Flash Floods Prediction Using Precipitable Water Vapor Derived From GPS Tropospheric Path Delays Over the Eastern Mediterranean

Shlomi Ziskin Ziv and Yuval Reuveni

***Abstract*— A flash flood is a rapid and intense response of a drainage area to heavy rainfall events. In the arid and semiarid parts of the Eastern Mediterranean (EM) region, the spatiotemporal distribution of rainfall is the most important factor for flash flood generation. A possible precursor to heavy rainfall events is the rise in tropospheric water vapor amount, which can be remotely sensed using ground-based global navigation satellite system (GNSS) stations. Here, we use the precipitable water vapor (PWV) derived from nine GNSS ground-based stations in the arid part of the EM region in order to predict flash floods. Our approach includes using three types of machine learning (ML) models in a binary classification task, which predicts whether a flash flood will occur given 24 h of PWV data. We train our models with 107 unique flash flood events and vigorously test them using a nested cross-validation technique. The results indicate a good agreement between all three types of models and across various score metrics. In addition, the models are further improved by adding more features such as surface pressure measurements. Finally, a feature importance analysis shows that the most important features are the PWV values from 2 to 6 h prior to a flash flood. These promising results indicate that it is possible to augment the current flash flood warning systems with a near real-time GNSS ground-based data-driven approach as demonstrated in this work.**

***Index Terms*— Eastern Mediterranean (EM), flash floods, global navigation satellite system (GNSS), machine learning (ML), path delays, precipitable water vapor (PWV).**

I. Introduction

FLASH floods are rapid and high-intensity flooding events, which are mainly caused by heavy rainfall. Since flash floods have a short response time of several hours, they are difficult to predict and cause damages and even casualties [1]. Among the factors that control the flash floods generation (e.g., soil saturation and surface cover), the spatiotemporal distribution of rainfall is the most significant one when analyzing hydrological models output [2], [3], [4], [5]. The rainfall regime in the arid and semiarid parts of the Eastern Mediterranean (EM) region is extremely variable and is mostly comprised of short duration, high-intensity events [6], [7], [8]. Therefore, in order to predict flood events, one must first account for the heavy rainfall events' location and timing, which can be achieved by remote sensing platforms (e.g., weather radar [9], [10], [11]). Another option is to measure the water vapor (WV) amount in the atmosphere in order to detect a mass moisture transport that is a perquisite to heavy rainfall events. One such method is the global navigation satellite system (GNSS) meteorology, which can continuously provide a near real-time estimate of the precipitable WV (PWV) above the ground-based receiver's location [12], [13].

GPS satellites that orbit the Earth at ≈20 200 km transmit navigation messages via radio waves to ground-based receivers. Each navigation message includes information that allows the ground-based user to find its position up to centimeter- or even millimeter-level accuracy [14], [15]. This precise point positioning (PPP) method, as it is called, can also be used to estimate the PWV content between each GPS satellite and the ground-based receiver. As they reach Earth, the transmitted radio waves are effectively dispersed by the ionosphere and absorbed by the troposphere [16]. These processes produce a measurable delay in the radio message upon arrival in the ground-based receiver. The ionospheric dispersion effect can be accounted for since GPS satellites transmit the radio waves in at least two frequency bands (e.g., $L1 = 1575.42$ MHz, $L2 = 1227.6$ MHz, and $L5 = 1176.45$ MHz) [17]. The tropospheric delay or the zenith tropospheric delay (ZTD)[1] consists of two major sources of absorption processes: hydrostatic delay or zenith hydrostatic delay (ZHD),[1] which is mainly due to the effect of atmospheric pressure [18] on the radio signal and wet delay that is due to the radio signal's interaction with water molecules [19]. The wet delay can be estimated by subtracting the ZHD from the ZTD.

In each ground-based receiver, the radio messages are stored as text files called Receiver Independent Exchange Format (RINEX) and can be processed by dedicated software such as NASA's JPL GipsyX [20]. The processing involves complex inversion algorithms [21], [22] and is used in order to solve the precise position of the ground-based receiver. From the position solution of the receiver, the ZTD can be extracted,

Manuscript received 4 November 2021; revised 2 June 2022 and 17 July 2022; accepted 17 August 2022. Date of publication 23 August 2022; date of current version 13 September 2022. This work was supported by the Israeli Ministry of Science, Technology and Space under Grant 3-15743. *(Corresponding author: Yuval Reuveni.)*

Shlomi Ziskin Ziv is with the Department of Physics, Ariel University, Ariel 40700, Israel, and also with the Eastern Research and Development Center, Ariel 40700, Israel (e-mail: shlomiziskin@gmail.com).

Yuval Reuveni is with the Department of Physics, Ariel University, Ariel 40700, Israel, also with the Eastern Research and Development Center, Ariel 40700, Israel, and also with the School of Sustainability, Interdisciplinary Center Herzliya, Herzliya 4610101, Israel (e-mail: yuvalr@ariel.ac.il).

Digital Object Identifier 10.1109/TGRS.2022.3201146

[1]The delay is given in zenith values by using mapping functions.

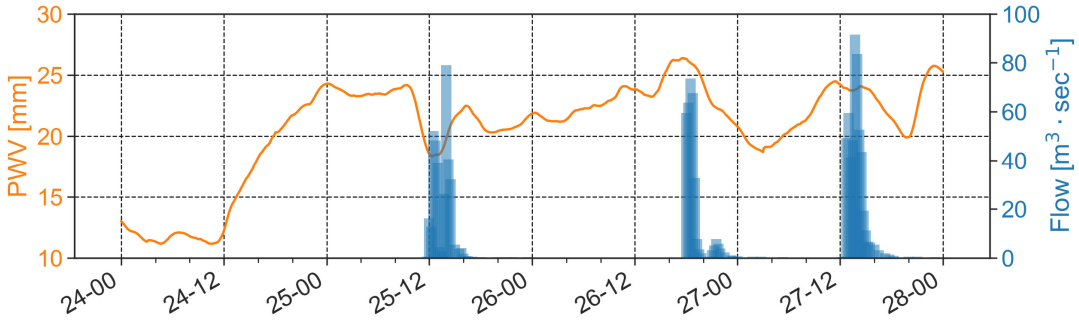

Fig. 1. PWV at Yerucham (YRCM) GNSS station superimposed on the water discharge (flow) at the Mamsheet hydrometric station located 12 km east of YRCM on April 24–27, 2018. Note the three major flash flood events on the 25th, the 26th, and the 27th. The PWV more than doubled during the second half of the 24th as a low-pressure system provided large quantities of moisture to the region.

while the ZHD is provided with the empirical mapping function model files (global mapping function (GMF)/GPT2/GPT3 that are based on climatological means) or from numerical weather data derived from, e.g., global six-hourly surface pressure measurements (Vienna Mapping Function 1—VMF1 data files) [23], [24]. Here, we used the GipsyX default option that is the GMF along with GPT2 data files to obtain the ZHD estimations. The obtained zenith wet delay (ZWD; ZTD-ZHD) is proportional to the total amount of WV in a vertical atmospheric column. Thus, the ZWD can be converted into PWV by using a WV mean atmospheric temperature [18].

In the past 30 years, GNSSderived PWV has been extensively compared to many other remote sensing platforms (e.g., Sun photometers and radiometers), radiosonde *in situ* measurements, and reanalysis products that resulted in an root mean square error (RMSE) ranging from 1 to 3 mm [25], [26], [27], [28], [29], [30], [31]. Furthermore, PWV maps can also be assimilated into modern numerical weather prediction models (e.g., WRF), which effectively lowers the WV prediction RMSE by more than 30% when compared to radiosonde measurements [32].

Using GNSS ground-based meteorology to monitor PWV before, during, and after heavy rainfall events is not new. Bonafoni *et al.* [33] reported at least six papers describing the rise in PWV values when the weather system enters the affected region. Moreover, the heavy precipitation begins only when the PWV reaches its peak value (e.g., [34], [35]). Moore *et al.* [36] showed that during the July 2013 summer monsoon in California, near real-time PWV can detect rapid moisture influx and issue a timely flood warning. Huelsing *et al.* [37] found that prior to the 2013 Colorado flood event, the PWV increased by about 10 mm and later remained almost constant due to the saturated atmosphere. These findings are consistent with our own analysis in the arid area of the EM. For example, Fig. 1 shows the PWV values for YRCM station located 12 km west of Mamsheet hydrometric station for April 24–27, 2018.[2] The PWV values more than doubled during the second half of the 24th event as a low-pressure system entered the region. Moreover, the flood events' peak discharges lag after the closest PWV peak values for the 26th and 27th events, while the 25th double peaked event shows a more complex behavior and can be the results of local and nonlocal coupled sources of humidity as suggested by Lynn *et al.* [38].

GNSS technology is a powerful tool for geoscience remote sensing and natural hazards forecasting, which requires permanent monitoring of the troposphere and ionosphere state on different spatial scales. Several studies in the field of machine learning (ML) associated with GNSS ionospheric total electron content (TEC) were focused on TEC time series prediction algorithms. Sun *et al.* [39] provided a long short-term memory (LSTM)-based model for predicting ionospheric vertical TEC above Beijing using a time sequence, consisting of the daily TEC vectors for their model input, and the output was TEC time series 24 h ahead. Liu *et al.* [40] used the LSTM with several input data, including historical time series of spherical harmonic (SH) coefficients, solar extreme ultra violet (EUV) flux, disturbance storm time index, and hour of the day, for predicting the 256 SH coefficients, traditionally used for constructing global ionospheric maps. Asaly *et al.* [41] used GNSS TEC data along with support vector machine (SVM) training set to build a solar flare X- and M-class predictor. Later on, they also used GNSS TEC data along with the SVM model for potentially predicting strong earthquake events [42]. Hsu [43] used an SVM classifier to separate the type of GNSS pseudorange measurement into three categories: clean, multipath, and nonline of sight, thus evaluating several features which were estimated from the GNSS raw data, including the received signal strength. In addition, he also proposed a new feature to indicate the consistency between measurements of pseudorange and Doppler shift. Linty *et al.* [44] used an ML decision tree and random forest (RF) algorithms, applied with big sets of 50-Hz postcorrelated GNSS data for automatic, accurate, and early detection of amplitude ionospheric scintillation events, reaching a detection accuracy of 98%.

