Satellite-based prediction of surface dust mass concentration in southeastern Iran using gradient boosting regression

Seyed Babak Haji Seyed Asadollah[1], Ahmad Sharafati[1]*, Davide Motta [2], Antonio Jodar-Abellan[3,4], Miguel Ángel Pardo[4,5]

[1] Department of Civil Engineering, Science and Research Branch, Islamic Azad University, Tehran, Iran.

[2] Department of Mechanical and Construction Engineering, Northumbria University, Wynne Jones Building, Newcastle upon Tyne NE1 8ST, United Kingdom

[3]Departamento de Análisis Geográfico Regional y Geografía Física, Facultad de Filosofía y Letras, Campus Universitario de Cartuja, University of Granada, 18071 Granada, Spain

[4]University institute of Water and Environmental Sciences. University of Alicante, Spain

[5]Department of Civil Engineering. University of Alicante, Spain

Corresponding authors: Ahmad Sharafati

Corresponding email: asharafati@srbiau.ac.ir, asharafati@gmail.com

Abstract

The southeastern section of Iran, especially the province of Khuzestan, experience severe air pollution levels, such as high values of Surface Dust Mass Concentration (SDMC). The province lacks accurate and well-placed ground observational stations, therefore the only viable approach for evaluating SDMC is via remote sensing. In this study, meteorological, hydrological and geological data on 11 input variables from Modern-Era Retrospective analysis for Research and Applications Version (MERRA-2), Global Precipitation Measurement (GPM) and Global Land Data Assimilation System (GLDAS) for the year 2018 are used for prediction of the SDMC values, also obtained from MERRA-2. For real-time prediction, Pearson's Correlation Coefficient (PCC) analysis shows that wind-related variables – surface wind speed, surface aerodynamic conductivity and surface pressure – are those with the highest correlation with SDMC. Using the Gradient Boosting Regression (GBR) algorithm, these three variables can simulate SDMC with good accuracy ($CC = 0.815\ ,\ N-RMSE = 0.605$). Future forecasting of SDMC requires knowledge of both wind-related and heat-related variables. However, SDMC predictions can be obtained with the GBR algorithm with adequate accuracy ($CC = 0.640\ .NRMSE = 0.781$) by just considering the surface pressure value observed four days before the forecasted day. This study shows that robust predictions of SDMC can be obtained using exclusively remote sensing data, without ground-based observations.

Keywords: Surface dust mass concentration, Gradient Boosting Regression, Prediction, Khuzestan

1. Introduction

Recent reports by the World Health Organization (WHO) revealed that nearly seven million people die in the world due to high levels of air pollution every year (WHO, 2014). Major air pollutants are carbon monoxide, sulfur, nitrogen dioxide, and surface-level ozone (Chen *et al.*, 2007, Council, 1992, Duan *et al.*, 2019, Sunyer *et al.*, 2003). Particulate Matter (PM) is also highlighted by WHO as the thirteenth mortality cause around the world, which makes it further hazardous (Anderson *et al.*, 2012). PM is a mixture of microscopic particles and liquid substances such as metals, organic materials, acids and dust (Planning, 1996, USEPA, 2019). PM can cause significant health problems, such as blood pressure, lung cancer, and cardiovascular diseases (Brook & Rajagopalan, 2009, Hamanaka & Mutlu, 2018, Raaschou-Nielsen *et al.*, 2016), therefore monitoring its concentration is critical for health and environmental purposes.

Numerous studies have been conducted on PM concentration measurement, using various sampling approaches and instruments (Amaral *et al.*, 2015, Kwasny *et al.*, 2010, Nakata *et al.*, 2013). Ground-based measurement, mostly done with monitoring stations, is the most commonly employed method. However, while the direct sampling approach is considered as the most accurate, it is neither cost- nor time-efficient, and station numbers are typically limited and station spatial distribution in generally irregular (M. Lee *et al.*, 2016, D. Liu & Li, 2015).

Recently, to overcome the disadvantages of direct sampling, Remote Sensing (RS) methods have become increasingly utilized for forecasting of hydrological and meteorological phenomena (Asadollah *et al.*, 2021, Ghozat *et al.*, 2022, Shiru *et al.*, 2022). RS covers a wide range of spatial and temporal data observations that can particularly benefit regions that lack direct observation stations (Campbell & Wynne, 2011, Davis & Swain, 1978), as in the case of Iran. To the authors' knowledge, there are no useful historical ground-based datasets regarding the concentration of air pollutants in any province of Iran.

Early satellite versions were unable to record near-surface PM concentration and instead provided a substitute parameter called Aerosol Optical Depth or AOD (Diao *et al.*, 2019). Considering a ray of light being radiated from a satellite source, the AOD is defined as the decay level of that light reflection from the surface, which is mainly caused by the presence of particles in the air column (Van Donkelaar *et al.*, 2010). The Multi-Angle Implementation of Atmospheric Correction (MAIAC) method, used in conjunction with the satellite-based sensor Moderate Resolution Imaging Spectroradiometer (MODIS), is widely used for predicting the AOD over different regions around the globe. (A. Chudnovsky *et al.*, 2013) evaluated the applicability of the MAIAC algorithm by comparing its predictions with ground-based PM2.5 observations from 84 monitoring stations across New England in the United States over the period 2002-2008. The results indicated that the AOD obtained from the MAIAC is correlated with the observed PM2.5 surface concentration. This research also showed that the MAIAC-AOD shows a better correlation with the in-situ PM2.5 compared to conventional MODIS-AQUA products.

Several studies focused on finding the relation between the AOD and ground-level PM concentration. (H. J. Lee *et al.*, 2012) developed a statistical method to predict the concentration of daily PM2.5 by combining the satellite AOD data with the ground-based ones. Specifically, the MODIS satellite outputs were used together with the observed data from the U.S. Environmental Protection Agency (EPA). The authors evaluated two groups of days consisting of days with or without satellite data availability during the period 2000-2008 for the New England. With a Pearson's Correlation Coefficient (PCC) of 0.91, their predictions were showed this to be a suitable approach, especially in urban areas. (A. A. Chudnovsky *et al.*, 2014) predicted the Fine Particulate Matter (FPM) in the air using high-resolution aerosol data obtained with the MAIAC algorithm using MODIS satellite observations. Several meteorological parameters (e.g., speed of wind and relative humidity), as well as the land use, were utilized to predict the daily-based FPM over New England. After calibrating the FPM satellite data with ground-based observation originated from the EPA, they used a novel interpolation approach, the Inverse Probability Weighting (IPW), to complete the prediction task. Their proposed model predicted real-time FPM with high accuracy. Analogously, (Just *et al.*, 2015) used the MODIS daily AOD values to predict PM2.5 over Mexico City from 2004 to 2014 using statistical modeling. With a correlation coefficient R value of 0.85, their model proved to be an accurate tool for predicting PM2.5 concentration an subcategory of AOD. (X. Zhang *et al.*, 2018) employed the MAIAC-MODIS satellite outputs to extract the AOD records and developed a multi-input statistical model based on geographical properties, climate variables (air temperature, wind speed, and visibility), and land use data to predict the ground-measured PM2.5 concentrations over Texas in the United States, between the years 2008 and 2013. Their proposed model provided accurate predictions with a correlation coefficient of 0.79~0.83.

In the last decade, technological advancement has led Artificial Intelligence (AI) to become the dominant regression and classification approach in many research fields (Mehdizadeh *et al.*, 2017, Nourani *et al.*, 2014, W.-C. Wang *et al.*, 2009). Compared to statistical and numerical methods, AI can achieve the desired target in a much faster and easier manner (Al-Othman *et al.*, 2022, Karandish & Šimůnek, 2016). Benchmark AI algorithms such as Artificial Neural Network (ANN) and Adaptive Neuro-Fuzzy Inference System (ANFIS) have been widely used for prediction in earth sciences. (Mirzaei *et al.*, 2019) investigated the relationship between the satellite-originated AOD values and ground measured PM2.5 concentrations over Tehran in Iran. A model known as Geographically and Temporally Weighted Regression (GTWR) was used to assess this relationship between the years of 2011 and 2017 and convert MODIS-AOD values to PM2.5 surface concentrations. Comparison of four different AI algorithms reveal that the Generalized Regression Neural Network (GRNN) algorithm performed better than its alternatives, ANN and ANFIS.

More advanced AI methods, based on Machine Learning (ML), generally provide better prediction performance compared to “classic” AI algorithms. The MLs show better task in reducing the prediction associated bias and variances, have better overfitting-elusive procedures and can be manually tuned more easily (Khanzode & Sarode, 2020, Müller & Guido, 2016, Wuest *et al.*, 2016). While the early ML models, such as Support Vector Machine (SVM) and Multivariate Adaptive Regression Splines (MARS) demonstrated acceptable performance, newer models called ensemble algorithms have shown superior applicability. Ensemble algorithms such as Ada-boost, Random Forest (RF), and Extreme Tree Regression (ETR) were successfully applied in various studies, outperforming the classic AI algorithms (F. Wang *et al.*, 2021, J. Zhang *et al.*, 2019, Zhu *et al.*, 2021).

The literature review by (Chu *et al.*, 2016) shows that multiple studies have adopted AI algorithms to predict aerosol levels over different regions of the world. For example, (Di *et al.*, 2016) predicted the AOD over the Unites States using ANN algorithms. (Nguyen *et al.*, 2015) employed Support Vector Regression (SVR) and Multiple linear regression (MLR) to simulate the organic carbon concentrations in Gosan, South Korea, between the years 2011 and 2012. Another related study (Lary *et al.*, 2014) utilized a machine learning regression to predict the worldwide aerosol concentration from 1997 to 2014. (Nabavi *et al.*, 2018) compared the performance of several ML algorithms for the prediction of monthly AOD in the western region of Asia using the MODIS outputs. They used wind characteristics, soil temperature, rainfall, drought index, and several other parameters as ML initial predictors, while the MODIS AOD value was used as the target variable. (Kianian *et al.*, 2021) employed RF as an ensemble ML algorithm to predict the spatial distribution of PM2.5, especially in regions that are prone to gaps in AOD coverage. They also used a statistical approach known as Lattice Kriging. Like many previous studies, they first calibrated the MODIS outputs with the EPA ground-based observations. Surface pressure, wind components, temperature, rainfall, relative humidity, radiation flux and many other variables were investigated as the meteorological parameters for PM2.5 prediction.

