DROUGHT FORECASTING USING A HYBRID NEURAL ARCHITECTURE FOR INTEGRATING TIME SERIES AND STATIC DATA

Julian Agudelo [*]
AgroParisTech, Agrial

Vincent Guigue [†]
AgroParisTech

Cristina Manfredotti
AgroParisTech

Hadrien Piot
Agrial

ABSTRACT

Reliable forecasting is critical for early warning systems and adaptive drought management. Most previous deep learning approaches focus solely on homogeneous regions and rely on single-structured data. This paper presents a hybrid neural architecture that integrates time series and static data, achieving state-of-the-art performance on the DroughtED dataset. Our results illustrate the potential of designing neural models for the treatment of heterogeneous data in climate related tasks and present reliable prediction of USDM categories, an expert-informed drought metric. Furthermore, this work validates the potential of DroughtED for enabling location-agnostic training of deep learning models. All the necessary code to reproduce the experiments is available at `https://github.com/JulAgu/drought_forecasting_HM`.

1 INTRODUCTION

Drought is a natural phenomenon characterized by a prolonged period of below-average precipitation, which causes significant hydrological imbalances that negatively impact land resources (Reichhuber et al., 2023). Droughts are typically classified into meteorological, agricultural, hydrological and socioeconomic types (Humphreys, 1931; Dracup et al., 1980; Wilhite et al., 2007), reflecting their broad impacts on different systems. As global warming intensifies, drought events' frequency, duration, and severity increase, exacerbating vulnerabilities. Given the multifaceted nature of droughts and their effects across various spatial and temporal scales, reliable forecasting is crucial for early warning systems and adaptive resources management. Effective predictions help mitigate impacts on water supply, agriculture, ecosystems, and communities (Aghelpour et al., 2020).

Several indices are commonly used to assess drought conditions. Some well-known examples are the Palmer Drought Severity Index (PDSI) (Palmer, 1968), the Standardized Precipitation Index (SPI) (Mckee et al., 1993), and the Standardized Precipitation Evapotranspiration Index (SPEI) (Vicente-Serrano et al., 2010). These indices depend on meteorological variables to quantify deviations from climatic norms. Alternatively, the U.S. Drought Monitor (USDM) categories provide a complete evaluation by integrating hydrological, climatic and weather data with experts' insights, capturing a holistic view of drought impacts (Svoboda et al., 2002). We use the USDM categories as our target variable for this study due to their comprehensive nature and integration into DroughtED (Minixhofer et al., 2021).

Various authors have used machine learning to predict drought indices (Nandgude et al., 2023). Traditionally, manual feature extraction has been used to feed classical machine learning algorithms (Gaikwad et al., 2015). However, in recent years, there has been a significant shift toward deep learning approaches, which employ representation learning to automatically extract meaningful features from data (Zheng & Casari, 2018). A considerable number of prior studies utilizing deep learning techniques for drought prediction have predominantly focused on homogeneous regions (Dikshit et al., 2022), and primarily employed uniform structured data, as images (Chaudhari et al., 2023) or time series (Lalika et al., 2024; Vijaya Shetty et al., 2023) alone.

[*] julian.agudeloacosta@agroparistech.fr, j.agudelo@agrial.com
[†] vincent.guigue@agroparistech.fr

This paper uses the DroughtED dataset introduced by Minixhofer et al. (2021) and presents a novel modeling approach for drought forecasting with heterogeneous data. The contributions of this research are threefold. (1) We introduce a neural architecture integrating time series and static data through FFNNs, LSTMs, categorical embeddings, and an attention mechanism. We benchmark our model against the dataset's baselines. (2) We conduct an ablation study to assess the contribution of each component within the proposed model. (3) We apply visualization techniques over latent states to perform model introspection.