Our goal in this work is to investigate the ability of GNSS-derived PWV to predict flash floods events in the arid part of the EM region using three types of ML models. Accordingly, Section II describes the PWV data and flood events used in this work along with all the ML methodology [e.g., preprocessing, metrics, and cross validation (CV)]. Section III presents the ML models' performance along with a feature importance analysis. We discuss the results in Section IV and our concluding remarks follow in Section V.

[2]These series of flash flood events claimed the lives of 15 people in Israel.

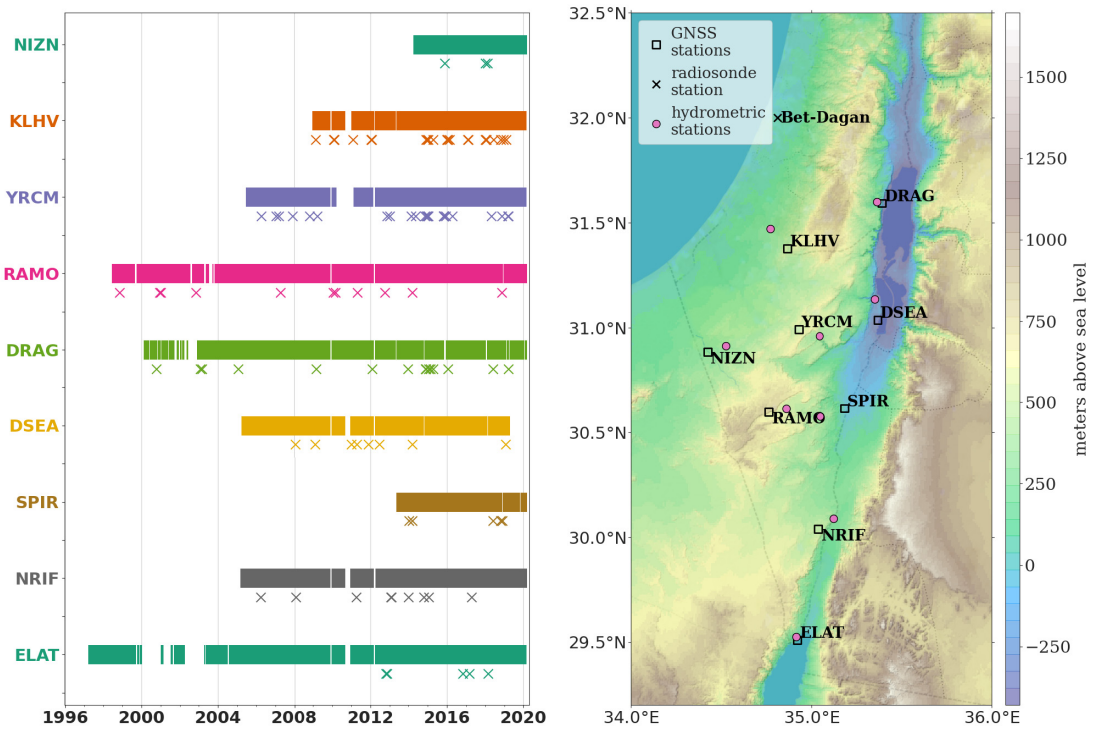

Fig. 2. (Left) PWV data availability for each of the SOI-APN stations in the southern part of Israel. The flash floods' unique events are plotted with x's under each nearest GNSS station. (Right) SOI-APN stations (black squares), Bet-Dagan IMS station (black x), and the hydrometric stations (pink) plotted on a height-filled contour map of the study area.

II. Data and Methodology

The GNSS ground station receiver's network used in this work is the Survey Of Israel Active Permanent Network (SOI-APN). We selected nine stations that are located in the arid climate of southern Israel in the EM. Their names and station IDs are presented in Table I and their locations are indicated on the map in Fig. 2.

Recently, Ziv *et al.* [45] used the SOI-APN stations to investigate the PWV diurnal variations in the EM region. They processed the RINEX files and produced the PWV time series for each station. In this work, we use the aforementioned PWV dataset and preprocess it along with the flood events database for the ML classification task.

Thus, we briefly outline the PWV derivation methods from ground-based receivers along with a description of the flood events dataset in Section II-A. Section II-B describes the insights we gather from analyzing and correlating the flood events with the PWV time series dataset. Finally, in Section II-C, we describe and elaborate upon the ML methodology used in this work, which includes data preprocessing, test score metrics, and the nested cross-validation technique.

Furthermore, in the spirit of reproducible science, we encourage the interested reader to explore the Python repository hosted on GitHub.com (https://github.com/ZiskinZiv/PW_from_GPS), which includes the procedures and ML methodology used in this study.

A. Datasets

The PWV dataset used in this work has been derived from the SOI-APN GNSS ground receivers. Ziv *et al.* [45] processed the daily RINEX files downloaded from the SOPAC/Garner GPS archive (http://garner.ucsd.edu/) using NASA's JPL GipsyX software [20]. The daily RINEX processing is done using NASA's JPL GipsyX [46] software via the PPP solution. We use a minimum cutoff elevation angle of 15°, GMF for the tropospheric model [24] and ocean loading for all of the stations. The full parameter tree used in this work is available at the Github.com repository (https://github.com/ZiskinZiv/PW_from_GPS/blob/master/my_trees/ISROcnld/ISROcnld_0.tree). The processing has resulted in ZWD that was translated into PWV using the following formula [13]:

$$\mathrm{PWV} = \Pi \times \mathrm{ZWD}. \tag{1}$$

Π is the dimensionless constant of proportionality and is mainly the function of the atmospheric mean temperature.

TABLE I

Geographical Coordinates, Altitude Above Sea Level, and the Names of the SOI-APN Stations in the Study Area

GNSS Station name	Station ID	Latitude [°N]	Longitude [°E]	Altitude [m a.s.l]
Nizana	NIZN	30.88	34.42	274
Kibutz Lahav	KLHV	31.38	34.87	498
Yerucham	YRCM	30.99	34.93	516
Mitzpe Ramon	RAMO	30.60	34.76	887
Metzoki dragot	DRAG	31.59	35.39	32
Dead-Sea Manufactories	DSEA	31.04	35.37	-361
Sapir	SPIR	30.61	35.18	12
Kibutz Neve Harif	NRIF	30.04	35.04	458
Eilat	ELAT	29.51	34.92	30

TABLE II
GEOGRAPHICAL COORDINATES, ALTITUDE ABOVE SEA LEVEL, AND THE NAMES OF THE HYDROMETRIC STATIONS ANALYZED IN THIS WORK

Hydrometric station name	Station ID	Latitude[°N]	Longitude[°E]	Altitude[m a.s.l]	Nearest GNSS station	Distance to GNSS station[km]	Flood events near GNSS station
Lavan - new nizana road	25191	30.91	34.53	251	NIZN	10	4
Shikma - Tel milcha	21105	31.47	34.77	202	KLHV	14	25
Mamsheet	55165	30.96	35.05	295	YRCM	12	25
Ramon	56140	30.61	34.86	480	RAMO	9	11
Draga	48125	31.60	35.37	-19	DRAG	3	15
Chiemar - down the cliff	48192	31.14	35.35	-320	DSEA	11	8
Nekrot - Top	56150	30.58	35.05	226	SPIR	14	5
Yaelon - Kibutz Yahel	60105	30.09	35.12	216	NRIF	10	9
Solomon - Eilat	60190	29.53	34.91	89	ELAT	2	5

Ziv *et al.* [45] used the Israeli Meteorological Service's (IMS) automated stations and radiosonde measurements [12] in order to estimate the atmospheric mean temperature, T_m, relationship to the surface temperature, T_s, in the study area: $T_m = 0.69T_s + 82$. All of these steps along with the PWV validation using the Bet-Dagan radiosonde station are described extensively in [45]. The final step in the PWV dataset preparation is the removal of the mean diurnal and annual variations. For each station, the resulting time series, which we call PWV anomalies, contains only the interdaily variability. Fig. 3 shows the mean diurnal and annual cycle for DSEA, RAMO, and ELAT stations. We can clearly see the difference between the stations' climatology, where, e.g., DSEA has the highest annual values since it is −361 m below sea level [47]. The diurnal variations are much smaller than the annual variations, and however, we can still spot them during summer (day of year (DoY): 152–244) where they are the strongest since the sea breeze mechanism is a dominant factor on the diurnal time scale [45]. The interested reader can refer to [45] or [47] for the full processing parameters and the PWV derivation and validation methodology as well as the diurnal, interannual, and long-term analysis.

The floods database has been received from the Israeli Water Authority (IWA, https://www.gov.il/en/departments/water_authority). The IWA manages and processes the measured data received from the hydrometric stations across Israel, which include the flood occurrence date times along with water level and water discharges for all recorded events. For each GNSS station, we searched for all available hydrometric stations located within a 15-km radius distance from the GNSS station location. We then selected the station with the highest amount of flood events, which we had the PWV data for, at least 24 h prior to the flood. Thus, we obtained an initial number of 151 flood events co-located with the respective GNSS stations. In Section II-C1, we discuss the data preprocessing and the subsequent trimming of the flood events to include only the unique events.

