

One-Day-Ahead Streamflow Forecasting in a Pluvial Mediterranean Catchment of Central Chile Using LightGBM, LSTM, CNN-LSTM and Metaheuristic-Tuned Gradient Boosting

Bahareh Pirzadeh ^{1*}✉

¹ Associate Professor, Department of Civil Engineering–Water, Faculty of Shahid Nikbakht Engineering, University of Sistan and Baluchestan, Zahedan, Iran

*Corresponding Authors: Email : (b.pirzadeh@gmail.com)

Received: 17 December 2025

Accepted: 10 January 2026

Available online: 15 February 2026

Abstract: Reliable daily streamflow forecasts are essential for flood warning and water allocation in the rain-fed catchments of central Chile, which have been affected by a decade-long megadrought. This study forecasts next-day discharge of the Cauquenes River at El Arrayán (622 km², CAMELS-CL station 7336001) for 1979–2019 using random forest (RF), light gradient boosting machine (LightGBM), long short-term memory (LSTM), a convolutional–LSTM network (CNN-LSTM) and two hybrid models in which LightGBM hyperparameters were optimized by Harris hawks optimization (LGBM-HHO) and particle swarm optimization (LGBM-PSO). Inputs were catchment-average precipitation, temperature and potential evapotranspiration from CR2MET and antecedent discharge, and the target was modelled in logarithmic space. Models were trained on 1979–2004, validated on 2005–2011 and tested on the megadrought period 2012–2019. All models outperformed persistence (test NSE = 0.69). The default LightGBM achieved the best test scores (NSE = 0.825, KGE = 0.855, RMSE = 5.22 m³/s), followed by LSTM (NSE = 0.808), whose validation performance was the best of all models. The metaheuristic searches reduced the validation loss only marginally, and the tuned models did not generalize better (LGBM-PSO NSE = 0.798; LGBM-HHO NSE = 0.755) while requiring about 20 min of computation each. The LSTM forecast high flows almost as well as LightGBM, but both deep learning models overestimated low flows during the drought. The results suggest that for daily forecasting with abundant data, well-configured gradient boosting and LSTM models are sufficient, and the added value of metaheuristic tuning is limited.

Keywords: Streamflow forecasting, CAMELS-CL, LightGBM, LSTM, Harris hawks optimization

1. Introduction

Short-term streamflow forecasts support flood warning, reservoir operation, irrigation scheduling and hydropower planning. Physically based rainfall–runoff models remain the backbone of operational hydrology, but data-driven models have become increasingly competitive during the last two decades (Yaseen et al., 2015). In particular, deep learning has shown that a single model trained on long records can match or exceed calibrated conceptual models (Kratzert et al., 2018, 2019). This success has also prompted a debate on the role of hydrological understanding in the age of machine learning (Nearing et al., 2021).

Among the data-driven approaches, gradient boosted decision trees and recurrent neural networks are currently the most widely used for daily streamflow prediction. Szczepanek (2022) compared XGBoost, LightGBM and CatBoost for a mountainous catchment and found that the differences between algorithms were smaller than the differences caused by hyperparameters. Cheng et al. (2020) showed that LSTM networks outperform feed-forward networks for long lead-time daily forecasting. Hybrid approaches that tune the hyperparameters of such models with metaheuristic algorithms, for example PSO-LSTM (Kilinc, 2022), are frequently reported to improve accuracy. However, the computational cost of these searches is rarely weighed

against their benefit, and the evaluation periods often do not contain the hydrological conditions that matter most for water management.

Central Chile provides a demanding test case. Its Mediterranean climate concentrates precipitation in a few winter months, and since 2010 the region has experienced an unprecedented sequence of dry years, the so-called megadrought (Garreaud et al., 2017, 2020). Catchments with a pluvial regime respond rapidly to frontal storms, produce very high peak flows, and recede to very low base flows in summer. Streamflow therefore varies over four orders of magnitude, and models trained on wetter decades must forecast under drier conditions. The CAMELS-CL dataset (Alvarez-Garretón et al., 2018) provides consistent, catchment-averaged forcings and streamflow for such basins and has made rigorous benchmarking possible.

The objectives of this study were (i) to develop one-day-ahead streamflow forecasting models for the pluvial Cauquenes River catchment using RF, LightGBM, LSTM and CNN-LSTM; (ii) to evaluate whether tuning LightGBM with Harris hawks optimization (HHO; Heidari et al., 2019) and particle swarm optimization (PSO; Kennedy and Eberhart, 1995) improves its accuracy; and (iii) to assess all models against persistence during the 2012–2019 megadrought, with

attention to peak flows, low flows and computational cost.

