Earth Science Informatics (2023) 16:3717–3728
https://doi.org/10.1007/s12145-023-01122-4

RESEARCH

Machine learning models to complete rainfall time series databases affected by missing or anomalous data

Andrea Lupi[1] · Marco Luppichini[1] · Michele Barsanti[2] · Monica Bini[1,3,4] · Roberto Giannecchini[1,3,5]

Received: 7 July 2023 / Accepted: 7 October 2023 / Published online: 20 October 2023

Abstract

In recent years, artificial intelligence in geosciences is spreading more and more, thanks to the availability of a large amount of data. In particular, the development of automatic raingauges networks allows to get rainfall data and makes these techniques effective, even if the performance of artificial intelligence models is a consequence of the coherency and quality of the input data. In this work, we intended to provide machine learning models capable of predicting rainfall data starting from the values of the nearest raingauges at one historic time point. Moreover, we investigated the influence of the anomalous input data on the prediction of rainfall data. We pursued these goals by applying machine learning models based on Linear Regression, LSTM and CNN architectures to several raingauges in Tuscany (central Italy). More than 75% of the cases show an R2 higher than 0.65 and a MAE lower than 4 mm. As expected, we emphasized a strong influence of the input data on the prediction capacity of the models. We quantified the model inaccuracy using the Pearson's correlation. Measurement anomalies in time series cause major errors in deep learning models. These anomalous data may be due to several factors such as temporary malfunctions of raingauges or weather conditions. We showed that, in both cases, the data-driven model features could highlight these situations, allowing a better management of the raingauges network and rainfall databases.

Keywords Climate time series · Rainfall prediction · Fill of missing data · Machine learning models

Introduction

Climate change is one of the most relevant issues for humanity in the Anthropocene era (Malhi et al. 2020). The average temperature on the mainland in the years 2006–2015 was 1.53 °C higher than that of the years 1850–1900 (IPCC 2019). It is now well known that higher temperature is causing severe changes in precipitation regimes as well, with increasingly extreme events (Hardwick Jones et al. 2010; Myhre et al. 2019; Tramblay et al. 2020; Luppichini et al. 2023b). It is also causing an alteration of the beginning and end of growing seasons, causing a general decrease in the regional crop yields and freshwater availability (Minoli et al. 2022). The biodiversity is further stressed and tree mortality increases (IPCC 2019). Understanding and modelling the past, present, and future climate are of fundamental importance to the issue of climate change and variability. Effective climate models represent one of our primary tools for projecting and adapting to climate change (Schmidt 2011).

Moreover, climate change has direct repercussion on the hydrogeological systems and groundwater resources, and the hydrogeological models are consequently part of the climate models (Amanambu et al. 2020; Li et al. 2022). Rainfall data, and precipitation in general, their variability, intensity, and duration, have paramount importance for the hydrogeological models (Sattari et al. 2017). Independently from the used model, several problems can still affect the

Communicated by: H. Babaie

✉ Marco Luppichini
marco.luppichini@dst.unipi.it

[1] Department of Earth Sciences, University of Pisa, Via S. Maria, 52, 56126 Pisa, Italy

[2] Department of Civil and Industrial Engineering, University of Pisa, Largo L. Lazzarino, 56122 Pisa, Italy

[3] CIRSEC Centro Interdipartimentale Di Ricerca Per Lo Studio Degli Effetti del Cambiamento Climatico Dell'Università Di Pisa, Via del Borghetto 80, 56124 Pisa, Italy

[4] Istituto Nazionale Di Geofísica E Vulcanologia (INGV), Via Vigna Murata 605, 00143 Rome, Italy

[5] Institute of Geosciences and Earth Resources, IGG-CNR, Via Moruzzi 1, 56124 Pisa, Italy

input rainfall datasets. These issues often concern missing or incorrect information that can lead models to misleading results. It is worth noting that the geographical distribution of raingauges is generally not uniform (e.g., for some areas there is a deficiency or lack of raingauges). Furthermore, the completeness of the time series is not always guaranteed due to, for example, non-continuous operation of raingauges during the monitoring period (Lebay and Le 2020).

Many physically based methods simplify the natural system features to predict its behaviour (Antonetti and Zappa 2018). However, the natural systems are inherently heterogeneous (Marçais and de Dreuzy 2017) and the physically based methods may show inherent limitations in reproducing natural phenomena. In recent years, the use of artificial intelligence (AI) and graphical processing units (GPUs) have enabled remarkable advances in machine learning (and especially in deep learning) applications such as techniques based on multilayer artificial neural networks (ANNs). Deep learning models have been successfully applied in many forecasting situations, including time series forecasting (Zheng et al. 2019; Yi et al. 2019; Fawaz et al. 2020; Nigro et al. 2022). Time series typically have chaotic and noisy problems and deep learning approaches are the most effective techniques for solving them (Livieris et al. 2020). Several authors use the rainfall dataset to create deep learning models available to replicate run-off processes (van Loon and Williams 1976; Marçais and de Dreuzy 2017; Kratzert et al. 2018; Boulmaiz et al. 2020; Sit et al. 2020; Tien Bui et al. 2020; Chattopadhyay et al. 2020; Luppichini et al. 2022a, 2023a). Long short-term memory (LSTM) and convolutional neural networks (CNNs) are two of the most popular, efficient, and used deep learning techniques (Zheng et al. 2019; Yi et al. 2019; Fawaz et al. 2020). In the last period, some works combined LSTM and CNN models for time series prediction (Kimura et al. 2019; Baek et al. 2020; Van et al. 2020; Xu et al. 2020). The benefits of the combined CNN-LSTM models are a consequence of the characteristic of LSTM of acquiring efficiently the information of sequence patterns, thanks to their peculiar architecture. The CNN layers filter out the noise in the input data to extract the most significant features needed for the final prediction model. Furthermore, standard CNN can identify spatial autocorrelation between data but is usually not suitable for a correct analysis of a complex temporal dependence over long times (Bengio et al. 2013; Livieris et al. 2020). Several works used deep learning models based on the LSTM networks to create run-off simulations (Kratzert et al. 2018; Le et al. 2019; Boulmaiz et al. 2020; Li et al. 2020; Liu et al. 2020; Nguyen and Bae 2020; Hu et al. 2020), whereas others based on CNN (Li et al. 2018; Huang et al. 2020; Kim and Song 2020; Hussain et al. 2020) or a combination of both (CNN-LSTM) (Kimura et al. 2019; Baek et al. 2020; Van et al. 2020; Xu et al. 2020). The performance of encoder-decoder LSTM layers (LSTM-ED) is great with sequential data like a time series. This architecture consists of two blocks: one to read the input sequence and encode it into a fixed-length vector, and a second one to decode the fixed-length vector and transmit the intended sequence (Sutskever et al. 2014).

Among several applications in hydrological modelling, deep learning models are also used for several additional applications, such as reconstructing missing data and predicting rainfall data (Gers et al. 2001). In these models, the input data must obviously be of high quantity and good quality. Furthermore, machine learning models are used to apply statistical regionalization procedures, which constitute a set of methodologies used to divide a geographical area into statistically homogeneous regions. The main aim of statistical regionalization is to simplify data understanding and analysis, allowing analysts to attain a more detailed and meaningful insight into local dynamics. Among the most common methodologies are cluster analysis, which groups similar geographic units based on relevant variables, and principal component analysis, which identifies common patterns of variation among the units (Yin et al. 2016; Alem et al. 2019; De Luca and Napolitano 2023).