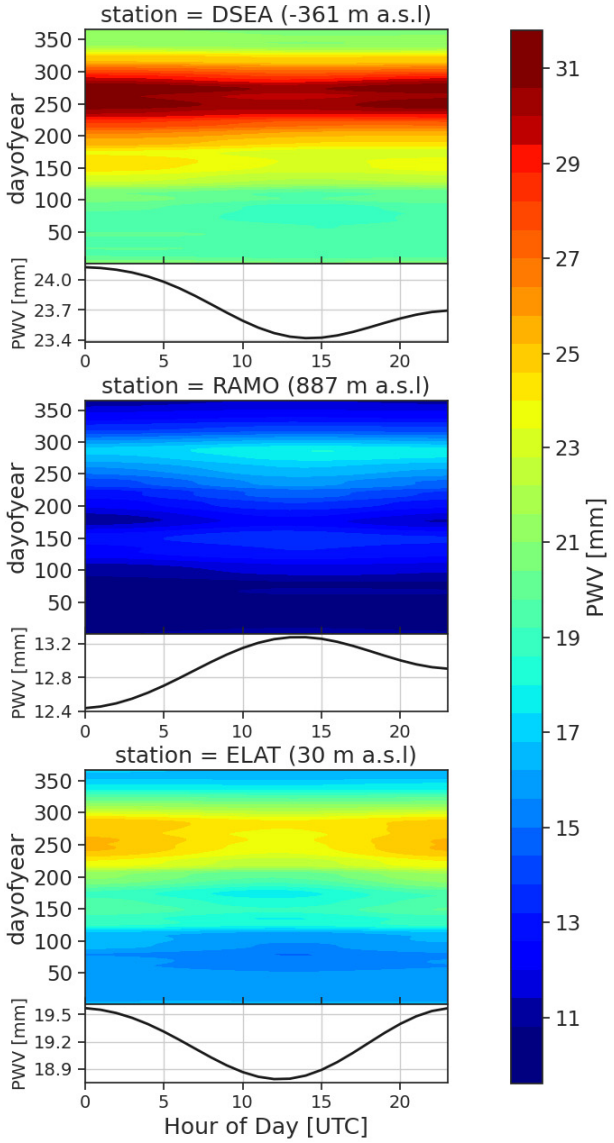

Fig. 3. PWV annual and diurnal climatology for (Top) DSEA, (Middle) RAMO, and (Bottom) ELAT stations. The diurnal annual mean is plotted under each filled contour panel.

B. Data Analysis

A first look at the left panel of Fig. 2 indicates that the flood events are quite rare, while the PWV data are mostly continuous. In order to detect the effect that PWV has on flood events, we averaged the PWV anomalies six days prior

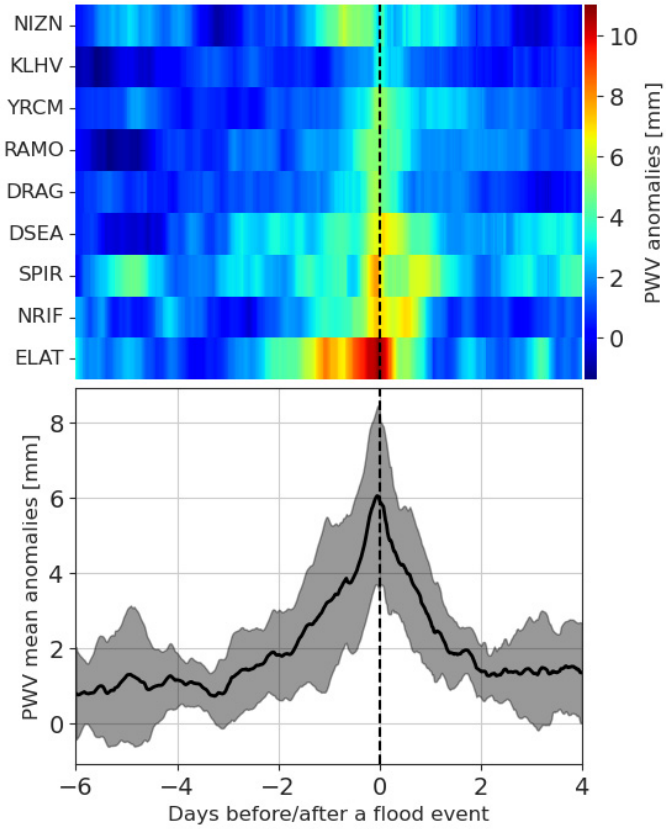

Fig. 4. (Top) PWV mean anomalies heatmap for the SOI-APN stations, presented in Fig. 2, with respect to a mean flood event. The average was calculated for various flood events (the rightmost column in Table II) per each station, from a total number of 151 events. (Bottom) Averaged PWV anomalies, along with its variability (indicated by the shaded gray strip), for the nine GNSS stations with respect to a ten-day time window around all the flood events (six days before and four days after the events, where the black dashed line is positioned at $t = 0$).

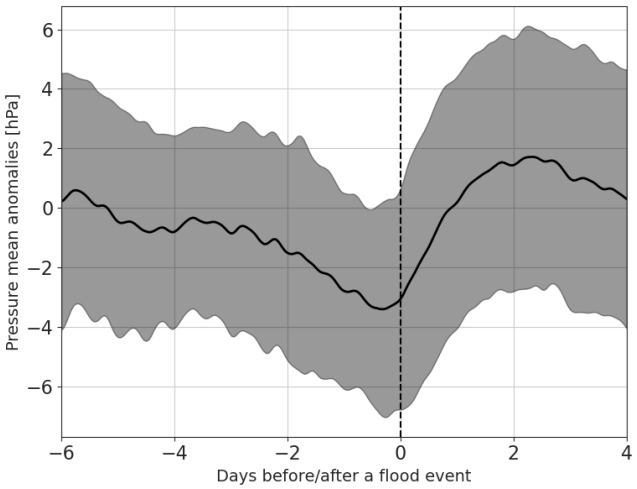

Fig. 5. Station averaged pressure anomalies with respect to a mean flood event (black dashed line at $x = 0$).

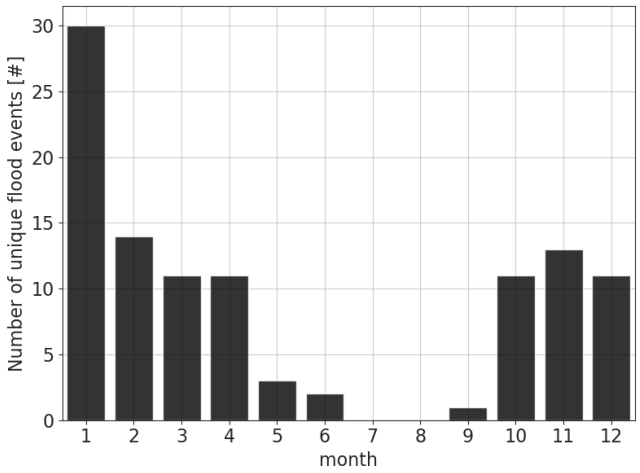

Fig. 6. Number of flood events per month in the arid climate of southern Israel for events which we have PWV data for.

and four days after a flood event. We repeated this step for all the GNSS stations and also averaged all the PWV anomalies stationwise. The top of Fig. 4 shows a heatmap for each GNSS station describing the averaged PWV anomalies before and after a co-located flood event. For example, we see that ELAT has the largest amplitude of PWV, with an almost 8-mm difference three days prior to a flood event. The bottom of Fig. 4 shows the station averaged PWV anomalies and the gray shading indicates its variability standard deviation (SD). We can see that the averaged station's PWV increases from about 1 mm three days prior to a flood to a peak of 6 mm in the flood event beginning. On average, the PWV doubles its value 24 h prior to a flood. After the flood event, we detect a drop to preflood PWV levels lasting for about 24 h. Interestingly enough, the PWV peak is reached almost exactly when the flood event begins, and thus, the drainage area's response to the rain event is probably very fast.

Since our main approach to flash flood prediction is mostly data-driven, we decided to add more features with a goal of increasing our model's performance. In particular, we added long-term hourly surface pressure measurements from the Bet-Dagan IMS station (see map in Fig. 2) and removed the diurnal and long-term climatology in the same manner as we did with the PWV data. Fig. 5 shows the mean pressure anomalies at Bet-Dagan station prior and after a flood event. As expected, the pressure drops before a flood event, representing a low-pressure system that produces precipitation events. The minimum pressure values are found about 6–8 h prior to a flood event. However, the variability is quite higher than the PWV dataset. This issue can be the result of using pressure data from only one station, which represents all the flood events. Unfortunately, we could not find enough surface pressure records that are co-located with the selected hydrometric stations for the same data period. Furthermore, since summer rain is very rare in the EM [48], we also added the DoY information as a feature to our PWV and surface pressure features. Fig. 6 shows the flood event count for each month of the year. It is clear that the most frequent month is January, with 30 events, while February–April and October–December have a mean of 11 events. May, June, and September have only a few events, while July and August have no flood events, as expected.