2. Materials and Methods

2.1. Study Catchment

The Cauquenes River at El Arrayán (station code BNA 7336001; 36.02° S, 72.38° W) drains a catchment of 622.1 km² in the Maule Region of central Chile (Fig. 1). Elevations range from 134 to 736 m a.s.l. within the Coastal Range, so snow plays a negligible role, and the flow regime is pluvial. The climate is Mediterranean. Mean annual precipitation during 1979–2019 was 959 mm, more than 80% of which fell between May and September, while mean annual Hargreaves potential evapotranspiration was 1157 mm (Fig. 3). Mean annual precipitation fell from 1002 mm in 1979–2009 to 823 mm in 2010–2019, reflecting the megadrought.

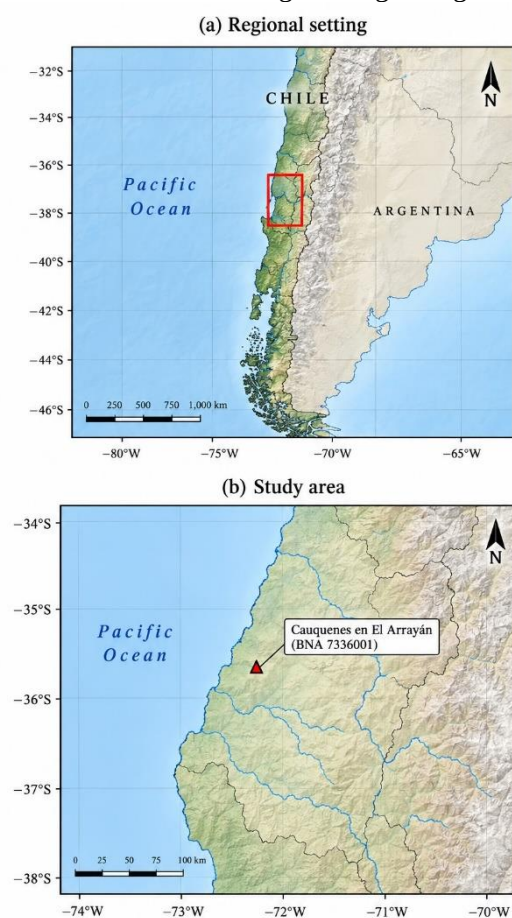


Figure 1. Location of the Cauquenes en El Arrayán streamflow station in central Chile

2.2. Data

Daily catchment-average precipitation (P), maximum and minimum air temperature ($T_{\max}$, $T_{\min}$) and potential evapotranspiration (PET) were taken from the CAMELS-CL dataset (Alvarez-Garreton et al., 2018). The meteorological variables are based on the CR2MET gridded product, and PET was computed with the Hargreaves–Samani equation (Hargreaves and Samani, 1985). Observed daily discharge (Q , m³/s) was measured

by the Chilean Dirección General de Aguas. The series for 1 January 1979 – 31 December 2019 (14,975 days) were obtained from the hydroTSM R package repository (Zambrano-Bigiarini, github.com/hzambran/hydroTSM), which redistributes the CAMELS-CL data for this catchment. Discharge was missing on 434 days (2.9%); these days were excluded from training and evaluation and were not interpolated. Descriptive statistics are listed in Table 1 and the series are shown in Fig. 2. Discharge is extremely skewed (skewness = 12.0). The median flow was 1.17 m³/s, whereas the maximum of 853 m³/s was recorded on 12 July 2006.

Table 1. Descriptive statistics of the daily variables (1979–2019)

Variable	Min	Max	Mean	SD	Skew	Kurt
P (mm/d)	0.00	111.63	2.62	7.69	4.48	27.12
$T_{\max}$ (°C)	8.16	37.16	19.19	4.87	0.17	-1.08
$T_{\min}$ (°C)	-2.93	18.87	7.47	3.52	0.12	-0.81
PET (mm/d)	0.79	7.75	3.17	1.60	0.22	-1.28
Q (m ³ /s)	0.01	853.00	7.95	26.77	12.02	225.26

