Projection of future forest fire events using remote sensing sources and multi-label ensemble classification algorithm: Case study Hircanian forest, Golestan, Iran

Abstract

Forest fire is considered as an environmental hazard which threats the various regions across the world. The frequent occurrence of fire incidents in forested areas at northern Iran forced significant socioeconomic casualties. In this study, a machine learning-based simulation and forecasting approach was applied on forest fire incidents in the Golestan province of Iran. The fire anomalies products of MODIS-MOD14A2 between the 2000- 2020 were used as the historical data. Two main MOD14A2 variables known as fire incidents with medium- (MC) and high-confidence (HC) were considered for further investigations. Based on 12 critical observation dates total of 915 study points have been investigated. Nine climate products of NASA's FLDAS satellite were used as the input variables. Three principal machine learning ensemble classifiers including Gradient Boosing (GBC), Random Forest (RFC) and Extremely randomized tree (ETC) algorithms were utilized and their classification accuracy in forest fire simulation have been measured. The ETC with AUC of 0.919 proved to be best classifier compare to its alternatives. Parameter sensitivity analysis also show that the combination of specific humidity, soil moisture content, net downward shortwave flux and air temperature is the most optimal input combination. In forecasting phase, first the **MIROC06-SSP370** prove to be the best historical period simulators of 4 optimal parameters. The products of these GCMs were used to forecast the fire events for the near-future period of (2030-2050). The projections reveal that the summer months are the crucial period of the year. Also, the Augusts by having mean fire event of 383.65 and 141.05 in MC and HC levels, respectively, proved to be the most critical month. Furthermore, the overall trend of forest fire occurrences shows both of decreasing and increasing pattern in MC and HC levels, respectively. The suggested projection method can be an effective tool in preventing occurrence of fire events and also can be a little contribution to preserve the valuable forest ecosystems.

1. Introduction

Forests, as one the most essential ecosystem elements, have an undeniable regulatory rule in decreasing climate change effects, managing carbon cycle as well as preserving the soil and water (Keenan et al., 2015; Morales-Hidalgo et al., 2015). Fire, as a hazardous factor, directly affects the hydrological and geomorphological characteristics of a region and also significantly affects the socioeconomic aspects of human life (Diakakis et al., 2017; Thom and Seidl, 2016). Due to higher impact of fire on biomass components of the ecosystems, it is considered as one of the predominant threats in the forest ecology (Mackey and Lindesay, 2002; Pastro et al., 2011).

In recent years, global warming increased the temperature in different regions across the globe, which make the susceptibility of forest fire occurrence more significant than before (Kasischke et al., 1995; Miranda et al., 1994). Beside by deteriorating the soil and vegetation conditions, the forest fire will further lead to faster change of regional climate condition due to the emission of aerosols and gases such as carbon dioxide (Oris et al., 2014). So, evaluating the causes of forest fire as well as its prediction have become one the focuses of environmental researchers in the last decade. However, the studies show that the forest fire is a complex phenomenon since it affected by numerous factors such as topography, weather, socioeconomic and vegetation condition (Avila-Flores et al., 2010; Guo et al., 2016).

Employing Artificial Intelligence (AI) approaches have become extremely popular in environmental sciences (Moghadam et al., 2021; Sharafati et al., 2021). The application of classic AI algorithms such as Artificial Neural Networks (ANNs) and Support Vector Machine (SVM) on prediction and preventation of forest fire occurrences are also reported in various literatures (Langford et al., 2018; Wang et al., 2009; Zhao et al., 2011). In a related study, (Soliman et al., 2010) proposed a smart early detection system based on ANN using temperature, smoke and light field data as inputs. A network of wireless sensors was responsible to observe the ground-based data and feed them to the ANN for the further process. (Sakr et al., 2011) compare the performance of ANN monthly based forest fire incidents prediction with SVM algorithm. Various parameters including minimum and maximum temperature, humidity, solar radiation, wind speed and precipitation were utilized. While the SVM performs better in binary classification, the ANN show better accuracy in multi-label categorization. (Satir et al., 2016) applied Multi-layer perceptron (MLP) on the ANN structure to predict the fire probability in Upper Seyhan Basin forest in Turkey. With accuracy coefficient of 0.83 the ANN-MLP was proved to be efficient predictive tool. Based on their analysis the landscape components such as tree cover, elevation and temperature were highly correlated with occurrence of fire incident in the case study area. (Bui et al., 2018) introduced a new hybrid AI that combines the ANN with mini-match backpropagation (MMBP) and Differential Flower Pollination (DEP) algorithms. The DEP algorithm was responsible for calibrating the ANN based on a database from Geographical Information System (GIS), while the MMBP carried out the spatial prediction of fire events. (Sevinc et al., 2020) developed a Bayesian network algorithm to forecast the possible origins of forest fire at Mugla Forestry area in Turkey between 2008 and 2018. Based on the multi-variable analysis, the month and temperature are the most important factors, respectively. Also, the human-induced origin was proved to have the highest rate compare to natural causes. (Zheng et al., 2020) mapped the regional burn severity caused by forest fire, using hybrid structure of semi-supervised Support Vector Regression (SSTCA-SVR) algorithm. The field-based survey observations were employed as the algorithm inputs and the SSTCA-SVR performance was proved to be marginally accurate over a diverse region across southwestern United States.

While these benchmark algorithms show acceptable performance, a newer branch of AI algorithms known as Ensemble models prove to have higher accuracy and lower deficiencies (Sharafati et al., 2020b, 2020a). These Ensemble models, which categorized as one the machine learning algorithms, utilized as the main predictive tool on diverse type of scientific fields and achieve excellent metrics (Asadollah et al., 2021a; Saha et al., 2021; Zhou et al., 2021). The application of ensemble algorithms in prediction of forest fire also proved to be promising. For example, (Sachdeva et al., 2018) evaluate the application of optimized Gradient Boosting (named as EO-GBDT), as one the most utilized ensemble techniques, in mapping forest fire possibility. Total of 18 parameters with different natures were selected as the input of EO-GBDT predictive model and forest fire was predicted on 702 locations across Chamoli, Bageshwar and Pithoragarh districts of India. Based on 10 accuracy metrics, the proposed EO-GBDT outperform other benchmark algorithm such as particle swarm optimized SVM, Neural Network (NN), decision trees (DT) and Logistic Regression (LR). In another related study, (Gigović et al., 2019) evaluate the applicability of an ensemble algorithm known as LogitBoost classifier in spatial simulation of forest fire. The Kernel logistic regression (KLR), RF and SVM were considered as the benchmark algorithms used for performance comparison. Based on several accuracy metrics such as area under the curve (AUC) and Kappa index LogitBoost outperformed its alternatives. The utilization of Ada-boost as another principal ensemble algorithm has also been investigated on several literatures and show superiority compare to classic AI algorithms (Rosadi et al., 2021, 2020). The utilization of Random Forest (RF), as another main ensemble approach, in term of predicting the forest fire are evidently more frequent than other ensemble algorithms. For example, in a comprehensive study (Mohajane et al., 2021) evaluated the RF hybridized with Frequently Ratio (FR) against four other robust AI technique, namely Frequency Ratio-Classification and Regression Tree (FR-CART), FR-MLP, FR-LR and FR-SVM in case of mapping susceptibility of forest fire. Considering north of Morocco as the case study, total of 510 study points were investigated using 10 fire-related variables. Based on AUC metric the FR-RFR outperformed all of its alternatives. Overall, these findings reveal that the application of Ensemble models will lead to better regression and classification results compare to their classical alternatives.

