IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received November 30, 2019, accepted December 16, 2019, date of publication January 3, 2020, date of current version January 10, 2020.

Digital Object Identifier 10.1109/ACCESS.2019.2963819

Flood Forecasting System Based on Integrated Big and Crowdsource Data by Using Machine Learning Techniques

SUPATTRA PUTTINAOVARAT[1] AND PARAMATE HORKAEW[2]
[1]Faculty of Science and Industrial Technology, Prince of Songkla University, Surat Thani Campus, Surat Thani 84000, Thailand
[2]School of Computer Engineering, Institute of Engineering, Suranaree University of Technology, Nakhon Ratchasima 30000, Thailand

Corresponding author: Paramate Horkaew (phorkaew@sut.ac.th)

This work was supported in part by the Agricultural Research Development Agency (Public Organization) under Grant PRP6105021450.

ABSTRACT Flood is one of the most disruptive natural hazards, responsible for loss of lives and damage to properties. A number of cities are subject to monsoons influences and hence face the disaster almost every year. Early notification of flood incident could benefit the authorities and public to devise both short and long terms preventive measures, to prepare evacuation and rescue mission, and to relieve the flood victims. Geographical locations of affected areas and respective severities, for instances, are among the key determinants in most flood administration. Thus far, an effective means of anticipating flood in advance remains lacking. Existing tools were typically based on manually input and prepared data. The processes were tedious and thus prohibitive for real-time and early forecasts. Furthermore, these tools did not fully exploit more comprehensive information available in current big data platforms. Therefore, this paper proposes a novel flood forecasting system based on fusing meteorological, hydrological, geospatial, and crowdsource big data in an adaptive machine learning framework. Data intelligence was driven by state-of-the-art learning strategies. Subjective and objective evaluations indicated that the developed system was able to forecast flood incidents, happening in specific areas and time frames. It was also later revealed by benchmarking experiments that the system configured with an MLP ANN gave the most effective prediction, with correct percentage, Kappa, MAE and RMSE of 97.93, 0.89, 0.01 and 0.10, respectively.

INDEX TERMS Flood forecasting system, big data, machine learning, crowdsource, deep learning.

I. INTRODUCTION

Natural flood is one of the most recurrent disasters [1]. Unlike stagnant water discharge, occasionally experienced in poorly planned cities, major flood incidents always cause considerable damages to properties and, more often than not, loss of lives. Several Asian countries, particularly Thailand, are subject to both southwest and northeast monsoons and accordingly facing seasonal deluge almost every year and in most parts of the countries [2]. Among notable causes, sudden and enduring heavy rain is the most pertinent one in Thailand [3], [4]. Furthermore, overflow from main rivers along shore sides to surrounding basins can greatly spread the damages [5]. Although being located further away from a river, an area with inappropriate land uses are unable to efficiently discharge accumulated precipitation, and hence are inevitably prone to even more frequent floods.

Regardless of causes, however, a flood is generally sudden and thus almost formidable for the general public and relevant organization to be adequately prepared for the incident. This is mainly due to the lack of an effective means of anticipating the disaster well in advance [6]. Despite the recent extensive development of computerized flood forecasting systems, they remained based primarily on present precipitation, monitored by rain stations or rain gauges. These facilities are normally owned by a meteorology department or similar organizations [7], [8]. Besides, they are scantly located in a few areas due to costly installation and maintenance. Hence, it is difficult to determine precipitation or predict flood accurately, especially in areas with no such facility [9]. To remedy this issue, precipitation in these areas were typically estimated either by inter- or extrapolation from those with rain stations

The associate editor coordinating the review of this manuscript and approving it for publication was Waleed Alsabhan.

present [10]–[12]. Due to a limited number of these stations and those readings in one area may not be a good representative to others. Therefore, estimated precipitation was insufficiently accurate to make a realistic forecast [13].

Conventional meteorological readings, e.g., precipitation, temperature, and humidity, etc., took really long time to measure, process, record, and transfer to relevant organizations [14], [15]. Analyses based on past precipitation were known to be associated with several shortcomings. For instance, they contribute to inaccurate [16], [17] and often outdated flood prediction. Limited sample size [18], inadequate computing capability, and inefficient prediction methods [16], [17] were all undermining the real potential of this scheme. Nonetheless, with the recent advances in distributed computing [19] and especially modern machine learning (ML), resembling human intelligence [16], [20]–[22], computerized flood forecasting, based on thematic factors has widely been investigated [23]–[25]. In addition, as the number of both open and proprietary data providers escalates, Big Data has now become a central source of information in such pursuits.

Thus far, according to recent surveys, most flood forecasting systems relied primarily on either monitored precipitation data or those obtained from a single source. Beside the mentioned limitations, existing systems remained lacking in other various aspects. For example, in a case where monitoring facilities or communication network of ones became malfunctioned, there would be no precipitation data available for imperative analyses. To the best of our knowledge, there was also no tool (software) that can accommodate area-specific forecasting well in advance. Furthermore, existing tools were highly dependent on demanding data preparation and compilation from various sources, including Big Data. As a consequent, automated and spontaneous notification of flood incidents to the public and authorities, or realistic anticipation of ones has remained a grand challenge.

In addition, there have been recent developments in flood forecasting systems based on ML. These systems embedded both attributes and crowdsourcing data into their ML frameworks. However, most existing systems operate by analyzing these data offline on premise before presenting their prediction results on various platforms [16], [20]–[22]. A typical practice was proposed in [20], where an ML was trained with real-time rainfalls, streamflow, and other data. It was unclear, nonetheless, how a prediction result was verified against an actual event, which was obtained from crowdsourcing.

Therefore, this paper proposes a novel flood forecasting system based on fusing meteorological, hydrological, geospatial, as well as crowdsourcing data, and integrating them into an ML framework. These data were compiled from various big data platforms, by using online application programming interfaces (API). The forecasting mechanism was driven by a machine learning strategy. To determine the most suitable one for the task, several state-of-the-art MLs, i.e., decision tree, random forest, naïve Bayes, artificial neural networks, support vector machine, and fuzzy logic, were compared. It is worth emphasizing that the novelty of this paper was not only to use different data in an ML, but also to enhance and verify its predictions based on crowdsourcing ones. It will be later shown in the experiments that the developed system was able to elevate known limitations and to enhance the effectiveness and efficiency of computerized flood forecasting.

This paper is organized as follow. The next section (II) surveys data, theories and practices relating to flood forecasting systems. Subsequently, section (III) describes the proposed scheme and the corresponding experiments. Then, the results of visual assessments and numerical evaluations on studied areas are reported and discussed (IV). Concluding remarks are given and prospective developments are suggested in the last section (V).

II. LITERATURE REVIEW

Reviews of related works consists of five main areas, which are, A) meteorological and hydrological data, B) geospatial data, C) application programming interface, D) crowdsource data, E) machine learning techniques for flood forecasting, and F) existing frameworks. The detailed discussion on relevant studies are provided in the following subsections.

A. METEOROLOGICAL AND HYDROLOGICAL DATA

Following prior investigations, it was concurred that meteorological data played a vital role in flood prediction [26], [27]. Particularly, amongst the most effective factors reported in recent works were, rainfall data [21], [26]–[29], rainfall duration [26]–[29], and stream networks [26]–[29]. In addition, other flood determinants engaged in its prediction included accumulated precipitation forecasting [29]–[31], flood hazard 100 years return period [32]–[33], and probability of precipitation [30]–[34].

Incorporating these determinants in forecasting was proven to increase its accuracy. Thus far, shortcomings of using meteorological and hydrological data were essentially two folds. Firstly, most data available are not real-time [14]–[18]. Secondly, these data were not of sufficiently high resolution [11], [28]. The most notable example of the latter is rainfalls being recorded at specific monitoring stations, normally located sparingly. For instance, there is normally only one station per city. Therefore, any area without one must rely on interpolation from those with actual readings. The resultant information is hence of low spatial resolution, inadequate as a representative instance for flood forecasting. This leads to remarkable errors. However, with the recent advances in remote sensing technology, radar-based rainfall monitoring [10], [28] has elevated such limitations. It can record data at more frequent rate, e.g., at every 6 minutes to 1 hours, and at higher spatial resolution than terrestrial stations.

This study, thus chose remotely sensed meteorological and hydrological data, provided by Thailand Meteorological Department (TMD) [35] and Global Flood [30]. These data were of high spatial and temporal resolution, suitable for the proposed objective.

B. GEOSPATIAL DATA

Literature survey on using geospatial data in flood analysis revealed that the majority of existing works involved both risk assessment (RA) and flood prediction. Depending on studied areas, these works employed various flood determining factors, the most common of which included height above sea level or elevation [5], [21], [37], slope [5], [21], [38], [39], land use and land cover [21], [40], [41], repeating flood [21], [27], and flow direction [5], [6], [21], [27]. Albeit these factors having less impact on flood than meteorological and hydrological ones, e.g., rainfall and stream network, etc. many studies have argued that without geospatial data, flood prediction would neither be so effective nor realistic. Some areas, despite heavy rainfalls, are not prone to flooding. This is because they may be subject to low to non-existent exposure to flood risk due to spatial advantages, e.g., being highly elevate, or having efficient draining passage, etc.

Accordingly, flood prediction in this study employed not only meteorological and hydrological data but also the geospatial ones. It was anticipated that with optimal combinations, the developed system would have better prediction accuracy, and hence being generalized in various areas.

C. APPLICATION PROGRAMMING INTERFACE (API)

API is a means of communication between websites, or specifically that between a service provider and clients. It allows transparent data interchange or cross-platform process execution. Provided a client being granted an access, updating or retrieving information can be completed programmatically in real-time.

Thanks to their versatility, a number of APIs have been deployed by various data service providers. In particular, Google Map API [42] is one of the prominent services worldwide. It offers access to geographical data in the forms of spatial maps for different proposes, including location finding (e.g., government buildings, tourist attraction points, or cooperate offices, etc.) [43], [44], environmental data displays (e.g., air pollution, gas concentration and water contaminations, etc.) [45], and pinpointing natural disasters (e.g., earthquake, wildfire, and flood, etc.) [46]. Currently, there are many providers offering API accesses to flood thematic data. These providers can be categorized into that based on proprietary and open-source platforms.

Unfortunately, the majority of these providers are often paid for, hence impeding entry to this technology, especially for those with limited financials. In order to develop a scheme that ensures greater public adoption, this study opted for those freely available, i.e., TMD [35] and Global Flood, whose system is called the Global Flood Awareness System (GLOFAS) [36]. The major advantages of these platforms were two folds. Firstly, they offered real-time readings. But most importantly, preliminary studies also confirmed that their historical data were accurate, corresponding to conditions and actual events. These characteristics would benefit accurate and instantaneous flood forecasting. Moreover, Google Map API supports interfaces consistent to these platforms and thus were used for visualizing flood forecasting results. Integration of Google MAP, TMD, and GLOFAS by API was thus a vital component in the proposed system.

D. CROWDSOURCE DATA

Crowdsource data have been adopted in many flood management and prevention schemes. To this end, relevant information is gathered from volunteers or participating groups. They are often used in coalition with primary data, conventionally collected by authorized entities. The main reason for such integration was to enhance both reliability of the information, and timing and economic efficiency of the compilation [47]–[52].

Currently, the most widely adopted approach is fetching data from social networks. Both plain text and multimedia data accounting the event are posted online by flood victims and maybe their relatives. They are simultaneously fetched, compiled and made accessible to authorized persons. Thus far, there are some limitations [53]. Utilization of crowdsource data has not yet been systemized. They are not used to assess, for instance, accuracies of flood prediction. Even then, there would be no means of verifying that these data were indeed of an affected area. Multimedia screening process is attentive and requires tremendous human resource, given vast amount of data involved. Without computerized automation, this process is often prohibitive during major flood event [48].

To our knowledge, crowdsourcing data were normally used in coordinating flood management and rescue operations, but not yet integrated with other thematic data in a forecasting system. To elevate these limitations, this study proposes exploiting crowdsource data in two ways. Firstly, uploaded photographs and flood levels were analyzed by Deep Learning (DL) neural network, to determine whether they were indeed of an actual event in a given area. Secondly, they served as input data, together with other relevant attributes, for the real-time forecasting system. Enhanced with these volunteer data, the system was anticipated to make a decision more accurately.

E. MACHINE LEARNING (ML) STRATEGIES

ML has increasingly been applied in flood forecasting systems. Therein, various ML strategies were explored. They included Decision Tree (DT) [16], Random Forrest (RF) [16], [54], Naïve Bayes, Artificial Neural Network (ANN) [16], [21], [22], Support Vector Machine (SVM) and Regression (SVR) [16], [54], [55] and Fuzzy Logic [16] [56], [21], etc. However, many of these studies focused upon conceptual frameworks, investigating feasibility of ML in flood forecasting. They did neither thoroughly validate their resultant forecasts against actual events, nor did they provide guidelines on relevant parameters settings. One exception was an extensive review by Mosavi *et al.* [16], who reported that the most efficient and promising ML strategies for this purpose were ANN, SVM, and SVR. Their overall accuracies were relatively high, ranging between 70% to 90%. It was construed that system accuracy was largely dependent on both learning strategies and attributes being considered.

Thus far, the majority of these ML based software demanded manual input data preparation. It hinders the systems from being automatically operational. Consequently, most existing systems were unable to instantaneous response to, for example, sudden flood. In addition, following such events, additional data would always be required for subsequent analyses. It was found in our survey that there existed a system attempting to remedy these issues. An online flood forecasting system based on Self-Organizing Map (SOM) was recently developed [20] on a .NET Framework. It was written in C# Language, interacting with an SQL Server and Google Map API. The system made its forecast based on rainfalls, streamflow, and other relevant attributes stored in the database, and displayed affected locations in real-time. A drawback of this system was that it relied on measurements from rain stations. A limited number and coverage of the active sites could undermine its performance. Furthermore, the forecast was made based on present data, restricting it from determining flood events much further ahead. The study did not however verify its forecasted results against the floods that actually happened. One was thus unable to assess the true performance of SOM in such framework. Nonetheless, a notable quality distinguishing this work from other similar attempts, was that the entire process was automatic. All data involved were directly acquired from monitoring sources and stored in a well-structured database. As such, it did not require any human intervention, either in data preparation or processing. Thanks to web-based design, the system was able to run on virtually all devices, regardless of their operating systems (OS).