2 DATA

DroughtED is a large-scale dataset designed to forecast drought conditions in the United States by integrating spatial and temporal features (Minixhofer et al., 2021).It includes historical meteorological time series, soil physical characteristics, and historical drought intensity information at the county level. The meteorological data are sourced from The NASA Prediction Of Worlwide Energy Resources (NASA POWER) project (Zhang et al., 2009), the soil properties are derived from the Harmonized World Soil Database (Nachtergaele et al., 2008) and the drought intensity evaluations are taken from the USDM (Svoboda et al., 2002).

Drought data are ordinal indicators measured locally. These indicators are then reduced to continuous average values and aggregated at the county level. The target values to be predicted are 6 continuous values at county level, corresponding to 6 consecutive weeks. We have data $\mathbf{Y} \in \mathbb{R}^{N \times 6}$ where N corresponds to the number of pairs (county, timestamp) noted (c, t) in the following. To predict these targets, we have static descriptors $\mathbf{S} \in \mathbb{R}^{C \times f}$, where f is the number of features describing the soil physical properties, and C the number of counties. Note that the descriptors will be divided into categorical $\mathbf{s_d} \in \mathbb{R}^{f_d}$ and numerical $\mathbf{s_n} \in \mathbb{R}^{f_n}$ features below. The meteorological data is represented as multivariate time series grouped in a tensor $\mathbf{X} \in \mathbb{R}^{C \times P \times M}$: for each county, we observe $M = 20$ different measurements over several years corresponding to P days.

We use all the descriptors from (Minixhofer et al., 2021): each local target to predict $\mathbf{y} \in \mathbb{R}^6$ (6 weeks following t in county c) is associated with static descriptors $\mathbf{s}_d$ and $\mathbf{s}_n$ as well as a multivariate time series $\mathbf{x} \in \mathbb{R}^{T \times M'}$. The period extracted from $\mathbf{X}$ corresponds to $T = 180$ days before the timestamp t for the county c, while we take the M available measurements plus the M measurements corresponding to the previous year over the same days to enable the model to build comparative features, thus $M' = 2M$.

Counties are indexed using the FIPS (NIST, 1990) identifier. We use the train, validation and test splits available in Kaggle[1].

3 THE PROPOSED MODEL

Despite the flexibility of neural network libraries, handling data with heterogeneous structures is still an open topic in deep learning (Guo et al., 2019; Kamm et al., 2023). We propose a hybrid neural model that combines four modules: Long-Short-Term Memory Recurrent Neural Networks (LSTM), Feedforward Neural Networks (FFNN), embedding layers and an attention mechanism. Figure 1 shows the schema of the proposed model.

The categorical features $\mathbf{s_d}$ are passed through the embedding layer E, which maps them to dense vectors $\mathbf{e} \in \mathbb{R}^z$. Then, the dense representations of all categorical features are concatenated and passed through a FFNN $\mathcal{F}$ that reduces its dimensionality, resulting in a vector $\mathbf{e}' = \mathcal{F}(E(\mathbf{s_d})) \in \mathbb{R}^{z'}$, where $z' < z$. In parallel, the multivariate time series $\mathbf{x}$ is fed into the LSTM, yielding hidden states at each time step t: $\{\mathbf{h}_t\}_{t=1:T} \in \mathbb{R}^{h \times T}$. The hidden states are further processed by the attention mechanism (detailed in Appendix A) to produce a context vector $\tilde{\mathbf{h}} \in \mathbb{R}^h$. The numerical features $\mathbf{s_n} \in \mathbb{R}^{f_n}$ remain unchanged.

The final representation is obtained by concatenating the context vector $\tilde{\mathbf{h}}$, the last hidden state $\mathbf{h}_T$, the continuous features $\mathbf{s_n}$, and the latent representation of the categorical features $\mathbf{e}'$.

[1] https://www.kaggle.com/datasets/cdminix/us-drought-meteorological-data

Table 1: Weekly results on the test set for the proposed model and the Minixhofer et al. (2021) LSTM model.