While the predictive algorithm is one of the essential components of AI-based researches, the desired performance is not reachable without the presence of quality input data (Cortes et al., 1994; Gudivada et al., 2017; Jain et al., 2020). Acquiring high quality ground-based hydrological and geological dataset is proved to be a very hard task in many regions across the world. To overcome this, utilization of Remote Sensing source such as satellite imagery obtained is become highly on demand. The studied related to forest fire are not excluded from this matter and satellite-based observations are significantly helps the detection and prevention of forest fire occurrences (Barmpoutis et al., 2020; Chuvieco and Congalton, 1989; Sunar and Özkan, 2001).

The Moderate Resolution Imaging Spectro radiometer (MODIS), Visible Infrared Imaging Radiometer Suite (VIIRS), Meteosat Second Generation- Spinning Enhanced Visible and InfraRed Imager (MSG-SEVIRI), Landsat 8 OLI, National Oceanic and Atmospheric Administration-Advanced Very High Resolution Radiometer (NOAA-AVHRR) and Worldview-2 satellite images are some of the most utilized satellite in case of monitoring the forest fire (Hua and Shao, 2017).

From all these sources the MODIS was proved to be the most utilized satellite source for predicting the forest fire in different regions of the world. For example, (Masocha et al., 2018) evaluate two fire products of MODIS satellite, e MCD14ML and MOD14A1, to detect the occurrence of fire in savannah regions over the Zimbabwe. (Shahramanyan et al., 2019) studied the impact of forest fire on economical aspects using MODIS satellite products. Using the images from MODIS satellite they first identify the burned area and then calculate the corresponding physical damages. (Kumari and Pandey, 2020) evaluated the relation between MODIS-based forest fire and several climate variables such as rainfall, solar radiation, land surface temperature and relative humidity. Their analysis indicated that from all these parameters the temperature has the highest relevancy with forest fire probability. (Fernandes et al., 2020) used rainfall and vegetation health index (VHI) parameters to forecast the fire events over Amazon forest in South America. The fire anomalies related products of MOD14A2 between the years of 2000 to 2019 were used as the target of the prediction. Their proposed forecasting methodology was able to detect the incidents of forest fire with time steps of 1 to 2 weeks prior to actual occurrence. (Shi et al., 2021) investigate the MCD64A1 biomass burning in several land types across south and southeast of Asia between 2001 to 2017. the results reveals that the fire events related to forest as well as woody savannas areas are the primary heat anomaly suppliers which dominate the emission characteristics.

MODIS product also proved to be useful inputs for AI algorithms in order to predict the possibility of fire incidents in forested regions. (Langford et al., 2018) used the Deep neural networks (DNN) to perform a binary classification of fire incidents over Alaska for the 2004 wildfire year. They considered a weight-selection approach for the DNN and applied it to several MODIS products to determine whether a remote sensing pixel has wildfire or not. By validating the DNN using ground-based fire observations, they proposed model with overall recall score of 0.96 proved to be highly efficient in prediction of 78,702 satellite-based pixels. In another study, (Maeda et al., 2009) forecast the Amazon forest regions considered to have higher risk of occurrences utilizing the remote images from MODIS as well as the ANN. The Feedforward structure with different setups has been considered as the principal algorithm. Also, several Studies focused on performance of Ensemble algorithms in forest fire prediction using MODIS products as the input. (Gigović et al., 2019) proposed a novel ensemble algorithm structure based on SVM and RF to measure the forest fire possibility in Serbia's Tara National Park. A comprehensive dataset was established based on MODIS satellite-, field- and aerial images-observations. Using the area under the curve (AUC) metric the Ensemble algorithm was proved to outperform the stand-alone SVM and RFR. Their results also reveal that the land slope, normalized dierence vegetation index (NDVI), soil characteristics as well as temperature are the most influential parameters. (Ma et al., 2020) evaluates the effectiveness of several parameters on MODIS's forest fire products using RF algorithm between the 2010 to 2016 over six different climate regions of China. Total of 21 different parameters which categorized under four main types of climatic, vegetation, topographic and socioeconomic have been investigated. Based on the outputs, the climatic and vegetation indicators were proved to be an important factor in all of six regions, while the other two factors were less efficient. From the climatic variables the soil moisture, temperature and humidity are proved the be the most influential parameters. The random forest also proved to be an excellent analytical tool in term of measuring the characteristics of forest fire events.

From previous sections it become apparent that the on real-time prediction of forest fire was the center of researchers concentration, however, have a concept of possible future fire incidents in forestry area seems a necessity. While futuristic occurrence of fire event can significantly help in preventing and managing the fire incidents, to the authors knowledge there are very few literatures that focused on this subject. (Busico et al., 2019) utilized the GIS and analytic hierarchy process (AHP) to forecast the forest fire risk up to 2024 in Southern section of Italy. In another related study, (Busico et al., 2019) employed the Community Land Model (CLM) to forecast the deforestation caused by draught and fire by 2047 as the near future horizon.

Iran is one of the countries that evidently sustained by severe negative impacts of climate change in recent decade (Abbaspour et al., 2009; Zarghami et al., 2011). One of the major concerns, which directly affected by global warming, is the frequent occurrences of forest fire specially in northern section of Iran (Najafabadi et al., 2015; Shafiei et al., 2010). Based on literatures, from all of northern provinces of Iran, the Golestan is the one that records the highest events regarding the forest fire (Adab et al., 2013; Eskandari, 2017; Pourtaghi et al., 2015).

This study aims to first, simulate the historical forest fire incidents in Golestan province and then project the possibility of future forest fire occurrence in the same region. To achieve this goal, several remote sensing sources including MODIS thermal anomalies and fire detection as well as 10 near surface climate parameters from the FLDAS satellite observation were employed as the input variables of classification tools. Three machine learning algorithms, namely GB, RF and extremely randomized tree were used as the ensemble classifiers. Next, as one the novelty of this study, the most optimal climate variables were extracted and subsequently utilized to forecast the forest fire probability in the future using ESGF's CMIP6 projections. It reveals that the proposed methodology can accurately forecast the future fire incidents by validating the projection algorithm with quality historical data. The results from this study can significantly aid the future forestry policy makers as well as work as a complementary tool in detection and management of future forest fire incidents.

2. Study area

Large quantity of fire occurrence in forested regions of Iran are recorded every year. Forest fires are mostly seen in northern, north-eastern and western sections of Iran that are mainly covered by burnable trees and savannas (Eskandari and Chuvieco, 2015). Manifestation of this hazardous phenomenon, force massive ecological and economical damages to the local regions, which subsequently require immediate managements measures. The Hyrcanian forest located in province of Golestan, is one of the regions which sustains from massive forest fire events specially in recent decade. Figure 1 show the Golestan province as well as the location of Hyrcanian forest.

