REGULAR ARTICLE

Enhancing Hydrological Model Daily Streamflow Predictions in Data-Scarce Watercourses by Integrating CNN-LSTM with Physical Processes

Kushagra Kulshreshtha[1,*] ✉, Nimesh Raj[2], Sohini Chowdhury[3], Yatika Gori[4], Anurag Shrivastava[5], A. Kakoli Rao[6], Akhil Sankhyan[7], P. William[8]

[1] *Institute of Business Management, GLA University, 281406 Mathura, Uttar Pradesh, India*
[2] *Centre of Research Impact and Outcome, Chitkara University, 140417 Rajpura, Punjab, India*
[3] *Centre of Research Impact and Outcome, Chitkara University, 140401 Rajpura, Punjab, India*
[4] *Department of Mechanical Engineering, Graphic Era Deemed to be University, Dehradun, India*
[5] *Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences, Chennai, TN, India*
[6] *Lloyd Institute of Engineering & Technology, Greater Noida, India*
[7] *Lloyd Law College, Greater Noida, India*
[8] *Department of Information Technology, Sanjivani College of Engineering, Kopargaon, MH, India*

(Received 25 August 2024; revised manuscript received 14 December 2024; published online 23 December 2024)

Daily streamflow prediction in data-sparse watercourses is significant for efficient water resource management and climate change variations. Especially in areas with sparse observational data, the geographical and temporal complexity of hydrological systems presents a significant challenge for traditional hydrological models. In this research, we offer an innovative approach for improving daily streamflow predictions by integrating the Convolutional Neural Network-Long Short-Term Memory (CNN-LSTM) architecture with physical processes and utilizing the Weather Research and Forecasting (WRF) model for the hydrological process. The objective is to improve the WRF model's capability of capturing the complex interactions between weather conditions and streamflow dynamics by combining this deep learning framework with the physical processes that define the model. For a more precise depiction of the hydrological system, the WRF model, well-known for its high-resolution atmospheric simulations, provides fine-grained meteorological inputs. The performance of the suggested method is evaluated using RMSE (5.14), MAE (6.85), MEDAE (5.97) as well as R^2 (12.05) metrics and they are compared to existing methods. The combination of CNN-LSTM and WRF offers a promising path for improving the accuracy and reliability of hydrological models, which is critical for informed decision-making in water resource management and climate resilience.

Keywords: Daily streamflow, Hydrological model, CNN-LSTM, WRF.

DOI: 10.21272/jnep.16(6).06021 PACS numbers: 92.40. – t, 92.70.Np

1. INTRODUCTION

Water scarcity is a worldwide concern that has been exacerbated by climate change, population increase and rising demand for freshwater supplies. Accurate stream flow estimates play a crucial role in sustainable water resource management, particularly in locations where hydrological data is sparse [1]. Water shortage is an important issue that has wide-ranging impacts on ecosystems, agriculture and human societies across many regions of the globe. The absence of complete hydrological data in places with limited data availability presents a substantial challenge to the efficient management of water resources [2]. The precise forecasting of daily Streamflow becomes of paramount significance to optimize water allocation, develop infrastructure and execute efficient conservation techniques. Hence, it is imperative to create hydrological models that are resilient and designed to address the distinct difficulties encountered in watercourses with limited data availability [3, 4]. The occurrence of data scarcity in the field of hydrology is attributed to various circumstances, including the presence of inadequate monitoring infrastructure, financial limitations and geopolitical considerations [5].

Obtaining precise and dependable data for the calibration and validation of models poses a significant problem in such settings. In these conditions, conventional hydrological models, which depend on comprehensive datasets, need help in delivering precise predictions [6]. The intricate nature of water systems, coupled with the scarcity of available observational data,

[*] Correspondence e-mail: kushagra.kulshrestha@gla.ac.in

Cite this article as: K. Kulshreshtha et al., *J. Nano- Electron. Phys.* 16 No 6, 06021 (2024) https://doi.org/10.21272/jnep.16(6).06021

highlights the need for novel methodologies to improve the precision of predictions. The integration of remote sensing data, machine learning techniques and data assimilation approaches in hydrological modeling has shown significant progress in enhancing predictions in watercourses with limited data availability [7, 8]. Remote sensing plays a crucial role in acquiring vital data pertaining to land cover, soil moisture and precipitation, hence bridging gaps in existing observational datasets. Machine learning algorithms possess the capability to boost the predictive potential of models by identifying intricate patterns.