While the majority of the previous studies used the AOD data obtained from the MODIS satellite, few have investigated other satellite-based aerosol diagnosis outputs such as those from the Modern-Era Retrospective analysis for Research and Applications Version 2, known as MERRA-2 (Gelaro *et al.*, 2017). (Sun *et al.*, 2019) compared the AOD outputs from MERRA-2 and MODIS and evaluated them against ground-based observations at 12 stations in China. Their findings suggested that the MERRA-2 outputs are in good agreement with both MODIS-based and ground-based values. (Gueymard & Yang, 2020) focused on global AOD data for a period of 15 years and showed that MERRA-2 performs better than the European Centre for Medium-Range Weather Forecasts (ECMWF)'s Copernicus Atmosphere Monitoring Service (CAMS). Besides AOD diagnosis outputs, MERRA-2 provides data on ambient air pollutants such as sulfate and dust concentration at the surface level.

This study focuses on surface dust, which is one of the main PM components, and aims to forecast the Surface Dust Mass Concentration (SDMC). The dust-related output of MERRA-2 was selected as the target parameter in this study, as done in several other studies too (Ukhov *et al.*, 2020, Veselovskii *et al.*, 2018, Xu *et al.*, 2020, Yao *et al.*, 2020). The Gradient Boosting Regression (GBR) was used here, because it is a robust and effective ensemble-based prediction algorithm (Johnson *et al.*, 2018, Srivastava *et al.*, 2018, Y. Zhang & Haghani, 2015). Differently from previous investigations, this study uses exclusively remote sensing data for dust concentration prediction. This study also presents, for the first time, a comprehensive analysis of correlation between various meteorological, hydrological, and geological variables with SDMC, and successfully forecasts SDMC few days in advance.

2. Materials and Methods

2.1. Study Area

Dust bowl-like storms have a significant socio-environmental impact in Iran (Salami *et al.*, 2021). Several studies investigated short- and long-term AOD patterns over Iran (Arkian & Nicholson, 2018, Sabetghadam *et al.*, 2018, Salami *et al.*, 2021, Yousefi *et al.*, 2020). Nearly all these studies pinpoint the province of Khuzestan as the region with the highest dust concentration. For example, (Rezaei *et al.*, 2019) evaluated Iran based on its spatial and temporal dust aerosol patterns utilizing the MODIS outputs between 2006 to 2015. Their results show that the Khuzestan and Sistan provinces are the most affected provinces among others. Similar results were obtained by (Mirakbari & Ebrahimi Khusfi, 2020) and (Dadashi-Roudbari & Ahmadi, 2020), which makes the Khuzestan province a good case study for evaluating the SDMC.

As shown in Figure 1, the Khuzestan province is located in the southwestern region of Iran. It has an approximate area of 63,000 km^2 and 4 million inhabitants. Based on the digital elevation map in Figure 1, Khuzestan's elevation below and above the sea level ranges between -105 and 3741 meters, respectively. This is associated with great diversity in climate conditions, from the cold temperatures in the north to tropical conditions in the south. Khuzestan's summer months are considered those from April to September, while October to March are the winter months. The annual average maximum and minimum temperature of this province are $50°C$ and $9°C$ in July and March, respectively. The annual precipitation rate varies from $\sim 200\ mm$ (sea coast in the south) to $\sim 1050\ mm$ (near the Zagros mountains in the north). The dominant wind direction in this province is from west to east and northwest to southeast (Zarasvandi *et al.*, 2011).

[Figure 1]

Dust storms, affecting air quality and impacting social life and economy, have become increasingly frequent in Khuzestan. These storms mainly happen in the summer months and originate from neighboring countries such as Iraq, with west winds into Iran (Daniali & Karimi, 2019). Based on reports from the Khuzestan meteorological stations, just in the year 2008 total of 1035 dust storm events were reported, which is considered a significant number (Zarasvandi, 2009).

2.2. Satellite Data

2.2.1. Modern-Era Retrospective Analysis for Research and Applications Version 2 (MERRA-2)

Due to the high improvement in assimilation structure, The MERRA-2 replaced the original MERRA. It has a more advanced system including hyperspectral radiance and microwave examination. It also includes the Goddard Earth Observing System (GEOS-5) upgrade and ozone samplings of NASA, which makes it an applicable climate evaluation tool (Gelaro *et al.*, 2017). MERRA-2 benefits from the employment of a Grid-point Statistical Interpolation (GSI) climate analysis program, which is structured based on an additive analysis procedure that evaluate the incremental the meteorological data every 6 hours (Gelaro *et al.*, 2017, MERRA, 2AD).

In this study, several types of MERRA-2 outputs were used, from different databases. First, the aerosol diagnosis from M2T1NXAER was used to extract surface dust mass concentration ($\frac{mg}{m^3}$) values over Iran. The M2T1NXAER has a temporal resolution of 1 hour and a spatial resolution is $0.5° \times 0.625°$ longitude and latitude, respectively. M2T1NXAER records black carbon, dry dust, organic carbon, sea salt, and sulfate aerosols in the air (Randles *et al.*, 2017). M2I1NXLFO, also used in this paper, mainly includes land surface data. It has a Same spatial and temporal resolution to M2T1NXAER and includes parameters such as surface layer height, pressure, air temperature, wind speed, and specific humidity (Reichle *et al.*, 2017).

2.2.2. Global Land Data Assimilation System (GLDAS)

GLDAS Version 2, used in this study, is structured in three components, GLDAS-2.0, -2.1, and -2.2. The former, GLDAS-2.0 is completely in congruity with the Princeton meteorological observations and covers the period 1948 to 2014. The year 2000 to present is covered by the 2.1 version. Unlike the two mentioned versions, the GLDAS-2.2 observations utilized data adjustment. GLDAS-2.1 has two major streams, one is associated with the Global Precipitation Climatology Project (GPCP) precipitation products, and one is operating without it. The reason behind this is the 3 to 4-months postponement of GPCP, which forced version 2.1 to represent a temporary data without it called early products. Once the GPCP product become accessible the GLDAS become synchronize with it and the early products become as archive. The data used in this study are from GLDAS-2.1 and have a temporal resolution of 3 hours and a spatial resolution of 0.25°. This product is simulated with version 7 of Land Information System (LIS) from model 3.6 of NOAH. In late 2020 the 3-hourly and monthly GLDAS-2.0 products were re processed with the land mask data from MODIS-MOD44W which corrected the previous version issues such as data missing (Rodell *et al.*, 2004).

2.3. Gradient Boosting Regression (GBR) Algorithm

The boosting technique is based on aggregating a set of simple predictors. This aggregation procedure is structured by focusing on errors originated in each step till a better predictor is constructed with the least outcome error (Nie *et al.*, 2021). Considering Y as the target variable and $X = \{X_1 . X_2 . \cdots . X_n\}$ as the input variables, the aim of the algorithm is to approximate $G'(X)$ as a branch of the original function $G(X)$ to map X to Y, so that the loss function $\mathcal{L}(Y . G(X))$ becomes minimum.

$$G'(X) = \operatorname{argmin} \mathcal{L}_{Y.X}(Y . G(X)) \tag{1}$$

By fitting the simple predictors to the $\mathcal{L}$ at each step of the regression procedure, the Gradient Boosting (GB) algorithm tries to reduce the errors characterizing the preceding steps. This error correctional strategy increases the prediction accuracy and simultaneously decreases the bias of the prediction model. Acknowledging the m ($m = 0. \ldots . M$) as the number of Stages which GB takes to properly train a tree, algorithm first employs an initial simple predictor $G_{m=0}(X)$ and then enforces a gradient so that the $\mathcal{L}$ is minimized. This gradient is computed as follows:

$$Y_i' = -\left[\frac{\partial \mathcal{L}(Y_i . G(X_i))}{\partial G(X_i)}\right]_{G(X)=G_{m-1}(X)} \tag{2}$$

Consider $\mathrm{T}(X_i.\alpha)$ as a regression tree and α as the simple predictor, a new tree can be structured by solving Equation (3), where α_m is the simple predictor parameter at each stage and Ψ is their corresponding weight:

$$\alpha_m = argmin \sum_{i=1}^{N} [Y_i' - \Psi \times \mathrm{T}(X_i.\alpha)]^2 \quad (3)$$

Considering β_m as each stage's optimal length, the $G_m(X)$ is updated at each iteration *m* as follows:

$$\beta_m = argmin \sum_{i=1}^{N} \mathcal{L}(Y_i - G_{m-1}(X_i) + \beta \mathrm{T}(X_i.\alpha_m)) \quad (4)$$

$$G_m(X) = G_{m-1}(X) + \beta_m \mathrm{T}(X_i.\alpha) \quad (5)$$

The iteration continues until the $\beta_m \, \mathrm{T}(X_i, \alpha)$ term in Equation (5) becomes its minimum possible value (Friedman, 2001). Figure shows a flowchart of the GBR algorithm.

[Figure 2]

In this study the GB algorithm was implemented using the ensemble sub-category of the Scikit-learn library of the Python Programming language. The provided Gradient Boosting Regression (GBR) algorithm has several parameters that need to be "tuned" so that the algorithm performs as its maximum accuracy. Table 1 shows the default and optimal values for each parameter, obtained with an iterative process developed by the authors.

Table 1: Gradient Boosting Regression (GBR) parameters.

Parameter	Description	Default Value	Optimal Value
loss	Enhancement approach for the loss function	Squared error	Absolute error
learning_rate	This parameter reduces the value of each tree's contribution	0.1	1.1
n_estimators	Number of boosting phases that must be accomplished	100	498
random_state	At each boosting step, this parameter sets the random seed delivered to each tree	None	34
max_depth	Individual regression estimators' maximum depth	3	1
min_samples_leaf	Number of minimum samples needed to consider a node a leaf node.	1	45

3. Results

3.1. Selection of Year and Region of Focus

MERRA-2-M2TMNXAER monthly SDMC data for Iran were downloaded from the NASA website for the period between 2009 and 2020 (*https://disc.gsfc.nasa.gov/datasets/M2TMNXAER_5.12.4/summary?keywords=M2TMNXAER*). This satellite provides monthly averaged aerosol diagnosis with spatial resolution of $0.5° \times 0.625°$. Considering this resolution there were total of 1056 observational grid cross-section points over the Iran. By calculating the maximum value of SDMC among the 12 months in each specific year and at each of 1056 points, Figure 3 shows the distribution of yearly maximum SDMC (mg/m^3) over Iran for the period considered. For better understanding of each specific year ranking based on SDMC level, thier Annual Maximum (A.M.) has been noted in the figure.