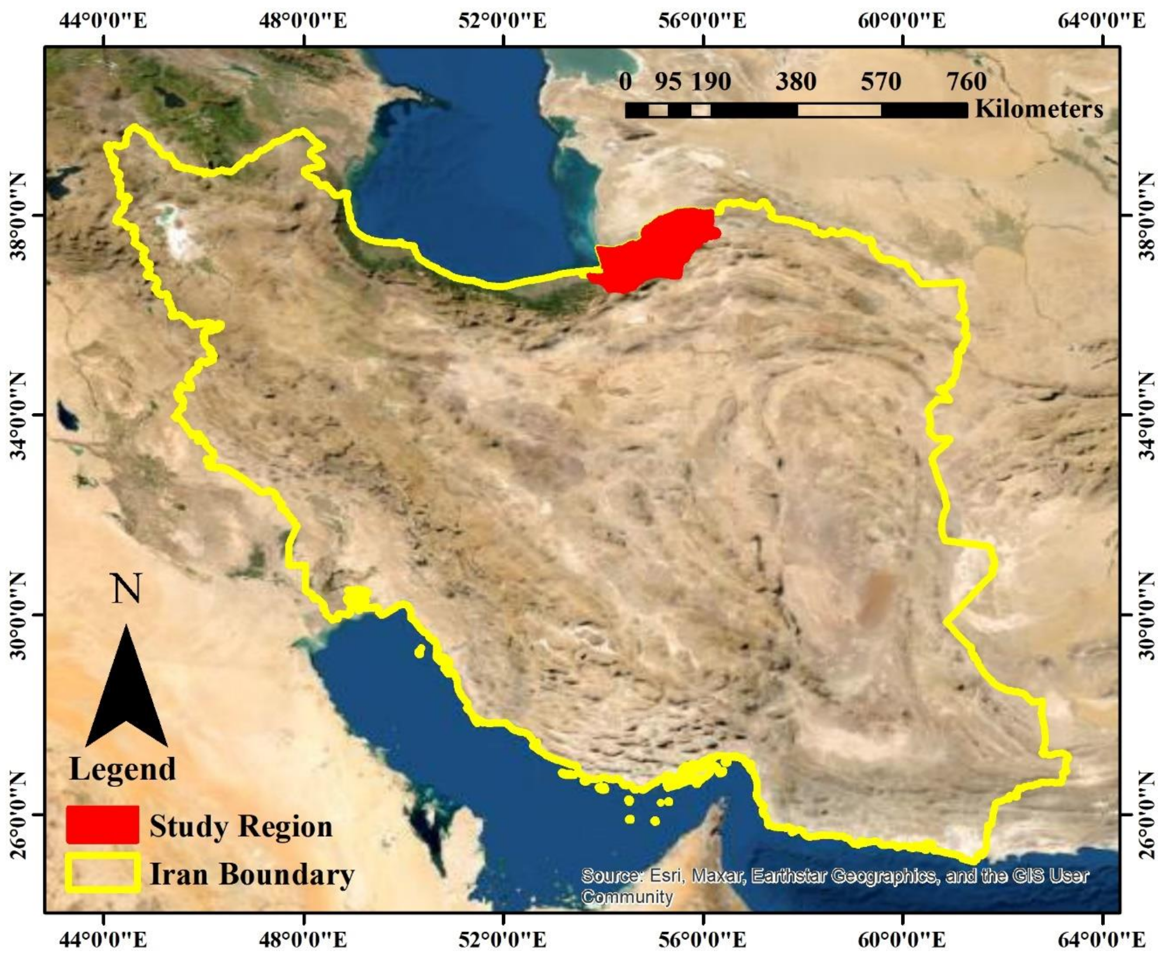


Figure 1: Location of Golestan province (Red indicated region) in Iran.

This forest, which is stretches nearly 850 kilometers along south-eastern margins of Caspain sea, is the home of nearly 180 bird and 58 mammal species. In the 2019, the Hyrcanian forest is inscribed as one the Iran's representatives in The United Nations Educational, Scientific and Cultural Organization (UNESCO) world heritage convention. The historical records show that most of fire event in this forest is taken place between August to the end of December due to the high wind speeds and low relative humidity (Adab, 2017; Adab et al., 2021). While based on the figure 1 the forest fires are concentrated on southern sections of Golestan province, the northern section also experience land fire events cause by low rainfall rate as well as high temperature (Veblen et al., 2000). The Golestan province has elevation range of 64 to 3828 meter above sea level and average annual rainfall rate of 450 millimeters. This province also includes both arid and semi-arid weather conditions as well as temperate mountainous which make it a region with diverse climatic situation (Eskandari et al., 2020).

3. Materials and Methods

3.1. Description of remote sensing sources

3.1.1. FLDAS

In current study, the land surface products of Famine Early Warning Systems Network (FEWS NET) Land Data Assimilation System (FLDAS) was utilized as the classifier inputs. The FLDAS is a joint project between NASA Goddard Space Flight Center (GSFC), U.S. Topographical Survey (USGS) Earth Resources Observation and Science (EROS) and University of California Santa Barbara (UCSB) Climate Hazards Group (CHG). The employed "FLDAS_NOAH001_G_CA_D" includes 23 variables modeled from the NOAH 3.6.1 and Land Information System (LIS7). One of the major advantage of this satellite, which makes it more suitable for the subject of this study, is its particular concentration on Central Asia region with specific spatial coverage of $21.0°\sim30.0°E$ and $56.0°\sim100.0°N$. FLDAS also benefits from very fine spatial resolution of 0.01° and daily temporal resolution which stretches from 2000-10-01 to the present day (https://disc.gsfc.nasa.gov/datasets/FLDAS_NOAH001_G_CA_D_001/summary).

3.1.2. MODIS

Moderate Resolution Imaging Spectroradiometer (MODIS) is a tool places on Aqua (EOS PM-1) and Terra (EOS AM-1) satellites. While the Aqua orbits around the earth by a horizontal manner, the Terra moves vertically crossing the equator. Terra- and Aque-MODIS are observing the earth bodies surface every 1 to 2 days gathering data in 36 spectral bands. The provided dataset will enhance the understanding of earth dynamics and occurring mechanisms on land, ocean and lower section of atmosphere. From various MODIS products, the thermal anomalies and fire masks are mainly obtained from MODIS-4 and -11 micro-meter radiances. The strategy of fire detection is based on unconditional discovery of fire events with sufficient strength, as well as finding anomalies in variables such as surface air temperature and sunlight reflection. MODIS applicability has been significantly developed to that the false alarms can be identified and rejected. There are total of two principal MODIS products, namely MOD14A2 and MYD14A2, that dedicated to active fire detection, however this study evaluates the application of former source. Products of MOD14A2, which consists of fire-mask and quality affirmation algorithm, has the spatial and temporal resolution of 1-kilometer and 8 days and provides gridded level-3 data in the Sinusoidal projection. Collection 5 of MODIS-Terra fire incidents are stage three validated, noting that the products uncertainties are adequately measured and resolve using proper references such as ground-based data.

