A Hybrid Surrogate Deep Learning Model for Actual Evapotranspiration Prediction

Amir Aieb [1,2], Alexander Jacob [2], Antonio Liotta [1], Muhammad Usman Liaqat[2]
[1]*Faculty of Engineering, Free University of Bozen-Bolzano,* 39100 Bolzano, Italy.
Email: amir.aieb@student.unibz.it (A.A.), antonio.liotta@unibz.it (A.L.)
[2]*Institute for Earth Observation, Eurac Research, Bozen-Bolzano*, 39100 Bolzano, Italy.
Email: alexander.jacob@eurac.edu (A.J.), muhammadusman.liaqat@eurac.edu (MU.L.)

***Abstract*—The estimation of hydrological components on a spatiotemporal scale poses a challenge for researchers as they develop data-driven tools that can be transferred to different regions with varying characteristics. In this study, we propose a hybrid architecture of a surrogate deep learning (DL) model based on the data obtained by the Wflow estimation. The choice of the target region is based on extensive lowlands and large variations in elevation, which makes it more challenging to improve the model's accuracy. General tasks are addressed in this paper related to geodata frame structuring and preprocessing, DL model enhancement, and multiscale evaluation. Our contribution focuses on proposing a novel combination of LSTM, MLP, CNN, and CVAE to achieve robust outcomes on a finer scale, followed by an integration of FCM clustering for a comparative evaluation of the models' performance. Training both LSTM and MLP using climate data and geophysical information of the catchment provides a performance comparable to the Wflow benchmark. In addition, thanks to the FCM clustering which classifies the basin in homogenous subregions, we can gain extra insights as to how the models perform in the region as a whole. Our findings show that MLP performs very well in the first subregion, whereas LSTM outperforms the rest of the catchment. The integration of spatial information provided from both models via CNN_Hyb and CVAE_Hyb significantly enhances the overall spatiotemporal prediction across the entire region. However, clustering-based evaluation reveals the influence of LSTM and MLP on CNN accuracy, indicating persistent biases in the third subregion. In contrast, the utilization of CVAE_Hyb effectively mitigates this bias, resulting in a performance increase from 0.85 to 0.93.**

Keywords*—*actual evapotranspiration, surrogate model, hybrid deep learning, clustering-based evaluation.

I. INTRODUCTION

Hydrological modeling has a crucial role in understanding and predicting the behaviors of water systems, providing valuable insights for water resource management, extreme events (flood and drought), and environmental planning [1]. The traditional models are often based on the physical aspect, which describes the movement and distribution of water through complex systems. In practice, these models require a substantial amount of data and lengthy calibration procedures [2]. The exploration of artificial intelligence tools in hydrology between 2017 and 2020 led to increased adoption by researchers and a significant rise in hydrologic applications. An advancement were made in areas such as soil moisture (SM), and actual evapotranspiration (AE), through the use of data-driven models (DDM). [3]. The main concept of the following model's category is to determine the relationship between input and target variables in the absence of a clear understanding of the physical process of a certain system [4]. Different model classes were distinguished from the DDM family such as tree-based, support vector-based, neuron network, hybrid, and fusion models [5]. In recent years, surrogate hydrological modeling has emerged as a transformative approach that serves as an efficient proxy for complex hydrological processes [6]. That allows bridging the gap between traditional physically based models and the demand for real-time, data-driven insights. Surrogate models are advantageous because they can assimilate a wide range of data sources, including satellite observations and ground measurements. Considering a wide range of influencing factors enhances the robustness of hydrological forecasts [7].

Several studies have recommended using deep learning (DL) models for hydrological prediction components on a spatiotemporal scale. Artificial neuron network, such as The Multi-Layer Perceptron (MLP) offers a straightforward and effective method for modeling nonlinear relationships at individual grid cells in hydrological data, primarily depending on how well the input data aligns with the target data [8]. In contrast, the Long Short-Term Memory (LSTM) model features a more intricate recurrent architecture with memory cells, specifically designed to manage spatiotemporal data effectively, and has demonstrated reliable accuracy in simulating hydrological variables such as SM and AE [9]. On the other hand, models such as Convolutional Neural Networks (CNNs), and Conditional Variational Autoencoders (CVAEs) represent a significant paradigm shift in predicting actual evapotranspiration, offering a robust framework for handling spatially distributed data in hydrological applications [10]. In terms of application, CNNs excel in capturing spatial dependencies directly from data, leveraging convolutional layers to detect intricate terrain features. Meanwhile, CVAEs offer a probabilistic approach, combining convolutional layers with variational inference to learn spatial representations while explicitly modeling uncertainty. CNNs focus on spatial pattern recognition, and CVAEs provide probabilistic outputs, enabling uncertainty quantification and the generation of synthetic scenarios [10]. Recent advancements in evapotranspiration prediction involve hybrid DL models, especially when calibration parameters are limited and only climate data is available. To tackle these challenges, a combination of Recurrent Neural Networks (RNN) and Convolutional Neural Networks (CNN), such as the CNN-BiLSTM model, has been proposed to enhance prediction accuracy by integrating spatial and temporal features with attention-based weighting [11]. Additionally, Conv-LSTM and CNN-LSTM models have been developed for estimating AE, offering improved performance over traditional empirical models by effectively managing the complexities of limited data and capturing intricate patterns in the data [12]. A critical challenge for data-driven models is the difficulty in managing high-dimensional input features, which can significantly increase modeling costs. To address this issue, employing adaptive sampling strategies is crucial for efficient model training. Surrogate deep learning (SDL) models in hydrology offer a powerful alternative to traditional methods by using neural networks to approximate complex hydrological processes. SDL are widely used due to their capability to incorporate high-dimensional parameter fields, without requiring the sampling of grids [13]. Yet, analyzing the performance of DL in hydrological modeling is a big challenge, due to the big variability when dealing with spatiotemporal scales. However, when constructing an SDL model to simulate hydrological parameters, the non-homogeneous geophysical data across spatial scales poses significant challenges, especially during the calibration phase. It refers to the variation in the statistical properties of the data across different locations, which can significantly impact the model's accuracy.