[Figure 3]

Figure 3 shows that the southern and southeastern sections of Iran, including the provinces of Khuzestan and Hormozgan as well as Sistan and Baluchestan, are highly prone to large dust concentrations. The Khuzestan province appears to be the most impacted province and is the focus of this investigation. Over the 12-year period considered, Iran was overall characterized by minimum, mean and maximum SDMC values of 0.04, 0.295 and 1.36 mg/m^3 respectively. The year 2018, as the year with highest SDMC observed value ($\sim 1.36\ mg/m^3$) was selected as the study period in current study.

3.2. Input Variable Correlation Analysis

Having selected the year and region of focus, SDMC data for the year 2018 for the Khuzestan province was extracted based on outputs extracted from MERRA-2-M2T1NXADG (*https://disc.gsfc.nasa.gov/datasets/M2T1NXADG_5.12.4/summary?keywords=M2T1NXADG*), consisting of 1-hour time averaged aerosol diagnosis data with the same spatial resolution as MERRA-2-M2TMNXAER. 1-hour data was converted to daily data by taking the maximum value within each 24-hour period, therefore obtaining 365 SDMC values.

As mentioned, the Khuzestan province is characterized by different types of climate conditions due to its significant variation in land elevation; therefore 18 Research Points (RPs) were considered in this study, equally distributed across the province (with latitudinal and longitudinal spacing of 0.5° and 0.625°, respectively, same as the resolution of MERRA2-M2T1NXADG). These include the cold region in the north as well as the extremely hot regions in the south.

Then, several hydrological, meteorological and geological outputs from Global Precipitation Measurement (GPM), Global Land Data Assimilation System (GLDAS), MERRA-2-M2I1NXLFO and MERRA-2-M2T3NVASM with various spatial and temporal distribution were extracted. Table 2 lists these variables, which are the 11 input variables considered for SDMC prediction in this study, with their units and corresponding satellite.

Table 2: Input variables considered for SDMC prediction.

Satellite	***Variable***	***Description***	***Units***
GLDAS	***Acond***	Aerodynamic conductance	m/s
	Esoil	Evaporation flux from soil	kg/m^2s
	Qh	Surface upward sensible heat flux	W/m^3
	Evap	Evapotranspiration	kg/m^2s
	SoilMoist	Surface soil moisture	kg/m^2
MERRA-2	***PS***	Surface pressure	Pa
	TLML	Surface air temperature	K
	SPEEDLML	Surface wind speed	m/s
	QLML	Surface specific humidity	–
	HLML	Surface layer height	m
GPM	***Precip***	Precipitation rate	mm

The Pearson’s Correlation Coefficient (PCC) was computed to quantify the correlation of each of the 11 variables in Table 2 with daily SDMC. While there are other alternatives, such as Spearman's or Kendall’s rank correlation, various studies found the efficiency and robustness of PCC, especially in handling data with nonapparent outliers and non-linearity (Chok, 2010, Hauke & Kossowski, 2011, Rebekić *et al.*, 2015).

Figure 4 contains the relative PCC heat map for the 18 RPs, including the average PCC values over all the RPs. The three parameters with the highest PCC value, surface wind speed SPEEDLML ($PCC = 0.61$), aerodynamic conductance ACOND ($PCC = 0.57$) and surface pressure PS ($PCC = 0.53$), were selected as input variables for real-time daily SDMC prediction with the GBR algorithm. PS denotes the atmospheric surface pressure that directly controls the movement of air masses from regions with low pressure to regions with higher pressure (Gomis *et al.*, 2008, Guo *et al.*, 2011). The surface aerodynamic conductance (ACOND) describes the effect of surface roughness on the movement of air masses (S. Liu *et al.*, 2007, Mallick *et al.*, 2018).

[Figure 4]

From Figure 4, the hydrological parameters rainfall, relative humidity and evapotranspiration have the least correlation with SDMC in the Khuzestan province.

3.3. Real-Time Prediction of SDMC

The three input variables for prediction, known at the current time (day) "t", were used in the GBR algorithm to predict the current time SDMC. This was done for all 18 RPs and the prediction performance was evaluated using four indices, PCC, Nash-Sutcliffe Efficiency (NSE), Normalized-Root Mean Squared Error (N-RMSE) and Normalized Mean Absolute Error (N-MAE). Results of these metrics which have been widely used in earth and water research fields (Jodar-Abellan *et al.*, 2019, Moriasi *et al.*, 2007, Pardo *et al.*, 2020), are shown in Table 3.

Table 3: SDMC real-time prediction performance for the 18 Research Points considered.

RPs	PCC	NSE	N-RMSE	N-MAE
P1	0.716	0.493	0.710	0.494
P2	0.771	0.585	0.642	0.453
P3	0.815	0.632	0.605	0.422
P4	0.648	0.397	0.774	0.446
P5	0.707	0.473	0.723	0.516
P6	0.793	0.621	0.613	0.437
P7	0.783	0.598	0.632	0.455
P8	0.706	0.457	0.735	0.518
P9	0.720	0.490	0.712	0.507
P10	0.686	0.451	0.738	0.517
P11	0.505	0.170	0.908	0.613
P12	0.701	0.481	0.718	0.494
P13	0.684	0.425	0.756	0.563
P14	0.525	0.223	0.878	0.628
P15	0.405	0.079	0.957	0.594
P16	0.541	0.253	0.862	0.618
P17	0.435	0.139	0.925	0.702
P18	0.398	0.095	0.948	0.649

Values in Table 3 vary across the RPs but, overall, the average PCC, NSE, N-RMSE and N-MAE are within an acceptable range. The P3 Research Point ($CC = 0.815$, $NSE = 0.632$, $N - RMSE = 0.605$ and $N - MAE = 0.422$) is characterized by the best prediction performance. Figure 5 highlights a clear pattern of better prediction performance in the southern regions of the province. This figure has been obtained using the Inverse Distance Weighting (IDW) interpolation method over the Khuzestan province based on the calculated performance indices of 18 research points.

[Figure 5]

3.4. Future Forecasting of SDMC

For the P3 Research Point (best real-time SDMC prediction, see previous section), future forecasting of SDMC was considered for lead times of ‘t-2’, ‘t-4’, ‘t-6’ and ‘t-8’ (input data from 2, 4, 6 and 8 days prior to the current time ‘t’ for which SDMC is forecasted). PCC values were calculated to quantify the correlation between the current time SDMC (at time “t”) and the 11 input variables from Table 2 for the four lead times. Figure 6 shows the PCC values in a circular bar chart.

[Figure 6]

From Figure 6 it can be seen that, analogously to the case of real-time prediction, precipitation and surface specific humidity have the lowest correlation with the SDMC. PS and surface upward sensible heat flux Qh and surface air temperature TLML correlations with SDMC show a significant increase when moving backward in time.

Based on the PCC analysis and using a procedure based on progressive elimination of input variables with lower and lower PCC value, a procedure applied in other previous studies (Sharafati, Asadollah, & Hosseinzadeh, 2020, Sharafati, Asadollah, & Neshat, 2020), combinations of the 11 input variables from Table 2 were constructed for future forecasting of SDMC (lead times of ‘t-2’, ‘t-4’, ‘t-6’ and ‘t-8’) using the GBR algorithm. The combinations considered are listed in Table 4.

Table 4: Input variable combinations for future forecasting of SDMC.

Combination	Input Variables
C1	*PS (t-4), PS (t-6), PS (t-2), PS (t-8), Qh (t-2), Qh (t-4)*
C2	*PS (t-4), PS (t-6), PS (t-2), PS (t-8), Qh (t-2)*
C3	*PS (t-4), PS (t-6), PS (t-2), PS (t-8)*
C4	*PS (t-4), PS (t-6), PS (t-2)*
C5	*PS (t-4), PS (t-6)*
C6	*PS (t-4)*
C7	*PS (t-2), Qh (t-2)*
C8	*PS (t-4), Qh (t-4)*
C9	*PS (t-6), TLML (t-6)*
C10	*PS (t-8), TLML (t-8)*

Figure 7 summarizes the prediction performance of the GBR algorithm for future forecasting of SDMC using the ten input variable combinations in Table 4, evaluated based on the PCC, NSE, N-RMSE and N-MAE indices.

[Figure 7]

From Figure 7, the prediction performance decreases when moving from lead time ‘t-2’ to ‘t-8’ for the input variables, as also observed in previous studies (Sharafati, Haji Seyed Asadollah, *et al.*, 2020). The C1 input variable combination, including the highest number of input variables (six), is associated with the best prediction performance ($PCC = 0.698$, $NRMSE = 0.733$). However, the C6 input variable combination, including only one variable (*PS (t-4)*) shows only an 8% reduction in accuracy ($PCC = 0.640$, $NRMSE = 0.781$), which is considered an insignificant performance reduction. C6 is therefore the optimal input variable combination, because it only requires the knowledge of a single variable (PS). The preferable use of lead time ‘t-4’ is also confirmed by the high performance of the C8 input variable combination ($PCC = 0.665$, $NRMSE = 0.759$).

4. Discussion and Conclusion

Being able to forecast Particulate Matter (PM) concentrations is essential, due to its effects on human life, economy and environment. This study aimed to simulate Surface Mass Dust Concentration using the Modern-Era Retrospective analysis for Research and Applications Version 2 (MERRA-2) aerosol satellite diagnosis. To do this, monthly SDMC data for the period 2009 to 2020 were downloaded from the MERRA-2 database.

The monthly evaluation of air dust distribution in Iran showed that the southeastern regions are characterized by higher dust concentrations, negatively affecting air quality. The province of Khuzestan is the most impacted, as also confirmed by other investigations (Sabetghadam *et al.*, 2018, Yousefi *et al.*, 2020), and was therefore selected as the case study. The year 2018 was specifically considered, because it was characterized by high SDMC and has complete recorded satellite observations. SDMC hourly data were obtained from MERRA-2 for 2018 for the Khuzestan province.