Furthermore, the utilization of data assimilation techniques facilitates the integration of up-to-date data into models, enhancing their adaptability to dynamic circumstances [9-10].

The following categories were used to group the study's remaining sections: In Section 2, we provide the related works. In Section 3, the method is explained. The performance analysis is provided in Section 4. Section 5 has the conclusion.

2. REALATED WORKS

Study [11] proposed a hybrid modeling framework that integrates the “Soil and Water Assessment Tool (SWAT+) model," which incorporates “glacial hydrological processes (GSWAT+)” with “Gated Recurrent Unit (GRU)” neural networks. The objective of the framework was to enhance the accuracy of the model and introduce a comprehensive approach for predicting both “high and low flows in glacial river” basins. The findings indicated that the hybrid model, namely the combination of the "GRU-GSWAT+" models, had superior performance. Article [12] forecasted the decline of groundwater levels by the utilization of a multidisciplinary methodology. A calibration and validation process was undertaken to ensure the accuracy of the SWAT model in predicting the state of water within the watershed under investigation, with a particular focus on stream flow. The approach yielded encouraging outcomes in the forecast of global water scarcity, demonstrating potential for application in other locations with limited data availability.

Author [13] presented “a fusion model based on few-shot learning," known as the LSTM-prototypical model. The model under consideration was calibrated and applied to forecast monthly runoff in "the Lancang River basin (LRB) and the source region of the Yellow River basin (SRYRB). "The results suggested that the developed model had great potential as a tool for predicting runoff in the two basins under investigation. Paper [14] developed an approach for the modeling of Streamflow in hilly basins that are characterized by limited availability of data. Their technique had shown strong performance in simulating Streamflow in mountainous basins with limited data availability, surpassing the accuracy of previously employed methods. Research [15-16] presented a new “deep learning model” for post-processing Streamflow simulations called "Self-activated and Internal Attention LSTM (SAINA-LSTM)." The model utilizes an “attention-based LSTM cell.” The findings of the comparative assessment indicate that the proposed technique demonstrates the ability to reduce the errors of daily flow.

The outcomes demonstrated that variations in runoff were influenced by temperature in elevated regions. In contrast, precipitation played a more significant role in driving changes in runoff in low-lying areas.

3. METHODOLOGY

3.1 Study Region

For this study, Nepal was selected as the test site due to its location in the middle Himalayan region and limited availability of data. Nepal is located in the center portion of the Himalayan area, spanning a longitudinal distance of around 940 km from east to west and a latitudinal distance of 175 to 260 km from north to south. The northern region of Nepal is characterized by its elevated topography, featuring prominent mountain ranges such as Mt Everest, Kanchenjunga and Annapurna. In contrast, the southern portion of the country is flat and characterized by plains. Approximately 80 % of the total yearly precipitation occurs in the summer monsoon season, with a discernible upward trajectory in the occurrence of intense rainfall events. The majority of streamflow stations are situated in lowland areas, with a comparatively smaller number of stations located in higher-elevation locations. It is essential to acknowledge that among the dataset of available stations in Nepal, there exists a limited number of stations situated at altitudes exceeding 3000 m. While the public streamflow records date back to 1950, it is essential to note that the majority of the monitoring stations were created during the period following the 1980s. The medium-sized irrigation projects employ a classification system to categorize river basins into seven distinct classes, utilizing topographical characteristics as a basis for estimating the monthly average Streamflow.