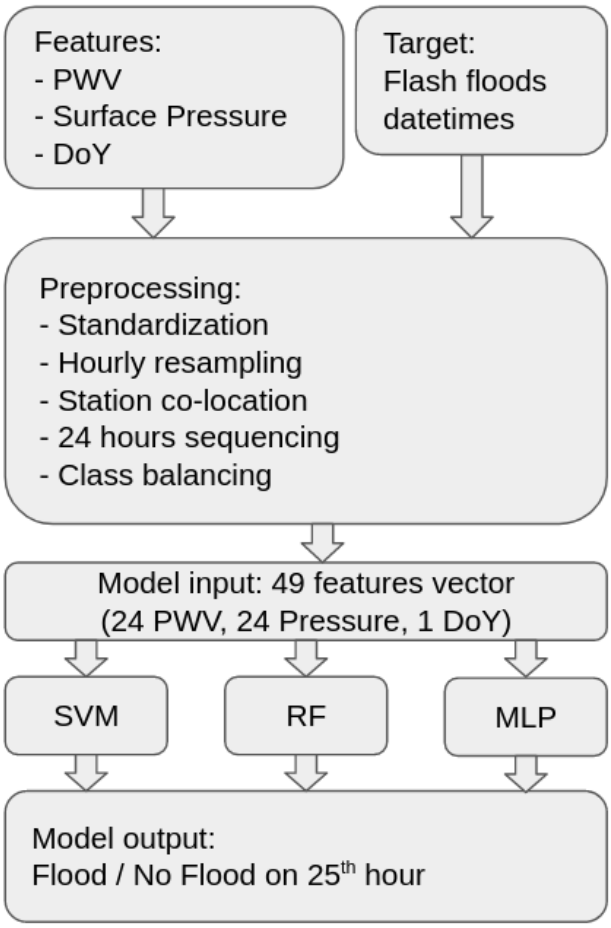

Fig. 7. Main ML methodology block diagram. The features are the PWV, surface pressure, and Doy, where the target is the flash floods datetimes. Preprocesing involves standardizing the PWV and surface pressure measurements, hourly resampling them, and colocating the GNSS and hydrometric stations. Finally, 24-h sequences are generated with class balancing. In the learning process, three general types of ML classifier models are optimized using CV: MLP, SVM, and RF. The final output of each model is whether or not a flash flood will occur in the 25th hour.

C. ML Methodology

Fig. 7 schematically describes all the ML methodological steps from the data preprocessing to producing the best model. We, therefore, elaborate on these steps in the following.

1) Preprocessing: Our data-driven approach to flood prediction considers a supervised learning task using binary classification. In particular, we ask the following question: given 24 h of PWV anomalies, surface pressure anomalies, and DoY, will there be a flood event in the following hour? When termed this way, we regard the PWV, surface pressure, and DoY data as features and the flood/nonflood events datetimes as the samples. Therefore, our preprocessing of the samples and features is given as follows. First, we removed from the flood database close events that are overlapping within a 24-h window. The idea was to find unique flood events as much as possible, without losing too many samples. This step leaves us with 107 flood events from an original 151 GNSS co-located events. The flood events are the positive class in our classification task. We then continued with the positive features, i.e., PWV and surface pressure that are resampled to hourly means. We then co-located each GNSS and hydrometric station and found 24 data points of PWV prior to each flood event. If half or more of the PWV data was missing, we dropped this event from our analysis. We used cubic interpolation to fill in the missing data points otherwise. We repeated this process with the surface pressure data, and however, in this case, we had only one surface pressure station (Bet Dagan) with the necessary data period and resolution. This step leaves us with 49 features (48 for PWV and pressure along with one for DoY). As for the negative class, we randomly searched for 24 h of PWV and pressure, which do not overlap the positive features, and we repeat this step only once for each flood event in each station, thus ensuring that the binary classification task is balanced. Our resulting matrix of features and samples is 214 (107 for each class) by 49. Finally, since two of our classifiers are sensitive to feature normalization, we use the standardized[3] version of the PWV and surface pressure anomalies for all the classifiers.

Our main goal is to use supervised learning classifiers in order to predict flash floods using PWV as the main input. Accordingly, we chose three common types of ML models: SVM, RF, and multilayered perceptron (MLP). All the models were implemented using the Scikit-Learn Python package [49].

The SVM classifier utilizes a linear hyperplane to separate each sample class [50]. Using the kernel trick, the hyperplane is transformed into a higher dimension, which gives the SVM more flexibility; however, the cost is a larger generalization error [51]. The RF classifier is a metaclassifier, which uses a number of decision trees on randomized selections of subset of features. The final output is produced by averaging all the individual decision tree classifiers [52]. The MLP classifier is a neural network algorithm, which includes multilayered nodes with weights [53]. Typically, the network architecture includes an input layer, any number of hidden layer, and an output layer where each layer's nodes are connected via activation functions (a so-called feedforward propagation). During the learning process, the weights are reevaluated using the backpropagation iterative algorithm [54] in order to decrease the cost function.

2) Score Metrics: We use six different metrics to evaluate the models' performance [55]. These metrics are: precision, recall, F1, accuracy, Heidke skill score (HSS), and true skill statistics (TSS). These metrics are defined in (2) and are a combination of the four possible outcomes of our classifier.

1) True positive (TP) is the correct prediction of a flood event.
2) True negative (TN) is the correct prediction of a no-flood situation.
3) False positive (FP or type I error or false alarm) is when the classifier predicts a flood but there was not any.
4) False negative (FN or type II error or simply miss) is when the classifier does not predict a flood but a flood

[3]Standardized anomalies are the removal of the long-term monthly mean from a time series and dividing it by the long-term monthly standard deviation.

occurs, hence the miss

$$\text{Fallout} = \text{FPR} = \frac{\text{FP}}{\text{FP} + \text{TN}} \tag{2a}$$

$$\text{Precision} = \text{PPV} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{2b}$$

$$\text{Recall} = \text{TPR} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{2c}$$

$$\text{F1} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{2d}$$

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{2e}$$

$$\text{TSS} = \frac{\text{TP}}{\text{TP} + \text{FN}} - \frac{\text{FP}}{\text{FP} + \text{TN}} = \text{Recall} - \text{Fallout} \tag{2f}$$

$$\text{HSS} = \frac{2 \times [\text{TP} \times \text{TN} - \text{FN} \times \text{FP}]}{(\text{TP} + \text{FN}) \times (\text{FN} + \text{TN}) + (\text{TP} + \text{FN}) \times (\text{FP} + \text{TN})}. \tag{2g}$$

The fallout or false positive rate (FPR), see (2a), measures the probability of false alarm (FPs). The precision or positive predictive value [see (2b)] measures the ability of the classifier not to produce false alarms. The recall also known as true positive rate (TPR), sensitivity, or hit rate [see (2c)] measures how successful the classifier is in predicting the positive class without missing (FN).

Unfortunately, precision and recall are always at tension with each other, where improving recall reduces the precision and vice versa. One way of dealing with this issue is to use the F1 score, which is the harmonic mean of the precision and recall [see (2d)]. The accuracy score [see (2e)] quantifies how well a classification test correctly identifies or excludes a condition (i.e., whether it is a TP or TN). The TSS [see (2f) [56] compares the probability of the true prediction, to the probability of false prediction or simply recall minus the fallout. Thus, a TSS no skill score is 0, while -1 means that the prediction labels should be reversed. The HSS [see (2g)], which is often used in weather and solar events prediction (e.g., [41]), quantifies the fractional improvement of the prediction accuracy relative to some set of control or reference predictions. It is normalized by the total range of possible improvement over the standard (i.e., it can be compared with different datasets). A perfect HSS score is 1, and a no skill score is 0, while an infinitely negative score is possible, suggesting that the prediction is worse than the reference prediction. An easy to implement and use formulation as presented here is available in [57].

Another widely used performance measurement visualization method is the receiver operating characteristics (ROC) curve, which illustrates the diagnostic ability of a binary classifier as its classification threshold is varied. The ROC curve is actually the recall or TPR plotted versus the fallout or FPR where, ideally, the TPR is maximized, while the FPR is minimized. The area under the ROC curve (ROC-AUC) can be used as a score metric where a no skill score is 0.5, while a perfect score is 1.

3) CV Strategy: Traditional CV or k-fold CV is a technique, which is often used to estimate the performance of ML models when making predictions with data not seen during training [58]. The data are divided into k segments or folds with the same size where each fold is being tested by the model and the other $k-1$ folds are used as training data. The process repeats k times, where the best score of each fold's validation procedure is used to select the best model. Since most ML models have hyperparameters (e.g., regularization coefficient) which need tuning, the CV step is often performed together with the hyperparameters tuning, a practice that can lead to overfitting [59]. A useful way of dealing with this issue is to separate the CV into two k-fold CV steps, which first tunes the model's hyperparameters and then evaluates the model's performance by estimating the generalization error. This procedure is called double CV or nested CV and is often implemented by using a nested loop, i.e., an inner loop that optimizes the hyperparameter space and an outer loop that estimates the generalization error [60]. Since nested CV uses a lot of computational time, we must balance the recommended number of folds [61] and the hyperparameter space with the computational time. Nevertheless, in order to quantify the bias that a particular selection of k can enter our generalization error, we run two nested CV configurations, one with four inner/outer folds and another with five inner/outer folds. This verification procedure is outlined in Section IV-A and concludes that there is little bias in selecting either $k = 4$ or $k = 5$. Thus, as shown in Fig. 8, we use five folds for hyperparameter tuning and validation (inner folds) and five folds for model selection (outer folds). The six scoring metrics are reported for each inner fold, where the model's hyperparameters are tuned, and for each outer fold, where the best model is chosen, thus estimating the model performance. Finally, since for each outer fold, we get a set of unique hyperparameters (i.e., essentially a different model), we plot the chosen hyperparameters as a function of a particular fold and metric in order to choose the best "mean" hyperparameters. This step measures the model sensitivity to the hyperparameters optimization and is also outlined in Section IV-A. These best "mean" hyperparameter sets, as will be shown in Section III, produce high enough scores and low k-fold variability in all models.

4) Permutation Test: We also subject our classifiers to the permutation test for labeled data [62]. This test, which has been extensively used in the field of computational biology, aims to address the following question: does the classifier detect a significant class structure, i.e., a real connection between the data and the class labels? We use a standard fivefold CV to estimate a null distribution by permuting the labels in the data and produce a "true" score without the permutations. The experimental p-value from these tests is calculated as follows:

$$p - \text{value} = \frac{S + 1}{n_{\text{permutations}} + 1} \tag{3}$$

where S is the number of permutations whose score $\geq$ the "true" score. Since ideally, S should be 0, the best possible p-value is $1/(n$ permutations $+ 1)$, and since we use 100 permutations, it is $1/101 = 0.0099$, while the worst p-value is when $S = n_{\text{permutations}}$, i.e., p-value $= 1.0$.

