METEOROLOGY HYDROLOGY AND WATER MANAGEMENT
IMGW-PIB Science Journal since 2013

VOL. 11 | ISSUE 2 | 2023

ISNN 2299-3835

Comparison and combination of interpolation methods for daily precipitation in Poland: evaluation using the correlation coefficient and correspondence ratio

Krystyna Konca-Kędzierska, Marta Gruszczyńska

Institute of Meteorology and Water Management – National Research Institute

Joanna Wibig

University of Lodz, Department of Meteorology and Climatology

Abstract

Interpolation of precipitation data is a common practice for generating continuous, spatially-distributed fields that can be used for a range of applications, including climate modeling, water resource management, and agricultural planning. To obtain the reference field, daily observation data from the measurement network of the Institute of Meteorology and Water Management – National Research Institute was used. In this study, we compared and combined six different interpolation methods for daily precipitation in Poland, including bilinear and bicubic interpolation, inverse distance weighting, distance-weighted average, nearest neighbor remapping, and thin plate spline regression. Implementations of these methods available in the R programming language (e.g., from packages akima, gstat, fields) and the Climate Data Operators (CDO) were applied. The performance of each method was evaluated using multiple metrics, including the Pearson correlation coefficient (RO) and the correspondence ratio (CR), but there was no clear optimal method. As an interpolated resulting field, a field consisting of the best interpolations for individual days was proposed. The assessment of daily fields was based on the CR and RO parameters. Our results showed that the combined approach outperformed individual methods with higher accuracy and reliability and allowed for generating more accurate and reliable precipitation fields. On a group of selected stations (data quality and no missing data), the precipitation result fields were compared with the fields obtained in other projects-CPLFD-GDPT5 (Berezowski et al. 2016) and G2DC-PLC (Piniewski et al. 2021). The variance inflation factor (*VIF*) was bigger for the resulting fields (~5), while for the compared fields, it was below 3. However, for the mean absolute error (*MAE*), the relationship was reversed – the *MAE* was approximately half as low for the fields obtained in this work.

Keywords

Precipitation, interpolation methods, daily gridded data, validation of precipitation gridded data set, observational data.

Submitted 30 December 2022, revised 22 June 2023, accepted 30 August 2023

DOI: 10.26491/mhwm/171699

1. Introduction

Climate change is undeniably confirmed in observational data, regardless of the perspective from which the data is analyzed (FAOSTAT 2022; NOAA 2022). Recognition of the necessity for adaptation and mitigation strategies is evident among governments and societies, a fact substantiated by the initiatives undertaken by the United Nations Framework Convention on Climate Change (UNFCCC). These strategies are rooted in climate data, and their effectiveness hinges on the precision of climate change forecasts. Certain existing climate scenarios fail to depict climate fluctuations for specific regions without accounting for localized conditions.

Precipitation is a meteorological phenomenon characterized by high spatiotemporal variability; unlike temperature or air pressure, this parameter is discontinuous. This makes interpolation or forecasting of precipitation challenging. This work attempts to create a reference field for Poland for the process of adjusting climate scenarios.

Based on The Fifth Assessment Report (IPCC 2014) prepared by the Intergovernmental Panel on Climate Change (IPCC), an international program, the Coordinated Regional Downscaling Experiment (CORDEX) (Giorgi et al. 2009) was established in the frame of the World Climate Research Program (WRCP). The CORDEX aimed to organize and coordinate a framework to produce improved regional climate change projections for 14 regional (CORDEX) domains. Among these 14 domains is EUR-11, which covers Europe. For this region, climate simulations were prepared and developed by the European branch of the international CORDEX initiative (EURO-CORDEX) (Jacob et al. 2014; Benestad et al. 2021). This project offers hindcast simulations, historical simulations, and climate scenarios (Benestad et al. 2017). Hindcast simulations cover 1989-2008, using initial and boundary data from global atmospheric reanalysis produced by the European Centre for Medium-Range Weather Forecasts (ERA-Interim reanalysis) (Dee et al. 2011). They are made for model makers to evaluate the model design and improve it. Historical simulations cover the period of 1951-2005. Regional Climate Models (RCM) use the initial and lateral boundary conditions from Global Climate Models (GCM), and their purpose is to analyze the correctness of the RCM-GCM scheme for a given region. Due to the higher resolution of the RCM model, it is possible to feed the RCM-GCM scheme with a more detailed description of the local environment. For example, height above sea level and terrain type are provided more accurately. As a result, this procedure should make the simulation results closer to real results. Various physical parametrizations and numerical methods are used in RCMs, which means that the quality of past climate reconstructions may change depending on the studied area and model. Comparing historical simulations with the local climate in the studied domain allows for appropriate selection and correction of RCM-GCM schemes. An ensemble of climate scenarios compatible with the domain conditions is necessary to study future climate changes and, in turn, create effective mitigation and adaptation programs.

Reliable and high-quality observational datasets are essential to create good-quality ensembles of climate scenarios. First, the full range of available data, including local observational data, should be used. The resources of many repositories contain gridded datasets, but regional analysis usually does not meet this condition. Within the European Climate Assessment & Dataset project, regular grid data E-OBS (high-resolution gridded mean/max/min temperature, precipitation, and sea level pressure for Europe and Northern Africa) (Cornes et al. 2018) with 0.1- and 0.25-degree spatial resolution and daily temporal resolution are available. The current list of stations from which the observational data are used to create E-OBS sets for rainfall for the territory of Poland includes 1688 items (including 40 synoptic stations provided by the Institute of Meteorology and Water Management – National Research Institute (IMWM-NRI)). However, the period over which individual observation series are available varies significantly. For the analyzed pe-

riod of 1976-2005, the minimum number of daily records is 89, with a median of 650 (these are significantly lower values than for the data used in this work, for which the minimum is 197 and the median 719). For 1826 days, the number of stations used to create E-OBS reanalysis is less than 197. This corresponds to approximately 17% of the total amount of data in the analyzed period, i.e., approximately five years. The use of more input data can help to describe precipitation more accurately.

Some meteorological services or institutes provide gridded precipitation data, but these have limitations when applied to Poland. The National Oceanic and Atmospheric Administration (NOAA) Physical Sciences Laboratory (PSL) website provides collections with gridded precipitation data ranging from 2.5 to 0.25 degrees. However, the high-resolution collections of gridded precipitation datasets cover only the United States of America (USA) area. The Swiss Federal Institute for Forest, Snow and Landscape Research (WSL) provides free access to high resolution (30 arcsecs, ~1 km) climate data – CHELSA (Climatologies at high resolution for the earth's land surface areas) (Karger et al. 2017). This website has available data based on a mechanistic statistical downscaling of global reanalysis data or global circulation model output. Unfortunately, the data from the reanalyzes covers the period from 1979, which is already intricate and highly processed.

There are studies where great importance is attached to the maximum use of available observational data, especially when extreme events are important in the analysis, as shown in Sheffield et al. (2006). The study conducted by Belo-Pereira et al. (2011) regarding rainfall data across the Iberian Peninsula demonstrated that accurate descriptions of region-specific meteorological situations require high-resolution datasets derived from a comprehensive measurement network. The global gridded datasets compared in Belo-Pereira et al. (2011) overestimated the number of days with precipitation while underestimating heavy precipitation events.

Two high–resolution daily gridded datasets were created in the climate change impact assessment for selected sectors in Poland (CHASE-PL) project. The work of Berezowski et al. (2016) described the first dataset with a resolution of 5 km covering the period of 1951-2013 for temperature and precipitation in the two largest Polish river basins. The work was carried out by increasing the resolution to 2 km, extending the period from 2013 to 2019, and extending the list of meteorological parameters, including humidity and wind speed (Piniewski et al. 2021). Data from the measurement network of the IMWM-NRI were used, as well as data from all neighboring countries, including the densest data network from Germany. The assessment of the quality of the prepared gridded datasets, included in the works of Berezowski et al. (2016) and Piniewski et al. (2021), also showed that they describe the local meteorological conditions well, but the projected Polish Geographic Coordinate System 1992 (PUWG-92) was used. Many studies (e.g., Herrera et al. 2018; Crespi et al. 2019) on the interpolation of observational data assess the causes of errors, pointing to the role of the density of measuring stations and the properties of interpolation methods. Daly et al. (2017) analyzed the uncertainty in gridded precipitation datasets for densely spaced rain gauge networks in the Appalachian Mountains in western North Carolina, USA. In this study, it was concluded that station density and misallocation are likely sources of errors. The sensitivity assessment of various interpolation

methods was included in Crespi et al. (2019), where the 1981-2010 monthly precipitation climatology for Norway at 1 km resolution was presented. In this paper, three interpolation algorithms were considered. The first algorithm, HCLIM+RK (the global historical climate database + Regression Kriging), was a combination of two methods, combining the output from a numerical model with *in-situ* observations. The second algorithm, MLRK (Multi-Linear Local Regression Kriging), resulted from the Multi-Linear Local Regression Kriging, and the third LWLR (Local Weighted Linear Regression) was the Local Weighted Linear Regression. Among other conclusions from the conducted research, the authors noted that the accuracy of MLRK and LWLR was more sensitive to the spatial variability of station distribution over the domain. Their interpolated fields were more affected by discontinuities and outliers, especially over those areas not covered by the rain-gauge network. A comprehensive analysis of the uncertainty in the gridded data was presented by Herrera et al. (2018). Three factors influencing the quality of the interpolated fields were analyzed: station density, interpolation methodology, and spatial resolution of the fields obtained. In the paper, an experiment was carried out for three interpolation methods and different levels of observational data density. This paper evaluated the experiment's results with a statistical analysis of variance (von Storch, Zwiers 1999; Deque et al. 2007; 2012). The authors stated that the station's density explained more than 60% of the variance of the interpolation procedures.

This study aimed to develop reference precipitation gridded datasets covering the territory of Poland. These datasets were planned to evaluate the RCM-GCM ensemble using measures presented in works such as Gleckler et al. (2008) and Konca-Kędzierska (2019), for the last available 30-year period (1976-2005) in the historical simulations on the EUR 0.11° grid.

Several of the works mentioned concerned gridded data for the area of Poland, but they did not correspond to the needs posed in this paper. For example, in Berezowski et al. (2016), Piniewski et al. (2021), and Cornes et al. (2018), a grid other than the EUR 0.11° was considered. In Herrera et al. (2018), the period did not cover the years 1976 and 1977. The undeniable influence of the quantity and quality of observation data on the quality of gridded fields has been confirmed in all these works.

2. Materials and methods

2.1. Data

Given our goal of producing a grid precipitation field for the 1976-2005 period (as part of the EURO-COREX project's historical scenarios), we opted to rely on the daily precipitation data publicly available from IMWM-NRI[1]. The IMWM-NRI observation network service department is accountable for ensuring data accuracy, verifying the suitability of station locations, and selecting measurement equipment in compliance with the Technical Regulations of the World Meteorological Organization (WMO). The measurement network includes three types of stations operating in different time regimes: synoptic, climatic, and

[1] https://danepubliczne.imgw.pl

rainfall. The amount of data for individual days of the period is variable, which affects the interpolated daily fields (Table 1).

Table 1. The statistics on N^{Day} – the amount of observational data per one day. Min. – minimum of N^{Day}, Mean – average value of N^{Day}, Max. – maximum value of N^{Day}, Q25, Q50, Q75 – 25th, 50th and 75th percentiles of N^{Day}, respectively.

Stations	Min.	Q25	Q50	Mean	Q75	Max.
Synoptic	57	60	60	60	61	63
Climatic	135	142	160	163	174	205
Rainfall	1	176	505	501	804	1166
All	197	397	719	717	1016	1429

In the analyzed period, there were 63 days when practically no data from rainfall stations were available, which is less than about 0.6% of the total number of days. In these cases, observations are mainly from synoptic and climatic stations. Although the minimum distance to the neighboring station varies from 3 to 75 km on the day with the lowest number of observations, compared to 0.5 to 33 km on the day with the highest number of observations, the data quality is sufficient, considering data comes from synoptic stations. The number of observations is increased at rainfall stations, densely located in mountainous areas, which undoubtedly significantly increases the quality of reproduction of precipitation by interpolated fields. The density of the observation network, especially in mountainous areas, is crucial for the quality of interpolated fields, e.g., Herrera et al. (2018). As shown in Figure 1, the spatial distribution of the observational data is irregular, which may have uneven effects on the interpolation methods used. The exceptionally high density of the measurement network is characteristic of the mountain and sub–mountain regions, where the spatial variability of precipitation is the highest.

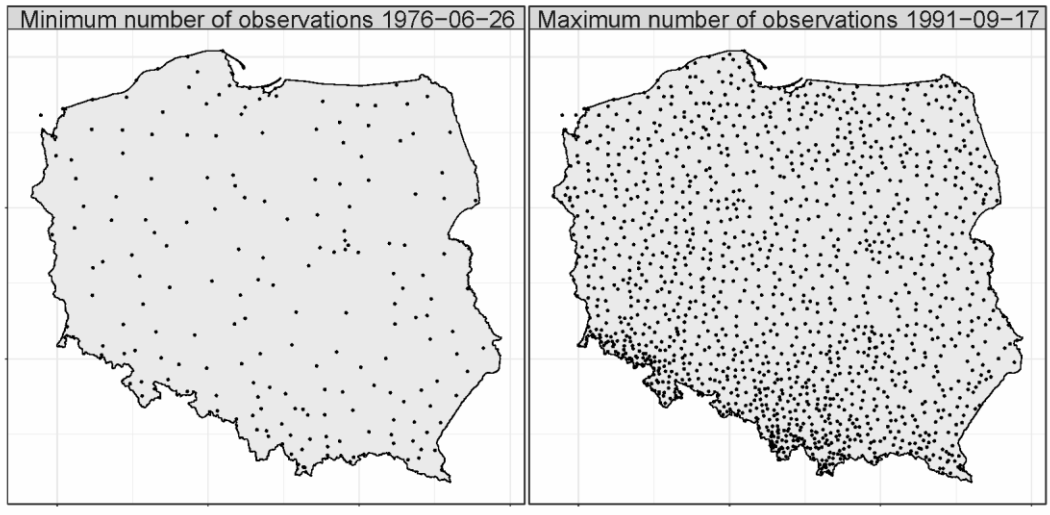


Fig. 1. Spatial distribution of observations for the days with an extreme number of observations.

The average annual number of observations per day varied throughout the years, with a median of 719 (ranging from 592 to 821). The variability in the number of observations regarding the type of stations is shown in Figure 2.

The decrease in the number of synoptic observations was compensated for by the increase in climatic observations, which was related to the change in the type of measurement stations. The number of synoptic stations ranged from 57 to 63; a decrease to 57 occurred in the last years of the period. In the analyzed period, the number of climatic stations increased by approximately 17%, ranging from 137 to 202. Following an increase in the mid-1980s, the number of stations stabilized at 160 by the end of the period. The number of rainfall stations ranged from 401 to 580, and these had the most significant impact on the total number of observations. In this work, we used observation data from IMWM-NRI, which are subject to routine quality control. The local data archive for the analyzed period 1976-2005 was created in November 2019.

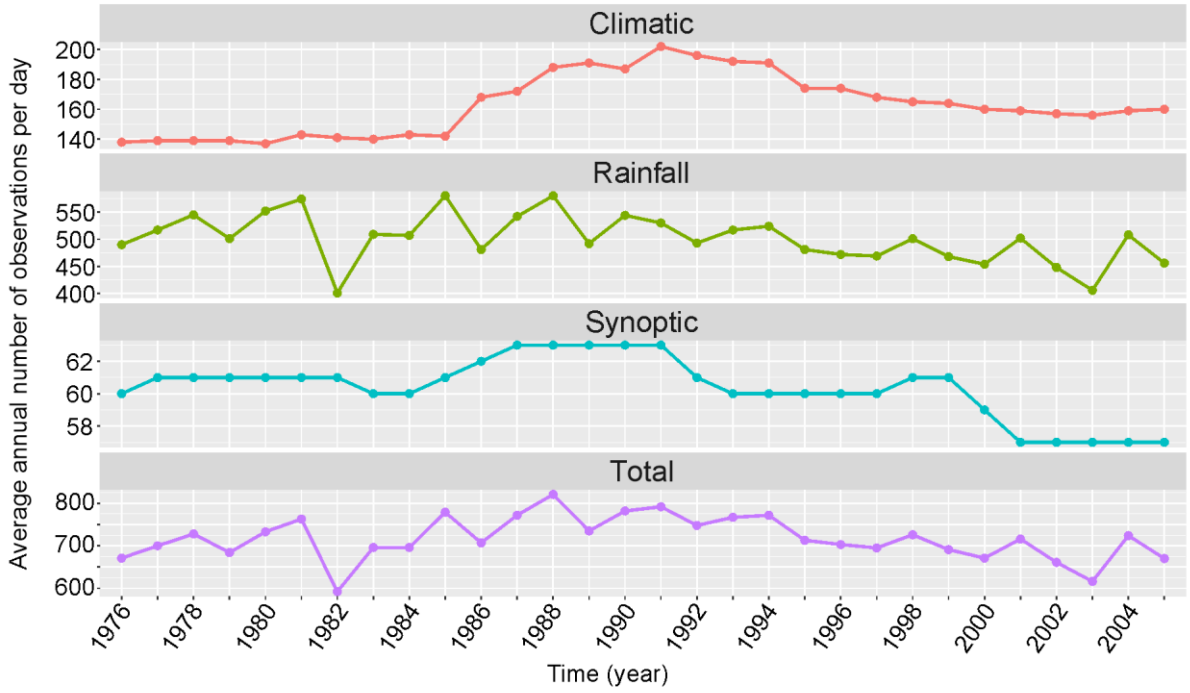


Fig. 2. The average annual amount of data for one day from 1976-2005 for climatic, rainfall, synoptic, and total observations.

2.2. Methods

We utilized the EURO-CORDEX simulation outcomes as-is, thus requiring a reference precipitation dataset that matched the node grid's shape but was tailored to Poland's region. Therefore, we selected a domain that is part of the EUR-11 grid. The spatial resolution of the EUR-11 rotated grid is 0.11 degrees, corresponding to a regular grid size of approximately 12.5 km. The analyzed area included only nodes located in Poland, of which there were 2,137. We analyzed the last 30-year period (1976-2005) available in the historical simulations, and will be applied these to correct climate scenarios.

The principle of using all historical data available in the IMWM-NRI database was applied. This caused the number of observations to change significantly on individual days. On the other hand, it is known that interpolation methods are varyingly sensitive to this factor. The combination of both premises suggests that instead of using one sophisticated interpolation method for the entire period, the interpolation method should be selected separately for each day. The allegation of methodological discontinuity can be compensated by obtaining a more realistic and reliable precipitation pattern. The fact that the selected domain is a regular grid with a small step of 0.11 degree influenced the choice of interpolation methods. Deterministic interpolation methods were used, such as bilinear, bicubic, inverse distance weighted, or nearest neighbor. Precipitation observation fields were prepared using the above-mentioned interpolation methods available in the R Project for Statistical Computing (R Core Team 2018) and in the Climate Data Operators tools (CDO) (Schulzweida 2019).

When evaluating interpolation through the leave-p-out cross-validation method (p is the size of the test set), the accuracy of the analysis may be impacted by high variability in the number of observations. Days with a small number of observations (minimum of 197) are assessed less thoroughly than days with many observations (median of 719). The division into training and test sets also introduces randomness into the evaluation process. This could be eliminated by taking as a constant test set the time series for 102 stations for which complete observational data are available. It was decided to abandon the method of dividing the observational data into the training and evaluation part, as it is usually done, for two reasons. Excluding the selected evaluation set from the interpolation process removes the data series most valuable for the interpolation. High variability in the number of observations per day (25% quantile is 397) for 25% of days of the analyzed period would mean resignation from 26% to 52% of the input data for interpolation methods. On the other hand, the leave-one-out computationally expensive method, is a measure of the additional errors incurred during its execution. For each point in the observational data on a given day of the period, interpolation is performed with the value for that point omitted from the input data. Furthermore, the considered point usually does not occur in the target grid, and its value must be somehow calculated (e.g., by choosing the nearest point from the target grid). The cross-validation error is the average of the sum of the additional error in obtaining the interpolated value and the error caused by not considering all points in the input data of the interpolation model. We conducted a comparative analysis of the obtained observation fields for localization, where the complete time series of observations were available during the analyzed period.

The degree of compliance of the obtained fields of observation was assessed based on statistical parameters such as the Pearson correlation coefficient (RO), Root Mean Squared Error (RMSE), and a normalized version of this parameter (NRMSE) as in Belo-Pereira et al. (2011) and Berezowski et al. (2016). In addition, the spatial compliance of the precipitation fields was also examined using the correspondence ratio (CR) (Belo-Pereira et al. 2011). For the obtained reference precipitation fields, the Mean Annual Cycle (MAC), as established in Belo-Pereira et al. (2011), was also compared. The results of the analyses did

not allow us to determine the best method among the methods listed in Table 2. Thus, the methods determining the best gridded dataset for particular days based on the correlation coefficient and the correspondence ratio were finally applied and presented in Section 3.10.

2.2.1. Interpolation methods

Areas of rainfall are unevenly distributed. The regions with no precipitation can border regions with intense precipitation. This makes precipitation quite a problematic parameter to interpolate. In geostatistics, kriging methods are commonly used, but such an interpolation process sometimes causes problems. In Prasad and Sushma's (2016) work, the result was satisfactory and encouraging for most of the data. However, where there was a small amount of data, the obtained values exceeded the range of observational data.

We tested several other interpolation methods available in the R environment (R Core Team 2018) and the Climate Data Operators (CDO) (Schulzweida 2019). Table 2 lists the six selected interpolation techniques and the names of the resulting sets with interpolated values.

Table 2. Interpolation procedures used.

Interpolated field	Interpolation procedures
v1	R (akima) – bilinear interpolation
v1BI	R (akima) – bicubic interpolation
v2dis	CDO – remapdis "Distance-weighted average remapping"
v2nn	CDO – remapnn "Nearest neighbor remapping"
v3IDW	R (gstat) – "Inverse Distance Weighted Interpolation"
v3TPS	R (fields) – "Thin Plate Spline regression"

The interpolation procedures in the Akima package (Akima, Gebhardt 2022) allow for regular and irregular grids for the input data. The method is based on the modified triangulation Akima code. A bilinear interpolation for regular grids was also added for comparison with the bicubic interpolation on regular grids. Calculations are made locally; thus, only neighboring points are considered. In rare cases, the interpolation with a linear function resulted in negative values corrected by the neighboring nodes' mean value.

The CDO interpolation procedures are based on the Spherical Coordinate Remapping and Interpolation Package (SCRIP) library developed at the Los Alamos National Laboratory (Jones 1998). Both are adapted to interpolate from an irregular grid, such as for measurement grid nodes. By default, the CDO operator 'remapdis' uses four values from the nearest neighborhood to interpolate the destination value. The result value is the weighted average of these values, the weights being the reciprocal distance between the points.

Nearest neighbor remapping 'remapnn' is the simplest spatial interpolation method. Every predicted point gets the value of the nearest measured point. The fields obtained by this method are not smooth, but they can provide good field interpolation where there is a sufficiently large number of observations.

The 'gstat' package is a very rich package for complex geostatic analysis with a scope defined by the title: "Spatial and Spatio-Temporal Geostatistical Modelling, Prediction and Simulation" (Pebesma 2004; Gräler

et al. 2016). However, the Inverse Distance Weighted predictions method was selected to prepare the reference field and preserve the similarity of the methods of obtaining the remaining compared fields. The 'gstat' package offers implementations of the Inverse Distance Weighted (IDW) method for different values of the power of the distance between nodes. The calculations were performed by utilizing the square of the distance between nodes to determine weights. Introducing this higher power value enhances the influence of values that are near the node. This contrasts with the CDO method known as 'remapdis' or 'Distance–weighted average remapping', where the distance is not squared.

The last of the selected interpolation methods is the 'Tps' procedure from the 'fields' package version 11.6 (Nychka et al. 2017). The 'Tps' procedure fits a thin plate spline surface to irregularly spaced data, uses a special type of piecewise polynomials, and is expected to give better results. It was applied assuming default parameter values.

2.2.2. Methods of evaluating the obtained reference fields

The obtained reference fields were assessed in three ways: assessment of the general fit based on annual and monthly characteristics; methods of analyzing daily data (in particular, extreme values assessed with number of days in the month when precipitation does not exceed 0.5 mm (LD05) and the 95th percentile (Q95)); and the overall assessment of the fit of fields using illustrative diagrams. Individual parameters were calculated for selected 102 synoptic and climatic stations for which complete data series for the studied period were available.

The first group includes field assessment of the annual sum of precipitation, the annual cycle of the monthly sum of precipitation, and the NRMSE for the monthly sum of precipitation. The overall level of matching reference fields was assessed based on basic statistical characteristics (extreme values, percentiles 1% and 99%, median and Interquartile Range (IQR)) of observational data and reference fields.

The annual cycle of the monthly sum of precipitation was used as described in Belo-Pereira et al. (2011), where the daily gridded precipitation dataset over mainland Portugal was assessed. For stations with complete sequences of monthly rainfall totals, 30–year averages were calculated and compared with the results of corresponding calculations for the reference fields. The percentage difference for the multi-year average of the monthly sum of precipitation allows assessment of the quality of the reconstruction of the annual variability by the reference fields.

To assess the compliance of the precipitation parameters in the interpolated fields, formulas such as the mean absolute error (MAE) were used:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \quad (1)$$

and the mean square error (MSE):

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \quad (2)$$

where y is the observation, $\hat{y}$ is the interpolated value of y, and n is the number of points included in the calculation.

Often, to calculate the error interpolation, the $RMSE$ is used. However, in the case of precipitation, which is a spatially variable parameter, $NRMSE$ may provide a better evaluation measure of interpolation error (Otto 2019).

$$NRMSE = \frac{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}}{\sigma} \tag{3}$$

where σ is the standard deviation of the sample of observations. This allows for assessing the interpolation method, making it independent of local precipitation variability. Using $NRMSE$, the monthly sum of precipitation, the maximum values, and the quantile of 95% of the daily sum of precipitation in a month were also analyzed. The analysis of cases of absence or very small precipitation was carried out using the number of days per month for which the precipitation was not more than 0.5 mm. The assessment of the spatial fit of reference fields was carried out using the CR, the idea of which was taken from the work Belo-Pereira et al. (2011):

$$CR = \frac{A_I}{A_U}100\% \tag{4}$$

A_I is a measure of the intersection of the area where the precipitation in the interpolated field and the field of observation exceeds a given threshold. A_U measures the sum of the areas where precipitation in the interpolated field or the field of observation values exceeds a given threshold. The areas for ten thresholds of daily total rainfall from 0.1 mm to 50 mm were analyzed. The number of stations that met this condition over the analyzed period was adopted as a representation of the measure of the investigated areas.

The RO was calculated for the entire period for stations with complete daily sequences of total precipitation. The Pearson correlation coefficient was also calculated for each day in the warm half of the –year (W – months from May to October) and in the cold half of the –year (C – months from November to April).

The mean values of $RMSE$ and MSE for the precipitation statistical parameters listed below were used to construct the ranking of the interpolation methods. For stations with complete daily sequences of total rainfall, the calculated values of MSE and MAE were used for the maximum rainfall in the month (MAX), the monthly sum (MS) of precipitation, the number of days in the month when precipitation does not exceed 0.5 mm (LD05), and the 95th percentile (Q95). The lowest error value gave the highest position in the ranking by summing up the position numbers for individual interpolation methods and all parameters considered (MAX, MS, LD05, Q95). This allowed us to determine which methods provided the best fit using this approach. However, this is too general and a simplified solution, which does not allow for a good interpolation choice in all cases. The interpolation result for individual days depends, to a large extent, on the quantity and quality of available observational data. It is even sensitive to the distribution of

data in the domain. The sensitivity of the methods to small amounts of data and not evenly distributed points is very different.

The variance inflation factor (*VIF*) was analyzed for the 30-year mean of the annual sum of precipitation. This parameter was analyzed in two cases: for interpolation methods (in 3.1 Average annual total of precipitation) and in the case of comparing the result fields of this work with fields from the CHASE projects (in 4. Discussion). For these purposes, a group of 96 stations was selected for which there was no missing data in the years 1976-2005. In addition, stations for which data from the CHASE project could not obtain a resolution of 5 km (Berezowski et al. 2016) were eliminated.

The formula (2.5) for the variance inflation factor (*VIF*) is based on the coefficient of determination, denoted R^2.

$$VIF = \frac{1}{1-R^2} \tag{5}$$

The coefficient of determination R^2 for a series of observation values (y_1, ..., y_n) each associated with a fitted value (f_1, ..., f_n) is defined as follows:

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}} \tag{6}$$

SS_{res} is the total sum of squares of residuals:

$$SS_{res} = \sum_i (y_i - f_i)^2 \tag{7}$$

SS_{tot} is the total sum of squares (proportional to the variance of the data):

$$SS_{tot} = \sum_i (y_i - y_{mean})^2 \tag{8}$$

The value of y_{mean} is the mean of the observed data. The *VIF* factor for various methods of obtaining fi approximations allows for the assessment of changes in variance in individual methods.

3. Results

The IMWM-NRI observational data for the studied period of 30 years (1976-2005) contained a complete series of daily sums of precipitation for 102 synoptic and climatic stations. Most of the analyses were carried out for the values of these stations. Sometimes, it was possible to use less stringent conditions for the completeness of the observational data, and more stations could be used (e.g. when analyzing the annual total of precipitation). However, when analyzing for MS and MAX, a complete series of observations was required, which limited the number of stations to 102.

3.1. Average annual total of precipitation

For the analysis of the average annual rainfall over the 30-year period (1976-2005), small gaps in observational data were negligible. It was decided to extend the number of analyzed locations to 53 synoptic stations and 67 climatological stations. For 99 stations, the complete sequence of observations was available, whereas for 21 stations, a single lack of data lasting up to three months was allowed. The weakening of selection criteria made it possible to analyze the mean annual sum of precipitation for more stations. The mean value and the median for the average annual total rainfall (YS) were overestimated for the interpolated data (the most for the v3IDW method). At the same time, the interquartile range for the observational data, amounting to 149, was much wider than for the interpolated data (from 69 to 112). This means the interpolated values were clustered in a narrower range (IRQ) around the higher values (median). Table 3 presents the essential precipitation statistics (YS min. – minimum, Q_0.01 percentile 1% (the probability that the value of the mean annual sum of precipitation is below this parameter is 0.01), YS Median – Median, YS Mean – Mean, Q_0.99 – percentile 99% (the probability that the value of the mean annual sum of precipitation is below this parameter is 0.99), YS Max. – maximum, IQR – Interquartile Range).

Table 3. The characteristics of the mean annual sum of precipitation (mm) from 1976-2005.

	YS Min.	Q_0.01	YS Median	YS Mean	Q_0.99	YS Max.	IQR
Observations	489	497	589	644	1202	1796	149
v1	512	528	631	658	1053	1615	103
v1BI	513	530	636	668	1109	1625	112
v2dis	524	544	643	668	1051	1689	102
v2nn	482	510	633	659	1053	1704	110
v3IDW	529	595	741	751	1017	1563	69
v3TPS	543	554	640	668	1067	1400	96

Analysis of these characteristics showed that the highest inconsistencies occurred for the v3IDW and v3TPS methods. For the v1BI method, there was an additional problem for three grid points located at the most southern end of the Polish border (Bieszczady National Park). For these three points, unrealistic values of the annual sum above 2600 mm were achieved (this is the maximum annual sum of precipitation for observational reference data), probably due to the extrapolation of data in the absence of a sufficient number of measurement points. As seen in Figure 3, for all methods except v3IDW, there is a middle band of lower annual precipitation totals and higher precipitation areas for the southern (mountain) and northern (coastal) extremes. For the v3IDW method, no such differentiation is apparent; the middle belt is not homogeneous (but it is impossible to distinguish), as is the case for other methods; and there are regular structures of lower and higher annual sums of precipitation. The remaining fields resemble the spatial distribution of annual rainfall totals for 2019 (IMGW-PIB 2020).

The analysis of the extreme values and the annual precipitation sum of percentiles provided a good result only for the maximum value. For the considered 120 stations, this value was 2600 mm, which occurred in

2001. All interpolation methods (for v1BI, the controversial three grid nodes were removed from the analysis) indicated the same year of occurrence and values ranging from 2097 mm to 2626 mm.

The variance inflation factor (*VIF*) was estimated for 96 stations for which time series of observations without missing data were available. The same stations were used in 4. Discussion for comparison of the resulting output methods of this work with data from the CHASE project (Berezowski et al. 2016; Piniewski et al. 2021). The results of the analysis are included in Table 4.

The lowest variance inflation factor *VIF* (formula 2.5) was achieved for the v3IDW method (R^2 is the lowest for this method). For the IDW method on the analyzed sample, the sum of residuals was the highest, i.e., the sum of squares of the difference between observations and interpolated values. Results showed the lower the value of *VIF* and R^2 (formula 2.6), the greater the SS_{res} (formula 2.7) of the sum of squares of the difference of observations and interpolated values. A small variance of the interpolated values does not necessarily mean a good fit for the observational data.

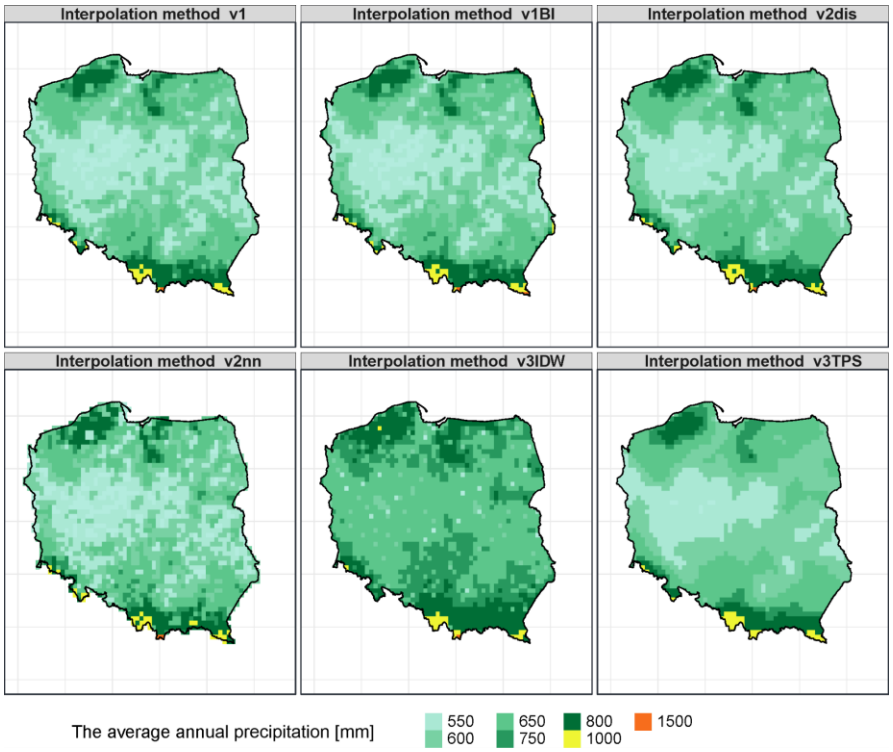


Fig. 3. Average annual total rainfall fields for interpolated daily rainfall fields from 1976-2005.

Table 4. The variance inflation factor, the coefficient of determination, and the sum of squares of residuals for the mean annual sum of precipitation from 1976 to 2005.

Interpolated field	Variance Inflation Factor (*VIF*)	Coefficient of determination (R^2)	Sum of squares of residuals (SS_{res})
v1	6.3	0.84	507254.9
v1BI	6.1	0.84	525839.6
v2dis	6.3	0.84	508471.1
v2nn	7.5	0.87	425769.5
v3IDW	3.2	0.68	1012278.5
v3TPS	4.3	0.77	745762.9

3.2. Mean annual cycle of precipitation

Based on the calculated 30-year average monthly sum of precipitation, the annual cycle of the monthly sum is presented in Figure 4. The overall assessment was positive for all models, and they all maintained the annual cycle. However, when calculating the percentage errors (in relation to the mean values for the observations) of the monthly averages, there was a division into two groups for better (methods v1, v1BI, v2nn) or worse (methods v2dis, v3IDW, v3TPS) reproduction of the annual cycle (Fig. 5). The lowest percentage error was achieved for the v2nn method for all months below 5%, and the highest for v3IDW for all months above 10%.