Motivated by [20], this study proposed a novel framework that made key improvements over that system. It used public crowdsource data, both as an input attributes to the system and for the verification purpose. Not only meteorological and hydrological data, but it also incorporated other relevant geospatial ones in training an ML system. Accumulated precipitation and probabilities of rainfall at particular levels could be estimated in advance. Prediction of relevant attributes and forecasted flood were validated against the actual situations. Detailed experiments comparing different strategies were carried out to evaluate their performance, and hence with available data, to determine a most suitable technique for area specific contexts. Verifications of the resultant flood forecasts were made against those obtained from both onsite expeditions and crowdsource data, downloaded via thaiflood.org. Numerical and *in situ* assessments indicated that the developed system can make the forecast more effectively, and hence much pertinent to public hazard management realm in larger scales.

F. FLOOD FORECASTING FRAMEWORK

Existing frameworks [57], [58] were implemented as application software on different platforms. In a first instance [57], a mobile application on ASP.NET framework written in C# language was developed. This software forecasted a flood event by using the HEC-HMS algorithm, based on factors stored in MySQL database. However, it was limited to only an Android operating system. Another more in-depth analysis of flood situations was developed on a web platform [58]. This system statistically analyzed floods based on the historical data of water levels, and seven points of water inlets and outlets. The resultant forecast was presented as point-wise flood levels, rendered on a two-dimensional map. Its shortcoming was that the analysis did not consider any geospatial data. In addition, forecasts could only be made at specific outlets, but not at those arbitrarily queried by users. Although it was developed as a web application, it did not support responsive web technology, and as such was unable to equally well satisfy user experience on all other devices apart from personal computers (PC).

In terms of mathematical models, it was found that current frameworks made their flood forecast based on three models, i.e., meteorological, hydrological, and ML ones. The first two models were used with many algorithms, e.g., HEC-HMS [57], HEC-RAS [59], Mike11 [60], Mike21 [60], Mike-Flood [60], and ECMWF [61], etc., while those using ML model were implemented with various methods (as discussed in the section I. E). Thus far, these systems were only experimental and, to our knowledge, not yet publicly distributed.

To elevate these limitations, this paper thus analyzed and designed a flood forecasting system that improved over the current ones. The aspects considered herein were supports of responsive web technology, automation of key processes, and availability and usability of the system. To this end, the proposed system was developed by using both meteorological and hydrological models in forecasting accumulated precipitation from data obtained from TMD big data and GLOFAS, and ML models in forecasting flood situations in given areas. The analyses were made based on meteorological, hydrological, geospatial and crowdsourcing data.

III. PROPOSED SCHEME AND EXPERIMENTS

A. STUDY AREA

To elucidate the merits of the proposed scheme, the experiments were carried out on two provinces, located in the South of Thailand. They were Surat Thani and Nakhon Si Thammarat. Their geographical illustrations are depicted in Figure 1.

The reason for considering these provinces in this study was the fact that, unlike other parts of the country, both areas are under influences of both southeast and northwest monsoons. As a consequence, they both are prone to heavy floods, almost every year. Nonetheless, evaluations and assessments

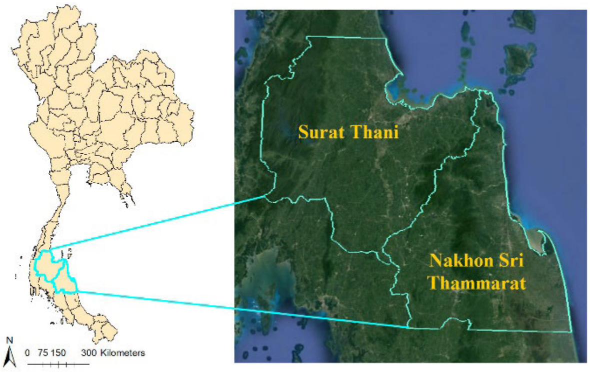


FIGURE 1. Study Area consisting of two cities in the South of Thailand. They were Surat Thani and Nakhon Si Thammarat Provinces.

later made in this paper would show that no restriction on data nor fundamental processes was imposed regarding these specific provinces. Therefore, the proposed scheme could be generalized and applied equally well to other areas.

B. DATA PREPARATION

In this study, data used in flood forecasting could be categorized into four main groups. They were 1) geospatial, 2) meteorological and hydrological, obtained from GLOFAS, 3) hourly rainfalls prediction from TMD Big Data platform, and 4) crowdsource (or volunteer) data. They were stored in geo-database and then processed by one of modern ML strategies. Data interchanges were done via four interface technologies, i.e., Web Feature Service (WFS), Web Map Service (WMS), TMD API, and Google MAP API. The resultant models were then employed in subsequent forecasting systems.

Since all data involved in this study were acquired from various sources, they were hence of different spatial and temporal resolutions. To normalize them into a common coordinate frame, a pre-processing step was required. In terms of spatial resolution, interpolation was made based on their geographic coordinates. It should be noted that this normalization was not intended nor did it able to increase their intrinsic information, but only to align corresponding positions for consistent sampling in ML. In this study, bilinear interpolation (BI) was used. On the other hand, interpolating these data temporally was not so trivial, especially when they were irregularly stored at diverse scales. The proposed system updated its forecasts on a daily basis. Therefore, data sampled at 24-hour intervals, e.g., rainfall, and precipitation, etc., would not require resampling. However, for historical records that do not as frequently change, e.g., 100-year return period, land use and land cover, etc., the most recent data available would be used. This is equivalent to applying nearest neighbor (NN) interpolation scheme. Detailed pre-processing on each dataset are listed in Table 1.

C. DATA INTEGRATION

Figure 2. depicts the conceptual diagram describing the proposed scheme. The key elements were data acquisition and interchange between the system and respective sources and their intelligence via MLs.

TABLE 1. Interpolations applied to data involved in this study.

Factors	ORG. Res.	NEW Res.	Temporal Point	Spatial Interp.
Elevation	5 m	30 m	Year 2017	BI
Slope	5 m	30 m	Year 2017	BI
Flow Direction	5 m	30 m	Year 2017	BI
Land use and Land cover	30 m	30 m	Year 2017	None
Repeating Flood	30 m.	30 m.	8 Years	None
Accumulated Precipitation	10 km	30 m	Daily	BI
Prob. of Precipitation (50 mm)	10 km	30 m	Daily	BI
Prob. of Precipitation (150 mm)	10 km	30 m	Daily	BI
Prob. of Precipitation (300 mm)	10 km	30 m	Daily	BI
Flood Hazard 100-year return period	10 km	30 m	Year 2019	BI
5 Year Return Period Exceedance	10 km	30 m	Year 2019	BI
Rainfall forecasting	2 km	30 m	Daily	BI
Rainfall Duration	Actual	Actual	Daily	None
Rainfall Intensity Level	Actual	Actual	Daily	None
Drainage Ability Problem	Actual	Actual	Daily	None

Thematic data acquisitions were divided into four groups, i.e., meteorological and hydrological data, hourly precipitation data, area specific geospatial data, and crowdsource data. Process in each group can be elaborated as follow.

Firstly, meteorological and hydrological data were acquired from the Global Flood server, called GLOFAS [36]. Meteorological data consisted of accumulated precipitation, and the probability of precipitation at different levels, predicted daily. The prediction is based on ECMWF (European Center for Medium-Range Weather Forecasts) model. Hydrological data consisted of flood hazard 100-year return period, and 5-year return period exceedance. These were acquired from GLOFAS via Web Map Service (WMS), with a program written in PHP language. Once downloaded, the data were stored in raster formats for subsequent analyses.

The second source of data was rainfall forecasting, which was acquired from big data repository, managed by TMD [35]. The acquisition via API was made by a web application developed with PHP and Leaflet JavaScript. With this platform, rainfalls and their accumulation could be predicted 48 hours in advance. Likewise, the current and predicted data were stored in our MySQL database for subsequent analyses.

In this study, MySQL was preferred to other spatial databases, thanks to its compatibility with involved crowd sources' interfaces, e.g., that of thaiflood.org, and integrability with a wide range of web-based administrative and flood management functionalities. In spite of not being specifically

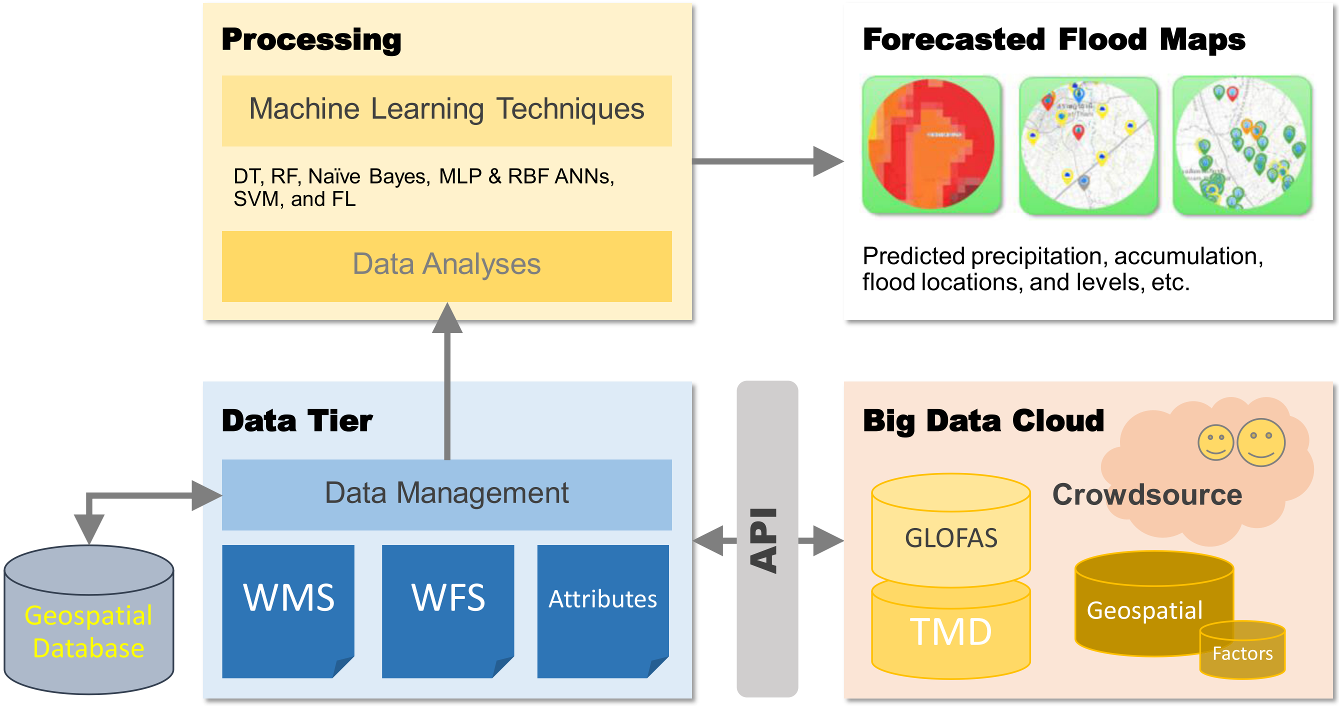


FIGURE 2. Conceptual diagram of the proposed scheme.

designed for big and geospatial data, its most recent update does support both platforms via Hadoop, Apache Sqoop and NoSQL, etc. Moreover, geographical processing was primarily not geometric operations but data driven MLs. Therefore, a dedicated geospatial database was not explicitly needed. Having said that, geographical attributes were indeed associated with thematic data via a well-defined relational structure. Accordingly, geographical data manipulation and processing was effectively handled by GeoJSON (RFC 7946), catered to web developers. Both MySQL and GeoJSON were opensource and fully supported by most web frameworks.

Thirdly, area-specific geospatial data consisted of height above sea level, slope, land use and land cover, 10-year repeating flood, and flow direction. These data were compiled from Thai authorities, i.e., the Land Development Department, and Informatics and Space Technology Development Agency (GISTDA). They were acquired by using Web Feature Service and stored in our database for further analyses.

The fourth group was crowdsourced data gathered from the public. They were further divided into two parts. Data in the first part were used in training the forecasting system. They were essentially real-time data accounting actual incidents. These real-time data consisted of area specific rainfall intensity levels, continuing rainfall durations, and drainage ability. Although many social media platforms provide these data, they were not cost effective, as fee would normally be charged for on-demand access to associated coordinates. This study thus developed, in-house software to gather these crowdsource data. To this end, an online reporting system was developed so that the participants may inform of their current situation. Once the system was launched, additional reports may be fed back to adjust ML model, making the forecast more realistic. The second part was verification data, characterizing floods happening in given areas. In addition to data reported by voluntary users, it also consisted of flood levels, fetched from thaiflood.org. In fact, the latter system was also developed in our previous work, with an intention to gather flood related information from the public for state's uses, e.g., devising flood management, planning and executing rescue mission, and reliving flood victims, etc.

For examples, geospatial data (i.e., elevation, slope, flow-direction, land use and land cover, and repeating flood) are shown in Figure 3, while those of meteorological and hydrological data (i.e., forecast of accumulated precipitation) are shown in Figure 4. Figure 5. illustrates six examples of probability of precipitation at 3 different levels. Finally, Figure 6(a) and (b) depicts the maps of flood hazard 100-year return period and 5-year return period exceedance. All data employed in this study are summarized in Table 2. The factors, their abbreviations, and ratings are referred by ML methods in the section III D.

The amount of data acquired from GLOFOS and TMD platforms by the proposed system were in both raster and vector types, and also of greater than 10TB in size. It is worth noted here that, although the amount actually arriving at, storing in, and passing through the system at a specific time was not categorically immense, they were of significantly high variety (structured and unstructured), and high velocity.