In this paper, we addressed the challenge of minimizing the SDL biases across the entire region by proposing a new hybrid SDL model architecture for predicting spatiotemporal daily actual evapotranspiration (DAE) provided by the Wflow model, as its global, grid-based approach delivers high-resolution and precise simulations of hydrological processes [14]. The new architecture is based mainly on the integration of CNN and CVAE models to capture the spatial information provided in turn by LSTM and MLP. Therefore, helps to increase the accuracy of the SDL model, particularly in the area where the calibration parameters provide a nonhomogeneous pattern compared to the DAE ground truth data. The choice of CNN and CVAE model is based on the different techniques employed for the convolutional layer, which are tailored specifically for spatial scale. Several contribution tasks will be discussed in this paper to better understand how the proposed surrogate model can enhance DAE prediction to capture a finer scale. The proposed methodology employs climate data, including precipitation, temperature, and potential evapotranspiration, to process the DAE across spatiotemporal scales, while geophysical information of the catchment helps in spatial calibration. Choosing the optimal features for training the SDL model is achieved through quality assessment. Subsequently, a comparative analysis of DL-integrated models within the surrogate modeling framework is conducted using Fuzzy C-Means (FCM) clustering. This method classifies the target basin into homogeneous subregions by leveraging ground truth parameters from the Wflow DAE dataset. Additionally, comparing these clusters with those derived from input feature classifications helps to identify areas prone to biases when using traditional models such as LSTM and MLP. The Adige catchment in northern Italy is selected due to its significant spatiotemporal variability in hydroclimate parameters, influenced by its diverse geomorphological characteristics, which poses a challenging task for DL models.

The paper is structured into three main sections, preceded by the introduction. Firstly, it presents a novel surrogate architecture for hydrological prediction. Subsequently, it discusses the results obtained from training and testing the models, focusing on the accuracy achieved through hyperparameter optimization. Finally, it integrates a clustering-based evaluation for a multiscale analysis of the model.

II. STUDY AREA

The Adige catchment is approximately 12,100 km^2. In terms of its discharge, this catchment exhibits the same flow behaviors as other Alpine catchments, with peak flow usually occurring between June and September during the snow-melting season [15]. Its main tributaries are located in the Alpine region between 45.8- and 46.6 degrees North latitude and 10.8 to 12.6 degrees East longitude (Fig 1), making them highly dependent upon snow dynamics.

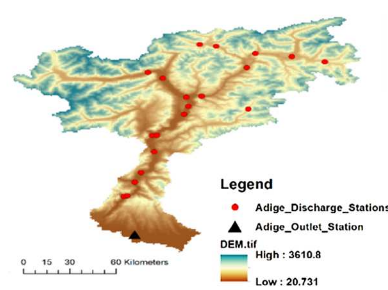

Fig. 1. Map of Adige catchment showing the elevations and discharge stations.

The effects of climate change have already had significant impacts on water resources management in this area, particularly in terms of hydropower generation and winter tourism [16]. Precipitation within the catchment is not uniform, with values ranging from 500 mm/yr in Val Venosta (located in the north-western part of the catchment) to 1600 mm/yr in the southern part of the basin, as reported by [17]. Since the catchment has a high elevation gradient, temperature also varies within it. The average monthly temperature ranges between 14°C in July and -4°C in January and December. The Adige catchment encompasses an area of approximately 12,100 km2 as reported by [17].

III. PROPOSED SURROGATE MODEL

The architecture of the proposed surrogate deep learning (SDL) model for predicting spatiotemporal daily actual evapotranspiration (DAE) is illustrated in Figure 2. This section provides a detailed overview of the hybrid model, beginning with data preparation and concluding with the acquisition of outcomes. The process involves generating target data for the surrogate model using the Wflow physical model, incorporating three dynamic climate-related features and 74 static parameters representing the catchment's geophysical characteristics. Following this, a preprocessing step is performed to enhance accuracy and reduce computational costs associated with the SDL algorithms. Typically, grid cell data from Netcdf files is utilized to train the SDL model, with dynamic features arranged in a 4D format (Longitude, Latitude, Time, climate parameters) and static features in a 3D format (Longitude, Latitude, geophysical variables) as shown in Figure 2. To ensure data quality, various tasks such as reshaping and masking are implemented to align with the dimensionality of the target data. Additionally, different strategies are applied to address missing data (MD) encountered during the masking process. Proper handling or exclusion of these gaps is crucial, as it can significantly influence the quality of hydrological models, depending on the extent of the MD.