	Week 1		**Week 2**		**Week 3**		**Week 4**		**Week 5**		**Week 6**	
Model	MAE	F_1	MAE	F_1	MAE	F_1	MAE	F_1	MAE	F_1	MAE	F_1
LSTM	0.150	81.6	0.229	71.6	0.286	64.5	0.347	57.4	0.394	54.2	0.432	49.6
HM	**0.126**	**82.2**	**0.169**	**74.7**	**0.209**	**68.6**	**0.244**	**64.0**	**0.269**	**58.6**	**0.294**	**51.0**

The resulting vector is $\mathbf{x}' = [\tilde{\mathbf{h}}, \mathbf{h}_T, \mathbf{e}', \mathbf{s}_n] \in \mathbb{R}^{2h+z'+f_n}$. This concatenated vector is then passed through an MLP $\mathcal{M}$, which outputs the prediction $\hat{y} = \mathcal{M}(\mathbf{x}') \in \mathbb{R}^6$.

To the best of our knowledge, no prior effort has employed a neural architecture combining these elements to predict USDM drought categories.

4 Experiments and Results

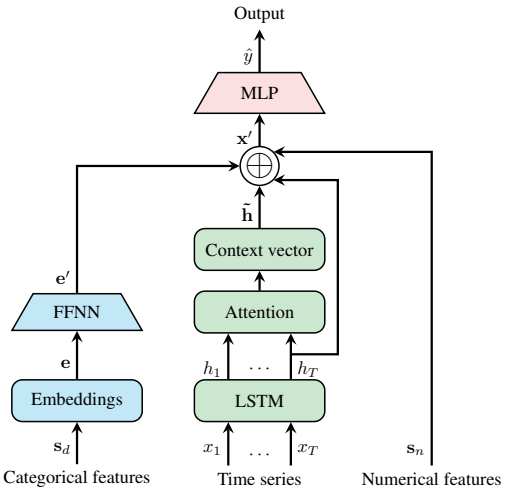


Figure 1: Schematic view of the proposed model.

Given the outlined framework, we conducted some experiments to investigate the following questions. a) How does the proposed hybrid architecture perform compared to the best baseline model established by Minixhofer et al. (2021) regarding predictive performance and generalization? b) What is the relative contribution of each architectural component to the overall performance? c) Does the proposed model sustain superior performance under a location-agnostic training compared to location-specific training? d) How do the attentional mechanism and latent embeddings interact to shape the learned representations within the model?

Hyperparameter optimization was conducted using Bayesian optimization via Optuna (Akiba et al., 2019), with the AdamW optimizer (Loshchilov & Hutter, 2019) and a cyclical learning rate schedule (Smith, 2017). We replicated the baseline LSTM model following the publicly available code and guidelines provided by Minixhofer et al. (2021), and we present results from our implementation to ensure consistency. The Transformer baseline model was excluded from the analysis since the LSTM demonstrated superior performance. This decision aligns with the findings of Minixhofer et al. (2021) and is further supported by Zeng et al. (2023), which highlights the limitations of transformer-based approaches in time series forecasting.

a) Predictive performance and generalization. After training the architecture with the optimal set of hyperparameters, we observed that the proposed Hybrid Model (HM) consistently yields better MAE and macro F_1 scores over the weeks, as shown in Table 1. Over the entire test set, the model shows relative improvements of 30% in the MAE, 9% in the F_1 and 7% in the multi-class weighted ROC-AUC score compared to the baseline LSTM. To estimate the expected prediction error more accurately, we performed 5-fold cross-validation (Appendix D) and conducted a paired t-test to evaluate whether HM significantly outperforms the LSTM. The results show a significant improvement in MAE, RMSE, and F_1, with p-values of 0.03, 0.04, and 0.02, respectively.

b) Ablation study. We evaluated the model under three ablation variables (Table 2). The best performances are achieved when the attention mechanism was included. Additionally, the ablation study founds that most of the knowledge is derived from meteorological data, which aligns with previous findings (Minixhofer et al., 2021).