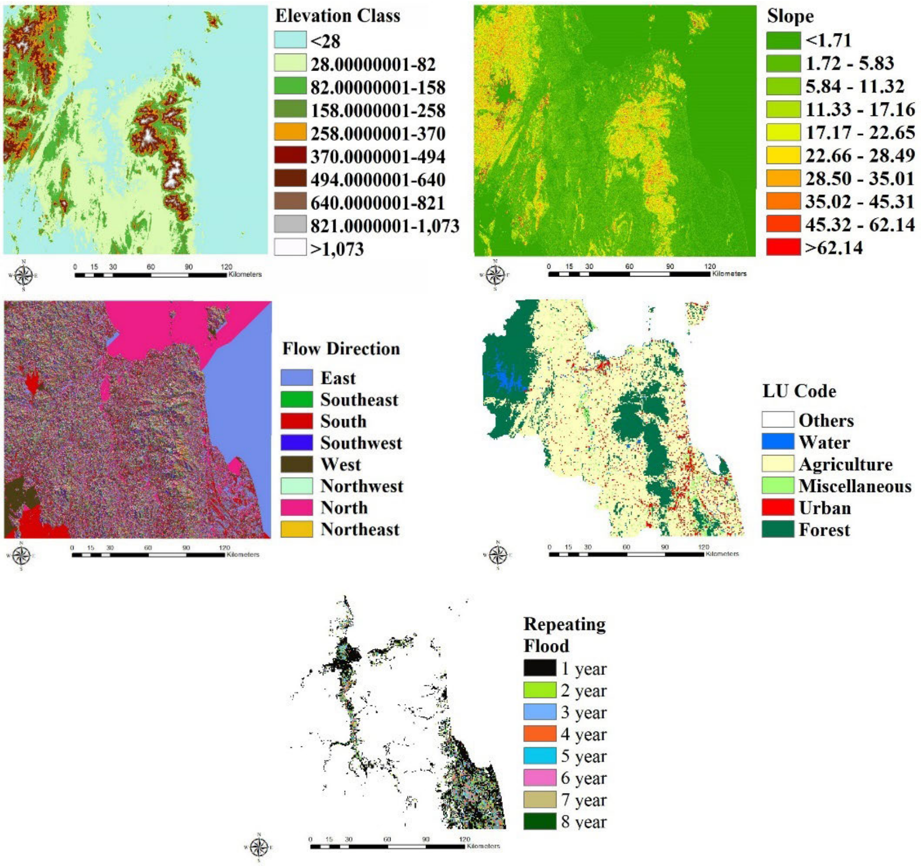


FIGURE 3. Spatial Factor maps consisting of elevation, slope, flow direction, land-use, and repeating flood.

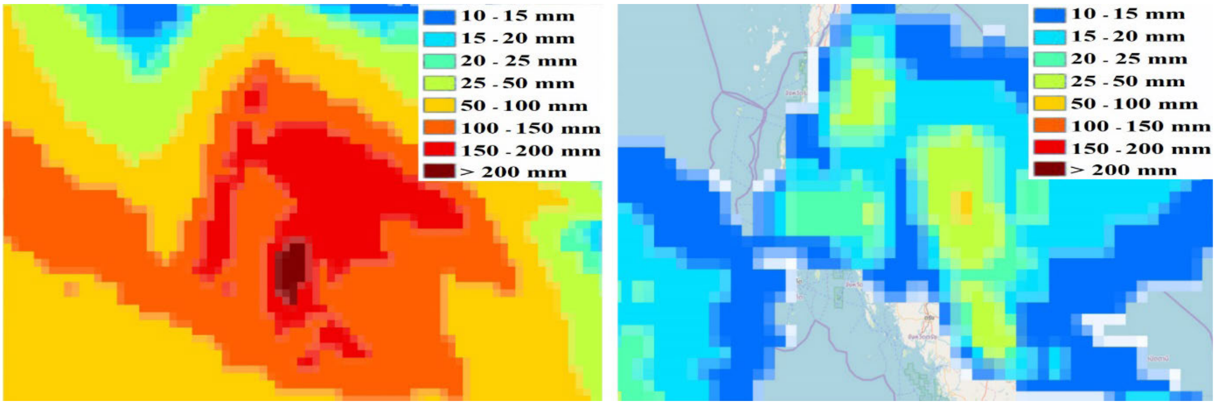


FIGURE 4. Accumulated precipitation forecast map of two selected areas.

These data, such as rainfalls and precipitations, were continuously varying during the system operations. As a consequence, conventional data handling was inadequate. Without big data operations, the forecasting could not be instantly responding to such heterogenous changes, especially when the proposed flood forecast system is scaled up to support wider areas of interest than those demonstrated herein. Furthermore, the system, in fact, utilized TMD big rainfalls data, processed by the Thai Meteorological Department, by using numerical weather forecast (NWP) and those by GLOFAS, by using ECMWF hydrological model. These data were also of not only high volume, variety, and velocity, but also inconclusive. These characteristics thus called for big and not conventional data handling. It would be later shown in the subsequent experiments that the rainfalls big data forecasted by ECMWF model yielded the most accurate results [61].

The analysis and design of the proposed system focused on its main functional requirements. They included gathering not only key geospatial factors from institutional data

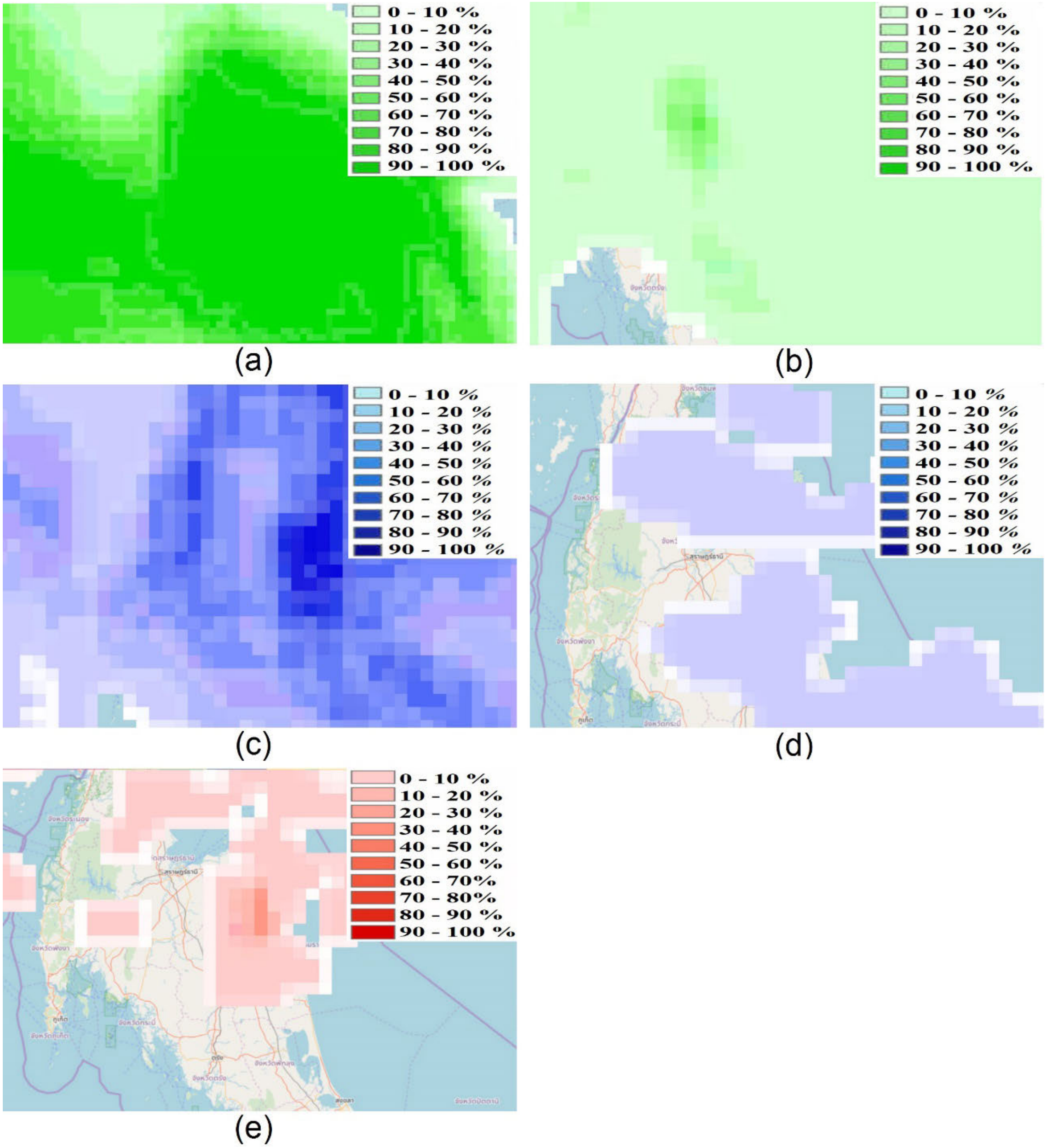


FIGURE 5. Precipitation maps for flood forecasting when the probabilities are high (a, c, and e), and low (b and d).

sources but also those crowdsourced from individuals. The main functional components of the system are illustrated as a use case diagram in Figure 7. Users are divided into two groups. Firstly, the one labelled as ''User,'' are the general public, community leaders, or government officers who are anticipating the event and need access to forecasts so as to prepare appropriate measure in accordance. The users are able to browse map data as well as relevant information geographically, such as observed and predicted precipitations. The prediction includes accumulated amount (in mm.) and probability (in percent) of precipitation, etc. Moreover, based on relevant factors within an area, they may query prediction of flood event, along with its likelihood and possible severity. The outcomes can then be verified against the flooding data, crowdsourced and reported via the Flood Mitigation System (thaiflood.org). Secondly, the other group labeled as ''Admin/Officer'' are authorized agents who are operating the services. In addition to generic functionalities, their administrative tasks include membership management as well as updating relevant data and forecasted results.

D. FLOOD FORECASTING BY MACHINE LEARNING

To determine a suitable ML strategy for flood prediction, this section compared the forecasting performances from

TABLE 2. Summary of the thematic data, their attributes, and units, employed in this study.

Factors	Classes	Rating	Sources
Elevation (DEM)	<28 m	1	Geospatial Factors
	28.00000001-82 m	2	
	82.00000001-158 m	3	
	158.0000001-258 m	4	
	258.0000001-370 m	5	
	370.0000001-494 m	6	
	494.0000001-640 m	7	
	640.0000001-821 m	8	
	821.0000001-1,073 m	9	
	>1,073 m	10	
Slope (Slope)	<1.716620802 (Degrees)	1	Geospatial Factors
	1.716620802 - 5.836510722 (Degrees)	2	
	5.836510723 - 11.32969728 (Degrees)	3	
	11.32969729 - 17.16620801 (Degrees)	4	
	17.16620802 - 22.65939457 (Degrees)	5	
	22.65939458 - 28.49590529 (Degrees)	6	
	28.49590530 - 35.01906433 (Degrees)	7	
	35.01906434 - 45.31878913 (Degrees)	8	
	45.31878914 - 62.14167298 (Degrees)	9	
	>62.14167298 (Degrees)	10	
Flow Direction (FlowDir)	1 = East	1	Geospatial Factors
	2 = Southeast	2	
	4 = South	3	
	8 = Southwest	4	
	16 = West	5	
	32 = Northwest	6	
	64 = North	7	
	128 = Northeast	8	
Land use and Land cover (LULC)	O = Others	1	Geospatial Factors
	W = Water	2	
	A = Agriculture	3	
	M = Miscellaneous	4	
	U = Urban	5	
	F = Forest	6	
Repeating Flood (RepeatFlood)	1 year	1	Geospatial Factors
	2 year	2	
	3 year	3	
	4 year	4	
	5 year	5	
	6 year	6	
	7 year	7	
	8 year	8	
Accumulated Precipitation (RainAcc)	10 - 15 mm	1	GLOFAS
	15 - 20 mm	2	
	20 - 25 mm	3	
	25 - 50 mm	4	
	50 - 100 mm	5	
	100 - 150 mm	6	
	150 - 200 mm	7	
	> 200 mm	8	
Probability of Precipitation (50 mm) (Rain_P50)	0 - 10 %	1	GLOFAS
	10 - 20 %	2	
	20 - 30 %	3	
	30 - 40 %	4	
	40 - 50 %	5	
	50 - 60 %	6	
	60 - 70 %	7	
	70 - 80 %	8	
	80 - 90 %	9	
	90 - 100 %	10	
Probability of Precipitation (150 mm) (Rain_P150)	0 - 10 %	1	GLOFAS
	10 - 20 %	2	
	20 - 30 %	3	
	30 - 40 %	4	
	40 - 50 %	5	
	50 - 60 %	6	
	60 - 70 %	7	
	70 - 80 %	8	
	80 - 90 %	9	

TABLE 2. *(Continued.)* Summary of the thematic data, their attributes, and units, employed in this study.