5) Imbalanced Dataset Test: Since flash floods are very rare events, we thus require a more realistic scenario for testing

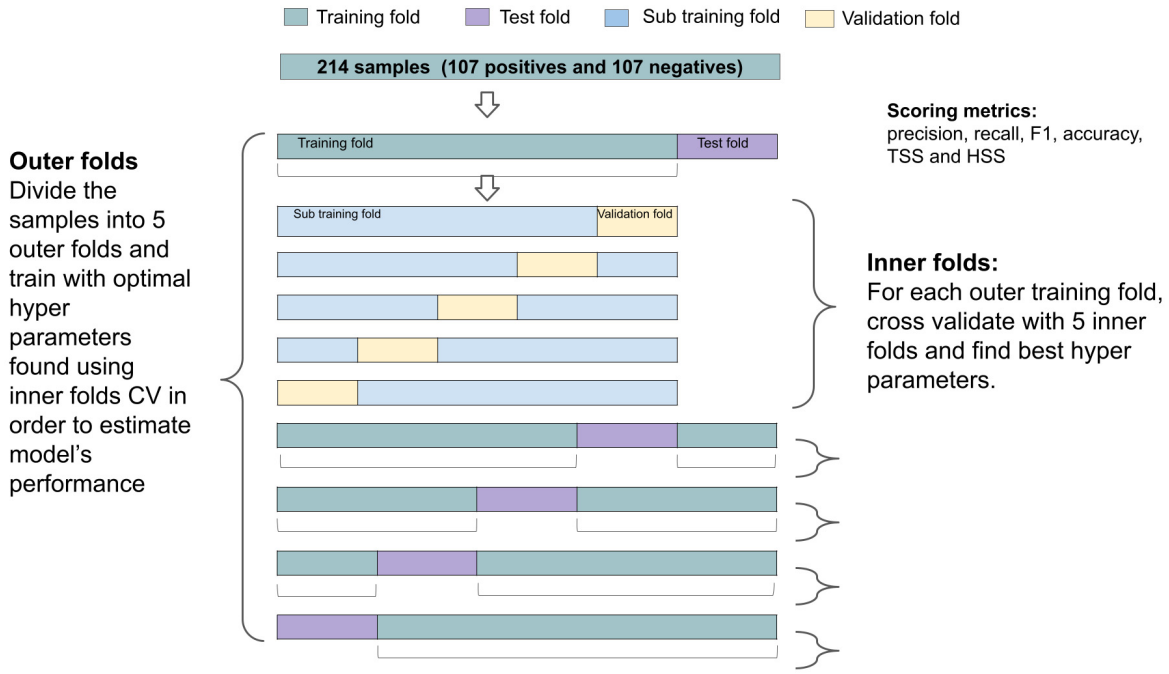

Fig. 8. Nested cross-validation strategy diagram used in the classification task in this work. It includes five splits in the outer loop for testing and five splits in the inner loop for hyperparameter tuning. For each fold, the samples are selected randomly from all the available data for training and testing/validating, and thus, the diagram oversimplifies the fold separation process for visualization purposes.

our classifier, which is trained with a balanced dataset. Therefore, we need to generate more negative samples from the PWV/pressure time series. From Fig. 2, we see that for all the stations (except ELAT), the minimum flash flood record (x's) starts approximately with the beginning of the PWV record (we simply have no earlier records for ELAT before 2012). Furthermore, there is a varying degree of flash flood events frequency within the stations. As a rough estimate, we divide the number of the total flash flood events ($\approx$100) with the total number of days of the largest time series (RAMO: $\approx$7500 days or $\approx$20.5 years) and reach a ratio of 1 flash flood event in 75 days or 1.3% positive ratio. Thus, we need to produce negative samples for each station that is complete (24 h) and do not coincide with a positive event. Unfortunately, with these constraints, we were able to find only 25 negative samples per a positive one or 4% positive ratio that is three times more frequent than the rough estimate. Nevertheless, we can use a specific data split in order to overcome this obstacle. The testing procedure for the imbalanced dataset is given as follows.

1) For each ML model, we train our classifiers with 66.66% of the balanced training set (71 positives and 71 negatives).
2) We evaluate the classifiers with the remaining 33.33% of the balanced dataset concatenated with all the remaining negative samples produced (36 positives and 2639 negatives) to receive a positive ratio of 1:73.3 or 1.36%, which is very close to our estimate.
3) We repeat the evaluation for each of the score metrics.

III. Experimental Results

Table III shows the best "mean" hyperparameters chosen for each model (e.g., SVM and RF). All the results shown in this section use the aforementioned set of hyperparameters. We encourage the interested reader to look at Section IV-A where the process of hyperparameters selection is outlined and discussed.

TABLE III

Best Hyperparameters Found Using CV for the SVM, RF, and MLP Classifiers Used in This Work. Other Hyperparameters Aside From the Ones Listed in the Table Were Used in Their Default Values (Python Scikit-Learn Ver 0.23.2)

SVM Classifier		RF Classifier		MLP Classifier	
Parameter	Best	Parameter	Best	Parameter	Best
kernel	rbf	n estimators	400	activation	relu
degree	NR*	max features	auto	hidden layer sizes	(10, 10, 10)
coef0	NR*	min samples leaf	1	learning rate	constant
C	1	min samples split	2	solver	lbfgs
gamma	0.02	max depth	5	alpha	0.1

* Not Relevant since degree and coef0 are only relevant for the poly kernel.

Fig. 9 shows the mean test scores and variability due to data splits selection for the SVM, RF, and MLP classifiers and for each metric. For most metrics, MLP has generally slightly worse scores than SVM and RF. For the feature groups, DoY performs poorly, followed by surface pressure that is only second to PWV, which has the highest scores for a single feature group. Adding pressure and DoY only slightly improves the scores for most models and metrics. DoY as a single feature has the highest fold selection variability, while all other features have lower variability.

Fig. 10 shows the mean ROC curves for the SVM, RF, and MLP classifiers where the variability due to fold selection is shown in shaded colors. The mean AUC scores and variability is shown in each panel's legend. The left panels show the mean ROC curves for only five data splits and the right panels show

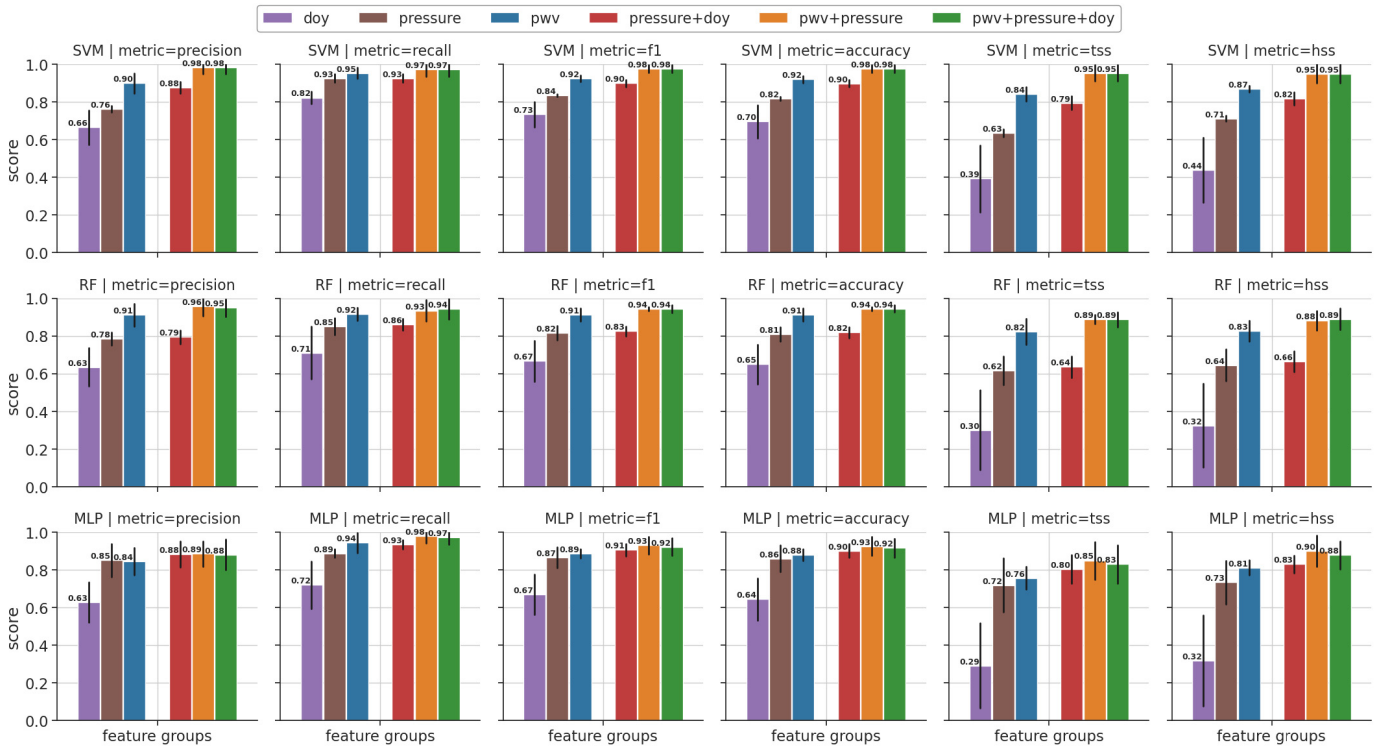

Fig. 9. Mean test scores for the SVM, RF, and MLP classifiers (row) and for each metric (column). The feature groups consist of DoY (purple), surface pressure (brown), PWV (blue), surface pressure and DoY (red), PWV and surface pressure (orange), and all three together (green). The mean scores are indicated to the top left of each bar and the SD of five data splits is represented by the error bar length.

the mean ROC curves where the negative class was resampled 25 times. For both panels, SVM and RF outperform MLP, where the left panels have slightly better AUC scores and lower variability than the right panels, as expected. Adding more features to PWV improves the AUC scores, although adding DoY to PWV + pressure is within the score variability, and thus, its effect cannot be distinguished.