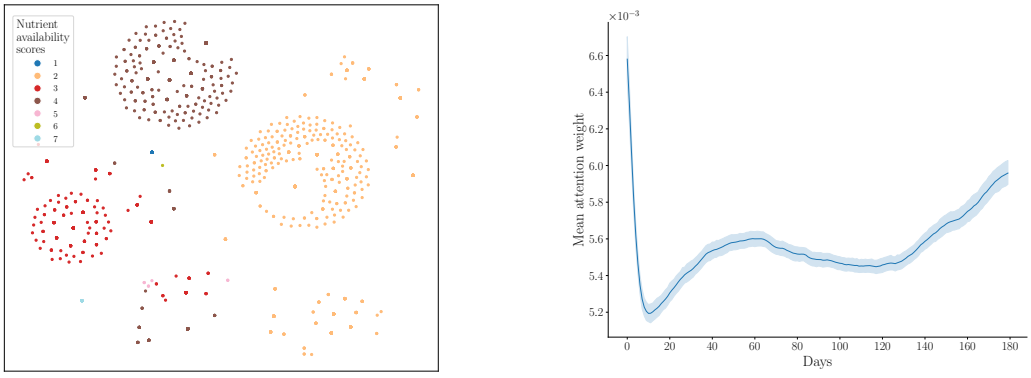
Table 2: Results of the ablation study for the proposed model (HM).

Ablation settings					
Static features	Time series	Attention mech.	MAE	RMSE	F1
✓	✓	✓	**0.217**	**0.377**	**66.3**
	✓	✓	0.267	0.419	56.2
	✓		0.271	0.420	56.6
✓	✓		0.280	0.427	57.1
✓			0.755	0.920	21.2

c) Location-agnostic vs location-specific training. As in the experiment carried out by Minixhofer et al. (2021) for the baseline LSTM, we consider the same 3 states obtained at random (Iowa, Montana and Oklahoma) and trained HM on each state alone and on all training data (Appendix F). The model trained on data from all counties —location-agnostic— demonstrated an average relative improvement of 9.3%. This indicates that when using HM, location-agnostic training outperforms location-specific training. In comparison, Minixhofer et al. (2021) reported an average relative improvement of 4.6% for the baseline LSTM.

d) Model introspection. We conduct a qualitative analysis of the model's intermediate representations. By using t-SNE dimensionality reduction (with a perplexity = 100 and 1000 iterations) (van der Maaten & Hinton, 2008), we examined how observations cluster according to each categorical feature. Our findings indicate that the embeddings closely align with the categories of the "Nutrient availability" feature (see Figure 2a). We consider this a favorable outcome, as previous work have found that droughts significantly influence the presence and accessibility of nutrients in the soil (He & Dijkstra, 2014; Bista et al., 2018).

In relation to the attentional weights, we processed the test set through the model. Then, we plotted the average attention weight for each day, along with a 95% confidence interval (Figure 2b). Attention is primarily concentrated on the first 10 days, with additional focus on the last 30 days of the look-back window.



(a) t-SNE colored by "Nutrient availability".

(b) Mean attention weights on the test set.

Figure 2: t-SNE over the embeddings and mean attention weights curve.

5 Conclusion and Future Work

We present a hybrid neural architecture and validate its effectiveness through empirical evaluation on the DroughtED dataset, achieving state-of-the-art performance in forecasting USDM drought categories.

Future improvements to the proposed architecture will prioritize refining the attention mechanism through two strategies: (1) calibrating attention weights sharpness via a learnable softmax temperature parameter and (2) leveraging expert-annotated labels to supervise attention training through

an auxiliary loss, mirroring the expert-guided process behind USDM categories. Beyond immediate performance gains, the architecture's modular design creates a flexible framework for extreme weather events forecasting and other tasks depending on static and temporal data interaction.