This work intends to use machine learning models to predict rainfall data taking advantage of a network of sensors. The models recreate precipitation time series by using data from nearby raingauges as inputs. The training data lacks temporal information, but each record is referred to a specific time. This allows the missing data to be entered into a rainfall database, allowing to complete time series for applying several types of study requiring the time series continuity (e.g., statistical methods, trend analysis, etc.). We also wanted to analyse the errors of the models investigating the role of anomalous data that can influence the performance of deep learning models. Indeed, all meteorological databases can have anomalous data caused by rare natural phenomena or anthropic factors (e.g., malfunctions of the sensor network). Understanding the answer of the machine learning models to the presence of these data is a key point for future applications of AI techniques in hydrological and meteorological studies.

We applied three machine learning models: the first one is a linear regression (LR), whereas the second and the third ones are based on CNN and LSTM. The first architecture of the deep learning models relies on a combination of CNN and LSTM layers (CNN-LSTM), whereas the second one relies on ED-LSTM. The dataset used is derived from 349 raingauges located in Tuscany (central Italy; Fig. 1), characterized by an extensive monitoring network and a wide variability of the mean annual precipitation (MAP), which is influenced by the morphology of the territory (Cantù 1977; Rapetti and Vittorini 1994; Fratianni and Acquaotta 2017). Tuscany is indeed very heterogeneous from a morphological and a geological point of view, characterized by

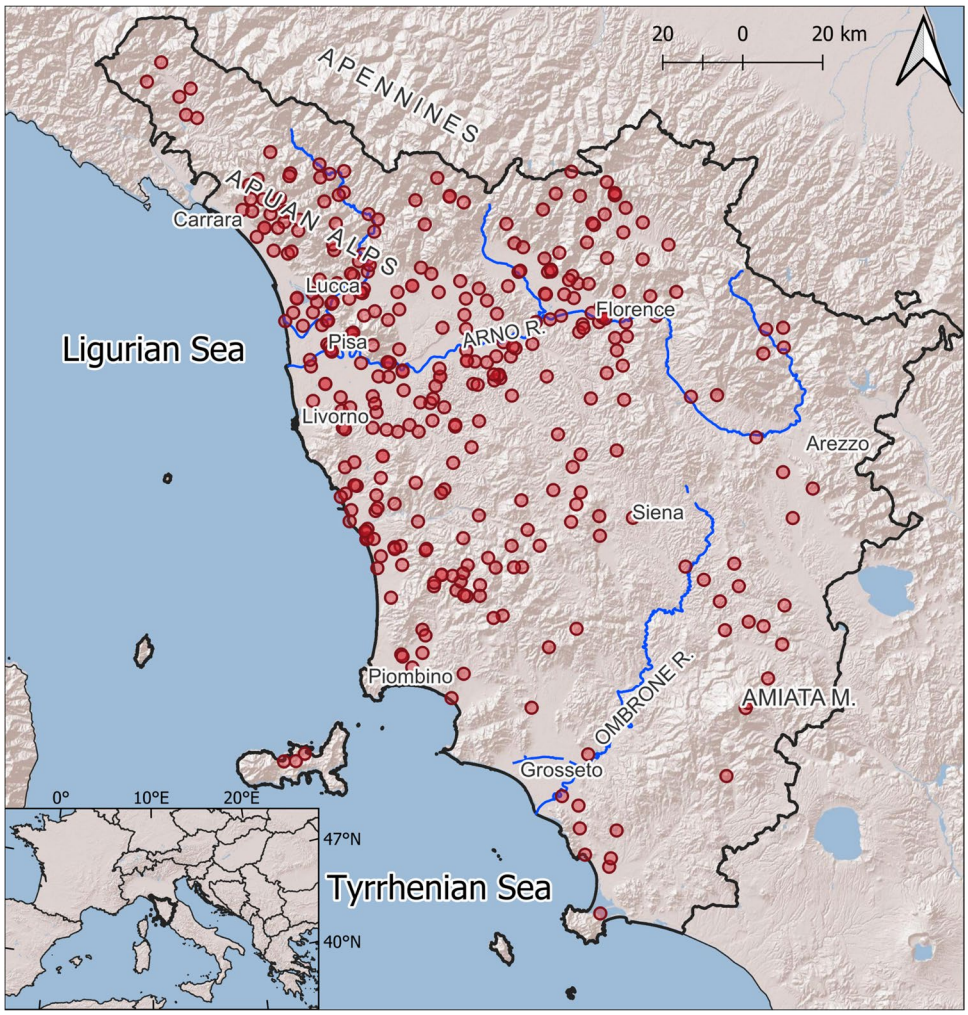

Fig. 1 Tuscany Region. The red points indicate the 349 raingauges managed by the Regional Hydrologic Service (SIR) and used in this work

mountain ranges, extensive hilly areas, and some relatively large plains (Carmignani et al. 2013; Baroni et al. 2015). In summary, the study area allows to apply the methodology and the investigations in an area characterized by a great climate variability and with a great number of raingauges. The manuscript is composed by the following paragraphs: material and methods, where we explain the methodology and the data used; results, where we show the products of this work; and finally, discussion and conclusions, where we analyse the results, and we propose the main consequences of this work.

Materials and methods

Database and data input pre-processing

The dataset used is provided by the Tuscany Region Hydrologic Service (SIR) and contains data acquired from several meteorological stations (https://www.sir.toscana.it/consistenza-rete). We collected the daily rainfall data by developing an automated download procedure through codes written in Python and HTTP protocol. The database derived from this procedure has also been used in different studies resulting reliable (Bini et al. 2021; Luppichini et al. 2021, 2022a, b, 2023b).

The monitoring activity in Tuscany started in 1910 and the entire rainfall dataset is today composed of 1103 time series. The number of raingauges increased from around 100 in the early 1900s to 350 in the 1940s, when the war slowed this growth. From the post-war period until the early 2000s the number of active raingauges was about 300 per year. In the last 20 years the network reached a peak of about 400 raingauges distributed on a region of almost 23,000 km^2. Each raingauge obviously has a different period of activity and some data may be missing within the time series. The territorial authority assigned a specific unique code to each sensor and when they move or change a sensor, they assign a different unique code. For this reason, each time series is assigned to a specific geographical condition and a specific sensor. The rainfall data used in this study are daily data, referring from 09:00

am to 09:00 am of the following day. The rainfall dataset is composed of validated and non-validated data (by SIR). The validated data is a subset derived from processing and checks that allow to remove any sampling errors and reduce the presence of inconsistency in the dataset. The non-validated data are raw measurements that have not yet been checked for integrity and correctness. In this work, we chose to use only the validated data to minimize errors. Some raingauges have time series with an insufficient amount of data available for the creation of a deep learning model. From tests carried out, we decided to use only raingauges which provide time series of at least six years. Each deep learning model predicts the missing data of the output raingauge using three (an arbitrary number) input raingauges. For each output raingauge, we chose the input raingauges based on the geographical distance and the difference in elevation between the sensors. The maximum distance and the maximum difference in altitude considered among the output raingauge and input raingauges was 10 km and 100 m, respectively. When more than three selected input raingauges were present, we did a manual screening choosing the best combination of input raingauges, representing a reasonable compromise between data completeness and the extensiveness of the dataset.