Factors	Classes	Rating	Sources
	90 - 100 %	10	
Probability of Precipitation (300 mm) (Rain_P300)	0 - 10 %	1	GLOFAS
	10 - 20 %	2	
	20 - 30 %	3	
	30 - 40 %	4	
	40 - 50 %	5	
	50 - 60 %	6	
	60 - 70 %	7	
	70 - 80 %	8	
	80 - 90 %	9	
	90 - 100 %	10	
Flood Hazard *100*-year return period (Flood100Year)	Shallow (<1m)	1	GLOFAS
	Moderate (1-3m)	2	
	Deep (3-10m)	3	
	Very Deep (>10m)	4	
5 Year Return Period Exceedance (PE5Year)	0 - 10 %	1	GLOFAS
	10 - 20 %	2	
	20 - 30 %	3	
	30 - 40 %	4	
	40 - 50 %	5	
	50 - 60 %	6	
	60 - 70%	7	
	70 – 80%	8	
	80 - 90 %	9	
	90 - 100 %	10	
Rainfall forecasting (RainLevel)	10 - 15 mm	1	TMD Big Data
	15 - 20 mm	2	
	20 - 25 mm	3	
	25 - 50 mm	4	
	50 - 100 mm	5	
	100 - 150 mm	6	
	150 - 200 mm	7	
	> 200 mm	8	
Rainfall Duration (RainDura)	None	1	Crowdsource
	Low (1 – 3 Hours)	2	
	Moderate (3 – 5 Hours)	3	
	High (>5 Hours)	4	
Rainfall Intensity Level (RainInt)	None	1	Crowdsource
	Low	2	
	Moderate	3	
	High	4	
Drainage Ability Problem (PDrainage)	None	1	Crowdsource
	Low	2	
	Moderate	3	
	High	4	

different MLs. Those considered in this study were DT (J48), RF, Naïve Bayes, ANN (both MLP and RBF architectures), SVM and Fuzzy Logic. On evaluating each ML, K-fold cross-validation was employed. The input data were divided into four groups, i.e., thematic spatial layers, meteorological and hydrological data obtained from GLOFAS, hourly predicted precipitation obtained from TMD big data, and crowdsourced (or volunteered) data. Altogether, they constituted to 15 variables (as shown in Table 1). For a given area, prediction outcomes were divided into 4 classes. They were 1) no flood is anticipated, 2) flood level were below 20 cm, 3) flood level were between 20 to 49 cm, and 4) flood level were 50 cm or above. The system provided flood forecasting on daily basis. Its lead time was hence 24 hours. The predicted results were subsequently validated against those retrieved from trusted agencies, onsite expeditions, and crowdsourced reports via the web application (developed by the authors and thaiflood.org). In this paper, the studied area covers Surat Thani and Nakhon Sri Thammarat provinces.

Unlike a physical based approach, data driven ML does not focus on insights into functional models, but intrinsic relationships between flood relevant factors and corresponding outcomes, learned from the past events. The design of ML models adopted in this framework and their characteristics are described as follow: Firstly, MLP is a configuration of ANN with multiple layers and is suitable for complicate learning tasks. An MLP network employs back propagation scheme, which consists of 2 reciprocal procedures, *i.e.*, forward and backward passes. The forward-pass traverses data presented at the input layers through the ones hidden in the network, while the backward pass iteratively adjusts their connecting weights such that the errors between actual and expected

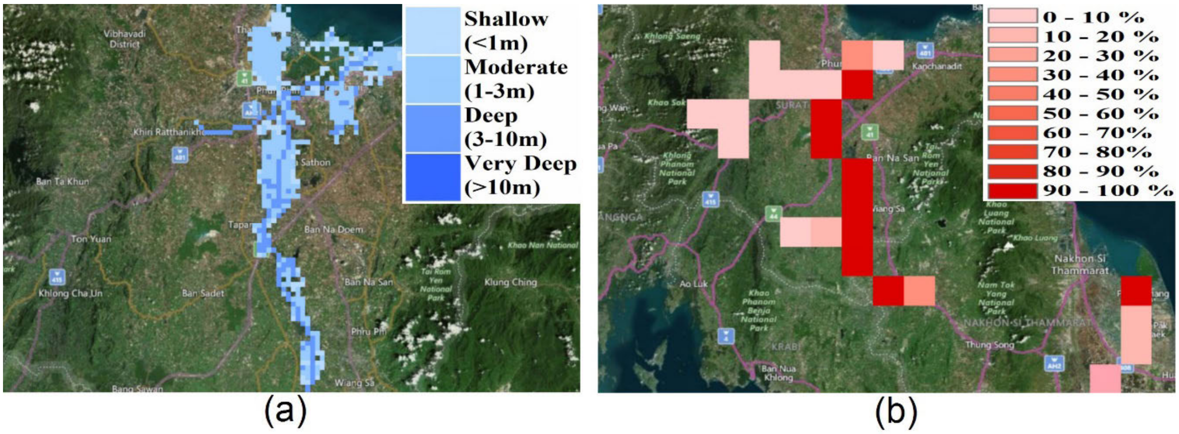


FIGURE 6. Maps of Flood hazard 100 years period (a) and 5-year return period exceedance (b).

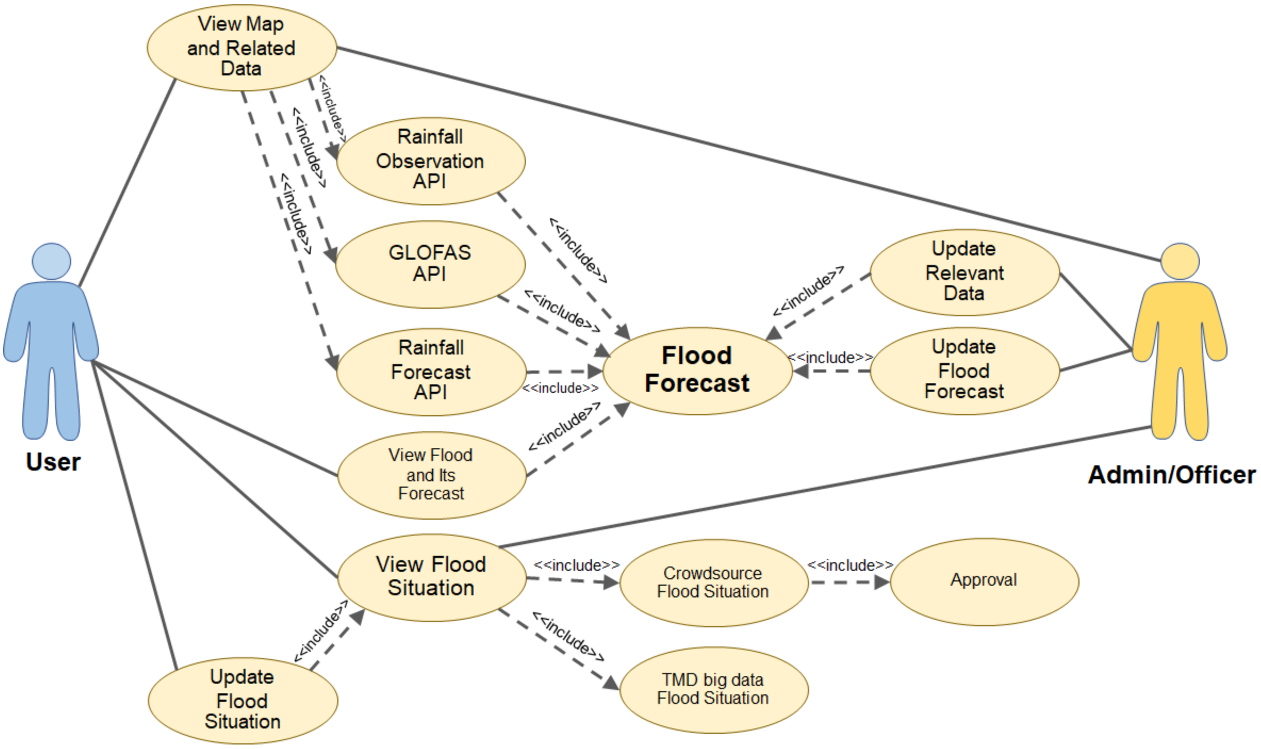


FIGURE 7. Use case diagram describing the functions of the proposed system.

responses are minimized. In our implementation, MLP was defined as per equations (1) and (2).

$$x_j = \sum_{i=1}^{n} x_i w_{ij} + b_j w_j \tag{1}$$

$$y_i = \frac{1}{1+e^{-x_i}} \tag{2}$$

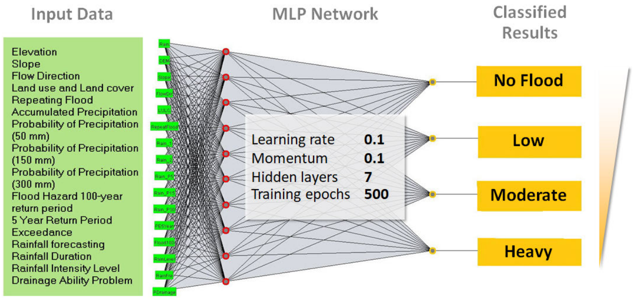
For a given dataset, n is the total number of input nodes, x_i is a sampled data point present at the i^{th} node, w_{ij} is the weight assigned to a link connecting the i^{th} and the j^{th} nodes, b_j and w_j are, respectively, bias and weight linking to the j^{th} node, and y_i is the response at the i^{th} node.

In the proposed framework, MLP consisted of three main parts, i.e., input data, the network, and classified results, as shown in Figure 8. The input data consisted of 16 attributes, listed in Table 2. The network parameters were empirically determined by preliminary trials. They were assigned to the ANN as follow: learning rate (0.1), momentum (0.1), number of hidden layers (7), and training epochs (500). The predicted classes which corresponded to four levels of flooding were defined as none-existing, low, moderate, and heavy.

SVM was first introduced in a binary classifier problem and later extended to multi-class ones [62]. It neither poses

TABLE 3. Resultant fuzzy rules applied during inference.

Rule No.	Fuzzy Rules	
	Antecedent (IF)	Consequence (THEN)
R1	RainDura is Moderate	Non-Flood
R2	RepeatFlood is Low **AND** Slope is High	Non-Flood
R3	PDrainage is Moderate **AND** PE5Year is Moderate	Non-Flood
R4	PDrainage is Low **AND** RainLevel is Low **AND** PE5Year is Moderate	Non-Flood
R5	PDrainage is Low **AND** LULC is Moderate	Non-Flood
R6	RainDura is Moderate **AND** Rain_P150 is High **AND** PDrainage is Low	Low Level Flood
R7	Rain_P300 is Low **AND** Rain_Level is Moderate **AND** RepeatFlood is Moderate	Low Level Flood
R8	RainDura is High **AND** Repeat Flood is Moderate **AND** PDrainage is Moderate	Moderate Flood
R9	RainDura is High **AND** RainLevel is High	Heavy Flood
R10	RainDura is High **AND** Slope is Low **AND** RainLevel is High	Heavy Flood



Learning rate 0.1
Momentum 0.1
Hidden layers 7
Training epochs 500

FIGURE 8. The MLP employed in this study and its parameter settings.

any assumption on samples distribution nor the geometry of their separation. This is because, in its general form, SVM creates decision boundaries on a hyper-plane based on a series of kernels. Let the training data consists of n vectors {(x_i, y_i), i = 1, 2, ... N}. A class value or target $y_i \in (-1, 1)$ is associated to each vector, where N is the number of training samples. This study defined a non-linear classifier, whose expression is given as follow.

$$f(x) = sgn\left[\sum\nolimits_{i=1}^{N} a_i y_i K\left(\mathbf{x}_i \cdot \mathbf{x}_j\right) + b\right], \tag{3}$$

where *sgn* is the sign function, K is the kernel function and the magnitude of a_i is determined by regularizing parameter C, · is inner product between two vectors, b is a bias value, and K is defined by a radial basis function, i.e.,

$$K\left(x_i \cdot x_j\right) = \exp\left(-\gamma \left\|x_i - x_j\right\|^2\right), \tag{4}$$

where γ is a kernel parameter, characterizing data dispersion and hence the extent of kernel supports. In the subsequent experiments, b, C and γ, were set to 0.0, 1.0 and 0.1, respectively.

DT (J48) classifier was also considered in this study. Based on a tree structure, this method classifies an instance at each node based on its attributes. Among different variants, a J48 uses Gini index to determine an appropriate attribute and criterion for a given node. In this study, the considered attributes were accumulated precipitation (**RainAcc**), probability of precipitation at 150 mm and 300 mm (**Rain_P150** and **Rain_P300**), drainage ability problem (**P_Drainage**), 5-year return period exceedance (**PE5 Year**), elevation (**DEM**), slope (**Slope**), rainfall duration (**RainDura**), and repeating flood (**RepeatFlood**). These attributes were categorized and given rating numbers, as specified in Table 2. The resultant rules for flood forecasting are illustrated in Figure 9. This DT reads as follow: On the left main branch, for instance, if the accumulated precipitation was no more than 25 mm (rated 3), the probability of precipitation at 300 mm would be next considered. If it was no greater than 10% (rated 1), no flooding would be anticipated. Otherwise, the drainage ability of the area would be then assessed. If it did not have any drainage difficulty (rated 1) and it had less than 10% (rated 1) of 5-year return period exceedance, the area did not either expect any flood. However, a low-level of flood (20-50 cm.) was forecasted, if any of these conditions were not satisfied. Forecasts on low, moderate, and heavy floods could similarly be made following the main branches on the right, in which case elevation, slope, and probability of precipitation at 150 mm, were taken into account.