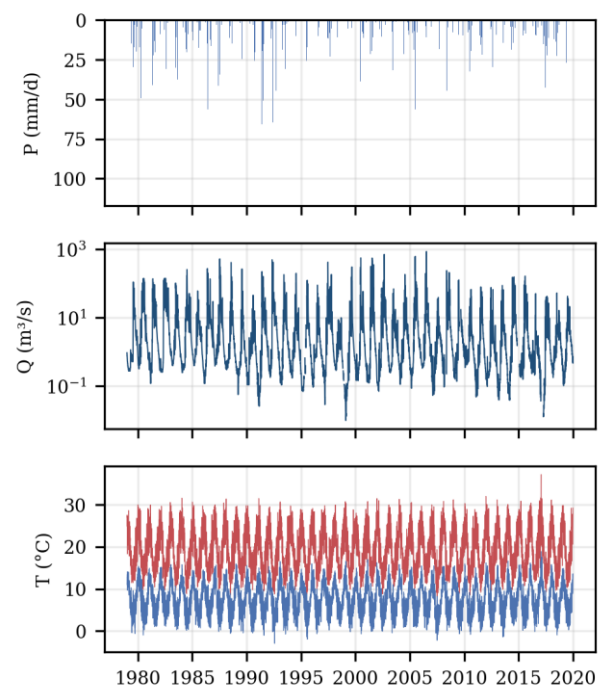


Figure 2. Daily precipitation, discharge (log scale) and maximum/minimum temperature of the Cauquenes catchment

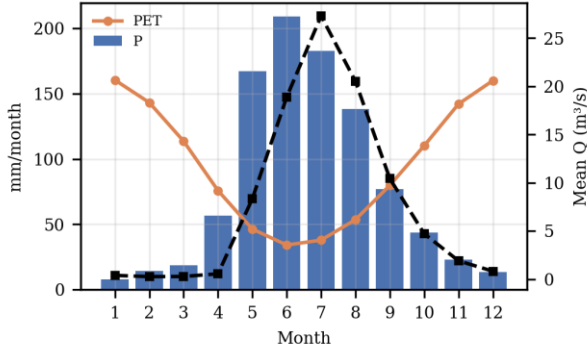


Figure 3. Mean monthly precipitation, potential evapotranspiration and discharge

2.3. Forecasting Setup and Inputs

The target was the mean discharge of day $t+1$, forecast at the end of day t . The tabular models (RF and LightGBM variants) used 17 inputs: precipitation of day $t+1$ (treated as a perfect quantitative precipitation forecast, a common assumption in hindcast experiments), precipitation of days t , $t-1$ and $t-2$, accumulated precipitation over 3, 7, 30 and 90 days, log-transformed discharge of days t , $t-1$ and $t-2$ and its 7-day mean, $T_{\max}$ and $T_{\min}$ of day t , the 7-day PET sum, and the sine and cosine of the day of year. The sequence models (LSTM and CNN-LSTM) received the previous 30 days of $[P, P(t+1), T_{\max}, T_{\min}, PET, \log Q, \sin, \cos]$. Because discharge spans four orders of magnitude, all models were trained on $z = \ln(1 + Q)$ and the forecasts were back-transformed before evaluation.

The record was split chronologically by target date into training (1979–2004, $n = 9109$), validation (2005–2011, $n = 2397$) and test (2012–2019, $n = 2742$) sets. The test period lies entirely within the megadrought, which makes it a demanding out-of-sample test. All inputs were standardized with training statistics. Persistence ($Q(t+1) = Q(t)$) was used as the benchmark, and all models were evaluated on the same days.

2.4. Tree-Based Ensemble Models

RF (Breiman, 2001) averaged 300 regression trees grown on bootstrap samples with a minimum leaf size of 3. **LightGBM** (Ke et al., 2017) is a gradient boosting framework that grows trees leaf-wise using histogram-based split finding. At iteration m the ensemble is updated as

$$F_m(x) = F_{m-1}(x) + \eta \cdot h_m(x) \quad (1)$$

where h_m is a regression tree fitted to the negative gradient of the squared loss and η is the learning rate. The default configuration used 400 trees, 31 leaves and $\eta = 0.05$.

2.5. Deep Learning Models

The **LSTM** network (Hochreiter and Schmidhuber, 1997) had one layer with 48 units, followed by dropout (0.2), a 24-unit ReLU dense layer and a linear output. The **CNN-LSTM** added a causal one-dimensional convolution (48 filters, kernel size 3, ReLU) and max pooling (size 2) before the LSTM layer. The convolution

extracts short-term patterns such as storm sequences before the recurrent layer models their temporal dependence (LeCun et al., 2015). Both networks were trained with Adam (learning rate 0.001, batch size 128, mean squared error) for up to 150 epochs with early stopping (patience 15). Three networks with different random seeds were averaged for each architecture.

