive accuracy, particularly during periods of intense rainfall and complex hydrological responses (Figure 4 and Table 1). Owing to its strong ability to learn and retain nonlinear relationships as well as the inherent time-lag between rainfall and the corresponding water-level response, the LSTM model outperformed other approaches. Therefore, an operational forecasting module was developed for this model.

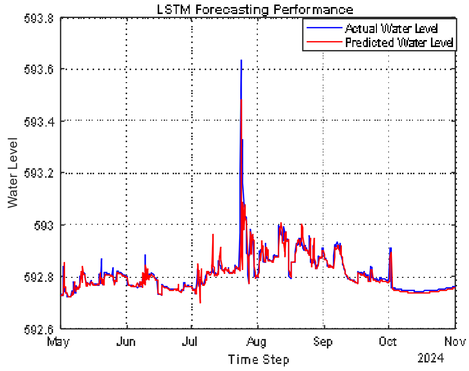

Figure 4. The water level during

Statistical metrics for different forecast lead times are presented in Table 1, showing that the 6-hour forecasts achieved the highest accuracy during training and testing period.

Table 1. Performance metrics (NSE, RMSE, MAE, R^2) of the LSTM model for 6 lead times

Case	RMSE	Bias	AME	neta	P
Calibration / Validation	0.03	0.00	0.02	0.97	98.91
Test Forecasts	0.04	-0.01	0.01	0.86	97.96

The input data for the module include observed water levels at the Hat Lot hydrological station and antecedent rainfall over the basin. These datasets are preprocessed and normalized prior to being used by the model. After training, the LSTM model is capable of forecasting water levels 6, 12, and 24 hours into the future, which is essential for early warning of flash floods and localized inundation. Selecting LSTM as the core model for operational implementation is appropriate given the characteristics of the Nam Pan basin-small catchment size, steep mountainous terrain, and short, concentrated rainfall events that generate rapid hydrological responses.

As shown in the figure below, the LSTM model more closely follows the actual dynamics of observed water levels, particularly during peak-rising periods caused by intense rainfall. The scatter plot further illustrates that LSTM predictions lie much closer to the ideal 1:1 line (Actual = Predicted), whereas the other models exhibit wider dispersion. This reinforces the advantage of LSTM in capturing nonlinear and lagged rainfall–runoff interactions typical of mountainous river basins.

Operational testing of the LSTM model at the Hat Lot station for 6-, 12-, and 24-hour forecast horizons shows consistent and stable performance across all time frames (Table 2). For the 6-hour forecast, the model achieved the lowest RMSE of 0.24 and a correlation coefficient of R^2 = 0.87, indicating high short-term predictive accuracy. The peak-water-level error was −0.95 m, with a peak timing error of 6 hours-an acceptable lag for operational forecasting (Table 2).

Table 2. Performance metrics of the LSTM-based water level forecasting model at Hat Lot during the 2024 flood season (testing phase)

Leading time	RMSE	Bias	AME	MAE	Scp	R2	P	Peak Value error	Peak Time_error
6 hours	0.24	0.038	0.12	2.22	0.250	0.87	90	0.95	0 hrs
12 hours	0.35	0.049	0.16	3.29	0.250	0.67	85	1.74	6hrs
24 hours	0.46	0.082	0.22	3.57	0.250	0.31	77	1.86	6 hrs

Results from the Hat Lot station further show that forecast accuracy decreases as the lead time increases from 6 to 24 hours, as reflected by an RMSE increase from 0.24 to 0.46 and an R^2 reduction from 0.87 to 0.31. MAE and Bias also increase correspondingly, indicating reduced precision for longer forecast windows. Interestingly, however, the peak value error for the 24-hour forecast is comparable to-or in some cases better than-that of the 6-hour forecast. This may be due to the 24-hour model being trained specifically for longer lead times and aligning more closely with the timing of peak events in the dataset. Meanwhile, the 6-hour iterative forecasting mode may accumulate noise over sequential predictions, thereby affecting peak accuracy.

Figure 5 also presents the rainfall-station contribution analysis and the LSTM-based trial forecasts for the 2024 flood season. Overall, the LSTM model demonstrates reliable and effective performance across different forecast horizons, making it particularly suitable for short- and medium-range flood-warning applications in small mountainous catchments such as the Nam Pan basin.

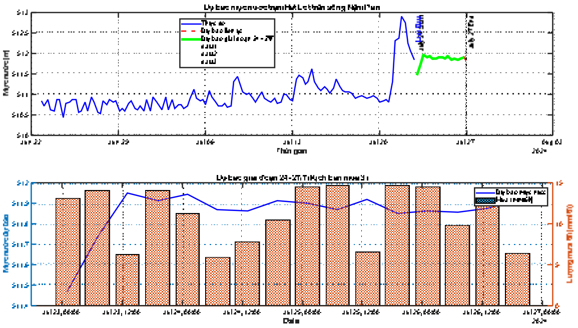

Figure 5. Water level forecasting module at Hat Lot station on Nam Pan River

4.2. Contribution of Cyclic Time Encoding and Lagged Water Levels

The integration of cyclical time-encoded variables (sine and cosine transformations) and lagged water levels significantly improved model performance. The sin–cos time encoding helped the model represent daily and seasonal hydrological cycles consistently, removing artificial discontinuities at temporal boundaries. This enhancement enabled the LSTM model to maintain accuracy even when rainfall time series contained short gaps or inconsistent measurement intervals-conditions commonly found in the Nam Pan basin.