Pearson's Correlation Coefficient (PCC) computation to evaluate the correlation between MERRA-2 SDMC and 11 meteorological, hydrological and geological parameters from Global Precipitation Measurement (GPM), Global Land Data Assimilation System (GLDAS) and another MERRA-2 database showed the wind-related variables - surface wind speed (SPEEDLML), surface aerodynamic conductivity (ACOND) and surface pressure (PS) - to be the most correlated with SDMC. Combinations of these three parameters were evaluated for real-time prediction of SDMC using the Gradient Boosting Regression (GBR) algorithm at 18 Research Points. The best prediction performance was obtained at Research Point 3 $(CC = 0.815\ and\ N - RMSE = 0.605)$, which was considered for the further forecasting analysis. The input variables are solely based on remote sensing (one of the elements of novelty of this study), therefore the related errors and uncertainties are expected to affect the SDMC prediction performance, compared to models using ground-based measurements. However, while for some Research Points the GBR algorithm produced predictions with PCC value in the moderate acceptance range ($0.3 \leq PCC \leq 0.7$) based on (Ratner, 2009), there are several Research Points where the prediction performance was strong ($0.7 \leq PCC \leq 1.0$). This outcome confirms the prediction potential of ensemble algorithms when using data affected by errors and modelling processes characterized by non-linearity. Comparing this study's results with those by (Nabavi *et al.*, 2018) shows that the GBR algorithm outperforms both SVM ($PCC = 0.81$) and MARS ($PCC = 0.80$). The GBR also appears to have better accuracy than ANN ($PCC = 0.62$), ANFIS ($PCC = 0.70$) and GRNN ($PCC = 0.71$) (Mirzaei *et al.*, 2019). From a spatial point of view, our results highlighted a better prediction performance for the southern low lands of the Khuzestan province, when compared with the higher and mountainous lands in the north. Predictions were also better in marshlands compared with rocky soils.

While the wind-related input variables govern the real time ('t') prediction of SDMC, the heat-related variables are also important for the future forecasting of SDMC (lead times of 't-2' to 't-8'). Considering Research Point 3 for the analysis, the PS variable allows for the better forecasting, closely followed by surface upward sensible heat flux (Qh) and surface air temperature (TLML). It is worth mentioning that, as the lead time goes from 't-2' to 't-8', the influence of TLML on SDMC forecasting becomes stronger than that of Qh. The evaluation of input variable combinations for future SDMC forecasting revealed that the use of a single input variable, *PS (t-4)*, with $PCC = 0.640\ and\ N - RMSE = 0.781$, is the optimal approach (most cost-efficient and applicable). To the authors' knowledge, there are no previous studies specifically aimed at forecasting future air pollutant concentrations. The future forecasting performance obtained here with the GBR algorithm is comparable with previously presented real-time predictions.

Declaration of interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

Al-Othman, A., Tawalbeh, M., Martis, R., Dhou, S., Orhan, M., Qasim, M. & Olabi, A. G. (2022) Artificial intelligence and numerical models in hybrid renewable energy systems with fuel cells: Advances and prospects. *Energy Conversion and Management* **253**, 115154. Elsevier.

Amaral, S. S., Carvalho, J. A. De, Costa, M. A. M. & Pinheiro, C. (2015) An overview of particulate matter measurement instruments. *Atmosphere* **6**(9), 1327–1345. Multidisciplinary Digital Publishing Institute.

Anderson, J. O., Thundiyil, J. G. & Stolbach, A. (2012) Clearing the air: a review of the effects of particulate matter air pollution on human health. *Journal of medical toxicology* **8**(2), 166–175. Springer.

Arkian, F. & Nicholson, S. E. (2018) Long-term variations of aerosol optical depth and aerosol radiative forcing over Iran based on satellite and AERONET data. *Environmental monitoring and assessment* **190**(1), 1–15. Springer.

Asadollah, S. B. H. S., Khan, N., Sharafati, A., Shahid, S., Chung, E.-S. & Wang, X.-J. (2021) Prediction of heat waves using meteorological variables in diverse regions of Iran with advanced machine learning models. *Stochastic Environmental Research and Risk Assessment* 1–16. Springer.

Brook, R. D. & Rajagopalan, S. (2009) Particulate matter, air pollution, and blood pressure. *Journal of the American Society of Hypertension* **3**(5), 332–350. Elsevier.

Campbell, J. B. & Wynne, R. H. (2011) *Introduction to remote sensing*. Guilford Press.

Chen, T.-M., Kuschner, W. G., Gokhale, J. & Shofer, S. (2007) Outdoor air pollution: nitrogen dioxide, sulfur dioxide, and carbon monoxide health effects. *The American journal of the medical sciences* **333**(4), 249–256. Elsevier.

Chok, N. S. (2010) Pearson’s versus Spearman’s and Kendall’s correlation coefficients for continuous data. University of Pittsburgh.

Chu, Y., Liu, Y., Li, X., Liu, Z., Lu, H., Lu, Y., Mao, Z., et al. (2016) A review on predicting ground PM2. 5 concentration using satellite aerosol optical depth. *Atmosphere* **7**(10), 129. Multidisciplinary Digital Publishing Institute.

Chudnovsky, A. A., Koutrakis, P., Kloog, I., Melly, S., Nordio, F., Lyapustin, A., Wang, Y., et al. (2014) Fine particulate matter predictions using high resolution Aerosol Optical Depth (AOD) retrievals. *Atmospheric Environment* **89**, 189–198. Elsevier.

Chudnovsky, A., Tang, C., Lyapustin, A., Wang, Y., Schwartz, J. & Koutrakis, P. (2013) A critical assessment of high-resolution aerosol optical depth retrievals for fine particulate matter predictions. *Atmospheric Chemistry and Physics* **13**(21), 10907–10917. Copernicus GmbH.

Council, N. R. (1992) *Rethinking the ozone problem in urban and regional air pollution*. National Academies Press.

Dadashi-Roudbari, A. & Ahmadi, M. (2020) Evaluating temporal and spatial variability and trend of aerosol optical depth (550 nm) over Iran using data from MODIS on board the Terra and Aqua satellites. *Arabian Journal of Geosciences* **13**(6), 1–23. Springer.

Daniali, M. & Karimi, N. (2019) Spatiotemporal analysis of dust patterns over Mesopotamia and their impact on Khuzestan province, Iran. *Natural Hazards* **97**(1), 259–281. Springer.

Davis, S. M. & Swain, P. H. (1978) Remote sensing: the quantitative approach. McGraw-Hill International Book Company New York.

Di, Q., Kloog, I., Koutrakis, P., Lyapustin, A., Wang, Y. & Schwartz, J. (2016) Assessing PM2. 5 exposures with high spatiotemporal resolution across the continental United States. *Environmental science & technology* **50**(9), 4712–4721. ACS Publications.

Diao, M., Holloway, T., Choi, S., O'Neill, S. M., Al-Hamdan, M. Z., Donkelaar, A. Van, Martin, R. V, et al. (2019) Methods, availability, and applications of PM2. 5 exposure estimates derived from ground measurements, satellite, and atmospheric models. *Journal of the Air & Waste Management Association* **69**(12), 1391–1414. Taylor & Francis.

Donkelaar, A. Van, Martin, R. V, Brauer, M., Kahn, R., Levy, R., Verduzco, C. & Villeneuve, P. J. (2010) Global estimates of ambient fine particulate matter concentrations from satellite-based aerosol optical depth: development and application. *Environmental health perspectives* **118**(6), 847–855. National Institute of Environmental Health Sciences.

Duan, Y., Liao, Y., Li, H., Yan, S., Zhao, Z., Yu, S., Fu, Y., et al. (2019) Effect of changes in season and temperature on cardiovascular mortality associated with nitrogen dioxide air pollution in Shenzhen, China. *Science of The Total Environment* **697**, 134051. Elsevier.

Friedman, J. H. (2001) Greedy function approximation: a gradient boosting machine. *Annals of statistics* 1189–1232. JSTOR.

Gelaro, R., McCarty, W., Suárez, M. J., Todling, R., Molod, A., Takacs, L., Randles, C. A., et al. (2017) The modern-era retrospective analysis for research and applications, version 2 (MERRA-2). *Journal of climate* **30**(14), 5419–5454.

Ghozat, A., Sharafati, A. & Hosseini, S. A. (2022) Satellite-based monitoring of meteorological drought over different regions of Iran: application of the CHIRPS precipitation product. *Environmental Science and Pollution Research* 1–18. Springer.

Gomis, D., Ruiz, S., Sotillo, M. G., Álvarez-Fanjul, E. & Terradas, J. (2008) Low frequency Mediterranean sea level variability: the contribution of atmospheric pressure and wind. *Global and Planetary Change* **63**(2–3), 215–229. Elsevier.

Gueymard, C. A. & Yang, D. (2020) Worldwide validation of CAMS and MERRA-2 reanalysis aerosol optical depth products using 15 years of AERONET observations. *Atmospheric Environment* **225**, 117216. Elsevier.

Guo, H., Xu, M. & Hu, Q. (2011) Changes in near-surface wind speed in China:

1969–2005. *International Journal of Climatology* **31**(3), 349–358. Wiley Online Library.

Hamanaka, R. B. & Mutlu, G. M. (2018) Particulate matter air pollution: effects on the cardiovascular system. *Frontiers in endocrinology* **9**, 680. Frontiers.

Hauke, J. & Kossowski, T. (2011) Comparison of values of Pearson's and Spearman's correlation coefficient on the same sets of data. Wydział Nauk Geograficznych i Geologicznych Uniwersytetu im. Adama Mickiewicza.

Jodar-Abellan, A., Valdes-Abellan, J., Pla, C. & Gomariz-Castillo, F. (2019) Impact of land use changes on flash flood prediction using a sub-daily SWAT model in five Mediterranean ungauged watersheds (SE Spain). *Science of the Total Environment* **657**, 1578–1591. Elsevier.

Johnson, N. E., Bonczak, B. & Kontokosta, C. E. (2018) Using a gradient boosting model to improve the performance of low-cost aerosol monitors in a dense, heterogeneous urban environment. *Atmospheric environment* **184**, 9–16. Elsevier.

Just, A. C., Wright, R. O., Schwartz, J., Coull, B. A., Baccarelli, A. A., Tellez-Rojo, M. M., Moody, E., et al. (2015) Using high-resolution satellite aerosol optical depth to estimate daily PM2. 5 geographical distribution in Mexico City. *Environmental science & technology* **49**(14), 8576–8584. ACS Publications.