For river basins, we gathered streamflow information from the DHM between 1985 and 2015 (Table 1). The estimation of the basin drainage region was conducted by outlining the boundaries of the basin using the "Advanced Spaceborne Thermal Emission and Reflection Radiometer (ASTER) digital elevation model (DEM) in the ArcGIS platform.” The “Asian Precipitation Highly Resolved Observational Data Integration towards Evaluation of water resources (APHRODITE)” precipitation product was utilized to calculate the yearly region average amount of precipitation (AP) for all basins within the period. It should be noted that better-resolution rainfall data is required beyond the geographical boundaries of Nepal. The identical product was employed, albeit with a reduced resolution, to cover the regions beyond Nepal in the context of transboundary basins, explicitly referring to the Chinese side. It was posited that no significant disparity exists, on average, among the data of both of these resolutions. The temperature data for the entire research area was obtained using the APHRODITE product with a consistent spatial resolution

Table 1 – Dataset description

No	River Name	Mean Slope (%)	Forest Cover (%)	Annual Precipitation (mm)	Basin area (Sq km)	Station location
1	Rapti	32.08	59.37	1456.47	3551.47	Bangaso-tigaon
2	Babai	20.54	49.31	1311.70	2592.44	Chepang
3	Trishuli	31.83	13.49	725.16	4624.00	Betrawati
4	Marikhola	31.20	63.33	1434.00	1938.00	Nayagaon
5	Sunkoshi	28.48	41.52	1190.00	10162.90	Khurkot

3.2 Hydrological Model

The Weather Research and Forecasting-Hydro (WRF-H) model was initially developed by the "National Center for Atmospheric Research (NCAR)" inside the Research Services Lab as a supplement program for hydrological purposes, intended to be integrated with the WRF model (Figure 1). The purpose of the design was to enhance the accuracy of simulations that depict the hydro process inside the WRF-H model. At present, the WRF-H possesses the inherent attributes of a hydrological modeling system that is distributed and grounded in physical principles. The model configuration employed in this work is the separated mode, which uses a "one-way process" utilizing meteorological triggering data sets. The WRF-H model can simulate the geographic distribution of discharge in three dimensions, encompassing subsurface regions, which represents a significant advancement compared to prior models that were limited to narrow discharge simulations. The physics choices that are accessible contain surface land parameterization, surface overland flow, etc. The WRF-H model can simulate hydrological processes and fluxes of energy at many spatial and temporal resolutions, utilizing the provided choices. In this research, we implemented the activation of surface overland, saturated subsurface and channels routed in the routing module.

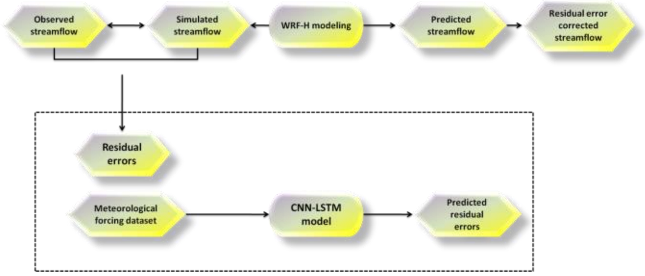

Fig. 1 – Flow of WRF-H

3.3 CNN-LSTM

The integration of Convolutional Neural Networks (CNN) with Long Short-Term Memory (LSTM) networks for improving the accuracy of daily streamflow predictions in hydrological models, especially in regions where there is limited availability of data. The proposed model integrates the advantageous features of CNN and LSTM networks to effectively use spatial patterns derived from physical processes and temporal dependencies inherent in streamflow data.

The initial phase of the architectural design, as displayed in Fig. 2, incorporates the utilization of three identity convolution layers in the CNN module, attempting to enhance the effectiveness of the feature. The utilization of convolutional layers in the model facilitates the extraction of intricate spatial linkages present in the hydrological data. This capability enables the model to identify and comprehend complex patterns that can have significant importance in achieving precise streamflow predictions. The process of feature is essential in the identification of pertinent information from the input data, enabling a thorough depiction of the underlying hydrological processes.

The following phase of the architectural design encompasses the utilization of an LSTM network. This network has been assigned with the analysis of the features that the preceding CNN layers have extracted. Its primary objective is to generate predictions for the subsequent streamflow data points. LSTM networks demonstrate exceptional proficiency in capturing extended dependencies in sequential data, rendering them highly suitable for simulating the underlying temporal dynamics observed in hydrological systems. This model leverages the advantageous characteristics of CNN, which excels in extracting spatial features and LSTM, is proficient in temporal analysis. The integration of several data sources shows potential for enhancing streamflow estimates in areas where data availability is limited.