The second step in the preprocessing phase involves feature reduction to mitigate potential negative impacts on the accuracy of the SDL model. This is achieved through various analyses, including correlation analysis, homogeneity assessment, and comparison of data distributions between input and target variables. Figure 2 outlines the sequential processing steps proposed for the hybrid surrogate model. Initially, LSTM and MLP models are used to predict daily actual evapotranspiration (DAE) by leveraging climate and geophysical information from the catchment. The LSTM model effectively captures the temporal variability in climate features, providing preliminary results that require further spatial calibration using additional catchment information. In contrast, the MLP model, a simpler artificial neural network (ANN), efficiently consolidates and retains relevant data from both climate and geophysical parameters at each pixel. In the subsequent processing step, CNN and CVAE models are integrated into the surrogate architecture to combine the outputs of the LSTM and MLP models. The evaluation analysis assesses the performance of both models to determine which convolutional layer most effectively integrates the spatiotemporal information from the LSTM and MLP models.

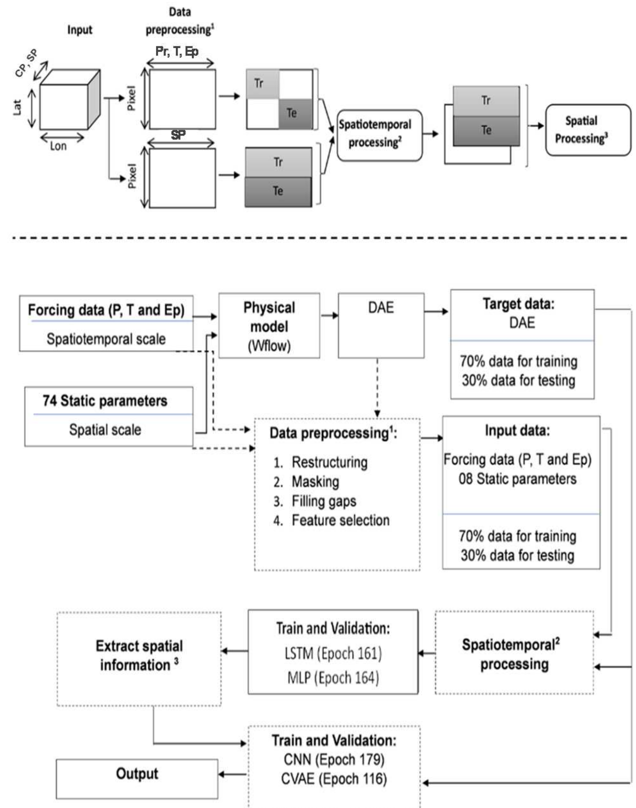

Fig. 2. Flowchart summarizing the new model architecture of hybrid surrogate-based deep learning model. Daily actual evapotranspiration (DAE); Precipitation (P); Temperature (T); Potential evapotranspiration (Ep); Climate parameter (CP); Static Parameter (SP); Time (T); Training (Tr); Testing (Te).

The CNN model uses a kernel function to capture the spatial variability of DAE, while the CVAE functions as a generative model that encodes and decodes high-dimensional data in a supervised manner, produces synthetic data samples that enhance the model's ability to perform spatiotemporal calibration.

IV. RESULTS AND DISCUSSION

This section undertakes a statistical evaluation and comparative analysis of accuracy among the DL models proposed for the surrogate architecture. It begins by delineating the spatiotemporal variability of both input and target data, followed by the application of spatial correlation for feature selection. Subsequently, we present the outcomes of training and testing, including the hyperparameters' optimization. Additionally, we introduce a pioneering evaluation approach employing FCM clustering to assess the performance of each model across various scales, encompassing spatial and temporal resolutions such as daily, mean daily, and seasonal. Several tests were used during the analysis, such as coefficient of determination (R^2), root mean square error (RMSE), mean absolute error (MAE), relative standard error (RSE), t-test, Kruskal-Wallis (Kwt), Kolmogorov-Smirnov (Ks), Hellinger Distance (HD), and Bhattacharyya Distance (BD) [18-24].

A. Data description and feature selection

In Figure 3, the spatiotemporal and spatial variability, along with the correlation analysis of both dynamic and static features selected for the proposed surrogate model, are depicted. In this analysis, a time of four years between 2019 and 2022 is considered for both climate and Wflow data. The findings are visually presented through histograms and trend curves.

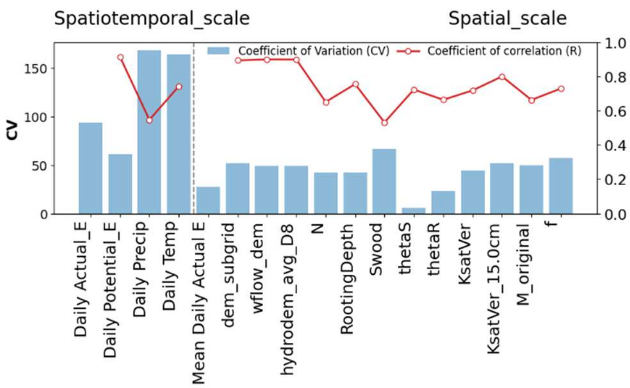

Fig. 3. Histograms of CV followed by trend plot of correlation coefficient (R), describing input features selection for surrogate-based DL model.