Fig. 11 shows the null distribution, best models scores, and p-values for the SVM, RF, and MLP classifier permutation tests for each metric. For all the models and nearly all metrics, the p-value is 0.0099, which is the highest score available (see Section II-C4 for definition), indicating that the models do a good job at detecting a class structure. However, looking at the MLP model, we can see that the best scores are lower than for the SVM and RF models, and for the recall metric, the MLP model almost fails completely with p-values of 0.06–0.11 for all feature groups (worst possible p-value is 1.0).

Which feature is useful for predicting flash floods? One way of understating which feature group is important to a classifier is to use the Scikit-Learn RF model's built-in feature importance's attribute, which is based on averaging the decrease in impurity over trees [63]. Fig. 12 shows the feature importance based on mean decrease impurity (MDI) for PWV, surface pressure, and DoY as the model uses all of them together. For all metrics, the RF classifier finds the PWV the most important feature with about 72% of the score, followed by surface pressure at about 27% and finally DoY with less than 1%. Interestingly enough, Fig. 12 also shows an hourly breakdown of the MDI-based feature importance where the highest PWV's hourly contribution is from 2 to 5 h prior a flood (totaling in roughly 20%). This finding is not very surprising since we expect that the PWV values close to the time of the flood would be the most relevant in the prediction of the flood. There are two more minor importance peaks that reside in around 13 and 20 h prior to a flood. However, their significance remains unclear and requires further investigation.

In many data scenarios, the MDI-based feature importance method contains biases and should not be relied upon [64]; however, it is not clear if this is the case in our work. Nevertheless, we decide to validate our findings using a game-theory-inspired method of feature importance based on the Shapley values [65]. We use the SHAP Python package [66] and calculate the mean SHAP values for the three feature groups where the RF is trained with its best HPs. The result is in Fig. 13, showing an almost similar picture as in Fig. 12, where the most important PWV values are from 2 to 6 h prior to a flood with two more smaller peaks in 14 and 19 h prior to a flood.

The imbalanced dataset test scores, as presented in Fig. 14, yield a drop in the precision and F1 metrics for all feature groups; however, for all other metrics, i.e., accuracy, TSS, HSS, and most importantly recall, the scores compared to the balanced test (Fig. 9) remain almost unchanged.

IV. Discussion

As to date and to the best of our knowledge, this work demonstrates for the first time the ability to directly predict flash flood events from GNSS-derived PWV using ML methodology. Thus, the first part of this discussion, which is presented in Section IV-A, is about the technical validity or

TABLE IV
HYPERPARAMETERS SEARCH SPACE FOR THE SVM, RF, AND MLP CLASSIFIERS USED IN THIS WORK

SVM Classifier		RF Classifier		MLP Classifier	
Parameter	Options	Parameter	Options	Parameter	Options
kernel	rbf, sigmoid, linear, poly	n estimators	100 to 1200	activation	identity, logistic, tanh, relu
degree	1, 2, 3, 4, 5	max features	auto, sqrt	hidden layer sizes	(10, 20, 10), (10, 10, 10), (10,)
coef0	0, 1, 2, 3, 4	min samples leaf	1 to 10	learning rate	constant, adaptive
C	0.01 to 100	min samples split	2 to 50	solver	adam, lbfgs, sgd
gamma	10^{-5} to 1	max depth	5 to 100	alpha	10^{-5} to 10

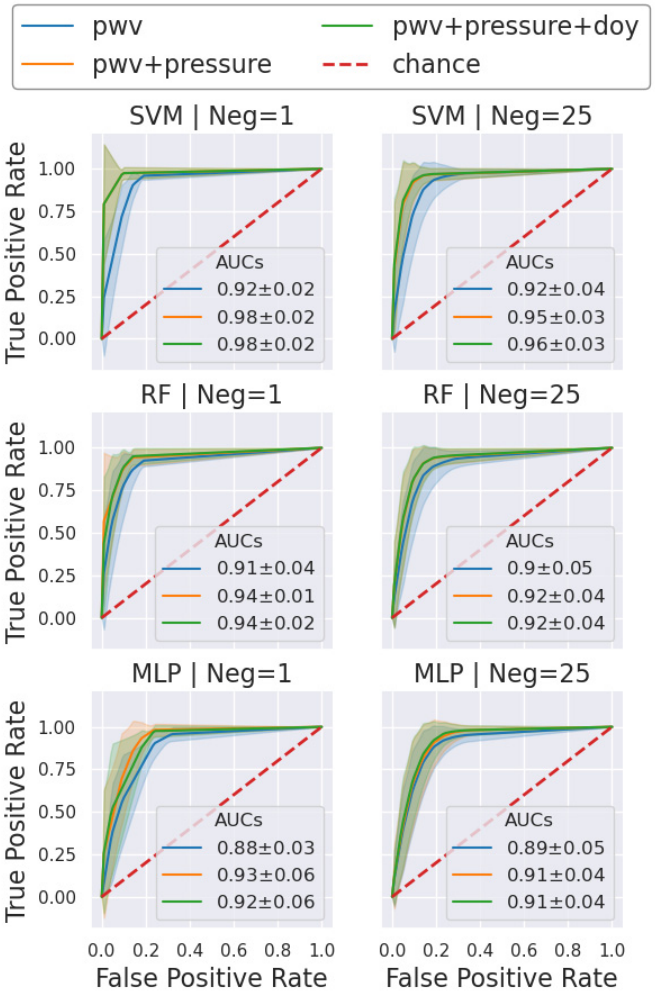

Fig. 10. Mean ROC curves for the SVM, RF, and MLP classifiers (row) with the best hyperparameters where the positive and negative classes are balanced (Neg = 1, left column). The feature groups consist PWV (blue), PWV and surface pressure (orange), and PWV + surface pressure + DoY (green) where the chance curve indicates a no skill curve (red dashed line). The right column is the same as the left except that we resampled the negative class 25 times (Neg = 25). The shaded area is the SD of 5 X Neg folds for each point in the ROC curve. The area under the curve is denoted in the legend with the SD of 5 X Neg folds as variability.

our results or, more specifically, the hyperparameter tuning and selection procedure. A comparison with other flash floods predicting works in the EM is discussed in Section IV-C.

A. Hyperparameters Tuning and Selection

Table III shows the best hyperparameters that were used in testing the models using different score metrics, producing ROC curves, and permutation testing. This set of hyperparameters was selected using a grid search for the various hyperparameters range, as shown in Table IV. The basic idea is to search for the best hyperparameters per each data split by using different metrics. If we select one set of hyperparameters for all of the data splits and show that the test scores do not vary too much, then we can justify our choice empirically. In addition, if we can also show that the same results hold for different score metrics, it will increase their robustness.

Figs. 15–17 show the optimal hyperparameters with a data split and metric breakdown for the SVM, RF, and MLP classifiers, respectively. For all the classifiers, we can detect a change in the hyperparameters for the recall metric, which raises a red flag over its usage. For the SVM classifier, except for the recall metric, the best kernel is radial basis function (rbf) with good estimates of gamma and C values of 0.02 and 1 respectively. For the RF classifier, min samples split parameter is very high for the recall metric compared to the other metrics, while all other parameters do not show any reaction when being optimized by recall. Interestingly, min samples leaf changes in some splits when the CV strategy is four inner folds as opposed to five inner folds. Finally, for the MLP classifier, the hidden layer sizes (NN architecture) parameter is almost evenly distributed, suggesting that this parameter is not important or our search grid is not large enough. For the activation parameter (activation function), the recall metric optimizes this parameter to be logistic as opposed to rectified linear unit (relu) for all other metrics.

In order to test how the hyperparameter optimization for each inner fold fares on the test fold, we use the same metrics on each inner fold optimized with the same metric. The results are shown in Figs. 18 and 19 for the four and five inner folds' CV strategies, respectively.

If we compare these figures to Fig. 9, we can see that for all metrics except recall, the mean test scores are within their data split variability (the error bars in the bar plot). For recall, SVM finds perfect scores for most features, while MLP finds a perfect score for some features and large variability for others. This apparent instability further lowers the recall metric's reliability in this task.

As with the score test, we can make ROC curves along with their AUC scores for each fold individually. Figs. 20 and 21 show the ROC curves for the four and five inner folds' CV strategies, respectively.