Fuzzy Logic is a computerized reasoning method that emulates complex human thoughts. Its strength is due to extension of Boolean logics to a fuzzy set of partial truths, whose values are continuously defined between 0 to 1. Fuzzy Logic consists of three main operations, as depicted in Figure 10. The first operation is Fuzzification which maps an input sample to a membership value by using a membership function. In this study, that of a triangular type was chosen. The second step is called inference, where fuzzified data are interpreted and analyzed based on a set fuzzy rules, specified in Table 3. Finally, Defuzzification assigns analyzed output variables with the exact decision. It was evident from the Fuzzy rules that the key determinant on flood forecasting were rain duration, rainfall forecasting, repeating flood, drainage ability problem, probability of precipitation (300 mm), probability

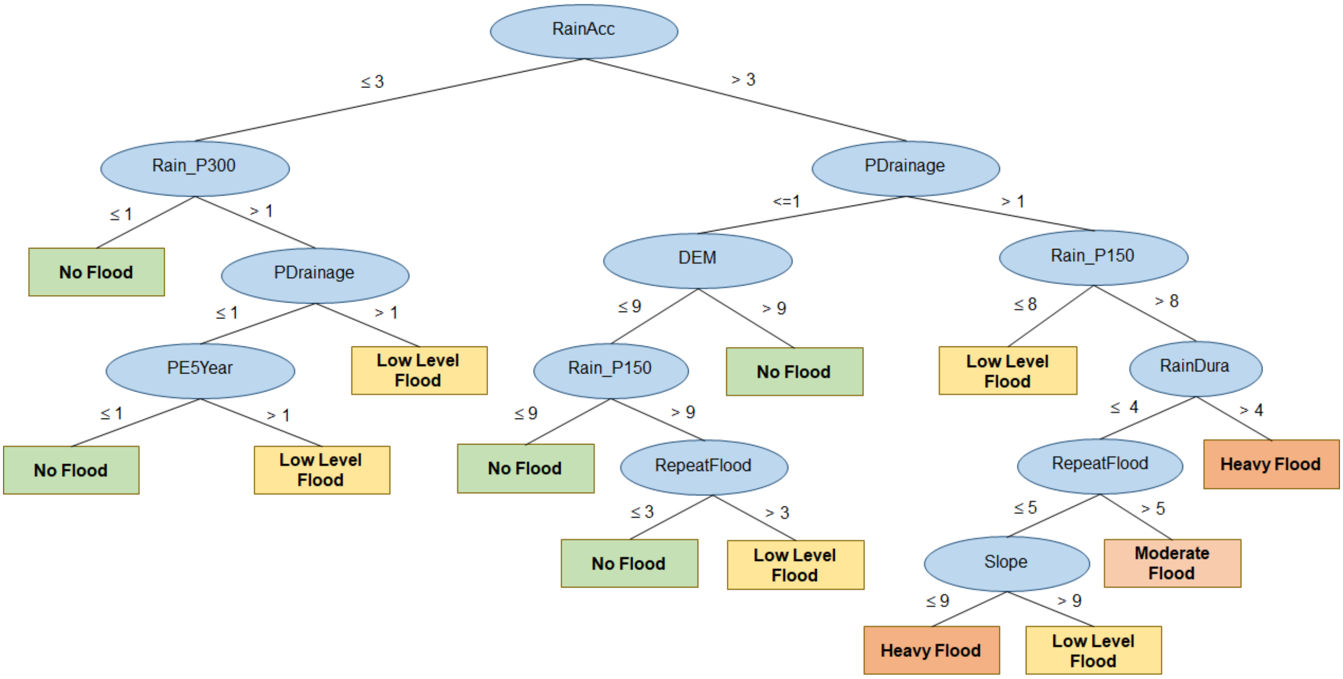


FIGURE 9. Resultant flood forecasting decision tree derived from the DT (J48) algorithm.

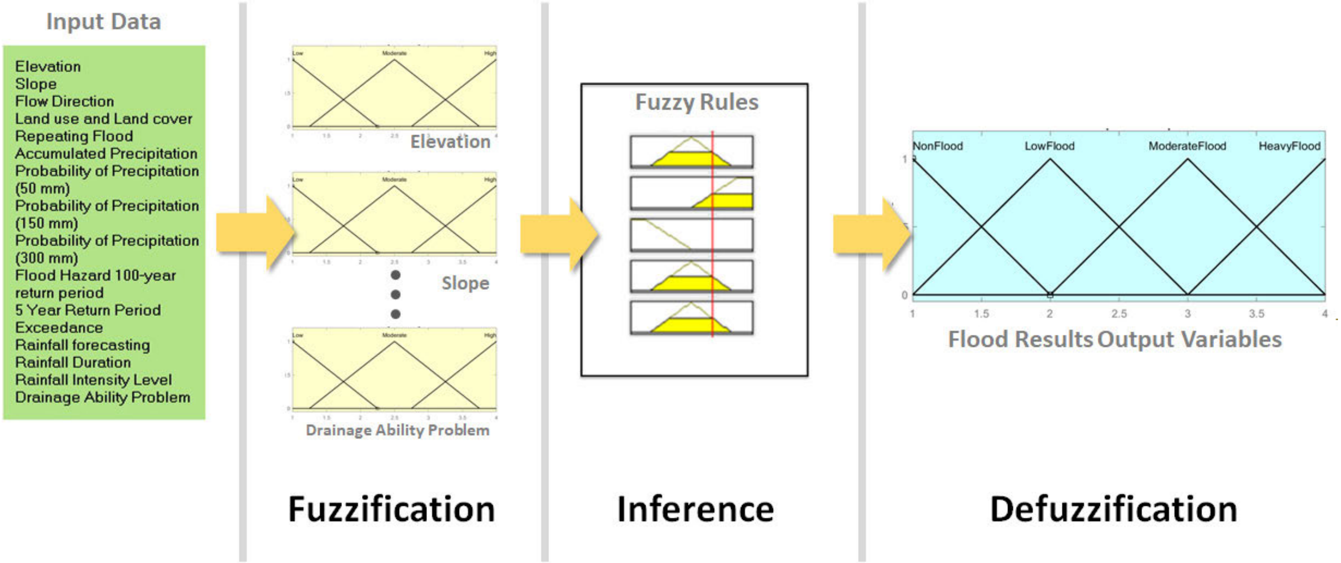


FIGURE 10. Diagram of the Fuzzy Logic employed in flood forecasting.

of precipitation (150 mm), land use and land cover and 5-year return period exceedance.

To address over-fitting issues 10-fold cross validation was used to assess all abovementioned ML algorithms, in the experiments.

E. ACCURACY ASSESSMENTS OF FLOOD FORECASTING

This paper employed standard accuracy metrics which were Corrected Classified Instances (CCI), Kappa, Mean Absolute Error (MAE), Root Mean Square Error (RMSE), True Positive (TP), False Positive (FP), Precision, Recall, F-Measure, and Area under ROC. Particularly, MAE and RMSE were defined as follow:

$$\mathrm{MAE} = \frac{1}{n}\sum\nolimits_{i=1}^{n} |x_i - x_i'| \tag{5}$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum\nolimits_{i=1}^{n} (x_i - x_i')^2} \tag{6}$$

TABLE 4. Evaluation metrics used in this studies and their expressions.

Evaluation Method	Expression	Description
TP Rate	TP rate = TP/(TP+FN)	TP is True Positive
FP Rate	FP Rate = FP/(FP+TN)	TN is True Negative
Corrected Classified Instances: CCI (Accuracy)	CCI = (TP+TN)/(TP+TN+FP+FN)	FP is False Positive FN is False Negative
Precision	Precision = TP/(TP+FP)	
Recall	Recall = TP/(TP+FN)	
F-measure	F-measure = (2×Precision×Recall)/(Precision+Recall)	

Kappa	K = [K0-Ke]/[1-Ke]	Ke= [(TN+FN) × (TN+FP)+(FP+TP) × (FN+TP)]/n^2
		K0 = (TN+TP)/n

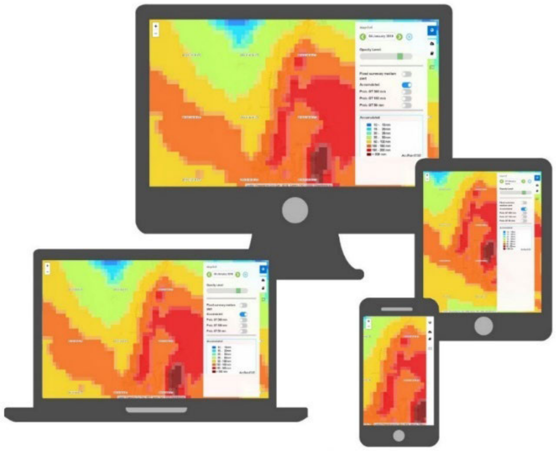


FIGURE 11. Examples of a screenshot of the developed web application fitted on various devices based on responsive Graphic User Interface (GUI) de-sign, based on Bootstrap framework.

where n was the number of data samples, x_i and $x_i^{'}$ were the actual and predicted values, respectively. All other relevant accuracy assessment metrics, namely, TP and FP rates, CCI, precisions, recall, F-measure, and Kappa and their expressions are listed in Table 4.

IV. RESULTS AND DISCUSSION

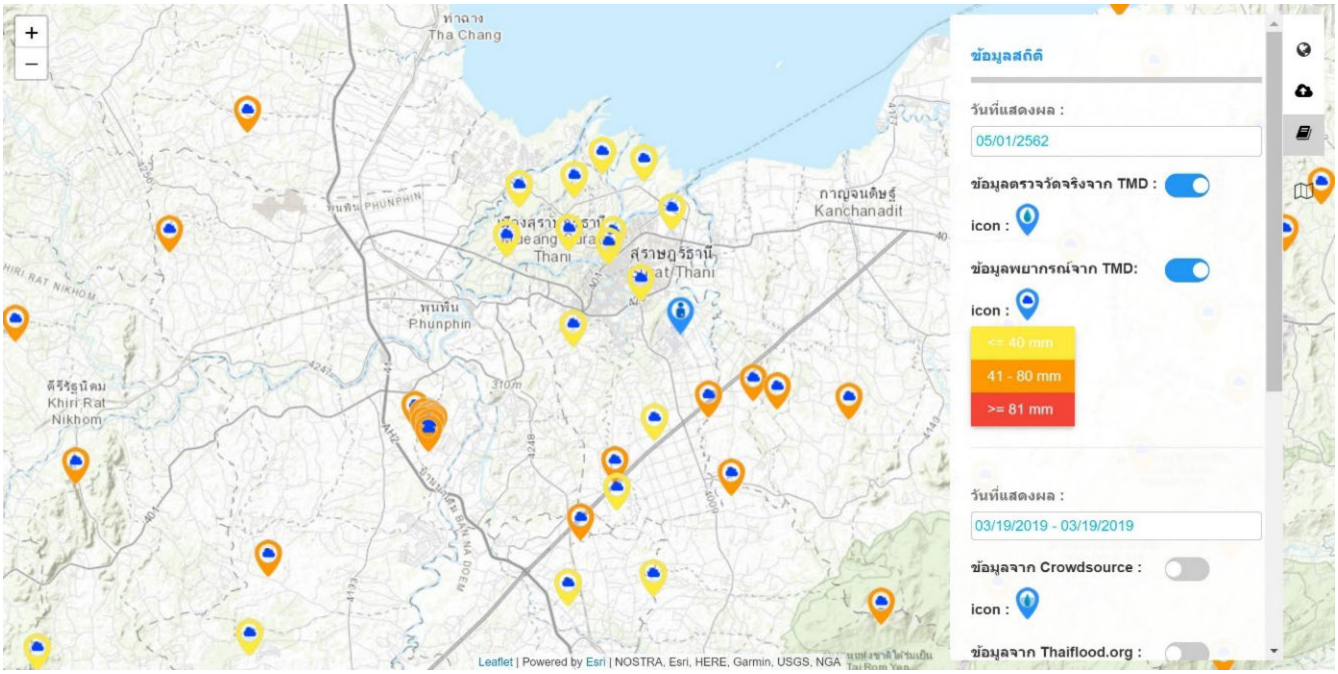
To ensure maximum versatility and most extensive coverage of crowdsourcing, this study adopted responsive web design paradigm in developing the website. The web application once deployed was able to support various devices, ranging from personal and portable computers to mobile phones and tablet computers, with varying screen sizes and resolutions. Examples of their screenshots are shown in Figure 11. By using the Bootstrap framework, the rendered CSS and JavaScript automatically aligned and adjusted the layout of components and controls to maintain uniform appearance and hence satisfying user experience (UX). Furthermore, the development cost was minimized as focuses were placed on core functionalities and customizable contents, instead of variations of frontend interfaces and layouts across platforms.

The results of the first module were rainfall data fetched via TMD API. They were stored in our geospatial database for further processing and partially displayed for a specific area on a user screen. Data consisted of two components, i.e., those estimated by the system and acquired from TMD. In Figure 12, they were represented by cloud and droplet pins, respectively. Each pin was color coded with yellow, orange, and red pins indicating rainfalls of the levels lower than 40 mm, 41 – 80 mm, and greater than 81 mm, respectively. Users are also able choose either or both components, and at a specific duration, simply by enabling the desired layers.

Meteorological and hydrological data that were acquired from GLOFAS are illustrated in Figure 13. In this figure, accumulated precipitation in each area was color coded. Specifically, blue, light green and dark red, represents low (10 – 25 mm), moderate (25 – 100 mm), and high (> 100 mm), respectively. Areas with neutral one indicated those with no precipitation. Likewise, the data would also be used in forecasting.

Another crucial information was the probability of precipitation. In this study, the respective probability in each area was color coded in three levels, i.e., 50 mm, 150 mm, and 300 mm. Examples of these levels are illustrated in Figures 14 to 16. In each figure, pixel intensities indicate the likelihood of precipitation (i.e., 0 to 1) at that location being of the respective levels. In Figure 14, for instance, light and dark green indicate low and high probability of 50 mm precipitation in a given location. Similarly, Figures 15 and 16 display the probability of precipitation of 150 mm and 300 in blue and red, respectively.

One of the main contributions of this paper was employing crowdsource data, not only in training but also in verification. The system obtained data from thaiflood.org, to which public users could send notifications of flood situations. It is worth noted that, without dedicate equipment, accurate quantification of relevant data is prohibitive in practice, especially when provided by the public. Instead of requesting an explicit number, the proposed system relied on GUI, by which a participant could provide the experienced factors. Particularly, they could choose, for instance, one out of four different rainfall levels, i.e., none, low, moderate, or heavy, as they had actually experienced. Their choices were later converted to rating numbers, as described in Table 2. Figure 17 illustrates an example of actual flood map. The crowdsource data were represented by pin icons. Each pin is coded in three different colors with respect to reported flood levels.