2.6. Hybrid LightGBM Models

In **HHO** (Heidari et al., 2019), hawks explore the search space by perching at random locations and then exploit it with soft or hard besiege strategies. The choice of strategy depends on the escaping energy of the prey,

$$E = 2E_0(1 - t/T) \quad (2)$$

where $E_0 \in (-1, 1)$ is the initial energy, t is the iteration and T the maximum number of iterations. For $|E| \geq 1$ the hawks explore; otherwise they besiege the best solution, optionally with Lévy-flight dives. **PSO** (Kennedy and Eberhart, 1995) moves particles with velocities that combine inertia, attraction to their personal best and attraction to the global best position:

$$v_i^{t+1} = w v_i^t + c_1 r_1 (p_i - x_i^t) + c_2 r_2 (g - x_i^t) \quad (3)$$

Seven LightGBM hyperparameters were optimized: number of trees (50–800), number of leaves (4–64), learning rate ($10^{-2.5}$ – $10^{-0.5}$), row and column subsampling ratios (0.5–1.0), minimum samples per leaf (5–60) and L2 regularization (10^{-2} – $10^{1.5}$). The fitness was the RMSE of standardized $\ln(1 + Q)$ on the validation period. Each optimizer used 12 agents and 25 iterations (MEALPY; Van Thieu and Mirjalili, 2023), and the final model was trained on the training period with the best configuration.

2.7. Evaluation

Performance was evaluated in the original units with R^2 , RMSE, MAE, NSE (Nash and Sutcliffe, 1970), KGE (Gupta et al., 2009) and Willmott's index:

$$NSE = 1 - \Sigma(Q_o - Q_s)^2 / \Sigma(Q_o - \bar{Q}_o)^2 \quad (4)$$

$$KGE = 1 - \sqrt{[(r - 1)^2 + (\alpha - 1)^2 + (\beta - 1)^2]} \quad (5)$$

where r is the correlation, α the ratio of standard deviations and β the ratio of means. $NSE > 0.75$ is usually considered very good for daily models (Moriassi et al., 2007). In addition, the RMSE of the highest 5% of flows (peak RMSE), the percent bias of the lowest 30% of flows (low-flow bias) and flow duration curves (Yilmaz et al., 2008) were used to evaluate the extremes.

3. Results

3.1. Hyperparameter Optimization

Table 2. LightGBM hyperparameters (default and optimized) and computation time

Model	Tree s	Leave s	η	Ro w	Col	Min leaf	λ	Time (min)
LightGBM	400	31	0.0500	1.00	1.00	20	0.0	0.05
LGBM-HHO	782	27	0.0557	0.79	1.00	46	31.6	23.4

LGBM-PSO	792	45	0.017 8	0.68	0.8 9	17	16. 3	18.8
----------	-----	----	------------	------	----------	----	----------	------

Both optimizers selected large ensembles (about 790 trees) with strong L2 regularization (Table 2). PSO chose more leaves and a slower learning rate, whereas HHO chose fewer leaves with a learning rate similar to the default. The validation fitness improved only slightly during the search, from 0.1139 to 0.1125 for HHO and from 0.1143 to 0.1108 for PSO (Fig. 4). The two searches took 23 and 19 min on a two-core CPU, compared with 3 s for a single LightGBM fit and 1.4 min for the three-member LSTM ensemble.

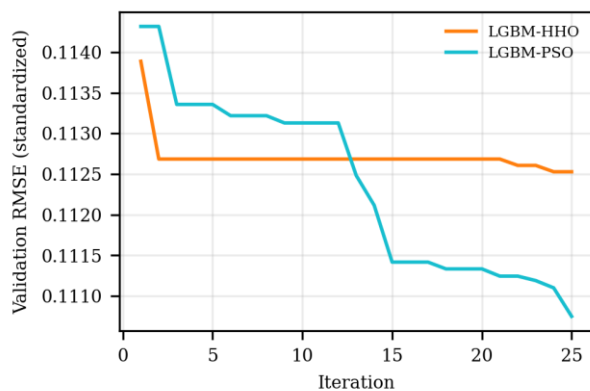


Figure 4. Convergence of HHO and PSO (validation RMSE of standardized $\ln(1+Q)$)

3.2. Overall Forecast Accuracy

Table 3. Performance of the one-day-ahead forecasts (RMSE and MAE in m^3/s)

Model	NSE (train)	NSE (val)	R^2 (test)	RMSE (test)	MAE (test)	NSE (test)	KGE (test)
Persistence	0.440	0.293	0.716	6.92	1.66	0.693	0.846
RF	0.880	0.681	0.802	5.56	1.12	0.802	0.836
LightGBM	0.965	0.683	0.825	5.22	1.06	0.825	0.855
LSTM	0.831	0.835	0.811	5.47	1.05	0.808	0.813
CNN-LSTM	0.824	0.736	0.786	5.78	1.18	0.786	0.827
LGBM-HHO	0.923	0.718	0.766	6.18	1.17	0.755	0.874
LGBM-PSO	0.894	0.700	0.799	5.61	1.09	0.798	0.866