Most of the ROC curves and the accompanying AUCs are consistent throughout the metrics except for recall. SVM and MLP perform poorly for all features with regard to this

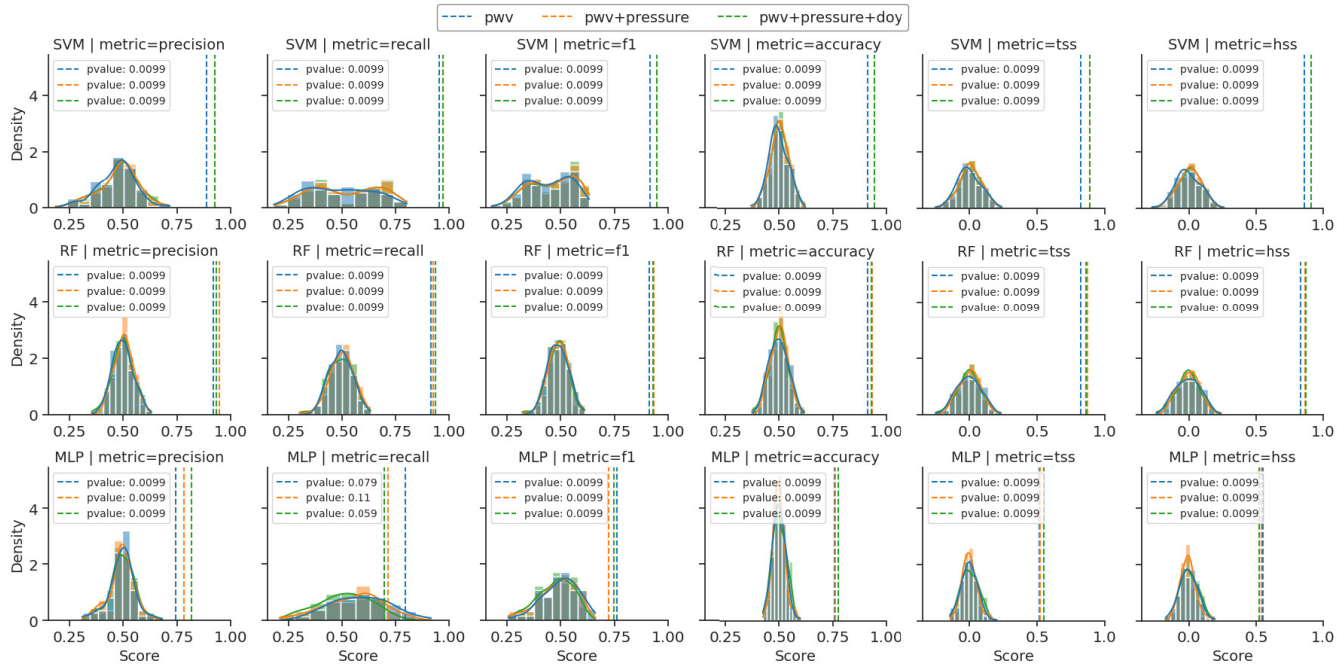

Fig. 11. Histogram panel of permutations, along with test scores and their experimental p-values for the SVM, RF, and MLP classifiers with the best hyperparameters. The histogram was obtained by permuting the labels for each feature 100 times.

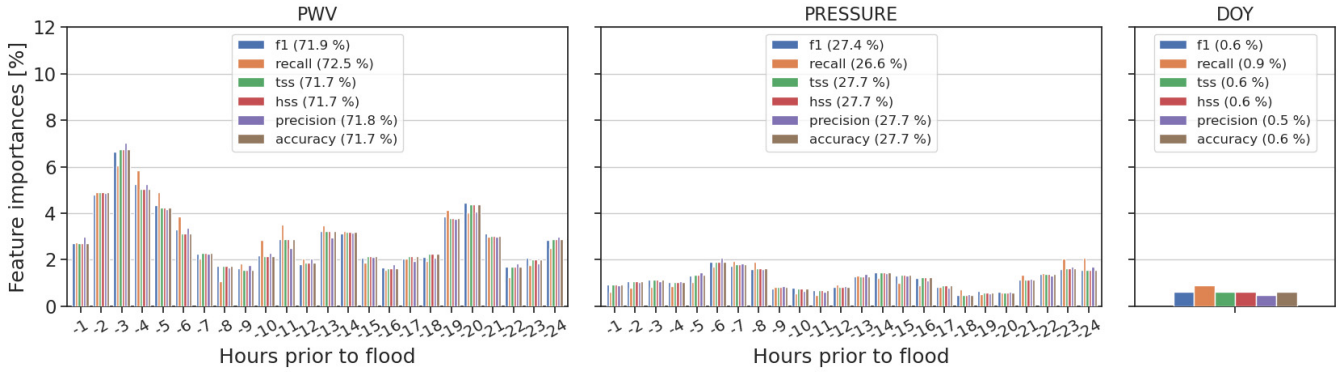

Fig. 12. Feature importance for the PWV, surface pressure, and DoY features as run together in the RF classifier. Each metric is presented in different colors. The importance is given in percent, when combined (per metric), they are equal to 100%.

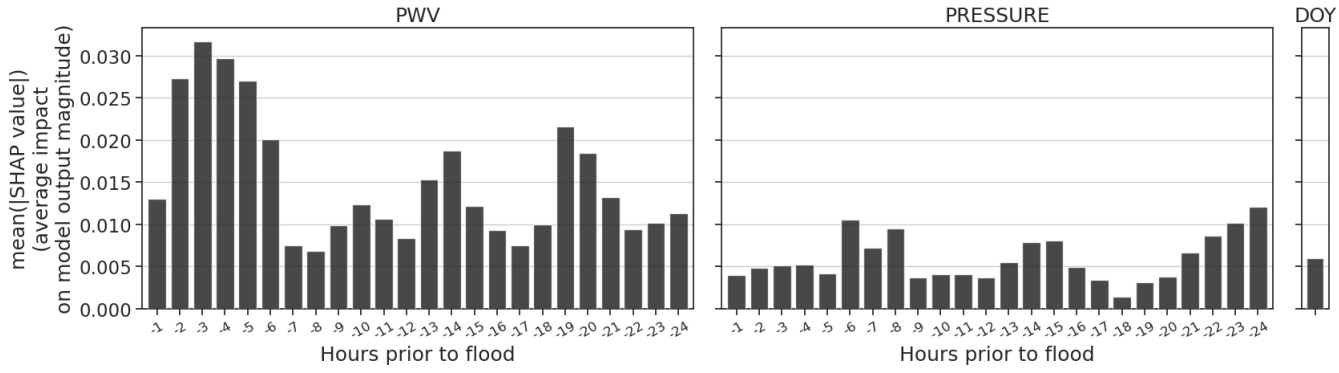

Fig. 13. Mean SHAP values for the PWV, surface pressure, and DoY features as run together in the RF classifier. The effect is symmetrical to each class, and thus, only the contribution to the positive is plotted.

metric with AUC scores of roughly 0.5 (random choice). Thus, we must conclude that for the purposes of hyperparameter optimization in our datasets, recall should not be used as a metric. Comparing these figures to the left column of Fig. 10 shows similar results, thus empirically justifying our choice in one best set of hyperparameters.

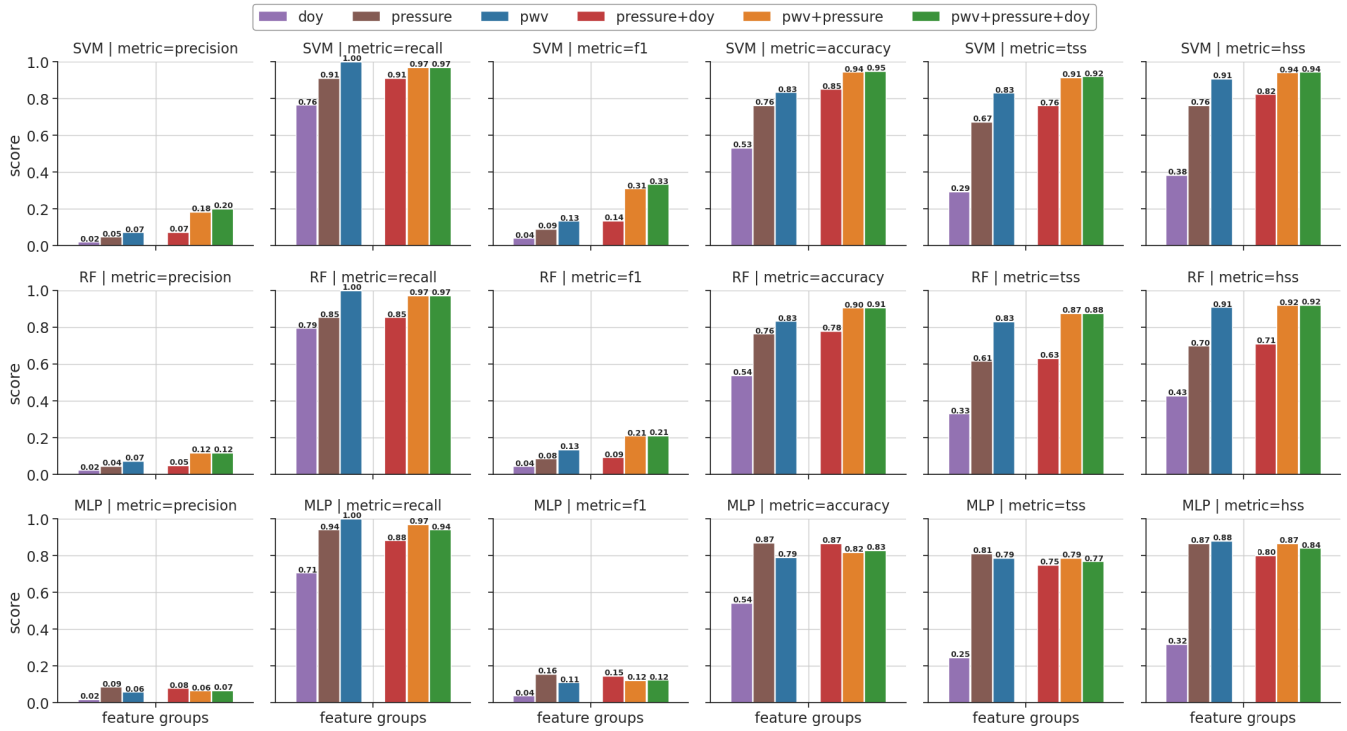

Fig. 14. Imbalanced dataset test scores for the SVM, RF, and MLP classifiers (row) and for each metric (column). The feature groups consist of DoY (purple), surface pressure (brown), PWV (blue), surface pressure and DoY (red), PWV and surface pressure (orange), and all three together (green). The scores are indicated above each bar for better readability.