ข้อมูลสถิติ

วันที่แสดงผล : 05/01/2562

ข้อมูลตรวจวัดจริงจาก TMD : ● on

icon :

ข้อมูลพยากรณ์จาก TMD: ● on

icon :

🟨 <= 40 mm
🟧 41 - 80 mm
🟥 >= 81 mm

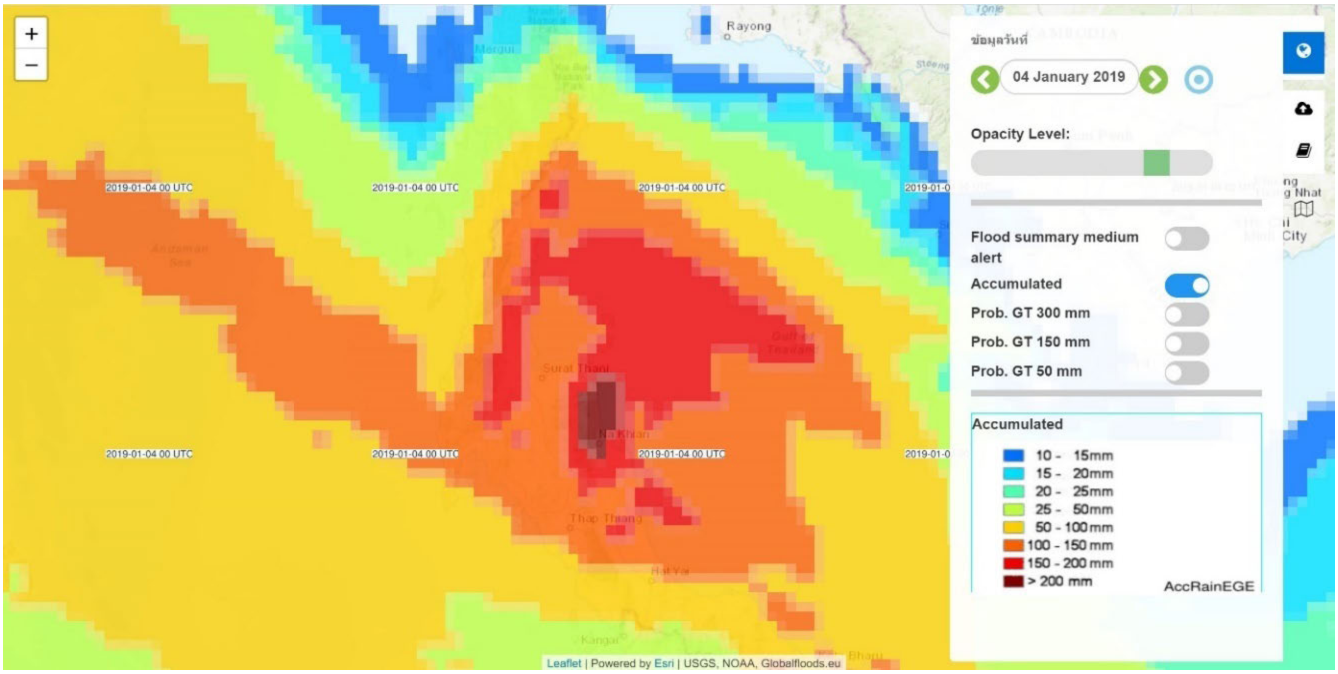
วันที่แสดงผล : 03/19/2019 - 03/19/2019

ข้อมูลจาก Crowdsource : ○ off

icon :

ข้อมูลจาก Thaiflood.org : ○ off

FIGURE 12. Rainfalls estimated by the system (cloud pins) and acquired from TMD (droplet pins) at a given date. Different levels of rainfalls are represented by yellow, orange, and red colors, respectively.



ข้อมูลวันที่

04 January 2019

Opacity Level:

Flood summary medium alert ○ off

Accumulated ● on

Prob. GT 300 mm ○ off

Prob. GT 150 mm ○ off

Prob. GT 50 mm ○ off

Accumulated

🟦	10 - 15mm
🟦	15 - 20mm
🟩	20 - 25mm
🟩	25 - 50mm
🟨	50 - 100mm
🟧	100 - 150 mm
🟥	150 - 200 mm
🟫	> 200 mm

AccRainEGE

FIGURE 13. Accumulated precipitation acquired from GLOFAS. Different levels of accumulation are represented by blue, light green, and dark red colors, respectively.

Yellow, orange, and red pins represented flood levels of less than 20 cm, 20 – 50 cm, and greater than 50 cm, respectively. Without crowdsourcing data, there would be no better means of acquiring rainfall duration and its intensity, as well as drainage problem, which were all crucial in, for examples, the DT (J48) and Fuzzy rules.

Flood forecasting was rendered based on learning of thematic data by an ML method. The resultant forecast was displayed on a web browser. With this platform, users may access this information from various devices. An example of forecasted flood is shown in Figure 18. The strength of its impacts was determined based its level, and accordingly

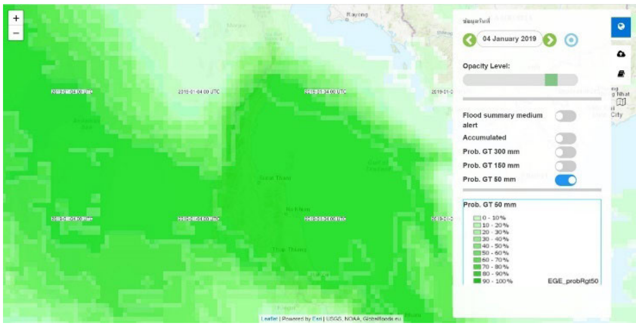


FIGURE 14. Probability map of precipitation level of 50 mm is color coded in green.

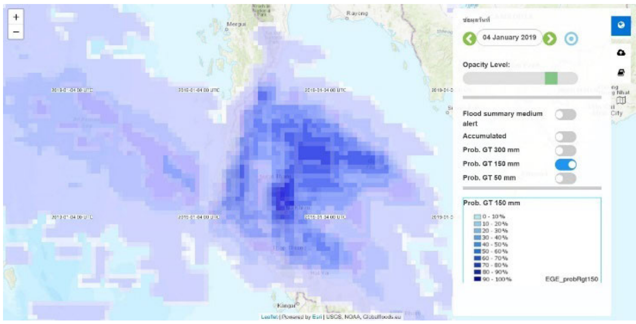


FIGURE 15. Probability map of precipitation level of 150 mm is color coded in blue.

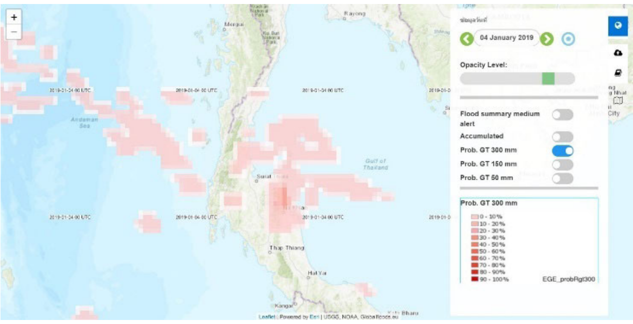


FIGURE 16. Probability map of precipitation level of 150 mm is color coded in red.

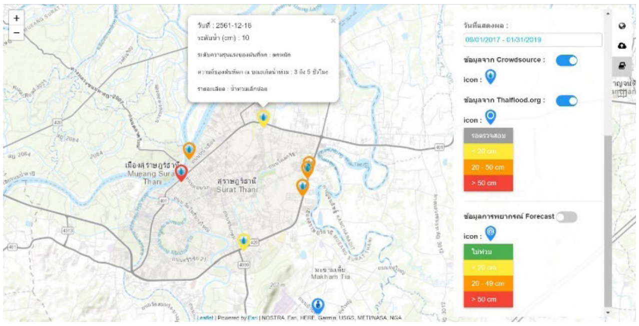


FIGURE 17. Crowdsource data, reported by thaiflood.org users, indicate different levels of actual flood incidents.

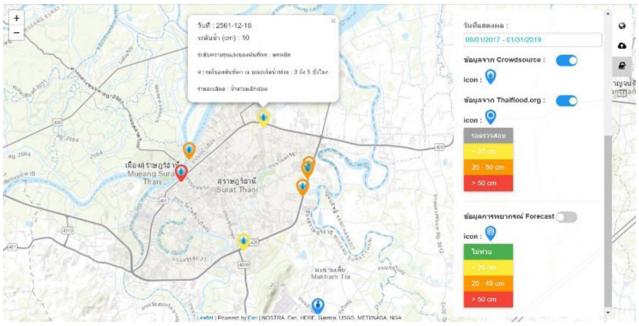


FIGURE 18. Example of forecasted flood, at a selected area and during specific dates.

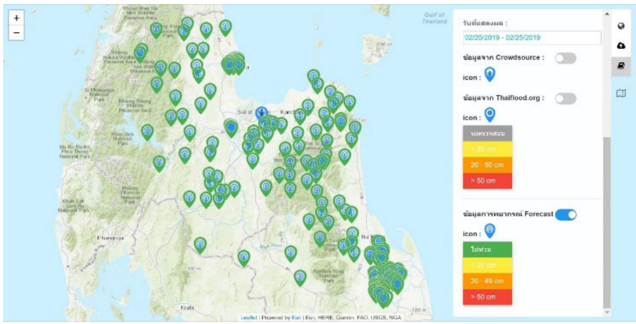


FIGURE 19. Example of when at a selected area, no flood was anticipated, during specific dates.

coded in different colors. Green pin indicate that the area was not at all affected by flood. On the contrary, yellow, orange, and red pin represents flooded areas, whose levels were less than 20 cm, 20 – 49 cm, and greater than 50 cm, respectively. Figure 19 shows an example when no flood incident was anticipated.

In addition to visual assessment and UX demonstration, numerical evaluations were also carried out. The forecasted floods were verified with reports made by informed authorized persons (in order to ensure validity). The metrics evaluated were correctly classified instances ratio, Kappa statistics, MAE, RMSE, TP, FP, precision, recall, F-measure, and ROC area. Comparisons of these metrics were made among various ML strategies, i.e., DT (J48), RF, Naïve Bayes, MLP ANN, RBF ANN, SVM, and fuzzy logic. Each ML was alternately trained and test in turn, while the other processing components, as well as training and verification data were kept unchanged.

The performance of these MLs are listed in Tables 5 and 6. It is evident that MLP ANN, SVM, and RF were placed in top three ranks, in terms of classification accuracies (i.e., 97.83%, 96.67%, and 96.67%) and Kappa coefficients (i.e., 0.89, 0.84, and 0.84). Statistically, Kappa values of greater than 0.8 indicate highly accurate forecast, while those between 0.4 – 0.8 were moderate performers. According to Table 5, DT, Naïve Bayes, and fuzzy logic fell in the latter category. These trends were similarly exhibited in MAE and RMSE, the closer to 0, the more accurate the forecast.

Taken into account the results of, not only flooded regions but also those unaffected by flood, the balance between relevant and irrelevant predicted samples had to

TABLE 5. Assessment of flood forecasting accuracy based on different machine learning strategies.

ML	Correctly Forecasted	Kappa	MAE	RMSE
Decision Tree (J48)	95.33%	0.77	0.02	0.14
Random Forest	**96.67%**	**0.84**	**0.02**	**0.10**
Naive Bayes	88.00%	0.59	0.61	0.24
Artificial Neural Network (MLP)	**97.83%**	**0.89**	**0.01**	**0.10**
Artificial Neural Network (RBF)	95.50%	0.80	0.02	0.14
Support Vector Machine	**96.67%**	**0.84**	**0.01**	**0.12**
Fuzzy Logic	95.67%	0.79	0.02	0.13

TABLE 6. Assessment of flood forecasting performance based on different machine learning strategies.

ML	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	PRC Area
Decision Tree (J48)	0.950	0.010	0.950	0.950	0.950	0.906	0.960
Random Forest	**0.967**	**0.117**	**0.966**	**0.967**	**0.966**	**0.997**	**0.990**
Naive Bayes	0.880	0.035	0.938	0.880	0.898	0.977	0.972
Artificial Neural Network (MLP)	**0.978**	**0.027**	**0.980**	**0.978**	**0.978**	**0.996**	**0.988**
Artificial Neural Network (RBF)	0.955	0.054	0.962	0.955	0.957	0.968	0.950
Support Vector Machine	**0.967**	**0.079**	**0.968**	**0.967**	**0.967**	**0.944**	**0.951**
Fuzzy Logic	0.957	0.105	0.956	0.957	0.955	0.961	0.959

be considered. More specifically, TP and FP rates, precision, recall, F-measure, ROC and PRC areas, were evaluated. The findings shown in Table 6 were consistent to the above accuracy assessments. The best three MLs, in terms of these balanced metrics, were MLP ANN, SVM and RF. Based on these observations, it is therefore safe to conclude that these are the most appropriate MLs for the proposed flood forecasting system. Since MLP ANN, SVM, and RF strategies gave the highest averaged performance overall, further detailed analyses were then performed. As a guideline on employing an ML strategy in practice, it should give not only high expectations but also consistent ones. Figure 20. shows Box-Whisker plots of four key metrics, i.e., MAE, RMSE, accuracy and Kappa. The metrics were evaluated on predicted flood levels, which were divided into four categories, i.e., no flood, low (< 20 cm), moderate (20 – 49 cm) and heavy (> 50 mm) flood. The correct instance means the forecasted level corresponded to the actual event, and vice versa.

It is evident from these graphs that, despite highly accurate forecasting results, the system with SVM exhibited relatively much wider variability across this dataset. On the contrary, with identical settings, the other MLs were more consistent and hence reliable in all four metrics, and thus are recommended in an actual implementation. In fact, depending on system requirements, technical preferences, and computing architecture involved, either MLP ANN or RF are equally applicable.

Additionally, suppose that an MLP ANN is preferred, due to its consistent favorable performance. Convergence analysis was performed on its RMSE measure. This value versus the number of iterations is plotted in Figure 21. It was evident from this graph that RMSE significantly improved up to approximately 250th round. After that, it started to converge until approximately 500th round, when no improvement was noticed. This result serves as a preliminary guideline on training the MLP ANN model. Likewise, convergence analyses on other parameter settings can follow the same suite, given a new set of areas, that may differ in terms of meteorological, hydrological, and geospatial characteristics than the two provinces, considered in this study.