All learning models outperformed persistence in all three periods (Table 3). In the test period, persistence reached $\text{NSE} = 0.693$, whereas the learning models achieved NSE values between 0.755 and 0.825. The default LightGBM was the most accurate model ($R^2 = 0.825$, $\text{RMSE} = 5.22 \text{ m}^3/\text{s}$, $\text{NSE} = 0.825$, $\text{KGE} = 0.855$), followed by the LSTM ($\text{NSE} = 0.808$), RF (0.802), LGBM-PSO (0.798), CNN-LSTM (0.786) and LGBM-HHO (0.755). The differences in RMSE between the first four models were within $0.4 \text{ m}^3/\text{s}$.

The ranking in the validation period (2005–2011), which contains the largest flood of the record ($853 \text{ m}^3/\text{s}$

in July 2006), was very different. There the LSTM was clearly the best model ($\text{NSE} = 0.835$), whereas the tree-based models reached only 0.68–0.72. Tree ensembles cannot predict values beyond the range of their training targets, which limits their skill for record floods. The LSTM, which uses a smooth learned mapping, extrapolated more gracefully. The tuned LightGBM models fitted the training period better than RF but not better than the default LightGBM (training $\text{NSE} = 0.965$), and they had higher test errors. The small improvement in validation fitness therefore did not translate into better forecasts.

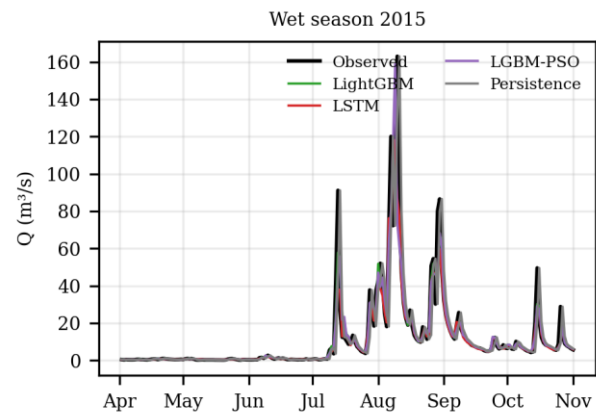


Figure 5. Observed and forecast hydrographs for the 2015 wet season (test period)

Fig. 5 shows the 2015 wet season, which contains the largest flood of the test period ($163 \text{ m}^3/\text{s}$ on 10 August 2015, following $130 \text{ m}^3/\text{s}$ on 9 August). All models reproduced the timing of most storm responses, while persistence lagged by one day, as expected. The two-day flood was difficult for all models. On 9 August the tuned LightGBM models and the CNN-LSTM overestimated the discharge ($159\text{--}193 \text{ m}^3/\text{s}$), and on 10 August all models underestimated the peak ($74\text{--}87 \text{ m}^3/\text{s}$). This shows the limited information content of daily inputs for the timing of flood crests. The scatter plots (Fig. 6) show a tight clustering around the 1:1 line over the full range of flows on a logarithmic scale, with larger relative scatter for the lowest flows.

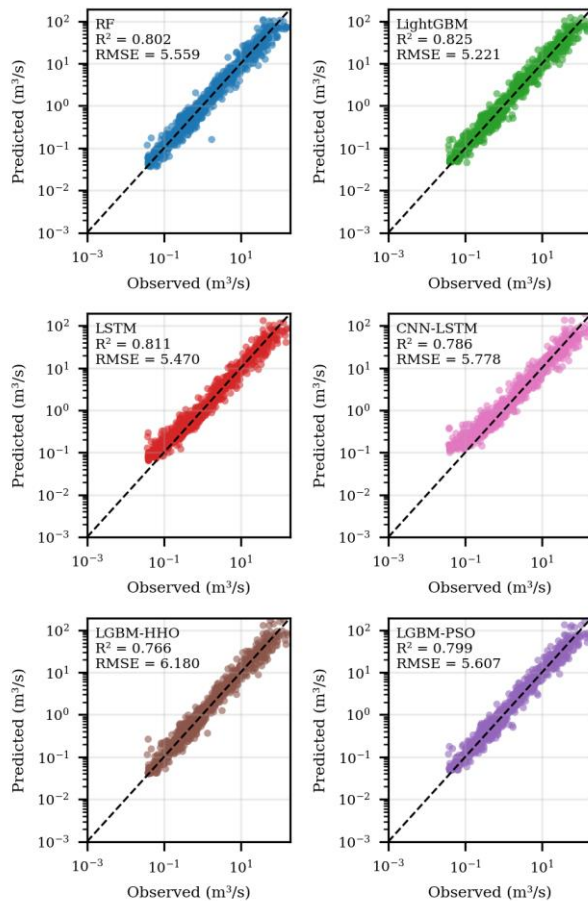


Figure 6. Observed versus forecast discharge in the test period (log scales)