3.1.3. Climate Projection Model Infrastructure phase 6 (CMIP6)

As one of the earth science researching infrastructures, the Earth System Grid Federation (ESGF) is responsible for archiving the climate models observation. The concept of ESGF is first established by Department of Energy (DOE) and gain further collaboration from other agencies such as NOAA, NSF and

NASA. ESGF's main objective is to store the values of weather components as well as publish the Coupled Model Intercomparison Projects (CMIPs) forecasts which extracted from General Circulation Models (GCM) observations (Covey et al., 2003; Meehl et al., 2005). The GCMs measurement in CMIP6, which is currently the latest phase, are evidently less prone to uncertainties and bias compare to the former CMIPs. Beside from higher temporal and spatial resolutions, the sixth phase of CMIP also includes more models sources, experiment ID's, variant labels and climate parameters (Asadollah et al., 2021b; Venkatramani Balaji et al., 2018). One of the major differences between CMIP6 and it former alternatives is the utilization of Shared socio-economic pathways (SSPs) instead of Representative Concentration Pathways (RCPs). While the RCPs only consider greenhouse gases and other sun radiation-related factors, the SSPs also considered socioeconomic factor in its future emissions. Based on the concentration of greenhouse gasses the RCPs are focused on level of global warming, whereas the SSPs are mainly fixated on measuring the amount of emissions reduction (Bhati and Rai, 2020; Sharaff and Gupta, 2019). In this study, the efficiency of different CMIP6 sources with various SSP scenarios on simulation and forecasting of forest fire have been investigated. The climate data were downloaded from ESGF exclusive CMIP6 platform (https://esgf-node.llnl.gov/search/cmip6/).

3.2. Ensemble algorithms

3.2.1. Gradient Boosting Classifier (GBC)

Since in this study, the Gradient boosting operates based on DT as simple learners, a short description about the DT seemed as necessity. DTs are considered as the most basic ML techniques which can handle both regression and classification problems. DT operates based on branching progression and recurrent partitioning. Each tree consists of a root node that followed by branches and eventually lead to final decision known as the leaf or terminal node. By utilizing the bagging or boosting procedure, ensemble techniques are tended to intensify the accuracy of DT models. In boosting, the simple models are built sequential manner meaning that each model trained based on errors of previous models. In GBC, a combination of DTs is developed and formed a set of weak learners. The first tree in that combination tries to predict the first label, and the output is compared with original label. By calculating the bias of first prediction, the second tree is trained so that it can reduced this bias. This iterative procedure continues until the level of bias satisfies the criteria (Livingston, 2005).

In order to compare the results of GBC, two other widely used machine learning algorithms known as Random Forest and Extra Tree classifiers are employed for comparison purposes only. To apply these three ensemble algorithms, the Scikit-Learn (SKlearn) machine learning library of Python programming language has been employed. The ensemble sub-category of SKlearn provides various regression and classification algorithm which operates based on tuning of several parameters. **Table 1** shows the name of tuning parameters of GBC, RFC and ETC, their description, default and optimal values.

Table 1: Default and optimized tuning values of GBC, RFC and ETC ensemble classification algorithms.

Variables	Default value	GBC Optimized values	RFC Optimized values	ETC Optimized values
N_Estimator	100	24	5	7
Max_Depth	3	6	11	13
Min_Samples_Split	2	33	2	2
Min_Samples_Leaf	2	15	1	1
Random_State	None	5	0	92
Learning_Rate	0.1	0.71	--	--

4. Results and discussion

The main course of this study is focused on two major phases of simulation and Projection. In the first stage, the corresponding classification algorithm is validated using the remote sensing data from the historical period, while, the second phase concentrates on forecasting the susceptibility of future fire events based on the calibrated algorithm and projected product of global climate models.

The Fire-mask products of MOD14A2 were extracted over the Golestan province of Iran for period of 2000 to 2021. As mentioned in previous section, MOD14A2 measures the level of fire incidents every eight days so there were total of 997 observations in selected temporal range. The product classifies every pixel with a number between 3 to 9 which denote no fire and fire with high certainty. Table 2 show these numerical indicators, their corresponding description and maximum value of occurrence per 997 days of observation.

Table 2: The numerical description of MODIS- products as well as their description.

Numerical indicators	Description	Maximum occurrence per 997 event
3	Non-fire water pixel	677
4	Cloud (land or water)	21747
5	Non-fire land pixel	30138
6	Unknown (land or water)	260
7	Fire (low confidence, land or water)	39
8	Fire (medium confidence, land or water)	149
9	Fire (high confidence, land or water)	49

By having total of 4189 and 796 records of fire events with medium and high confidence, this study only considered the 8 and 9 indicators as the target of classification task. Other indicators (e.g. 3, 4, 6 and 7) were eliminated due to their irrelevancy to our region of study or inadequate number of observation between 2000 to 2021. Also, the points with 5 numerical value which indicates the No Land Fire (NLF) situation, has been considered in this study to give the multi-label classification more legitimacy. Besides from fire level classification, the selection of NLF points will also enable this study to further investigate the fire/ no-fire conditions.

4.1. Selection of the date with most fire events

Figure 2 depicts the annual-based aggregation of fires with medium confidence (MC) and high confidence (HC) as well as their average and 3rd quartile ($Q75\%$) values. Based on the statistics from MC ($Average = 190.227$ and $Q75\% = 248.75$) and HC ($Average = 36.182$ and $Q75\% = 58.25$), it became apparent that the years of 2003, 2006, 2010 and 2016 have simultaneously higher recorded fire events than $Q75\%$ as a criteria. Also, the year 2004 and 2002 respectively show higher rate of fire occurrences in specific levels of medium and high confidence. Meanwhile, the year 2010 with approximately 122% and 81% higher recorded annual fire incidents in MC and HC, respectively, is the most critical fire-year in the Golestan province.