Karandish, F. & Šimůnek, J. (2016) A comparison of numerical and machine-learning modeling of soil water content with limited input data. *Journal of Hydrology* **543**, 892–909. Elsevier.

Khanzode, K. C. A. & Sarode, R. D. (2020) Advantages and Disadvantages of

Artificial Intelligence and Machine Learning: A Literature Review. *International Journal of Library & Information Science (IJLIS)* **9**(1), 3.

Kianian, B., Liu, Y. & Chang, H. H. (2021) Imputing satellite-derived aerosol optical depth using a multi-resolution spatial model and random forest for PM2. 5 prediction. *Remote Sensing* **13**(1), 126. Multidisciplinary Digital Publishing Institute.

Kwasny, F., Madl, P. & Hofmann, W. (2010) Correlation of air quality data to ultrafine particles (UFP) concentration and size distribution in ambient air. *Atmosphere* **1**(1), 3–14. Molecular Diversity Preservation International.

Lary, D. J., Faruque, F. S., Malakar, N., Moore, A., Roscoe, B., Adams, Z. L. & Eggelston, Y. (2014) Estimating the global abundance of ground level presence of particulate matter (PM2. 5). *Geospatial health* **8**(3), S611–S630.

Lee, H. J., Coull, B. A., Bell, M. L. & Koutrakis, P. (2012) Use of satellite-based aerosol optical depth and spatial clustering to predict ambient PM2. 5 concentrations. *Environmental research* **118**, 8–15. Elsevier.

Lee, M., Kloog, I., Chudnovsky, A., Lyapustin, A., Wang, Y., Melly, S., Coull, B., et al. (2016) Spatiotemporal prediction of fine particulate matter using high-resolution satellite images in the Southeastern US 2003–2011. *Journal of exposure science & environmental epidemiology* **26**(4), 377–384. Nature Publishing Group.

Liu, D. & Li, L. (2015) Application study of comprehensive forecasting model based on entropy weighting method on trend of PM2. 5 concentration in Guangzhou, China. *International journal of environmental research and public health* **12**(6), 7085–7099. Multidisciplinary Digital Publishing Institute.

Liu, S., Lu, L., Mao, D. & Jia, L. (2007) Evaluating parameterizations of aerodynamic resistance to heat transfer using field measurements. *Hydrology and earth system sciences* **11**(2), 769–783. Copernicus GmbH.

Mallick, K., Wandera, L., Bhattarai, N., Hostache, R., Kleniewska, M. & Chormanski, J. (2018) A critical evaluation on the role of aerodynamic and canopy–surface conductance parameterization in SEB and SVAT models for simulating evapotranspiration: A case study in the upper biebrza national park wetland in poland. *Water* **10**(12), 1753. Multidisciplinary Digital Publishing Institute.

Mehdizadeh, S., Behmanesh, J. & Khalili, K. (2017) Evaluating the performance of artificial intelligence methods for estimation of monthly mean soil temperature without using meteorological data. *Environmental Earth Sciences* **76**(8), 1–16. Springer.

MERRA, G. (2AD) tavgU_2d_lnd_Nx: 2d, diurnal, time-averaged, single-level, assimilation, land surface diagnostics V5. 12.4. *EarthData GES DISC NASA*.

Mirakbari, M. & Ebrahimi Khusfi, Z. (2020) Investigation of spatial and temporal changes in atmospheric aerosol using aerosol optical depth in Southeastern Iran. *Journal of RS and GIS for Natural Resources* **11**(3), 87–105. Bushehr Branch, Islamic Azad University.

Mirzaei, M., Amanollahi, J. & Tzanis, C. G. (2019) Evaluation of linear, nonlinear, and hybrid models for predicting PM 2.5 based on a GTWR model and MODIS AOD data. *Air Quality, Atmosphere & Health* **12**(10), 1215–1224. Springer.

Moriasi, D. N., Arnold, J. G., Liew, M. W. Van, Bingner, R. L., Harmel, R. D. & Veith, T. L. (2007) Model evaluation guidelines for systematic quantification of accuracy in watershed simulations. *Transactions of the ASABE* **50**(3), 885–900. American society of agricultural and biological engineers.

Müller, A. C. & Guido, S. (2016) *Introduction to machine learning with Python: a guide for data scientists*. ‘ O’Reilly Media, Inc.’

Nabavi, S. O., Haimberger, L., Abbasi, R. & Samimi, C. (2018) Prediction of aerosol optical depth in West Asia using deterministic models and machine learning algorithms. *Aeolian research* **35**, 69–84. Elsevier.

Nakata, M., Sano, I., Mukai, S. & Holben, B. N. (2013) Spatial and temporal variations of atmospheric aerosol in Osaka. *Atmosphere* **4**(2), 157–168. Multidisciplinary Digital Publishing Institute.

Nguyen, D. L., Kim, J. Y., Ghim, Y. S. & Shim, S.-G. (2015) Influence of regional biomass burning on the highly elevated organic carbon concentrations observed at Gosan, South Korea during a strong Asian dust period. *Environmental Science and Pollution Research* **22**(5), 3594–3605. Springer.

Nie, P., Roccotelli, M., Fanti, M. P., Ming, Z. & Li, Z. (2021) Prediction of home energy consumption based on gradient boosting regression tree. *Energy Reports* **7**, 1246–1255. Elsevier.

Nourani, V., Baghanam, A. H., Adamowski, J. & Kisi, O. (2014) Applications of hybrid wavelet–artificial intelligence models in hydrology: a review. *Journal of Hydrology* **514**, 358–377. Elsevier.

Pardo, M. Á., Riquelme, A. J., Jodar-Abellan, A. & Melgarejo, J. (2020) Water and energy demand management in pressurized irrigation networks. *Water* **12**(7), 1878. Multidisciplinary Digital Publishing Institute.

Planning, U. S. E. P. A. O. of A. Q. (1996) *Review of the national ambient air* *quality standards for particulate matter: Policy assessment of scientific and technical information*. DIANE Publishing.

Raaschou-Nielsen, O., Beelen, R., Wang, M., Hoek, G., Andersen, Z. J., Hoffmann, B., Stafoggia, M., et al. (2016) Particulate matter air pollution components and risk for lung cancer. *Environment international* **87**, 66–73. Elsevier.

Randles, C. A., Silva, A. M. Da, Buchard, V., Colarco, P. R., Darmenov, A., Govindaraju, R., Smirnov, A., et al. (2017) The MERRA-2 aerosol reanalysis, 1980 onward. Part I: System description and data assimilation evaluation. *Journal of climate* **30**(17), 6823–6850.

Ratner, B. (2009) The correlation coefficient: Its values range between+ 1/− 1, or do they? *Journal of targeting, measurement and analysis for marketing* **17**(2), 139–142. Springer.

Rebekić, A., Lončarić, Z., Petrović, S. & Marić, S. (2015) Pearson's or Spearman's correlation coefficient-which one to use? *Poljoprivreda* **21**(2), 47–54. Fakultet agrobiotehničkih znanosti Osijek i Poljoprivredni institut Osijek.

Reichle, R. H., Draper, C. S., Liu, Q., Girotto, M., Mahanama, S. P. P., Koster, R. D. & Lannoy, G. J. M. De. (2017) Assessment of MERRA-2 land surface hydrology estimates. *Journal of Climate* **30**(8), 2937–2960.

Rezaei, M., Farajzadeh, M., Mielonen, T. & Ghavidel, Y. (2019) Analysis of spatio-temporal dust aerosol frequency over Iran based on satellite data. *Atmospheric Pollution Research* **10**(2), 508–519. Elsevier.

Rodell, M., Houser, P. R., Jambor, U. E. A., Gottschalck, J., Mitchell, K., Meng, C.-J. J., Arsenault, K., et al. (2004) The global land data assimilation system.

Bulletin of the American Meteorological Society **85**(3), 381–394. American Meteorological Society. doi:10.1175/BAMS-85-3-381

Sabetghadam, S., Khoshsima, M. & Alizadeh-Choobari, O. (2018) Spatial and temporal variations of satellite-based aerosol optical depth over Iran in Southwest Asia: Identification of a regional aerosol hot spot. *Atmospheric Pollution Research* **9**(5), 849–856. Elsevier.

Salami, H., Khorami, S., Yazdani, S. & Saleh, I. (2021) Economic Evaluation of the Damages of Dust Bowl on Crop Yield by Choice Experiment Method in Khuzestan Province of Iran. *International Journal of Agricultural Management and Development* **11**(3). Islamic Azad University, Rasht Branch.

Sharafati, A., Asadollah, S. B. H. S. & Hosseinzadeh, M. (2020) The potential of new ensemble machine learning models for effluent quality parameters prediction and related uncertainty. *Process Safety and Environmental Protection*. Elsevier.

Sharafati, A., Asadollah, S. B. H. S. & Neshat, A. (2020) A new artificial intelligence strategy for predicting the groundwater level over the Rafsanjan aquifer in Iran. *Journal of Hydrology* **591**, 125468. Elsevier.

Sharafati, A., Haji Seyed Asadollah, S. B., Motta, D. & Yaseen, Z. M. (2020) Application of newly developed ensemble machine learning models for daily suspended sediment load prediction and related uncertainty analysis. *Hydrological Sciences Journal*. Taylor & Francis.

Shiru, M. S., Shahid, S., Chae, S.-T. & Chung, E.-S. (2022) Replicability of Annual and Seasonal Precipitation by CMIP5 and CMIP6 GCMs over East Asia. *KSCE Journal of Civil Engineering* 1–12. Springer.

Srivastava, C., Singh, S. & Singh, A. P. (2018) Estimation of air pollution in Delhi using machine learning techniques. *2018 International Conference on Computing, Power and Communication Technologies (GUCON)*, 304–309. IEEE.

Sun, E., Xu, X., Che, H., Tang, Z., Gui, K., An, L., Lu, C., et al. (2019) Variation in MERRA-2 aerosol optical depth and absorption aerosol optical depth over China from 1980 to 2017. *Journal of Atmospheric and Solar-Terrestrial Physics* **186**, 8–19. Elsevier.