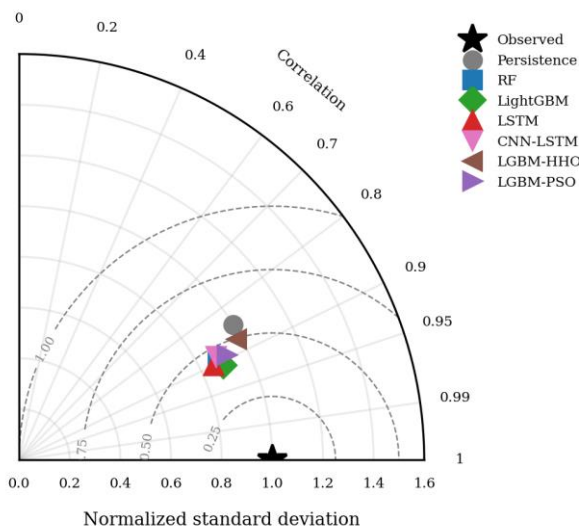


Figure 7. Taylor diagram of the test-period forecasts

3.3. Peak and Low Flows

Table 4. Performance for high and low flows in the test period

Model	Peak RMSE (m³/s)	Low-flow bias (%)	Max. forecast (m³/s)
Persistence	29.7	+5.1	163
RF	24.3	+6.5	109

LightGBM	22.9	+7.9	123
LSTM	24.0	+25.9	131
CNN-LSTM	25.3	+44.4	193
LGBM-HHO	27.1	+6.9	177
LGBM-PSO	24.5	+8.1	159

Peak flows: observed Q above the 95th percentile (21.0 m³/s); low flows: below the 30th percentile (0.47 m³/s). Observed maximum: 163 m³/s.

Table 4 and the flow duration curves (Fig. 8) summarize the extremes. LightGBM had the lowest peak-flow error (22.9 m³/s), followed by LSTM (24.0 m³/s) and RF (24.3 m³/s). All learning models were better than persistence (29.7 m³/s). For low flows, the tree-based models had small positive biases of 6–8%. The LSTM (+26%) and especially the CNN-LSTM (+44%) clearly overestimated the summer recession flows. The flow duration curves confirm that the deep learning models deviate mainly in the low-flow segment (exceedance > 70%). Because the networks were trained mostly on the wetter pre-2005 period, they seem to have learned a base-flow level that was too high for the megadrought years. Tree ensembles, which anchor their forecasts to the observed discharge of the previous day, were less affected.

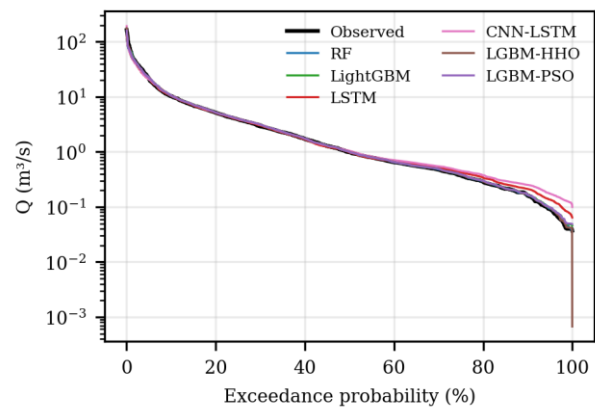


Figure 8. Flow duration curves of observed and forecast discharge in the test period

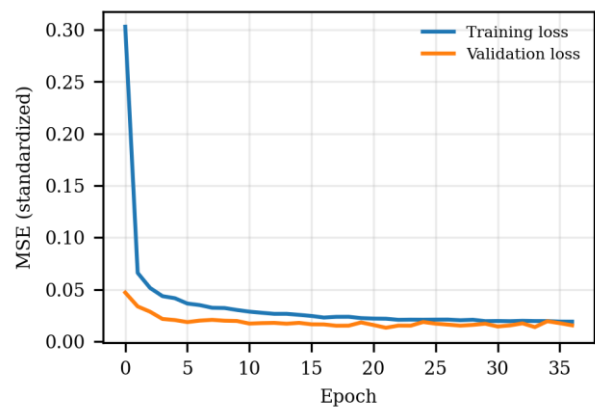


Figure 9. Training and validation loss of the LSTM (first ensemble member)