A robust correlation is observed on the spatiotemporal scale, between the daily potential evapotranspiration (DEP) and temperature (DT) compared to DAE, shown by an R score ranging from 0.75 to 0.90. While, a remarkable variability in DT compared to the target data is evident, potentially impacting the accuracy of the hydrological model on the spatiotemporal scale. Within this analysis, we identified 12 out of 74 static parameters that exhibit a strong correlation with the mean DAE. Notably, dem_subgrid, Wflow_dem, hydrodem_avg_D8 and KsarVer_15.0cm demonstrate the highest correlation, exceeding 0.85 on the spatial scale.

B. Deep learning models' performance

The graphical visualization of the training and testing deep learning models to predict DAE between 2019 and 2022 is summarized in Figure 4, complemented by statistical results shown in Table I. In this analysis, the used dataset is partitioned as follows: the period from 2019 to 2020 is dedicated to training both LSTM and MLP models, while data obtained in 2021 serve for testing both LSTM and MLP. Concurrently, the same 2021 dataset is used for training and testing both CNN_Hyb and CVAE_Hyb, with a focus on refining outcomes from previous processing phases. The period of 2022, is dedicated to validating and comparing the accuracy of all DL models across the entire region, maintaining uniformity in scale (Table I). During this analytical phase, various tests are conducted to evaluate the models' performance for both training and testing. Which is the assessment of absolute residual values via heatmaps, density curve analysis, and examination of loss function for 200 epochs. Additionally, other metrics including R^2, RMSE, HD, and BD provided in Figure 4 offer further insight into accuracy and similarity concerning data distribution in comparison to the physical assessment of DAE provided by Wflow. Notably, CNN_Hyb and CVAE_Hyb exhibit superior performance over LSTM and MLP across both training and testing phases. Particularly, CVAE_Hyb demonstrates exceptional results and enhanced stability throughout all periods, boasting R2 values of 0.95 and 0.94, respectively. On the contrary, LSTM exhibits slightly better performance compared to MLP, as indicated by the RMSE and R^2 results presented in Figure 4. This distinction becomes more apparent when examining the heatmaps and density curves, particularly during the testing phase. Notably, the data generated by CVAE_Hyb is closely similar to the distribution of Wflow data, displaying minimal deviation from the actual data with HD and BD values of 10.30 and 11.97, respectively. These results were attained following the optimization of each DL model's hyperparameters using 200 epochs. The loss score illustrates the robustness and stability of the CVAE_Hyb model, with errors becoming more similar after training and testing using a few epochs. Additionally, Figure 4 shows the optimal hyperparameters for LSTM, MLP, CNN_Hyb, and CVAE_Hyb were identified at epochs 161, 164, 179, and 116, respectively.

A statistical analysis of the predicted DAE data distribution and variability by each proposed DL model is shown in Table I, using different metrics. CVAE_Hyb consistently outperforms the other models, as evidenced by its highest R^2_{Adj} (0.950) and lowest MAE (0.130), indicating strong predictive accuracy. Additionally, both t-tests and KS tests statistic parameters reveal significant performance and the minimum data distribution differences, with CNN_Hyb and CVAE_Hyb consistently showing the lowest p-values, implying distinct superiority over LSTM and MLP. Overall, CVAE_Hyb emerges as the most accurate and statistically robust model, demonstrating its effectiveness in capturing complex patterns and accurately representing data distributions in hydrological modeling.

C. FCM clustering-based evaluation

The multiscale evaluation of the DL model's performance proposed for the new surrogate model architecture is extensively analyzed in this subsection, using FCM clustering. Various techniques have been applied in this step, including the detection of the middle pixel of each homogeneous regions which serve as reference points for analyzing the entire subregion for comparison.

