
MCANN: A Mixture Clustering-Based Attention Neural Network for Multivariate Time Series Forecasting

David C. Anastasiu

Computer Science and Engineering
Santa Clara University
Santa Clara, CA 95053
danastasiu@scu.edu

Yanhong Li

Computer Science and Engineering
Santa Clara University
Santa Clara, CA 95053
yli20@scu.edu

Abstract

Forecasting time series with sparse extreme values remains a challenging problem in fields such as hydrology, energy, and finance. Traditional attention-based models often suffer from diluted focus and entangled feature representations when trained on skewed distributions. We propose MCANN (Mixture Clustering-Based Attention Neural Network), a novel framework that leverages statistical distribution separation and mixture-based attention. MCANN improves forecasting accuracy on long-horizon time series with extreme fluctuations. It dynamically partitions input features into distinct statistical clusters and assigns attention weights within and across these components. This design allows the model to learn disjoint feature subspaces that enhance representation disentanglement. Our experiments on real-world reservoir inflow datasets demonstrate that MCANN consistently outperforms state-of-the-art models.

1 Introduction

Long-term time series forecasting plays a critical role in decision-making for natural resource planning and disaster mitigation [4, 9, 2]. In hydrological domains, reservoir water level often exhibit heavy-tailed, sparse, and skewed characteristics due to seasonal changes and rare extreme events. Conventional machine learning models [1, 12, 3] struggle with long-term prediction and complex distribution inputs. Classic deep learning models [10, 11, 13], including LSTM and Transformer variants, typically assume a homogeneous latent space, which limits their ability to model distributional heterogeneity. To address this, we propose MCANN, a clustering-guided neural network architecture designed to separate latent dynamics into statistically distinct regions. Our design is inspired by the observation that time series features often arise from multiple generative processes, each requiring separate attention strategies. Unlike standard attention mechanisms, which tend to conflate dissimilar patterns, MCANN introduces a clustering-aware attention layer that guides attention computation based on distributional similarity.

2 Methodology

MC-ANN is designed to capture the intrinsic structure and distributional diversity in univariate reservoir water level time series that exhibit high variance and sporadic extreme values. By internalizing the role of statistical separation previously handled via an external classifier, MC-ANN automatically learns to adaptively focus on and weight different regimes within the time series using a mixture modeling framework and a clustering-aware attention mechanism. Our key contributions in this work are as follows:

- **A unified end-to-end architecture:** Our model integrates trend modeling, regime separation, and prediction into a single framework, enabling joint optimization for improved generalization and robustness.
- **Mixture clustering-enhanced sampling mechanism:** MC-ANN uses a soft clustering-based sampling strategy to highlight underrepresented yet critical segments (e.g., spikes, transitions), improving gradient flow and focus without manual rules.
- **Clustering-aware attention loss:** The model integrates a distinctive attention-based Loss, blending point-wise and segment-wise clustering approaches, which has been validated in ablation studies to enhance accuracy by upwards of 20%.
- **Deployed in real-world applications:** MC-ANN has demonstrated strong utility in practice. We integrated the model into *FlowView*, a web-based hydrological forecasting platform developed in collaboration with the Santa Clara Valley Water District. The system delivers daily updated, three-day-ahead forecasts of reservoir water levels, supporting operational decision-making in water distribution, flood prevention, and hydropower scheduling.

3 Experiments

We compared MCANN with state-of-the-art models, including NEC+ [7], DNN-U [5], A-LSTM [6], Informer [15], iTransformer [8], FEDformer [16], NLinear, and DLinear [14]. The experiment results are shown in Tables 1. Overall, MCANN consistently delivers superior performance compared to all baselines, achieving improvements ranging from approximately 10% up to nearly 45% in rolling prediction settings. Importantly, this performance gain is attained without compromising the accuracy of short-term (3-day) forecasts.

Table 1: Effectiveness comparisons against state-of-the-art methods on rolling 8-hour prediction.

Datasets	Metric	MC-ANN	NEC+	iTransformer	Informer	NLinear	DLinear	NBEATS	DNN-U	A-LSTM
Almaden	RMSE	7.412	10.580	65.683	211.241	16.199	23.009	23.229	<u>9.676</u>	18.040
	MAPE	0.002	0.002	0.016	0.204	0.005	0.006	0.009	<u>0.004</u>	0.016
Coyote	RMSE	45.373	<u>64.590</u>	755.083	7437.162	385.546	346.678	159.024	117.000	1282.913
	MAPE	0.002	0.002	0.020	0.653	0.013	0.012	0.006	<u>0.004</u>	0.126
Lexington	RMSE	255.739	<u>303.510</u>	1600.916	9565.245	645.141	798.529	468.829	318.024	660.354
	MAPE	0.003	0.003	0.048	0.773	0.023	0.024	0.011	<u>0.005</u>	0.068
Stevens Creek	RMSE	7.382	14.977	48.256	714.468	27.380	48.692	34.998	<u>13.363</u>	117.497
	MAPE	0.002	0.002	0.0136	0.589	<u>0.005</u>	0.012	0.008	0.006	0.104
Vasona	RMSE	5.137	<u>5.775</u>	15.269	19.580	7.045	12.544	10.572	11.370	23.587
	MAPE	0.004	0.004	0.020	0.028	<u>0.006</u>	0.013	0.011	0.016	0.049

Table 2: Effectiveness comparisons against state-of-the-art methods on 3-days prediction.