After this procedure, we selected 349 output raingauges distributed in the study area. Figure 1 shows how the density of the stations is higher in northwestern Tuscany than in southeastern one, due to the greater variability of rainfall, which needs a more effective monitoring network.

The mathematical expression of the models, representative of all the investigated raingauges, can be defined as:

$$\widehat{R} = f\left(X_t\right) = f(R1_t, R2_t, R3_t) \quad (1)$$

where R is the predicted output raingauge rainfall at time t; $R1_t$,$R2_t$ and $R3_t$ are the rainfall values of the three input raingauges at time t. An example of the input dataset is shown in Table 1.

Table 1 Example of data input: Input raingauge 1 (Empoli), Input raingauge 2 (San Miniato) and Input raingauge 3 (Vinci) (R1, R2, R3) are used to predict the rainfall of the Output raingauge (Cerreto Guidi) (R0)

Data	Input raingauge 1	Input raingauge 2	Input raingauge 3	Output raingauge
05/06/2013	0.2	0.0	5.8	0.0
06/06/2013	0.0	0.0	0.0	0.6
07/06/2013	0.0	0.0	0.0	0.0
08/06/2013	0.0	0.0	0.0	0.0
09/06/2013	4.4	4.6	4.8	0.0
24/02/2021	0.0	0.0	0.0	0.0
25/02/2021	0.0	0.0	0.0	0.4
26/02/2021	0.0	0.4	0.0	0.2
27/02/2021	0.0	0.0	0.0	0.0

Model development

To accomplish the deep learning models of this study, we mainly used the open-source framework Tensorflow (Abadi et al. 2015) and the libraries Numpy, Pandas, Scikit-Learn, and Keras (Chollet 2015) in Python language. The LR model is developed using the scikit-learn framework. We specifically selected two model architectures. The first one is composed of two CNNs layers of 64 and 128 filters (with a kernel size of 2 and stride length of 1), respectively, followed by a Max Pooling layer with size 1, an LSTM layer of 200 units, a dense layer of 50 neurons, and an output layer of one neuron (Fig. 2A). Usually, CNNs layers precede a pooling layer, which helps to reduce the size of the information while keeping the information unblemished. One pooling technique often used in CNN design is the Max Pooling (Zhou and Chellappa 1988). The second selected architecture is an encoder-decoder LSTM, with two LSTM nodes. Both the encoder and the decoder consist of a pair of sequence layers (LSTM) of 32 and 16 units followed by a repeat vector node for the encoder and 16 and 32 units for the decoder followed in turn by a time-distributed dense node (Fig. 2B). Each first LSTM layer returns the whole output sequence to the second one, instead the last ones return only the last hidden state. To evaluate the discrepancy between predicted and actual values, we used a loss function measured on each observation, which allowed us to calculate the cost function. We needed to minimize the cost function by identifying the optimized values for each weight. Thanks to multiple iterations, the optimization algorithms transfer the identification of the weights that minimize the cost function. In our implementation, we used the Adam optimizer (Kingma and Ba 2014), which is an adaptive learning speed method, namely it computes individual learning rates for several parameters (Kingma and Ba 2014). The activation function used for CNN-LSTM and ED-LSTM models is rectified linear units (ReLU) function (Agarap 2018). To stop the training, we used API of Keras and specifically the "early stopping" method, setting a number of epochs with no improvement after which the training is stopped at 200. This method allows the training procedure to stop when the monitored metric has stopped improving. The monitored metric was the value of the cost function. Given all the possible hypotheses, we wanted to find the best one (called "optimal"), namely the one that allowed us to make more precise estimates, always based on data in our possession. For each model, the input dataset is divided into three subsets called training, validation, and test datasets. The training and validation datasets are used

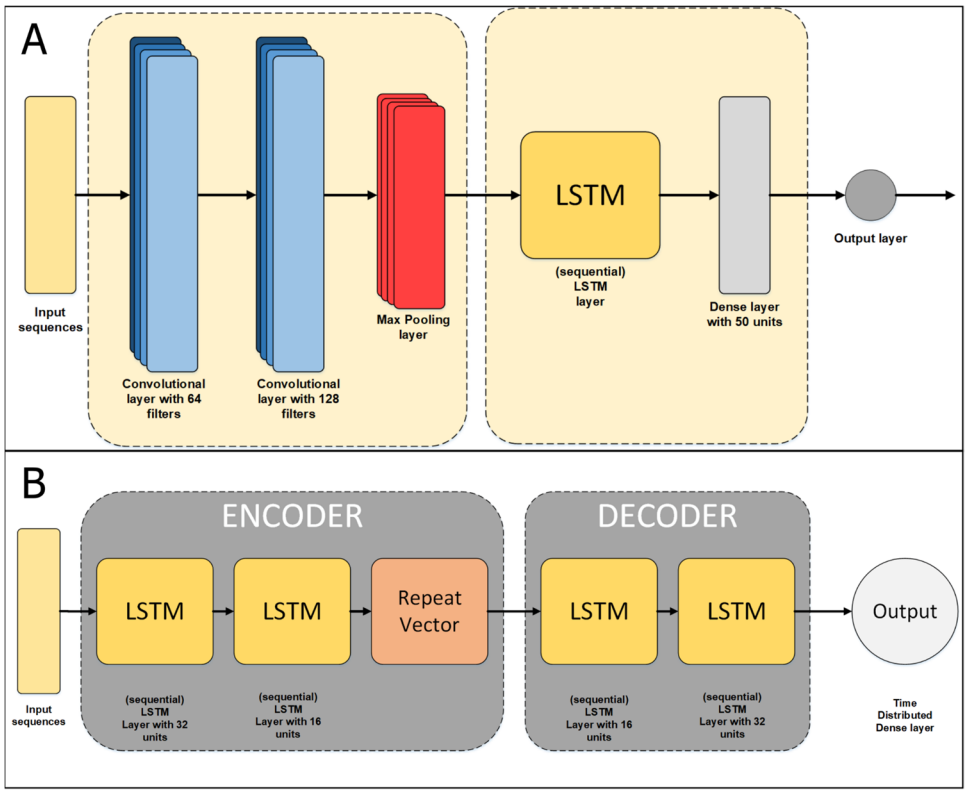

Fig. 2 Architecture of the deep learning models used in this study. **A**) CNN-LSTM; **B**) ED-LSTM

during the learning phase. The test dataset is used afterwards to evaluate the quality of the model. In this way, we can determine the ability of the model to predict new cases not used during the learning phase. The training dataset is 60% of the primary dataset, whereas the test and validate datasets include the remaining 20% and 20%, respectively. This type of splitting is commonly used in the supervised training of deep learning models, allowing sufficient data for training and model quality verification (Gholami et al. 2015). The models were fitted using a division for batches. For the training of the LR models we used the same training and test datasets used to the CNN-LSTM and ED-LSTM models. In this way, we can compare the results of the three models. In our models, the cost function used was the mean absolute error (MAE) calculated on the training dataset during the resolution of each batch and on the validation dataset at the end of each epoch. This procedure allowed to minimize the overfitting effect on the training set.