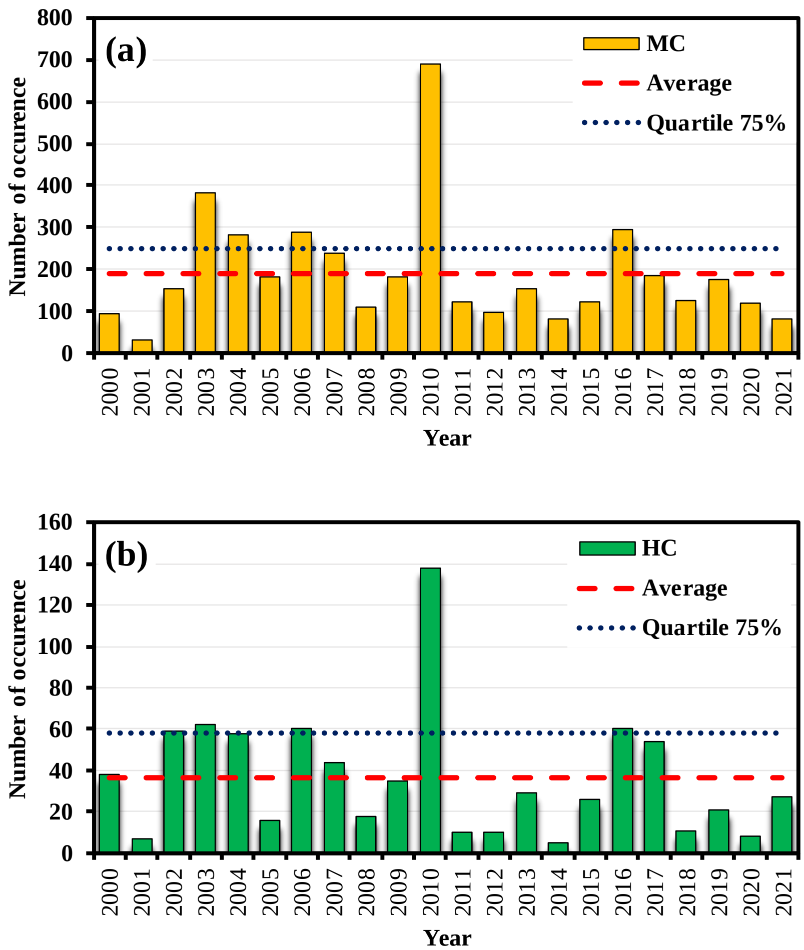


Figure 2: Bar-chart of annual fire incidents in province of Golestan, Iran, with a) medium and b) high occurrence confidences.

Based on these identified critical years, total of 900 randomly selected points, which equally extracted from MC, HC and NLF, were selected and used as the study points. These 900 points were chosen from 12 different dates between 2002 and 2016 and proven to be randomly distributed across the Golestan province.

4.2. Extraction of climate variables

Since MOD14A2 has resolution of 0.01 degree, a satellite product with similar resolution must be selected so that the classification task could be performed in its most optimal condition. To achieve this task, the remote sensing sources from NASA's GES DISC platform have been evaluated. Based on tuning the spatial and temporal resolution to the most similar with the MOD14A2, it becomes apparent that the FLDAS Noah Land Surface Model which concentrates on the central Asia is the best remote sensing source. The FLDAS is the only satellite in the GES DISC which has the spatial resolution of $0.01° \times 0.01°$ and also contains various climate and geological parameters. While there were other sources which have nearly similar spatial resolution, their representative products were irrelevant to the subject of this study so their utilization has not been considered. Table 3 demonstrate the name of FLDAS products as well as their description and units.

Table 3: The FLDAS employed climate parameters, their corresponding indicator and units.

No.	Value indicator	Long name	Unit
1	Evap	Total evapotranspiration	$kg/m^2.s$
2	Psurf	Surface pressure	Pa
3	Qair	Specific humidity	$kg/m^2.s$
4	Qsb	Surface runoff	$kg/m^2.s$
5	Radt	Surface radiative temperature	K
6	SoilMoi	Soil moisture content	$kg/m^2.s$
7	SoilTemp	Soil temperature	K
8	Swnet	Surface net downward shortwave radiation flux	W/m^2
9	Tair	Air temperature	K

4.3. Point-based classification of Fire events

By having the longitude and latitude of selected 900 points from 12 chosen dates, the values of 9 FLDAS productsfor each point have been extracted and structured an input Dataframe. In next stage, the established Dataframe is used as the input of three noted Ensemble algorithms and the classification task has been carried out. Total of 900 points were randomly distributed between the training and testing phases with corresponding ratio of 0.75 to 0.25, respectively. Table 4 show the overall classification accuracy of these three Ensemble algorithms based on four highly utilized metrics.

Table 4: The overall performance of GBC, RFC and ETC based on five different classification metrics.

Phase	Classifiers	Precision	Recall	F1-score	Area Under the Curve (AUC)
Testing	GBC	0.773	0.771	0.772	0.907
	RFC	0.674	0.674	0.674	0.861
	ETC	0.804	0.802	0.802	0.917

From the table, it is apparent that the simulation performance in training phase is significantly better than the testing. While the GBC shows an slightly better classification metrics in the training phase, it has been replaced by the Extremely randomized Tree algorithm in the testing phase with $Precision_{testing} = 0 \cdot 782.$ $Recall_{testing} = 0 \cdot 80.$ $F1S_{testing} = 0 \cdot 780\ and\ AUC_{testing} = 0.919$. Overall the performance of GBC and ETC was highly similar, while the RFC shows the worst result in both phases.

Previous results, which shown in Table 4, are acquired by considering all of the 9 products of FLDAS for simulation of Fire-mask at 900 study points. While the performance metrics are in the satisfactory range, the number of inputs must be reduced to its most minimal value so that the current methodology can be considered as an optimal approach. To achieve this goal, nearly 500 different arrangements of these 9 parameters have been evaluated. Table 5 shows the metric values of 10 top combinations, their corresponding metrics and parameters.

Table 5: Classification metrics and parameters description of top ten input combinations.

Number of parameters	Precision	Recall	F1-S	AUC	Average	Parameters
4	0.831	0.828	0.828	0.918	0.851	[Qair, SoilMoi, Swnet, Tair]
5	0.822	0.819	0.820	0.918	0.845	[Qair, Qsb, SoilMoi, Swnet, Tair]
4	0.823	0.819	0.819	0.911	0.843	[Qair, RadT, SoilMoi, Swnet]
6	0.816	0.815	0.815	0.916	0.840	[Qair, Qsb, SoilMoi, SoilTemp, Swnet, Tair]
6	0.810	0.811	0.810	0.922	0.838	[Psurf, Qair, Qsb, RadT, SoilMoi, Swnet]
6	0.811	0.811	0.810	0.919	0.838	[Evap, Qair, RadT, SoilMoi, Swnet, Tair]
5	0.814	0.811	0.810	0.912	0.837	[Qair, Qsb, RadT, SoilMoi, Swnet]
6	0.811	0.811	0.810	0.916	0.837	[Evap, Qair, Qsb, SoilMoi, Swnet, Tair]
6	0.810	0.811	0.810	0.916	0.837	[Evap, Qair, Qsb, SoilMoi, SoilTemp, Swnet]
4	0.811	0.811	0.810	0.910	0.835	[Qair, SoilMoi, SoilTemp, Swnet]

The results reveal that the inputs arrangement of $[Qair.SoilMoi.Swnet.Tair]$ with metrics of $Precision = 0 \cdot 831.\ Recall = 0 \cdot 828.\ F1S = 0 \cdot 928\ and\ AUC = 0.918$, is the best combination which simultaneously has the close to highest accuracy as well as lowest number of inputs.

4.4. Mapping the FLDAS to CMIP6 sources

In order to find the products from CMIP6's sources which has the highest correlation with FLDAS, the daily climate products from the year 2015 was selected. Based on the revealed optimal combination, the CMIP6 products of "huss", "rsds", "mrso" and "tas" which are representative of specific humidity, total soil moisture content, surface net downward shortwave flux and air temperature in the ESGF system has been utilized. To have better evaluation of future conditions, the GCM products were downloaded based on different types of SSP's (SSP1, 2 and 3) and compared with the FLDAS outputs. Table 6 show the name of the utilized GCM sources, their description as well as spatial resolution.