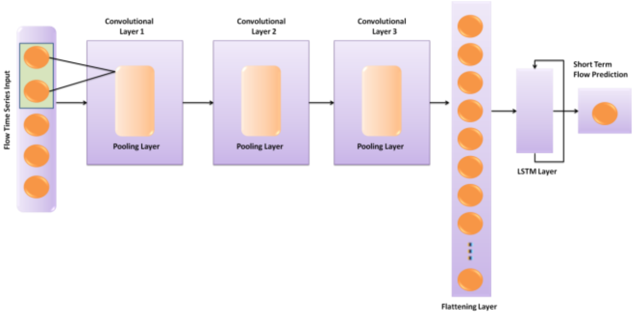

Fig. 2 – Flow of CNN-LSTM

4. RESULT AND DISCUSSION

The proposed approach was implemented in the Python tool (v 3.10). The existing methods are the support vector machine (SVM) [20], XGBoost [20] and linear regression (LR) [20], assessed with the proposed approach in terms of "Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), Median Absolute Error (MEDAE) and R-squared (R^2)".

RMSE signifies enhanced accuracy in the forecast of daily Streamflow, hence indicating the model's proficiency in minimizing errors associated with predictions. The RMSE performance for the suggested and current techniques is shown in Figure 3 yet in comparison to the existing methodologies; SVM, XGBoost as well as LR achieved performance scores of 10.99, 6.28 and 10.67, respectively. In contrast, the proposed method, CNN-LSTM, achieved a performance score of 5.14. This finding provides evidence that the proposed methodology is better than the other approaches.

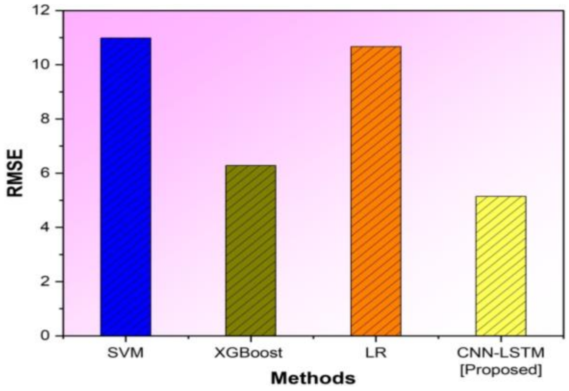

Fig. 3 – RMSE comparison

The MAE serves as a quantitative indicator of the extent of prediction discrepancies, providing valuable information regarding the average magnitude of flaws observed in daily streamflow estimates. The MAE performance of the existing and suggested techniques is shown in Figure 4. SVM, XGBoost as well as LR performed at 9.39, 7.93 and 11.94 when compared to existing techniques; the suggested method (CNN-LSTM) performed at 6.85. This indicates that the proposed approach is superior to the other existing systems.

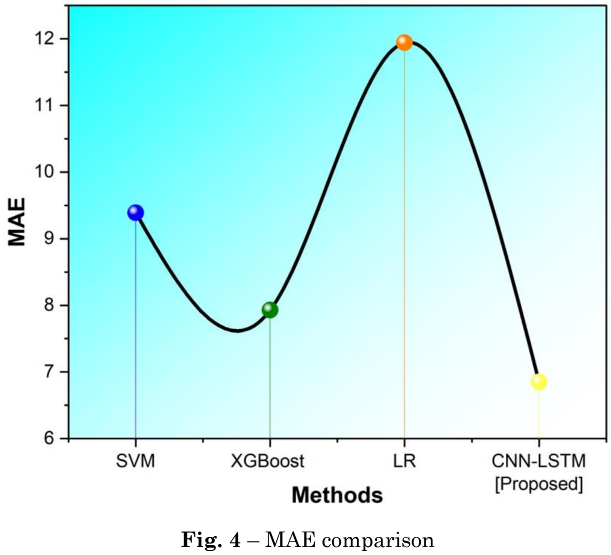

Fig. 4 – MAE comparison

The MEDAE statistic is a measure that represents the median of the overall difference between predicted and actual results. The MEDAE exhibits a higher level of resistance to the influence of outliers compared to the MAE. This characteristic makes MEDAE a reliable measure of the central tendency in the forecast errors associated with daily Streamflow. MEDAE is well-suited for situations when there is limited data available for watercourses. Figure 5 depicts the MEDAE comparison of the existing methods and the proposed approach. When compared to existing algorithms, SVM, XGBoost and LR, each performed at 7.18, 7.12 as well as 12.25, respectively, whereas the suggested method (CNN-LSTM) performed at 5.97. This shows that compared to the alternatives, the proposed solution is better.