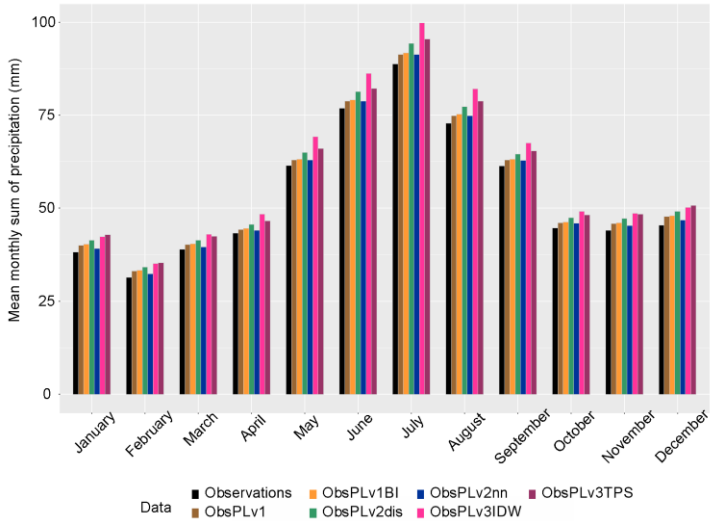


Fig. 4. Annual cycle of the monthly total precipitation for 1976-2005 reconstructed in interpolated data.

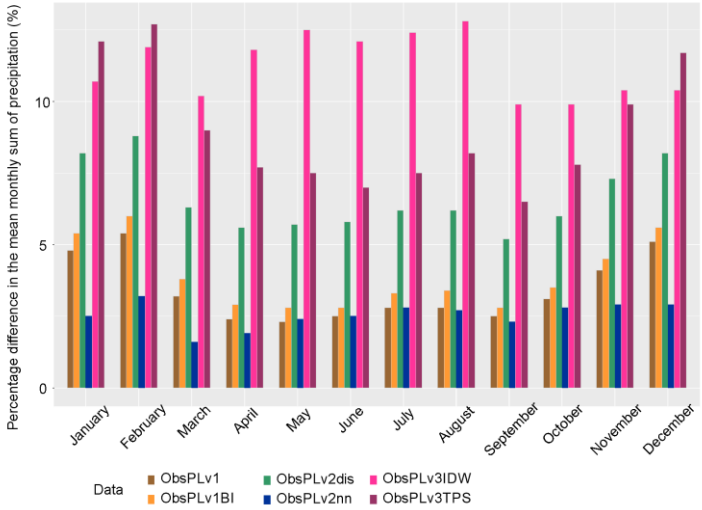


Fig. 5. Percentage error for the reconstruction of the annual cycle of the monthly rainfall total reconstructed in interpolated data.

3.3. Monthly sum of precipitation

The analysis of the *NRMSE* and *MAE* values for the monthly precipitation sum is included in Table 5. The mean value of the *NRMSE* varied from 0.22 to 0.36, while the maximum value for all methods was at level

2. The mean value of *MAE* changed from 5.06 to 9.47. The large spread was reached by the maximum values of *MAE* from 30.8 to 48.3 mm, while the standard deviation *SD* of the *MAE* showed moderate variability ranging from 5.5 mm to ~8 mm. The *NRMSE* rating indicated the v2nn method, but the mean *MAE* exceeded 40 mm. The minimum value of the *MAE* was achieved by the v3IDW method, although this had the highest average error value.

Table 5. Statistics for the error of the monthly sum of precipitation.

Interpolated field	*NRMSE*	Max *NRMSE*	Min *NRMSE*	*MAE*	Max *MAE*	Min *MAE*	*SD*
v1	0.23	1.19	0.02	5.91	48.30	0.47	5.89
v1BI	0.23	1.19	0.04	6.06	48.25	0.92	6.00
v2dis	0.26	1.19	0.04	6.80	40.57	1.00	6.51
v2nn	0.20	1.20	0.00	5.06	41.69	0.00	5.56
v3IDW	0.36	1.22	0.03	9.47	30.80	0.75	7.94
v3TPS	0.33	1.21	0.23	8.64	32.03	3.93	7.83

3.4. Maximum daily sum of precipitation in a month

A similar analysis was carried out for the maximum daily sum of precipitation. The mean value of the *NRMSE* (Fig. 6) ranged from 0.26 to 0.42. The mean value of *MAE* ranged from 1.78 to 2.82.

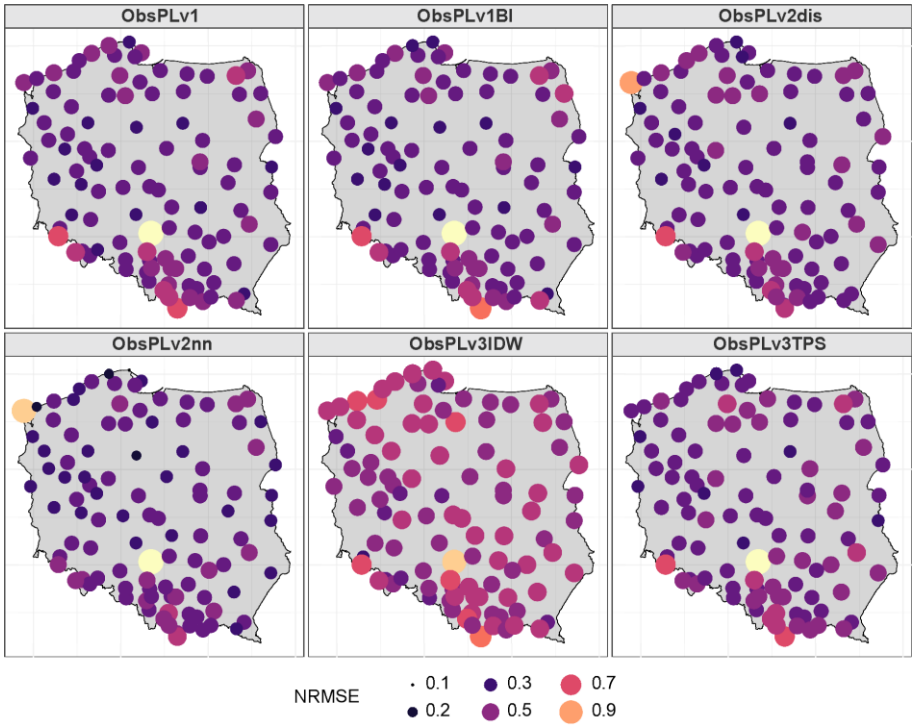


Fig 6. Map of the Normalized Root Mean Squared Error (*NRMSE*) for the maximum daily sum of precipitation in a month for the warm half of the –year (May to October) from 1976 to 2005.

3.5. The 95th percentile of the daily total precipitation in a month

The maximum values depend on random outliers, but the reconstruction of the range of high daily total precipitation can be assessed based on the 95th percentile (Q95).

The assessment according to the Q95 parameter was more homogeneous for the interpolation methods (Table 6). The mean value of *NRMSE* varied from 0.40 to 0.48, the mean value of *MAE* did not exceed 2 mm, and the standard deviation *SD* of *MAE* changed from 1.89 to 2.27. Moreover, the lowest error parameters were achieved for the v3IDW method.

Table 6. Statistics for the error of the 95th percentile of the daily total precipitation in a month.

Interpolated field	*NRMSE*	Max *NRMSE*	Min *NRMSE*	*MAE*	Max *MAE*	Min *MAE*	*SD*
v1	0.43	0.84	0.28	1.56	5.54	0.98	1.89
v1BI	0.43	0.89	0.28	1.58	5.52	0.97	1.91
v2dis	0.44	0.98	0.29	1.62	6.25	0.99	2.00
v2nn	0.46	1.19	0.22	1.67	7.22	0.98	1.89
v3IDW	0.40	0.78	0.24	1.47	4.29	0.82	1.92
v3TPS	0.48	0.83	0.36	1.77	5.26	1.19	2.27

3.6. The number of days with the daily sum of precipitation below 0.5 mm

Situations without precipitation or small daily total precipitation can be described using the number of days (LD05) in a month when the threshold was set to 5 mm.

Table 7 presents the basic statistics of two commonly used metrics, *NRMSE* and *MAE*, for assessing the accuracy of an interpolated field of LD05. The table shows that v2nn had the lowest *NRMSE* (0.43) and the lowest *MAE* (1.22), suggesting it may be the most accurate method for interpolating the field. Conversely, v3IDW had the highest *NRMSE* (1.50) and the highest *MAE* (5.26), indicating that it may be the least accurate method. For the v3IDW and v3TPS methods, the value of *NRMSE* was greater than 1. This suggests that these methods may overestimate the number of days with precipitation. Additionally, the *MAE*'s standard deviation (*SD*) for these two methods was above 2, while it did not exceed 2 for the other methods. This indicates that the errors for v3IDW and v3TPS are more variable and less consistent than the errors for the other methods. Overall, these findings suggest that v3IDW and v3TPS may not be the most accurate methods for interpolating rainfall data in this context.

Figure 7 illustrates that, with a few exceptions, the *NRMSE* for the v3IDW method was greater than that of the other methods throughout the entire domain.

Table 7. Statistics for the error of LD, the number of days in a month with a threshold of 5 mm.

Interpolated field	*NRMSE*	Max *NRMSE*	Min *NRMSE*	*MAE*	Max *MAE*	Min *MAE*	*SD*
v1	0.62	1.12	0.18	1.94	3.89	0.36	1.57
v1BI	0.65	1.15	0.23	2.05	3.92	0.53	1.63
v2dis	0.76	1.54	0.23	2.48	5.40	0.54	1.81
v2nn	0.43	1.07	0.00	1.22	3.69	0.00	1.22
v3IDW	1.50	2.42	0.24	5.26	9.02	0.56	2.78
v3TPS	1.01	1.73	0.43	3.39	5.91	1.50	2.21

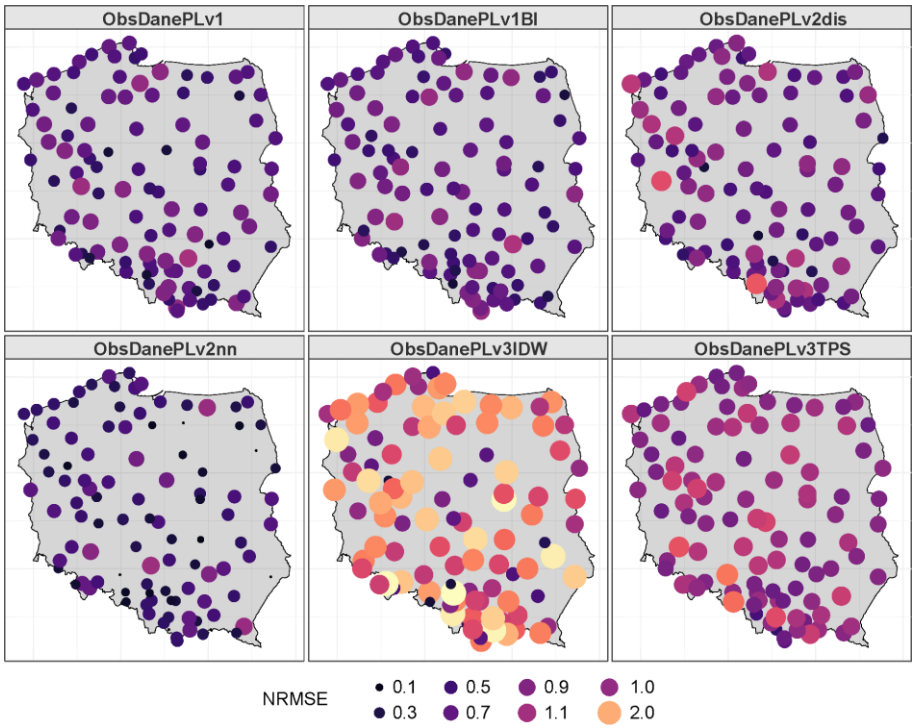


Fig. 7. Map of the *NRMSE* value for the number of days in a month with daily total precipitation less or equal to 0.5 mm.

3.7. Correspondence ratio (CR) for the daily sum of precipitation

The CR measures the accuracy of spatial mapping by interpolation methods for areas with daily precipitation exceeding a threshold. The threshold of 0.1 mm indicates that even a small amount of fallout has been detected, which is referred to as a 'trace' amount. Certain areas were recreated correctly at 80% using different methods. However, two methods (v3IDW and v3TPS) did not perform as well as the others, achieving only 64% and 69% accuracy, respectively.

Figure 8 shows the analysis results for eight selected thresholds: 0.1, 1, 5, 10, 20, 30, 40, and 50 mm. The v2nn method achieved the best results, with close to 90% agreement for the 0.1 mm threshold and above 80% for the 1 mm threshold. For the 50 mm threshold, the agreement for the v2nn method is above 60%, while it ranges from 36% to 49% for the other methods.

3.8. Correlation coefficient (RO) for the daily sum of precipitation

The RO value was analyzed in two ways. RO was calculated for the entire period, i.e., for the time series of interpolated values and observations. For all analyzed locations, the RO value was above 0.8, results are shown in Figure 9.

In addition, the RO (the Pearson correlation coefficient) factor was calculated for individual days. In the absence of precipitation for more than 80% of observations, the RO was replaced with a measure of agreement in two-way tables. The basic characteristics calculated for these values indicated a slight differentiation when dividing the period into a warm and a cold half-year. For the warm half-year, the 25% RO quantile for the v3TPS method was 0.6; for the v2nn method, it was 0.9; and for the others, it was 0.8.

The median value of RO for this half-year was 0.9. For the cold half-year, the 25% quantile of the RO value for the v3TPS method was 0.6; for the remainder, it was 0.8. The median in the cold half-year for RO for the v3TPS method was 0.8, and for the other methods was 0.9.

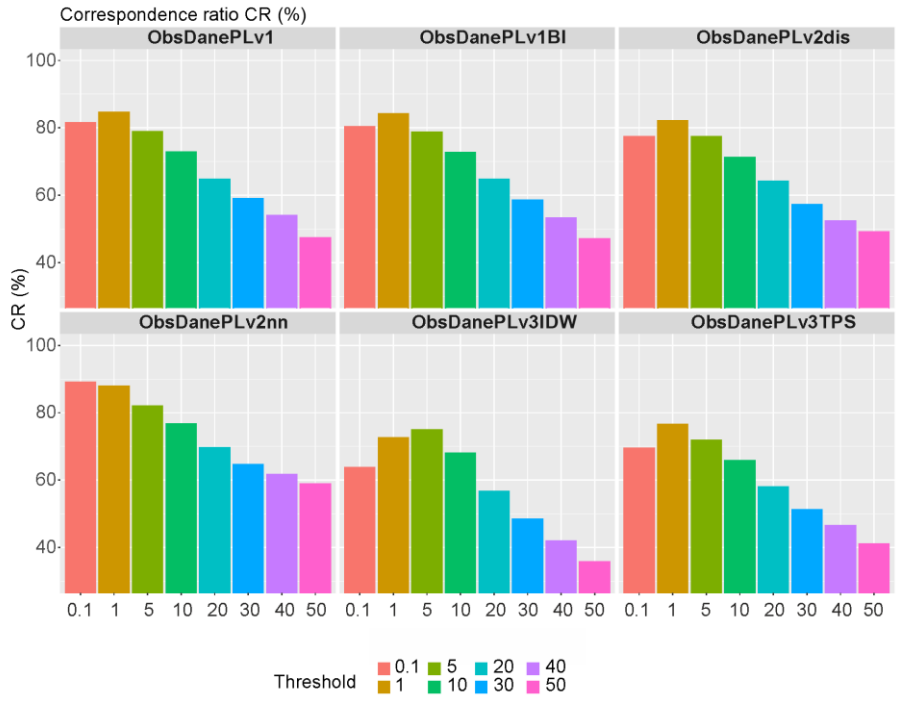


Fig. 8. Correspondence ratio (CR) for the daily sum of precipitation.

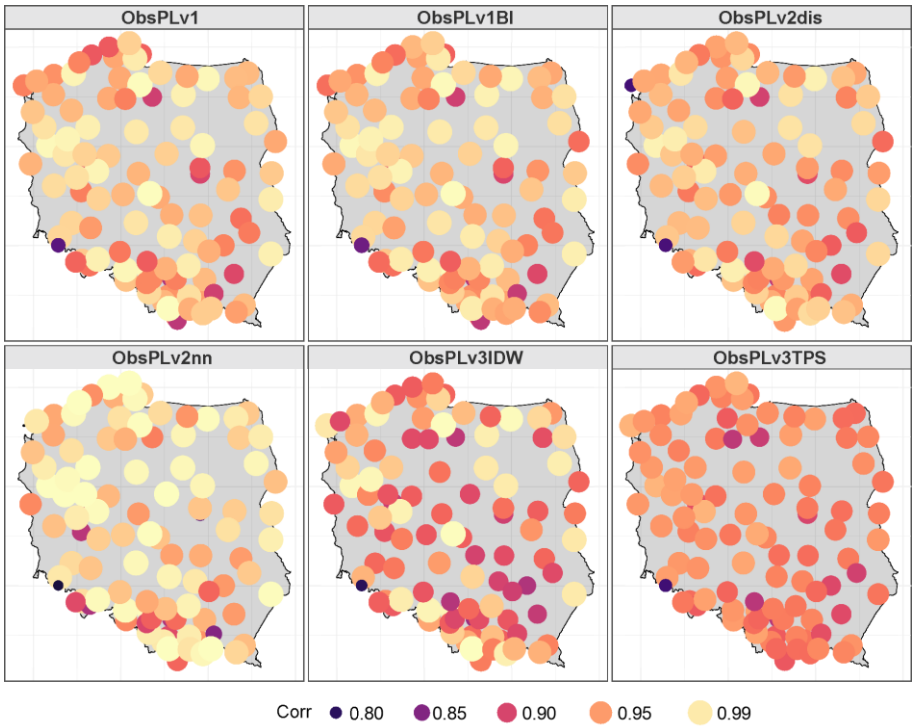


Fig. 9. Map of the correlation coefficient of the daily sum of precipitation for the period 1976-2005.

3.9. Ranking of interpolation methods based on the average error

The calculated *NRMSE*/*RSME* and the *MAE* values allow ordering of the considered interpolation methods. A ranking of the methods was constructed considering these values for the maximum daily sum of precipitation in a month, the monthly sum of precipitation, the 95th percentile of the daily total precipitation, and the number of days with precipitation less than or equal to 0.5 in a month. The order of the methods in this ranking were v1, v2nn, v1BI, v2dis, v3IDW, and v3TPS (Table 8). There are six possible orders or rankings based on a certain parameter. Each number from 1 to 6 represents a different ranking, with 1 indicating the highest rank and 6 indicating the lowest rank.

Table 8. Table of points for the ranking for *NRMSE* and *MAE*.

Method	MAX				MS				PCTL95				LD				SUM
	NRMSE mean	*MAE* mean	*NRMSE* max	*MAE* max	*NRMSE* mean	*MAE* mean	*NRMSE* max	*MAE* max	*NRMSE* mean	*MAE* mean	*NRMSE* max	*MAE* max	*NRMSE* mean	*MAE* mean	*NRMSE* max	*MAE* max	
v1	2	2	2	2	2	2	1	2	2	2	3	2	2	2	2	2	32
v1BI	3	3	3	3	3	3	2	3	3	3	4	3	3	3	3	3	48
v2dis	4	4	4	4	4	4	3	4	4	4	5	4	4	4	4	4	64
v2nn	1	1	6	1	1	1	4	1	5	5	6	5	1	1	1	1	41
v3IDW	5	5	1	5	6	6	6	6	1	1	1	1	6	6	6	6	68
v3TPS	6	6	5	6	5	5	5	5	6	6	2	6	5	5	5	5	83

3.10. The resulting method for determining interpolated fields of daily precipitation totals

The choice of the reference field is essential because historical simulations show a differentiated fit to the field of observations, which translates into the selection and correction of climate scenarios and, thus, the conclusions resulting from the selected climate simulations.

None of the interpolation methods were outrightly superior. Most calculations indicated the v2nn method as the most appropriate interpolation method. However, the v2nn method failed when there was too little data around the node. There were unacceptable situations in the data for this method when interpolated monthly sums of precipitation in several nodes were zero for several years. Some interpolation methods overestimated the number of precipitation situations, and the resulting field of interpolated values was non-non-realistically smooth. However, such methods are irreplaceable in the case of missing observations when the alternative is the inability to perform the interpolation. The solution may be to select the interpolation method for individual days based on, for example, the CR and/or the RO value.

Analyses were performed for three datasets, in which the interpolation method was selected for each day. The adopted three selection criteria were based on the RO, the average correspondence ratio (CR_SR) for the thresholds of 0.1, 1, 5,10, and 20 mm, and based on both indicators together (RO_CR_SR). Three sets of complexes interpolated observational data RO, CR_SR, and RO_CR_SR were obtained by joining the appropriate fields interpolated for individual days. For the RO and CR_SR sets, the highest coefficient

value on a given day determined the choice of the interpolation method. For the RO_CR_SR set, the interpolation method was selected based on rankings of the daily values of the RO and CR_SR parameters for the considered interpolation methods. Table 9 presents the statistics for each of the chosen interpolation methods.

Table 9. Statistics (the correlation coefficient – RO, the average correspondence ratio – CR_SR, the correlation coefficient, and the average correspondence ratio – RO_CR_SR) for each chosen interpolation method for the sets of composite data from 1976-2005.

Percentage of days for which the method was selected						
Interpolated obs.	v1	v1BI	v2dis	v2nn	v3IDW	v3TPS
RO	11.7	13.4	8.6	47.0	14.1	5.2
CR_SR	7.7	5.8	4.8	72.7	3.6	5.4
RO_CR_SR	26.3	9.4	7.6	51.3	3.8	1.6

For each of the adopted criteria, the v2nn method – the nearest neighborhood – was chosen most often, in the case of the CR_SR criterion, as much as 72.7% of the time. For the datasets obtained in this way, an analysis of the fit was performed for the monthly sum (MS) of rainfall and the maximum rainfall in the month MAX for 102 stations.

For the MS of precipitation for all complex sets, the 75th percentile of *NRMSE* was below 0.3.

Table 10. Statistics (the correlation coefficient – RO_MS, the average correspondence ratio – CR_SR_MS, the correlation coefficient of the average correspondence ratio – RO_CR_SR_MS) for the monthly sum (MS) of precipitation in composite sets.

NRMSE						
Interpolated obs.	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
RO_MS	0.03	0.16	0.24	0.24	0.28	1.19
CR_SR_MS	0.02	0.14	0.23	0.24	0.30	1.19
RO_CR_SR_MS	0.02	0.15	0.23	0.24	0.28	1.19
MAE						
Interpolated obs.	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
RO_MS	0.6	3.4	5.7	6.5	7.3	50.1
CR_SR_MS	0.5	2.9	5.2	6.2	7.0	50.2
RO_CR_SR_MS	0.5	3.2	5.4	6.3	7.1	49.8

The maximum value for the *NRMSE* was greater than 1, but the value of NRSME=1.19 was only for one station (Czestochowa, 350190550), for the rest of this value is less than 0.76 (Table 10). The *MAE* statistics improved significantly, and the mean value was approximately 6 (previously ranging from 5 to 9.5). The maximum value of the *MAE* of approximately 50 mm was reached for the mountain station (Kasprowy Wierch, 1987 m above sea level). For the remaining analyzed stations, it did not exceed 27 mm.

Table 11 contains statistics *NRMSE* and *MAE* for the maximum daily sum of monthly precipitation. The maximum *NRMSE* achieved for the Czestochowa station slightly exceeded 1; otherwise, this value was less

than 0.71 (previously 1.78 to 2.82). The maximum values of *MAE* were around 11 mm, previously ranging from 8 to 12. The 75th percentile for *MAE* did not exceed 3 mm, and the median and mean values were around 2 mm.

Table 11. Statistics of the fit error for the maximum daily sum of precipitation MAX for composite interpolation sets.

NRMSE						
Interpolated obs.	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
RO_MAX	0.04	0.20	0.29	0.31	0.42	1.04
CR_SR_MAX	0.03	0.17	0.28	0.31	0.43	1.04
RO_CR_SR_MAX	0.03	0.19	0.30	0.31	0.42	1.03
MAE						
Interpolated obs.	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
RO_MAX	0.2	1.3	1.9	2.2	2.7	11.0
CR_SR_MAX	0.2	1.2	1.9	2.2	2.7	10.8
RO_CR_SR_MAX	0.2	1.4	1.9	2.2	2.7	10.9

The average value of the CR for complex sets concerning individual interpolation methods was also analyzed. Figure 10 shows differences in the mean values of CR.

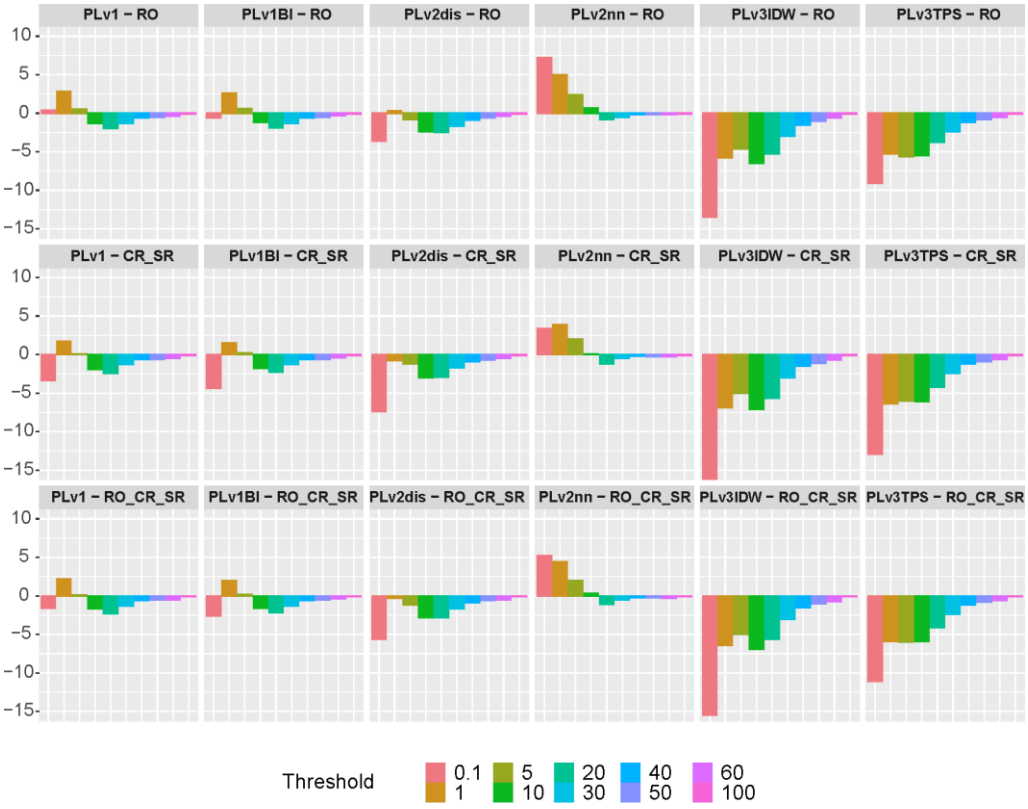


Fig. 10. Differences in the mean correspondence ratio (CR) value for interpolation methods and complex sets RO, CR_SR, and RO_CR_SR. Rows refer to the complex sets RO, CR_SR, and RO_CR_SR. Columns correspond to the interpolation methods.

In most cases, the CR was greater for the sets RO, CR_SR, and RO_CR_SR than for individual interpolation methods. The exception was the nearest neighborhood method v2nn, for which the difference CR was positive for the thresholds 0.1, 1, 5, and 10. For the v3IDW and v3TPS methods, which were selected the least frequently in the result set according to the criteria based on RO and CR, the difference in the mean CR value was always negative. For the linear methods v1 and v1BI, the CR value was a few percent higher than for the complex sets only for the 1 mm threshold.

4. Discussion

The choice of the reference field is essential because the historical simulations show a differentiated fit to the field of observations. This translates into the selection and correction of climate scenarios, and thus, the conclusions resulting from the selected climate simulations.

The series of observational data differ in the amount and quality of data for each day of the period under consideration. In turn, interpolation methods react differently to this input data variability, and some are not disturbed by even a very small number of observations. As shown in the works cited (Sheffield et al. 2006; Prasad, Sushma 2016; Herrera et al. 2018; Crespi et al. 2019), these two factors significantly impact uncertainty in gridded precipitation datasets.

A comparative analysis using different statistical parameters for several selected interpolation methods for the entire period did not yield a 'best method'. The v2nn method was chosen most frequently; but, for several years, for several nodes, the monthly sum of daily precipitation was zero. Other interpolation methods significantly overestimated the area of precipitation occurrence, inflated the extreme values, or the resulting field of interpolated values was non-non-realistically smooth. Correlation and correspondence ratio are important indicators of the quality of interpolated precipitation data, so a composite method based on both indicators was chosen. Three gridded datasets, RO, CR_SR, and RO_CR_SR, were prepared, in which the interpolation method was selected based on the values of the daily coefficients RO, CR, RO, and CR, respectively. The analysis comparing the monthly rainfall sums, the maximum daily sum in a month, and the correspondence ratio showed that the sets RO, CR_SR, and RO_CR_SR allowed for constructing more reliable data than the sets obtained using individual interpolation methods.

This approach allowed for the determination of gridded data based on different ways of selecting interpolation methods. Precipitation analysis concerns both the amount of precipitation and the area of occurrence of this phenomenon, which can be described using indicators such as averaged CR and RO.

Interpolated precipitation data with a resolution of 5 km and 2 km (denoted as CHASE_5KM and CHASE_2KM, respectively) were provided in the CPLFD-GDPT5 (Berezowski et al. 2016) and G2DC-PLC (Piniewski et al. 2021) projects. For selected locations, data from 1976-2005 were chosen and compared with analogous time series for datasets discussed in this paper (denoted as Comp_RO, Comp_CR_SR, and Comp_RO_CR_SR). The comparison was based on the MAE value for the annual

sum of daily precipitation. The values of this error for all years and models ranged from 35.4 mm to 136.4 mm. The *MAE* values averaged over the period 1976-2005 are presented in Table 12.

The average *MAE* value for CHASE_2KM was 6.6 mm smaller than that for CHASE_5KM. The values for the 'Comp' group of models were approximately 50% lower than those for the 'CHASE' group. The comparison of the *MAE* for the annual sum of precipitation for each year of the period is presented in Figure 11.

Figure 11 shows that for all years, the error for the 'Comp' models was smaller than that for the 'CHASE' models. This suggests that choosing more appropriate basic interpolation methods for each day, based on the values of the daily coefficients RO and CR, can significantly improve the resulting datasets.

The variance inflation factor of the average annual totals for compared fields was also analyzed (Table 13). The group of 96 stations was characterized by a high squared deviation from the mean, proportional to the variance value, SS_{tot} (2.8) was 3 203 554.

As a result, the obtained *VIF* values ranged from 2.3 to 4.9, with values above 4.5 referring to the data obtained in this study. The application of the interpolated field selection method for individual days of the period resulted in the decrease of the *VIF* value below the value of 5.

Fig. 11. Mean absolute error (*MAE*) for the annual sum of precipitation for each year from 1976-2005.

Table 12. Mean absolute error (*MAE*) for the annual sum of precipitation averaged over the period 1976-2005.

Model	CHASE_2KM	CHASE_5KM	Comp_CR_SR	Comp_RO	Comp_RO_CR_SR
Mean of *MAE*	99.4	106.0	49.1	53.3	50.1

Table 13. The variance inflation factor, the coefficient of determination, and the sum of squares of residuals for the mean annual sum of precipitation for resulting output methods of this work and data from the CHASE projects for 1976-2005.

Interpolated field	Variance Inflation Factor (*VIF*)	Coefficient of determination (R^2)	Sum of squares of residuals (SS_{res})
CHASE_2KM	2.9	0.66	1 089 791.4
CHASE_5KM	2.3	0.56	1 413 943.4
Comp_CR_SR	4.9	0.80	652 693.7
Comp_RO_	4.7	0.79	681 438.2
Comp_RO_CR_SR	4.9	0.80	647 471.0

5. Conclusions

This study introduced an innovative algorithm to determine the most appropriate interpolation method for each day based on the field fit characteristics at selected points. The selection procedure was based on the CR values (assessing compliance of areas with a given daily precipitation threshold) and /or RO (assessing the linear correlation). It was assumed that the potential uncertainty related to the different interpolation approaches used for each day would be compensated by greater compliance with rainfall areas and by maintaining a linear correlation. Among other things, the values of the variance inflation factor (*VIF*) were compared for the interpolation methods (Table 2), the three sets of complexes interpolated observational data (RO, CR_SR, and RO_CR_SR), and datasets CHASE from the CPLFD-GDPT5 (Berezowski et al. 2016) and G2DC-PLC (Piniewski et al. 2021) projects. The range of *VIF* values for the interpolation methods was from 3.2 to 6.3. For the sets of complexes interpolated observational data, *VIF* values were less than 5, which is lower than for most interpolation methods. For CHASE datasets, *VIF* was 2.3 for 5 km resolution and 2.9 for 2 km resolution. However, when comparing the latter two groups of datasets using the mean absolute error *MAE*, the average *MAE* error for the data resulting from this work was found to be 50% smaller. Our results provide evidence that expanding the range of available interpolation methods for each day and expanding the selection algorithm based on precipitation characteristics can significantly improve the resulting precipitation fields.

Author Contributions: Conceptualization, K.K., J.W.; methodology, K.K J.W.; software, K.K.; validation, K.K.; formal analysis, K.K.; investigation, K.K.; resources, K.K; data curation, K.K.; writing – original draft preparation, K.K.; writing – review and editing, K.K., JW, M.G.; visualization, K.K., M.G.; supervision, J.W.; project administration, K.K., M.G.; funding acquisition, M.G, K.K.

Funding: The study was performed as part of the research project "Methods of precipitation verification for operational applications, climatological studies and in climate change modeling (DS–4/2021)", financed by the Ministry of Science and Higher Education (Poland), the statutory activity of the Institute of Meteorology and Water Management–National Research Institute in 2021.

Data Availability Statement: Publicly available datasets were analyzed in this study. These data can be found here: https://danepubliczne.imgw.pl/datastore (accessed on 2 November 2021).

References

Akima H., Gebhardt A., 2022, Interpolation of Irregularly and Regularly Spaced Data, Package "akima", version 0.6-3.4, available at https://cran.r-project.org/web/packages/akima/index.html (data access 08.09.2023).

Belo-Pereira M., Dutra E., Viterbo P., 2011, Evaluation of global precipitation data sets over the Iberian Peninsula, Journal of Geophysical Research, 116, D20101, DOI: 10.1029/2010JD015481.

Benestad R., Buonomo E., Gutiérrez J.M., Haensler A., Hennemuth B., Illy T., Jacob D., Keup-Thiel E., Katragkou E., Kotlarski S., Nikulin G., Otto J., Rechid D., Remke T., Sieck K., Sobolowski S., Szabó P., Szépszó P., Teichmann C., Vautard R., Weber T., Zsebeházi G., 2021, Guidance for EURO-CORDEX climate projections data use, Version 1.1 – 2021.02, available at https://www.euro-cordex.net/imperia/md/content/csc/cordex/guidance_for_euro-cordex_climate_projections_data_use__2021-02_1_.pdf (data access 08.09.2023).

Benestad R., Haensler A., Hennemuth B., Illy T., Jacob D., Keup-Thiel E., Kotlarski S., Nikulin G., Otto J., Rechid D., Sieck K., Sobolowski S., Szabó P., Szépszó P., Teichmann C., Vautard R., Weber T., Zsebeházi G., 2017, Guidance for EURO-CORDEX climate projections data use, Version 1.0 – 2017.08, available at https://www.euro-cordex.net/imperia/md/content/csc/cordex/euro-cordex-guidelines-version1.0-2017.08.pdf (data access 08.09.2023).

Berezowski T., Szcześniak M., Kardel I., Michałowski R., Okruszko T., Mezghani A., Piniewski M., 2016, CPLFD-GDPT5: High-resolution gridded daily precipitation and temperature data set for two largest Polish river basins, Earth System Science Data, 8 (1), 127-139, DOI: 10.5194/essd-8-127-2016.

Cornes R.C., van der Schrier G., van den Besselaar E.J.M., Jones P.D., 2018, An ensemble version of the E-OBS temperature and precipitation data sets, Journal of Geophysical Research: Atmospheres, 123 (17), 9391-9409, DOI: 10.1029/2017JD028200.

Crespi A., Lussana C., Brunetti M., Dobler A., Maugeri M., Tveito O.E., 2019, High resolution monthly precipitation climatologies over Norway (1981-2010): Joining numerical model data sets and in situ observations, International Journal of Climatology, 39 (4), 2057-2070, DOI: 10.1002/joc.5933.

Daly C., Slater M.E., Roberti J.A., Laseter S.H., Swift L.W., 2017, High-resolution precipitation mapping in a mountainous watershed: ground truth for evaluating uncertainty in a national precipitation dataset, International Journal of Climatology, 37 (S1), 124-137, DOI: 10.1002/joc.4986.

Dee D.P., Uppala S.M., Simmons A.J., Berrisford P., Poli P., Kobayashi S., Andrae U., Balmaseda M.A., Balsamo G., Bauer P., Bechtold P., Beljaars A.C.M., van de Berg L., Bidlot J., Bormann N., Delsol C., Dragani R., Fuentes M., Geer A.J., Haimberger L., Healy S.B., Hersbach H., Hólm E.V., Isaksen L., Kållberg P., Köhler M., Matricardi M., Mcnally A.P., Monge-Sanz B.M., Morcrette J.-J., Park B.-K., Peubey C., de Rosnay P., Tavolato C., Thépaut J.-N., Vitart F., 2011, The ERA-Interim reanalysis: configuration and performance of the data assimilation system, Quarterly Journal of the Royal Meteorological Society, 137 (656), 553-597, DOI: 10.1002/qj.828.