Evaluation of models

Each model is associated with some errors, the evaluation of which provides information on the performance of the model itself. In this work, we used the Mean Absolute Error (MAE). The MAE is an arithmetic mean of the absolute errors, and is one of the methods used to assess the model performance (Willmott and Matsuura 2005):

$$MAE = \frac{\sum_{t=1}^{n} \left| \widehat{R}_t - R_{0_t} \right|}{n} \tag{2}$$

In addition to the MAE, the average relative error was calculated for each raingauge:

$$RMAE = \frac{1}{n} \times \sum_{t=1}^{n} \left| \frac{\widehat{R}_t - R_{0_t}}{\widehat{R}_t} \right| \tag{3}$$

The models were moreover evaluated by the parameter R2, an index measuring the link between the variability of data and correctness of the statistical model used. Another method was used to estimate the performance of the models, which also made it possible to try to understand the cause of the errors of the prediction models. This method consists of taking the models errors for each raingauge:

$$X = \widehat{R}_t - R_{0_t} \tag{4}$$

and correlating them with the value derived from the average of the rainfall at the three input raingauges (R_1, R_2,R_3) minus the rainfall amount of the output raingauge (R_0) for the same day of the model errors:

$$Y = \left(\frac{R_{1_t} + R_{2_t} + R_{3_t}}{3} \right) - R_{0_t} \quad (5)$$

To get this correlation, the Pearson Correlation Coefficient (PCC, Kirch 2008), which highlights any linear relationship two statistical variables, was used:

$$\rho XY = \frac{\sigma XY}{\sigma X \sigma Y} \quad (6)$$

According to the Cauchy–Schwarz inequality, PCC ranges between + 1 and -1, where + 1 corresponds to perfect positive linear correlation, 0 corresponds to no correlation, and -1 corresponds to perfect negative linear correlation (Lee Rodgers and Alan Nice Wander 1988).

Results

Figure 3A shows the boxplots of MAE of the deep learning models of the 349 raingauges time series simulated. The median of the errors is about 3 mm for CNN-LSTM and ED-LSTM models, while LR model has a median of about 1 mm.

The R^2 values are reported in Fig. 3B in which we can recognize that the two deep learning models have similar R2, while LR model shows the best values of error metric. Figure 3C shows the average relative error for six daily rainfall bands considering all raingauges (0–1 mm, 1–3 mm, 3–5 mm, 5–10 mm, 10–30 mm, 30–50 mm). The figure shows the relative errors for the three model architectures are also very similar in these cases.

The analysis of the spatial distribution of errors does not denote a clustering or a specific spatial distribution (a case of MAE of CNN-LSTM models is reported in Fig. 4).

The errors on the training and validation dataset are monitored during the training time, but we can compare the MAE calculated on the test and validation dataset (training dataset for the LR model; Fig. 5), during the post-training phase, to evaluate if the models are not subject to overfitting. The MAEs calculated on the validation dataset and on the test dataset are comparable; the difference is present in a small range around 0, indicating a low degree of overfitting (Fig. 5).

Model errors are higher when the difference among input and output data is higher. Each model shows a strong correlation (median PCC values of 0.9) between the difference among the input and output values and the absolute error (Fig. 6).

Discussion

The errors of CNN and ED architectures are very similar, but the LR model is the best. If we compare the MAE of the two architectures for each raingauge, we can observe that they almost overlap (Fig. 7). The correlation between the two errors by the Pearson method gives a value of 0.93.

Comparing the errors of the models proposed in this study with those derived from other AI-based works or different approaches (e.g., mathematical, statistical) is

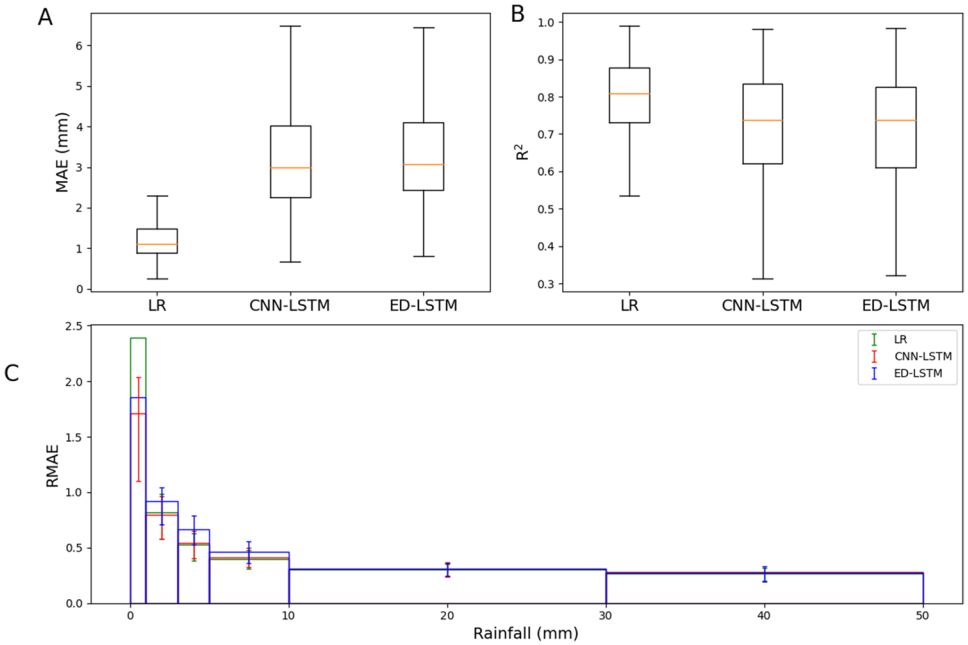

Fig. 3 Absolute and relative errors of CNN-LSTM and ED-LSTM models: **A**) Mean absolute error (MAE); **B**) R^2; **C**) Relative Mean Absolute Error (RMAE). The boxes represent the interval between the 25th and 75th percentiles (Q1 and Q3). IQR is the interquartile range Q3-Q1. The upper whisker will extend to the last datum lower than Q3 + 1.5 × IQR. Similarly, the lower whisker will reach the first datum higher than Q1 – 1.5 × IQR