3.4. Importance of Inputs

The permutation importance of LightGBM (Fig. 10) was dominated by the current discharge ($Q(t)$), followed at a much lower level by same-day precipitation ($P(t)$), $Q(t-1)$ and $P(t-1)$. The long-term precipitation accumulation ($P90$) was also among the top five inputs, which reflects the role of catchment wetness in the runoff response. Next-day precipitation had almost no importance. This indicates that, at the daily resolution of the data, the discharge response of this 622 km² catchment is mainly driven by precipitation already observed on the issue day.

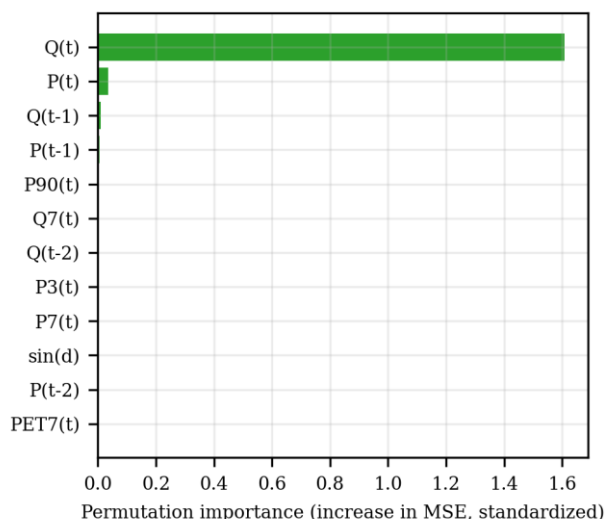


Figure 10. Permutation importance of the LightGBM inputs (test period)

4. Discussion

With more than 9,000 daily training samples, a default LightGBM configuration was already close to the best achievable accuracy, and neither HHO nor PSO improved it in the test period. This agrees with Szczepanek (2022), who noted that gradient boosting models with reasonable defaults perform similarly and that tuning mainly avoids poor configurations. In our case, both optimizers were guided by a validation period containing the largest flood of the record. Minimizing the error in that period pushed the models towards configurations that were not optimal for the dry test period. The cost–benefit ratio was clearly unfavourable: about 20 minutes of search per model produced no gain, whereas the default model took seconds. Hybrid models should therefore be evaluated against well-configured standalone models and under changing hydroclimatic conditions, not only against poorly configured baselines.

The LSTM was the most robust model across periods. It had the best validation performance, including the 2006 flood, and the second-best test performance. Its main weakness was a positive bias in low flows during the megadrought. This bias could be reduced by training on multiple catchments (Kratzert et al., 2019), by using loss functions that weight low flows, or by including drought-period data in training. The

CNN-LSTM did not improve on the plain LSTM. The additional convolutional layer seems to offer little benefit when the input sequence is short and the relevant signals (recent precipitation and discharge) are already explicit.

The study has several limitations. Observed next-day precipitation was used as a perfect forecast, so operational skill would be lower. In practice, however, next-day precipitation contributed very little in this catchment. Only one catchment was analysed, and results for snow-influenced or larger basins may differ. Finally, the models were deterministic. Probabilistic forecasts and multi-day lead times would be needed for operational flood warning.

5. Conclusions

Daily one-day-ahead streamflow forecasts were developed for the pluvial Cauquenes River catchment in central Chile with RF, LightGBM, LSTM, CNN-LSTM and two metaheuristic-tuned LightGBM models. The main conclusions are: (1) all learning models clearly outperformed persistence, and the default LightGBM gave the best test-period accuracy ($NSE = 0.825$, $KGE = 0.855$); (2) the LSTM was the most robust model across hydroclimatic conditions and best predicted the record flood of the validation period, but it overestimated low flows during the megadrought; (3) HHO and PSO tuning improved the validation fitness only marginally, did not improve the test performance and increased computation time by several orders of magnitude; and (4) the current discharge and same-day precipitation carried almost all the predictive information. For operational use, an ensemble of LightGBM and LSTM forecasts is recommended, together with training data that cover both wet and drought conditions.

6. Data Availability

CAMELS-CL data (Alvarez-Garretton et al., 2018; camels.cr2.cl) for station 7336001 were obtained from the hydroTSM repository (github.com/hzambran/hydroTSM). The code used for the analysis is available from the authors on request.