Deque M., Rowell D.P., Luthi D., Giorgi F., Christensen J.H., Rockel B., Jacob D., Kjellstrom E., de Castro M., Van Den Hurk B., 2007, An intercomparison of regional climate simulations for Europe: assessing uncertainties in model projections, Climatic Change, 81 (1):53-70, DOI: 10.1007/s10584-006-9228-x.

Deque M., Somot S., Sanchez-Gomez E., Goodess C.M., Jacob D., Lenderink G., Christensen O.B., 2012, The spread amongst ENSEMBLES regional scenarios: regional climate models, driving general circulation models and interannual variability, Climate Dynamics, 38 (5), 951-964, DOI: 10.1007/s00382-011-1053-x.

FAOSTAT, 2022, Climate Change Data: Annual Surface Temperature Change, available at https://climatedata.imf.org/pages/climatechange-data, (data access 08.09.2023).

Giorgi F., Jones C., Asrar G.R., 2009, Addressing climate information needs at the regional level: the CORDEX framework, WMO Bulletin, 58 (3).

Gleckler P.J., Taylor K. E., Doutriaux C., 2008, Performance metrics for climate models, Journal of Geophysical Research, 113 (D6), DOI: 10.1029/2007JD008972.

Gräler B., Pebesma E., Heuvelink G., 2016, Spatio-Temporal Interpolation using gstat, The R Journal, 8 (1), 204-218, DOI: 10.32614/RJ-2016-014.

Herrera S., Kotlarski S., Soares P.M.M., Cardoso R.M., Jaczewski A., Gutiérrez J.M., Maraun D., 2018, Uncertainty in gridded precipitation products: Influence of station density, interpolation method and grid resolution, International Journal of Climatology, 39 (9), 3717-3729, DOI: 10.1002/joc.5878.

IMGW-PIB, 2020, Bulletin of National Meteorological and Hydrological Services, (in Polish), Institute of Meteorology and Water Management – National Research, available at https://danepubliczne.imgw.pl/data/dane_pomiarowo_obserwacyjne/Biuletyn_PSHM/Biuletyn_PSHM_2019_ROCZNY.pdf (data access 08.09.2023).

IPCC, 2014, Climate Change 2014: Synthesis Report, Contribution of Working Groups I, II and III to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change, Core Writing Team, R.K. Pachauri, L.A. Meyer (eds.), IPCC, Geneva, Switzerland, 151 pp.

Jacob D., Petersen J., Eggert B., Alias A., Bøssing Christensen O., Bouwer L.M., Braun A., Colette A., Déqué M., Georgievski G., Georgopoulou E., Gobiet A., Menut L., Nikulin G., Haensler A., Hempelmann N., Jones C., Keuler K., Kovats S., Kröner N., Kotlarski S., Kriegsmann A., Martin E., van Meijgaard E., Moseley C., Pfeifer S., Preuschmann S., Radermacher C., Radtke K., Rechid D., Rounsevell M., Samuelsson P., Somot S., Soussana J.-F., Teichmann C., Valentini R., Vautard R., Weber B., Yiou P., 2014, EURO-CORDEX: new high-resolution climate change projections for European impact research, Regional Environmental Change, 14, 563-578, DOI: 10.1007/s10113-013-0499-2.

Jones P.W., 1998, A User's Guide for SCRIP: A Spherical Coordinate Remapping and Interpolation Package, available at https://raw.githubusercontent.com/wiki/SCRIP-Project/SCRIP/files/SCRIPusers.pdf (data access 08.09.2023).

Karger D.N., Conrad O., Böhner J., Kawohl T., Kreft H., Soria-Auza R.W., Zimmermann N.E., Linder H.P., Kessler M., 2017, Climatologies at high resolution for the earth's land surface areas, Scientific Data, 4, 170122, DOI: 10.1038/sdata.2017.122.

Konca-Kędzierska K., 2019, Evaluation of the precipitation field reconstructed in the climate models from the EURO-CORDEX project for the Polish domain in the 1978-2005 period, (in Polish), [in:] Współczesne problem klimatu Polski, L. Chojnacka-Ożga, H. Lorenc (ed.), Warszawa, IMGW-PIB, 173-185,

NOAA, 2022, Climate Change: Global Temperature, available at https://www.climate.gov/news-features/understanding-climate/climate-change-global-temperature (data access 08.09.2023).

Nychka D., Furrer R., Paige J., Sain S., 2017, fields: Tools for spatial data, R package version 11.6, DOI: 10.5065/D6W957CT, available at https://github.com/NCAR/Fields (data access 08.09.2023).

Otto S.A., 2019, How to normalize the RMSE, available at www.marinedatascience.co/blog/2019/01/07/normalizing-the-rmse/ (data access 08.09.2023).

Pebesma E.J., 2004, Multivariable geostatistics in S: the gstat package, Computers & Geosciences, 30 (7), 683-691, DOI: 10.1016/j.cageo.2004.03.012.

Piniewski M., Szcześniak M., Kardel I., Chattopadhyay S., Berezowski T., 2021, G2DC-PLC: a gridded 2 km daily climate dataset for the union of the Polish territory and the Vistula and Odra basins, Earth System Science Data, 13, 1273-1288, DOI: 10.5194/essd-13-1273-2021.

Prasad M.S.G., Sushma N., 2016, Spatial prediction of rainfall using universal kriging method: a case study of Mysuru District, International Journal of Engineering Research & Technology – Geospatial, 4 (20), DOI: 10.17577/IJERTCONV4IS20010.

R Core Team, 2018, R: A language and environment for statistical computing, R Foundation for Statistical Computing, Vienna, Austria, available at https://www.R-project.org/ (data access 08.09.2023).

Schulzweida U., 2019, CDO User Guide (Version 1.9.8), DOI: 10.5281/zenodo.3539275.

Sheffield J., Goteti G., Wood E.F., 2006, Development of a 50-year high-resolution global dataset of meteorological forcings for land surface modeling, Journal of Climate, 19, 3088-3111, DOI: 10.1175/JCLI3790.1.

von Storch H., Zwiers F.W., 1999, Statistical Analysis in Climate Research, Cambridge University Press, Cambridge, DOI: 10.1017/CBO9780511612336.

Hydrologic drought characteristics of selected basins in various climate zones of Lebanon

El Tayara-Zobaida

Lebanese University

Abstract

River basins in Lebanon recently have experienced increasing droughts, which has prompted this study to characterize drought temporally and spatially. The study describes and analyzes hydrologic and precipitation conditions in seven river basins, representing most flow directions in various climatic zones. The characteristics of hydrologic and rainfall drought were discussed and analyzed, depending on available data from five climatic zones and fourteen hydrometric stations distributed in the river, allowing for a detailed analysis of drought. The Standardized Precipitation Index (*SPI*) and the Streamflow Drought Index (*SDI*) were calculated at 6-month intervals (first and second 6-months) using the DRINC program. The hydrologic and rainfall drought characteristics maps generated in the GIS platform may help to identify the degree of drought in the study areas. The investigation was carried out by examining the strength of relationships between *SDI* and *SPI* using bivariate correlation analysis. The significance of the correlation coefficient is used in this study to decide whether linear relationships between the *SPI* and *SDI* occurred in the first and second six months. Calculating the correlation coefficient for these variables based on hydrologic and rainfall data reveals an inconsistent correlation over different periods.

Keywords

Drought characteristics mapping, streamflow drought index, standardized precipitation index, basin, Lebanon.

Submitted 22 January 2023, revised 18 August 2023, accepted 5 October 2023

DOI: 10.26491/mhwm/174194

1. Introduction

Lebanon has a Mediterranean climate and is vulnerable to hydrologic and rainfall droughts in most of its basins. Hydrologic and rainfall droughts will be some of the biggest natural disasters that future generations will face because average rainfall in Lebanon has declined from 740 mm (1901-1930) to 638 mm (1991-2020). This reduction caused a significant decrease in runoff; the annual flow in Lebanon (total streamflow) was 4,300 Mm^3/year in 1970, dropping to 3,171 Mm^3/year in 2009, accompanied by depletion of groundwater levels and declines in many wells and springs[1].

The purpose of the present study is not to show the onset of drought in the Lebanese basins. As Tannehill (1947) wrote more than seventy years ago: "Drought has no beginning of it to have an end." Rather, the purpose is to examine the drought characteristics in Lebanon through *SDI* and *SPI* indices and to test the significance of the correlations between them, over seven selected basins that represent the different climatic zones. The adoption of data in certain basins of Lebanon does not imply that the drought began on any

[1] https://unfccc.int/sites/default/files/resource/lebanon_snc.pdf.

specific date. However, relying on secular data can enable projections for the future by using arithmetic methods and models.

Hydrological drought[2] is the decrease of water in all elements of the hydrological cycle. It means less water flowing through rivers and stored in lakes, but also lower levels of groundwater flow. Hydrological droughts are complex, recurring hazards that can cause water shortages in streams or storage entities such as reservoirs, lakes, and groundwater. Hydrological drought occurs when low water supply becomes evident, especially in streams and groundwater, usually after many months of meteorological drought. Meteorological drought[3] is defined as a shortfall of precipitation over some period of time. It is measured by comparing the amount of precipitation, e.g., for a season or a year, to an average over a longer time series (e.g., many seasons or years).

Hydrological drought is based on the impact of rainfall deficits (including snowfall) in a specific area (as a basin) and time (as month, season, and year) on the water supply, such as streamflow, reservoir and lake levels, and groundwater table (Wilhite, Glantz 1985). The climatic zone must be considered because rainfall varies widely from one climatic zone to another (The National Weather Service).

The question is, to what extent has hydrological drought affected the basins of Lebanon over the years? To answer this question, this research entails discussing and analyzing hydrologic pluvial droughts in some Lebanese basins and climatic zones during periods when data are available. The droughts are characterized by calculating the correlation between hydrologic pluvial droughts and their spatiotemporal variability based on basin characteristics.

An extensive rainfall dataset for five climate zones (1901 to 2020) was available. Water measurements are not available over such long periods; some of these measurements began in 1939, and others more recently. The overall range of drainage measurement in the selected basins is from 1939 to 2020, beginning in 1939 (El Litany), 1949 (El Awali western drainage), 1966 (El Bared), 1991 (El Assi), 1994 (Ed Damour), 1995 (Abu Ali), and 2002 (El Hassbani). All are based on thematic information for each basin, e.g., topography, mean flow, and precipitation.

2. Study area

The basins were selected based on their different flow directions (north, south, and east) and their representation of different climatic zones (Fig. 1)[4]. The flow direction of El Bared, Abou Ali (located in the northern climatic zone), El Awali, and Ed Damour (located in the Mount Lebanon climatic zone) is westward, and that of El Assi (located in the Beqaa climatic zone) is northward. The El Hassbani and El

[2] University of Nebraska (NDMC) National Drought Mitigation Center.
[3] Https://edo.jrc.ec.europa.eu/edov2/html/1001.html#:~.
[4] NB: The research relied only on hydrologic data from the stations that had a long measurement period and have continued to date (Table 3).

Litani (located in the Beqaa, Nabatieh climatic zones, and South climatic zone for El Litani only), flow southward, and then the Litani turns westward. The river basins selected all have permanent flows (Table 1).

Table 1. Some characteristics of the 7 selected basins in Lebanon.

Variable	El Bared	Abu Ali	Ed Damour	El Awali	El Assi	El Litany	El Hasbani
Basin area (km^2)	262	476	302	298	1370	2153	578
Max altitude (m)	2877	3093	1863	1945	2750	2543	2807
Alt. of stream gauge (m)	560	1092	360	500	918	1304	1000
Slight slope (%)	6.3	14.3	26.9	0.87	35.1	33.7	15.9
Strong slope (%)	18.5	21.1	22.8	0.198	24.3	23.7	24.3
Very strong slope (%)	41.6	37.4	32.6	0.418	27.3	29.1	40.2
Very steep slope (%)	33.6	27.2	17.7	0.298	13.3	13.4	19.6
Mean stream slope (m/km)	13	54	51	36	19	5	14
Drainage density (km/km^2)	1.00	1.75	5.30	14.35	1.27	0.84	1.14
Average annual precipitation (Mm^3/year)	225	505	335	320	1254	2078	598
Average winter rainfall (Dec.-Mar.) 1901-2020 (mm)	636	636	656	656	582	582	613
Source (m^3/s)	Shoukkar	Kadisha	Al-Safa	1 & 2	3 & 4	5 & 6	7 & 8
	1.15	0.72	1.42	0.48, 1.58	2.44, 0.95	31.8, 38.5	1.21, 1.9
Recording of hydrological data	1966/67 – 2019/20	1965/66 – 2019/20	1994/95 – 2019/20)	1979/80 – 2019/20	(1990/91 – 2018/19)	1939/40 – 2017/18	1992/93 – 2017/18
Discharge at mouth (Mm^3/year)	134.8	208.5	183	393.9	372.1*	379.3**	193.5
Average Base Flow Index (BFI = Base flow volume/Total flow volume)	Sea Mouth 0.19	Abou Samra 0.18 Kousba 0.15	Jisr El Qadi 0.09 Es Safa 0.24 Sea Mouth 0.01	Sea Mouth 0.35 Marj Bisri 0.11	Hermel 0.72	Qasmieh 0.02	Qaroun 0.04 Khardale 0.27

1 – Jezzin; 2 – El-Barouk; 3 – Ain Ez-Zarqa; 4 – Labouh; 5 – Anjar; 6 – Ez-Zarqa, 7 – Hasbani; 8 – Wazzani.

*El Assi at Hermel; **Litani at Khardale; ***After Wazzani spring.

NB: "Flows are never natural due to human interventions and manipulations: pumping, influence of dams, over- or underestimated, or calculated, interpolated flows". (FAO 1974).

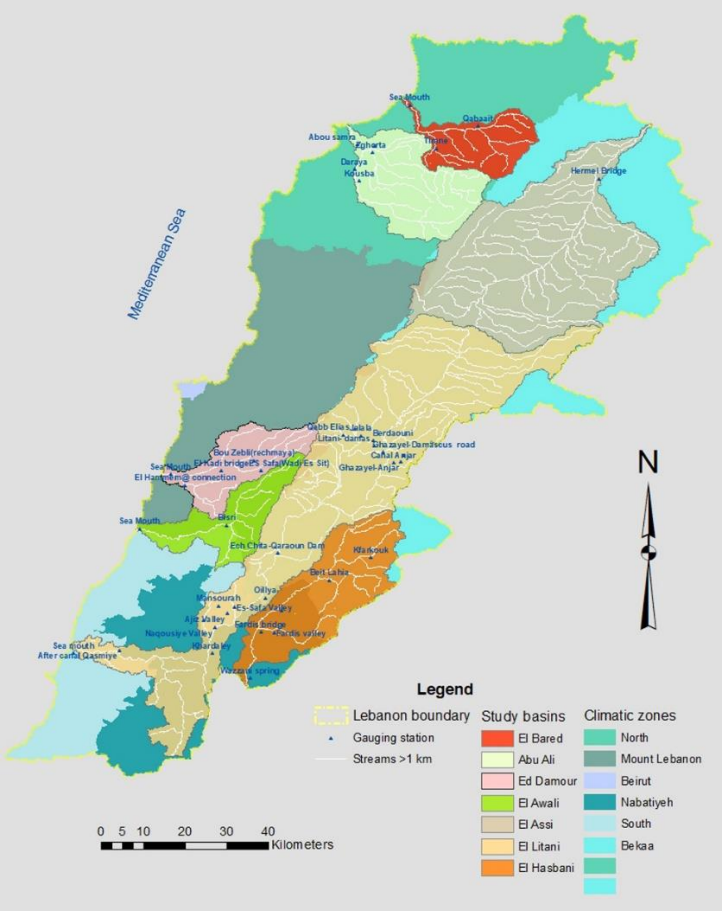


Fig. 1. Locations of the studied basins and climate zones of Lebanon.

2.1. Physical characteristics of the selected basins

The topography of the major river basins is described by the Digital Elevation Model (DEM), which was obtained from "ASTER GDEM is a product of METI and NASA" with a spatial resolution of 30 m × 30 m. The river network and basin boundary are automatically extracted from the DEM itself with ArcGIS software. Accordingly, the maximum altitude of the basins is from 3,093 m at Abu Ali to 1,863 m at Ed Damour. Slight slopes are 35.1% and 33.7% at El Litani and El Assi, respectively. El Bared, Abu Ali, El Damour, El Awali, and El Hasbani have strong slopes, ranging from 32.6 to 41.8%.

A brief description of the characteristics of rivers of the selected basins follows. El Bared originates from Jurd al-Danniyeh. The three main springs are the Sukkar (1.15 m^3/s), Sanubar, and Brisa. The river has a confluence with the Mussa River. Abu Ali originates from the Qadisha (0.72 m^3/s) and Thahre El Qadib in Jabal El Makmel.

Ed Damour descends from the springs of El Safa (1.42 m^3/s) and those of Jabal El-Barouk and is characterized by permanent flow and many valleys. El Awali arises mainly from El Barouk Spring (1.58 m^3/s) in Jebal El Barouk-Niha. It also receives water from the artificial Lake Qanan.

El Litani is the longest Lebanese river (174 km, entirely inside Lebanon), extending over the Beqaa Plain and to the south, then heading westward to the sea. The principal springs of El Litani are Ez Zarka (38.5 m^3/s) and Anjar (31.8 m^3/s). Hasbani is a transboundary river, a major tributary of the Jordan River, descended from Jabal Hermon and runs about 25 km in Lebanon. The principal springs are El Hasbani (1.21 m^3/s) and El Wazzani (1.9 m^3/s); other springs are located in the southeastern part of Beqaa and flow southwards.

El Assi is also a transboundary river shared by Lebanon, Syria, and Turkey, it is about 33 km in length in Lebanon, and the main springs are Ain Zarqa (2.44 m^3/s) and Labouh (0.95 m^3/s).

2.2. Hydrological drought history of the basins studied

The signs of drought include decreases in the minimum average of the base flow index, such as those recorded in Table 1. El Damour is an example at the Sea Mouth gauge with an index of 0.01.

The historical drought study is based on the available data for hydrological measurement periods ranging from 15 to 79 years (Table 2). This table shows that, in downstream coastal rivers, drought events occurred 4 to 23 times, with flows falling an average of 30 to 88%. For the interior rivers, drought events occurred 7-36 times, with the flow rate dropping 30-91% below average. It must be said that flow measurements, at most gauging stations, are "observed, not natural" flows due to human interventions and manipulations such as pumping and the influence of dams that lead to under- or over-estimated and calculated flows (FAO 1974).

Table 2. Hydrological drought history of 7 rivers selected in Lebanon.

	Name	Station	Periods	Number of years	Average (m^3/s)	Number of years with 30% below the average	Average below 30 to 91% (m^3/s)	Range of average below 30% (%)
Coastal rivers	El Bared	Sea mouth	1966/67-1972/73 1995/96-2019/20	7 25	4.61	10	2.9-1.60	36-65
	Abu Ali	Abu Samra	1995/96-2019-20	25	6.98	4	4.2-2.57	38-62
	Ed Damour	Sea mouth	1994/95-2019/20	26	5.39	7	3.7-1.39	31-74
	El Awali	Saida	1949/50-1972/73 1991/92-2019-20	24 29	11.1	23	7.8-1.33	30-88
Interior rivers	El Assi	Hermel	1991/92-2018/19	28	11.78	8	8.2-4.67	30-60
	El Hasbani	Aft. spring	2002/03-2017/18	15	3.91	7	2.3-0.99	40-75
	El Litani	Khardale	1939/40-2017/18	79	12.03	36	8.2-1.10	32-91

3. Availability of hydrologic and rainfall data in the study area

The *SDI* index was computed for fourteen hydrological stations with different measurement periods ranging from 1931 to 2020, and the *SPI* was calculated for five climatic zones with data covering 1901-2020. For comparison of annual drainage and precipitation in a basin, the periods were matched for the two types of data.

3.1. Rainfall characteristics of 7 climatic zones

Climate zone identification and the rainfall dataset are based on the work of the World Bank[5], which presents high-level information on Lebanon's climate zones and the seasonal cycle of precipitation for the latest climate data 1901-2020. Climate zone classifications are derived from the Koppen-Geiger climate classification system based on seasonal rainfall and temperature patterns, which separates six main climatic areas: Beirut, Mt. Lebanon, North, Beqaa, South, and Nabatieh.

The average annual rainfall declined over a century from 719 to 651 mm, with 60% falling in the rainy season (Nov.-May). According to FAO (2018), the rainfall has decreased by 40-50%, and as a result, many springs and wells have dried up. It should be noted that the decrease in rainfall affects all seven climatic zones but to different extents. Many studies attribute the decrease in runoff rates to a reduction in rainfall and snowfall (Haddad et al. 2014) that affects the recharge of the groundwater table and the springs.

3.2. Annual and seasonal average values of rainfall for climatic zones

Statistical analysis of the annual and seasonal average rainfall for the climatic zones in Lebanon shows that the second season (December, January, and February) has the most abundant precipitation (Fig. 2). Note that the annual and seasonal average rainfall in the Beqaa climatic zone has the lowest average without calculating the amount of equivalent snow water. The lowest average is recorded in June, July, and August. The annual and monthly rainfall for all zones has the characteristics of a Mediterranean climate, with the highest total rainfall recorded in Jan. The lowest rainfall total is observed in Aug. About 70% of the annual rainfall occurs in the winter months, from December to February, and 21-24% in spring and autumn.

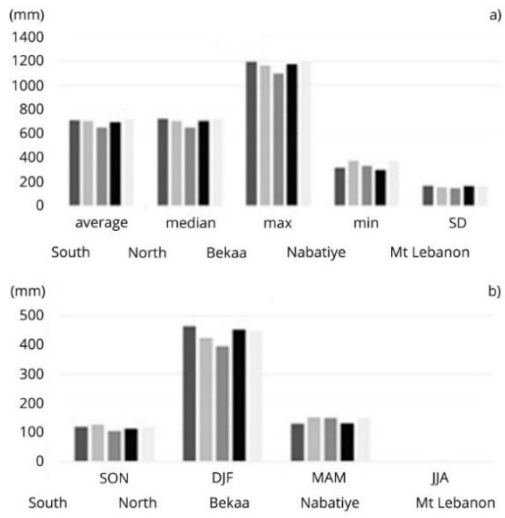


Fig. 2. Annual statistics (a) and seasonal average (b) rainfall of climatic zones in Lebanon (1931-2020) (reference: World Bank: Climate Change Knowledge Portal).

[5] https://climateknowledgeportal.worldbank.org/.

3.3. Hydrological statistics of the river basins

The Hydrological Service of Office National of El Litani provides hydrometric data for different periods at the stations of seven basins (Table 3). Fourteen of the 29 gauging stations were accredited by the long measurement period, despite a gap for some years (especially 1975-1989), and 15 were rejected due to the short duration of the measurement. A period of 25 to 70 years was selected for the hydrological drought study based on data quality, recorded length, and area coverage. For each basin, a discharge time series was available for various periods between 1931 and 2020 (Fig. 3).

Table 3. The average annual flow of the gauging stations and the corresponding recording date of 7 rivers in Lebanon.

Basin	Stations[6]	Area (km^2)	Elva (m)	Latitude DD	Longitude DD	Recorded years	Discharge (Mm^3)	Runoff (mm)
El Bared	Tirane	40	510	34.405	36.03	1967-1969	116.2	2916
	Qabaait	161	390	34.4533	36.12	1967-1969	114	708
	Sea Mouth	**262**	**0.5**	**34.4983**	**35.9733**	**1960-2020**	**149.07**	**569**
Abu Ali	Tirane	40	510	34.405	36.03	1967-1969	116.2	2916
	Qabaait	161	390	34.4533	36.12	1967-1969	114	708
	Sea Mouth	**262**	**0.5**	**34.4983**	**35.9733**	**1960-2020**	**149.07**	**569**
Ed Damour	**EL Safa**	**40**	**518**	**33.7133**	**35.6533**	**1960-1979 1990-2001**	**30.33**	**758**
	Rachmaya	52	447	33.735	35.6383	1960-2001	55.43	1066
	El Hamam	77	45	33.6816	35.49	1966-1973	36.78	478
	Jisr El Qadi	**185**	**250**	**33.7133**	**35.5666**	**1960-2001**	**127.52**	**689**
	Sea Mouth	**302**	**0**	**33.7061**	**35.4594**	**1994-2020**	**157.39**	**521**
El Awali	**Bisri**	**222**	**385.8**	**33.4233**	**35.5616**	**1950-1989 2001-2020**	**115**	**518**
	Saida -Sea M.	**298**	**3.5**	**33.4116**	**35.4083**	**1940-2020**	**352**	**1181**
El Assi	**Hermel**	**1370**	**585**	**34.34**	**36.38**	**1931-1979 1990-2019**	**369.00**	**269**
	Wazzani Spring	**-**	**271.2**	**33.2683**	**35.63**	**1960-1969 2003-2018**	**102.95**	-
	Before spring	340	548	33.4133	35.6966	1960-1979	25.73	76
	Fardis Bridge	448	494.6	33.3666	35.6533	1960-1979 2002-2018	60.29	135
	Aft. Wazzani spr.	**526**	**183**	**33.2572**	**35.6216**	**2002-2018**	**127.30**	**273**
El Litani	Qabb elias	19	914	33.79	35.83	1960-1979 1990-2001	18.76	987
	Jelala	22	890	33.7883	35.8683	1960-1979	6.67	303
	Berdaouni	77	866	33.7783	35.895	1950-2001	33.32	433
	Ghzayel	126	867	33.7533	35.9166	1990-2001	107.84	856
	Mansourah	1345	859.3	33.68	35.6616	1931-1969	247.05	184
	Qaroun dam	**1545**	**866**	**33.7666**	**35.8966**	**1939-1947 1969-2011**	**327.56**	**212**
	Quelieh	1680	521	33.44	35.6633	1930-2011	358.63	213
	Khardale	**1808**	**239.2**	**33.3216**	**35.5483**	**1931-2018**	**379.30**	**210**
	Ghandourah	2066	859.3	33.68	35.6616	1960-2012	298.50	144
	Qasmieh	**2153**	**0**	**33.3216**	**35.25**	**1960-1970 1990-2018**	**279.62**	**130**

[6] The research relied on the hydrological data of the stations in **bold**, which had a relatively long measurement period and have continued to date, to compare them with the rainfall data for the same period.

The streamflow stations in El Bared and Ed Damour basins have more missing data, and it was difficult to analyze drought with minimum data availability. For these two basins, the gauging stations used are El Bared at Sea Mouth, Abu Ali at Kousba, and Abou Samra, which have a minimum of 25 years of data. But the other five basins fulfill the minimum data length required (30 years) for drought analysis: Ed Damour at EsSafa, Jisr El-Qadi, and Sea Mouth. El Awali at Bisri and Sea Mouth, El Assi at Hermel, El Hasbani at Wazzani Spring and Fardis Bridge, El Litani at Qaroun Dam, Khardale, and Qasmieh.

The annual discharge at sea mouths of rivers El Bared, Abu Ali, Ed Damour, El Awali, and El Litani ranges between 149.07 and 379.3 Mm^3. The annual discharges of El-Hasbani and El-Assi before leaving the Lebanese territories are respectively between 127.3 and 369 Mm^3.

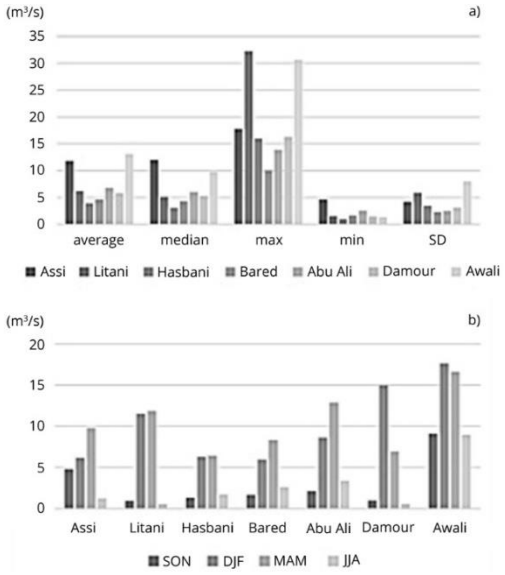


Fig. 3. Annual statistics (a) and seasonal average (b) of discharges of 7 basins in Lebanon (Cf. Table 3). The stations with long periods of measurement were adopted.

Comparing measuring stations, El Litani at Khardale has a high flow, followed by El Assi at El Hermel. El Litani is the only one of these rivers on which a dam is located (at Qaraoun). Nevertheless, the discharge at El Khardale is the most susceptible, as it has been decreasing dramatically (as shown later).

In four of seven basins (El Litani, El Hasbani, El Damour, and El Awali), the average runoff has one peak in the winter (in January, February or March) before snowmelt. In the three others (El Assi, El Bared, and Abu Ali), the average runoff has one peak in the spring (April). The lowest runoff for all rivers occurs in the summer, from July to September, except in El Hasbani and El Assi, where minimum flows are from October to December.

The drainage of rivers depends on seasonal rainfall, and it varies widely from the rainy to the dry season (Ed Damour, El-Hassbani). The drainage of rivers that depend on melting snow and water springs decreases the

difference between seasons (El Assi and El Bared). The drainage of the El Awali and El Litani rivers cannot be relied upon due to pumping (El Awali) or retention of the discharge (Qaraoun Dam) for technical reasons.

For each basin, one close-by climatic zone was chosen that seemed representative of the climatic conditions of the basin, and that was the most representative of topographical factors such as elevation and exposition.

4. Methodology

Managing water resources requires establishing drought characteristics in the form of published indices that are easy to interpret and can be applied to improve knowledge of the intensity and severity of drought.

Despite various ways of evaluating and defining hydrological drought, all are focused on the same issues with a time-step evaluation of the phenomenon (day, decade, month, season, and year). By accepting the definition of low flow for a period, the criterion for the threshold selection and separation of hydrological droughts from a series of low-flow events followed the classification proposed by Dracup et al. (1980).

Hence, the threshold level in this study depends on deviation from normal mean discharge for an average period of months, seasons, and years. Another distinguishing feature of drought is its duration, taking two to three months to become established, and possibly continuing for months or years. In a country with a small area, such as Lebanon, the entire country may be affected since droughts are usually regional phenomena; they result from large-scale anomalies in atmospheric circulation patterns that become established and persist over periods of months, seasons, or longer.

Drought is quantified in the form of simple indicators. Among these indicators[7], are two that will fit this threshold and have advantages:

1. The standardized precipitation index (*SPI*) is a statistical indicator comparing the total precipitation received at a particular location during n months with the long-term rainfall distribution of the same period at that location. It applies to all climate regimes.
2. The streamflow drought index (*SDI*) uses monthly streamflow values to monitor and identify drought events represented by a particular gauge. The advantage of this indicator is that missing data is allowed, that is, although some years may be missing (Table 3), as in this study, the longer the streamflow record, the more accurate the results. As with *SPI*, various timescales can be examined. The streamflow drought index is defined as a decrease in the amount of available water in all of its forms.

[7]Many indices of drought are in widespread use today, which is gaining increasing popularity in the USA, is the Standardized Precipitation Index (SPI) developed by McKee et al. (1993).

4.1. Assessment of hydrological and rainfall droughts in seven selected basins in Lebanon

4.1.1. Rainfall drought analysis using Standardized Precipitation Index (*SPI*)

The Standardized Precipitation Index (*SPI*) has been recognized as a standard index that should be used for quantifying and reporting rainfall drought; *SPI* is based on time series drought characteristics:

$$SPI = \frac{x_i - \bar{x}}{\sigma}$$

where: x_i is the monthly precipitation for the climatic zone (1901-2020); $\bar{x}$ is the mean of the monthly precipitation; σ is the standard deviation of the monthly precipitation.

The calculation of *SPI* reveals temporal and spatial relationships of series, allowing quantification and comparison with different durations and providing information on differences in drought behavior among climatic zones.

The monthly precipitation data were analyzed and processed by calculating the *SPI* using the Software DrinC[8]. This software for *SPI* is used to derive 6-months *SPI* values for each climatic zone, where the 6-monthly value gives intermediate-term drought. The drought categories are defined by the classification based on *SPI* (Edossa et al. 2009) and *SDI* (Hong et al. 2014). The result obtained from software based on 6-months *SPI* time series, the maximum drought severity that occurred in the two climatic zones of Nabatieh and south, was respectively –4.67 and –4.37 in the second 6-months (Apr.-Sep.) of the year 1989. The drought frequency (%) in the third category shows that the extreme drought (<–2.0) occurred in the first 6-months (Oct.-Mar.) in Nabatieh and in the second 6-months in the Beqaa (Table 4). The maximum moderate and severe drought frequency ranges between 0.62 (North) and 0.3 (Beqaa) in the first 6-months of the year.

Table 4. Drought frequency[9] in the severity category according to *SPI* of the climatic zones in Lebanon.

Climatic zones	Moderate (–1.0, –1.49)		Severe (–1.5, –1.99)		Extreme (<–2.0)	
	Oct.-Mar.	Apr.-Sept.	Oct.-Mar.	Apr.-Sept.	Oct.-Mar.	Apr.-Sept.
North	0.62	0.50	0.14	0.25	0.24	0.25
Mt Lebanon	0.55	0.44	0.25	0.31	0.20	0.25
Beqaa	0.50	0.38	0.30	0.31	0.20	0.31
South	0.53	0.38	0.29	0.38	0.18	0.25
Nabatieh	0.50	0.47	0.19	0.27	0.31	0.27

The probability of occurrence of drought, calculated in this study, for climate zones in South, Beqaa, Nabatieh, Mount Lebanon, and North, relative to the year and severity of events is:

Oct.-Mar. Apr.-Sept.

[8] www.ewra.net/drinc.
[9] Drought frequency of 1.

Moderate	1 in 15 to	1 in 20 years
Severe	1 in 20 to	1 in 40 years
Extreme	1 in 24 to	1 in 40 years

During 120 years in all climatic zones, drought ranged from 30% to 14%. In the north, a drought recurrence was recorded in 30.8% of the years, followed by Mt. Lebanon and the Beqaa with 29.8% of the years as a recurring drought. In the South and Nabatieh, 14% and 25.7% of drought recurrences were recorded.

The maximum of extreme drought recurrence was observed in the Beqaa for 4 years:1927-28,1965-66,1984-85 and 2009-10 (Apr.-Sep.), in the North for 5 years: 1932-33, 1959-60, 1972-73, 2007-08 and 2013-14 (Oct.-Mar.) and in the Nabatieh also for five years: 1932-33, 1950-51, 1959-60, 1978-79 and 1998-99 (Oct.-Mar.).

4.1.2. Hydrological drought analysis using the Streamflow Drought Index

The *SDI* method, developed by Nalbantis (2008), was used to characterize the hydrological drought events for the studied area. To compute the *SDI* values, the monthly observed flows of the time series are assumed. Its calculation is similar to *SPI* and, therefore, has the same characteristics of simplicity and efficiency. The *SDI* is based on monthly observed streamflow volumes at different time scales and thus offers the advantage of representing streamflow drought in the short, medium, and long term. The formula is (Gumus, Algin 2017):

$$SDI_{ij} = \sum_{j=6(k-1)+1}^{6k} Q_{ij} \qquad k = 1,2$$

where: Q_{ij} is river discharge for hydrological year (i), and month (j) within that hydrological year (October through September). Based on these series, the cumulative streamflow volume is computed where $k=1$ and $k=2$ are the first 6 months (Oct.-Mar.) and second 6 months (Apr.-Sept.) periods, respectively in a hydrological year.

The *SDI* values have been classified by (Hong et al. 2014) into eight classes that vary from extreme wet to extreme drought. The classes of wetness and dryness of *SDI* range between –2 and +2.

Hydrological drought is characterized by duration (D), severity (S), magnitude (M), and relative frequency (RF). Drought duration is the time between consecutive drought events (onset and end of drought). The duration (D) is from the initiation of a negative *SDI* until the flow returns to a positive *SDI* value (Tareke, Awoke 2022).

The relative drought frequency is the ratio of the number of droughts (n) with negative *SDI* in drought duration and the total number of drought years in the analysis (N), and RF is defined as (Tareke, Awoke 2022):

$$RF = \frac{n}{N} \cdot 100$$

4.1.3. Spatial distribution of droughts in the selected basins and climatic zones using Inverse Distance Weighting

The spatial distribution maps are rendered using the method of inverse distance weighting (IDW). This method is a powerful deterministic technique for the spatial interpolation of results, and it is an inherent advantage that is relatively fast in computation, which improves the ease of interpretation (Shepard 1968; Lu, Wong 2008). Commercially available software (ArcGIS) was used to obtain the spatial distribution, and the maps were generated using ArcGIS 10.7 for the studied area. The IDW technique assumes that unmeasured near points have a higher probability of weighting than far points (Gumus, Algin 2017).