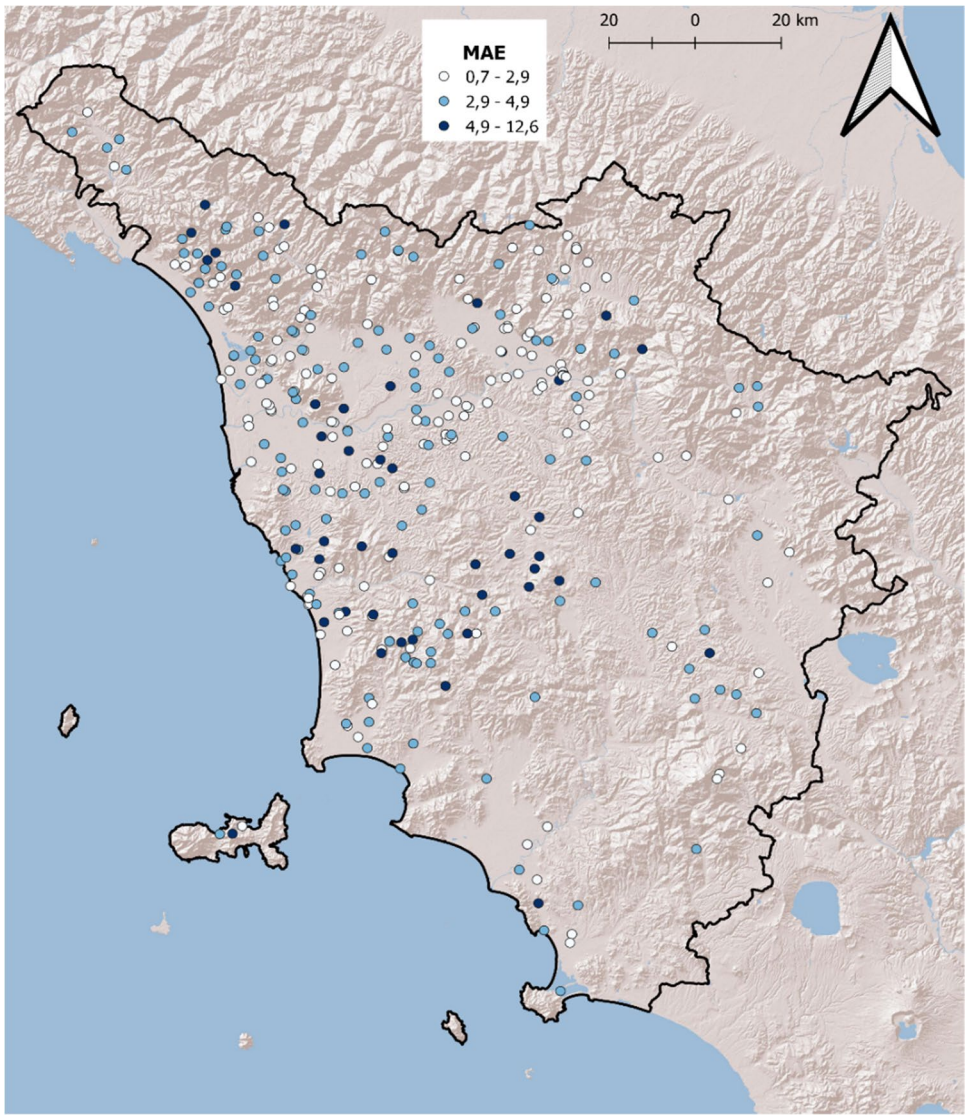

Fig. 4 Spatial location of Mean Absolute Errors (MAE) of the CNN-LSTM

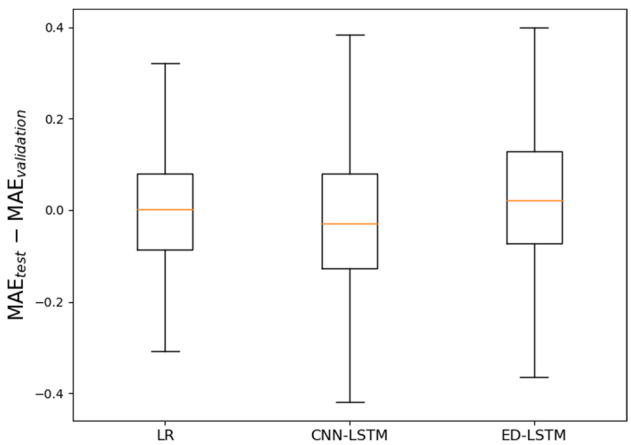

Fig. 5 Difference in MAEs calculated on the test dataset and on the validation dataset for CNN-LSTM and ED-LSTM while for LR the MAEs are calculated using test dataset and training dataset

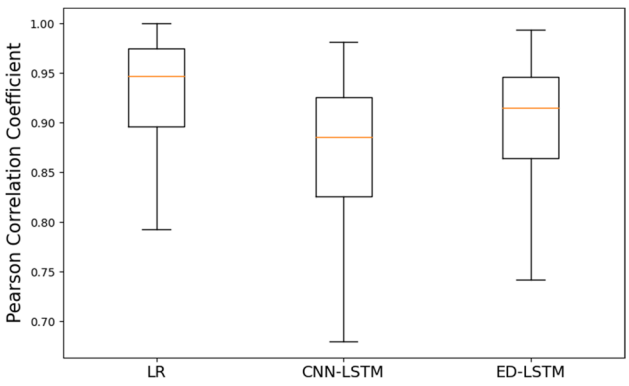

Fig. 6 Boxplots represent the distribution of Pearson Correlation Coefficient calculated by comparing the absolute errors and the difference at the same time (t) of the mean input rainfall and the rainfall output

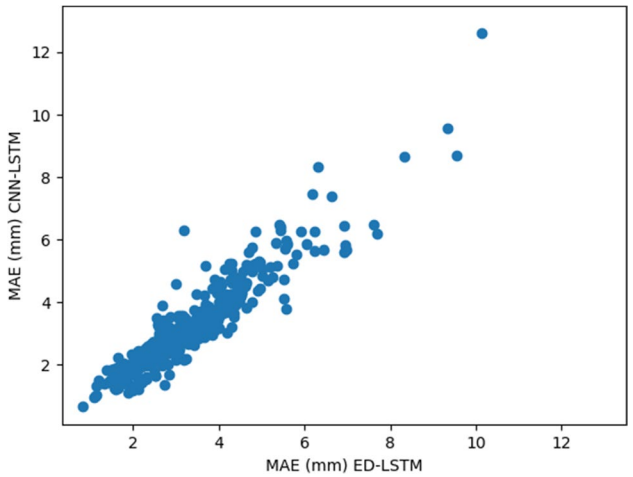

Fig. 7 The plot shows the ratio between the MAE obtained for the two types of model architectures

very complex. The differences depend on several factors such as the features of the study area, the spatial distribution of the raingauges, and the rainfall distribution and amount (de Silva et al. 2007). However, considering this, we compared our model errors with those derived from other methods (Beauchamp et al. 1989; Gyau-Boakye and Schultz 1994; Abebe and Price 2003; Coulibaly and Evora 2007; Caldera et al. 2016; Balcha et al. 2023), recognizing a reasonable comparability. For example, Coulibaly and Evora (2007) compared six ANN structures to predict missing precipitation data using three input raingauges, and their models have MAE in the range 1.5–2 mm and an R^2 of 0.75. On the other hand, using statistical methods Balcha et al. (2023) obtained a MAE ranging from 2 to 8 mm. Analyzing these cases, we recognized that AI techniques could have higher accuracy than traditional methods. This result could be attributed to the non-stationary behaviour of rainfall models and to the capacity of AI models to work with not linear relations (Creutin et al. 1997).

We can divide the models into two groups considering the spatial relationship between input and output raingauges: group A is composed of the cases for which the output raingauges are located out of the triangle; group B is composed of the cases for which the output raingauges are spatially located in the triangle with the input raingauges as vertices (see Fig. 8). The analysis of the two cases allows us to understand if the MAE values are related to these two configurations of the input and output. Group A counts 26% of the raingauges, whereas group B includes the remaining 74%. Both groups have an average MAE of about 3.2 mm, and the percentage of raingauges with an error greater than 5 mm is 12% for both cases. This analysis shows that the model errors are not related to the two types of relative positions between the output and input raingauges.

The errors of the model are strongly correlated with the difference between the mean rainfall of the input raingauges and the value of the output raingauges (Eq. 5). The greater the difference between input and output data, the greater is the model prediction error, as shown in Fig. 6. Deepening this result (de Silva et al. 2007) in each of the time series used, we can find some records exhibiting a high difference between the input and output values at the same time *t*. In other cases, these differences are also in the input dataset. For each time series, we identified and counted the records in which one of the following conditions occurs:

i) the output raingauge measured more than 5 mm (rainy day) and all three input raingauges measured 0 mm (no rainy day);
ii) the output raingauge measured 0 mm (no rainy day) and the input raingauges measured more than 5 mm (rainy day);
iii) the output raingauge recorded 0 mm and the average rainfall of the three input raingauges is greater than 5 mm (rainy day).