REFERENCES

Pouya Aghelpour, Hadigheh Bahrami-Pichaghchi, and Ozgur Kisi. Comparison of three different bio-inspired algorithms to improve ability of neuro fuzzy approach in prediction of agricultural drought, based on three different indexes. *Computers and Electronics in Agriculture*, 170:105279, 2020. ISSN 0168-1699. doi: https://doi.org/10.1016/j.compag.2020.105279. URL https://www.sciencedirect.com/science/article/pii/S0168169919324354.

Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2019.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *CoRR*, abs/1409.0473, 2014. URL https://api.semanticscholar.org/CorpusID:11212020.

Deepesh R. Bista, Scott A. Heckathorn, Dileepa M. Jayawardena, Sasmita Mishra, and Jennifer K. Boldt. Effects of drought on nutrient uptake and the levels of nutrient-uptake proteins in roots of drought-sensitive and -tolerant grasses. *Plants*, 7(2), 2018. ISSN 2223-7747. doi: 10.3390/plants7020028. URL https://www.mdpi.com/2223-7747/7/2/28.

Shilpa Chaudhari, Vandana Sardar, and Prosenjit Ghosh. Drought classification and prediction with satellite image-based indices using variants of deep learning models. *International Journal of Information Technology*, 15(7):3463–3472, Oct 2023. ISSN 2511-2112. doi: 10.1007/s41870-023-01379-4. URL https://doi.org/10.1007/s41870-023-01379-4.

Abhirup Dikshit, Biswajeet Pradhan, and M. Santosh. Artificial neural networks in drought prediction in the 21st century–a scientometric analysis. *Applied Soft Computing*, 114:108080, 2022. ISSN 1568-4946. doi: https://doi.org/10.1016/j.asoc.2021.108080. URL https://www.sciencedirect.com/science/article/pii/S1568494621009819.

John A. Dracup, Kil Seong Lee, and Edwin G. Paulson Jr. On the definition of droughts. *Water Resources Research*, 16(2):297–302, 1980. doi: https://doi.org/10.1029/WR016i002p00297. URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/WR016i002p00297.

Nikhil Gaikwad, Gaurav Chavan, Hemant Palivela, and Preeja Ravishankar Babu. Study of various feature extraction and selection techniques for drought prediction in precision agriculture. *Advances in Intelligent Systems and Computing*, 340:611–619, 01 2015. doi: 10.1007/978-81-322-2247-7_62.

Wenzhong Guo, Jianwen Wang, and Shiping Wang. Deep multimodal representation learning: A survey. *IEEE Access*, 7:63373–63394, 2019. doi: 10.1109/ACCESS.2019.2916887.

Mingzhu He and Feike A. Dijkstra. Drought effect on plant nitrogen and phosphorus: a meta-analysis. *New Phytologist*, 204(4):924–931, 2014. doi: https://doi.org/10.1111/nph.12952. URL https://nph.onlinelibrary.wiley.com/doi/abs/10.1111/nph.12952.

W. J. Humphreys. How droughts occur. *Bulletin of the American Meteorological Society*, 12 (1):18–22, 1931. ISSN 00030007, 15200477. URL http://www.jstor.org/stable/26262429.

Simon Kamm, Sushma Sri Veekati, Timo Müller, Nasser Jazdi, and Michael Weyrich. A survey on machine learning based analysis of heterogeneous data in industrial automation. *Computers in Industry*, 149:103930, 2023. ISSN 0166-3615. doi: https://doi.org/10.1016/j.compind.2023.103930. URL https://www.sciencedirect.com/science/article/pii/S0166361523000805.

Christossy Lalika, Aziz Ul Haq Mujahid, Mturi James, and Makarius C.S. Lalika. Machine learning algorithms for the prediction of drought conditions in the wami river sub-catchment, tanzania. *Journal of Hydrology: Regional Studies*, 53:101794, 2024. ISSN 2214-5818. doi: https://doi.org/10.1016/j.ejrh.2024.101794. URL https://www.sciencedirect.com/science/article/pii/S2214581824001423.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

T. B. Mckee, Nolan J. Doesken, and John R. Kleist. The relationship of drought frequency and duration to time scales. In *8th Conference on Applied Climatology*, 1993. URL https://api.semanticscholar.org/CorpusID:129950974.