Sunyer, J., Ballester, F., Tertre, A. Le, Atkinson, R., Ayres, J. G., Forastiere, F., Forsberg, B., et al. (2003) The association of daily sulfur dioxide air pollution levels with hospital admissions for cardiovascular diseases in Europe (The Aphea-II study). *European heart journal* **24**(8), 752–760. Oxford University Press.

Ukhov, A., Mostamandi, S., Silva, A. Da, Flemming, J., Alshehri, Y., Shevchenko, I. & Stenchikov, G. (2020) Assessment of natural and anthropogenic aerosol air pollution in the Middle East using MERRA-2, CAMS data assimilation products, and high-resolution WRF-Chem model simulations. *Atmospheric Chemistry and Physics* **20**(15), 9281–9310. Copernicus GmbH.

USEPA, O. A. R. (2019) Health and environmental effects of particulate matter (PM). Retrieved.

Veselovskii, I., Goloub, P., Podvin, T., Tanre, D., Silva, A. Da, Colarco, P., Castellanos, P., et al. (2018) Characterization of smoke and dust episode over West Africa: comparison of MERRA-2 modeling with multiwavelength Mie–Raman lidar observations. *Atmospheric measurement techniques* **11**(2), 949–969. Copernicus GmbH.

Wang, F., Wang, Y., Zhang, K., Hu, M., Weng, Q. & Zhang, H. (2021) Spatial heterogeneity modeling of water quality based on random forest regression and model interpretation. *Environmental Research* **202**, 111660. Elsevier.

Wang, W.-C., Chau, K.-W., Cheng, C.-T. & Qiu, L. (2009) A comparison of performance of several artificial intelligence methods for forecasting monthly discharge time series. *Journal of hydrology* **374**(3–4), 294–306. Elsevier.

WHO. (2014) 7 million premature deaths annually linked to air pollution. Retrieved March 25, 2014, from https://www.who.int/news/item/25-03-2014-7-million-premature-deaths-annually-linked-to-air-pollution

Wuest, T., Weimer, D., Irgens, C. & Thoben, K.-D. (2016) Machine learning in manufacturing: advantages, challenges, and applications. *Production & Manufacturing Research* **4**(1), 23–45. Taylor & Francis.

Xu, X., Wu, H., Yang, X. & Xie, L. (2020) Distribution and transport characteristics of dust aerosol over Tibetan Plateau and Taklimakan Desert in China using MERRA-2 and CALIPSO data. *Atmospheric Environment* **237**, 117670. Elsevier.

Yao, W., Che, H., Gui, K., Wang, Y. & Zhang, X. (2020) Can MERRA-2 reanalysis data reproduce the three-dimensional evolution characteristics of a typical dust process in East Asia? A case study of the dust event in May 2017. *Remote Sensing* **12**(6), 902. Multidisciplinary Digital Publishing Institute.

Yousefi, R., Wang, F., Ge, Q. & Shaheen, A. (2020) Long-term aerosol optical depth trend over Iran and identification of dominant aerosol types. *Science of The Total Environment* **722**, 137906. Elsevier.

Zarasvandi, A. (2009) Environmental impacts of dust storms in the Khuzestan province. *Environmental Protection Agency (EPA) of Khuzestan province, Internal Report, 375p*.

Zarasvandi, A., Carranza, E. J. M., Moore, F. & Rastmanesh, F. (2011) Spatio-temporal occurrences and mineralogical–geochemical characteristics of airborne dusts in Khuzestan Province (southwestern Iran). *Journal of geochemical exploration* **111**(3), 138–151. Elsevier.

Zhang, J., Ma, G., Huang, Y., Aslani, F. & Nener, B. (2019) Modelling uniaxial compressive strength of lightweight self-compacting concrete using random forest regression. *Construction and Building Materials* **210**, 713–719. Elsevier.

Zhang, X., Chu, Y., Wang, Y. & Zhang, K. (2018) Predicting daily PM2. 5 concentrations in Texas using high-resolution satellite aerosol optical depth. *Science of the Total Environment* **631**, 904–911. Elsevier.

Zhang, Y. & Haghani, A. (2015) A gradient boosting method to improve travel time prediction. *Transportation Research Part C: Emerging Technologies* **58**, 308–324. Elsevier.

Zhu, X., Zhang, P. & Xie, M. (2021) A Joint Long Short-Term Memory and AdaBoost regression approach with application to remaining useful life estimation. *Measurement* **170**, 108707. Elsevier.

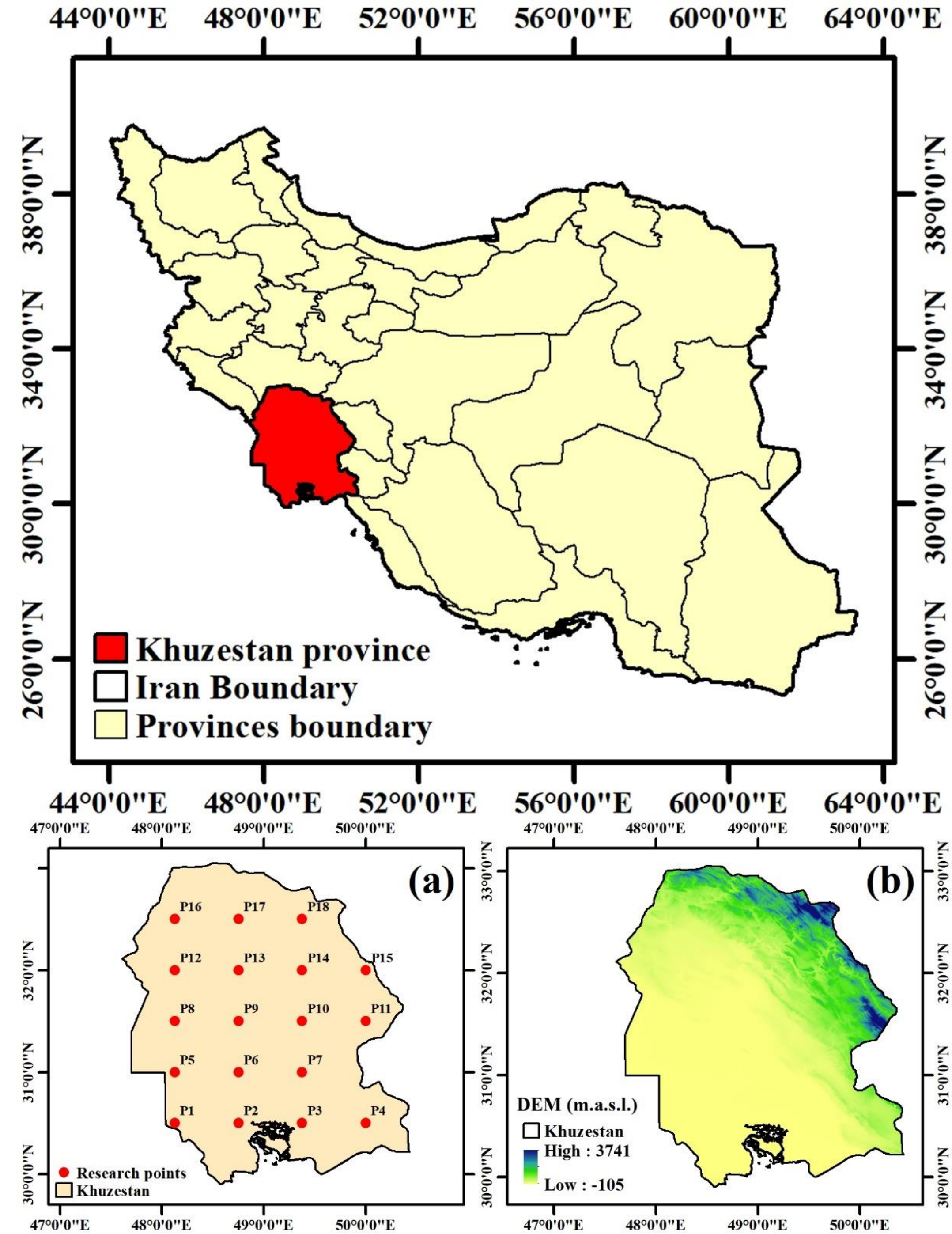

Figure 1: Location of the Khuzestan province in Iran with (a) Research Point distribution and (b) Digital Elevation Model (DEM).

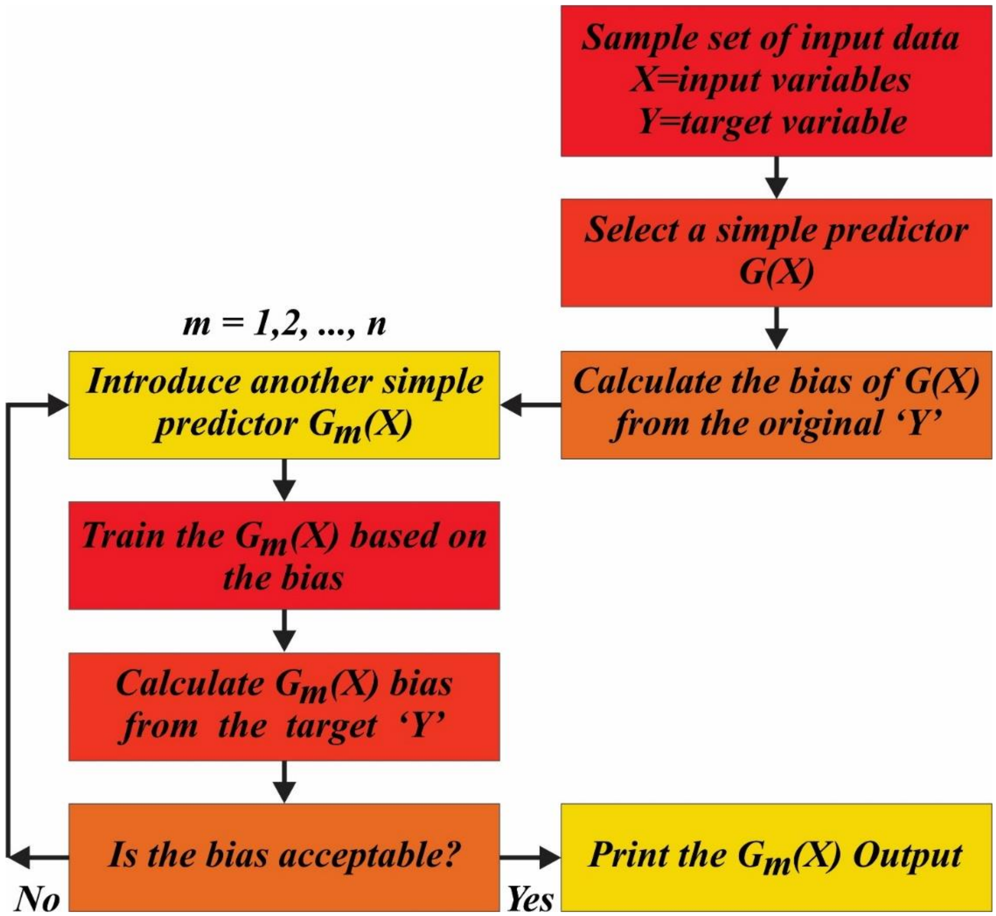

Figure 2: Flowchart of Gradient Boosting Regression.