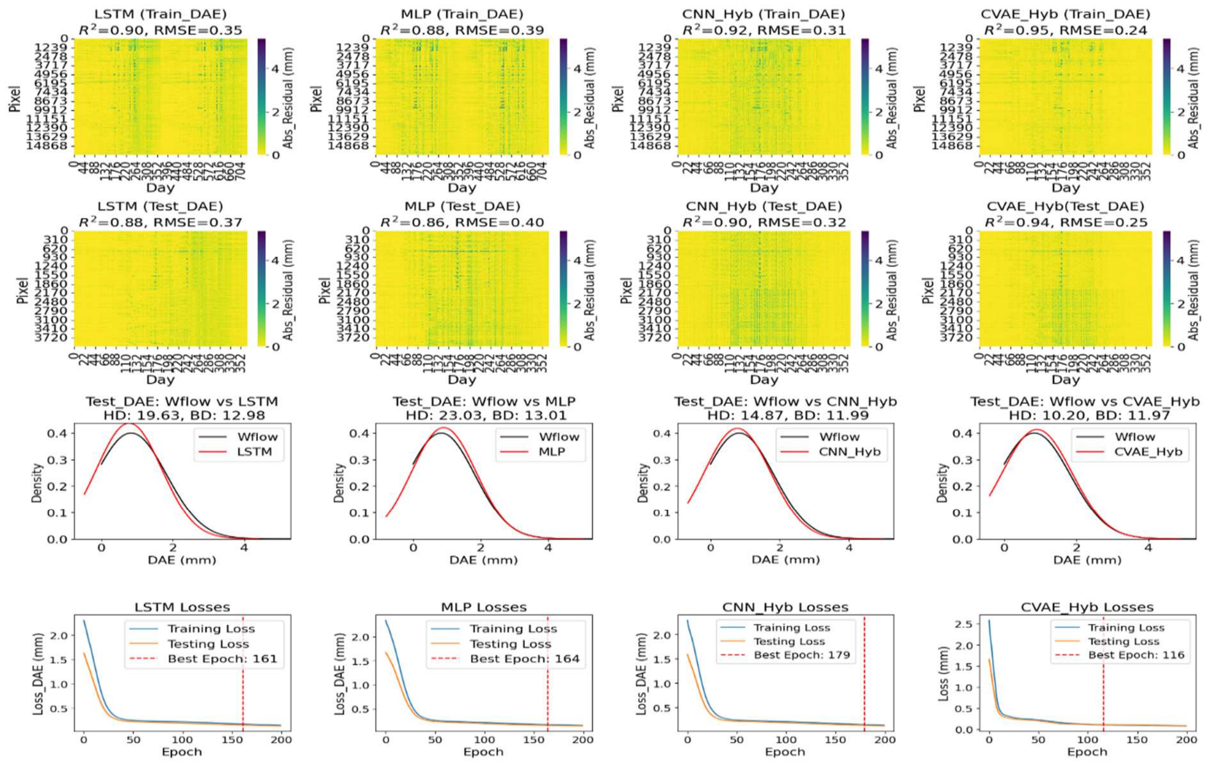

Fig. 4. Statistical assessment of DL model training and testing for daily actual evapotranspiration (DAE) prediction during 2019 and 2021, including heat maps of absolute residuals, density curves, and loss plots. Hellinger Distance (HD); Bhattacharyya Distance (BD).

Table I. Statistical performance of surrogate-based DL models for daily actual evapotranspiration prediction in 2022.

Index	LSTM	MLP	CNN_Hyb	CVAE_Hyb
R^{2Adj}	0.890	0.850	0.920	0.950
MAE	0.220	0.250	0.190	0.130
RSE	0.290	0.320	0.270	0.200
ttest_Pvalue	0.040	0.009	0.005	0.001
ttest_statistic	163.830	37.910	32.870	17.280
kwt_Pvalue	0.040	0.030	0.005	0.001
kwt_statistic	5293.060	4072.160	17082.410	112924.000
KS_Pvalue	0.021	0.035	0.009	0.001

Adjusted coefficient of Determination (R^2_{Adj}); Mean Absolute Error (MAE); Root Squared Error (RSE); Kruskal-Wallis (kwt). Kolmogorov-Smirnov test (KS).

It represents a consistent visualization for all the surrounding pixels. FCM, as a model, can capture spatial scale; hence, we integrate a set of statistical information about forcing climate data and Wflow_DAE, such as the min, mean, max, median, Q1, Q3, and skewness values. This helps in providing comprehensive information about spatiotemporal variability. Different tests have been applied in this evaluation under daily, mean daily (shown in Figure 6), and seasonal (shown in Figure 7) scales. Figure 5 shows the regional classification obtained by applying the FCM model using input features and Wflow_DAE. The results highlight how preprocessing features can impact model accuracy. Typically, the data distribution in the 1st and 2nd clusters shows notable similarity.

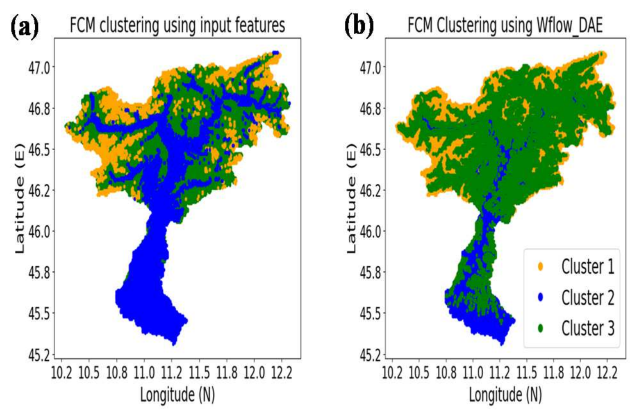

Fig. 5. Comparative assessment of FCM clustering within the Adige basin, using statistical parameters of input features (a) and target features (b). Input features: climate and static features; Target feature: actual evapotranspiration.

In contrast, when using the selected features, both LSTM and MLP models fail to produce satisfactory outcomes under spatial scale in the southern part of the Adige basin. This creates challenges for our work, particularly in analyzing and proposing a more accurate surrogate model to predict data in this subregion.