In addition to ML evaluations, those on the developed flood forecasting framework were also performed. To this end, a questionnaire querying satisfactions on using the software was handed out to two user groups, i.e., general users (participants/ flood victims) and authorized officers (including rescue teams, heads of communities, and government personnel). There were 100 subjects, in total, answering this survey. They consisted of 85 general users and 15 officers, respectively. The questionnaire items addressed three aspects of the software, i.e., functionality, efficiency and security, and decision making, flood preparedness and response supports, in five Likert scales. The answered formed were evaluated separately for each group. The resultant scores were listed

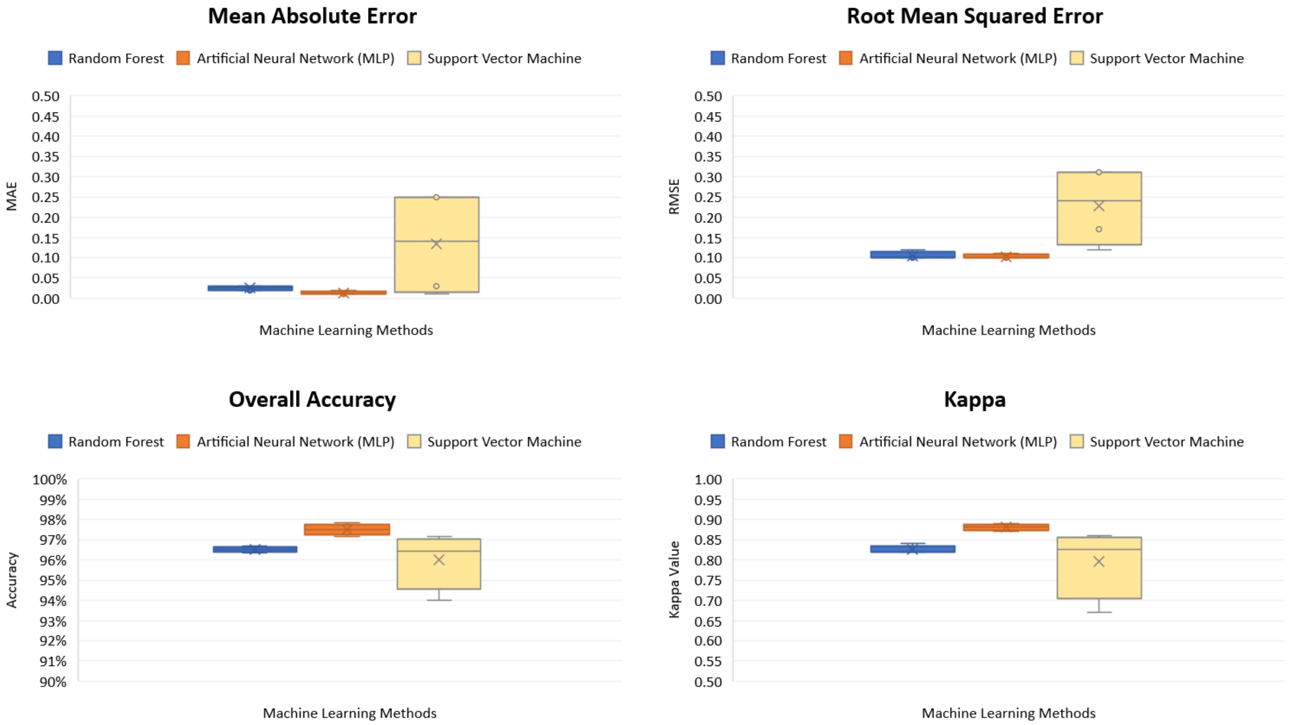


FIGURE 20. MAE, RMS, accuracy and Kappa variabilities in three best performing MLs.

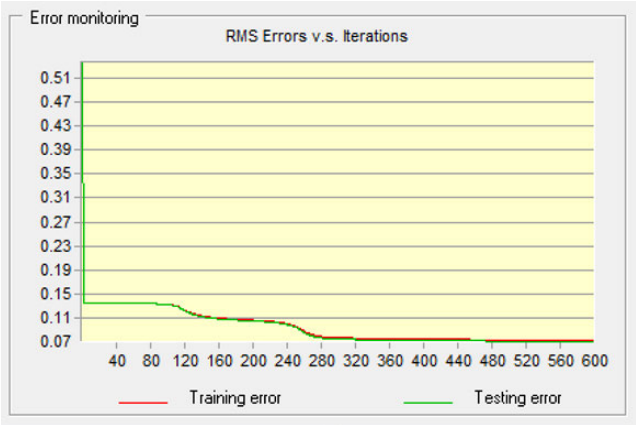


FIGURE 21. RMSE produced by MLP ANN.

in Table 7 and 8, for authorized officers and general users, respectively.

In general, it was revealed that the overall evaluations were very good. The average scores were 4.93 ± 0.12 and 4.76 ± 0.44, as acknowledged by authorized officer and general user groups, respectively. Their views on each of these aspects also reflected their general opinion. It was, however, worth pointing out that, the officers gave the highest score to the system efficiency and security, while the general users did to decision making and flood preparedness and response support. This evident implies primary concerns, expectations, and hence satisfactions, inherently exhibited in each group.

TABLE 7. Framework evaluations results, answered by authorized officers.

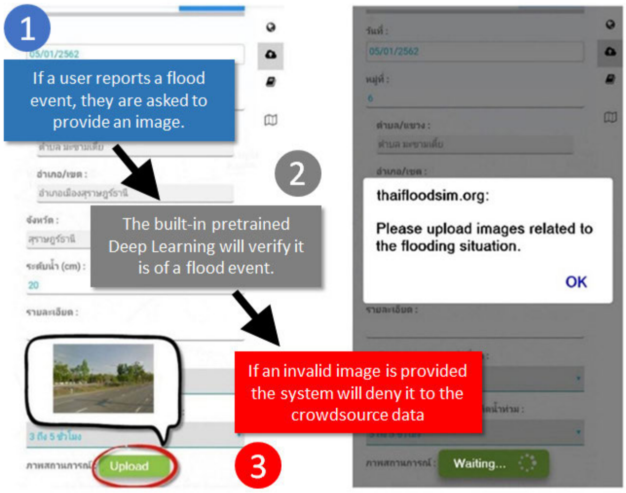
Software Aspects	$\bar{x}$	SD	Meaning
Functionality	4.91	0.13	Very Good
Efficiency and Security	4.96	0.08	Very Good
Decision making and flood preparedness and response, support	4.93	0.14	Very Good
Average	**4.93**	**0.12**	**Very Good**

TABLE 8. Framework evaluations results, answered by participant and flood victims.

Software Aspects	$\bar{x}$	SD	Meaning
Functionality	4.74	0.46	Very Good
Efficiency and Security	4.76	0.45	Very Good
Decision making and flood preparedness and response, support	4.79	0.40	Very Good
Average	**4.76**	**0.44**	**Very Good**

It also served as a guideline on the key aspects, on which one needs to focus, when developing flood forecasting system.

It is worth noted that, in addition to official geospatial, meteorological and hydrological data, this study took into account crowdsource ones for enhancing its classification. The accounts of the actual event were gathered from the



If a user reports a flood event, they are asked to provide an image.

The built-in pretrained Deep Learning will verify it is of a flood event.

If an invalid image is provided the system will deny it to the crowdsource data

thaifloodsim.org:

Please upload images related to the flooding situation.

FIGURE 22. Crowdsource data pre-screening module based on deep learning.

public in real-time. The crowdsource rainfalls, for instance, served as a confirmation of that forecasted by TMD system. Moreover, drainage issues given by crowdsource at specific time could enhance that estimated only by geospatial calculations. Having staid that, the less trustworthy of crowdsource data, the much adverse effect on the system performance was caused. To elevate this issue, the developed system authorized heads of their community to verify the data provided by their members. Moreover, to facilitate this screening, a user was asked to provide a photograph of flooding, taken exactly when the event was notified. An off-the-shelf deep image learning module (illustrated in Figure 22) was subsequently used to verify this photo and hence associated report.

The deep image learning module was implemented by using Clarify API [63]. The API was based on ZFNet architecture, which was improvement over the AlexNet. It was trained by 200 images of equally flood and non-flood events. This process was employed to ensure the reliability of crowdsourcing.

V. CONCLUSION

This paper proposed a novel distributed flood forecasting system, based on integrating meteorological, hydrological, geospatial, and crowdsource data. Big data made available by prominent agencies were acquired by means of various cross-platform APIs. Forecasting was performed based on these data learned by modern ML strategies. They were decision tree, RF, Naïve Bayes, MLP and RBF ANN, SVM, and fuzzy logics. Evaluation results on studied areas indicated that the system could forecasted flood events highly accurately. Three best performing MLs were MLP ANN, SVM, and RF, respectively. It was elucidated empirically that the developed system could be used to alert the public and authorities alike of not only a current flood but also future ones. This system also enhanced user experience via responsive graphical interfaces, interoperable on different computing devices including mobiles. This advantage effectively encouraged greater contribution of crowdsource data from the public, enriching data aggregation and hence increasing system accuracy and reliability. As such, the developed system is adaptive, in a sense that as the system became more "experienced" (i.e., through learnings), the forecasting gets more realistic.

In prospects, the system can be readily employed in existing floods management schemes, e.g., those led by government agencies or non-profit organizations. Moreover, thanks to distributed architecture, the system can reach wider public, and therefore serves as an effective means of communicating with them (and especially the flood victims), regarding current status and development of the disaster. Future improvements of the system include initial flood representation and its extent being adapted to the current location of the device, so that they can be instantly made aware of by its user. Moreover, flooded location pined by an icon may be augmented with color-coded regions, so that the conditions (e.g., levels and extents) of affected areas may be better comprehended.

Some data considered in this study, such as GLOFAS, were not of intrinsically high spatial resolution, However, they were accurate. The accumulated precipitations and their probability at different levels, for instance, corresponded to the actual event. Their API system was also reliable. These characteristics were favored by the proposed system. Their resolution shortcomings were remedied by incorporating other more detailed layers as well as crowdsource factors into the ML framework. Possible improvements for this issue include involving the Internet of Things (IoT) in measuring actual meteorological data with preferred coverage.

ACKNOWLEDGMENT

The authors would like to thank Geo-Informatics and Space Technology Development Agency (Public Organization), Thailand Meteorological Department, Google Inc., and the GLOFAS website for providing the remotely sensed data (Meteorological and Hydrological Data) used in the preparation of this article.

REFERENCES

[1] *Natural Disaster Data Book*, ADRC, Hyōgo, Japan, 2012.

[2] *Natural Disaster Data Book*, ADRC, Hyōgo, Japan, 2015.

[3] D. Komori, S. Nakamura, M. Kiguchi, A. Nishijima, D. Yamazaki, S. Suzuki, A. Kawasaki, K. Oki, and T. Oki, "Characteristics of the 2011 Chao Phraya River flood in central Thailand," *Hydrolog. Res. Lett.*, vol. 6, no. 6, pp. 41–46, 2012.

[4] P. Promchote, S.-Y. S. Wang, and P. G. Johnson, "The 2011 great flood in Thailand: Climate diagnostics and implications from climate change," *J. Climate*, vol. 29, no. 1, pp. 367–379, Jan. 2016.

[5] S. Puttinaovarat, P. Horkaew, K. Khaimook, and W. Polnigongit, "Adaptive hydrological flow field modeling based on water body extraction and surface information," *J. Appl. Remote Sens*, vol. 9, no. 1, Dec. 2015, Art. no. 095041.

[6] S. K. Jain, P. Mani, S. K. Jain, P. Prakash, V. P. Singh, D. Tullos, S. Kumar, S. P. Agarwal, and A. P. Dimri, "A Brief review of flood forecasting techniques and their applications," *Int. J. River Basin Manage.*, vol. 16, no. 3, pp. 329–344, Jul. 2018.

[7] N. Belabid, F. Zhao, L. Brocca, Y. Huang, and Y. Tan, "Near-real-time flood forecasting based on satellite precipitation products," *Remote Sens.*, vol. 11, no. 3, p. 252, Jan. 2019.

[8] A. K. Kar, A. Lohani, N. Goel, and G. Roy, "Rain gauge network design for flood forecasting using multi-criteria decision analysis and clustering techniques in lower Mahanadi river basin, India," *J. Hydrol., Regional Stud.*, vol. 4, pp. 313–332, Sep. 2015.

[9] M. Dembélé and S. J. Zwart, "Evaluation and comparison of satellite-based rainfall products in Burkina Faso, West Africa," *Int. J. Remote Sens.*, vol. 37, no. 17, pp. 3995–4014, Sep. 2016.

[10] A. Foehn, J. G. Hernández, B. Schaefli, and G. De Cesare, "Spatial interpolation of precipitation from multiple rain gauge networks and weather radar data for operational applications in Alpine catchments," *J. Hydrol.*, vol. 563, pp. 1092–1110, Aug. 2018.

[11] F. Cecinati, A. Moreno-Ródenas, M. Rico-Ramirez, M.-C. Ten Veldhuis, and J. Langeveld, "Considering rain gauge uncertainty using kriging for uncertain data," *Atmosphere*, vol. 9, no. 11, p. 446, Nov. 2018.

[12] S. Stisen and M. Tumbo, "Interpolation of daily raingauge data for hydrological modelling in data sparse regions using pattern information from satellite data," *Hydrolog. Sci. J.*, vol. 60, no. 11, pp. 1911–1926, Sep. 2015.

[13] Q. Hu, Z. Li, L. Wang, Y. Huang, Y. Wang, and L. Li, "Rainfall spatial estimations: A review from spatial interpolation to multi–source data merging," *Water*, vol. 11, no. 3, p. 579, Mar. 2019.

[14] C. Berndt and U. Haberlandt, "Spatial interpolation of climate variables in Northern Germany—Influence of temporal resolution and network density," *J. Hydrol., Regional Stud.*, vol. 15, pp. 184–202, Feb. 2018.