Researchers assume that the IDW method provides a similar output map to that of the Gaussian process regression method (Kriging), which is useful for smaller areas or in the case of high station density (Gemmer et al. 2004). The power parameter (which controls the values of the interpolated sample over the expected location of a searched radius) is the main factor affecting the accuracy of the IDW method. Higher power values result from nearby samples, which affect estimation, and the resulting spatial interpolation surfaces become more detailed.

4.2. Correlation between hydrologic and rainfall drought indicators

Pearson's correlation coefficient (r) describes the strength and direction of a linear relationship between two quantitative variables, although the interpretation of the strength of this relationship varies between disciplines. The strength of r values has been interpreted as "greater than +0.5 (strongly positive) and less than –0.5 (strongly negative)." As an inferential statistic, r is used to test statistical hypotheses of significance for a linear relationship between two variables. However, the reliability of the linear model also depends on the length of data in the sample.

The significance of the correlation coefficient is used in this study to decide whether the linear relationship between the *SPI* and *SDI* is strong, moderate, or weak. The hypothesis t-test, comparing *SPI* and *SDI* correlation coefficients, addresses whether there is a linear relationship between them by using the t observed and the r-value.

5. Results

The study of *SPI* and *SDI* indices based on the hydrologic and rainfall data of the seven river basins and the five climatic zones, which were calculated on timescales of 6-months, clarifies how the hydrologic and rainfall droughts in the critical situation are achieved.

5.1. Hydrologic and rainfall drought statistical analysis

The spatial distributions covering climatic zones based on the *SPI* values demonstrate the highest severe and extreme drought events observed in the first 6-months for the years: 1902/03, 1921/22, 1927/28, 1965/66, 1986/87, 1988/89, 2009/10, and for the second 6-months for the years: 1932/33, 1959/60, 1972/73, 1978/79, 1998/99, 2007/08, 2013/14 (Fig. 4). The spatial distributions based on the *SDI* values (Fig. 5) indicate the highest severe and extreme drought events for the first and second 6-months are encountered in the years 1985/86, 1989/90, 1998/99, 2000/01, 2007/08, 2013/14 (at most of the stations for this year), and 2017/18, with an occurrence of 4 to 18%.

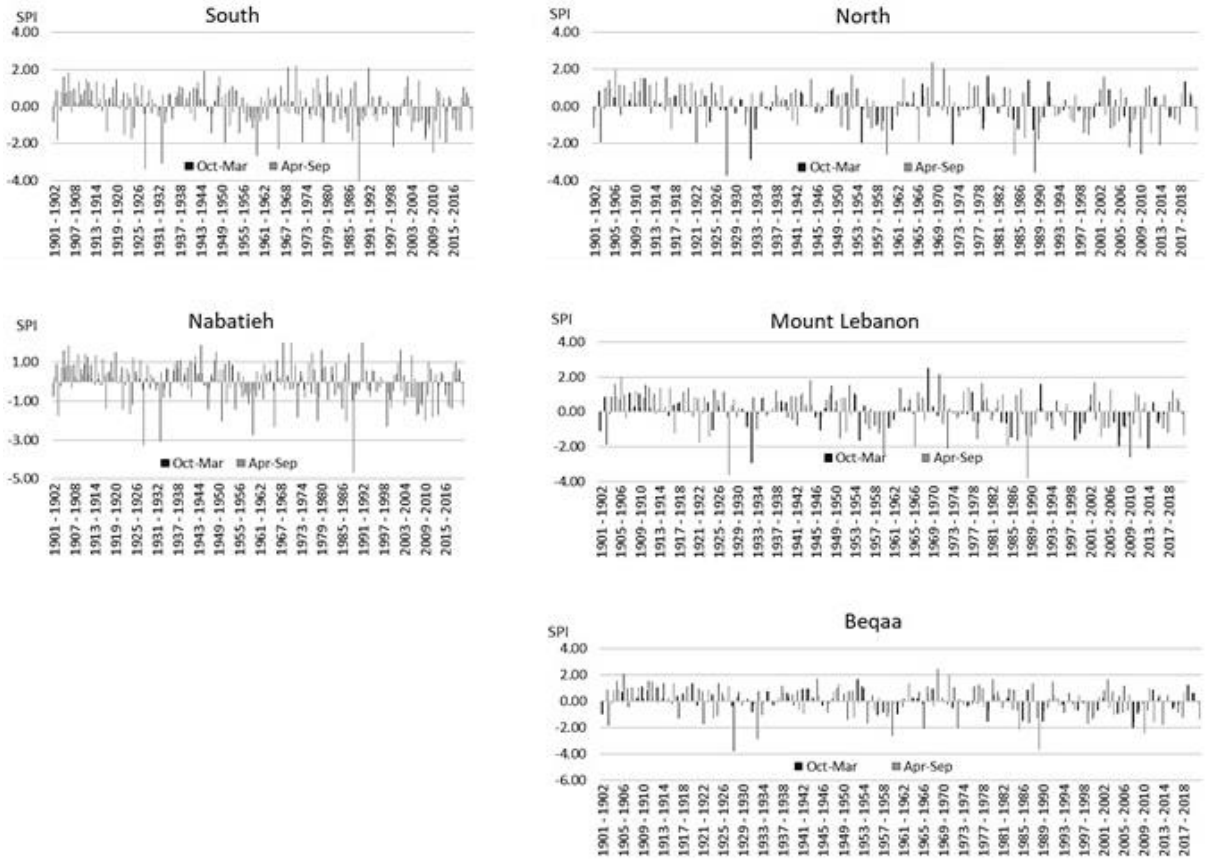


Fig. 4. Temporal standardized precipitation index (*SPI*) values of the five climatic zones in Lebanon according to two seasonal periods (Oct.-Mar. and Apr.-Sept.) (1901/1902-2019/2020).

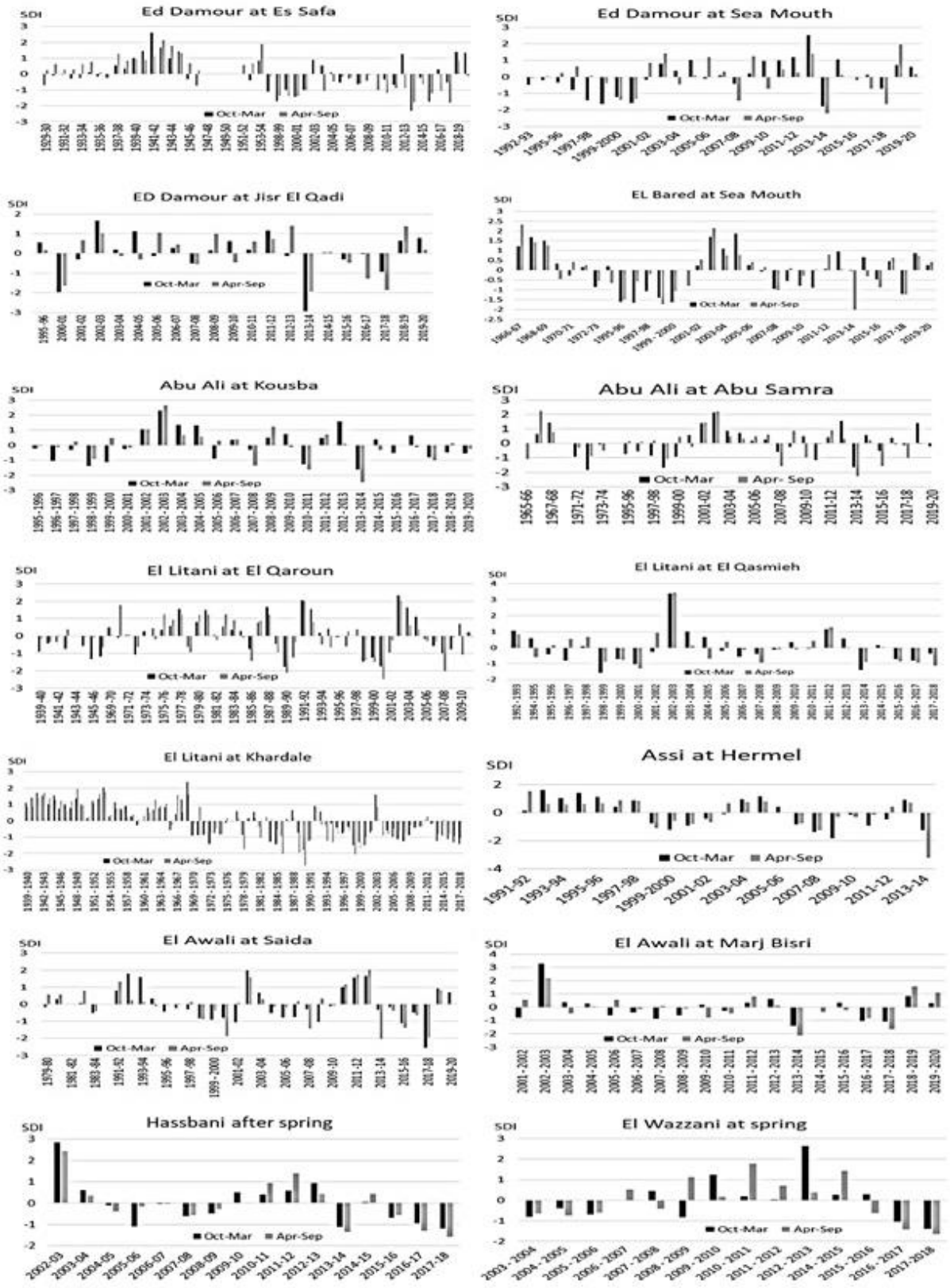


Fig. 5. Temporal drought index (*SDI*) values of selected stations in river Lebanon according to two seasonal periods (Oct.-Mar. and Apr.-Sept.).

5.1.1. Statistical Analysis of the Standardized Precipitation Index

The *SPI* distributions demonstrate that some climatic zones (Tables 4 and 5) exceed the threshold of classes of extreme drought (*SPI* < –2) and extreme wet (*SPI* > 2). The variation in *SPI* mean values shows that the second 6-months period of the years 1927/28, 1932/33, and 1988/89 are characterized as extreme drought events.

The median of the *SPI* (Table 6) of the first and second 6-months periods at climatic zones in Lebanon from 1901 to 2020 is about mildly wet. The recurrence of extreme drought in the south, in Beqaa, Nabatieh, Mt. Lebanon, and North, ranged from 3 to 5 times over 120 years.

Table 5. Severity category according to *SPI* of climatic zones supplying studied basins (1901-2020).

Climatic zones that affect the basins	Moderate drought (%)		Severe drought (%)		Extreme drought (%)		The sum of the three drought classes (%)	
	first 6-months	second 6-months	first 6-months	second 6-months	first 6-months	second 6-months	first 6-months	second 6-months
North (El Bared & Abu Ali)	10.8	6.7	2.50	3.3	4.17	3.3	17.5	13.3
Mt Lebanon (Ed Damour)	9.17	5.8	4.17	4.2	3.33	3.3	16.7	13.3
Mt Lebanon-south (El Awali)	8.34	5.4	4.17	4.6	2.92	3.3	15.4	13.3
Beqaa (El Azzi)	8.33	5.4	5.00	4.6	3.33	3.3	16.7	13.3
Beqaa-Nabatieh (El Hassbani)	7.50	5.4	3.75	4.1	3.75	3.7	15	13.2
Beqaa-Nabatieh-South (El Litani)	7.48	5.3	3.89	4.5	3.31	3.6	14.7	13.4

Table 6. The median of *SPI* at climatic zones in Lebanon (1901-2020).

Climatic zone	Median of *SPI*		Description	Value
	Oct.-Mar.	Apr.-Sep.		
South	0.12	0.15	Mildly wet	–0.99, 0.99
North	0.01	0.09		
Nabatieh	0.08	0.18		
Mt Lebanon	0.05	0.09		
Beqaa	0.02	0.09		

In the first 6-months period, the greatest drought is recorded in the north climatic zone, and in the second 6-months period, it is recorded in the northeastern Beqaa zone.

The percentage of severe and extreme droughts ranged between 2.5 and 5% in the five climatic zones. The frequency of normal conditions was 52.5% in the first 6-months periods and 55.8% in the second 6-months period over all five zones.

The median *SPI* for the first and second 6-months was calculated for each climate zone from 1901/02 to 2019/20.

The median of the five climatic zones (Table 6) for the first and second 6-months periods *SPI* is mildly wet. It ranged between 0.01 in the first 6 months (North climatic zone) and 0.18 in the second 6 months (Nabatiyeh climatic zone). The probability of the median is at least 50% will be less or greater than or equal to –0.99 and +0.99, which means mildly wet conditions should be expected to occur in one out of two years.

5.1.2. Statistical Analysis of the Streamflow Drought Index

Since *SDI* is a point or site-specific drought indicator, the discussion here considers the fourteen gauging stations of the basin. Hydrological drought is progressing gradually due to the scarcity of rainfall, which stops or decreases for some consecutive months from April to November. Therefore, this study focuses on two seasons based on the first and second *SDI*-6 of hydrological drought analysis that can summarize the annual drought conditions. The result indicates that the frequency of drought based on *SDI* is similar for both 6-months seasons and seven studied basins(Table 7), even though the drought is not very strong. The temporal and spatial variation of streamflow drought in the study area using *SDI* indicates that the lowest negative *SDI* value (–2) was registered for one year at nine measurement stations out of 14, as shown in Figure 5.

Table 7. Frequency of hydrologic droughts in 7 basins of Lebanon.

Basin	Hydrologic gauging station	Moderate hydrological drought (%)		Severe hydrological drought (%)		Extreme hydrological drought (%)		The sum of the three drought classes (%)	
		first 6-months	second 6-months	first 6-months	second 6-months	first 6-months	second 6-months	first 6-months	second 6-months
El Bared	Sea Mouth	9.1	9.1	9.1	6.1	0	3.0	18.2	15.2
Abu Ali	Abu Samra	3.2	6.5	9.7	6.5	0	3.2	12.9	16.2
	Kousba	16.0	12.0	4.0	4.0	0	4.0	20	20
Ed Damour	Es Safa	7.0	16.3	4.7	4.7	2.3	0	14	21
	Jisr El Qadi	0	4.8	4.8	14.3	4.8	0	9.6	19.1
	Sea mouth	7.4	11.1	11.1	3.7	0	3.7	18.5	18.5
El Awali	Marj Bisri	15.8	0	0	5.3	0	5.3	15.8	10.6
	Saida	9.1	6.1	0	6.1	3.0	3.0	12.1	15.2
El Assi	Hermel	13.0	8.7	4.3	0	0	4.3	17.3	13
El Hasbani	Wazzani	14.3	7.1	0	7.1	0	0	14.3	14.2
	Hassbani	18.8	12.5	0	6.3	0	0	18.8	18.8
El Litani	Qaroun	12.2	6.1	4.1	2.0	0	6.1	16.3	14.2
	Qasmieh	8.0	8.0	4.0	0	0	0	12	8.0
	Khardale	15.2	6.3	2.5	3.8	0	3.8	17.7	13.9

Moderate, severe, and extreme droughts, occurring several times over the years of record at various gauging stations, indicate that the El Litani River basin was highly affected by three extreme droughts, followed by the El Awali River basin at Saida with two extreme droughts. The analysis implies that El Damour, El Bared, and Abu Ali at Abu Samra and El Awali have had similar drought occurrence seasons at least in the past ten years. The drought analysis shows that El Hassbani and the lower and central parts of El Litani recorded a constant drought for several years.

For consecutive years, the majority of the river basins were affected by severe to extreme drought. Besides these drought events, some periods were dominated by moderate drought conditions at most stations, and some other stations were affected by extreme drought in both the first and second *SDI*.

The sum of drought events ranges from 8% for the second 6 months (El Litani at Qasmieh) to 20% for both the first and the second 6-months periods (Abu Ali at Kousba). Although 80% of the averages of all events were in the normal or near-normal range of wetness, 30% experienced a hydrological drought (Table 7).

According to the *SDI*, 10 out of 14 monitoring stations were affected by a near-drought once every two years in the first season, and five out of 14 monitoring stations were affected by near-drought in the first and second seasons. Other than that, the median of the *SDI* did not exceed near-normal or mildly wet (Table 8).

The median *SDI* of the first and second 6-months periods for each hydrological gauge station was calculated for different periods (Table 3).

The median of the first and second 6-months of *SDI* for 14 hydrological gauging stations (Table 8) was mildly wet, ranging from –0.24 in the first 6-months period (Kousba) to +0.45 in the second 6-months period (El Hermel). The probability of mildly wet conditions in *SDI* is 50%, which means they occur one out of two years.

Table 8. The median of the Streamflow Drought Index at the hydrological gauging stations studied (reference of date to Table 3).

Hydrological gauge station	Median of *SDI*		Description	Value
	Oct.-Mar.	Apr.-Sept.		
Es Safa	–0.19	0.02	Mildly wet	–0.99, 0.99
Abu Samra	–0.11	0.11		
El Hermel	–0.09	0.45		
Bared S.M.	0.05	0.06		
Damour S.M.	–0.04	0.11		
Hassbani	–0.05	–0.08		
Jisr El Qadi	0.15	0.17		
Khardale	–0.03	0.05		
Kousba	–0.24	0.02		
Marj Bisri	–0.03	–0.10		
Qaroun	0.02	–0.03		
Qasmieh	–0.21	–0.09		
Saida	–0.22	0.05		
Wazzani	0.01	–0.12		

5.2. Extreme hydrologic drought events

The *SDI* values of the hydrological drought events indicate that some stations encounter extreme drought conditions. Extreme drought events occurred in the second 6-months period at El Damour Sea Mouth (2013/14), Khardale (1985/86, 1989/90, and 1998/99), Kousba, and Marj Bisri (2013/14), Qaroun (1989/90, 2007/08), and Saida (2013/14). Accordingly, the effect of extreme drought in rainfall appears as a severe hydrological drought for the same timescale in the following year (1989/90). Commonly, the most extreme streamflow drought year for all gauging stations in the basins of the study area was obtained in hydrological years 2013/14 (Abu Samra, El Hermel, El Bared Sea Mouth, Ed Damour Sea Mouth, Es Safa, Kousba, Marj Bisri), 1985/86, 1989/90 and 1998/99 (Khardale), 1989/90, 2000/01 and 2007/08 (Qaroun) and in 2017/18 (Saida).

Extreme drought in the study basins recurred a maximum of 3 times over 49 years (Qaroun) and over 79 years (El Khardale), which means the severity of events is 1/16 years at Qaroun and 1/26 years in Khardale.

5.3. Spatial distribution of hydrologic and rainfall drought

The spatial distribution maps obtained using the IDW method are shown in Figures 6 and 7. The figures indicate the occurrence of drought events based on the values of *SPI* and *SDI* by considering classes from normal to extreme drought. Figure 6 shows the occurrence rates of moderate, severe, and extreme drought in the entire climatic zones, in the range for the first 6-months period were 7.5-11%, 2.5-5%, and 2.9-4.2%, respectively. The *SPI* for the second 6-months period ranged from 5-6.7%, 3.3-4.6%, and 3.3-4.2% for moderate, severe, and extreme drought, respectively. The greatest rainfall drought level, 3.8-4.2%, was observed in the northern climatic zone, which affects the basin of El Bared and Abu Ali, based on *SPI* for the first 6-months period, and 4-4.2 in the climatic zone of the Beqaa, which affects the basin of El Assi, based on *SPI* for the second 6-months period.

Considering the entire basins at all timescales, the lowest drought occurrence was 17.5% (North) for *SPI* in the first 6-months period and 20% (Abu Ali) for *SDI*, also in the first 6 months. Percentage drought occurrence with varying intensities is shown in Figures 6 and 7 for *SPI* and *SDI* for the first and second 6-months periods. Extreme drought events, per *SPI*, were observed in the zone of the basins of El Bared, Abu Ali, and El Assi. Per *SDI*, extreme drought was encountered in the northwest of El Damour basin, West of El Awali, north of El Assi, and middle of El Litani.

The spatial distributions of the most drought events occurring for various periods are shown in Figure 7 for *SDI*. The figure shows that the highest severe and extreme drought events (first 6-months period) are observed at Ed Damour stations Sea Mouth and Marj Bisri, as well as in El Bared at Sea Mouth, Abu Ali at Abu Samra, and in El Assi at El Hermel.

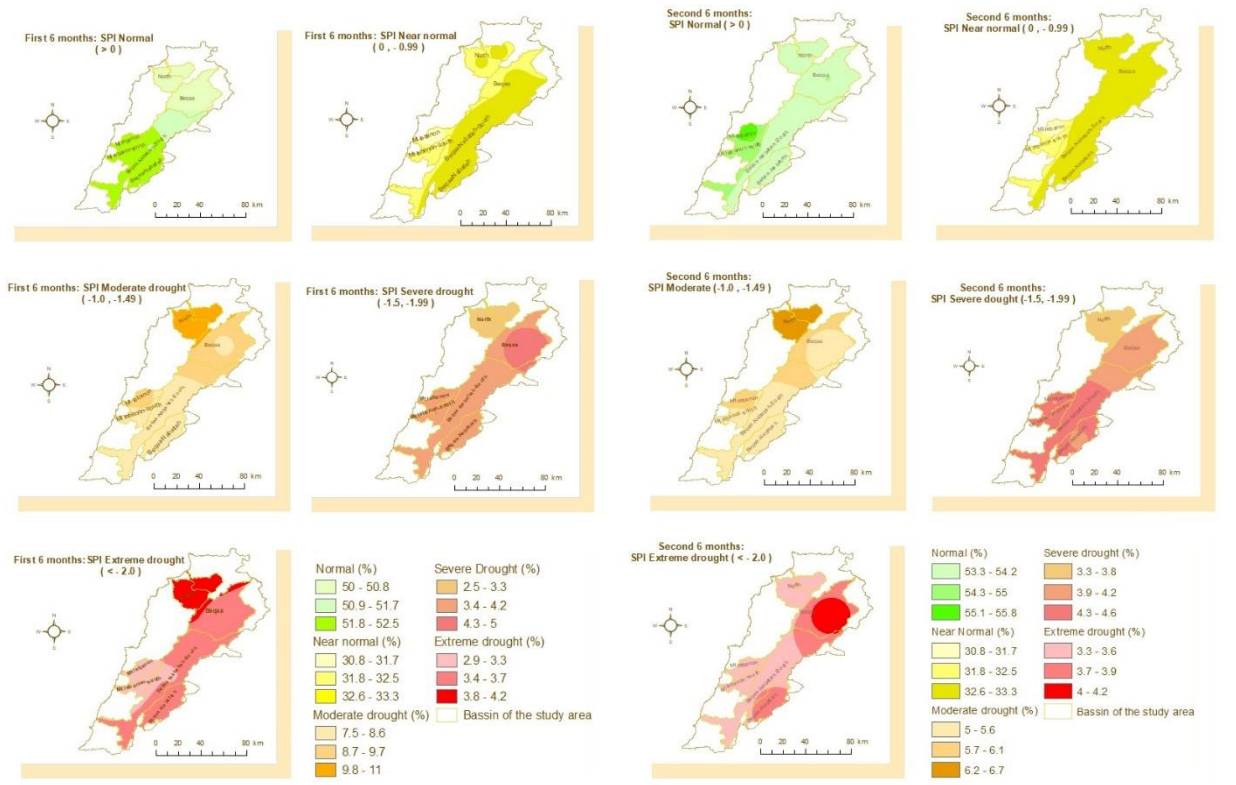


Fig. 6. *SPI* for 6-months periods in climatic zones which affect the basin in Lebanon (1901-2020).

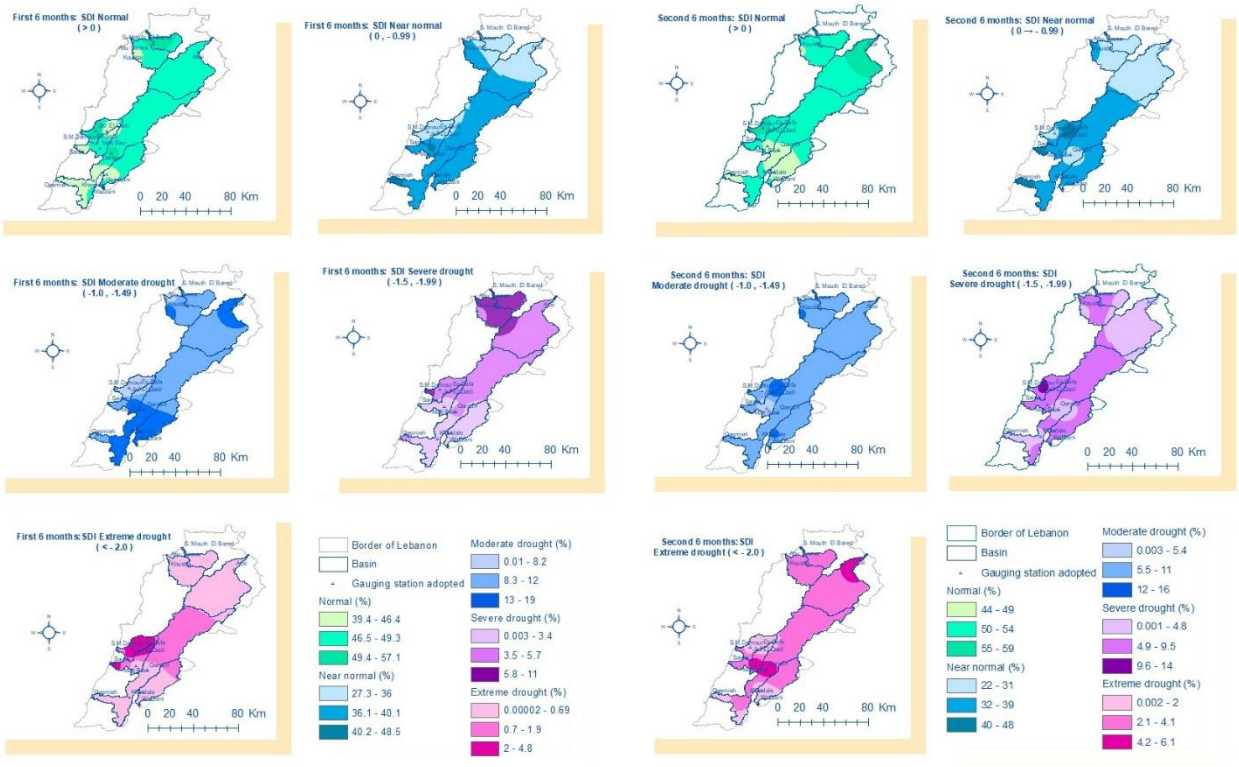


Fig. 7. *SDI* for 6-months periods at the selected streamflow sites in Lebanon (dates according to Table 3).

Based on *SDI*, drought intensities of moderate, severe, and extreme in most of the basins are respectively in the range of 0.01-19%, 0.003-11%, and from 0-4.8% (first 6-months period) and 0-16%, 0-14%, and 0-6.1% for the second 6-months period. Severe drought events were observed in the climatic zones of Beqaa, Mt Lebanon, South, and Nabatieh, as in the basins of El Bared, most of the Abu Ali, and the middle and mouth of Ed Damour basin.

Extreme droughts based on *SPI* ranged between 2.9-4.2% and 3.3-4.2% (first and second 6-months periods), and, based on *SDI,* were between 0-4.8 and 0.002-6.1 for the study basins. For combined 6-months periods, severe droughts were 2.5-5% per *SPI*, and for most of the study basins per *SDI*, they were in the range of 0-14%.

5.4. Correlation between *SDI* and *SPI* indices

The investigation has been analyzed by examining the strength of the relationships between *SDI* and *SPI* using bivariate correlation analysis (Table 9). The correlation coefficients (Pearson's *r*) based on the values of *SPI* and *SDI* for the various timescales[10] are provided in Figure 8.

Given that the critical thresholds for a strong correlation are –0.50 or +0.50, there is a strong positive relationship between *SPI* and *SDI* for the first 6-months period (Oct.-Mar.) for 12 stations, ranging from 0.86 (El Hassbani after spring) to 0.57 (Ed Damour at Sea Mouth). are weak positive relationships for two stations: 0.45 (El Assi at El Hermel) and 0.33 (El Hassbani at El-Wazzani). The correlation between *SPI* and *SDI* in the second 6-months period (Apr.-Sept.) is very weak, with values ranging from a minimum of –0.01 to –0.15 to a maximum of 0.02 to 0.40 for 12 stations. Correlations are modest (0.40) for the two remaining stations (El-Safa and El-Qaraoun).

The significance of *r* from a set of data points that appear to have a linear relationship is presented in Table 9. Correlations between *SDI* and *SPI* were not significant ($p > 0.05$), except for El Safa, which is statistically significant ($p = 3.79$).

[10]For comparison, the same years were adopted for the rain and water data (Table 3).

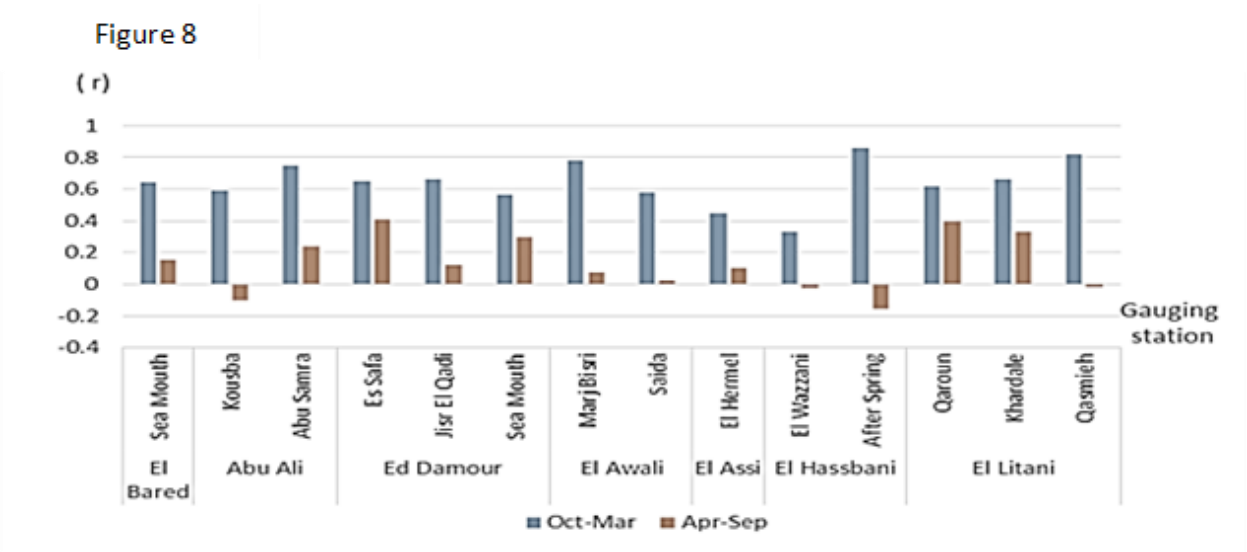


Fig. 8. Correlation *SDI-SPI* for the first and second 6-months periods in basins of Lebanon (reference of date to Table 3).

Table 9. *SDI* and *SPI* correlation in selected basins in various climate zones of Lebanon (reference of date to Table 3).

Climatic zones	Basin	Station	Number of years	First 6-months (Oct.-Mar.)		Second 6-months (Apr.-Sept.)		*t* critical
				r	*t* observed	*r*	*t* observed	
North	El Bared	Sea Mouth	33	0.64	0.71	0.16	0.14	2.042
	Abu Ali	Kousba	25	0.59	1.12	–0.10	0.93	2.069
		Abu Samra	31	0.75	1.08	0.24	0.68	2.045
Mt Lebanon	Ed Damour	Es Safa	27	0.65	1.08	0.41	3.79	2.08
		Jisr El Qadi	21	0.66	0.92	0.12	0.66	2.093
		Sea Mouth	27	0.57	1.30	0.30	0.72	2.06
	El Awali	Marj Bisri	19	0.78	0.76	0.08	0.61	2.11
		Saida	29	0.58	1.19	0.02	0.55	2.052
Beqaa	El Assi	El Hermel	23	0.45	1.16	0.10	0.87	2.08
Beqaa, Nabatieh	El Hassbani	El Wazzani	14	0.33	1.48	–0.02	0.75	2.179
		Aft. Spring	16	0.86	1.25	–0.15	0.83	2.145
Beqaa, Nabatieh,	El Litani	Qaroun	45	0.62	1.81	0.40	0.65	2.021
		Khardale	79	0.66	0.85	0.33	0.60	1.994
		Qasmieh	25	0.82	1.54	–0.01	0.74	2.069

6. Discussion and conclusion

To answer the research question that deals with the characteristics of hydrologic and rainfall drought and their spatial and temporal distribution in the basins of different climatic zones in Lebanon, the results show an unevenly, mildly wet to mild drought conditions, with a 14.5% decline in rainfall from 1901 to 2020. The results for rainfall drought are similar to those from previous studies, which range from 12 to 16% (Shaban 2015). The decrease is larger in the case of rivers, where the hydrologic drought ranges from 21.3% (El Bared at the Sea Mouth (1966-2020) to 58.5% (El Litani at Khardale, 1939-2018), which differs from Shaaban (2018), who reported a difference between 23 and 29%.

The temporal variation for the first and the second 6-months periods in median values of *SPI*, represents mildly wet events in climatic zones in Lebanon occur with a probability of one out of two years.

The study indicates that the greatest drought for the first season was recorded in the northern climatic zone, and in the northeastern Bekaa climatic zone for the second season. The climatic zones most vulnerable to rain drought are Bekaa and the north.

While severe drought is frequent in most zones, this result can be compared with Faour's (2015) studies, which specify moderate to extreme drought in the Bekaa Valley, and a moderate to severe drought in the Amioun area (located in the northern climatic zone).

The temporal variation in median *SDI* based on the first and second 6-months periods oscillates between mild drought (-) and wet (+) events in basins studied in Lebanon.

Measurement stations have generally been affected by moderate, and, in recent years, occasionally severe and extreme droughts. It must be said that each basin constitutes a case of moderate drought due to the difference in the measurement period and the lack of data for several years. Unfortunately, it is not possible to compare the results of the hydrological drought index for the study basins of Lebanon with previous ones, where similar studies are not available in the literature.

The spatial and temporal maps displaying the distribution of 14 water metering stations contributed to this study to understand the ratio and locations vulnerable to hydrologic and rainfall drought. The climatic zones with severe drought, for the first 6-months periods, are in the north, and for the second 6-months periods are in the northern Bekaa. The basins with severe drought, for the first 6-months periods, are in El Damour and El Awali, and for the second 6-months periods are in northern and central Bekaa.

The relationship between *SPI* and *SDI* fined in the first and second 6-months periods is weak, so it is not statistically significant. Since the values of *SPI* and *SDI* are not matched, one might assume the correlation between *SPI* and *SDI* values is weak. Calculating the correlation coefficient for these variables based on hydrologic and rainfall data reveals an inconsistent correlation over different periods.

These insignificant relationships reinforce the result of this study that the decline in discharge at the studied measurement stations is not only associated with the decrease in precipitation but may be related to other factors, including human intervention.

In all cases, the status is quite alarming and demands immediate water management plans to conserve water resources in the study basins, which are heavily populated and cover large investment areas. It is important to conduct additional investigations into hydrological drought and to develop drought early warning systems. Therefore, this study gives important information to decision-makers, about hydrologic and rainfall drought

conditions for five climatic zones and seven river basins representing most flow directions to mitigate these effects.

Acknowledgment

All gratitude to the Lebanese University for its funding and support of this research. I also appreciate the helpfulness of the hydrological service managers of the Department of the Litany National Authority in Lebanon for providing the hydrological data needed to carry out the current research.

References

Beran M.A., Rodier J.A., 1985, Hydrological aspects of drought: a contribution to the International Hydrological Programme, Studies and reports in hydrology, UNESCO-WMO, Paris, 149 pp.

Dracup J.A., Lee K.S., Paulson Jr. E.G., 1980 On the definition of droughts, Water Resources Research, 16 (2), 297-302, DOI: 10.1029/WR016i002p00297.

Edossa D.C., Babel M.S., Gupta A.D., 2009, Drought analysis in the Awash River Basin, Ethiopia, Water Resources Management, 24 (7), 1441-1460, DOI: 10.1007/s11269-009-9508-0.

FAO, 1974, Hydro-Agricultural Development Project – Farms in Southern Lebanon – Basic Statistics, National Litani Office – UNDP – FAO, 132 pp.

FAO, 2018, Drought characteristics and management in North Africa and the Near East. Rome, FAO Water Reports No 45, 265 pp.

Faour G., Mario M., Sandra A.N., 2015, Regional LANDSAT-based drought monitoring from 1982 to 2014, Climate, 3 (3), 563-577, DOI: 10.3390/cli3030563.

Gavrilov N.B., Markovic S., 2015, Comment on "Analysis of changes in meteorological variables using Mann-Kendall and Sen's slope estimator statistical tests in Serbia" by Gocic and Trajkovic (2013), DOI: 10.13140/RG.2.1.4264.4322.