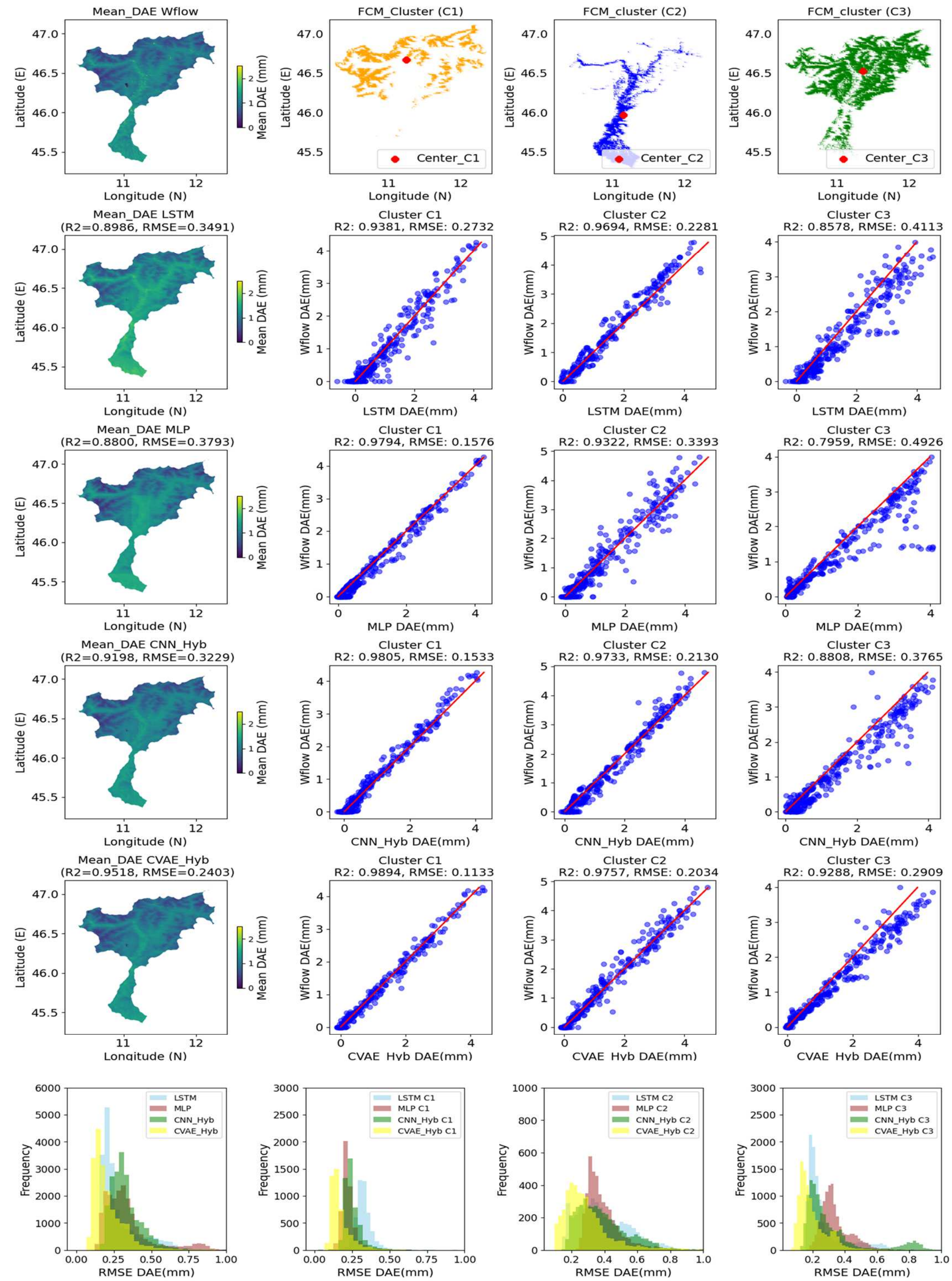

Fig. 6. An integrated evaluation of DL models' accuracy in predicting daily actual evapotranspiration (DAE) through FCM clustering-based assessment for 2022. Coefficient of Determination (R^2); Root Mean Square Error (RMSE).

Figure 6, displays the results of predicting DAE in 2022 using all the SDL models proposed for this paper. The analysis is based mainly on the results provided by all models in each cluster under a daily scale, given by regression plots between predicted and ground truth data, and heatmaps of mean DAE for the entire region.

The results reveal a strong performance of the CVAE_Hyb model in all subregions, where all metrics indicate its stability and robustness in predicting DAE throughout the whole region, given by R^2 ranging between 0.93 and 0.98. The model effectively addresses issues presented by the non-homogenous input features in the 3rd subregion due to its ability to capture high resolution via generating synthetic data. By contrast, the other models exhibit some outliers, particularly in the 3rd cluster. Additionally, CNN_Hyb demonstrates an improvement compared to LSTM and MLP; however, in terms of accuracy, CVAE_Hyb proves to be more accurate. These results are further shown by the RMSE frequencies, where in all cases, CVAE_Hyb exhibits the minimum interval values of RMSE distribution, shown graphically with a left-side skewness, indicating that most values are lower than the mean RMSE. In the first cluster, characterized by high elevation (> 2000 m), the MLP outperforms the LSTM throughout the year. However, the LSTM proves more effective in predicting DAE in areas with elevations below 1000 m, as shown in Figures 1 and 6.

The spatiotemporal visualization of the predicted DAE distribution from various DL models presented in this paper is compared to the Wflow results using scatter plots. Figure 7 illustrates these comparisons, aiding in the diagnostic of data trends and variability relative to the previously demonstrated results. This analysis significantly contributes to understanding the accuracy of DL models across different scales. During the autumn period, results reveal a bias from the LSTM model in all regions, where its outcomes tend to overfit the Wflow data. In contrast, the CVAE_Hyb model demonstrates stability and robustness in capturing all scales, showing a data distribution similar to the ground truth. Additionally, the results indicate that during the dry period, particularly in summer, there is a similarity among all outcomes provided by the DL models across the entire region.