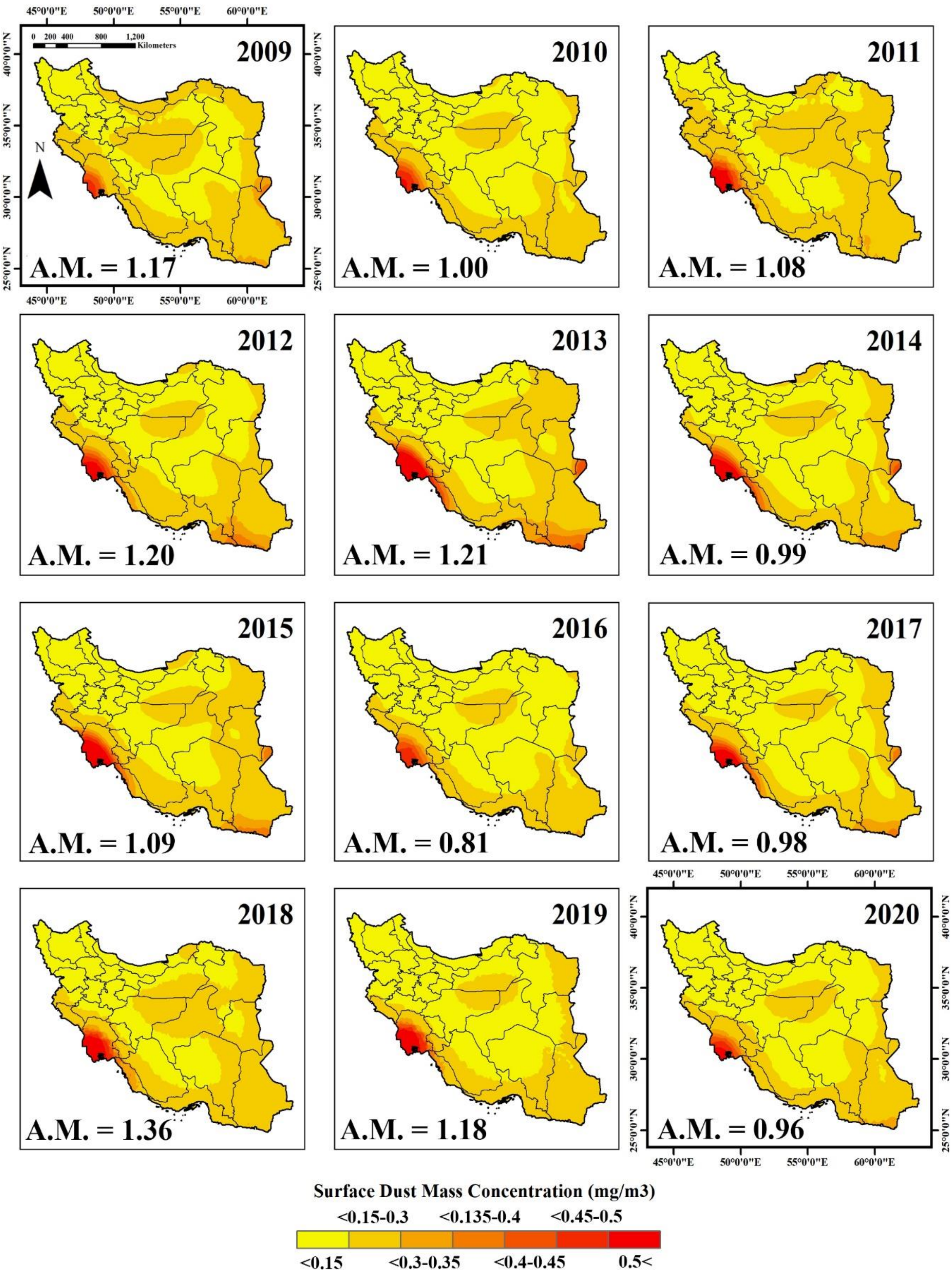

Figure 3: Yearly maximum Surface Dust Mass Concentration (SDMC) between 2009 and 2020 in Iran.

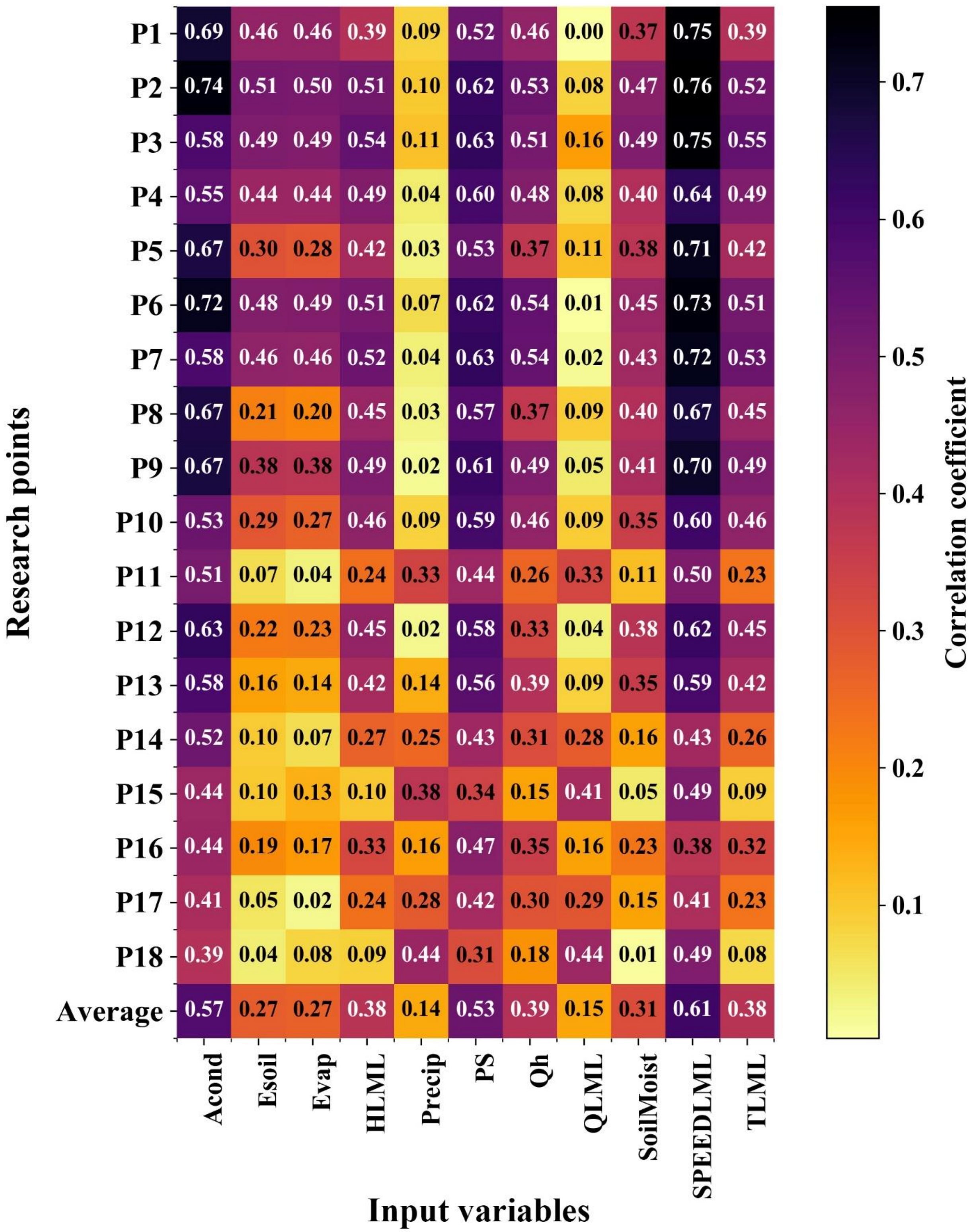

Figure 4: Heat map of Pearson's Correlation Coefficient (PCC) between the 11 input variables considered and daily SDMC for the 18 Research Points.