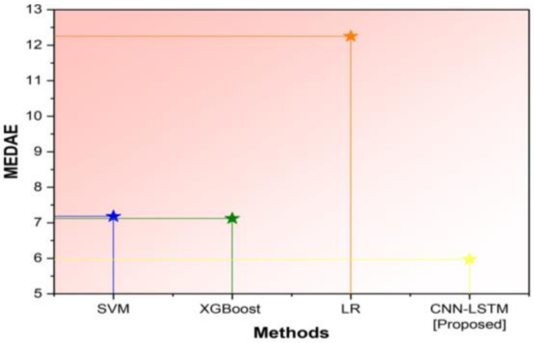

Fig. 5 – MEDAE comparison

4.1 Discussion

Integrating CNN with LSTM networks is an effective alternative to Support Vector Machines (SVM), XGBoost and Linear Regression for daily streamflow predictions. The hybrid approach, CNN-LSTM, capitalizes on the respective advantages of both architectures. The CNN module demonstrates efficacy in capturing spatial dependencies associated with hydrological data, hence facilitating the identification of complicated patterns and spatial correlations in Streamflow. This statement pertains to the constraint of LR in managing intricate relationships. The LSTM component has strong proficiency in capturing temporal relationships, enabling the model to comprehend the sequential characteristics of daily streamflow data.

5. CONCLUSION

In this research, we highlight the significant significance of daily streamflow forecast in the context of water resource management and climate change adaptation, with a specific focus on watercourses that need more data. Hydrological systems are inherently complicated, both geographically and temporally. This is apparent in regions where observational data is scarce. Hence, novel approaches are required to improve the accuracy of predictions. The integration of the CNN in LSTM architecture with the Weather Research and Forecasting (WRF) model is a remarkable development. This study presents a methodology for regionalizing the flow duration curve to forecast daily Streamflow in the data-scarce area of the Central Himalayas. The assessment of the proposed approach utilizing performance indicators, including RMSE (5.14), MAE (6.85), MEDAE (5.97) and R2 (12.05), demonstrates its potential advantage over existing methodologies. The model's performance is dependent on the degree to which the training data mirrors the actual distribution of the streamflow data. If the training data lacks accuracy, the model can encounter difficulties in generalizing its predictions to novel and unknown data. In further research, it could be beneficial to examine the potential of data augmentation techniques and the incorporation of synthetic data generation methods to improve the inclusiveness of training data for the CNN-LSTM model in the context of daily streamflow predictions.