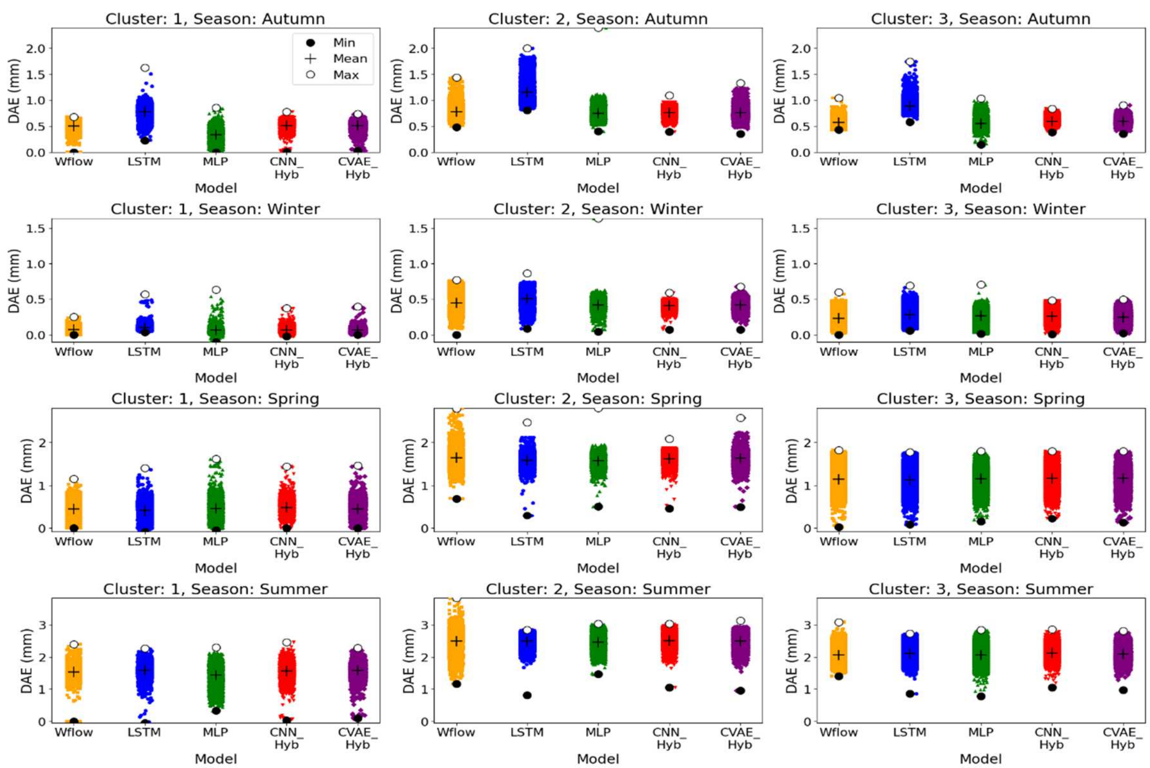

Fig. 7. Scatter plots displaying the seasonal daily actual evapotranspiration predicted by different DL models during 2022, across three specific subregions delineated by FCM clustering. Daily actual evapotranspiration (DAE).

V. CONCLUSIONS AND FUTURE WORK

Our overarching aim is to predict spatiotemporal daily actual evapotranspiration (DAE) using a small set of input features. This investigation proposes a novel surrogate model architecture. A preprocessing step was applied in this work to select the most efficient input features using for models' prediction. The application of FCM clustering shows a crucial approach to anticipate information on the impact of input features on the models' accuracy. A notable degree of similarity was observed within the first cluster when comparing the distribution provided by the input data and the ground truth DAE. High performance was achieved in this subregion using relatively simple models such as LSTM and MLP, yielding R^2 values of 0.94 and 0.98, respectively.

Conversely, in the rest of the catchment, particularly in the third subregion where the input data are non-homogeneous compared to the DAE ground truth, some biases were observed when using both models, requesting the integration of CNN and CVAE. The selection of these models aims to assess their convolutional characteristics in capturing spatial scales using the information provided by the preceding models. Furthermore, results demonstrate an improvement in DAE prediction accuracy when employing both CNN_Hyb and CVAE_Hyb. However, CNN_Hyb performance depends upon the quality of results provided by MLP and LSTM. Contrariwise, CVAE_Hyb emerges as the best model, exhibiting robustness and stability throughout the entire region, with a noteworthy 8% improvement in accuracy observed within the third region, given by an R2 value of 0.93.

Looking ahead, we plan to enhance the proposed model's applicability by incorporating a preprocessing step. Refining the calibration parameters is essential for increasing the accuracy of the SDL model by analyzing the importance of features within each subregion and making them more homogeneous to perform effectively across the entire area. Furthermore, employing more advanced DL techniques tailored to different climate scenarios will enable the model to better manage the complex patterns of climate in various subregions.

ACKNOWLEDGMENTS

This project has been supported by European Union PNRR Funding under Italian DM 352/2022.

REFERENCES

[1] M. I. Brunner, L. Slater, L. M. Tallaksen, and M. Clark, "Challenges in modeling and predicting floods and droughts: A review," *Wiley Interdisciplinary Reviews: Water,* vol. 8, p. e1520, 2021.

[2] C. Paniconi and M. Putti, "Physically based modeling in catchment hydrology at 50: Survey and outlook," *Water Resources Research,* vol. 51, pp. 7090-7129, 2015.

[3] P. Aghelpour, V. Varshavian, M. Khodamorad Pour, and Z. Hamedi, "Comparing three types of data-driven models for monthly evapotranspiration prediction under heterogeneous climatic conditions," *Scientific Reports,* vol. 12, p. 17363, 2022.