Gemmer M., Becker S., Jiang T., 2004, Observed monthly precipitation trends in China 1951-2002, Theoretical and Applied Climatology, 77, 39-45, DOI: 10.1007/s00704-003-0018-3.

González E.A., Gutmann E., Aalstad K., Fayad A., Bouchet M., Gascoin S., 2021, Snowpack dynamics in the Lebanese mountains from quasi-dynamically downscaled ERA5 reanalysis updated by assimilating remotely sensed fractional snow-covered area, Hydrology and Earth System Sciences, 25 (8), 4455-4471, DOI: 10.5194/hess-25-4455.

Gumus V., Algin H.M., 2017, Meteorological and hydrological drought analysis of the Seyhan–Ceyhan River Basins, Turkey, Meteorological Applications, 24 (1), 62-73, DOI: 10.1002/met.1605.

Haddad E.A., Farajalla N., Camargo M., Lopes R.L., Vieira F.V., 2014, Climate change in Lebanon: higher-order regional impacts from agriculture, Region, 1 (1), 9-24.

Hong X., Guo S., Zhou Y., Xiong, 2014, Uncertainties in assessing hydrological drought using streamflow drought index for the upper Yangtze River basin, Stochastic Environmental Research and Risk Assessment, 29, 1235-1247, DOI: 10.1007/s00477-014-0949-5.

Lu G.Y., Wong D.W., 2008, An adaptive inverse-distance weighting spatial interpolation technique, Computers & Geosciences, 34 (9), 1044-1055, DOI: 10.1016/j.cageo.2007.07.010.

McKee T., Doesken N., Kleist J., 1993, The relationship of drought frequency and duration of time scales, 8th Conference on Applied Climatology, 17-22 January 1993, Anaheim, California.

Nalbantis I., 2008, Evaluation of a Hydrological Drought Index, European Water, 23/24, 67-77.

Shaban A., 2008, Impact of Climate Change on Water Resources of Lebanon: Indications of Hydrological Droughts, [in:] Climatic Changes and Water Resources in the Middle East and North Africa. Environmental Science and Engineering, F. Zereini, H. Hötzl (eds.), Springer, Berlin, Heidelberg, DOI: 10.1007/978-3-540-85047-2_11.

Shaban A., 2008, Indicators and aspects of hydrological drought in Lebanon, Water Resources Management, 23, 1875-1891, DOI 10.1007/s11269-008-9358-1.

Shaban A., Houhou R., 2015, Drought or humidity oscillations? The case of coastal zone of Lebanon, Journal of Hydrology, 529, 1768-1775, DOI: 10.1016/j.jhydrol.2015.08.010.

Shepard D., 1968, A two-dimensional interpolation function for irregularly spaced data, Proceedings of the 23rd ACM National Conference, DOI: 10.1145/800186.810616.

Tallaksen L.M., van Lanen H.A.J., 2004, Hydrological drought. Processes and estimation methods for streamflow and groundwater. Developments in Water Science No 48, Elsevier, 579 pp.

Tannehill I.R., 1947, Drought, Its Causes and Effects, Princeton University Press.

Tareke K.A., Awoke G.A., 2022, Hydrological drought analysis using Streamflow Drought Index (SDI) in Ethiopia, Advances in Meteorology. Special Issue: Computational Algorithms for Climatological and Hydrological Applications, DOI: 10.1155/2022/7067951.

Turney S., 2022, Pearson Correlation Coefficient (r) | Guide & Examples.

Wilhite D.A., Glantz M.H., 1985, Understanding the drought phenomenon: the role of definitions, Water International, 10 (3), 111-120, DOI: 10.1080/02508068508686328.

Wilhite D.A., Sivakumar M.V.K., Pulwarty R., 2014, Managing drought risk in a changing climate: the role of national drought policy, Weather and Climate Extremes, 3, 4-13, DOI: 10.1016/j.wace.2014.01.002.

Assessing the efficiency of a random forest regression model for estimating water quality indicators

Maryam Zavareh, Viviana Maggioni

Department of Civil, Environmental, and Infrastructure Engineering, George Mason University

Xinxuan Zhang

Department of Civil, Environmental, and Infrastructure Engineering, George Mason University

Eversource Energy Center, University of Connecticut, Storrs, CT

Abstract

This work evaluates the efficiency of Random Forest (RF) regression for predicting water quality indicators and investigates factors affecting water quality in 11 watersheds in Virginia, District of Columbia, and Maryland. Ten years of daily water quality data along with hydro-meteorological information (such as precipitation) and watershed physiology and characteristics (e.g., size, soil type, land use) are used to predict dissolved oxygen (DO), specific conductivity (K), and turbidity (Tu) across the selected watersheds. The RF regression model is developed for six scenarios, with an increasing number of predictors introduced in each scenario. The first scenario contains the smallest amount of information (water quality indicators DO, K and Tu), while scenario 6 contains all the available variables. The RF model is evaluated based on three statistical metrics: the relative root mean square error, the correlation coefficient, and the percentage of variance explained. In addition, the degree of importance for each predictor is used to rank their importance within each scenario. The model shows excellent performance for DO as the predicted variable. The model predicting K slightly outperforms the one predicting Tu. Scenario 4 (built based on water quality indicators, hydro-meteorological data, watershed physiology and land cover information) provided the best tradeoff between performance and efficiency (quantified in terms of the amount of information needed to develop the model). In conclusion, based on the RF model, land cover plays a significant role in predicting water quality indicators. In addition, the developed RF regression model is adaptable to watersheds in this region over a range of climates.

Keywords

Random Forest, water quality, hydro-meteorological information.

Submitted 11 December 2023, revised 30 January 2024, accepted 6 February 2024

DOI:

1. Introduction

Monitoring surface water quality provides important information that can be used for actions to sustain ecological systems, as well as to protect human health and livelihoods. Assessing temporal and spatial changes in water quality is fundamental for controlling and preventing water pollution. Several approaches have been investigated over the years to analyze such changes. Traditional methods based on statistical and numerical models are structurally complex, costly, time consuming, and require substantial data and detailed information (Jadhav et al. 2015). In addition, traditional models are not capable of reflecting the sophisticated interaction between chemical, physical, and biological properties of water quality (Chen et al. 2018). Furthermore, traditional models often require data pre-processing and assumptions regarding statistical distribution of data, which is usually unknown (Najah et al. 2019).

Recent developments in computer science, especially in Artificial Intelligence (AI), overcome most limitations of traditional modeling and has shown potential for handling water quality data (Tiyasha, Yaseen 2020). Machine learning (ML) is a branch of AI that enables computers to learn without explicit programming (Mitchell 2013). ML has been widely used in many fields, including medicine (Long et al. 1993), engineering (Hulten 2018), finance (Mezrich 1994), ecology (Kijewski et al. 2019), as well as environmental and water resources engineering (Chen et al. 2018; Norouzi, Moghaddam 2020). One of the powerful features of ML is its capability to identify non-linear and complex relationships between input and output data (Najah et al. 2019). Several ML models have been applied to water quality studies over the past two decades, including neural networks (Yu et al. 2020), artificial neural networks (Jeong et al. 2001; Amiri, Nakane 2009; Imani et al. 2021), adaptive neuro-fuzzy inference systems (Najah et al. 2019), support vector regression models (Wang et al. 2017), and rough set theory (Zavareh, Maggioni 2018). Some ML algorithms, including factor analysis (Akoto, Abankwa 2014), principal component analysis (PCA) and granger causality (Zavareh et al. 2021), have also been explored for data dimension reduction and to identify causal relationships. However, none of these techniques is perfect. For example, artificial neural networks require large amounts of data for training and often overfit data (Tiyasha, Yassen 2020). On the other hand, approaches like rough set and fuzzy set theories cannot handle and/or process quantitative data (Dubois, Prade 1992). Data dimension reduction techniques, like PCA, can make it difficult to interpret principal components (Karamizadeh et al. 2013).

Within ML forecasting models, RF is appealing because (Díaz-Uriarte, Alvarez de Andrés 2006; Boulesteix et al. 2012): (a) RF handles quantitative as well as qualitative data; (b) it does not overfit data; (c) its predictive performance is high compared to other modeling approaches; (d) it can directly process high dimensional data without dimensional reduction; (e) it does not need pre-processing; and (f) it can capture non-linear dependencies between predictor and predicted variables.

RF has been employed in water resources science and engineering in recent years (Parkhurst et al. 2005; Chen et al. 2017; Tyralis et al. 2019; Li et al. 2020). For instance, RF models have proven successful in generating groundwater potential maps (Golkarian et al. 2018; Sameen et al. 2019), stream flow forecasting (Papacharalampous, Tyralis 2018), predicting groundwater level (Wang et al. 2018), analyzing effects of urbanization on hydrological variables (Saadi et al. 2019), urban water consumption forecasting (Chen et al. 2017), as well as for predicting water inrush rate in coal mines (Zhao et al. 2018) and soil infiltration rate (Singh et al. 2017). RF is particularly suitable when non-linear relationships exist, which is the case for the majority of processes in water science (Kijewski et al. 2019; Tyralis et al. 2019).

RF has also become popular for predicting water quality indicators (Papacharalampous, Tyralis 2018). For instance, Devi (2019) investigated the application of an RF classification model to water quality prediction in Kadapa district, India. The study examined water quality indicators, including pH, total dissolved solids, elec-

trical conductivity, and chloride concentration to build a Water Quality Index (WQI) for drinking water assessment, revealing that total dissolved solids was the most important variable affecting WQI, whereas pH was least important. The model classified drinking water in the region with 94% accuracy and a 6.3% error rate. Another study investigated the application of an RF classification model on water quality (Tesoriero et al. 2017) to predict redox-sensitive contaminant concentration (nitrate, iron, and arsenic) in groundwater in northeastern Wisconsin. Their RF classification showed a high potential for assessing aquifer and stream vulnerability at regional and national scales. Furthermore, Wang et al. (2021) developed an RF regression model to predict water quality distribution in China's Taihu Lake basin. Their model used watershed features and climate variables as predictor variables of three water quality parameters, permanganate index (CODMn), total phosphorus (TP), and total nitrogen (TN). The RF models showed that TN concentration was affected by agricultural non-point sources, while the CODMn and TP were impacted by agricultural and domestic sources.

The present work builds upon these past studies and develops an RF regression model to assess water quality indicators in selected watersheds within Chesapeake Bay basin in the Eastern United States. Different scenarios are proposed to evaluate the effect of different groups of predictors on model performance and to rank their importance in estimating several major water quality indicators: dissolved oxygen concentration, specific conductivity, and turbidity. Finally, an independent watershed is used to assess the transferability of the proposed RF model to other watersheds having similar climate, size, and topography.

2. Study area and dataset

Eleven watersheds across the District of Columbia, Maryland, and Virginia (known as the DMV region) were selected for this study. The DMV region is particularly vulnerable to hydro-meteorological hazards, which are exacerbated by sea level rise because of its vicinity to the coast (Solakian et al. 2020). In addition, excessive algal growth, poor water clarity, and low dissolved oxygen related to eutrophication have been issues in the Chesapeake Bay area for the past few years (Zhang et al. 2018). Thus, researchers, local organizations, and governmental agencies have increased their efforts to collect and interpret water quality data to promote the health of the DMV watersheds that feed into the bay (Zhang et al. 2018).

Data for this work are extracted from 11 United States Geological Survey (USGS) stations located at the outlet of each watershed, as shown in Figure 1. These data contain water quality indicators, including dissolved oxygen (DO) in milligram per liter (mg l^{-1}), specific conductivity (K) in microsiemens per centimeter at 25 degrees Celsius (μS cm^{-1} at 25°C), turbidity (Tu) in Formazin Nephelometric Units (FNU), and water temperature (WT) in degrees Celsius (°C). Additional information is also considered here, including precipitation, discharge, air temperature, watershed size, and length of rivers running across watersheds, along with watershed land cover, soil type, and livestock count. These data are mainly extracted from USGS, National Aeronautics

and Space Administration (NASA), North America Land Data Assimilation System (NLDAS), and National Land Cover Database (NLCD). For more information regarding the data and the watersheds, we refer the reader to Zavareh et al. (2021).

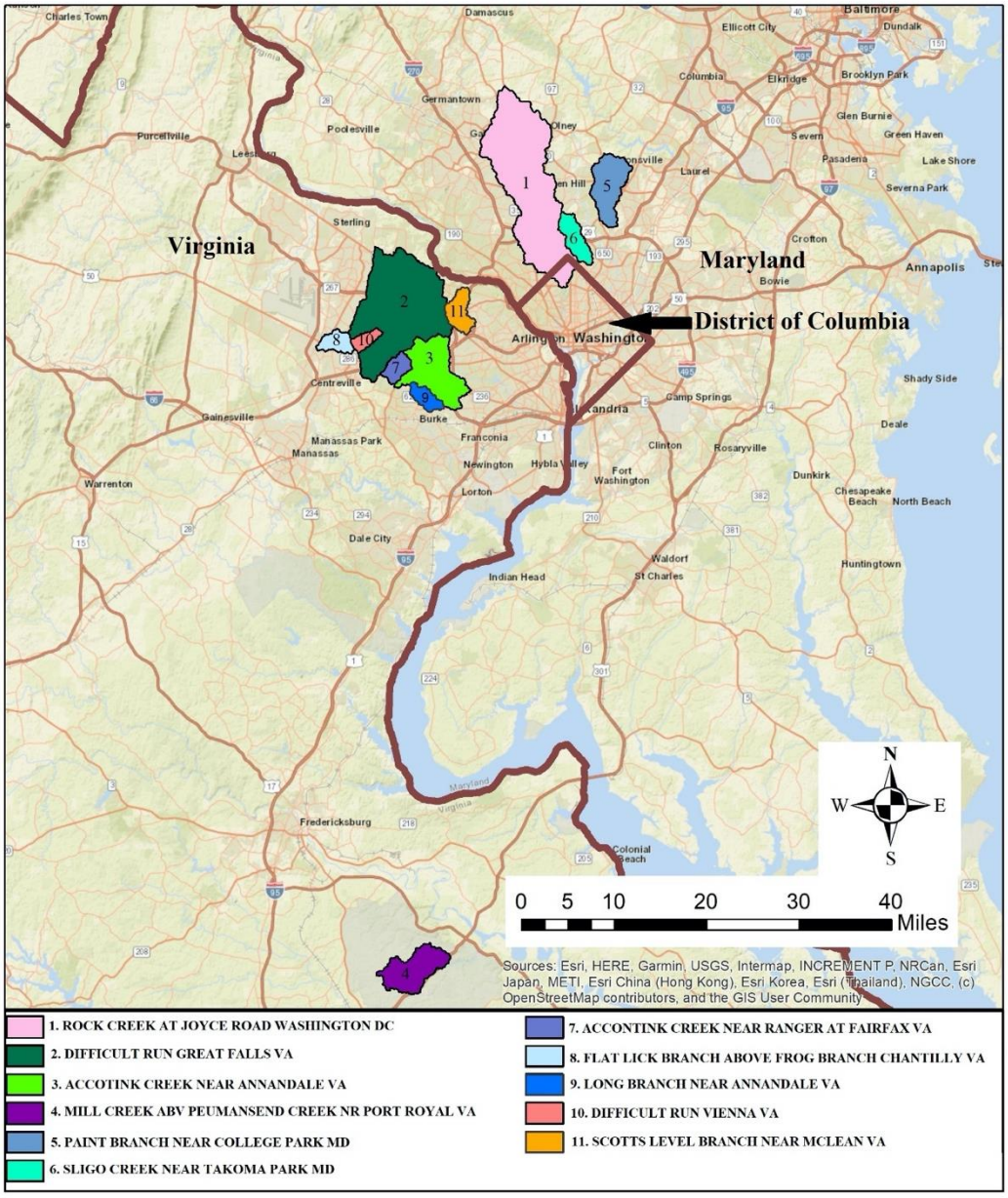


Fig. 1. Location of the 11 watersheds selected for this study across the DMV region.

Table 1 displays watershed characteristics, including watershed physiology (size of watershed and total length of rivers in a watershed), land cover, soil type, and livestock head count for all watersheds in this study. Watersheds 1-10 are used for developing the RF model, whereas Scotts Level Branch (watershed 11) is used as an independent watershed for assessing model performance in the validation phase of this study.

Watershed size varies between 7 and 169 km², while the total length of rivers ranges between 5 and 132 km. Land cover is summarized into five main groups, including wetland, developed, barren, forest, shrubland, and reported as percentages. Most watersheds are highly urbanized, with more than 50% of the total area being developed, except for watersheds 4 and 10. Watershed 4 is least developed, with only 8% of its total area classified as developed; watershed 6 is the most developed, with 87% of the total area classified as developed. Four watersheds (1, 4, 5, and 6) are mainly characterized by soil type B with moderate infiltration, whereas there is a prevalence of soil type C with slow infiltration in all other watersheds. Soil type is A least common in all watersheds. Land use and soil type affect infiltration rates, stream flow, and stormwater runoff (carrying contaminants), and can be particularly useful for interpreting relationships among water quality indicators and environmental characteristics (Zavareh et al. 2021).

The minimum and maximum livestock head counts were 2 and 885, respectively. As shown in Table 1, even highly urbanized watersheds contain livestock (e.g., watershed 6 is the most urbanized watershed and has a headcount of 89 livestock). The livestock head count is included because the manure and waste from concentrated animal feeding operations have been a long-standing concern in contamination of water runoff as a potential non-point source of water quality degradation (Burkholder et al. 2007; Dufour et al. 2012).

Table 1. Characteristics of watersheds in this study. Watershed area and total length of rivers are in km and km², respectively, whereas land use and soil type are in percent.

Watershed features	1	2	3	4	5	6	7	8	9	10	11
Area	169	149	62.0	37.0	34.0	17.0	10.0	10.0	10.0	7.00	9.00
Total length of rivers	103	132	56.0	30.2	23.9	9.30	9.20	10.0	8.70	5.00	8.20
Wetland, open water	2.10	4.70	2.70	6.90	2.70	0.10	0.10	1.00	2.30	3.80	0.00
Developed	69.2	53.5	74.2	7.90	61.1	87.8	85.4	86.0	70.6	44.0	82.3
Barren	0.20	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.20	00.0
Forest	20.9	38.9	22.8	77.6	29.1	11.7	14.3	11.6	27.1	51.0	13.8
Shrubland, Herbaceous, Planted	7.50	2.80	0.40	7.40	7.00	0.30	0.20	1.30	0.10	0.80	3.90
Soil Type A	0.70	2.90	1.20	0.00	1.00	0.00	0.00	0.00	0.70	4.00	0.00
Soil Type B	73.6	29.9	18.1	99.8	76.2	81.2	6.00	4.30	20.5	29.4	51.0
Soil Type C	16.0	66.7	80.7	0.20	14.5	11.1	93.6	89.7	78.9	66.5	36.7
Soil Type D	9.8	0.50	0.10	0.00	8.30	7.70	0.30	6.00	0.00	0.00	12.2
Livestock count	885	152	46.0	65.0	185	89.0	2.00	8.00	5.00	5.00	75.0

3. Methodology

3.1. The Random Forest Model

RF is an ensemble method, first developed by Breiman (2001), that uses multiple decision tree algorithms to produce repeated predictions of the same phenomenon. The ensemble combines predictions from multiple

learning models to obtain better accuracy than the individual models (Rokach 2010). One of the advantages of the RF method is that there is no need to pre-process or normalize data.

RF can be used for classification purposes and as a regression method depending on the nature of the dependent predicted variable (Tyralis et al. 2019). In regression models, the dependent variable is continuous (quantitative), whereas in classification algorithms it is categorical. RF models for regression are formed by growing trees depending on numerical values as opposed to class labels (Breiman 2001). In the present case, since the nature of predicted variables is continuous, we use an RF regression model. In this approach, RF grows a forest from many regression trees. A Regression Tree (RT) is a set of restrictions or conditions which are hierarchically structured, and which are successively applied from a root to a terminal node or leaf of the tree (Breiman et al. 1993; Zabihi et al. 2016).

The first step in developing an RF model is bootstrapping, in which data is randomly sampled from the entire dataset with replacement (i.e., data can be picked more than once). Each RT is grown in a bootstrapped subsample of a training dataset, which is known as bagging (Lagomarsino et al. 2017). The remaining data are called Out Of Bag (OOB), and they are used to estimate the prediction error and the importance of the variables (Han et al. 2016). Predictions based on the OOB set prevent overfitting (Lagomarsino et al. 2017). Overfitting may also result from extremely large trees, where lower branches introduce modeling noise. To avoid overfitting, the RT needs to be pruned. Pruning trees generates a simpler tree by deleting redundant variables. The second step is feature (variable) selection. In order to determine a split at each node in a decision tree, variables are randomly selected as features (Breiman 2001). Feature selection helps to build uncorrelated trees. Additional, feature selection introduces an extra layer of randomness to the model. The third step is to repeat steps 1 and 2 to build a forest with many trees, with each tree trained with different data. Consequently, two important parameters need to be selected in every RF model: the number of trees and the number of splits at each node.

In this work, the RF model is developed based on data from 10 watersheds across the study area to estimate three water quality indicators: DO, K, and Tu. When one indicator is assigned to be the predicted variable, the other two are used as predictor variables. From the original data, 70% is dedicated to train the model, and the remaining 30% is used for testing (verification). The model is then validated using an independent watershed, i.e., Scotts Level Branch.

The RF model built based on all information (water quality indicators in addition to information listed in Table 1) is trained with different numbers of trees: 50, 100, 200, 300, 400, 500, and 600 (Fig. 2). The optimal number of trees is chosen based on the value minimizing the relative Root Mean Square Error (*rRMSE*), which is a measure of the relative misfit between modeled variables (DO, K, and Tu) and the corresponding observed values. This study uses 500 trees, the value at about which *rRMSE* reaches a plateau. This estimate is

consistent with the default values used in prior studies (Boulesteix et al. 2012; Devi 2019; Saadi et al. 2019; Al-Abadi et al. 2021).

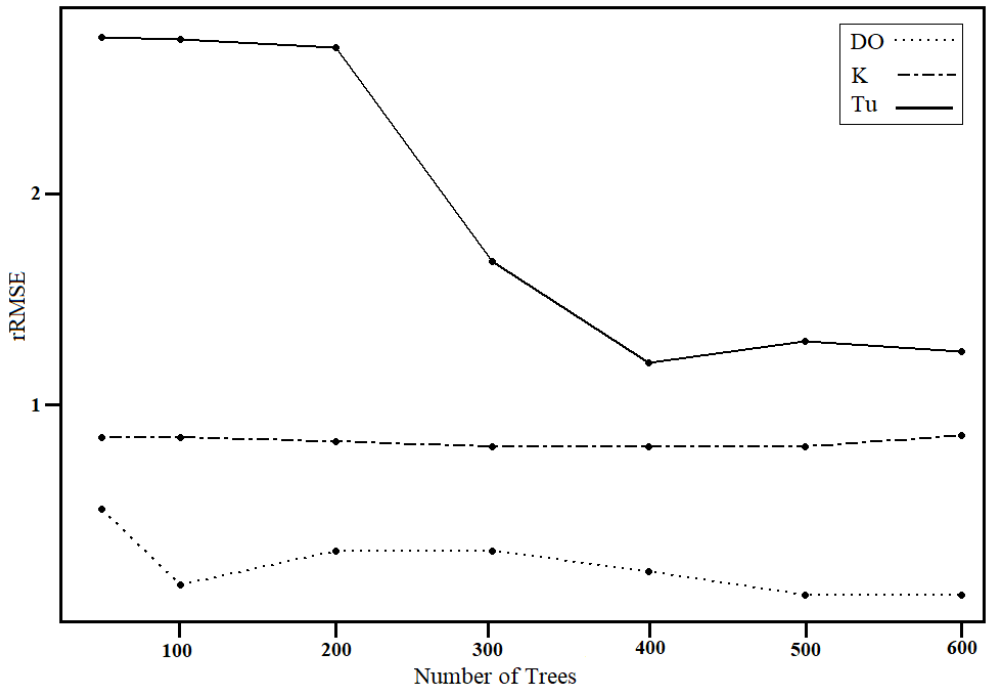


Fig. 2. Values of *rRMSE* for modeled DO, K, and Tu with respect to their corresponding observed values as a function of the number of trees used in the RF regression model.

Another parameter to calibrate when building an RF model is the number of variables at each split (*mtry*). For a regression RF model, *mtry* is suggested to be approximately one third of the number of variables in the dataset (Díaz-Uriarte, Alvarez de Andrés 2006; Boulesteix et al. 2012; Fox et al. 2020). This value was chosen for the present study. Here, *mtry* values are selected based on the number of variables in each of the six scenarios described in the Section 3.2.

3.2. Model scenarios and performance evaluation

The RF is developed for six scenarios, shown in Table 2. The number of variables increases moving from scenario 1 to scenario 6. The first scenario contains only four water quality indicators, i.e., DO, K, Tu, and WT. In the second scenario, hydrologic characteristics of the watersheds, namely precipitation, discharge, and temperature, are added to the variables considered in scenario 1. In the third scenario, watershed physiology (watershed area and the total length of rivers in each watershed) is included. Land cover information is included in scenario 4, soil type is added to scenario 5, and livestock head count in each watershed is incorporated in scenario 6. As mentioned previously, the number of *mtry* for each scenario is one third of the number of variables in each scenario: *mtry* is 2 for scenarios 1 and 2, 3 for scenarios 3, 4 for scenario 4, and 6 for scenarios 5 and 6.

Table 2. Scenarios and number of predictor variables.

	Scenario					
	1	2	3	4	5	6
Water quality (DO, K, Tu, WT)	X	X	X	X	X	X
Hydrology (precipitation, discharge, temperature)		X	X	X	X	X
Watershed physiology (watershed area and length of rivers)			X	X	X	X
Land cover information				X	X	X
Soil type information					X	X
Livestock headcount						X
Total number of variables	3	6	8	13	17	18

Three statistical metrics are used to analyze model performance of each scenario: correlation coefficient (R), relative Root Mean Square Error ($rRMSE$), and percentage variance explained ($\%\,Var$).

The correlation coefficient between observed and predicted values is:

$$R = \sum_{i=1}^{n}\left(V_i - \overline{V}\right)\left(P_i - \overline{P}\right) \Big/ \sqrt{\sum_{i=1}^{n}\left(V_i - \overline{V}\right)^2 \sum_{i=1}^{n}\left(P_i - \overline{P}\right)^2} \quad (1)$$

where: V_i are the measured values of variables, P_i are the predicted variable values, n is the number of variables in testing data, and $\overline{V}$ and $\overline{P}$ are the means of measured data variables and model predicted data, respectively (Wu et al. 2020).

The $RMSE$ indicates the overall misfit between the modeled and observed variables (Yu et al. 2020). This is a common metric to evaluate the performance of prediction results. A perfect prediction model would have zero $RMSE$. Since the errors are squared before they are averaged, it is very sensitive to large errors in the measured data (Wang et al. 2018). As a result, this study uses $rRMSE$ to assess model misfit. Its calculation formula is:

$$rRMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(P_i - V_i)^2} \Big/ \overline{V} \quad (2)$$

where: V_i are variables from measured testing data, P_i are predicted values of a variable, n is the number of variables in testing data, and $\overline{V}$ and $\overline{P}$ are the mean of variables in testing data and model predicted data, respectively (Wu et al. 2020).

The $\%\,Var$ is a measure to show how well out-of-bag predictions explain the predicted variance of the training set. The percent variation is the explained variation divided by total variation. In other words:

$$\%\,Var = \sum_{i=1}^{n}(o_i - \bar{o})\,(b_i - \bar{b}) \Big/ \sum_{i=1}^{n}(o_i - \bar{o}) + (b_i - \bar{b}) \quad (3)$$

where: o_i is a variable from OOB data, b_i is a variable from bootstrap data, and $\bar{o}$ and $\bar{b}$ are the mean of OOB and bootstrap data.

The importance measure is used to estimate how much the prediction error increases when OOB data for that variable are permuted, while all others are unchanged (Liaw, Wiener 2002). The importance measures are computed to rank all predictors: if the importance measure of a variable is lower relative to others, that variable contributes minimally to the prediction process and can be potentially excluded. The importance measure is computed as the Mean Decrease in Accuracy (MDA):

$$MDA = \frac{1}{ntree}\sum_{t=1}^{ntree}(EP_{tj} - E_{tj}) \quad (4)$$

where: $ntree$ is the number of trees, EP_{tj} is the OOB error on tree t after permuting the values of X_j, and E_{tj} is the OOB error on the tree t before permuting the value of X_j (Han et al. 2016). Permutation-based importance is crucial since it avoids allocating high importance to features that may not be predictive for unseen data due to overfitting (Pedregosa et al. 2011).

4. Results

4.1. RF Model Evaluation

The three-performance metrics (R, $\%\,Var$, and $rRMSE$) are calculated for each scenario when either DO, K, or Tu, is the predicted variable (Fig. 3). The best performance in terms of all three statistics is observed when estimating DO, based on the other water quality indicators. Minimal changes are observed when more predictors are included in the RF model, with slight improvement in $\%\,Var$ and $rRMSE$ when moving from scenario 1 to scenario 2, which added information about watershed hydrology. The effect of urbanization was also significant when DO was granger caused by K and Tu, as shown by Zavareh et al. (2021).

When predicating K and Tu, R values improve when moving to more complex scenarios. This is particularly evident when estimating Tu after hydrological information is added in scenario 2. This can be associated with K and Tu being strongly affected by precipitation and discharge.

In terms of $\%\,Var$, increases of 20% and 45% are shown for K and Tu, respectively, when information on watershed hydrology is included. In addition, increases of 12% and 10% for K and Tu are detected when watershed physiology is added to scenario 2. This suggests that hydrological information and watershed physiology highly improve prediction of data variance. However, adding watershed characteristics of land cover or soil type does not improve $\%\,Var$.

A slight improvement in the *rRMSE* of DO is observed when hydrologic information is added to the model. When K (Tu) is the predicted variable, *rRMSE* decreases by more than 50% (140%) when watershed physiology and land cover are added to the model.

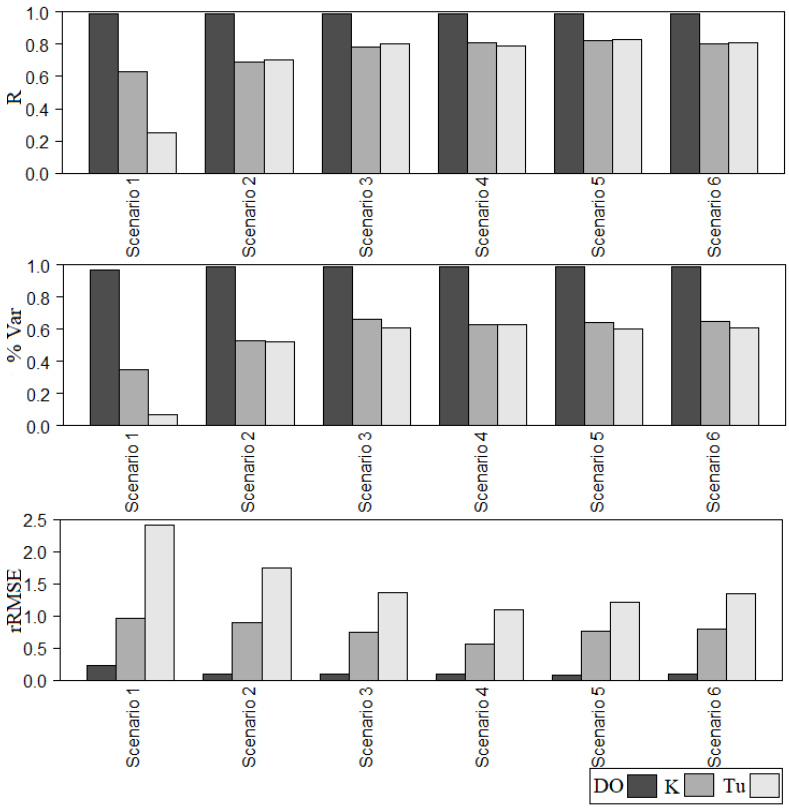


Fig. 3. Correlation coefficient (top), Explained Variance (middle), and *rRMSE* (bottom) of DO, K, and Tu with respect to their corresponding observed values for the model scenarios in Table 2.

Based on these results, the model based on scenario 4, which considers water quality, hydrologic information, watershed size, length of rivers, and land cover, outperforms the other models when considering both the statistical metrics shown in Figure 2 and model efficiency, i.e., the amount of required information. Thus, adding information regarding soil type and livestock count does not improve *R*, *%Var*, and/or *rRMSE* enough to justify the collection of these data, which can be time consuming and expensive in an operational setting. As a result, scenario 4 is selected for further investigation and recommended as the best compromise between performance and efficiency.

4.2. Predictor importance

The importance measures (*MDA*) for each predicted variable are calculated for every scenario. Higher *MDA* values indicate when a predictor variable plays a more important role in estimating the predicted variable. In

other words, if the accuracy of the RF model decreases due to exclusion of a certain predictor, the predictor is critical in developing the RF model.

Figure 4 shows the *MDA* values for scenario 4. When predicting DO, WT is the most important variable, followed by discharge and developed area. It is well known that DO and WT are highly correlated (Galloway 2002). A higher volume of water moves faster and increases the flow turbulence, which results in more oxygen dissolving in the water (Kelly 1997). Also, urbanization results in less impervious surfaces, which increase runoff and can elevate the amount of organic matter in water. Consequently, urbanization alters DO concentration due to organic matter decomposition (Smith et al. 1992).

Precipitation is the most important variable for predicting K. This is expected as precipitation increases runoff that can carry saline-polluted water, resulting in higher K. In addition, it is important to note that discharge, WT, and T are also highly predictive of K. This is consistent with findings from Zavareh et al. (2021). The most important watershed characteristic for predicting K is the area of developed land (urbanization). Like precipitation, urbanization contributes to K, as it decreases the possibility of salinity absorption into the soil and increases salinity in surface water.

Discharge is the most important predictor of Tu. Higher water volume increases the speed of its movement, stirring up the water and increasing turbidity (Dalwadi, Padole 2019). The levels of K and WT are the second and third most important variables predicting Tu. This is in line with past studies that have shown strong Granger causality relationships between WT (cause) and Tu (effect) (Zavareh et al. 2021).

In summary, discharge plays a very important role when predicting DO, K, and Tu. Additionally, the volume of discharge is directly affected by land cover. If the land cover of a watershed changes, the overall water yield (runoff) of the watershed changes, which affects water quality (Kumar et al. 2018). This explains why scenario 4 outperforms scenarios 1-3 (which lack information regarding land use, which may have a critical effect on water quality).

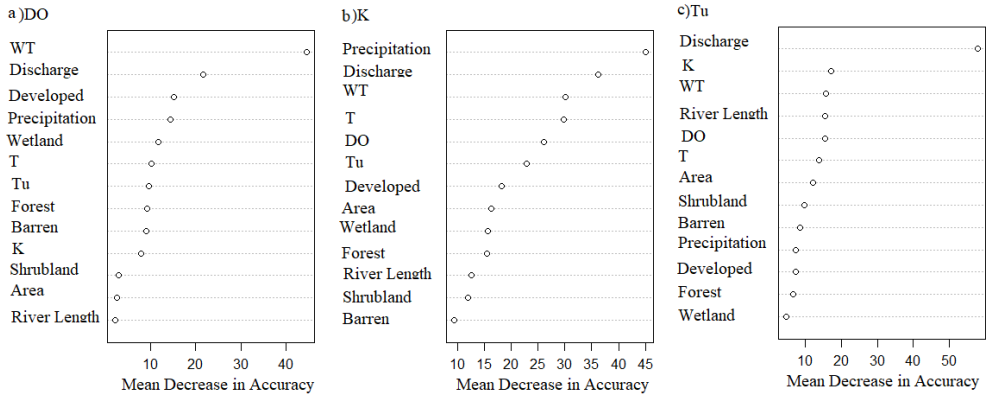


Fig. 4. Mean decrease in accuracy for predictors of the RF regression model built for Scenario 4 for predicting a) DO, b) K, and c) Tu.

4.3. Model validation

In order to assess the applicability of the RF model, we evaluate its performance across an independent watershed, Scotts Level Branch, for which four months of data are available (January 2020 to April 2020). Information on DO was unavailable for this watershed.

Figure 5 shows time series of predicted and corresponding measured values of K and Tu for Scotts Level Branch. Model estimates are presented for the 6 scenarios as an ensemble envelope bounded by the minimum and maximum values obtained across all 6 models.

Observed K values fall within the model ensemble bounds, showing that the model encapsulates the actual values of K and well reproduces its variability over time. However, the model identifies a peak in late February that was not captured by in-situ measurements. This can be either due to an overestimation by the model during a specific precipitation event, or it could be an event missed by the observations. Similarly, some peaks in modeled Tu are not present in the observed time series. Nevertheless, Tu variability during the period of interest is well captured within the model envelope.