The procedure highlighted that each time series has a number of these cases, variable with a maximum number of more than 8%. We defined these as anomalous cases because it is very complex to understand the causes of these measured differences. We remember that the data used in this work are validated by the SIR, the

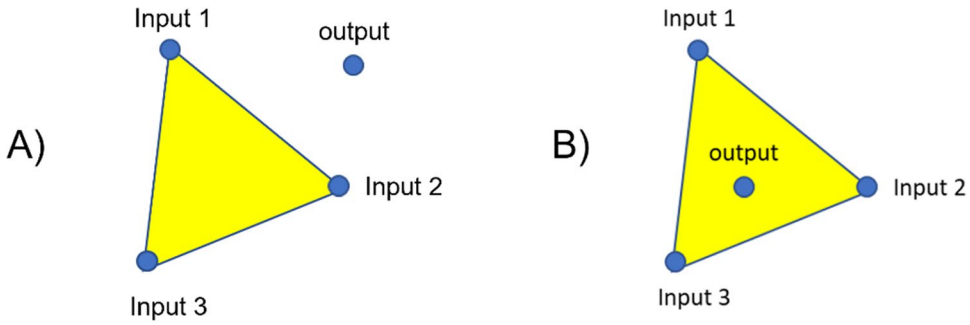

Fig. 8 Relation between input raingauges area (yellow) and respective output raingauges outside (**A**) and inside (**B**)

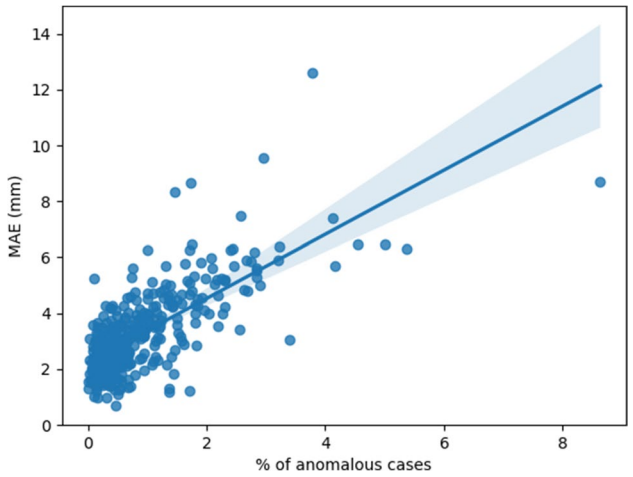

Fig. 9 Scatterplot showing the ratio between anomalous data amount in the 349-time series and MAE of the CNN-LSTM models

meteorological service which provided them. We cannot assert that these records are wrong because we cannot exclude a meteorological factor influencing the measure. At the same time, we are not sure that the cause is, at least completely, due to meteorological conditions because the stations are spatially very near and with a little difference in altitude.

The models show a higher error concerning the number of anomalous cases (Fig. 9). This relation is quantified by a PCC value of 0.71 for both the used architectures. This result highlights the data-driven behaviour of AI techniques and can be used to emerge these particular cases from the database.

This research will allow the creation of a control procedure on the time series to improve their knowledge and understand if the causes are meteorological or instrumental and improve the management of the database.

Conclusions

This work demonstrates that deep learning models can predict rainfall values using the time series of nearby raingauges as input. The errors of the models are comparable to those obtained by other works which used similar or different techniques. These models can be applied in the analysis of the rainfall time series, for instance, to compute the missing data. This problem is one of the main issues that afflict the meteorological time series, such as other types of environmental monitoring parameters.

However, this study also demonstrated that the deep learning model performances are strongly influenced by the input data, confirming the data-driven behaviour of these techniques. Major errors correspond to major differences among the input data or with the output values. The causes of these anomalies can be different. We cannot exclude meteorological factors, but we suppose that the main cause could be linked to raingauges malfunctions, given the specific selection of the input. This problem can affect all rainfall networks and improving knowledge on the identification of these anomalous data can allow a better management of the measurement network and validation procedures.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1007/s12145-023-01122-4.

Acknowledgements The authors are grateful to the Tuscan Regional Hydrologic Service for providing the data used in this work.

Author contributions Conceptualization, M.L., Mi.Ba, and R.G.; methodology, M.L. and Mi.Ba; software, A.L., M.L. and Mi.Ba; validation, A.L., M.L. and M.B.; formal analysis, A.L.; investigation, A.L., M.L. and Mi.Ba; resources, Mo.Bi and R.G…; data curation, A.L. and M.L..; writing—original draft preparation, A.L and M.L..; writing—review and editing,. Mi.Ba, Mo.Bi, R.G.; visualization, Mo.Bi and R.G..; supervision, R.G.; project administration, Mo.Bi and R.G..; funding acquisition, Mo.Bi and R.G.. All authors have read and agreed to the published version of the manuscript.

Funding Open access funding provided by Università di Pisa within the CRUI-CARE Agreement. This research was funded by project CUP F54J16000020001 "Autorità di Bacino Distrettuale dell'Appennino Settentrionale – Misure di prevenzione tese a supportare ed ottimizzare la pianificazione di gestione, la programmazione e realizzazione degli interventi di cui al PGRA" (Resp. M. Bini and R. Giannecchini) and by the project: "Valutazione di scenari dei deflussi superficiali di alcuni selezionati corpi idrici superficiali, basandosi su una analisi dettagliata di banche date meteorologiche e l'applicazione di tecniche Intelligenza Artificiale" Consorzio di Bonifica Toscana Nord, (Resp. M. Bini and R. Giannecchini).

Data availability You can contact Marco Luppichini (marco.luppichini@dst.unipi.it) for data and materials.

Declarations

Competing interests The authors declare no competing interests.

References

Abadi M, Agarwal A, Barham P, Brevdo E, Chen Z, Citro C, Corrado GS, Davis A, Dean J, Devin M, Ghemawat S, Goodfellow I, Harp A, Irving G, Isard M, Jia Y, Jozefowicz R, Kaiser L, Kudlur, M, … Zheng X (2015) TensorFlow: large-scale machine learning on heterogeneous systems. http://tensorflow.org/

Abebe AJ, Price RK (2003) Managing uncertainty in hydrological models using complementary models. Hydrol Sci J 48(5):679–692. https://doi.org/10.1623/hysj.48.5.679.51450

Agarap AF (2018) Deep learning using rectified linear units (relu). ArXiv Preprint ArXiv:1803.08375

Alem AM, Tilahun SA, Moges MA, Melesse AM (2019) Chapter 9 - A regional hourly maximum rainfall extraction method for part of Upper Blue Nile Basin, Ethiopia. In: Melesse AM, Abtew W, Senay G (eds) Extreme Hydrology and Climate Variability. Elsevier, pp 93–102

Amanambu AC, Obarein OA, Mossa J et al (2020) Groundwater system and climate change: Present status and future considerations. J Hydrol (Amst) 589:125163. https://doi.org/10.1016/j.jhydrol.2020.125163

Antonetti M, Zappa M (2018) How can expert knowledge increase the realism of conceptual hydrological models? A case study based on the concept of dominant runoff process in the Swiss Pre-Alps. Hydrol Earth Syst Sci 22:4425–4447. https://doi.org/10.5194/hess-22-4425-2018

Baek S-S, Pyo J, Chun JA (2020) Prediction of water level and water quality using a CNN-LSTM combined deep learning approach. Water 12(12). https://doi.org/10.3390/w12123399