[4] T. Kim, T. Yang, S. Gao, L. Zhang, Z. Ding, X. Wen, J. J. Gourley, and Y. Hong, "Can artificial intelligence and data-driven machine learning models match or even replace process-driven hydrologic models for streamflow simulation?: A case study of four watersheds with different hydro-climatic regions across the CONUS," *Journal of Hydrology,* vol. 598, p. 126423, 2021.

[5] R. Zhang, R. Zen, J. Xing, D. M. S. Arsa, A. Saha, and S. Bressan, "Hydrological process surrogate modelling and simulation with neural networks," in *Advances in Knowledge Discovery and Data Mining: 24th Pacific-Asia Conference, PAKDD 2020, Singapore, May 11–14, 2020, Proceedings, Part II 24*, 2020, pp. 449-461.

[6] S. Razavi, B. A. Tolson, and D. H. Burn, "Review of surrogate modeling in water resources," *Water Resources Research,* vol. 48, 2012.

[7] W.-P. Tsai, D. Feng, M. Pan, H. Beck, K. Lawson, Y. Yang, J. Liu, and C. Shen, "From calibration to parameter learning: Harnessing the scaling effects of big data in geoscientific modeling," *Nature communications,* vol. 12, p. 5988, 2021.

[8] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation,* vol. 9, pp. 1735-1780, 1997.

[9] O. Kaissi, S. Belaqziz, M. H. Kharrou, S. Erraki, C. El Hachimi, A. Amazirh, and A. Chehbouni, "Advanced learning models for estimating the spatio-temporal variability of reference evapotranspiration using in-situ and ERA5-Land reanalysis data," *Modeling Earth Systems and Environment,* pp. 1-25, 2023.

[10] A. Singh and A. Haghverdi, "Development and evaluation of temperature-based deep learning models to estimate reference evapotranspiration," *Artificial Intelligence in Agriculture,* vol. 9, pp. 61-75, 2023.

[11] S. Wang, X. Yu, Y. Li, S. Wang, and C. Meng, "Application of a hybrid deep learning approach with attention mechanism for evapotranspiration prediction: a case study from the Mount Tai region, China," *Earth Science Informatics,* vol. 16, pp. 3469-3487, 2023.

[12] G. Sharma, A. Singh, and S. Jain, "A hybrid deep neural network approach to estimate reference evapotranspiration using limited climate data," *Neural Computing and Applications,* pp. 1-20, 2022.

[13] R. Sun, B. Pan, and Q. Duan, "A surrogate modeling method for distributed land surface hydrological models based on deep learning," *Journal of Hydrology,* vol. 624, p. 129944, 2023.

[14] W. J. van Verseveld, A. H. Weerts, M. Visser, J. Buitink, R. O. Imhoff, H. Boisgontier, L. Bouaziz, D. Eilander, M. Hegnauer, and C. ten Velden, "Wflow_sbm v0. 6.1, a spatially distributed hydrologic model: from global data to local applications," *Geoscientific Model Development Discussions,* vol. 2022, pp. 1-52, 2022.

[15] G. Chiogna, P. Skrobanek, T. S. Narany, R. Ludwig, and C. Stumpp, "Effects of the 2017 drought on isotopic and geochemical gradients in the Adige catchment, Italy," *Science of the Total Environment,* vol. 645, pp. 924-936, 2018.

[16] B. Majone, F. Villa, R. Deidda, and A. Bellin, "Impact of climate change and water use policies on hydropower potential in the south-eastern Alpine region," *Science of the Total Environment,* vol. 543, pp. 965-980, 2016.

[17] L. Laiti, S. Mallucci, S. Piccolroaz, A. Bellin, D. Zardi, A. Fiori, G. Nikulin, and B. Majone, "Testing the hydrological coherence of high-resolution gridded precipitation and temperature data sets," *Water Resources Research,* vol. 54, pp. 1999-2016, 2018.

[18] S. S. M. Nadoushani, N. Dehghanian, and B. Saghafian, "A fuzzy hybrid clustering method for identifying hydrologic homogeneous regions," *Journal of Hydroinformatics,* vol. 20, pp. 1367-1386, 2018.

[19] C. Hagquist and M. Stenbeck, "Goodness of fit in regression analysis–R 2 and G 2 reconsidered," *Quality and Quantity,* vol. 32, pp. 229-245, 1998.

[20] T. Chai and R. R. Draxler, "Root mean square error (RMSE) or mean absolute error (MAE)?–Arguments against avoiding RMSE in the literature," *Geoscientific model development,* vol. 7, pp. 1247-1250, 2014.

[21] M. A. Khan, S. A. Memon, F. Farooq, M. F. Javed, F. Aslam, and R. Alyousef, "Compressive strength of fly-ash-based geopolymer concrete by gene expression programming and random forest," *Advances in Civil Engineering,* vol. 2021, pp. 1-17, 2021.

[22] T. K. Kim, "T test as a parametric statistic," *Korean journal of anesthesiology,* vol. 68, p. 540, 2015.

[23] P. E. McKight and J. Najab, "Kruskal-wallis test," *The corsini encyclopedia of psychology,* pp. 1-1, 2010.

[24] V. W. Berger and Y. Zhou, "Kolmogorov–smirnov test: Overview," *Wiley statsref: Statistics reference online,* 2014.