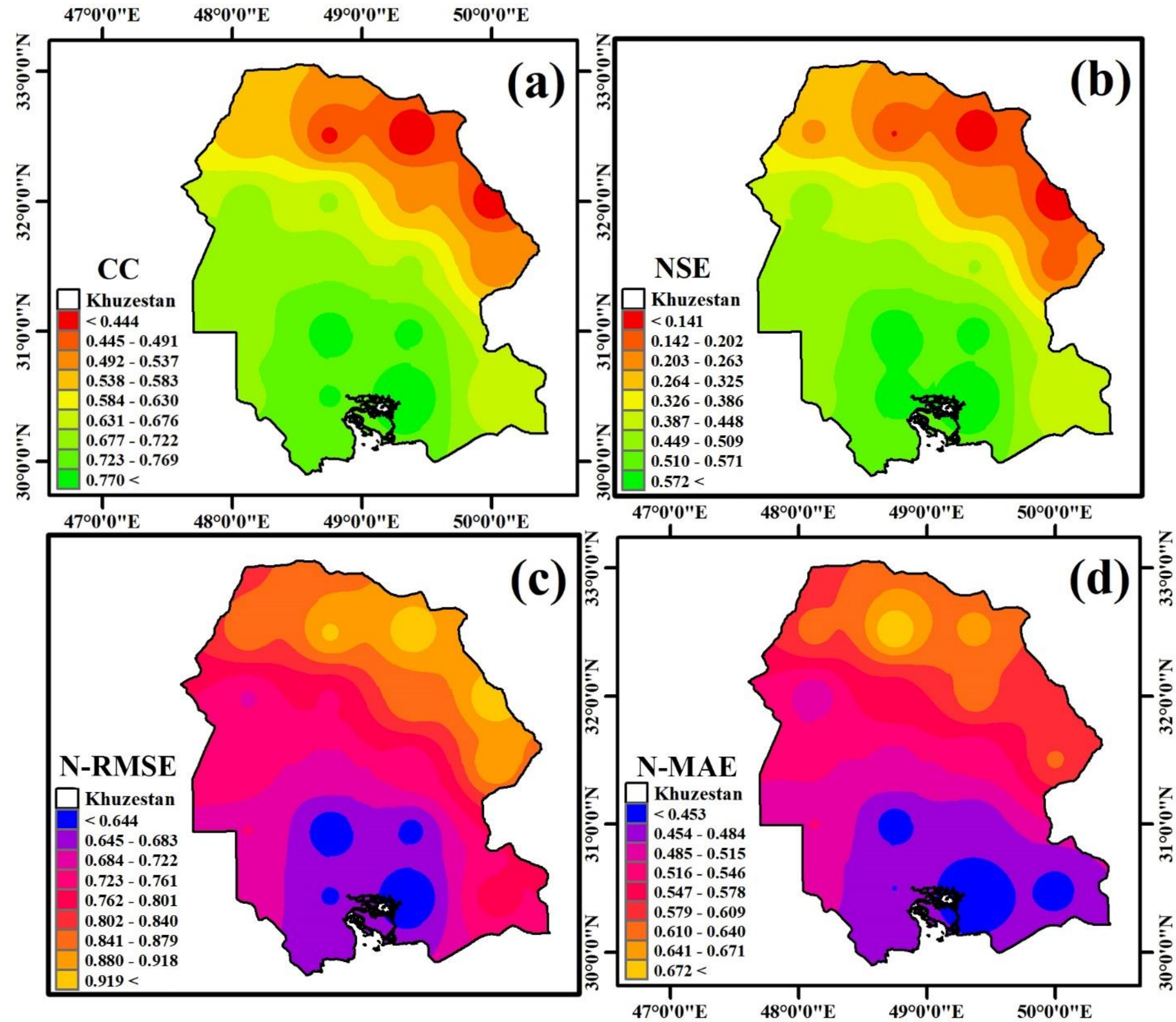

Figure 5: SDMC real-time prediction performance over the Khuzestan province.

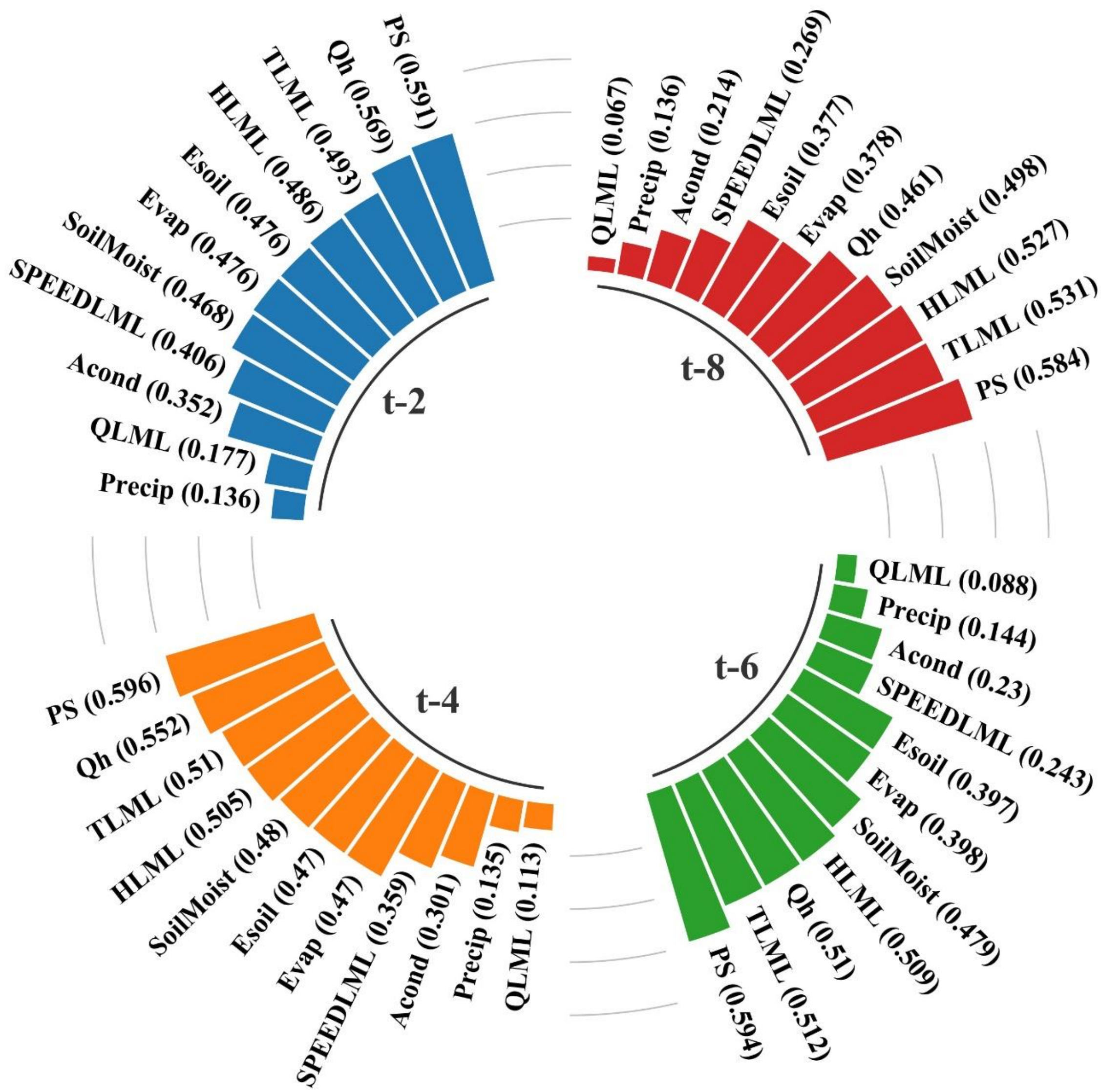

Figure 6: Pearson's Correlation Coefficient for correlation between the 11 input variables considered and SDMC for different lead times.

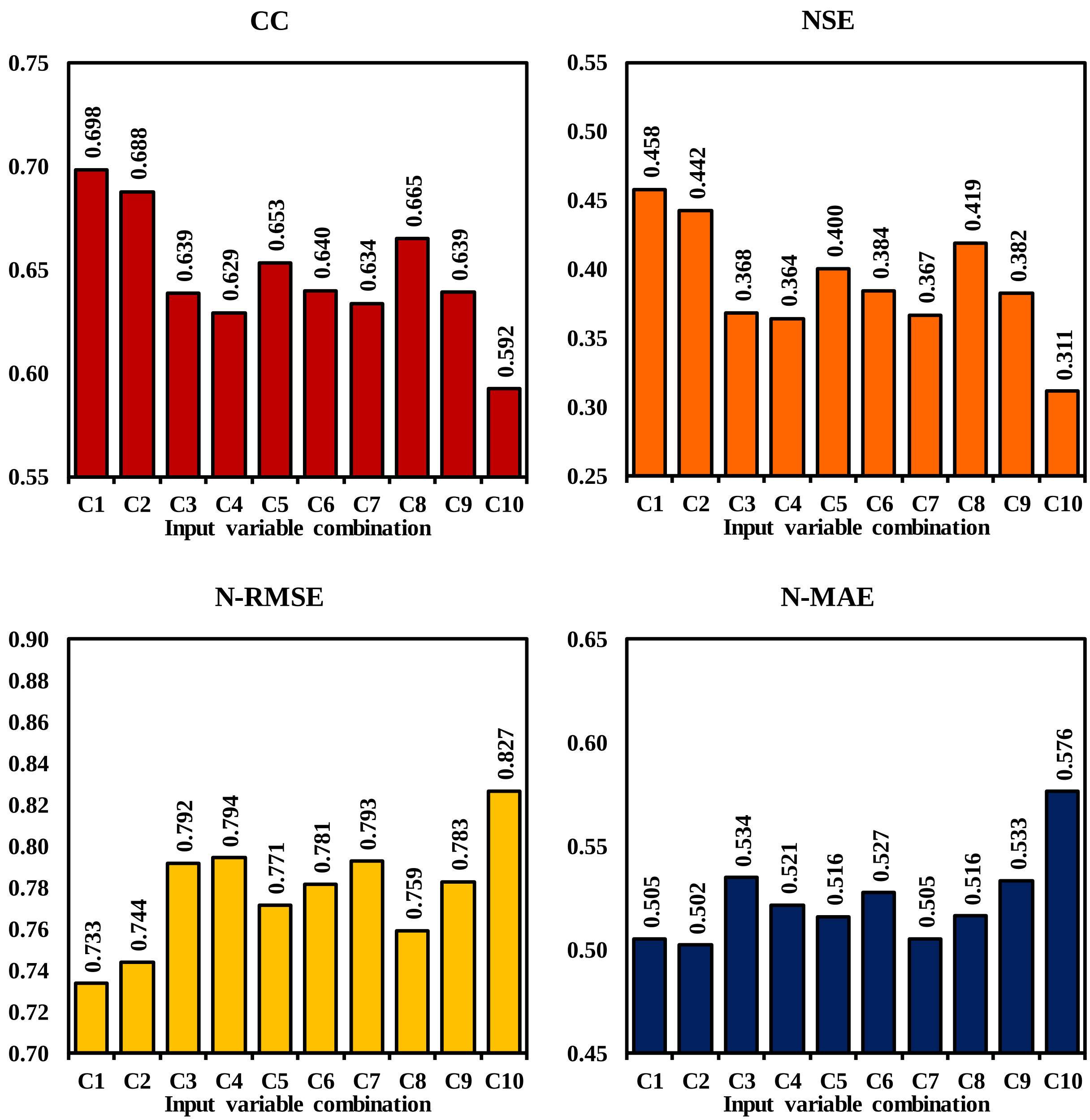

Figure 7: Prediction performance for future forecasting of SDMC for the ten input variable combinations considered.