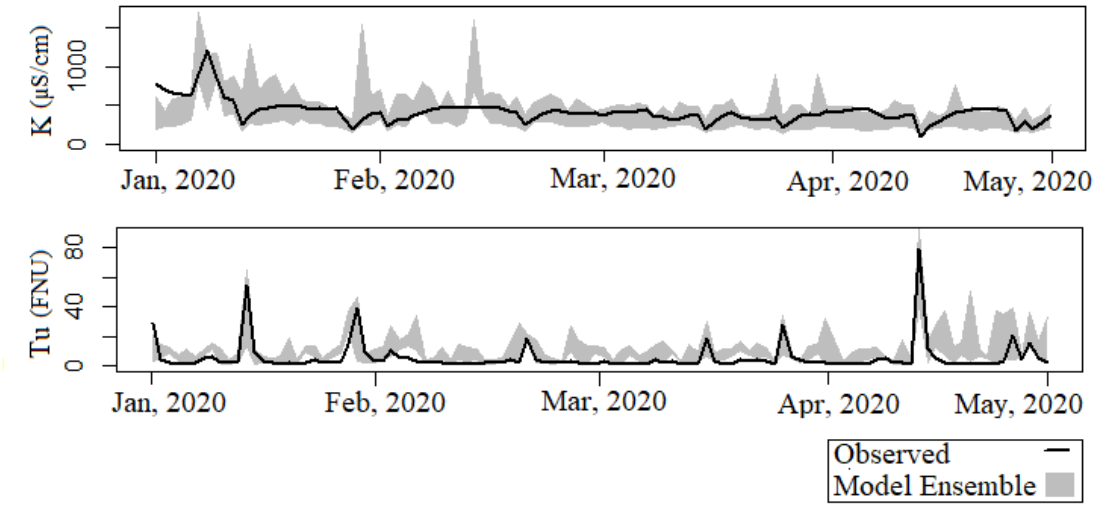


Fig. 5. Time series of modeled and observed K and Tu for Scotts Level Branch. The ensemble of modeled values is shown as a shaded area enveloped between the minimum and maximum values obtained from the models built on the 6 scenarios.

Table 3 shows the results for the three statistical metrics used in this study to evaluate the RF model performance for K and Tu in the validation watershed for the 6 scenarios. Correlation coefficients more than doubled when adding hydrology information to scenario 1 for both K and Tu. The R value improves when more information is added to the model, and as in the training phase, it increases sharply when hydrology information is included in the model (i.e., moving from scenario 1 to 2). The $\%Var$ values for K and Tu more than doubled and tripled when hydrology and watershed physiology information are added (i.e., scenario 1 vs. scenario 3). Conversely, the results of $rRMSE$ do not consistently increase or decrease as more information is

added to the model. However, scenario 4 shows relatively low *rRMSE* compared to other scenarios. In general, when comparing the three statistical metrics, scenario 4 shows the best performance for predicting K and Tu. This is in line with the results for the RF model, as previously discussed.

Table 3. Correlation coefficient (*R*), Explained Variance (*%Var*), and *rRMSE* for predicted and observed K and Tu values in the Scotts Level Branch watershed.

Scenario	K			Tu		
	R	*%Var*	*rRMSE*	*R*	*%Var*	*rRMSE*
1	0.17	0.27	0.51	0.18	0.12	0.53
2	0.57	0.46	0.43	0.79	0.45	0.89
3	0.58	0.64	0.80	0.9	0.56	0.48
4	0.52	0.65	0.41	0.94	0.58	0.45
5	0.54	0.65	0.37	0.89	0.51	0.60
6	0.50	0.64	0.53	0.90	0.60	1.01

Scatterplots of actual and predicted K and Tu for scenario 4 are presented in Figure 6. Although the dispersion around the 1:1 line is consistent, the modeled K values are overall well aligned to K observed in the watershed during the 4-month validation period, with a correlation coefficient of 0.52. In terms of Tu, the model well reproduces large Tu values (correlation coefficient of 0.94), but overestimates observed values of Tu below 10 FNU. This can be potentially improved by considering a larger sample size and verifying the model for a longer time series and/or in a different watershed.

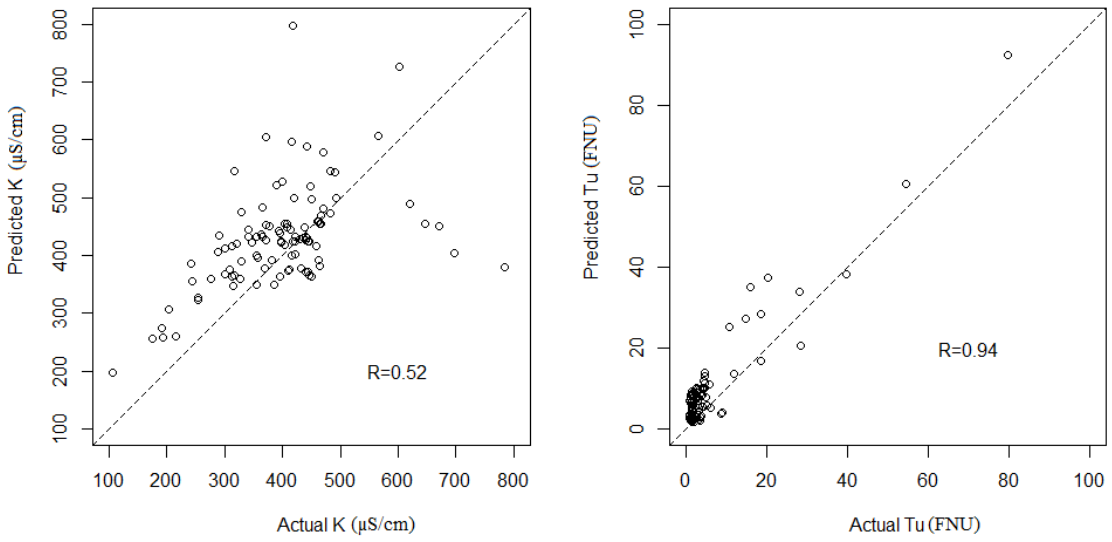


Fig. 6. Scatterplots of observed and predicted K (left) and Tu (right) in the Scotts Level Branch watershed.

5. Conclusions

This study investigates the efficiency of RF regression for predicting water quality indicators (DO, K, and Tu) and provides insight into factors affecting stream water quality. The RF models are built based on information

from 10 watersheds in the DMV region, with one independent watershed used for assessing model applicability. The RF model performance is analyzed based on three statistical metrics, R, $\%Var$, and $rRMSE$. In addition, degree of importance is calculated for each scenario to rank relative contribution of predictors in estimating water quality.

The RF models to predict DO show the highest performance (average $R = 0.99$, average $\%Var = 0.98$, average $rRMSE = 0.11$) when modeling the 10 watersheds. The RF models predicting K (average $R = 0.75$, average $\%Var = 0.57$, and average $rRMSE = 0.82$) slightly outperform the models that predict Tu (average $R = 0.69$, average $\%Var = 0.50$, and average $rRMSE = 1.62$). However, when comparing the scenario performances for DO, K, and TU and taking into account the amount of information needed for developing each model, scenario 4 is the most efficient option. This highlights the importance of land cover information in predicting water quality.

The most important measure for predicting DO is WT, which is expected due to their strong correlation (Galloway 2002). The second and third most important measures of DO are discharge and urbanization. In comparison, precipitation and discharge are the most important measures for predicting K. Among all watershed characteristics, urbanization plays the most important role in predicting K, as it results in greater area of impervious land, which increases runoff volume and the concentration of total dissolved solids (Kumar et al. 2018). When predicting Tu, discharge is the most important measure, as more discharge yields more suspended solids, which increases turbidity. The second most important measure is K, as increased dissolved solids concentration contributes to higher Tu.

An independent watershed is used to assess the performance of the developed models and evaluate their applicability to a different region. Model performance is similar to that observed in the training phase, with scenario 4 (which includes water quality data, hydrology information, watershed size, length of rivers in watersheds, and land cover information) outperforming other scenarios. However, longer time series and different watersheds should be considered to verify these results.

In conclusion, along with watershed physiology and hydrological characteristics, urbanization plays an important role in predicting DO, K, and Tu. In general, land cover highly impacts the production and transportation of sediments and organic matter (Inserillo et al. 2017). This emphasizes the vulnerability of surface water and streams to anthropogenic changes.

It is important to mention that there are limitations in using RF models in water quality data analysis. For instance, extrapolation beyond the training data requires implementing techniques or procedures to mitigate the risks associated with extrapolation, such as using appropriate model validation methods, considering uncer-

tainty estimates, and potentially applying domain knowledge to make informed decision. Additionally, the selection of relevant variables significantly impacts model performance. A comprehensive elucidation of fitting methodologies is imperative to avoid inaccuracy in drawing predictive conclusions.

Future work should extend this study to other regions to verify the effects of climate on the relationships between hydrometeorology and water quality. Additionally, finer temporal resolutions can be considered to investigate rates of hydrological response, especially in watersheds of different sizes. Additional water quality indicators like pH and nitrate concentration would help generalize the results of this work and make the proposed analyses more useful for water quality management. Finally, extreme weather events should be analyzed to understand how they impact model outcomes.

Acknowledgments

Water quality data are provided by the U.S. Geological Survey. The authors thank Ishrat Jahan Dollan for providing Figure 1.

References

Akoto O., Abankwa E., 2014, Evaluation of Owabi Reservoir (Ghana) water quality using factor analysis, Lakes & Reservoirs: Science, Policy and Management for Sustainable Use, 19 (3), 174-182, DOI: 10.1111/lre.12066.

Al-Abadi A.M., Fryar A.E., Rasheed A.A., Pradhan B., 2021, Assessment of groundwater potential in terms of the availability and quality of the resource: a case study from Iraq, Environmental Earth Sciences, 80 (12), DOI: 10.1007/s12665-021-09725-0.

Amiri B.J., Nakane K., 2009, Comparative prediction of stream water total nitrogen from land cover using artificial neural network and multiple linear regression, Polish Journal of Environmental Studies, 18 (2), 151-160.

Boulesteix A.-L., Janitza S., Kruppa J., König I.R., 2012, Overview of random forest methodology and practical guidance with emphasis on computational biology and bioinformatics, WIREs Data Mining and Knowledge Discovery, 2 (6), 493-507, DOI: 10.1002/widm.1072.

Breiman L., 2001, Random forests, Machine Learning, 45 (1), 5-32, DOI: 10.1023/A:1010933404324.

Breiman L., Friedman J.H., Olshen R.A., Stone C.J., 1993, Classification and Regression Trees, Wadsworth Statistics/Probability Series, Chapman & Hall, New York, N.Y., 368 pp.

Burkholder J., Libra B., Weyer P., Heathcote S., Kolpin D., Thorne P.S., Wichman M., 2007, Impacts of waste from concentrated animal feeding operations on water quality, Environmental Health Perspectives, 115 (2), 308-312, DOI: 10.1289/ehp.8839.

Chen G., Long T., Xiong J., Bai Y., 2017., Multiple random forests modelling for urban water consumption forecasting, Water Resources Management, 31 (15), 4715-4729, DOI: 10.1007/s11269-017-1774-7.

Chen S., Fang G., Huang X., Zhang Y., 2018, Water quality prediction model of a water diversion project based on the improved artificial bee colony-backpropagation neural network, Water, 10 (6), DOI: 10.3390/w10060806.

Dalwadi N., Padole M., 2019, The Internet of Things based water quality monitoring and control, Smart Innovation, Systems and Technologies. Innovations in Computing, 141, 409-417, DOI: 10.1007/978-981-13-8406-6_39.

Devi G., 2019, Random forest advice for water quality prediction in the regions of Kadapa District, International Journal of Innovative Technology and Exploring Engineering, 8 (6S4), 1464-1466, DOI: 10.35940/ijitee.F1298.0486S419.

Díaz-Uriarte R., Alvarez de Andrés A., 2006, Gene selection and classification of microarray data using random forest, BMC Bioinformatics, 7 (1), DOI: 10.1186/1471-2105-7-3.

Dubois D., Prade H., 1992, Putting rough sets and fuzzy sets together, [in:] Intelligent Decision Support, R. Słowiński (ed.), Springer Netherlands, Dordrecht, 203-232, DOI: 10.1007/978-94-015-7975-9_14.

Dufour A., Bartram J., Bos R., 2012, Animal Waste, Water Quality and Human Health, IWA Publishing, London, 489.

Fox E.W., Ver Hoef J.M., Olsen A.R., 2020, Comparing spatial regression to random forests for large environmental data sets, PLOS ONE, 15 (3), e0229509, DOI: 10.1371/journal.pone.0229509.

Galloway J.M., 2002, Simulation of Hydrodynamics, Temperature, and Dissolved Oxygen in Norfork Lake, Arkansas, 1994-1995, Water-Resources Investigations Report 02, Little Rock, Ark: USDeptof the Interior, USGeological Survey.

Golkarian A., Naghibi S.A., Kalantar B., Pradhan B., 2018, Groundwater potential mapping using C5.0, random forest, and multivariate adaptive regression spline models in GIS, Environmental Monitoring and Assessment, 190 (3), 1-16, DOI: 10.1007/s10661-018-6507-8.

Han H., Guo X., Yu H., 2016, Variable selection using Mean Decrease Accuracy and Mean Decrease Gini based on Random Forest, [in:] 7th IEEE International Conference on Software Engineering and Service Science (ICSESS), 219-224, DOI: 10.1109/ICSESS.2016.7883053.

Hulten G., 2018, Building Intelligent Systems: A Guide to Machine Learning Engineering, Apress, New York, 339 pp., DOI: 10.1007/978-1-4842-3432-7.

Imani M., Hasan M.M., Bittencourt L.F., McClymont K., Kapelan Z., 2021, A novel machine learning application: water quality resilience prediction model, Science of The Total Environment, 768, DOI: 10.1016/j.scitotenv.2020.144459.

Inserillo E.A., Green M.B., Shanley J.B., Boyer J.N., 2017, Comparing catchment hydrologic response to a regional storm using specific conductivity sensors, Hydrological Processes, 31 (5), 1074-1085, DOI: 10.1002/hyp.11091.

Jadhav M.S., Khare K.C., Warke A.S., 2015, Water quality prediction of Gangapur Reservoir (India) using LS-SVM and genetic programming, Lakes & Reservoirs: Science, Policy and Management for Sustainable Use, 20 (4), 275-284, DOI: 10.1111/lre.12113.

Jeong K.-S., Joo G.-J., Kim H.-W., Ha K., Recknagel F., 2001, Prediction and elucidation of phytoplankton dynamics in the Nakdong River (Korea) by means of a recurrent artificial neural network, Ecological Modelling, 146 (1-3), 115-129, DOI: 10.1016/S0304-3800(01)00300-3.

Karamizadeh S., Abdullah S.M., Manaf A.A., Zamani M., Hooman A., 2013, An overview of principal component analysis, Journal of Signal and Information Processing, 4 (3B), 173-175, DOI: 10.4236/jsip.2013.43B031.

Kelly V.J., 1997, Dissolved oxygen in the Tualatin River, Oregon, during winter flow conditions, 1991 and 1992, United States Geological Survey Water-Supply Paper, 2465, U.S. Geological Survey, 74 pp., DOI: 10.3133/ofr95451.

Kijewski T., Zbawicka M., Strand J., Kautsky H., Kotta J., Rätsep M., Wenne R., 2019, Random forest assessment of correlation between environmental factors and genetic differentiation of populations: case of marine mussels Mytilus, Oceanologia, 61 (1), 131-142, DOI: 10.1016/j.oceano.2018.08.002.

Kumar S., Moglen G.E., Godrej A.N., Grizzard T.J., Post H.E., 2018, Trends in water yield under climate change and urbanization in the US Mid-Atlantic region, Journal of Water Resources Planning and Management, 144 (8), DOI: 10.1061/(ASCE)WR.1943-5452.0000937.

Lagomarsino D., Tofani V, Segoni S, Catani F., Casagli N., 2017, A tool for classification and regression using random forest methodology: applications to landslide susceptibility mapping and soil thickness modeling, Environmental Modeling & Assessment, 22 (3), 201-214, DOI: 10.1007/s10666-016-9538-y.

Li M., Zhang Y., Wallace J., Campbell E., 2020, Estimating annual runoff in response to forest change: a statistical method based on random forest, Journal of Hydrology, 589, DOI: 10.1016/j.jhydrol.2020.125168.

Liaw A., Wiener M., 2002, Classification and regression by randomForest, R News, 2-3, 18-22.

Long W.J., Griffith J.L., Selker H.P., D'Agostino R.B., 1993, A comparison of logistic regression to decision-tree induction in a medical domain, Computers and Biomedical Research, 26 (1), 74-97, DOI: 10.1006/cbmr.1993.1005.

Mezrich J.J., 1994, When is a tree a hedge?, Financial Analysts Journal, 50 (6), 75-81, DOI: 10.2469/faj.v50.n6.75.

Mitchell T.M., 2013, Machine Learning, McGraw-Hill Series in Computer Science, McGraw-Hill, New York.

Najah A.A., Othman F.B., Afan H.A., Ibrahim R.K., Fai C.M., Hossain M.S., Ehteram M., Elshafie A., 2019, Machine learning methods for better water quality prediction, Journal of Hydrology, 578, DOI: 10.1016/j.jhydrol.2019.124084.

Norouzi H., Moghaddam A.A., 2020, Groundwater quality assessment using random forest method based on groundwater quality indices (case study: Miandoab plain aquifer, NW of Iran), Arabian Journal of Geosciences, 13 (18), DOI: 10.1007/s12517-020-05904-8.

Papacharalampous G.A., Tyralis H., 2018, Evaluation of random forests and Prophet for daily streamflow forecasting, Advances in Geosciences, 45, 201-218, DOI: 10.5194/adgeo-45-201-2018.

Parkhurst D.F., Brenner K.P., Dufour A.P., Wymer L.J., 2005, Indicator bacteria at five swimming beaches—analysis using random forests, Water Research 39 (7), 1354-1360, DOI: 10.1016/j.watres.2005.01.001.

Pedregosa F., Varoquaux G., Gramfort A., Michel V., Thirion B., Grisel O., Blondel M., Prettenhofer P., Weiss R., Dubourg V., Vanderplas J., Passos A., Cournapeau D., Brucher M., Prrot M., Duchesnay E., 2011, Scikit-learn: machine learning in Python, The Journal of Machine Learning Research, 12 (85), 2825-2830.

Rokach L., 2010, Ensemble-based classifiers, Artificial Intelligence Review, 33 (1-2), 1-39, DOI: 10.1007/s10462-009-9124-7.

Saadi M., Oudin L., Ribstein P., 2019, Random forest ability in regionalizing hourly hydrological model parameters, Water, 11 (8), DOI: 10.3390/w11081540.

Sameen M.I., Pradhan B., Lee S., 2019, Self-learning random forests model for mapping groundwater yield in data-scarce areas, Natural Resources Research, 28 (3), 757-775, DOI: 10.1007/s11053-018-9416-1.

Singh B., Sihag P., Singh K., 2017, Modelling of impact of water quality on infiltration rate of soil by random forest regression, Modeling Earth Systems and Environment, 3 (3), 999-1004, DOI: 10.1007/s40808-017-0347-3.

Smith D.E., Leffler M., Mackiernan G., 1992, Oxygen Dynamics in the Chesapeake Bay: A Synthesis of Recent Research, technical report, College Park, Md: Maryland Sea Grant College in cooperation with the Virginia Sea Grant College.

Solakian J., Maggioni V., Godrej A.N., 2020, On the performance of satellite-based precipitation products in simulating streamflow and water quality during hydrometeorological extremes, Frontiers in Environmental Science, (8), DOI: 10.3389/fenvs.2020.585451.

Tesoriero A.J., Gronberg J.A., Juckem P.F., Miller M.P., Austin B.P., 2017, Predicting redox-sensitive contaminant concentrations in groundwater using random forest classification, Water Resources Research, 53 (8), 7316-7331, DOI: 10.1002/2016WR020197.

Tiyasha T.M.T., Yaseen Z.M., 2020, A survey on river water quality modelling using artificial intelligence models: 2000-2020, Journal of Hydrology, 585, DOI: 10.1016/j.jhydrol.2020.124670.

Tyralis H., Papacharalampous G., Langousis A., 2019, A brief review of random forests for water scientists and practitioners and their recent history in water resources, Water, 11 (5), 910, DOI: 10.3390/w11050910.

Wang F., Wang Y., Zhang K., Hu M., Wenig Q., Zhang H., 2021, Spatial heterogeneity modeling of water quality based on random forest regression and model interpretation, Environmental Research, 202, DOI: 10.1016/j.envres.2021.111660.

Wang X., Liu T., Zheng X., Peng H., Xin J., Zhang B., 2018, Short-term prediction of groundwater level using improved random forest regression with a combination of random features, Applied Water Science, 8 (5), DOI: 10.1007/s13201-018-0742-6.

Wang X., Zhang F., Ding J., 2017, Evaluation of water quality based on a machine learning algorithm and water quality index for the Ebinur Lake watershed, China, Scientific Reports, 7 (1), DOI: 10.1038/s41598-017-12853-y.

Wu D., Wang H., Seidu R., 2020, Smart data driven quality prediction for urban water source management, Future Generation Computer Systems, 107, 418-432, DOI: 10.1016/j.future.2020.02.022.

Yu X., Shen J., Du J., 2020, A machine-learning-based model for water quality in coastal waters, taking dissolved oxygen and hypoxia in Chesapeake Bay as an example, Water Resources Research, 56 (9), DOI: 10.1029/2020WR027227.

Zabihi M., Pourghasemi H.R., Pourtaghi Z.S., Behzadfar M., 2016, GIS-based multivariate adaptive regression spline and random forest models for groundwater potential mapping in Iran, Environmental Earth Sciences, 75 (8), DOI: 10.1007/s12665-016-5424-9.

Zavareh M., Maggioni V., 2018, Application of rough set theory to water quality analysis: a case study, Data, 3 (4), DOI: 10.3390/data3040050.

Zavareh M., Maggioni V., Sokolov V., 2021, Investigating water quality data using principal component analysis and granger causality, Water, 13 (3), DOI: 10.3390/w13030343.

Zhang Q., Murphy R.R., Tian R., Forsyth M.K., Trentacoste E.M., Keisman J., Tango P.J., 2018, Chesapeake Bay's water quality condition has been recovering: insights from a multimetric indicator assessment of thirty years of tidal monitoring data, Science of the Total Environment, 637-638, 1617-1625, DOI: 10.1016/j.scitotenv.2018.05.025.

Zhao D., Wu Q., Cui F., Xu H., Zeng Y., Cao Y., Du Y., 2018, Using random forest for the risk assessment of coal-floor water inrush in Panjiayao coal mine, Northern China, Hydrogeology Journal, 26 (7), 2327-2340, DOI: 10.1007/s10040-018-1767-5.

A novel hybrid framework to model the relationship of daily river discharge with meteorological variables

Maha Shabbir, Sohail Chand
University of the Punjab, Pakistan
Farhat Iqbal
Imam Abulrahman Bin Faisal University, Saudi Arabia

Abstract

River discharge is affected by many factors, such as water level, rainfall, and precipitation. This study proposes a new hybrid framework named LAES (LASSO-ANN-EMD-SVM) to model the relationship of daily river discharge with meteorological variables. This hybrid framework is a composite of the least absolute shrinkage and selection operator (LASSO), an artificial neural network (ANN), and an error correction method. In the first stage, LASSO identifies meteorological variables that have a significant influence on the generation of river discharge. Next, the ANN model is used to predict river discharge using meteorological variables selected by LASSO, and the error series is determined. The error series is decomposed into intrinsic mode functions and residuals using empirical mode decomposition (EMD). The EMD components are modeled using the support vector machine (SVM) model, and the error predictions are aggregated. In the last stage, the LASSO-ANN predictions and the predicted error series are aggregated as the final discharge prediction. The proposed hybrid framework is illustrated on the Kabul River of Pakistan. The performance of the proposed hybrid framework is compared with six models using various performance measures and the Diebold-Mariano test. These models include multiple linear regression (MLR), SVM, ANN, LASSO-MLR, LASSO-SVM, and LASSO-ANN models. The findings reveal that the proposed hybrid model outperforms all other models considered in the study. In the testing phase, the root mean square error (*RMSE*), mean absolute percentage error (*MAPE*), and mean absolute error (*MAE*) of the proposed LAES hybrid model are 337.143 m^3/s, 32.354%, and 218.353 m^3/s which are smaller than all other models compared in the study. Our proposed hybrid system is an efficient model for river discharge prediction that will be helpful in water management and protection against floods. Long-term prediction can help to identify the major effects of climate change and to make evidence-based environmental policies.

Keywords

LASSO, river discharge, ANN, SVM, EMD.

Submitted 5 September 2023, revised 22 April 2024, accepted 24 April 2024
DOI: 10.26491/mhwm/187899

1. Introduction

Water is necessary for the survival of all living organisms in the world. Water is life, and demand for it is increasing due to rapid increases in population, urbanization, and industrialization. Moreover, water is a primary need for domestic, industrial, and agricultural activities (Mehta et al. 2022). Thus, it is essential to carefully manage and plan water resources to reduce loss of life and property damage caused by drought, floods, or heat waves (Ali, Shahbaz 2020; Mangukiya et al. 2022). Climate changes influence the hydrological cycle globally; the resulting variations in weather and climate have increased the risks of drought and floods because weather changes, variations in precipitation, peak flows, and extreme temperatures have impacts on river discharge (Mehmood et al. 2021). The amount of discharge generated from a catchment depends on various factors such as duration, meteorological variables, velocity, and water level (Gleason et

al. 2014; Saidi et al. 2018; Malik et al. 2020). Therefore, it is necessary to model river discharge using information on the weather at the relevant hydrological station (Dariane, Azimi 2018).

In the past thirty years, stochastic, physical, black box (machine learning and statistical), and conceptual models have been widely applied in hydrological studies. Physical models have been used for hydrological modeling, but their successful application is bound to the complexity of governing equations and the difficulty in measuring the parameters involved (Yousuf et al. 2017). Statistical models try to determine the relationships within the actual data. Their application is limited when data have unique and complex characteristics such as non-linearity, multicollinearity, volatility, irregularities, noise, outliers, and more. In the past two decades, machine learning models have gained importance in hydrology due to their flexibility in handling datasets with unique characteristics (Ravindran et al. 2021; Elbeltagi et al. 2022). Rasouli et al. (2012) applied a support vector machine (SVM), Bayesian neural network, and Gaussian process to predict non-linear river discharge in North America using climate and weather variables. Ali and Shahbaz (2020) applied an artificial neural network (ANN) to predict river discharge in the upper Jhelum River basin of Pakistan.

Although data-driven (statistical and machine learning) models are applied to predict river discharge, there is no single model that can predict river discharge without bias or with utmost certainty (Mehmood et al. 2021). Literature shows that researchers have developed hybrid models by combining two or more techniques to improve the prediction ability of the models (Shabbir et al. 2024). Wang and Li (2018) introduced a hybrid framework based on an error correction approach using the generalized autoregressive conditionally heteroscedastic (GARCH) model when inherent correction and heteroscedasticity of errors cannot be ignored. Zhang et al. (2018a) developed an error-correction-based hybrid framework using an autoregressive (AR) model to predict water levels with improved accuracy. Luo et al. (2019) suggested a hybrid framework based on a composition of factor analysis, decomposition of time series, data regression, and error suppression to predict river discharge. Yan et al. (2020) combined a generalized additive model (GAM) with principal component analysis (PCA) to model the relationship between water level and macroinvertebrate diversity index in the Baiyandian Lake of China. Mehr and Gandomi (2021) suggested a hybrid model by integrating a multi-stage genetic programming (MSGP) model with the least absolute shrinkage and selection operator (LASSO) for improved prediction of river flow. Emadi et al. (2022) modeled river water using a hybrid evolutionary data-driven approach.

River discharge estimation is challenging in hydrological studies because its generation depends on various factors such as rainfall patterns, spatial-temporal irregularities, climatic changes, and many more (Cheng et al. 2019; Hu et al. 2022). In literature, much discussion is on the time series prediction of river discharge (see Luo et al. 2019; Mehr, Gandomi 2021; Adnan et al. 2022). There is an essential need to develop new methods to evaluate the possible influence of different factors on the generation of river discharge. Keeping in view this gap, this study aims to develop a new hybrid approach to examine the relationship between river discharge and meteorological variables.

A new hybrid framework named LAES (LASSO-ANN-EMD-SVM) is proposed in this study based on a combination of feature selection, an ANN model, and an error correction method. In the first stage, LASSO is used to identify meteorological variables that have significant relationships with river discharge. The variables identified by LASSO are then used as input variables to the ANN model to obtain the discharge predictions, and then the error series is computed. Further, the empirical mode decomposition (EMD) technique is used to decompose error series into intrinsic mode functions and residuals. These components are modeled using the SVM model, and their predictions are aggregated. The final discharge prediction is obtained by adding the LASSO-ANN discharge predictions with EMD-SVM error predictions. Application of the proposed LAES hybrid framework is demonstrated for the Kabul River of Pakistan, and its prediction performance is compared with different models.

The proposed hybrid framework is novel as it efficiently predicts river discharge by considering the influence of meteorological variables that have a significant impact on river discharge using LASSO. In addition, the error correction approach in the proposed LAES hybrid model helps to enhance the prediction of discharge by capturing the randomness and volatility of the error series. It provides reliable estimates of river discharge and can be helpful in the management of water supply and flood control.

2. Methods

2.1. Multiple linear regression

The multiple linear regression (MLR) model is a simple and widely used modeling technique. The MLR model is given as:

$$y_j = \beta_0 + \beta_1 x_{1j} + \beta_2 x_{2j} + \cdots + \beta_p x_{pj} + u_j, \quad j = 1,2,\ldots,n \tag{1}$$

where y_j is the dependent (output) variable, β_j are the regression coefficients, x_j are the independent (input) variable, n is the number of observations, p is the number of independent variables, and u_j is the residual term.

2.2. Least absolute shrinkage and selection operator

Tibshirani (1996) introduced the least absolute shrinkage and selection operator (LASSO) as a variable-selection approach for regression models. The method minimizes the residual sum of squares subject to the absolute values of the regression coefficients. LASSO performs variable selection and regularization simultaneously to enhance the interpretability and precision of statistical models (Tibshirani 1996). This study applies LASSO to determine important meteorological variables for predicting river discharge.

Assuming a sample contains M events where each event has p number of independent variables and one dependent variable, let $\mathbf{y_i}$ be the dependent (output) variable, and $\mathbf{x_i} = \left(x_1, x_2, \ldots, x_p\right)^T$ be the vector of i^{th} independent (input) variables, then the objective function of LASSO is:

For all $\sum_{j=1}^{p}|\beta_j| \leq \lambda$, find the $\min_{\beta} \frac{1}{M}\sum_{i=1}^{M}\left(\boldsymbol{y_i} - \mathbf{x_i^T}\boldsymbol{\beta}\right)^2$ (2)

where λ is a pre-determined parameter that determines the regularization degree and $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_p)$ is the vector of regression coefficients. Let $\mathbf{X}$ be the matrix of independent variables, i.e. $\mathbf{X_{ij}} = (x_i)_j$, where $i = 1, 2, \ldots, M, j = 1, 2, \ldots, p$ and x_i^T is the i^{th} row of $\mathbf{X}$. Then, the above formula in a compact form can be written as:

For all $\|\boldsymbol{\beta}\|_1 \leq \lambda$, calculate $\min_{\beta}\left\{\frac{1}{M}\|\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2\right\}$ (3)

where $\|\boldsymbol{\beta}\|_p = \left(\sum_{i=1}^{M}|\beta_i|_p\right)^{1/p}$ is the standard L_p norm, $\mathbf{1_M}$ is a column vector of M dimensions with entries 1. In this study, LASSO is employed using the optimal glmnet library in R language and the optimal value of the LASSO parameter using this library is obtained using a 5-fold cross-validation approach.

2.3. Artificial neural network

The artificial neural network (ANN) is a robust modeling tool in which information processing is a representation of biological systems (Kachrimanis et al. 2003). The network is constructed from interconnected neurons, which can determine values from the inputs through network processing. The neuron receives input signals and provides the output signal that mainly depends on the neuron processing function. The ANN architecture consists of a series of interlinked neuron layers. Every layer is linked with another layer through neurons, which transfer information between these layers. Through this processing, the information reaches the output (dependent variable) layer. The ANN mechanism follows four assumptions:

a) Inputs are handled by neurons.
b) Through the connection of neurons, the information of inputs is passed on to the adjacent layers.
c) Each neuron has a weight, and the output from the neuron is the product of its input and its associated weight.
d) The transmitted inputs are passed via the activation of neurons to obtain the output.

Figure 1a shows the architecture of the ANN model, and Figure 1b presents the structure of a neuron where every input (independent variable) comes from other neurons and are multiplied by their weights $(w_j; j = 1,2, \ldots, n)$ respectively and then aggregated with the bias ($\mathbf{b}$) vector. This aggregated input (s) is passed using the transfer or activation function (f) to obtain the output (a) of a specific neuron. Letting $\mathbf{x}$ be the vector of independent (input) variables, the neural network maps into another output vector $\mathbf{a}$ through:

$$\mathbf{a} = f(\mathbf{x}.w + \mathbf{b}) \quad (4)$$

The mean squared error (*MSE*) is computed and using the back-propagation process, the weights of the entire network are modified in the training process. The accuracy of the ANN depends on the quality and amount of data in training.

In this study, the ANN algorithm is trained by a back-propagation technique where the output and input variables are applied in the network. The Broyden-Fletcher-Goldfarb-Shanno (BFGS) optimization is employed in a three-hidden-layer network. In the input layer of the ANN algorithm, the activation function is applied with 1000 iterations in the hidden layers. In this study, the ANN algorithm is applied using the validant library in the R programming language.

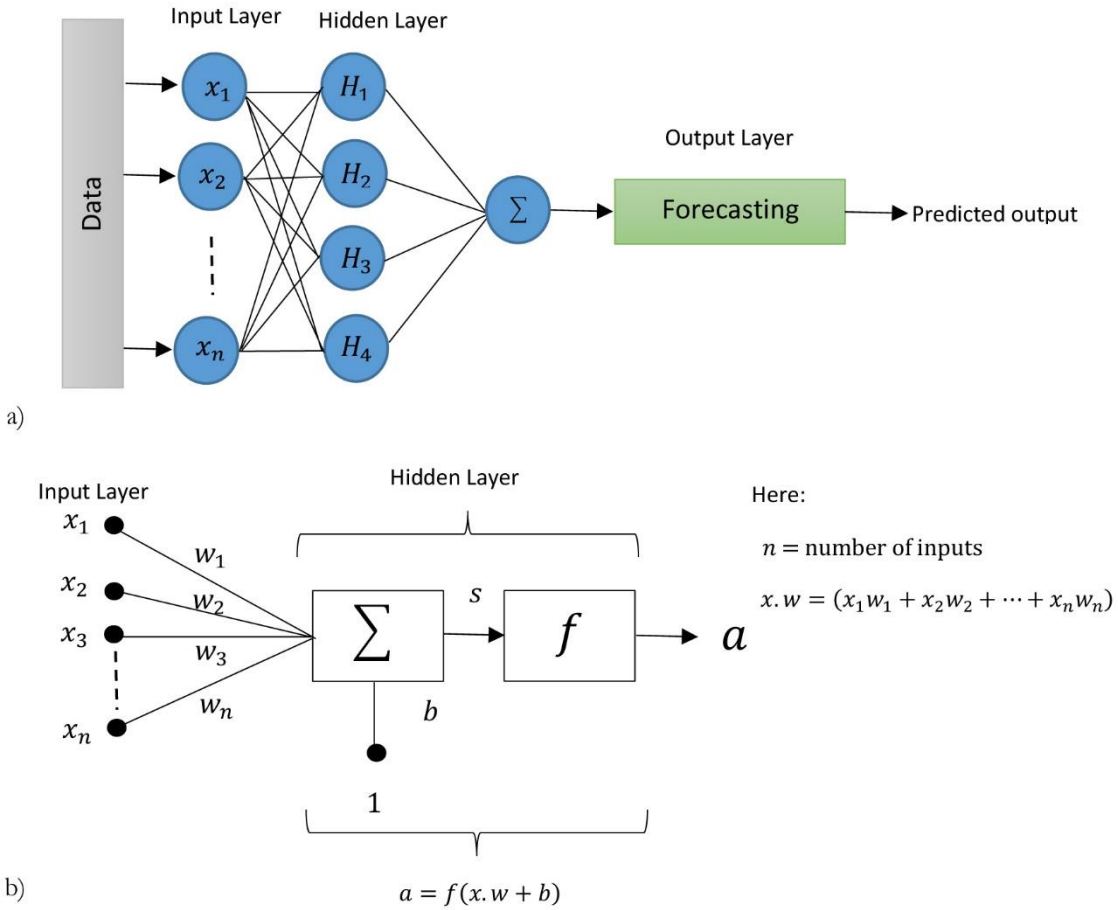


Fig. 1. The mathematical model of ANN (a) and systematic representation of a neuron (b).