Table 6: The overall performance of GBC, RFC and ETC based on five different classification metrics.

Source ID	Full name of the institution	Nominal resolution
CanESM5	The Canadian Earth System Model Version 5.0	$100 \times 100\ km$ $500 \times 500\ km$
EC-Earth3-Veg-LR	Europe-wide consortium Earth Observations Version 3.0	$100 \times 100\ km$ $200 \times 200\ km$
GFDL-ESM4	Geophysical Fluid Dynamics Laboratory Earth System Model Version 4.0	$50 \times 50\ km$ $100 \times 100\ km$
INM-CM4 & 5	Institute for Numerical Mathematics Climate Model Version 4.0 & 5.0	$100 \times 100\ km$
MIROC6	Model for Interdisciplinary Research on Climate Version 6.0	$50 \times 50\ km$ $100 \times 100\ km$ $250 \times 250\ km$
MRI-ESM2	Meteorological Research Institute Earth System Model Version 2.0	$100 \times 100\ km$ $200 \times 200\ km$ $500 \times 500\ km$
UKESM1-LL	U.K. Earth System Model Version 1.0	$100 \times 100\ km$ $250 \times 250\ km$

In order to find the relevancy between the products of GCMs and FLDAS, four statistical metrics namely correlation coefficient (R), Root Mean Square Error (RMSE), Mean Absolute Error (MAE) and Nash-Sutcliff Efficiency (NSE) were calculated. Since these metrics are in different scales, first their values are normalized between 0 to 1 which are the indicators of worst to best results, respectively. The normalized values were then employed as the inputs of Heat-map accuracy visualization approach which can be seen in Figure 4.

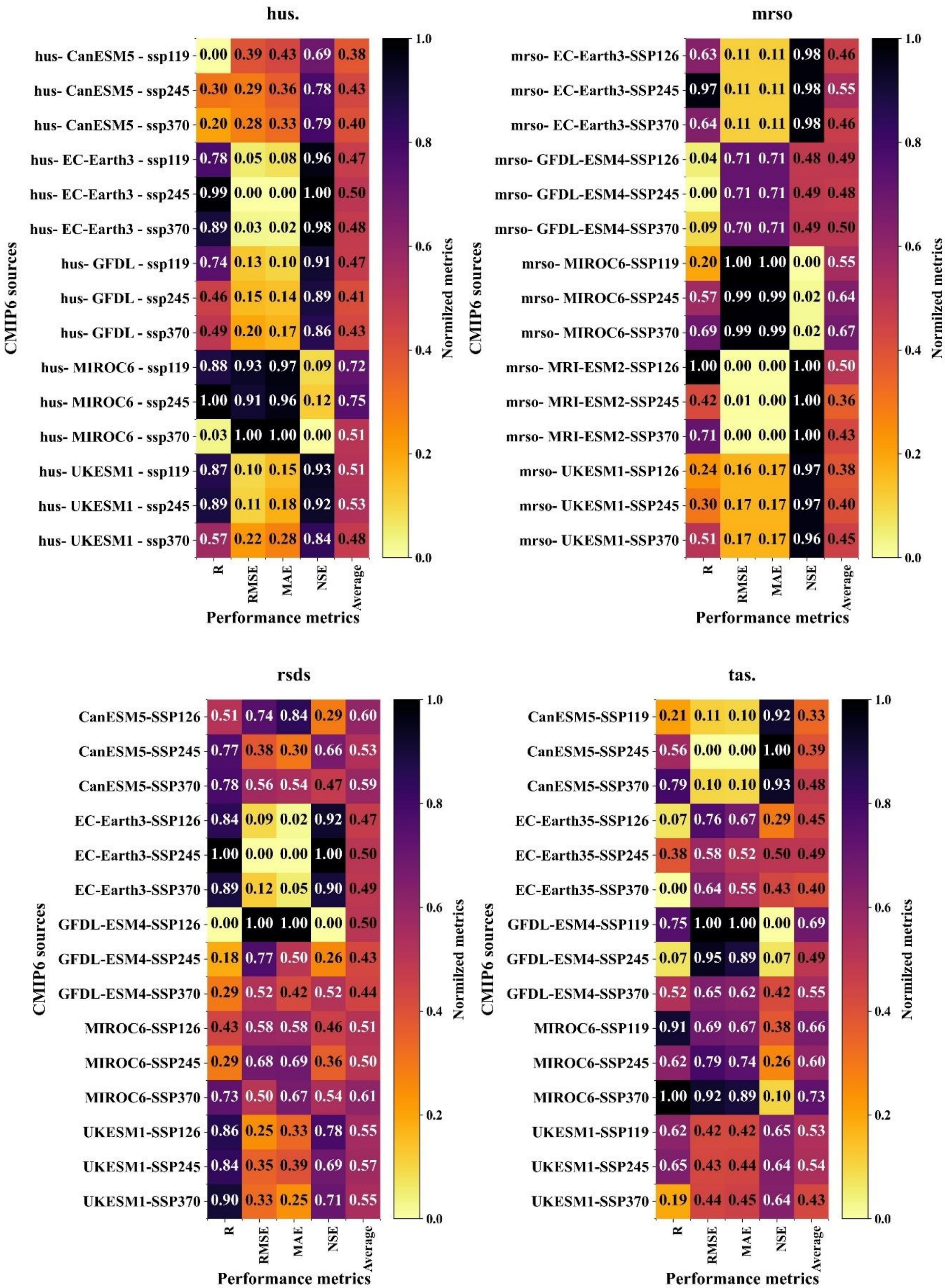


Figure 4: Heat-map of normalized R, RMSE, MAE and NSE performance metrics obtained from evaluating the products of CMIP6 and FLDAS.

Based on the outputs of Heat-maps, the MIROC06-SSP370 was the best simulator of Qair, Swnet, SoilMoi and Tair for the historical period of 2015. The outputs reveal to be in the acceptable range which make them fit for projecting the respected variables in future horizons.

4.5. Future projection of fire events

The findings in previous sections reveals the four optimal parameters and their corresponding best GCMs based on records of historical period. In the next stage, the future projections of “huss”, “rsds”, “mrso” and “tas” for the near future period of 2030-2050 have been extracted and used as the input of trained ETC to forecast the possible future fire occurrences in our study region. To achieve this goal, we only considered the 305 points with the high fire certainty over 12 chosen fire events. From Figure 5 demonstration, it can be understood these points are fairly distributed across the Golestan province which make them good candidate for further evaluation.