REFERENCES

1. D. Heras, C. Matovelle, *Revista Ambiente & Agua* **16**, e2708 (2021).
2. L. Lin, C. Tang, Q. Liang, Z. Wu, X. Wang, S. Zhao, *J. Hydrol.* **617**, 128758 (2023).
3. X. Zhang, Y. Qi, H. Li, S. Sun, Q. Yin, *Sci. Rep.* **13** No 1, 17168 (2023).
4. M. Musie, S. Sen, P. Srivastava, *J. Hydrol.* **579**, 124168 (2019).
5. V. Kumar, N. Kedam, K.V. Sharma, D.J. Mehta, T. Caloiero, *Water* **15** No 14, 2572 (2023).
6. K.W. Ng, Y.F. Huang, C.H. Koo, K.L. Chong, A. El-Shafie, A.N. Ahmed, *J. Hydrol.* **625**, 130141 (2023).
7. A. Bhusal, U. Parajuli, S. Regmi, A. Kalra, *Hydrology* **9** No 7, 117 (2022).
8. E.K. Siabi, Y.T. Dile, A.T. Kabo-Bah, M. Amo-Boateng, G.K. Anornu, K. Akpoti, C. Vuu, P. Donkor, S.K. Mensah, A.B. Incoom, E.K. Opoku, *Appl. Artificial Intelligence* **36** No 1, 2138130 (2022).
9. Y. Long, W. Chen, C. Jiang, Z. Huang, S. Yan, X. Wen, *J. Hydrol.: Reg. Stud.* **47**, 101420 (2023).
10. H. Dastour, Q.K. Hassan, *Hydrology* **10** No 4, 95 (2023).
11. C. Yang, M. Xu, S. Kang, C. Fu, D. Hu, *J. Hydrol.* **625**, 129990 (2023).
12. A. Rafik, Y.A. Brahim, A. Amazirh, M. Ouarani, B. Bargam, H. Ouatiki, Y. Bouslihim, L. Bouchaou, A. Chehbouni, *J. Hydrol.: Reg. Stud.* **50**, 101569 (2023).
13. M. Yang, Q. Yang, J. Shao, G. Wang, W. Zhang, *Environ. Model. Software* **162**, 105659 (2023).
14. M. Fan, J. Xu, Y. Chen, W. Li, *Sci. Total Environ.* **790**, 148256 (2021).
15. B. Alizadeh, A.G. Bafti, H. Kamangir, Y. Zhang, D.B. Wright, K.J. Franz, *J. Hydrol.* **601**, 126526 (2021).
16. J. Choi, J. Lee, S. Kim, *Ecological Eng.* **182**, 106699 (2022).

Покращення гідрологічної моделі щоденних прогнозів стоку в водотоках з дефіцитом даних шляхом інтеграції CNN-LSTM з фізичними процесами

Kushagra Kulshreshtha[1], Nimesh Raj[2], Sohini Chowdhury[3], Yatika Gori[4], Anurag Shrivastava[5], A. Kakoli Rao[6], Akhil Sankhyan[7], P. William[8]

[1] *Institute of Business Management, GLA University, 281406 Mathura, Uttar Pradesh, India*
[2] *Centre of Research Impact and Outcome, Chitkara University, 140417 Rajpura, Punjab, India*
[3] *Centre of Research Impact and Outcome, Chitkara University, 140401 Rajpura, Punjab, India*
[4] *Department of Mechanical Engineering, Graphic Era Deemed to be University, Dehradun, India*
[5] *Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences, Chennai, TN, India*
[6] *Lloyd Institute of Engineering & Technology, Greater Noida, India*
[7] *Lloyd Law College, Greater Noida, India*
[8] *Department of Information Technology, Sanjivani College of Engineering, Kopargaon, MH, India*

Щоденне прогнозування потоків у водотоках із невеликою кількістю даних має важливе значення для ефективного управління водними ресурсами та зміни клімату. Особливо в районах з рідкісними даними спостережень географічна та часова складність гідрологічних систем становить значну проблему для традиційних гідрологічних моделей. У цьому дослідженні ми пропонуємо інноваційний підхід для покращення щоденних прогнозів стоку шляхом інтеграції архітектури згорткової нейронної мережі з довготривалою короткочасною пам'яттю (CNN-LSTM) із фізичними процесами та використання моделі дослідження та прогнозування погоди (WRF) для гідрологічного процесу. Мета полягає в тому, щоб покращити здатність моделі WRF фіксувати складну взаємодію між погодними умовами та динамікою потоку шляхом поєднання цієї основи глибокого навчання з фізичними процесами, які визначають модель. Для більш точного зображення гідрологічної системи модель WRF, добре відома своїм моделюванням атмосфери з високою роздільною здатністю, надає детальні метеорологічні дані. Ефективність запропонованого методу оцінюється за допомогою показників RMSE (5,14), MAE (6,85), MEDAE (5,97), а також R2 (12,05) і порівнюється з існуючими методами. Поєднання CNN-LSTM і WRF пропонує багатообіцяючий шлях для підвищення точності та надійності гідрологічних моделей, що має вирішальне значення для прийняття обґрунтованих рішень щодо управління водними ресурсами та стійкості до клімату.

Ключові слова: Добовий стік, Гідрологічна модель, CNN-LSTM, WRF.