2.4. Empirical mode decomposition

Huang et al. (1998) introduced empirical mode decomposition (EMD) as an adaptive method for signal analysis. The EMD is designed to analyze non-linear series. The EMD approach assumes that a signal contains different intrinsic mode functions ($IMFs$) of oscillations. Every mode has the same number of extrema and zero-crossings. There is a single extremum between successive zero-crossings. In this way, the signal is decomposed into different $IMFs$ and residuals. A component is an IMF if it satisfies two conditions: (i) the number of extrema and the number of zero-crossings must be equal to one or differ at most by one, and (ii) at any point, the average of the envelope is zero (Huang et al. 1998). Any original signal $y(t)$ can be decomposed using the EMD algorithm as follows (Lei et al. 2003; Jungsheng et al. 2006):

a) Find the local minima and maximum through the cubic spline line as the upper envelope and lower envelope, respectively.

b) Find the mean (m_1) of upper and lower envelopes.

c) The difference between the $y(t)$ and the 1^{st} component m_1 is the first component denoted as h_1 i.e. $h_1 = y(t) - m_1$. If h_1 is an IMF, then it is said to be the first IMF component of $y(t)$.

d) If h_1 is not an IMF, then it is treated as an original signal, and the steps (a)-(c) are repeated, then $h_1 - m_{11} = h_{11}$.

After repeating the sifting process k times, h_{1k} becomes an IMF , i.e. $h_{1(k-1)} - m_{1k} = h_{1k}$, then it is termed as:

$$c_1 = h_{1k} \tag{5}$$

The first IMF component from the data.

e) Next, subtract c_1 from $y(t)$ to obtain $u_1 = y(t) - c_1$ where u_1 denotes the treated data, and the process is repeated n times to get n $IMFs$ of $y(t)$. Then,

$$\left.\begin{array}{c} u_1 - c_2 = u_2 \\ \vdots \\ u_{n-1} - c_n = u_n \end{array}\right\} \tag{6}$$

At the end of the process, we have $IMFs$ $(c_j; j = 1,2, \dots, n)$ and residual (u_j). By summation of all the components, the original signal $y(t)$ can be obtained as:

$$y(t) = \sum_{j=1}^{n} c_j + u_n \tag{7}$$

The EMD method is implemented using the *EMD* library in R language in this study.

2.5. Support vector machine

Support vector machine (SVM) is a popular modeling technique for classification and regression problems. The SVM algorithm maps complex high-dimensional data into high-feature space (Vapnik 1995). We assume a training set with n observations, $\{x_d, y_d\}, d = 1,2, \dots, n, x_d \in R, y_d \in R$, where y_d denotes the estimated value of the dependent (output) variable, x_d is the corresponding lagged values of the dependent variable, and n is the sample size. Then, the SVM is developed as:

$$f(x) = \boldsymbol{\omega}^{\mathbf{T}} \varphi(x) + b \tag{8}$$

where $f(x)$ is the estimated dependent variable, $b \in R$ is the bias, and $\boldsymbol{\omega} \in R$ represents the vector of weights. The transfer function $\varphi(x)$ maps input data into high-dimensional space. The Eq. (8) is solved by risk minimization as follows:

$$\text{Minimum: } \left(\frac{\|\boldsymbol{\omega}^2\|}{2} + c\sum_{d=1}^{n}(\xi^* + \xi)\right) \text{ subject to: } \begin{cases} f(x_d) - y_d \leq \varepsilon + \xi^* \\ y_d - f(x_d) \leq \varepsilon + \xi \\ \xi, \xi^* \geq 0 \end{cases} \quad (9)$$

where $c > 0$ represents the penalty parameter, ξ and $\xi^*$are slack variables that show the upper and lower constraint of $f(x)$, and ε denotes the insensitive loss function. Further, the Lagrangian function is used as the non-linear regression function, which replaces $\varphi(x)$ and $\boldsymbol{\omega}$ in Eq. (8) as:

$$f(x_d) = \sum_{d=1}^{n}(\alpha_d - \alpha_d^*)k(x, x_d) + b \quad (10)$$

where $k(x, x_d) = \langle \varphi(x), \varphi(x_d) \rangle$ is the kernel function. The α_d^* and α_d represents the Lagrange coefficients.

In this study, SVM is applied to capture the features of the error series using the radial basis function (RBF) kernel, i.e. $k(x, x_d) = e^{\frac{\|x - x_d\|}{2g^2}}$, where g is the width of RBF (Baydaroğlu et al. 2018). The SVM algorithm is applied in this study using the R language e1071 library with unit cost and $g = 1/m$ where m is the number of input variables.

3. Proposed hybrid framework

In this paper, we propose a novel LASSO-ANN-EMD-SVM (LAES) hybrid framework to predict daily river discharge based on its relationship with the meteorological variables. The proposed LAES hybrid framework is displayed in Figure 2.

The steps of the LAES framework are:

a) LASSO is applied for the selection of meteorological variables that influence discharge (y) of the river.

b) Next, the ANN model is employed to model river discharge using meteorological variables as independent variables and the predictions of river discharge $(\hat{y}_{LA})$ are obtained. Further, the error (i.e. $\hat{e} = y - \hat{y}_{LA}$) is computed.

c) Using EMD, the error is decomposed into sub-series, and then the SVM model is used to predict each sub-series. By aggregating them, the predicted error $(\hat{e}_{ES})$ is obtained.

d) The final river discharge prediction is obtained using the predicted error series to correct the predicted river discharge in stage II $(\text{i.e. } \hat{y}_{LAES} = \hat{y}_{LA} + \hat{e}_{ES})$.

The proposed LAES hybrid method is a unique combination of the feature selection method with the ANN model and error correction approach. To the best of our knowledge, there is no hybrid model in the literature that integrates LASSO with an error correction approach for modeling non-linear and high-dimensional data sets.

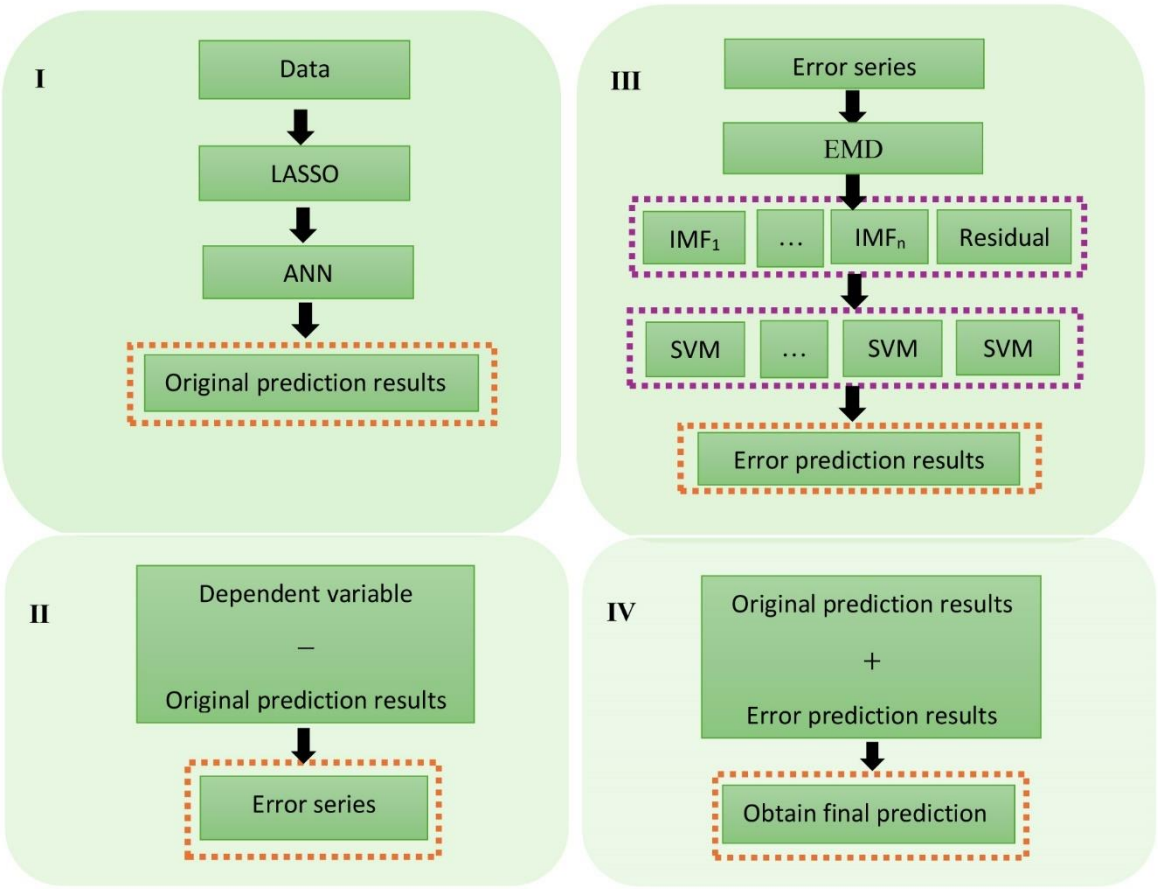


Fig. 2. The proposed LAES hybrid framework.

3.1. Limitations of LAES hybrid framework

The efficiency of the LAES hybrid framework depends on the optimal choice of parameters of the LASSO approach. This framework works efficiently when the independent variables are selected using the optimal value of the LASSO parameter and the information loss by dropping variables is minimal. A high value of the LASSO parameter can contribute toward a loss of information, which may result in poor model fit. Secondly, the performance of the proposed hybrid method depends on the availability of data variables that may vary in different regions of the world due to differences in weather characteristics. The performance of the LAES hybrid model may vary with respect to changes in region (or location) of study and climatic conditions.

4. Application

Data and performance measures are described in this section. The codes of this study were written in R language version 4.1.0. The complete analysis is performed on a personal computer with an Intel Core i9-9900 CPU (32GB RAM).

4.1. Description of data

The Khyber Pakhtunkhwa province is a mountainous region, including the Tirich Mir, Lalazar, Hindu Kush, and some other mountain ranges. The changing climate of this region affects air temperature, water

flows, precipitation, and groundwater resources for irrigation systems and domestic use. These conditions make the northern area of Pakistan prone to drought or flooding due to changing environment and weather conditions.

The Kabul River begins at the Unai pass base from the Hindu Kush mountains in Afghanistan, flowing toward the east and spanning 700 km to drain into the Indus River of Pakistan (Mehmood et al. 2021). The Kabul River at Nowshera station is located at a latitude of 34°0'25"N and longitude of 71°58'50"E. The hydrometeorological regime is characterized by rain in the spring and snow in the winter. The melting of glaciers in summer is increasing each year due to high temperatures, leading to rising water levels in the river (Rasouli 2022). In addition, rainfall in the monsoon season also affects water levels in the river. The Kabul River is influenced by varying climatic conditions, which may lead to hydrometeorological hazards (i.e., heatwaves, floods or drought).

Figure 3 shows the location of the Kabul River in Pakistan. Kabul River data was collected from the Surface Water Hydrology Project (SWHP) Department of the Water and Power Development Authority of Pakistan (WAPDA) from 1st January 2005 to 31st December 2017. The data contain river discharge and meteorological variables. The meteorological variables include air temperature (minimum and maximum), pan water (minimum and maximum), relative humidity (8 AM and 5 PM), dew point (8 AM and 5 PM), evapotranspiration, and wind speed. Average temperature and precipitation have high variability across the basin. River flow has been high during the monsoon period in Pakistan, particularly in July and August. In the midst of 2005, 2010, and 2015, there was extensive flooding due to high temperatures and heavy rainfall in the region. The discharge had some missing values, which were replaced with the monthly average (mean) value. Outliers present in the data were also replaced by median of the respective month. The number of observations for each variable is 4748, approximately 365 daily values for 13 years.

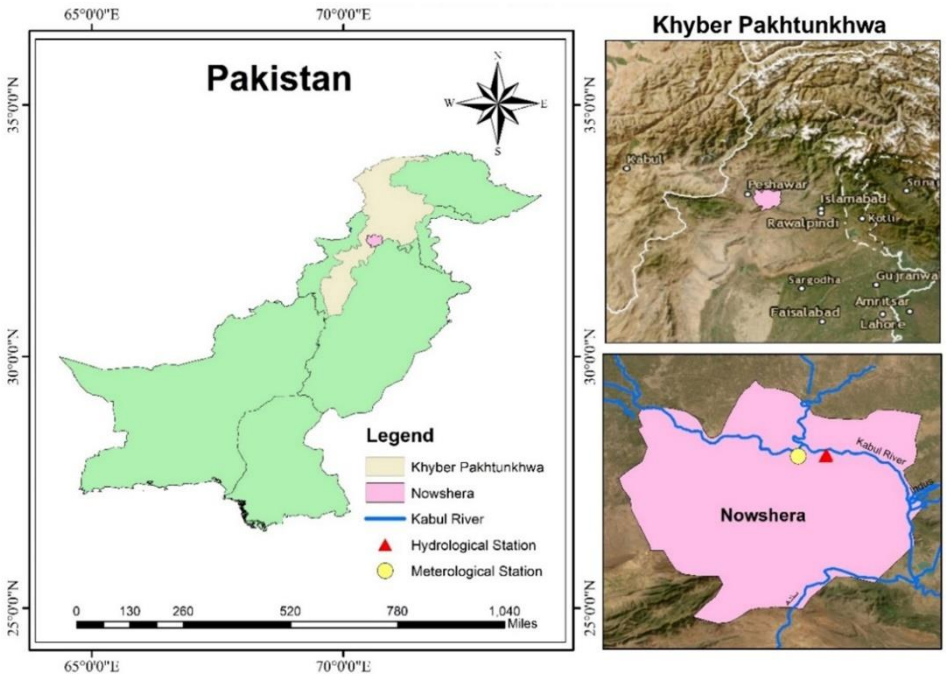


Fig. 3. Location of Kabul River in Pakistan.

Table 1 shows summary descriptions of all the variables of the Kabul River data. The air temperature (maximum), air temperature (minimum), pan water (maximum), pan water (minimum), dew point (8 AM and 5 PM), relative humidity (8 AM and 5 PM) have negatively skewed distributions, while river discharge, wind speed, evapotranspiration, precipitation and rainfall have positively skewed distributions. The average discharge in the Kabul River is 871.8 m^3/s. Figure 4 shows the Kabul River discharge series. It shows that there are non-linear relationships between river discharge and all meteorological variables.

Table 1. Descriptive summary of variables.

Variables	Units	Variables	Mean	Minimum	Maximum	Standard Deviation	Skewness
River discharge	m^3/s	y	871.8	68.7	4724.0	750.7	1.4
Air Temperature Maximum	°F	x_1	85.0	5.0	122.0	15.4	–0.3
Air Temperature Minimum	°F	x_2	64.0	5.0	110.0	13.9	–0.1
Pan Water Maximum	°F	x_3	79.9	8.0	112.0	14.3	–0.3
Pan Water Minimum	°F	x_4	72.8	16.0	106.0	13.1	–0.1
Dew point 8 AM	°F	x_5	61.2	–9.0	93.0	13.6	–0.1
Dew point 5 PM	°F	x_6	70.1	12.0	110.0	15.5	–0.1
Relative Humidity 8 AM	%	x_7	81.7	4.0	100.0	14.4	–1.7
Relative Humidity 5 PM	%	x_8	70.9	1.0	100.0	15.9	–0.9
Wind Speed	mph	x_9	30.6	0.0	170.0	24.3	1.3
Evapotranspiration	mm d^{-1}	x_{10}	5.1	0.0	27.9	5.1	0.9
Precipitation	mm d^{-1}	x_{11}	2.7	0.0	91.0	8.2	4.7
Rainfall	mm d^{-1}	x_{12}	3.3	0.0	161.0	11.5	6.1

The data variables were normalized using the following (Duan et al. 2021):

$$z_{normal} = \frac{z - z_{min}}{z_{max} - z_{min}} \tag{11}$$

where z is the original data variable, z_{normal} is the normalized data variable, z_{min} is the minimum value, and z_{max} is the maximum value of the original data variable. After normalization, the dataset is divided into two parts, where 80% of the data are used for training and the remaining 20% for testing (Kisi et al. 2021; Shabbir et al. 2022). The performance of models is evaluated by 5-fold cross-validation using different performance evaluation measures and the average results of these indicators for training and testing data.

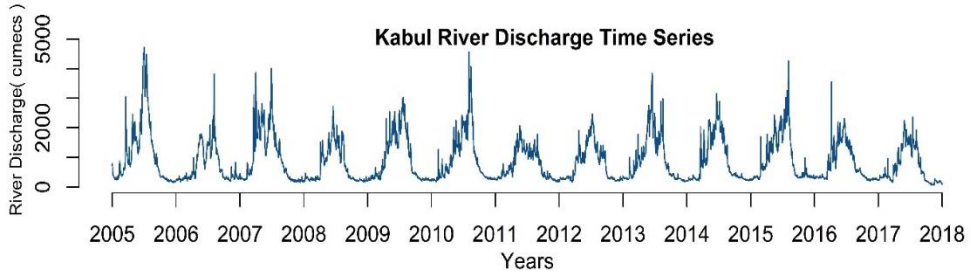


Fig. 4. Kabul River discharge series.

4.2. Performance evaluation measures

The prediction performance of the proposed hybrid framework is evaluated on both training and testing datasets. A 5-fold cross-validation approach and different goodness-of-fit measures are selected to assess the performance of models. These measures include root mean square error (***RMSE***), mean absolute percentage error (***MAPE***), root-relative square error (***RRSE***), mean absolute error (***MAE***) and coefficient of determination (R^2). These measures are given as follows (Zeinali et al. 2020; Shabbir et al. 2023):

$$RMSE = \sqrt{\frac{1}{n}\sum_{j=1}^{n}\left(y_j - \hat{y}_j\right)^2} \tag{12}$$

$$MAPE = \frac{100}{n}\sum_{j=1}^{n}\left|\frac{y_j - \hat{y}_j}{y_j}\right| \tag{13}$$

$$RRSE = \sqrt{\frac{\sum_{j=1}^{n}\left(y_j - \hat{y}_j\right)^2}{\sum_{j=1}^{n}\left(\hat{y}_j - \bar{\hat{y}}\right)^2}} \tag{14}$$

$$MAE = \frac{1}{n}\sum_{j=1}^{n}\left|y_j - \hat{y}_j\right| \tag{15}$$

$$R^2 = 1 - \left(\frac{\sum_{j=1}^{n}\left(y_j - \hat{y}_j\right)^2}{\sum_{j=1}^{n}\left(y_j - \bar{y}\right)^2}\right) \tag{16}$$

where n denotes the total number of observations, y_j denotes the actual observation and $\hat{y}_j$ denotes the predicted values. The terms $\bar{y}$ and $\bar{\hat{y}}$ denote the average of observed and predicted values, respectively.

To compare the performance of the different models for river discharge prediction, the improvement percentages of ***RMSE***, ***MAPE***, ***RRSE***, and ***MAE*** are also used and are given as:

$$P_{RMSE} = \frac{\left(RMSE_i - RMSE_j\right)}{RMSE_i} \times 100 \tag{17}$$

$$P_{MAPE} = \frac{\left(MAPE_i - MAPE_j\right)}{MAPE_i} \times 100 \tag{18}$$

$$P_{RRSE} = \frac{\left(RRSE_i - RRSE_j\right)}{RRSE_i} \times 100 \tag{19}$$

$$P_{MAE} = \frac{\left(MAE_i - MAE_j\right)}{MAE_i} \times 100 \tag{20}$$

$$P_{R^2} = \frac{\left(R_i^2 - R_j^2\right)}{R_i^2} \times 100 \tag{21}$$

where subscript i denotes the competing model and subscript j indicates the proposed LAES hybrid model. These quantities indicate the degree of improvement in the prediction performance of one model relative to another model (Duan et al. 2021).

The Diebold-Mariano (DM) test has been widely used in literature to compare the forecast accuracy of two models (Silva et al. 2021; Shabbir et al. 2022). The null and alternative hypotheses are:

$$H_0: E[d_t] \geq 0 \quad (22)$$

$$H_1: E[d_t] < 0$$

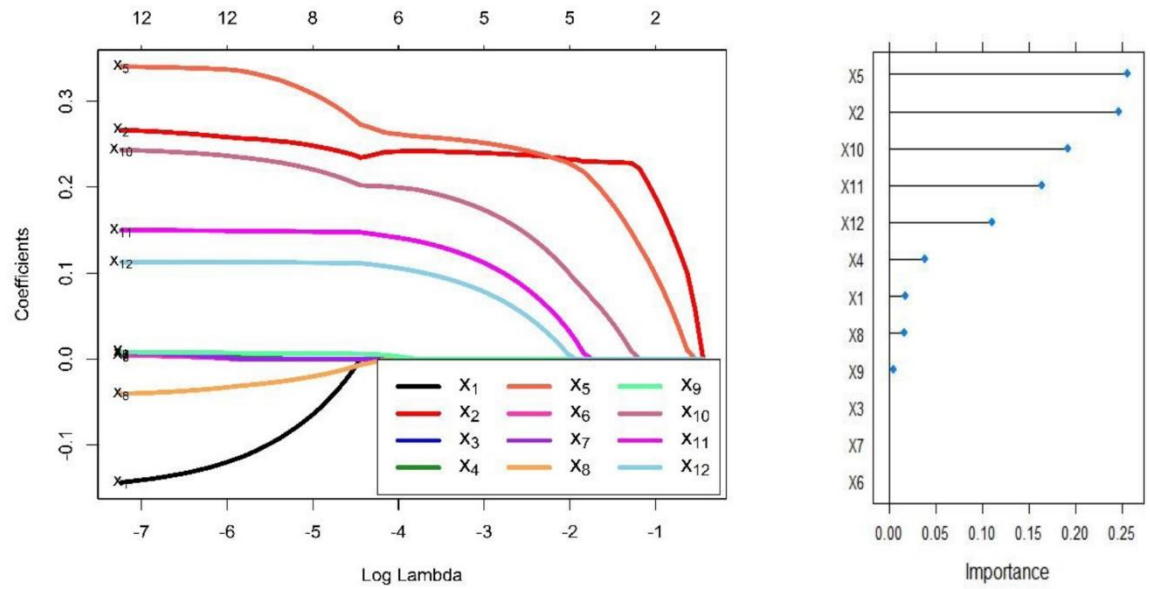
where d_t is the difference loss function, i.e., $d_t = e_{1t} - e_{2t}$, e_{1t} and e_{2t} denotes the set of prediction errors of two competing models. The test statistic is $DM = \frac{\bar{d}}{\left(2\pi \hat{f}_d(0)/m\right)^{1/2}}$, where m is the length of prediction errors, $\bar{d} = \frac{1}{m}\sum_{t=1}^{m}(d_t)$ is the average loss differential between two predictions, The DM statistic follows the standard normal distribution (i.e. $N(0,1)$) and $\hat{f}_d(0)$ is the spectral density. The $2\pi\hat{f}_d(0)$ is the consistent estimator of the asymptotic variance. The null hypothesis (H_0) is rejected if $DM < -Z_\alpha$, where Z is the standardized normal percentile with probability α.

In this study, a one-sided DM test is used to compare the prediction accuracy of the LAES model with six models. This test uses subscript 1 for the proposed LAES model and subscript 2 for the competing models. This test is applied using the squared differences loss function to compare models at a 1% significance level. If $DM < -2.326$, we will reject the null hypothesis. The proposed LAES hybrid model is compared with MLR, SVM, ANN, LASSO-MLR, LASSO-SVM and LASSO-ANN models in this study.

5. Results and discussion

In the proposed hybrid framework, LASSO is employed to choose meteorological variables that have significant roles in predicting Kabul River discharge. This step eliminates insignificant variables and constructs a better prediction model. Using LASSO, we retain only important input variables that influence the river discharge of the Kabul River. The results of the LASSO using $\lambda = 0.010$ are shown in Figure 5a. LASSO eliminates three meteorological variables, i.e., pan water (maximum), relative humidity (8 AM) and relative humidity (5 PM). The air temperature (minimum and maximum), dew point (8 AM), relative humidity (5 PM), rainfall, precipitation, wind speed, and evapotranspiration are significant variables for prediction of river discharge. These variables. $\{x_1, x_2, x_4, x_5, x_8, x_9, x_{10}, x_{11}, x_{12}\}$ are used as inputs to LASSO-based models. Bui et al. (2019) stated that dew point is a component of the temperature variable. The precipitation and rainfall factors are dependent on the air temperature and are indirectly associated with the dew point.

Figure 5b shows that dew point (8 AM) is the most significant variable for predicting river discharge. These variables selected by LASSO are used as inputs to the ANN model in the proposed hybrid framework. The prediction results by LASSO-ANN in the first round of the training phase are demonstrated in Figure 6a. The results of the remaining rounds are given in supplementary materials.



a) b)

Fig. 5. The variable screening (a) and variable importance (b) results from LASSO on Kabul River data.

After ANN model training, the predictions and error series are obtained. Stationarity of the error series is checked using an augmented Duckey-Fuller (ADF) test. The Dickey-Fuller statistic is –3.3875, indicating that the error series in the first round is non-stationary at the 5% level of significance. The results of ADF tests of the remaining rounds are provided in the supplementary materials. Next, the EMD decomposes the error series into $IMFs$ and residuals as shown in Figure 6b. Then, the SVM is applied to model each component of the decomposed error series. The sub-series predictions are obtained and aggregated as the final error prediction shown in Figure 6c. The final prediction of river discharge is computed by adding the predicted errors and predicted river discharge. Lastly, the actual predicted values of river discharge are obtained by anti-normalization using Eq. 11. Figure 7 shows the predicted discharge plot in the testing phase in the first round. It reveals that the proposed LAES hybrid models have the closest predictions to the observed river discharge.

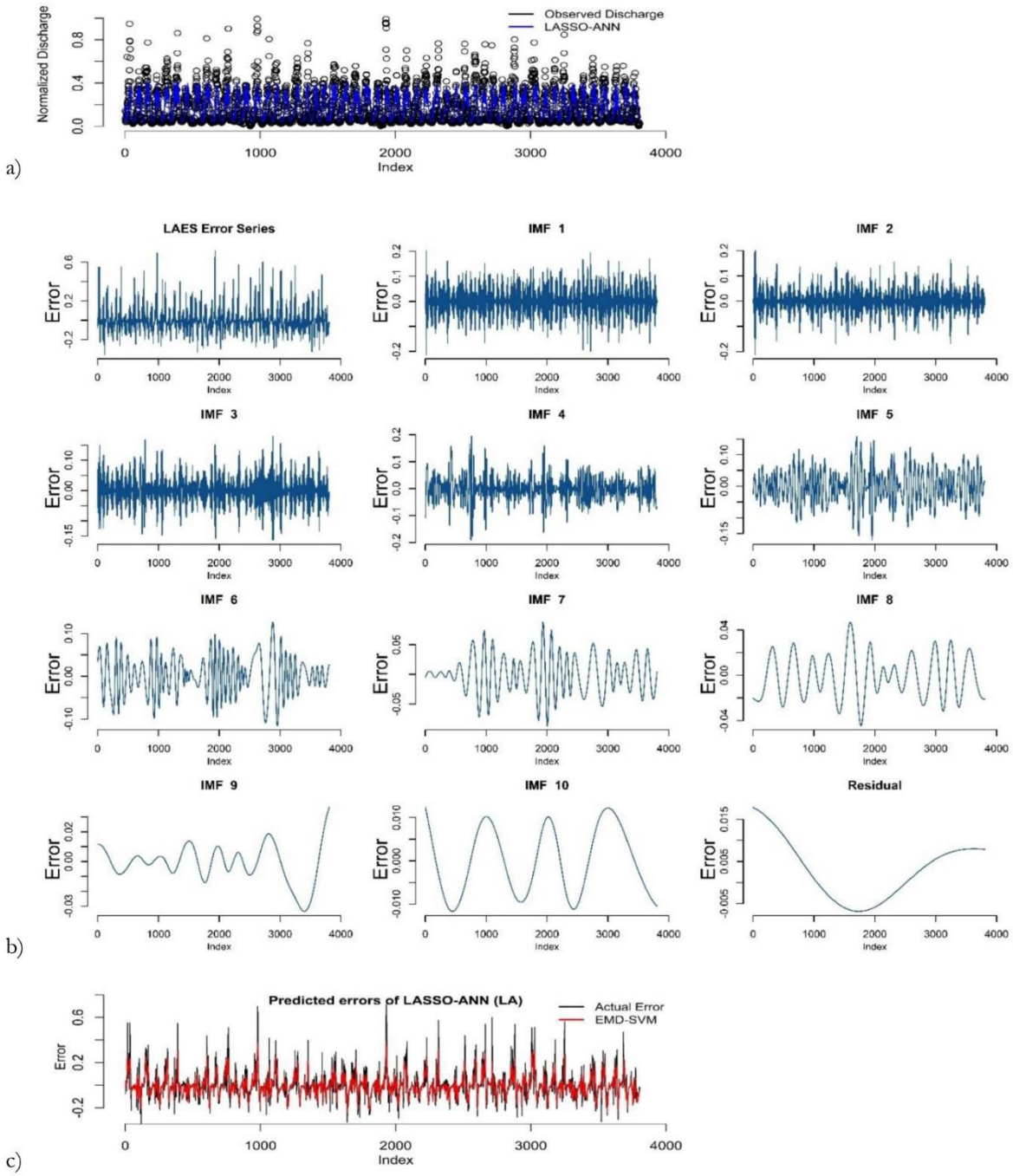


Fig. 6. Prediction results of LASSO-ANN: (a) error decomposition using EMD; (b) modeling of decomposed components (c) in the first round of training the phase for the Kabul River.

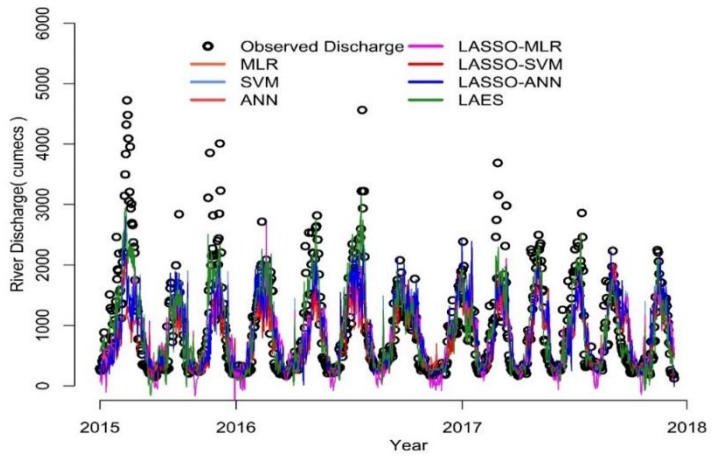


Fig. 7. Prediction plot of Kabul River discharge on test data of first round.

5.1. Comparison of model accuracy

The daily river discharge was estimated against various meteorological variables. Table 2 presents the training and testing phase results of daily river discharge prediction. In the training phase, the MLR model is the worst performer among all models (*RMSE* = 533.822 m^3/s, *MAE* = 378.003 m^3/s, *RRSE* = 0.711, *MAPE* = 66.786% and R^2 = 49.4%). However, the SVM and ANN models performed relatively better than the MLR model. For example, in the training phase, the *RMSE* for MLR, SVM and ANN models is 533.822 m^3/s, 511.262 m^3/s and 507.015 m^3/s, respectively. Similar to this study, Zhang et al. (2018b) found that the MLR model is the worst performer for predicting river discharge in the East River basin of China. Some other studies found that the non-linear features of river discharge are captured well by SVM and ANN models (see Poul et al. 2019 and Meng et al. 2021).

Comparing the performance of models based on meteorological variables selected by LASSO, we found that the performance of all models is improved in most of the instances. The performance of the LASSO-MLR model is better than the MLR model in the testing phase (*RMSE* = 543.559 m^3/s, *MAE* = 381.889 m^3/s, *RRSE* = 0.725, *MAPE* = 67.758% and R^2 = 47.4%). However contrary results are obtained in the training phase, in which the LASSO-MLR model has a similar fit to the MLR model. The prediction ability of LASSO-ANN and LASSO-SVM is better than ANN and SVM models respectively. Mehr and Gandomi (2021) found that LASSO improved the predictive ability of a multi-stage genetic programming model by reducing the number of genes for predicting river discharge in the Sedre River of Turkey. In the training phase, the proposed LAES hybrid model has the best fit for river discharge data based on various performance criteria (*RMSE* = 302.952 m^3/s, *MAE* = 201.022 m^3/s, *RRSE* = 0.404, *MAPE* = 30.494% and R^2 = 83.7%).

Comparing the results in the testing phase, the MLR model has the poorest performance when all the meteorological variables were used as inputs (*RMSE* = 554.277 m^3/s, *MAE* = 383.541 m^3/s, *RRSE* = 0.739, *MAPE* = 68.134% and R^2 = 45.3%). The use of LASSO for dimension reduction enhanced the performance of MLR, SVM, and ANN models in the testing phase. Judging by *RMSE*, *RRSE* and R^2, the

LASSO-ANN model is a better performer than the LASSO-SVM and LASSO-MLR models. However, comparing *MAE* and *MAPE*, the LASSO-SVM model performs better than the LASSO-MLR and LASSO-ANN hybrid models (*MAE* = 307.124 m^3/s and *MAPE* = 39.394%). The proposed LAES model outperforms all competing models in the testing phase (i.e., *RMSE* = 337.143 m^3/s, *MAE* = 218.353 m^3/s, *RRSE* = 0.449, *MAPE* = 32.354% and R^2 = 79.8%). Overall, the proposed LAES hybrid model has higher prediction accuracy than single and LASSO-based ANN, SVM, and MLR models.

Figure 8a presents the goodness-of-fit measure values of all the models considered in the study in both training and testing data. It shows that the proposed LAES hybrid model has the highest accuracy among all models considered in the study. The Taylor diagram in Figure 8b shows that the proposed LAES model is the most efficient among all models considered in predicting daily river discharge based on its relationship with meteorological variables.

Table 2. Performance analysis of the proposed model with different models.

Models	*RMSE* (m^3/s)	*MAE* (m^3/s)	*RRSE*	*MAPE* (%)	R^2
	Training				
MLR	533.822	378.003	0.711	66.786	0.494
SVM	511.262	309.783	0.681	39.372	0.536
ANN	507.015	334.263	0.676	50.508	0.542
LASSO-MLR	534.091	378.263	0.712	66.878	0.494
LASSO-SVM	469.381	280.664	0.625	35.972	0.609
LASSO-ANN	456.981	302.596	0.609	45.686	0.629
LAES	**302.952**	**201.022**	**0.404**	**30.494**	**0.837**
	Testing				
MLR	554.277	383.541	0.739	68.134	0.453
SVM	527.427	324.443	0.702	41.814	0.505
ANN	524.117	342.108	0.699	51.618	0.511
LASSO-MLR	543.559	381.889	0.725	67.758	0.474
LASSO-SVM	499.947	307.124	0.666	39.394	0.556
LASSO-ANN	497.256	324.178	0.664	48.056	0.559
LAES	**337.143**	**218.353**	**0.449**	**32.354**	**0.798**
Note: Bold values represent minimum values in each column					

The improvements of the proposed LAES hybrid model are shown in Table 3 in terms of P_{RMSE}, P_{MAE}, P_{RRSE}, P_{MAPE} and P_{R^2} for both training and testing phases. The proposed LAES hybrid model has 43.3%, 40.7% and 40.3% lower *RMSE* than the MLR, SVM, and ANN models, respectively, in the training phase. The findings indicate that the MLR model is least efficient for non-linear data, consistent with the findings of Zhang et al. (2018b).