Similarly, the inclusion of multiple water-level lags (6, 12, 18, and 24 hours) provided crucial information on antecedent hydrological states, such as basin storage and soil saturation, which strongly influence runoff generation and flood dynamics. These lagged features enabled the model to better reproduce the timing and magnitude of rapid flood rises, reducing forecast errors during high-flow periods.

Compared with LSTM configurations that excluded these additional features, the enhanced model demonstrated higher NSE and R^2 values, reduced RMSE, and more reliable peak-flow predictions. This confirms that time encoding and lagged hydrological states are essential components for effectively modeling non-homogeneous and discontinuous hydrometeorological time series.

4.3. Discussion

The analysis indicates that the LSTM model is well suited for hydrological forecasting in small mountainous basins such as Nam Pan, where data scarcity, steep terrain, and rapid hydrological responses create challenges for traditional modelling approaches. The model's ability to handle incomplete, temporally inconsistent datasets and to learn nonlinear relationships between rainfall and runoff is particularly valuable in these environments.

However, some limitations remain. Forecast accuracy decreases for lead times beyond 24 hours, primarily due to the basin's highly dynamic rainfall regime and rapid response characteristics. Additionally, the model relies heavily on the availability of accurate rainfall inputs; thus, improvements in rainfall measurement networks-or the integration of radar or satellite rainfall products-could further enhance model performance. Future research may also explore hybrid deep-learning architectures (e.g., CNN–LSTM, Attention-LSTM, Transformers) and the inclusion of additional physical variables such as soil moisture or upstream runoff to enhance forecast reliability.

Overall, the results confirm that LSTM is a powerful tool for handling non-homogeneous, fragmented time series in hydrological forecasting and can provide reliable operational forecasts for flash-flood early warning systems in mountainous regions of Vietnam.

5. Conclusion

This study demonstrates that the Long Short-Term Memory (LSTM) neural network is a suitable and effective approach for forecasting water levels in the Nam Pan River basin, a small mountainous catchment characterized by discontinuous, non-homogeneous hydrometeorological datasets and rapid hydrological responses. By integrating multi-station rainfall, lagged water levels, and cyclically encoded time variables, the model successfully captured the nonlinear rainfall–runoff dynamics and reproduced both the timing and magnitude of major flood events across multiple forecast horizons.

The LSTM model delivered stable performance for 6-, 12-, and 24-hour predictions, achieving high values of NSE, low RMSE, and strong agreement with observed water levels. The test forecasts for the 2024 flood season further confirmed the model's robustness and its ability to adapt to complex hydrological patterns driven by intense rainfall and fast basin response times. These results underscore the practical applicability of LSTM-based forecasting for real-time flood early-warning systems in small mountainous basins where conventional hydrological modeling is constrained by limited data availability.

The operational forecasting module developed in this study provides a functional framework that can be integrated directly into ongoing flood-monitoring systems. Its automated workflow-from data acquisition and preprocessing to prediction and visualization-offers a scalable tool for supporting early-warning operations in the Nam Pan basin.

Future research may focus on enriching the input dataset using radar or satellite rainfall products, incorporating additional hydrological variables such as soil moisture or upstream flows, and exploring advanced deep-learning architectures such as CNN-LSTM, Attention-based models, or Transformer networks to further enhance forecasting accuracy and stability.

References

Vietnam Meteorological and Hydrological Administration (2020–2024), *Rainfall and water level datasets from meteorological–hydrological stations in Son La Province,* Ministry of Natural Resources and Environment.

Vietnam Academy for Water Resources (2022), *Overview report on the Da River basin and hydrological system of the Northwest region*, Ministry of Agriculture and Rural Development.

C. Shen (2018), *"A transdisciplinary review of deep learning research and its relevance for water resources scientists,"* Water Resources Research, vol. 54, no. 11, pp. 8558–8593.

F. Kratzert, D. Klotz, C. Brenner, K. Schulz, and M. Herrnegger (2019), *"Rainfall–runoff modelling using Long Short-Term Memory (LSTM) networks,"* Hydrology and Earth System Sciences, vol. 23, no. 10, pp. 5089–5112.

C. Hu, Q. Wu, H. Li, S. Jian, L. Wang, and J. Chen (2021), *"Deep learning approaches for hydrological time series prediction: A review,"* Hydrological Sciences Journal, vol. 66, no. 8, pp. 1271–1290, 2021.

P. Bai, X. Liu, and S. Hu (2023), *"Integrating CNN–LSTM networks for streamflow prediction in data-scarce basins,"* Journal of Hydrology, vol. 617, p. 128136.

S. Hochreiter and J. Schmidhuber (1997), *"Long short-term memory,"* Neural Computation, vol. 9, no. 8, pp. 1735–1780.

D. P. Kingma and J. Ba (2015), *"Adam: A method for stochastic optimization,"* in Proceedings of the 3rd International Conference on Learning Representations (ICLR).