Balcha SK, Hulluka TA, Awass AA, Bantider A, Ayele GT (2023) Comparison and selection criterion of missing imputation methods and quality assessment of monthly rainfall in the Central Rift Valley Lakes Basin of Ethiopia. Theor Appl Climatol 154(1):483–503. https://doi.org/10.1007/s00704-023-04569-z

Baroni C, Pieruccini P, Bini M, Coltorti M, Fantozzi PL, Guidobaldi G, Nannini D, Ribolini A, Salvatore MC (2015) Geomorphological and neotectonic map of the Apuan Alps (Tuscany, Italy). Geografia Fisica e Dinamica Quaternaria 38(2):201–227. https://doi.org/10.4461/GFDQ.2015.38.17

Beauchamp JJ, Downing DJ, Railsback SF (1989) Comparison of regression and time-series methods for synthesizing missing streamflow records. JAWRA J Am Water Resour Assoc 25(5):961–975. https://doi.org/10.1111/j.1752-1688.1989.tb05410.x

Bengio Y, Courville A, Vincent P (2013) Representation Learning: A Review and New Perspectives. IEEE Trans Pattern Anal Mach Intell 35:1798–1828. https://doi.org/10.1109/TPAMI.2013.50

Bini M, Casarosa N, Luppichini M (2021) Exploring the relationship between river discharge and coastal erosion: An integrated approach applied to the pisa coastal plain (italy). Remote Sens 13(2). https://doi.org/10.3390/rs13020226

Boulmaiz T, Guermoui M, Boutaghane H (2020) Impact of training data size on the LSTM performances for rainfall–runoff modeling. Model Earth Syst Environ 6:2153–2164. https://doi.org/10.1007/s40808-020-00830-w

Caldera HPGM, Piyathisse VRPC, Nandalal KDW (2016) A comparison of methods of estimating missing daily rainfall data. Engineer: Journal of the Institution of Engineers, Sri Lanka, 49(4):1-8. https://doi.org/10.4038/engineer.v49i4.7232

Cantù V (1977) The climate of Italy. In: Wallen CC (ed) Climate of central and southern Europe. Elsevier, pp 127–184

Carmignani L, Conti P, Cornamusini G, Pirro A (2013) Geological map of Tuscany (Italy). J Maps 9:487–497. https://doi.org/10.1080/17445647.2013.820154

Chattopadhyay A, Nabizadeh E, Hassanzadeh P (2020) Analog forecasting of extreme-causing weather patterns using deep learning. J Adv Model Earth Syst 12(2):e2019MS001958. https://doi.org/10.1029/2019MS001958

Chollet F (2015) Keras. GitHub. https://github.com/fchollet/keras

Coulibaly P, Evora ND (2007) Comparison of neural network methods for infilling missing daily weather records. J Hydrol 341(1):27–41. https://doi.org/10.1016/j.jhydrol.2007.04.020

Creutin JD, Andrieu H, Faure D (1997) Use of a weather radar for the hydrology of a mountainous area. Part II: Radar measurement validation. J Hydrol 193(1):26–44. https://doi.org/10.1016/S0022-1694(96)03203-9

De Luca DL, Napolitano F (2023) A user-friendly software for modelling extreme values: EXTRASTAR (EXTRemes Abacus for STAtistical Regionalization). Environ Modell Softw 161:105622. https://doi.org/10.1016/j.envsoft.2023.105622

De Silva RP, Dayawansa NDK, Ratnasiri MD (2007) A comparison of methods used in estimating missing rainfall data. J Agric Sci - Sri Lanka 3(2):101–108. https://doi.org/10.4038/jas.v3i2.8107

Fawaz HI, Forestier G, Weber J, Fawaz HI, Forestier G, Weber J, Idoumghar L, Muller P (2020) Deep learning for time series classification : a review To cite this version : HAL Id : hal-02365025 Deep learning for time series classification : a review

Fratianni S, Acquaotta F (2017) The Climate of Italy. In: Soldati M, Marchetti M (eds) Landscapes and Landforms of Italy. Springer International Publishing, Cham, pp 29–38

Gers FA, Eck D, Schmidhuber J (2001) Applying LSTM to time series predictable through time-window approaches. https://doi.org/10.1007/3-540-44668-0_93

Gholami V, Chau KW, Fadaee F et al (2015) Modeling of groundwater level fluctuations using dendrochronology in alluvial aquifers. J Hydrol (amst) 529:1060–1069. https://doi.org/10.1016/j.jhydrol.2015.09.028

Gyau-Boakye P, Schultz GA (1994) Filling gaps in runoff time series in west africa. Hydrol Sci J 39(6):621–636. https://doi.org/10.1080/02626669409492784

Hardwick Jones R, Westra S, Sharma A (2010) Observed relationships between extreme sub-daily precipitation, surface temperature, and relative humidity. Geophys Res Lett 37(22). https://doi.org/10.1029/2010GL045081

Hu Y, Yan L, Hang T, Feng J (2020) Stream-flow forecasting of small rivers based on LSTM

Huang C, Zhang J, Cao L et al (2020) Robust Forecasting of River-Flow Based on Convolutional Neural Network. IEEE Transactions on Sustainable Computing 5:594–600. https://doi.org/10.1109/TSUSC.2020.2983097

Hussain D, Hussain T, Khan A et al (2020) A deep learning approach for hydrological time-series prediction: A case study of Gilgit river basin. Earth Sci Inform 13:1–13. https://doi.org/10.1007/s12145-020-00477-2

IPCC (2019) Climate Change and Land: an IPCC special report on climate change, desertification, land degradation, sustainable land management, food security, and greenhouse gas fluxes in terrestrial ecosystems [Shukla PR, Skea J, Calvo Buendia E, Masson-Delmotte V, Pörtner H-O, Roberts DC, Zhai P, Slade R, Connors S, van Diemen R, Ferrat M, Haughey E, Luz S, Neogi S, Pathak M, Petzold J, Portugal Pereira J, Vyas P, Huntley E, Kissick K, Belkacemi M, Malley J (eds.)]. In press

Kim DY, Song CM (2020) Developing a discharge estimation model for ungauged watershed using CNN and hydrological image. Water 12(12). https://doi.org/10.3390/w12123534

Kimura N, Yoshinaga I, Sekijima K et al (2019) Convolutional Neural Network Coupled with a Transfer-Learning Approach for