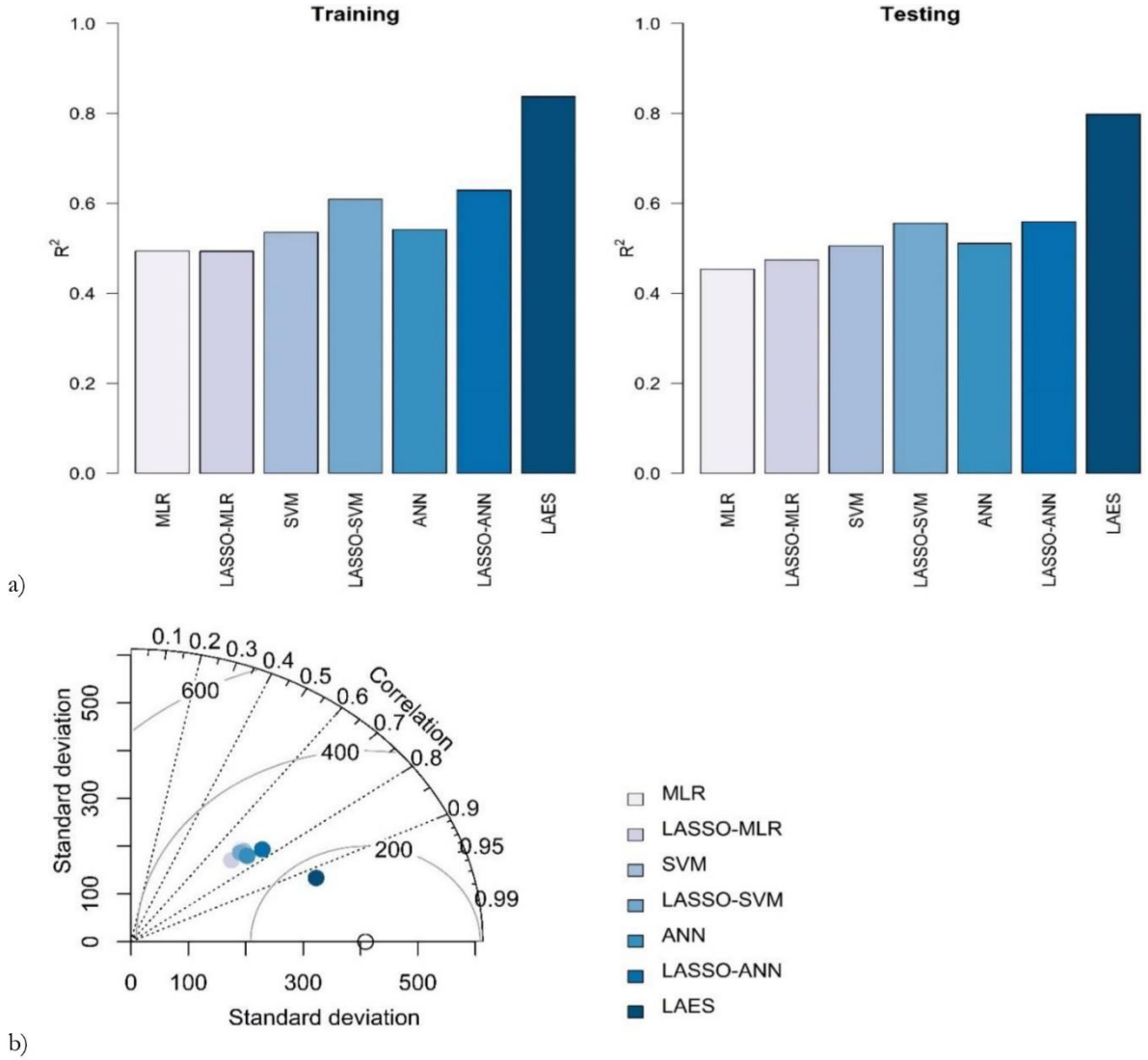


Fig. 8. Prediction results of models in training and testing phase (a) and Taylor diagram (b) for the Kabul River.

Comparing the LAES model to LASSO-based models, we found that their promoting improvements were lower compared to single MLR, SVM, and ANN models in the majority of the scenarios. During testing, the reduction in *RMSE* by LASSO-MLR and MLR models is 38% and 39.2%, respectively. Similarly, the improvements by the LAES model vs. the SVM model (36.1%) are higher than the LAES model vs. the SVM model (32.6%). The proposed LAES hybrid model has 68.2%, 43.6%, and 42.7% better prediction accuracy than the LASSO-MLR, LASSO-SVM, and LASSO-ANN models. Kang et al. (2023) also stated that LASSO helps enhance the predictive performance of monthly run-off, which is influenced by meteorological events.

Generally, the proposed LAES hybrid model has promising predictions compared to all six models. During the training phase, the MAE of LAES compared to MLR, SVM, ANN, LASSO-MLR, LASSO-SVM, and LASSO-ANN decreased by 46.8%, 35.1%, 39.9%, 46.9%, 28.4%, and 33.6% respectively. These results are in agreement with the findings of Duan et al. (2021). They reported that the decomposition-based error correction approach significantly improves the accuracy of models.

Table 3. Improved percentage (%) of proposed model versus other models.

Models	Training					Testing				
	P_{RMSE}	P_{MAE}	P_{RRSE}	P_{MAPE}	P_{R^2}	P_{RMSE}	P_{MAE}	P_{RRSE}	P_{MAPE}	P_{R^2}
LAES vs. MLR	43.3	46.8	43.3	54.3	–69.4	39.2	43.1	39.2	52.5	–76.1
LAES vs. SVM	40.7	35.1	40.8	22.6	–56.2	36.1	32.7	36.0	22.6	–57.9
LAES vs. ANN	40.3	39.9	40.3	39.6	–54.5	35.7	36.2	35.7	37.3	–56.1
LAES vs. LASSO-MLR	43.3	46.9	43.3	54.4	–69.6	38.0	42.8	38.0	52.3	–68.2
LAES vs. LASSO-SVM	35.5	28.4	35.5	15.2	–37.5	32.6	28.9	32.6	17.9	–43.6
LAES vs. LASSO-ANN	33.7	33.6	33.7	33.3	–33.0	32.2	32.6	32.3	32.7	–42.7

The DM test results on the testing data of Kabul River discharge are given in Table 4. The null hypothesis for all competing models is rejected at a 1% significance level. Thus, the prediction accuracy of the proposed hybrid LAES model is higher than the six benchmark models. Therefore, the DM test confirms that the proposed LAES hybrid model has higher prediction accuracy than the competing models in predicting river discharge.

Table 4. DM test of proposed hybrid model versus different models on the testing dataset.

Model	MLR	SVM	ANN	LASSO-MLR	LASSO-SVM	LASSO-ANN
DM-value	–9.118***	–8.688***	–10.256***	–10.702***	–8.434***	–8.299***
*** significant at a 1% significance level						

6. Conclusion

In this study, a new hybrid framework named LAES (LASSO-ANN-EMD-SVM) is introduced for modeling river discharge using information from several meteorological variables. The proposed hybrid model is a composite of a variable selection approach with an artificial neural network and error correction method. The application of the LAES hybrid framework is illustrated using the data from the Kabul River in Pakistan. The effectiveness and predictive ability of the proposed framework are compared with six models using different performance measures. The numerical findings reveal that the LAES hybrid model has better prediction performance than the single and LASSO-based MLR, SVM, and ANN models. Judging by *RRSE*, the LAES hybrid model has 43.3%, 40.8%, 40.3%, 43.3%, 35.5%, and 33.7% lower prediction errors than MLR, SVM, ANN, LASSO-MLR, LASSO-SVM and LASSO-ANN models respectively. The Diebold-Mariano test shows that the proposed LAES model has higher prediction accuracy than all competing models in the study. The proposed LAES model can serve as a successful tool for river discharge prediction by considering the impact of meteorological variables. In this study, we have used the LAES hybrid model for regression modeling only, but it can be applied for time series prediction of hydrological variables (such as river inflow and monthly run-off). For future research, new hybrid models can be developed by considering (i) relevance vector machine (RVM) or deep learning models such as multilayer perceptron (MLP) in modeling; and (iii) using decomposition techniques such as ensemble EMD, complete EEMD (CEEMD), and variational mode decomposition (VMD) methods in the error correction stage. The proposed LAES model can serve as a successful tool for river discharge prediction of catchment areas of different areas of the world for efficient planning of water resources.

Acknowledgments

We acknowledge the SWHP department of WAPDA, Pakistan, for providing the data required for this research work.

References

Adnan R.M., Mostafa R.R., Elbeltagi A., Yaseen Z.M., Shahid S., Kisi O., 2022, Development of new machine learning model for streamflow prediction: case studies in Pakistan, Stochastic Environmental Research and Risk Assessment, 36, 999-1033, DOI: 10.1007/s00477-021-02111-z.

Ali S., Shahbaz M., 2020, Streamflow forecasting by modeling the rainfall–streamflow relationship using artificial neural networks, Modeling Earth Systems and Environment, 6, 1645-1656, DOI: 10.1007/s40808-020-00780-3.

Baydaroğlu Ö., Koçak K., Duran K., 2018, River flow prediction using hybrid models of support vector regression with the wavelet transform, singular spectrum analysis and chaotic approach, Meteorology and Atmospheric Physics, 130, 349-359, DOI: 10.1007/s00703-017-0518-9.

Bui A., Johnson F., Wasko C., 2019, The relationship of atmospheric air temperature and dew point temperature to extreme rainfall, Environmental Research Letters, 14 (7), DOI: 10.1088/1748-9326/ab2a26.

Cheng K., Wei S., Fu Q., Li T., 2019, Adaptive management of water resources based on an advanced entropy method to quantify agent information, Journal of Hydroinformatics, 21 (3), 381-396, DOI: 10.2166/hydro.2019.007.

Dariane A., Azimi S., 2018, Streamflow forecasting by combining neural networks and fuzzy models using advanced methods of input variable selection, Journal of Hydroinformatics, 20 (2), 520-532. DOI: 10.2166/hydro.2017.076.

Duan J., Zuo H., Bai Y., Duan J., Chang M., Chen B., 2021, Short-term wind speed forecasting using recurrent neural networks with error correction, Energy, 217, DOI: 10.1016/j.energy.2020.119397.

Elbeltagi A., Nunno F.D., Kushwaha N.L., Marinis G.D., Granata F., 2022, River flow rate prediction in the Des Moines watershed (Iowa, USA): a machine learning approach, Stochastic Environmental Research and Risk Assessment, 36, 3835-3855, DOI: 10.1007/s00477-022-02228-9.

Emadi A., Sobhani R., Ahmadi H., Boroomandnia A., Zamanzad-Ghavidel S., Azamathulla H.M., 2022, Multivariate modeling of river water withdrawal using a hybrid evolutionary data-driven method, Water Supply, 22 (1), 957-980, DOI: 10.2166/ws.2021.224.

Gleason C.J., Smith L.C., Lee J., 2014, Retrieval of river discharge solely from satellite imagery and at-many-stations hydraulic geometry: Sensitivity to river form and optimization parameters, Water Resources Research, 50 (12), 9604-9619, DOI: 10.1002/2014WR016109.

Hu J., Wu Y., Sun P., Zhao F., Sun K., Li T., Sivakumar B., Qiu L., Sun Y., Jin Z., 2022, Predicting long-term hydrological change caused by climate shifting in the 21st century in the headwater area of the Yellow River basin, Stochastic Environmental Research and Risk Assessment, 36, 1651-1668, DOI: 10.1007/s00477-021-02099-6.

Huang N.E., Shen Z., Long S.R., Wu M.C., Shih H.H., Zheng Q., Yen N.C., Tung C.C., Liu H.H., 1998, The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis. Proceedings of the Royal Society A. London, 454 (1971), 903-995, DOI: 10.1098/rspa.1998.0193.

Jungsheng C., Dejie Y., Yu Y., 2006, A fault diagnosis approach for roller bearings based on EMD method and AR model, Mechanical Systems Signal Processing, 20 (2), 350-362, DOI: 10.1016/j.ymssp.2004.11.002.

Kachrimanis K., Kamaryan V., Malamataris S., 2003, Artificial neural networks (ANNs) and modeling of powder flow, International Journal of Pharmaceutics, 250 (1), 13-23, DOI: 10.1016/S0378-5173(02)00528-8.

Kang Y., Cheng X., Chen P., Zhang S., Yang Q., 2023, Monthly runoff prediction by a multivariate hybrid model based on decomposition-normality and Lasso regression, Environmental Science and Pollution Research, 30, 27743-27762, DOI: 10.1007/s11356-022-23990-x.

Kisi O., Alizamir M., Shiri J., 2021, Conjunction model design for intermittent streamflow forecasts: extreme learning machine with discrete wavelet transform, [in:] Intelligent Data Analytics for Decision-Support Systems in Hazard Mitigation, Springer, Singapore, 171-181, DOI: 10.1007/978-981-15-5772-9_9.

Lei Y., He Z., Zi Y., 2003, Application of the EEMD method to rotor fault diagnosis of rotating machinery, Mechanical Systems and Signal Processing, 23 (4), 1327-1338, DOI: 10.1016/j.ymssp.2008.11.005.

Luo X., Xiaohui Y., Zhu S., Xu Z., Meng L., Peng J., 2019, A hybrid support vector regression framework for streamflow forecast, Journal of Hydrology, 568, 184-193, DOI: 10.1016/j.jhydrol.2018.10.064.

Malik A., Tikhamarine Y., Souag-Gamane D., Kisi O., Pham Q.B., 2020, Support vector regression optimized by meta-heuristic algorithms for daily streamflow prediction, Stochastic Environmental Research and Risk Assessment, 34, 1755-1773, DOI: 10.1007/s00477-020-01874-1.

Mangukiya N.K., Mehta D.J., Jariwala R., 2022, Flood frequency analysis and inundation mapping for lower Narmada basin, India, Water Practice and Technology, 17(2), 612-622, DOI: 10.2166/wpt.2022.009.

Mehmood A., Jia S., Lv A., Zhu W., Mehmood R., Saifullah M., Adnan M.R., 2021, Detection of spatial shift in flood regime of the Kabul river basin in Pakistan, causes, challenges, and opportunities, Water, 13 (9), 1296-1301, DOI: 10.3390/w13091276.

Mehr A.D., Gandomi A.H., 2021, MSGP-LASSO: An improved multi-stage genetic programming model for streamflow prediction, Information Sciences, 561, 181-195, DOI: 10.1016/j.ins.2021.02.011.

Mehta D.J., Eslamian S., Prajapati K., 2022, Flood modelling for a data-scare semi-arid region using 1-D hydrodynamic model: a case study of Navsari Region, Modeling Earth Systems and Environment, 8 (2), 2675-2685, DOI: 10.1007/s40808-021-01259-5.

Meng E., Huang S., Huang Q., Fang W., Wang H., Leng G., Wang L., Liang H., 2021, A hybrid VMD-SVM model for practical streamflow prediction using an innovative input selection framework, Water Resources Management, 35, 1321-1337, DOI: 10.1007/s11269-021-02786-7.

Poul A.K., Shourian M., Ebrahimi H., 2019, A comparative study of MLR, KNN, ANN and ANFIS models with wavelet transform in monthly stream flow prediction, Water Resources Management, 33, 2907-2923, DOI: 10.1007/s11269-019-02273-0.

R Core Team, 2022, R: A language and environment for statistical computing, R Foundation for Statistical Computing, Vienna, Austria, https://www.R-project.org/.

Rasouli H., 2022, Climate change impacts on water resource and air pollution in Kabul sub-basins, Afghanistan, Advances in Geological and Geotechnical Engineering Research, 4 (1), 11-27, DOI: 10.30564/agger.v4i1.4312.

Rasouli K., Hsieh W.W., Cannon A.J., 2012, Daily streamflow forecasting by machine learning methods with weather and climate inputs, Journal of Hydrology, 414-415, 284-293, DOI: 10.1016/j.jhydrol.2011.10.039.

Ravindran S.M., Bhaskaran S.K., Ambat S.K., 2021, A deep neural network architecture to model reference evapotranspiration using a single input meteorological parameter, Environmental Processes, 8, 1567-1599, DOI: 10.1007/s40710-021-00543-x.

Saidi H., Dresti C., Manca D., Ciampittiello M., 2018, Quantifying impacts of climate variability and human activities on the streamflow of an Alpine river, Environmental Earth Sciences, 77, DOI: 10.1007/s12665-018-7870-z.

Shabbir M., Chand S., Iqbal F., 2022, A Novel Hybrid Method for River Discharge Prediction. Water Resources Management, 36, 253-272. DOI: https://doi.org/10.1007/s11269-021-03026-8.

Shabbir M., Chand S., Iqbal F., 2023, Prediction of river inflow of the major tributaries of Indus river basin using hybrids of EEMD and LMD methods, Arabian Journal of Geosciences, 16, 257, DOI: 10.1007/s12517-023-11351-y.

Shabbir M., Chand S., Iqbal F., 2024, Novel hybrid and weighted ensemble models to predict river discharge series with outliers, Kuwait Journal of Science, 51 (2), DOI: 10.1016/j.kjs.2024.100188.

Silva R.G., Ribeiro M.H., Moreno S.R., Mariani V.C., Coelho L.D., 2021, A novel decomposition-ensemble learning framework for multi-step ahead wind energy forecasting, Energy, 216, DOI: 10.1016/j.energy.2020.119174.

Tibshirani R., 1996, Regression Shrinkage and Selection via the Lasso, Journal of Royal Statistical Society: Series B (Methodological), 58 (1), 267-288, DOI: 10.1111/j.2517-6161.1996.tb02080.x.

Vapnik V., 1995, The nature of statistical learning theory, Springer, New York, 188 pp., DOI: 10.1007/978-1-4757-2440-0.

Wang J., Li Y., 2018, Multi-step ahead wind speed prediction based on optimal feature extraction, long short term memory neural network and error correction strategy, Applied Energy, 230, 429-443, DOI: 10.1016/j.apenergy.2018.08.114.

Yan S., Wang X., Zhang Y., Liu D., Yi Y., Li C., Liu Q., Yang Z., 2020, A hybrid PCA-GAM model for investigating the spatiotemporal impacts of water level fluctuations on the diversity of benthic macroinvertebrates in Baiyangdian Lake, North China, Ecological Indicators, 116, DOI: 10.1016/j.ecolind.2020.106459.

Yousuf I., Ghumman A.R., Hashmi H.N., 2017, Optimally sizing small hydropower project under future projected flows, KSCE Journal of Civil Engineering, 21, 1964-1978, DOI: 10.1007/s12205-016-1043-y.

Zeinali M., Azari A., Heidari M., 2020, Multiobjective optimization for water resource management in low-flow areas based on a coupled surface water–groundwater model, Journal of Water Resources Planning and Management, 146 (5), DOI: 10.1061/(ASCE)WR.1943-5452.0001189.

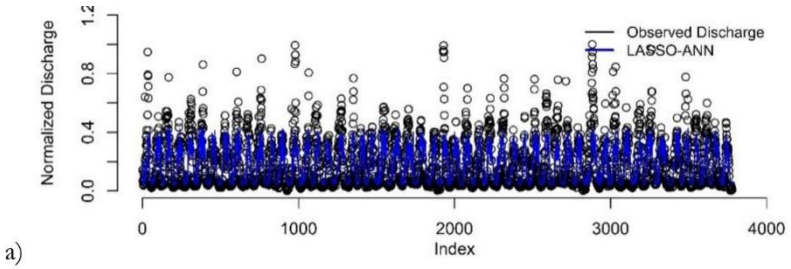
Zhang X., Liu P., Zhao Y., Deng C., Li Z., Xiong M., 2018a, Error correction-based forecasting of reservoir water levels: Improving accuracy over multiple lead times, Environmental Modelling and Software, 104, 27-39, DOI: 10.1016/j.envsoft.2018.02.017.

Zhang Z., Zhang Q., Singh V.P., 2018b, Univariate streamflow forecasting using commonly used data-driven models: Literature review and case study, Hydrological Sciences Journal, 63 (7), 1091-1111, DOI: 10.1080/02626667.2018.1469756.

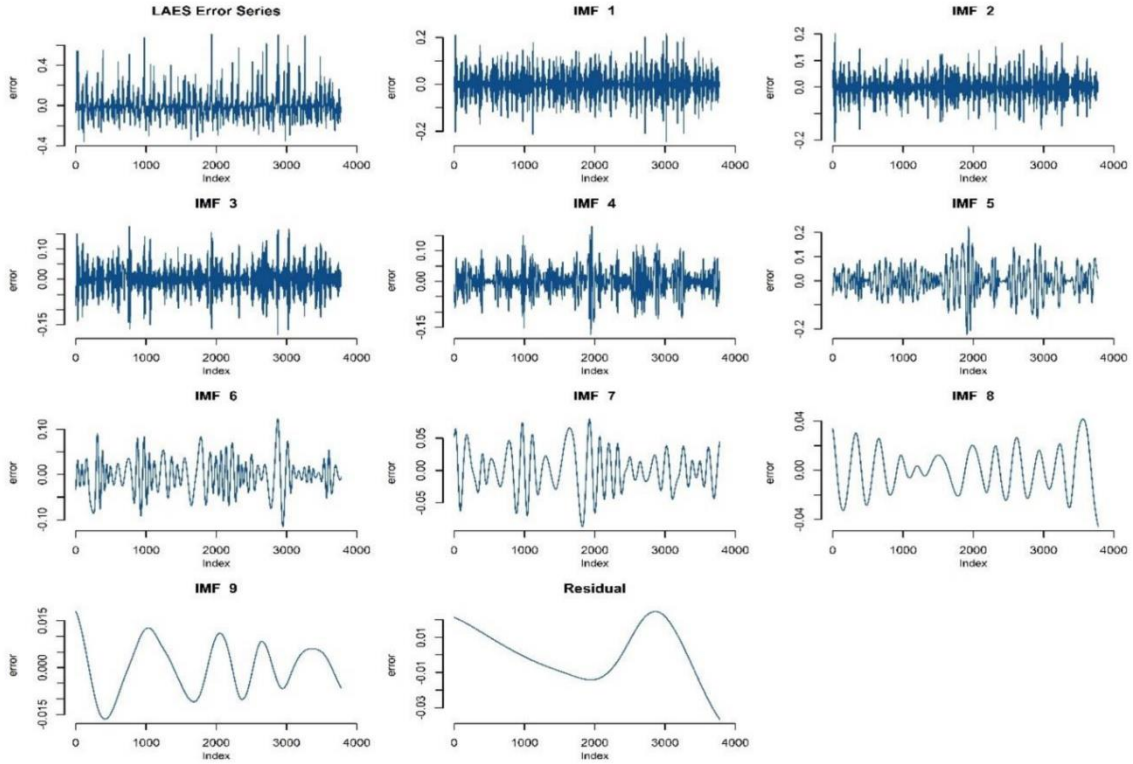
Appendix

ADF test results. H_0 – the time series contains unit root and is non-stationary; H_1 – the time series is stationary.

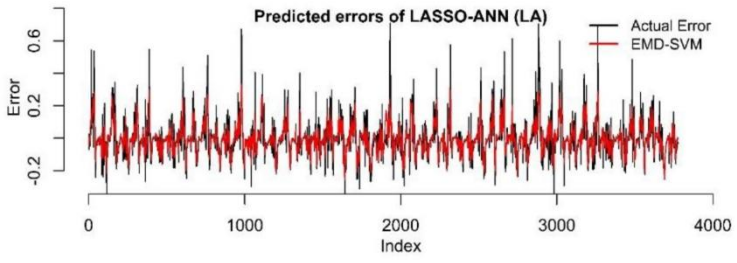
Fold	2	3	4	5
Duckey-Fuller Statistic	–3.2011*	–3.1671*	–3.0869*	–3.3423*
p-value	0.08769	0.0935	0.1327	0.06334



a)

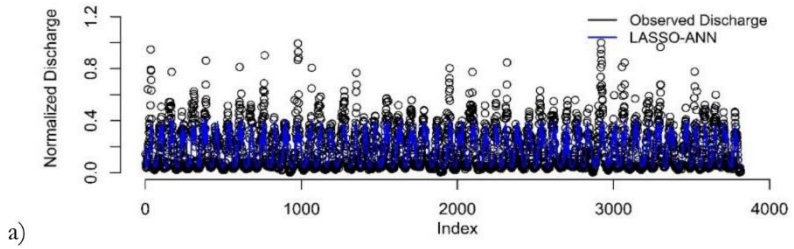


b)

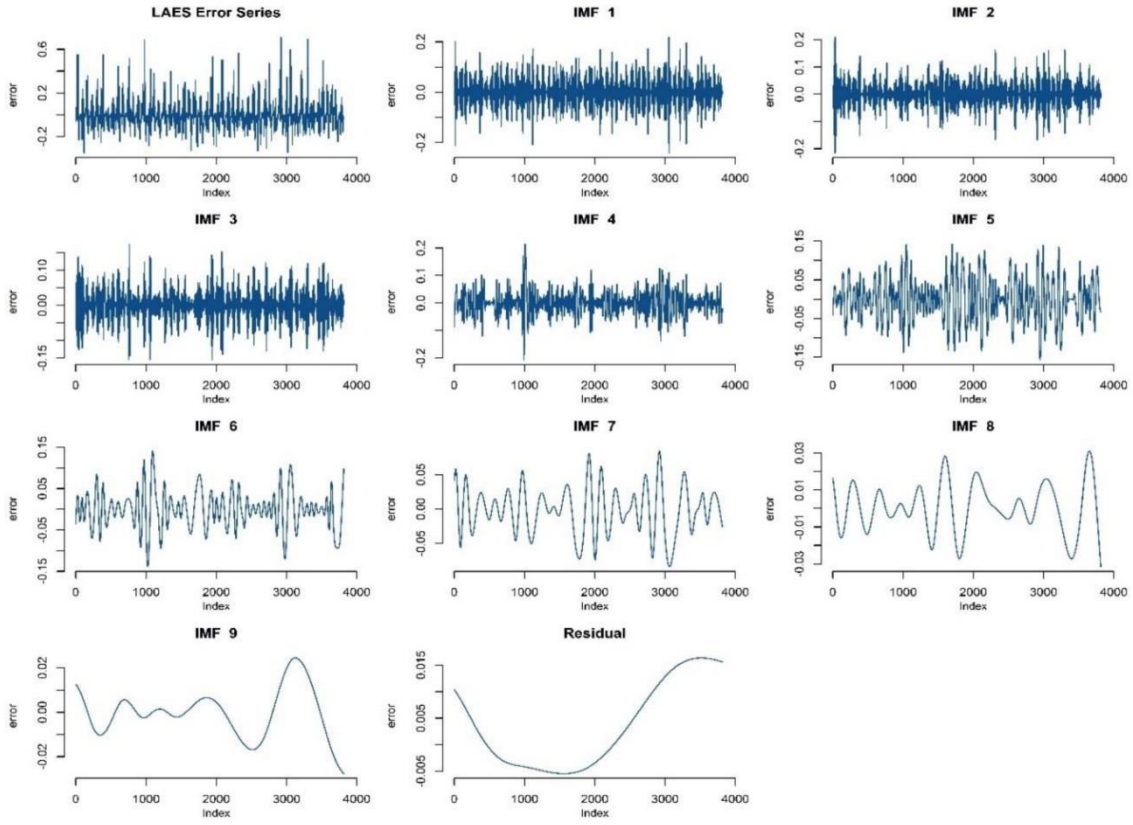


c)

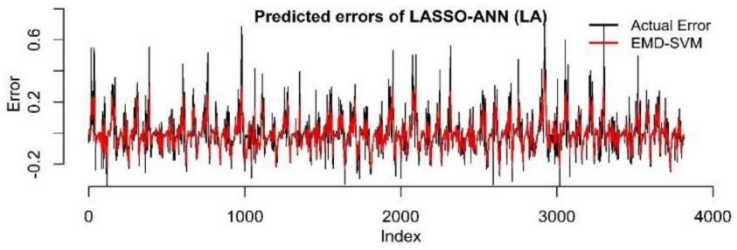
Fig. S1. Prediction results of LASSO-ANN (a) Error decomposition using EMD (b), modeling of decomposed components (c) in the second fold of training phase of Kabul River.



a)

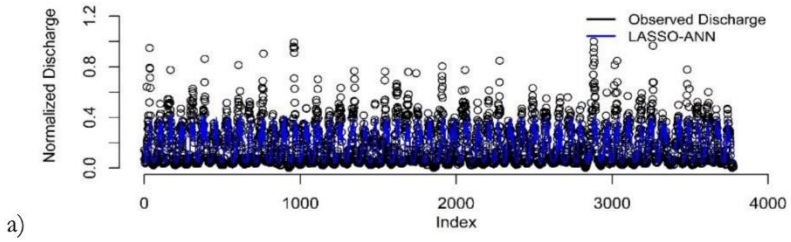


b)

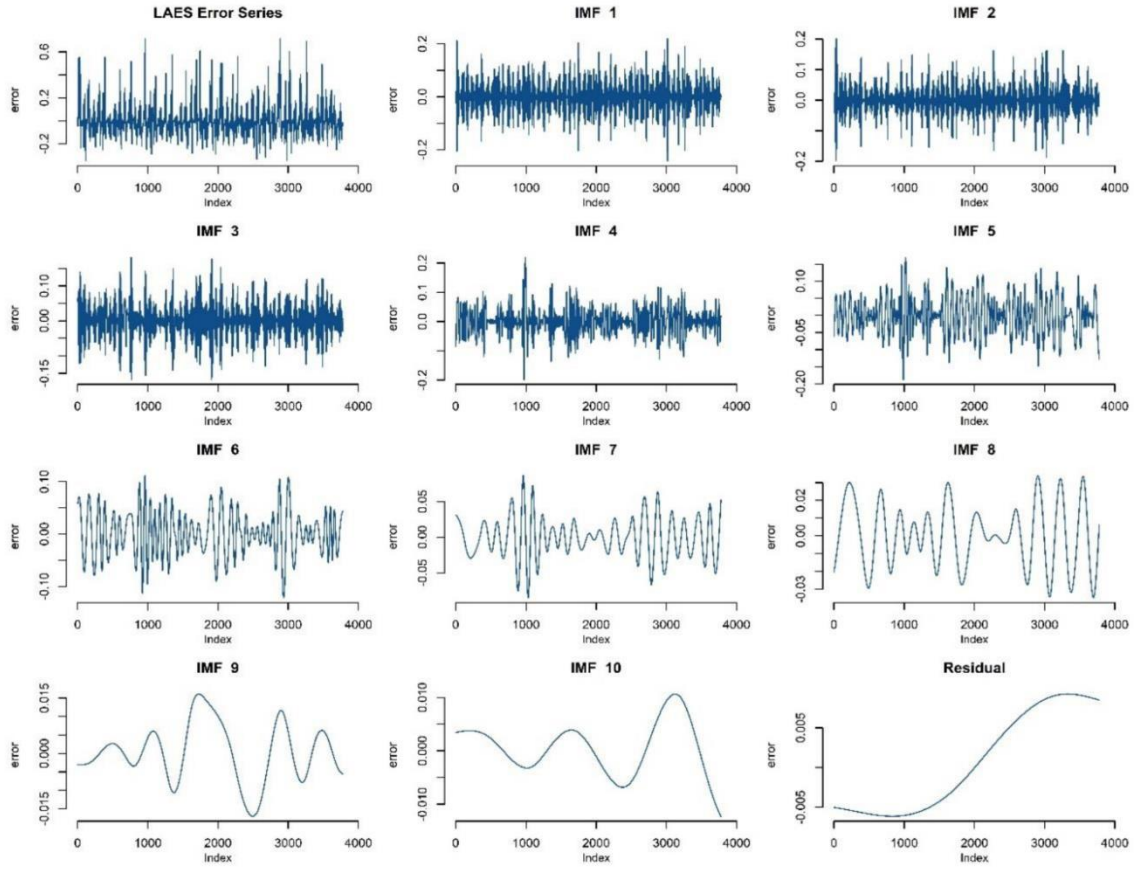


c)

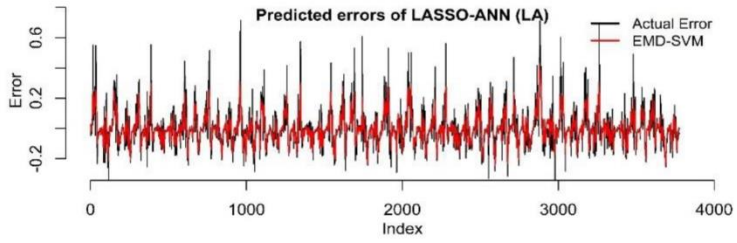
Fig. S2. Prediction results of LASSO-ANN (a) Error decomposition using EMD (b), modeling of decomposed components (c) in the third fold of training phase of Kabul River.



a)

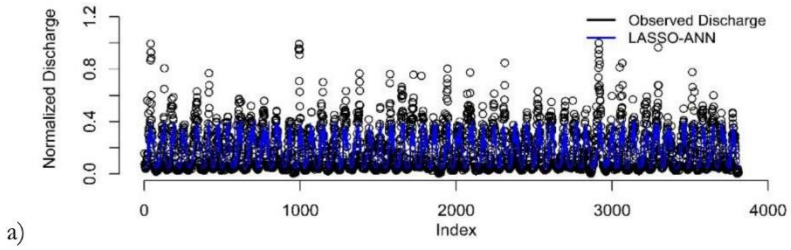


b)

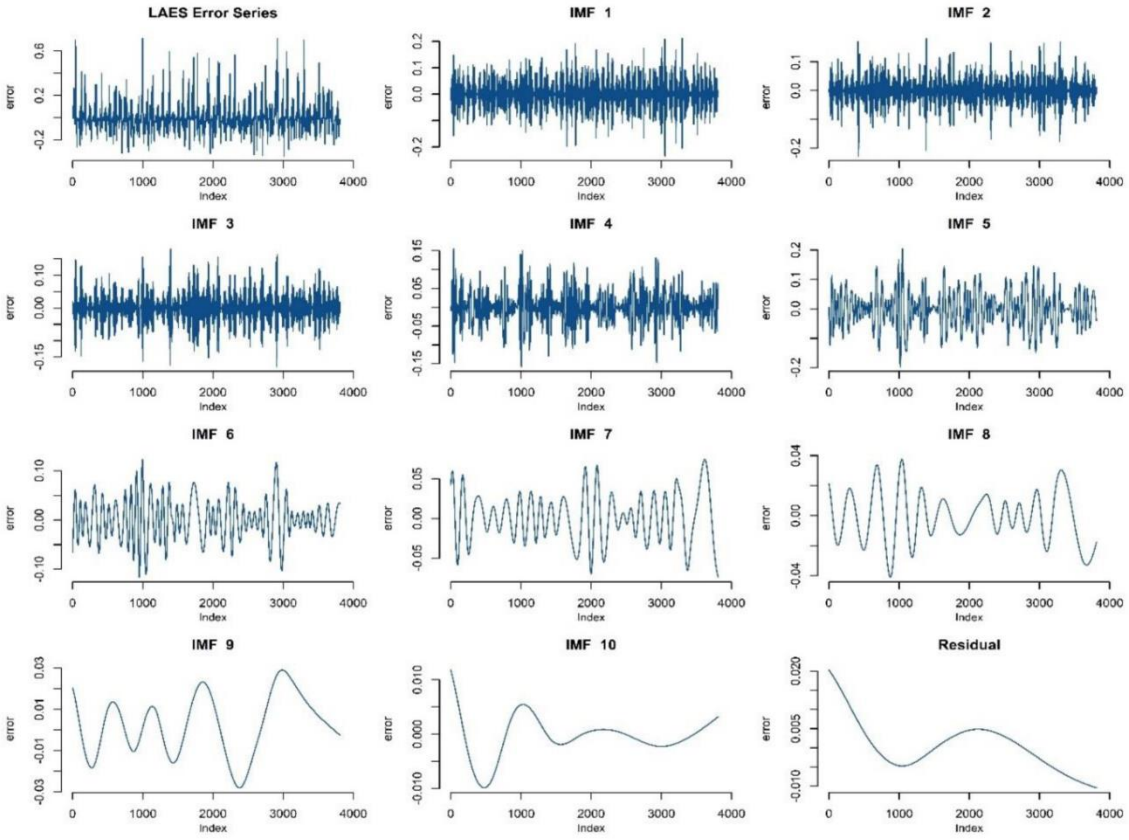


c)

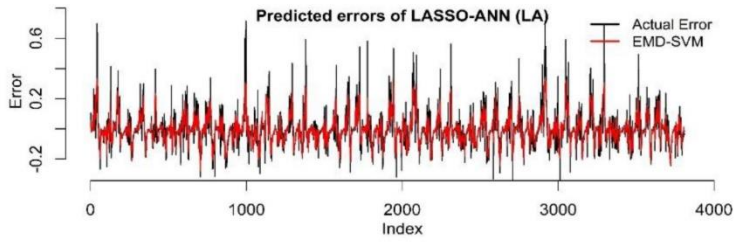
Fig. S3. Prediction results of LASSO-ANN (a) Error decomposition using EMD (b), modeling of decomposed components (c) in the fourth fold of training phase of Kabul River.



a)



b)



c)

Fig. S4. Prediction results of LASSO-ANN (a) Error decomposition using EMD (b), modeling of decomposed components (c) in the fifth fold of training phase of Kabul River.

Reviewers cooperating with Editorial Board of Meteorology Hydrology and Water Management Magazine in 2023

Katarzyna Baran-Gurgul

Wojciech Ciężkowski

Nelson Cordero

Bartosz Czernecki

Adam Froń

Wiesław Gądek

Tapas Karmaker

Krzysztof Kochanek

Tomasz Kolerski

Kacper Kotulak

Maciej Kryza

Ewa Szalińska van Overdijk

Bogusław Pawłowski

María-Teresa Sebastiá-Frasquet

Amin Shaban

Tamara Tokarczyk

Edmund Tomaszewski

Vazha Trapaidze

Paweł Wilk

THANK YOU

Reviewers cooperating with Editorial Board of
Meteorology Hydrology and Water Management Magazine in 2023

Katarzyna Baran-Gurgul

Wojciech Ciężkowski

Nelson Cordero

Bartosz Czernecki

Adam Froń

Wiesław Gądek

Tapas Karmaker

Krzysztof Kochanek

Tomasz Kolerski

Kacper Kotulak

Maciej Kryza

Ewa Szalińska van Overdijk

Bogusław Pawłowski

María-Teresa Sebastiá-Frasquet

Amin Shaban

Tamara Tokarczyk

Edmund Tomaszewski

Vazha Trapaidze

Paweł Wilk

THANK YOU

METEOROLOGY HYDROLOGY AND WATER MANAGEMENT

IMGW-PIB Science Journal since 2013