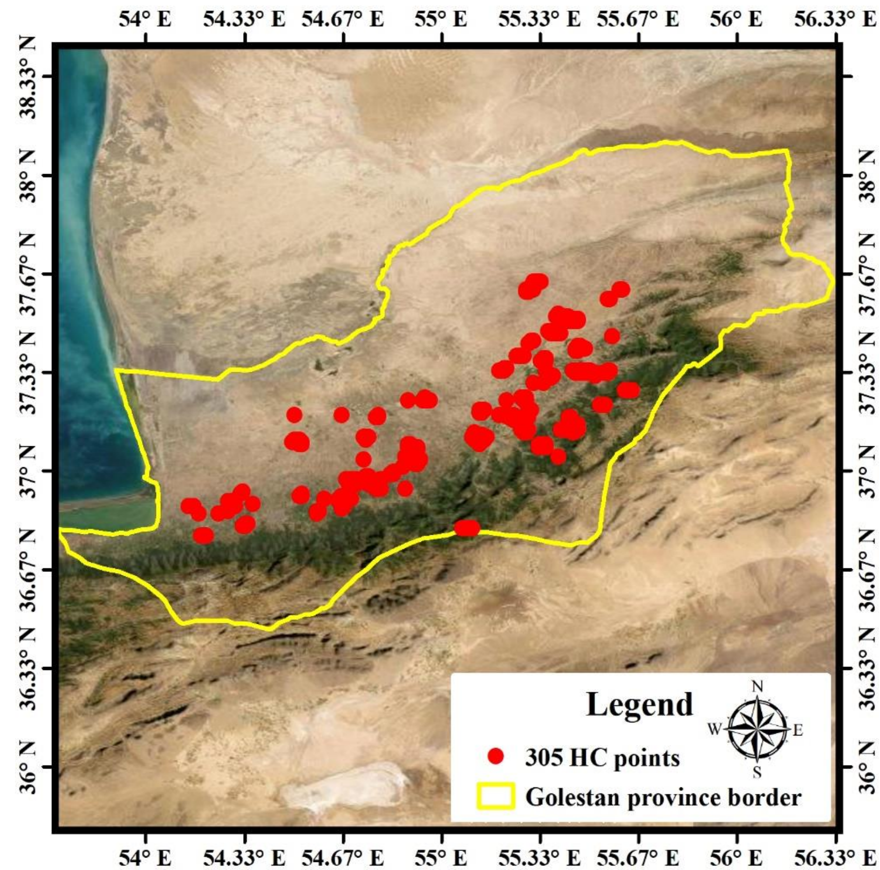
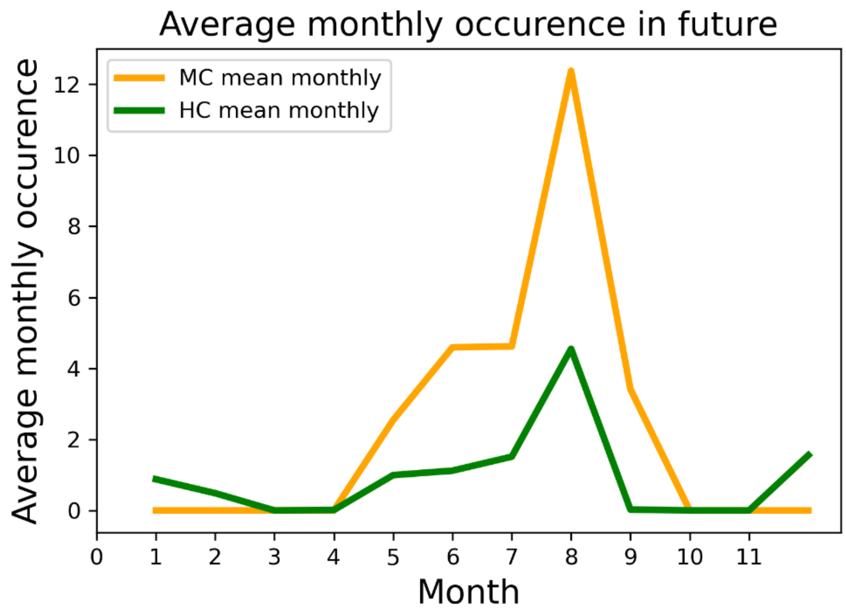


Figure 5: Distribution of 305 points with the high confidence fire events across the Golestan province, Iran.

Based on the daily temporal resolution of CMIP6 products, the ETC forecasting process led to total of 7305 (1st of January 2030 to 30th of December 2049) fire events records in each of 305 points. In order to post-process the ETC outputs, first the number of points with the level of MC and HC in each of the 7305 days have been measured. Next, this daily-based calculated values were divided based on their corresponding year and month of occurrence. Finally, the average of MC and HC occurrences in the studied 20 years (between the 2030 and 2050) in each of the months were calculated and the results are demonstrated in figure 6.



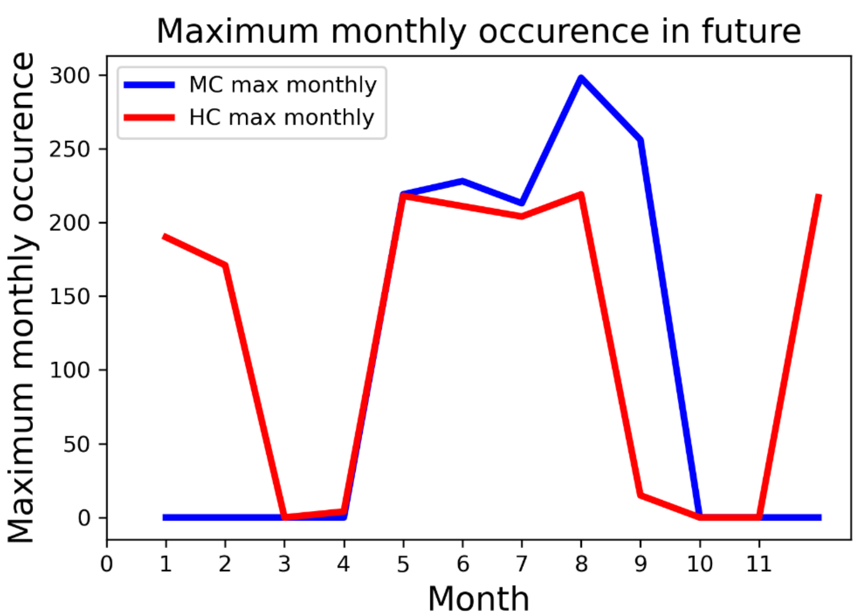
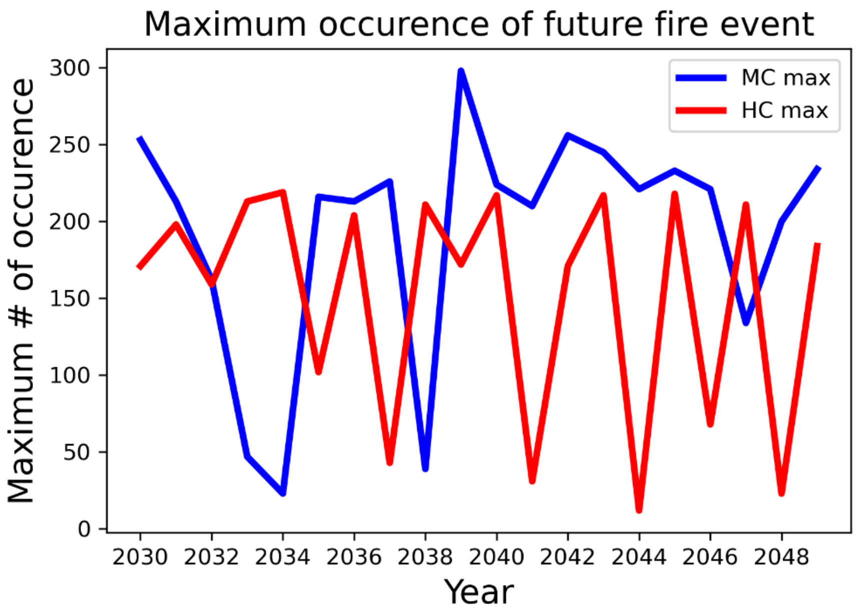


Figure 6: Monthly-based demonstration of (a) medium and (b) high confident fire events between the 2030 to 2060.