7. References

- Alvarez-Garretton, C., Mendoza, P. A., Boisier, J. P., Addor, N., Galleguillos, M., Zambrano-Bigiarini, M., Lara, A., Puelma, C., Cortes, G., Garreaud, R., McPhee, J., & Ayala, A. (2018). The CAMELS-CL dataset: Catchment attributes and meteorology for large sample studies – Chile dataset. *Hydrology and Earth System Sciences*, 22(11), 5817–5846.
- Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32.
- Cheng, M., et al. (2020). Long lead-time daily and monthly streamflow forecasting using machine learning methods. *Journal of Hydrology*, 590, 125376.

- Garreaud, R. D., Alvarez-Garreton, C., Barichivich, J., Boisier, J. P., Christie, D., Galleguillos, M., LeQuesne, C., McPhee, J., & Zambrano-Bigiarini, M. (2017). The 2010–2015 megadrought in central Chile: Impacts on regional hydroclimate and vegetation. *Hydrology and Earth System Sciences*, 21(12), 6307–6327.
- Garreaud, R. D., Boisier, J. P., Rondanelli, R., Montecinos, A., Sepúlveda, H. H., & Veloso-Aguila, D. (2020). The Central Chile Mega Drought (2010–2018): A climate dynamics perspective. *International Journal of Climatology*, 40(1), 421–439.
- Gupta, H. V., Kling, H., Yilmaz, K. K., & Martinez, G. F. (2009). Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling. *Journal of Hydrology*, 377(1–2), 80–91.
- Hargreaves, G. H., & Samani, Z. A. (1985). Reference crop evapotranspiration from temperature. *Applied Engineering in Agriculture*, 1(2), 96–99.
- Heidari, A. A., Mirjalili, S., Faris, H., Aljarah, I., Mafarja, M., & Chen, H. (2019). Harris hawks optimization: Algorithm and applications. *Future Generation Computer Systems*, 97, 849–872.
- Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, 9(8), 1735–1780.
- Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T.-Y. (2017). LightGBM: A highly efficient gradient boosting decision tree. *Advances in Neural Information Processing Systems*, 30, 3146–3154.
- Kennedy, J., & Eberhart, R. (1995). Particle swarm optimization. In *Proceedings of ICNN'95 – International Conference on Neural Networks* (Vol. 4, pp. 1942–1948). IEEE.
- Kilinc, H. C. (2022). Daily streamflow forecasting based on the hybrid particle swarm optimization and long short-term memory model in the Orontes Basin. *Water*, 14.
- Kratzert, F., Klotz, D., Brenner, C., Schulz, K., & Herrnegger, M. (2018). Rainfall–runoff modelling using Long Short-Term Memory (LSTM) networks. *Hydrology and Earth System Sciences*, 22(11), 6005–6022.
- Kratzert, F., Klotz, D., Shalev, G., Klambauer, G., Hochreiter, S., & Nearing, G. (2019). Towards learning universal, regional, and local hydrological behaviors via machine learning applied to large-sample datasets. *Hydrology and Earth System Sciences*, 23(12), 5089–5110.
- LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, 521, 436–444.
- Moriasi, D. N., Arnold, J. G., Van Liew, M. W., Bingner, R. L., Harmel, R. D., & Veith, T. L. (2007). Model evaluation guidelines for systematic quantification of accuracy in watershed simulations. *Transactions of the ASABE*, 50(3), 885–900.
- Nash, J. E., & Sutcliffe, J. V. (1970). River flow forecasting through conceptual models part I — A discussion of principles. *Journal of Hydrology*, 10(3), 282–290.
- Nearing, G. S., Kratzert, F., Sampson, A. K., Pelissier, C. S., Klotz, D., Frame, J. M., Prieto, C., & Gupta, H. V. (2021). What role does hydrological science play in the age of machine learning? *Water Resources Research*, 57(3), e2020WR028091.
- Szczepanek, R. (2022). Daily streamflow forecasting in mountainous catchment using XGBoost, LightGBM and CatBoost. *Hydrology*, 9(12), 226.
- Van Thieu, N., & Mirjalili, S. (2023). MEALPY: An open-source library for latest meta-heuristic algorithms in Python. *Journal of Systems Architecture*, 139, 102871.
- Yaseen, Z. M., El-Shafie, A., Jaafar, O., Afan, H. A., & Sayl, K. N. (2015). Artificial intelligence based models for stream-flow forecasting: 2000–2015. *Journal of Hydrology*, 530, 829–844.
- Yilmaz, K. K., Gupta, H. V., & Wagener, T. (2008). A process-based diagnostic approach to model evaluation: Application to the NWS distributed hydrologic model. *Water Resources Research*, 44(9), W09417.