Christoph D Minixhofer, Mark Swan, Calum McMeekin, and Pavlos Andreadis. Droughted: A dataset and methodology for drought forecasting spanning multiple climate zones. In *ICML 2021 Workshop on Tackling Climate Change with Machine Learning*, 2021. URL https://www.climatechange.ai/papers/icml2021/22.

F.O. Nachtergaele, H. van Velthuizen, L. Verelst, N.H. Batjes, J.A. Dijkshoorn, V.W.P. van Engelen, G. Fischer, A. Jones, L. Montanarella, M. Petri, S. Prieler, E. Teixeira, D. Wilberg, and X. Shi. *Harmonized World Soil Database (version 1.0)*. FAO, Italy, 2008. Food and Agric Organization of the UN (FAO); International Inst. for Applied Systems Analysis (IIASA); ISRIC - World Soil Information; Inst of Soil Science-Chinese Acad of Sciences (ISS-CAS); EC-Joint Research Centre (JRC).

Neeta Nandgude, T. P. Singh, Sachin Nandgude, and Mukesh Tiwari. Drought prediction: A comprehensive review of different drought prediction models and adopted technologies. *Sustainability*, 15(15), 2023. ISSN 2071-1050. doi: 10.3390/su151511684. URL https://www.mdpi.com/2071-1050/15/15/11684.

National Institute of Standards and Technology NIST. Federal information processing standards publication: counties and equivalent entities of the united states, its possession, and associated areas. *reports of the National Technical Information Service*, 1990. URL https://nvlpubs.nist.gov/nistpubs/Legacy/FIPS/fipspub6-4.pdf.

Wayne C. Palmer. Keeping track of crop moisture conditions, nationwide: The new crop moisture index. *Weatherwise*, 21(4):156–161, 1968. doi: 10.1080/00431672.1968.9932814. URL https://doi.org/10.1080/00431672.1968.9932814.

A. Reichhuber, M. Svoboda, C. King-Okumu, A. Mirzabaev, S.M. Vicente-Serrano, R. Srinivasan, K. Ehlert, X. Jia, A. Karnib, R. Lal, B. Mislimshoeva, N.H. Ravindranath, A. López Santos, L. Schipper, R. Stefanski, A. Vuković, and H. Zhang. Multiscale approaches for the assessment and monitoring of social and ecological resilience to drought. a report of the science-policy interface. Technical report, United Nations Convention to Combat Desertification, Bonn, Germany., 2023.

Leslie N. Smith. Cyclical learning rates for training neural networks. In *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 464–472, 2017. doi: 10.1109/WACV.2017.58.

Mark Svoboda, Doug LeComte, Mike Hayes, Richard Heim, Karin Gleason, Jim Angel, Brad Rippey, Rich Tinker, Mike Palecki, David Stooksbury, David Miskus, and Scott Stephens. The drought monitor. *Bulletin of the American Meteorological Society*, 83(8):1181 – 1190, 2002. doi: 10.1175/1520-0477-83.8.1181. URL https://journals.ametsoc.org/view/journals/bams/83/8/1520-0477-83_8_1181.xml.

Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(86):2579–2605, 2008. URL http://jmlr.org/papers/v9/vandermaaten08a.html.

Sergio M. Vicente-Serrano, Santiago Beguería, and Juan I. López-Moreno. A multiscalar drought index sensitive to global warming: The standardized precipitation evapotranspiration index. *Journal of Climate*, 23(7):1696 – 1718, 2010. doi: 10.1175/2009JCLI2909.1. URL https://journals.ametsoc.org/view/journals/clim/23/7/2009jcli2909.1.xml.