As the figures demonstrate, in both studied fire confidence levels, August (Month #8) with 383.26 and 141.050 has the highest average recordings of fire with medium and high certainties, respectively. Investigations on monthly-averaged records reveals that large portion of fire events at both MC and HC levels are occurred between the months of May to August, which are the summer time in Iran. These months are mostly described by high air temperature and low rainfall rate which make them more prone to forest fire events.

To conduct further analysis, comparing the number of fire incidents in historical (2000-2020) and near-future (2030-2050) periods on annual-basis seemed a necessity. To do this, the 7305 daily projected fire events has been reduced to 920 events which resulted by considering the exact future observation dates of MOD14A2 with the 8-day temporal resolution. This procedure will evidently lead to more consistency between two datasets and also make the projected fire incidents statistical more comparable with the one from MOD14A2. Next, the quantitative values of MC and HC fire events were aggregated on base on year of occurrence and the results are depicted in Figure 7.



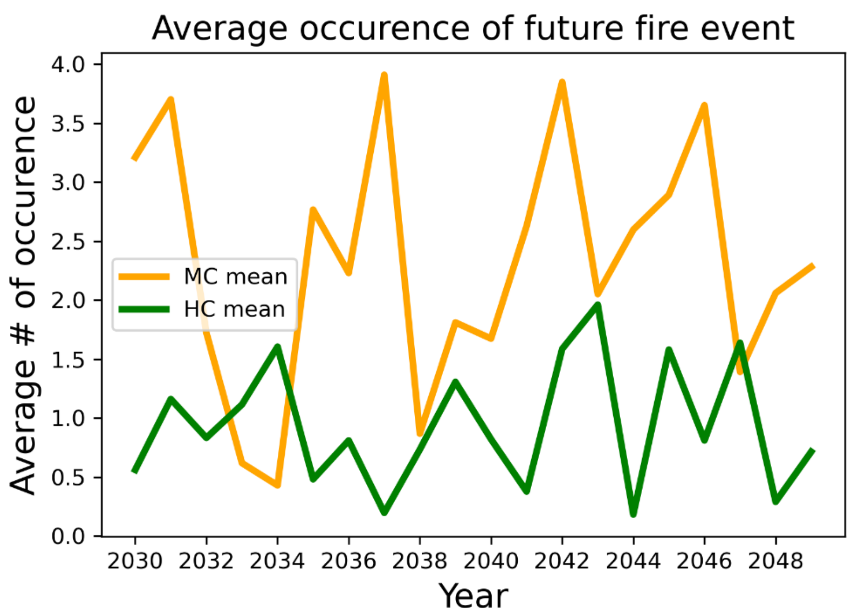


Figure 7: Comparison between yearly-based summation of fire events with (a) medium and (b) high confidences between the 2030 to 2060 and the historical period average (Red-line) obtained from MOD14A2 between 2000 to 2020.

As the figure illustrates, there are high fluctuations in total number of fire events between 2030 to 2050, however the fluctuation is more significant in medium confidence level. As the figure 7-a shows, while the total number of fires with MC level sustains heavy fluctuations, their overall occurrence average is decreased compare to the calculated value from the historical period. The years of 2042 and 2039 with 314 and 302 fire events are the years with highest possibility of MC fire incidents. However, based on pattern from Figure 7-b, the mean rate of future fire incidents with high certainty nearly doubled compare to the results obtained from historical period. In case of fire events with HC level, the years 2033, 2034 and 2036 with total fire incidents of 284, 219 and 219, respectively, are the most critical years in near future. Furthermore, in both MC and HC cases the year 2035 shows the lowest possibility of fire event.

5. Conclusion

This study focuses on simulating and forecasting the occurrence of forest fire incidents in the Golestan province of Iran. Initially, the MODIS-MOD14A2 fire products between 2000 to 2020 have been extracted over the study region. From all of MOD14A2 variables, only those indicating fires with medium- and high-certainties as well as the one denoting no land fire have been selected for further investigation. Based on fire records from 997 MODIS observations dates, total of 12 date which has higher occurrences value than the total 3rd quartile have been selected as the most critical days. Based on these 12 dates, total of 915 study points were extracted which has similarly 305 points with medium and high fire confidences as well as no land fire. By having the exact location of these 915 study points, their corresponding near surface climate variables have been obtained from the NASA's FLDAS satellite. The FLDAS with the spatial resolution of $0 \cdot 01°$ has the closest resolution with the MODIS products which make it the most suitable option among other similar satellites. 9 climate variables including total evapotranspiration, surface pressure, specific humidity, surface runoff, soil moisture content, soil temperature, net downward shortwave radiation flux and air temperature were extracted from the FLDAS database and constructed an input dataset for the three ensemble learning classifiers. Gradient Boosing, Random Forest and Extremely randomized tree algorithm were compared and their performance in simulating the forest fire have been evaluated. Based on four classification metrics, including Precision, recall, balanced accuracy, F1 and Area under the curve (AUC), the Extremely randomized tree obtains the highest accuracy. Also, by considering different arrangements of these 9 variables, total of 500 combinations were assessed and it revealed that the combination with four principal parameters including Qair, SoilMoi, Swnet and Tair is the most optimal selection. After concluding the simulation phase, products of various CMIP6 GCM sources for the year of 2015 were assessed and ones which has the highest correlation with the FLDAS have been selected. Based four widely employed relevancy metrics, namely R, RMSE, MAE and NSE, it became evident that the MIROC06-SSP370 has the most consistency with our extracted climate variables. Using these CMIP6 sources the specific humidity, soil moisture content, air temperature and net downward shortwave radiation flux for the future period of 2030 to 2050 were obtained and used in the ETC for the forecasting purposes. The average monthly evaluation of future projections shows higher probability of forest fire events in summer months with the August being the most critical month in both fires with medium and high confidence levels. Also, the summation of fire occurrences on annual-basis revealed that the overall rate of medium confidence fire events decreased compare to the rate from 2000 to 2020. In contrast, the total value of forest fire incidents with high certainty nearly got twice compare to the historical period. Meanwhile, the year 2042 and 2033 demonstrate the most recodes of forest fire events with medium and high confidence levels, respectively. The proposed simulation and projection method can be significantly useful in management of fire as well as preserve the valuable forestry regions specially in arid and semi-arid areas of the world.