Time-Series Flood Predictions. Water (basel) 12:96. https://doi.org/10.3390/w12010096
Kingma DP, Ba J (2014) Adam: a method for stochastic optimization
Kirch W (ed) (2008) Pearson's correlation coefficient. In: Encyclopedia of Public Health. Springer, Dordrecht. https://doi.org/10.1007/978-1-4020-5614-7_2569
Kratzert F, Klotz D, Brenner C, Schulz K, Herrnegger M (2018) Rainfall – runoff modelling using Long Short-Term Memory (LSTM) networks, pp 6005–6022
Le XH, Ho H, Lee G, Jung S (2019) Application of Long Short-Term Memory (LSTM) Neural Network for Flood Forecasting. Water (basel) 11:1387. https://doi.org/10.3390/w11071387
Lebay M, Le M (2020) Edition 1 | Article 1036 ScienceForecast Publications LLC., | https: Citation: Egigu ML. Techniques of filling missing values of daily and monthly rain fall data: a review. SF Journal of Environmental and Earth Science 3:1036
Lee Rodgers J, Alan Nice Wander W (1988) Thirteen ways to look at the correlation coefficient. Am Stat 42(1):59–66. https://doi.org/10.1080/00031305.1988.10475524
Li W, Kiaghadi A, Dawson C (2020) High temporal resolution rainfall–runoff modeling using long-short-term-memory (LSTM) networks. Neural Comput Appl. https://doi.org/10.1007/s00521-020-05010-6
Li J, Zhou Y, Wang W et al (2022) Response of hydrogeological processes in a regional groundwater system to environmental changes: A modeling study of Yinchuan Basin. China. J Hydrol (Amst) 615:128619. https://doi.org/10.1016/j.jhydrol.2022.128619
Li X, Du Z, Song G (2018) A method of rainfall runoff forecasting based on deep convolution neural networks. In: 2018 Sixth International Conference on Advanced Cloud and Big Data (CBD). pp 304–310. https://doi.org/10.1109/CBD.2018.00061
Liu D, Jiang W, Mu L, Wang S (2020) Streamflow Prediction Using Deep Learning Neural Network: Case Study of Yangtze River. IEEE Access 8:90069–90086. https://doi.org/10.1109/ACCESS.2020.2993874
Livieris IE, Pintelas E, Pintelas P (2020) A CNN–LSTM model for gold price time-series forecasting. Neural Comput Appl 32:17351–17360. https://doi.org/10.1007/s00521-020-04867-x
Luppichini M, Barsanti M, Giannecchini R, Bini M (2021) Statistical relationships between large-scale circulation patterns and local-scale effects: NAO and rainfall regime in a key area of the Mediterranean basin. Atmos Res 248:105270
Luppichini M, Barsanti M, Giannecchini R, Bini M (2022a) Deep learning models to predict flood events in fast-flowing watersheds. Sci Total Environ 813:151885. https://doi.org/10.1016/j.scitotenv.2021.151885
Luppichini M, Bini M, Barsanti M et al (2022b) Seasonal rainfall trends of a key Mediterranean area in relation to large-scale atmospheric circulation: How does current global change affect the rainfall regime? J Hydrol (Amst) 612:128233. https://doi.org/10.1016/j.jhydrol.2022.128233
Luppichini M, Bini M, Giannecchini R (2023a) CleverRiver: an open source and free Google Colab toolkit for deep-learning river-flow models. Earth Sci Inform. https://doi.org/10.1007/s12145-022-00903-7
Luppichini M, Bini M, Giannecchini R, Zanchetta G (2023b) High-resolution spatial analysis of temperature influence on the rainfall regime and extreme precipitation events in north-central Italy. Sci Total Environ 880:163368. https://doi.org/10.1016/j.scitotenv.2023.163368
Malhi Y, Franklin J, Seddon N, Solan M, Turner MG, Field CB, Knowlton N (2020) Climate change and ecosystems: Threats, opportunities and solutions. In: Philosophical Transactions of the Royal Society B: Biological Sciences (vol 375, issue 1794). Royal Society Publishing. https://doi.org/10.1098/rstb.2019.0104
Marçais J, de Dreuzy J-R (2017) Prospective Interest of Deep Learning for Hydrological Inference. Groundwater 55:688–692. https://doi.org/10.1111/gwat.12557
Minoli S, Jägermeyr J, Asseng S et al (2022) Global crop yields can be lifted by timely adaptation of growing periods to climate change. Nat Commun 13:7079. https://doi.org/10.1038/s41467-022-34411-5
Myhre G, Alterskjær K, Stjern CW et al (2019) Frequency of extreme precipitation increases extensively with event rareness under global warming. Sci Rep 9:16063. https://doi.org/10.1038/s41598-019-52277-4
Nguyen DH, Bae D-H (2020) Correcting mean areal precipitation forecasts to improve urban flooding predictions by using long short-term memory network. J Hydrol (Amst) 584:124710. https://doi.org/10.1016/j.jhydrol.2020.124710
Nigro M, Ambrosio M, Fagioli MT, Curcio C, Giannecchini R (2022) Analysis of fragmented piezometric levels records: the ARTE (Antecedent Recharge Temporal Effectiveness) approach. Acque Sotterranee - Italian Journal of Groundwater 11(4):21–32. https://doi.org/10.7343/as-2022-566
Rapetti F, Vittorini S (1994) Le precipitazioni in Toscana: osservazioni sui casi estremi. Riv Geogr Ital 101:47–76
Sattari MT, Rezazadeh-Joudi A, Kusiak A (2017) Assessment of different methods for estimation of missing data in precipitation studies. Hydrol Res 48:1032–1044. https://doi.org/10.2166/nh.2016.364
Schmidt G (2011) Climate change and climate modeling. Eos, Transactions American Geophysical Union 92(23):198–199. https://doi.org/10.1029/2011eo230012
Sit M, Demiray BZ, Xiang Z et al (2020) A comprehensive review of deep learning applications in hydrology and water resources. Water Sci Technol. https://doi.org/10.2166/wst.2020.369
Sutskever I, Vinyals O, Le Qv (2014) Sequence to sequence learning with neural networks. http://arxiv.org/abs/1409.3215
Tien Bui D, Hoang N-D, Martínez-Álvarez F et al (2020) A novel deep learning neural network approach for predicting flash flood susceptibility: A case study at a high frequency tropical storm area. Sci Total Environ 701:134413. https://doi.org/10.1016/j.scitotenv.2019.134413
Tramblay Y, Llasat MC, Randin C, Coppola E (2020) Climate change impacts on water resources in the Mediterranean. Reg Environ Change 20:83. https://doi.org/10.1007/s10113-020-01665-y
Van SP, Le HM, Thanh DV et al (2020) Deep learning convolutional neural network in rainfall–runoff modelling. J Hydroinf 22:541–561. https://doi.org/10.2166/hydro.2020.095
van Loon H, Williams J (1976) The Connection Between Trends of Mean Temperature and Circulation at the Surface: Part I. Winter Mon Weather Rev 104:365–380. https://doi.org/10.1175/1520-0493(1976)104%3c0365:TCBTOM%3e2.0.CO;2
Willmott CJ, Matsuura K (2005) Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. Clim Res 30:79. https://doi.org/10.3354/cr030079
Xu W, Jiang Y, Zhang X et al (2020) Using long short-term memory networks for river flow prediction. Hydrol Res 51:1358–1376. https://doi.org/10.2166/nh.2020.026
Yi A, Li Z, Gan M et al (2019) A deep learning approach on short-term spatiotemporal distribution forecasting of dockless bike-sharing system. Neural Comput Appl 31:1–13. https://doi.org/10.1007/s00521-018-3470-9
Yin Y, Chen H, Xu C-Y et al (2016) Spatio-temporal characteristics of the extreme precipitation by L-moment-based index-flood method in the Yangtze River Delta region, China. Theor Appl Climatol 124:1005–1022. https://doi.org/10.1007/s00704-015-1478-y
Zheng J, Fu X, Zhang G (2019) Research on Exchange Rate Forecasting Based on Deep Belief Network. Neural Comput Appl 31:573–582. https://doi.org/10.1007/s00521-017-3039-z

Zhou YT, Chellappa R (1988) Computation of optical flow using a neural network. In: IEEE 1988 International Conference on Neural Networks, pp 71–78. https://doi.org/10.1109/ICNN.1988.23914