S Vijaya Shetty, M S Bharath, R Chandan, V Divya, and S Sushma Nayak. Prediction of drought - a machine learning approach using time series data. In *2023 International Conference on Applied Intelligence and Sustainable Computing (ICAISC)*, pp. 1–6, 2023. doi: 10.1109/ICAISC58445.2023.10199898.

Donald A. Wilhite, Mark D. Svoboda, and Michael J. Hayes. Understanding the complex impacts of drought: A key to enhancing drought mitigation and preparedness. *Water Resources Management*, 21(5):763–774, May 2007. ISSN 1573-1650. doi: 10.1007/s11269-006-9076-5. URL https://doi.org/10.1007/s11269-006-9076-5.

Ailing Zeng, Muxi Chen, Lei Zhang, and Qiang Xu. Are transformers effective for time series forecasting? In *Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence and Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence and Thirteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'23/IAAI'23/EAAI'23. AAAI Press, 2023. ISBN 978-1-57735-880-0. doi: 10.1609/aaai.v37i9.26317. URL https://doi.org/10.1609/aaai.v37i9.26317.

Taiping Zhang, William S. Chandler, James M. Hoell, David Westberg, Charles H. Whitlock, and Paul W. Stackhouse. A global perspective on renewable energy resources: Nasa's prediction of worldwide energy resources (power) project. In D. Yogi Goswami and Yuwen Zhao (eds.), *Proceedings of ISES World Congress 2007 (Vol. I – Vol. V)*, pp. 2636–2640, Berlin, Heidelberg, 2009. Springer Berlin Heidelberg. ISBN 978-3-540-75997-3.

Alice Zheng and Amanda Casari. *Feature Engineering for Machine Learning: Principles and techniques for Data scientists*. O'Reilly Media, 2018.

A DETAIL ON THE ATTENTION MECHANISM

Since the proposed model is designed as a flexible framework, a simple implementation of the attention mechanism is used. This implementation resembles Bahdanau attention (Bahdanau et al., 2014). While their implementation aims to calculate the alignment scores between the sequences of an encoder and a decoder, the proposed approach focuses on differentially weighting the hidden states from the LSTM.

Given the set of hidden states $H = [h_1, h_2, ..., h_T]$ at the output of the LSTM, we calculate the scores using a linear layer:

$$s_t = Wh_t + b \tag{1}$$

where W and b are learnable matrix of weights and bias.

Then we calculate the attention weights by passing the scores through a softmax function:

$$\alpha_t = softmax(s_t) = \frac{e^{s_t}}{\sum_{i=1}^{T} e^{s_i}} \tag{2}$$

the context vector $\tilde{\mathbf{h}}$ is computed as the weighted sum of the LSTM hidden states, using the attention weights.

$$\tilde{\mathbf{h}} = \sum_{t=1}^{T} \alpha_t h_t \tag{3}$$

B SELECTED HYPERPARAMETERS FOR EACH MODEL

Table 3: Hyperparameters for the baseline models (Minixhofer et al., 2021) and for the proposed Hybrid Model (HM). Where applicable, the notation remains consistent with that used in the body of the article.

Hyperparameter	**Notation**	**LSTM**	**Transformer**	**Hybrid Model (HM)**
LSTM or Transformer Number of layers		2	4	2
LSTM Hidden size	h	512	512	490
Initial embedding size	z	-	256	27
Reduced embedding size (after FFNN)	z'	-	-	6
Final MLP number of layers		-	-	2
FFNN inner hidden size		-	4096	-
Attention Heads		-	2	-
Batch size		128	128	128
Dropout probability		0.1	0.1	0.1
Embeddings dropout probability		-	-	0.4
Weight Decay		0.01	0.01	0.01
Learning rate		7e-5	7e-5	7e-5
Number of epochs		7	7	9

C RESULTS ON THE TEST SET

Table 4: Results on the test set for the LSTM baseline and the proposed model.