Datasets	Metric	MC-ANN	NEC+	iTransformer	Informer	NLinear	DLinear	NBEATS	DNN-U	A-LSTM
Almaden	RMSE	53.539	58.117	59.272	217.641	60.516	64.596	64.764	58.648	<u>57.649</u>
	MAPE	0.014	0.014	<u>0.015</u>	0.162	0.017	0.021	0.018	0.017	0.021
Coyote	RMSE	<u>433.571</u>	466.276	608.228	8507.417	631.056	730.057	535.886	417.24	1338.622
	MAPE	0.011	0.011	0.015	0.619	0.019	0.022	0.013	<u>0.012</u>	0.128
Lexington	RMSE	774.209	<u>794.842</u>	926.294	11878.486	1019.081	1082.898	931.356	832.329	1050.5
	MAPE	<u>0.015</u>	0.014	0.020	0.930	0.030	0.033	0.023	0.018	0.078
Stevens Creek	RMSE	71.303	91.810	93.042	1052.549	90.084	99.598	89.918	<u>76.090</u>	156.591
	MAPE	<u>0.013</u>	0.011	0.015	0.888	0.014	0.024	0.015	0.016	0.128
Vasona	RMSE	18.474	20.893	18.264	22.051	20.157	<u>20.021</u>	21.405	20.683	32.245
	MAPE	0.018	<u>0.020</u>	<u>0.020</u>	0.031	<u>0.020</u>	0.024	0.023	0.027	0.062

4 Conclusion and future work

MCANN presents a new direction for integrating statistical awareness into attention-based sequence models. Future extensions include: (1) applying this approach to multimodal time series such as climate-satellite fusion and energy market signals, (2) adapting MCANN to classification problems like multimodal emotion recognition from video, and (3) incorporating external features such as location embeddings, calendar events, and textual metadata to further improve predictive power in sparse-regime forecasting.

Acknowledgments

GPU hardware for our research was provided by NVIDIA and Supermicro. Computing resources were also made possible by the Santa Clara University HPC Center.

References

- [1] George.E.P. Box and Gwilym M. Jenkins. *Time Series Analysis: Forecasting and Control*. Holden-Day, , 1976.
- [2] Yuning Chen, Kang Yang, Zhiyu An, Brady Holder, Luke Paloutzian, Khaled M. Bali, and Wan Du. Marlp: Time-series forecasting control for agricultural managed aquifer recharge. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '24, page 4862–4872, New York, NY, USA, 2024. Association for Computing Machinery.
- [3] Jianan Han, Xiao-Ping Zhang, and Fang Wang. Gaussian process regression stochastic volatility model for financial time series. *IEEE Journal of Selected Topics in Signal Processing*, 10(6):1015–1028, 2016.
- [4] Pradeep Hewage, Marcello Trovati, Ella Pereira, and Ardhendu Behera. Deep learning-based effective fine-grained weather forecasting model. *Pattern Analysis and Applications*, 24(1):343–366, 2021.
- [5] Sebastian C. Ibañez, Carlo Vincienzo G. Dajac, Marissa P. Liponhay, Erika Fille T. Legara, Jon Michael H. Esteban, and Christopher P. Monterola. Forecasting reservoir water levels using deep neural networks: A case study of angat dam in the philippines. *Water*, 14(1), 2022.
- [6] Yan Le, Changwei Chen, Ting Hang, and Youchuan Hu. A stream prediction model based on attention-lstm. *Earth Science Informatics*, 14:1–11, 06 2021.
- [7] Yanhong Li, Jack Xu, and David C Anastasiu. An extreme-adaptive time series prediction model based on probability-enhanced lstm neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 8684–8691, 2023.
- [8] Yong Liu, Tengge Hu, Haoran Zhang, Haixu Wu, Shiyu Wang, Lintao Ma, and Mingsheng Long. itransformer: Inverted transformers are effective for time series forecasting. *arXiv preprint arXiv:2310.06625*, 2023.
- [9] Saloni Mohan, Sahitya Mullapudi, Sudheer Sammeta, Parag Vijayvergia, and David C. Anastasiu. Stock price prediction using news sentiment analysis. In *2019 IEEE Fourth International Conference on Big Data Computing Service and Applications (BigDataService)*, BDS 2019, pages 205–208, , April 2019. IEEE.
- [10] Boris N Oreshkin, Dmitri Carpov, Nicolas Chapados, and Yoshua Bengio. N-beats: Neural basis expansion analysis for interpretable time series forecasting. *arXiv preprint arXiv:1905.10437*, 2019.
- [11] David Salinas, Valentin Flunkert, Jan Gasthaus, and Tim Januschowski. Deepar: Probabilistic forecasting with autoregressive recurrent networks. *International Journal of Forecasting*, 36(3):1181–1191, 2020.
- [12] Zhi-Yu Wang, Jun Qiu, and Fang-Fang Li. Hybrid models combining emd/eemd and arima for long-term streamflow forecasting. *Water*, 10(7), 2018.
- [13] Haixu Wu, Tengge Hu, Yong Liu, Hang Zhou, Jianmin Wang, and Mingsheng Long. Timesnet: Temporal 2d-variation modeling for general time series analysis. *arXiv preprint arXiv:2210.02186*, 2022.
- [14] Ailing Zeng, Muxi Chen, Lei Zhang, and Qiang Xu. Are transformers effective for time series forecasting? *arXiv preprint arXiv:2205.13504*, 2022.
- [15] Haoyi Zhou, Shanghang Zhang, Jieqi Peng, Shuai Zhang, Jianxin Li, Hui Xiong, and Wancai Zhang. Informer: Beyond efficient transformer for long sequence time-series forecasting. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 11106–11115, 2021.
- [16] Tian Zhou, Ziqing Ma, Qingsong Wen, Xue Wang, Liang Sun, and Rong Jin. Fedformer: Frequency enhanced decomposed transformer for long-term series forecasting. In *International Conference on Machine Learning*, pages 27268–27286. PMLR, 2022.