[15] S. Zhang, J. Zhang, Y. Liu, and Y. Liu, "A mathematical spatial interpolation method for the estimation of convective rainfall distribution over small watersheds," *Environ. Eng. Res.*, vol. 21, no. 3, pp. 226–232, Sep. 2016.

[16] A. Mosavi, P. Ozturk, and K.-W. Chau, "Flood prediction using machine learning models: Literature review," *Water*, vol. 10, no. 11, p. 1536, Oct. 2018.

[17] H. Cloke and F. Pappenberger, "Ensemble flood forecasting: A review," *J. Hydrol.*, vol. 375, nos. 3–4, pp. 613–626, Sep. 2009.

[18] T. E. Adams and T. C. Pagano, *Flood Forecasting: A Global Perspective*. New York, NY, USA: Academic, 2016.

[19] Q. Liang, Y. Xing, X. Ming, X. Xia, H. Chen, X. Tong, and G. Wang, "An open-source modelling and data system for near real-time flood," in *Proc. 37th IAHR World Congr.* Tyne, U.K.: Newcastle Univ., 2017, pp. 1–10.

[20] L.-C. Chang, F.-J. Chang, S.-N. Yang, I.-F. Kao, Y.-Y. Ku, C.-L. Kuo, and I. Amin, "Building an intelligent hydroinformatics integration platform for regional flood inundation warning systems," *Water*, vol. 11, no. 1, p. 9, Dec. 2018.

[21] S. Puttinaovarat, P. Horkaew, and K. Khaimook, "Conguring ANN for inundation areas identication based on relevant thematic layers," *ECTI Trans. Comput. Inf. Technol. (ECTI-CIT)*, vol. 8, no. 1, pp. 56–66, May 2014.

[22] G. Tayfur, V. Singh, T. Moramarco, and S. Barbetta, "Flood hydrograph prediction using machine learning methods," *Water*, vol. 10, no. 8, p. 968, Jul. 2018.

[23] C. Choi, J. Kim, J. Kim, D. Kim, Y. Bae, and H. S. Kim, "Development of heavy rain damage prediction model using machine learning based on big data," *Adv. Meteorol.*, vol. 2018, pp. 1–11, Jun. 2018.

[24] J. A. Pollard, T. Spencer, and S. Jude, "Big data approaches for coastal flood risk assessment and emergency response," *WIREs Climate Change*, vol. 9, no. 5, p. e543, Sep. 2018.

[25] N. Tkachenko, R. Procter, and S. Jarvis, "Predicting the impact of urban flooding using open data," *Roy. Soc. Open Sci.*, vol. 3, no. 5, May 2016, Art. no. 160013.

[26] K. Sene, *Flash Floods: Forecasting and Warning*. Amsterdam, The Netherlands: Springer, 2012.

[27] *Manual on Flood Forecasting and Warning*, World Meteorolog. Org., Geneva, Switzerland, 2011.

[28] X. Chen, L. Zhang, C. J. Gippel, L. Shan, S. Chen, and W. Yang, "Uncertainty of flood forecasting based on radar rainfall data assimilation," *Adv. Meteorol.*, vol. 2016, pp. 1–12, Aug. 2016.

[29] L. Alfieri, M. Berenguer, V. Knechtl, K. Liechti, D. Sempere-Torres, and M. Zappa, "Flash flood forecasting based on rainfall thresholds," in *Handbook of Hydrometeorological Ensemble Forecasting*. Berlin, Germany: Springer-Verlag, 2019, pp. 1223–1260.

[30] J. Ye, Y. Shao, and Z. Li, "Flood forecasting based on TIGGE precipitation ensemble forecast," *Adv. Meteorol.*, vol. 2016, pp. 1–9, Nov. 2016.

[31] M. Santos and M. Fragoso, "Precipitation thresholds for triggering floods in the Corgo Basin, Portugal," *Water*, vol. 8, no. 9, p. 376, Aug. 2016.

[32] F. Dottori, M. Kalas, P. Salamon, A. Bianchi, L. Alfieri, and L. Feyen, "An operational procedure for rapid flood risk assessment in Europe," *Nat. Hazards Earth Syst. Sci.*, vol. 17, no. 7, pp. 1111–1126, Jul. 2017.

[33] C. Li, X. Cheng, N. Li, X. Du, Q. Yu, and G. Kan, "A framework for flood risk analysis and benefit assessment of flood control measures in urban areas," *Int. J. Environ. Res. Public Health*, vol. 13, no. 8, p. 787, Aug. 2016.

[34] D. Cane, S. Ghigo, D. Rabuffetti, and M. Milelli, "Real-time flood forecasting coupling different postprocessing techniques of precipitation forecast ensembles with a distributed hydrological model. The case study of May 2008 flood in western Piemonte, Italy," *Natural Hazards Earth Syst. Sci.*, vol. 13, no. 2, pp. 211–220, Feb. 2013.

[35] Thailand Meteorological Department. (2019). *TMD Big Data*. [Online]. Available: http://www.rnd.tmd.go.th/bigdata.php

[36] Globalfloods. (2019). *GLOFAS*. [Online]. Available: http://www.globalfloods.eu/accounts/login/?next=/glofas-forecasting/

[37] D. D. Konadu and C. Fosu, "Digital elevation models and GIS for watershed modelling and flood prediction—A case study of Accra Ghana," in *Appropriate Technologies for Environmental Protection in the Developing World*. Dordrecht, The Netherlands: Springer, pp. 325–332, Feb. 2009.

[38] J. García-Pintado, D. C. Mason, S. L. Dance, H. L. Cloke, J. C. Neal, J. Freer, and P. D. Bates, "Satellite-supported flood forecasting in river networks: A real case study," *J. Hydrol.*, vol. 523, pp. 706–724, Apr. 2015.

[39] M. Shafapour Tehrany, F. Shabani, M. Neamah Jebur, H. Hong, W. Chen, and X. Xie, "GIS-based spatial prediction of flood prone areas using standalone frequency ratio, logistic regression, weight of evidence and their ensemble techniques," *Geomatics, Natural Hazards Risk*, vol. 8, no. 2, pp. 1538–1561, Dec. 2017.

[40] S. Ramly, W. Tahir, and S. N. H. S. Yahya, "Enhanced flood forecasting based on land–use change model and radar–based quantitative precipitation estimation," in *Proc. ISFRAM*. Singapore: Springer, 2015, pp. 305–317.

[41] R. E. Emerton, E. M. Stephens, F. Pappenberger, T. C. Pagano, A. H. Weerts, A. W. Wood, P. Salamon, J. D. Brown, N. Hjerdt, C. Donnelly, C. A. Baugh, and H. L. Cloke, "Continental and global scale flood forecasting systems," *WIREs Water*, vol. 3, no. 3, pp. 391–418, May 2016.

[42] Google. (2019). *Google Map API*. [Online]. Available: https://developers.google.com/maps/documentation/javascript/tutorial

[43] A. Pejic, S. Pletl, and B. Pejic, "An expert system for tourists using Google maps API," in *Proc. 7th Int. Symp. Intell. Syst. Inform.*, Sep. 2009, pp. 317–322.

[44] Y. Wang, G. Huynh, and C. Williamson, "Integration of Google maps/earth with microscale meteorology models and data visualization," *Comput. Geosci.*, vol. 61, pp. 23–31, Dec. 2013.

[45] P. Aplin, G. Priestnall, G. Harvey, and N. Mount, *Representing, Modeling, and Visualizing the Natural Environment*. Boca Raton, FL, USA: CRC Press, 2008.

[46] S. Nedkov and S. Zlatanova, "Google maps for crowdsourced emergency routing," *Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci.*, vol. XXXIX-B4, pp. 477–482, Aug. 2012.

[47] J. P. De Albuquerque, M. Eckle, B. Herfort, and A. Zipf, "Crowdsourcing geographic information for disaster management and improving urban resilience: An overview of recent developments and lessons learned," in *European Handbook of Crowdsourced Geographic Information*. London, U.K.: Ubiquity Press, Aug. 2016, pp. 309–321.

[48] J. Fohringer, D. Dransch, H. Kreibich, and K. Schröter, "Social media as an information source for rapid flood inundation mapping," *Natural Hazards Earth Syst. Sci.*, vol. 15, no. 12, pp. 2725–2738, Dec. 2015.

[49] G. Schimak, D. Havlik, and J. Pielorz, "Crowdsourcing in crisis and disaster management—Challenges and considerations," in *Environmental Software Systems. Infrastructures, Services and Applications*. Cham, Switzerland: Springer, 2015, pp. 56–70.

[50] T. Simon, A. Goldberg, and B. Adini, "Socializing in emergencies—A review of the use of social media in emergency situations," *Int. J. Inf. Manage.*, vol. 35, no. 5, pp. 609–619, Oct. 2015.

[51] F. E. A. Horita, L. C. Degrossi, L. F. G. de Assis, A. Zipf, and J. P. De Albuquerque, "The use of volunteered geographic information (VGI) and crowdsourcing in disaster management: A systematic literature review," in *Proc. 19th Americas Conf. Inf. Syst.*, 2013, pp. 1–10.

[52] M. Zook, M. Graham, T. Shelton, and S. Gorman, "Volunteered geographic information and crowdsourcing disaster relief: A case study of the haitian earthquake," *World Med. Health Policy*, vol. 2, no. 2, pp. 6–32, Jan. 2010.

[53] L. See, P. Mooney, G. Foody, L. Bastin, A. Comber, J. Estima, S. Fritz, N. Kerle, B. Jiang, M. Laakso, H.-Y. Liu, G. Milčinski, M. Nikšič, M. Painho, A. Põdör, A.-M. Olteanu-Raimond, and M. Rutzinger, "Crowdsourcing, citizen science or volunteered geographic information? The current state of crowdsourced geographic information," *ISPRS Int. J. Geo-Inf.*, vol. 5, no. 5, p. 55, Apr. 2016.

[54] P.-S. Yu, T.-C. Yang, S.-Y. Chen, C.-M. Kuo, and H.-W. Tseng, "Comparison of random forests and support vector machine for real-time radar-derived rainfall forecasting," *J. Hydrol.*, vol. 552, pp. 92–104, Sep. 2017.

[55] M. Dehghani, B. Saghafian, F. Nasiri Saleh, A. Farokhnia, and R. Noori, "Uncertainty analysis of streamflow drought forecast using artificial neural networks and Monte–Carlo simulation," *Int. J. Climatol.*, vol. 34, no. 4, pp. 1169–1180, Mar. 2014.

[56] A. K. Lohani, N. Goel, and K. Bhatia, "Improving real time flood forecasting using fuzzy inference system," *J. Hydrol.*, vol. 509, pp. 25–41, Feb. 2014.

[57] M. Azam, H. S. Kim, and S. J. Maeng, "Development of flood alert application in Mushim stream watershed Korea," *Int. J. Disaster Risk Reduction*, vol. 21, pp. 11–26, Mar. 2017.

[58] J. Xiaoming, H. Xiaoyan, D. Liuqian, L. Jiren, L. Hui, C. Fuxin, and R. Minglei, "Real-time flood forecasting and regulation system of Poyanghu Lake Basin in China," *EPiC Ser. Eng.*, vol. 3, pp. 2368–2374, 2018.

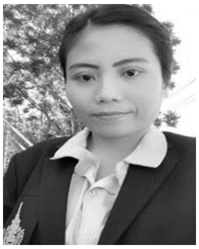
[59] N. A. Sulaiman, N. F. A. Aziz, N. M. Tarmizi, A. M. Samad, and W. Z. W. Jaafar, "Integration of geographic information system (GIS) and hydraulic modelling to simulate floodplain inundation level for Bandar Segamat," in *Proc. IEEE 5th Control Syst. Graduate Res. Colloq.*, Aug. 2014, pp. 114–119.

[60] S. Ghosh, S. Karmakar, A. Saha, M. P. Mohanty, S. Ali, S. K. Raju, and P. L. N. Murty, "Development of India's first integrated expert urban flood forecasting system for Chennai," *Current Sci.*, vol. 117, no. 5, pp. 741–745, 2019.

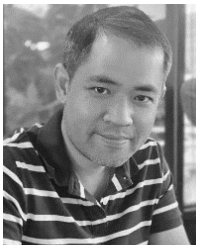
[61] Q. Ran, W. Fu, Y. Liu, T. Li, K. Shi, and B. Sivakumar, "Evaluation of quantitative precipitation predictions by ECMWF, CMA, and UKMO for flood forecasting: Application to two Basins in China," *Natural Hazards Rev.*, vol. 19, no. 2, May 2018, Art. no. 05018003.

[62] E. Mayoraz and E. Alpaydin, "Support vector machines for multiclass classification," in *Proc. Int. Work-Conf. Artif. Neural Netw.* Berlin, Germany: Springer, 1999, pp. 833–842.

[63] E. R. Davies, *Computer Vision: Principles, Algorithms, Applications, Learning*. New York, NY, USA: Academic, 2017.



SUPATTRA PUTTINAOVARAT received the B.B.A. degree (Hons.) and the M.S. degree in management of information technology from the Prince of Songkla University, Thailand, in 2007 and 2010, respectively, and the Ph.D. degree in information technology from the Suranaree University of Technology, Thailand, in 2016. She is currently an Assistant Professor with the Department of Applied Mathematics and Informatics, Faculty of Science and Industrial Technology, Prince of Songkla University, Surat Thani Campus, Surat Thani, Thailand. Her research interests include flood modeling, geographic information systems, remote sensing, and machine learning.



PARAMATE HORKAEW received the B.Eng. degree (Hons.) in telecommunication engineering from the King Mongkut's Institute of Technology, Ladkrabang, Thailand, in 1999, and the Ph.D. degree in computer science from Imperial College London, University of London, London, U.K., in 2004. He is currently an Assistant Professor with the School of Computer Engineering, Suranaree University of Technology, Thailand. His main research interests include computational anatomy, digital geometry processing, computer vision, and graphics.