Model	**MAE**	**RMSE**	**F_1**	**ROC-AUC**
LSTM	0.306	0.478	61.9	80.6
HM	0.218	0.378	67.3	85.9

D 5-FOLDS CV

Table 5: Extensive results of the 5-fold Cross-validation for the baseline LSTM and the proposed model.

	LSTM			**HM**		
Fold	MAE	RMSE	F1	MAE	RMSE	F1
1	0.347	0.553	58.34	0.244	0.433	60.22
2	0.365	0.570	42.79	0.302	0.519	59.67
3	0.272	0.444	66.22	0.254	0.404	75.22
4	0.332	0.548	44.82	0.266	0.433	59.84
5	0.310	0.504	63.88	0.299	0.502	71.06

Table 6: Mean and standard deviation for each metric on the 5-fold Cross-validation results.

Model	**MAE** $(\overline{x} \pm \sigma)$	**RMSE** $(\overline{x} \pm \sigma)$	**F_1** $(\overline{x} \pm \sigma)$
LSTM	0.325 ± 0.036	0.524 ± 0.051	55.2 ± 0.108
HM	0.273 ± 0.026	0.458 ± 0.050	65.2 ± 0.074

E DETAIL ON THE ABLATION STUDY

Table 7: Weekly results of the ablation study on the test set.

	Week 1		**Week 2**		**Week 3**		**Week 4**		**Week 5**		**Week 6**	
Model	MAE	F_1	MAE	F_1	MAE	F_1	MAE	F_1	MAE	F_1	MAE	F_1
HM	**0.126**	**82.2**	**0.169**	**74.7**	**0.209**	**68.6**	**0.244**	**64.0**	**0.269**	**58.6**	**0.294**	**51.0**
TS+Att	0.134	65.9	0.189	62.3	0.250	56.3	0.307	51.2	0.360	51.6	0.361	50.8
TS	0.136	61.9	0.192	62.3	0.253	56.3	0.312	56.4	0.364	51.6	0.368	50.8
SF+TS	0.144	73.9	0.203	62.3	0.262	56.3	0.320	51.2	0.374	49.3	0.375	50.8
SF	0.779	20.4	0.746	22.7	0.752	25.3	0.713	18.9	0.754	19.5	0.787	17.0

F LOCATION-AGNOSTING VS LOCATION-SPECIFIC TRAINING EXPERIMENT

Table 8: Weekly results for the HM on county vs national training data. The selected counties are Iowa (IA), Montana (MT) and Oklahoma (OK).

		Week 1		**Week 2**		**Week 3**		**Week 4**		**Week 5**		**Week 6**	
Train	Eval.	MAE	F_1	MAE	F_1	MAE	F_1	MAE	F_1	MAE	F_1	MAE	F_1
IA	IA	0.101	86.7	0.179	67.7	0.214	69.5	0.287	63.5	0.298	60.9	0.272	59.4
MT	MT	0.203	52.3	0.314	49.1	0.339	50.9	0.341	38.6	0.377	37.0	0.407	35.9
OK	OK	0.156	75.8	0.230	59.7	0.269	56.9	0.327	61.6	0.352	59.1	0.387	57.8
	IA	0.086	89.3	0.122	79.2	0.151	78.4	0.189	71.9	0.214	73.9	0.235	66.4
all	MT	0.144	59.5	0.168	52.9	0.178	50.0	0.209	45.5	0.237	44.4	0.265	38.6
	OK	0.096	83.1	0.160	75.8	0.196	77.9	0.209	77.5	0.260	73.5	0.298	66.3

Table 9: Results on the test set using county vs national training data.

Train	**Eval.**	**MAE**	**RMSE**	**F_1**
IA	IA	0.201	0.383	73.8
MT	MT	0.301	0.354	46.7
OK	OK	0.278	0.402	63.1
	IA	0.166	0.315	76.6
all	MT	0.200	0.320	48.2
	OK	0.218	0.378	67.3