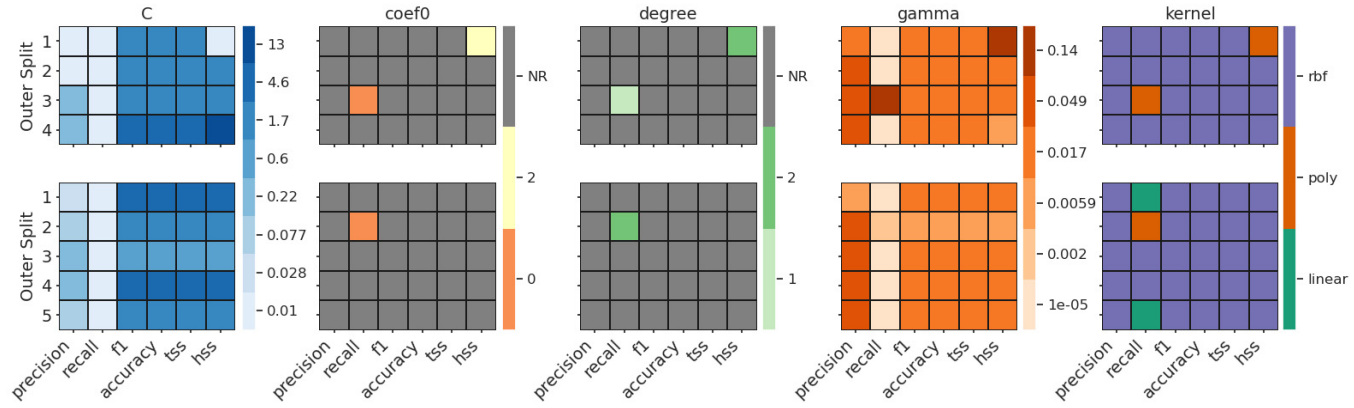

Fig. 15. Panel showing the optimal hyperparameters that were found using grid search CV for the SVM classifier. Each set of hyperparameters was found for each outer split (row) and each metric (column). Moreover, the hyperparameters are also optimized when considering (Top) four and (Bottom) five inner folds. Each hyperparameter name is denoted in the title of each top panel and its values are indicated in the accompanied colorbar. NR means not relevant since degree and coef0 hyperparameters are relevant only when the kernel is poly.

B. Realistic Flash Flood Scenario Test

The imbalanced dataset test, which simulates a more rare flash flood occurrence than previously examined, is presented in Fig. 14. As estimated in Section II-C5, this scenario simulates a 1 in 75 days flash flood frequency and represents a flash flood occurrence for the study terrain in the EM area. For most metrics, the classifiers performed admirably, and however, there is a significant drop in the precision and the F1 metrics' performances. For a more moderate imbalanced dataset (1 in 40), the metric scores, e.g., SVM model, show a 30% mean improvement (not shown), suggesting that increasing the dataset for both training and testing might improve the classifier's performance for possibly more rare event occurrence. Furthermore, since the F1 metric is very sensitive to both the recall and the precision metrics, a drop in either lowers F1 considerably, and thus, the F1 scores are expected since our precision dropped as well. As expected, the recall metric did not suffer the same decrease as the precision, and thus, we conclude that the classifier performs well at minimizing FNs, i.e., when flash floods occur but no warning has been issued. This low miss rate is extremely important for an early warning system which our classifiers have demonstrated through this test.

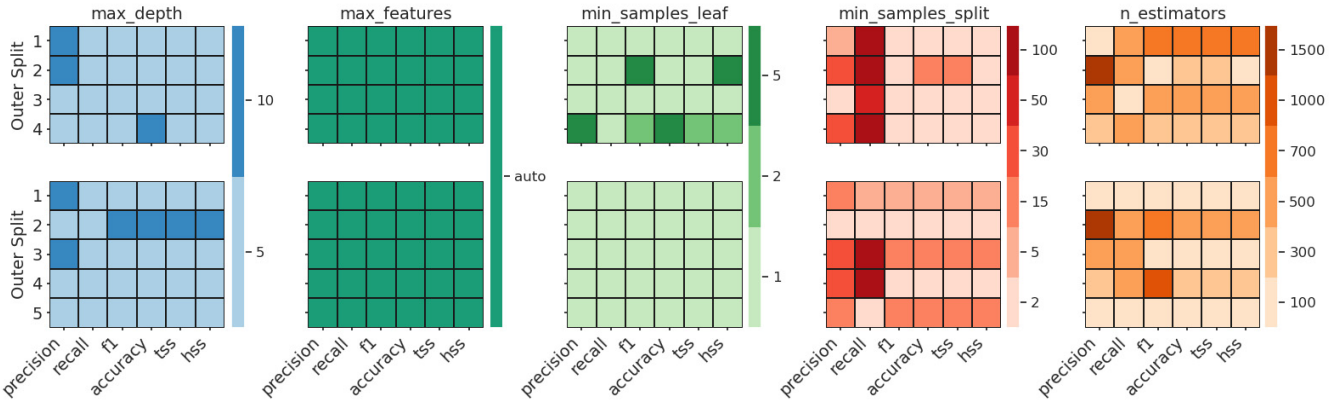

Fig. 16. Panel showing the optimal hyperparameters that were found using grid search CV for the RF classifier and its hyperparameters. Each set of hyperparameters was found for each outer split (row) and each metric (column). Moreover, the hyperparameters are also optimized when considering (Top) four and (Bottom) five inner folds. Each hyperparameter name is denoted in the title of each top panel and its values are indicated in the accompanied colorbar.

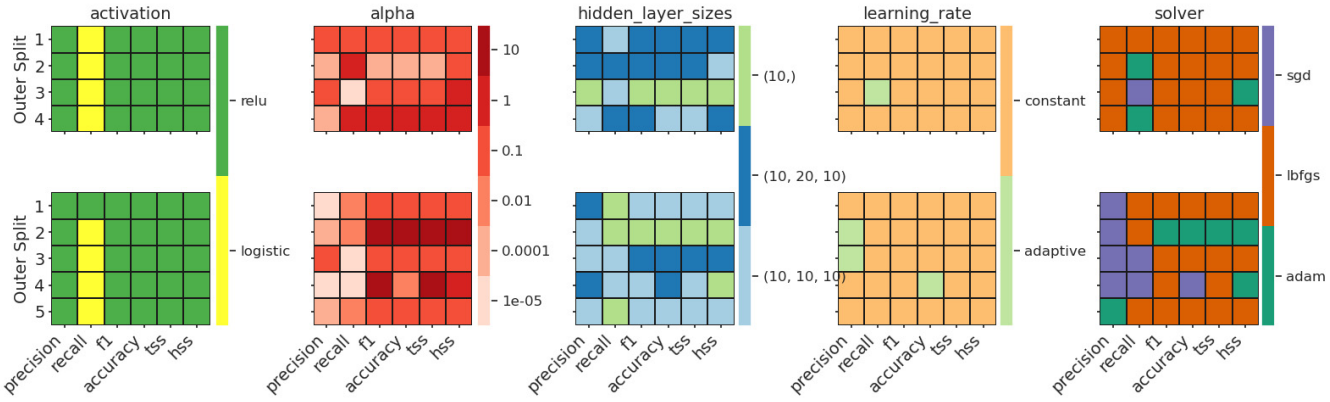

Fig. 17. Panel showing the optimal hyperparameters that were found using grid search CV for the MLP classifier and its hyperparameters. Each set of hyperparameters was found for each outer split (row) and each metric (column). Moreover, the hyperparameters are also optimized when considering (Top) four and (Bottom) five inner folds. Each hyperparameter name is denoted in the title of each top panel and its values are indicated in the accompanied colorbar.

C. Comparison With Other Studies in the EM

In the EM area, we found two major studies, which aim to predict flash floods and produce results that can be translated into the metrics reported in this work. However, both studies have used hydrological-based models without PWV as input and thus consist of different datasets and, as such, are not considered a valid comparison to our work. Nevertheless, reporting their results should give us a rough estimate of the current flash flood prediction ability in the EM area.

Morin *et al.* [10] used rainfall radar data along with a hydrological model in order to predict flash floods in two catchments that drain into the Dead Sea area that is located in the arid part of the EM. They achieved a TPR [also known as recall, see (2c)] ranging from 0.41 to 0.82 and an FPR [also known as fallout, see (2a)] ranging from 0.21 to 0.25.

Rozalis *et al.* [3] used somewhat similar methodology as in [10]; however, the study area is located $\approx$117 km north of the Dead Sea and large enough (27^2 km) to include the Mediterranean, semiarid, and arid climates. Furthermore, the prediction algorithm they used was made to predict three levels of peak discharge ($<$14 m^3/s, 14–50 m^3/s, and $>$50 m^3/s), which can be translated into ML terms as a multiclass classification task. Since in this work, we solve a binary classification task, we will present the aggregated results from [3] to get an estimate of their model's performance. Thus, from a total of 20 events, 12 are correctly predicted (TP), 1 is missed (FN), and 7 were false alarms (FP). Since Rozalis *et al.* [3] did not include the TNs, we can only consider their TPR that is 0.92.

Finally, hydrological models are not meant only for flood prediction but rather a deeper understanding of the underlying physics that drive the flash floods. Unfortunately, our approach here is mostly data-driven and does not present a clear and better understanding of the flash floods phenomenon. Nevertheless, by using ML methodology, we were able to maximize the impact of the small amount of physics that is hidden in the PWV time series and produce a successful flash flood predictor, which can be used as the basis of an early warning system.

V. Summary and Conclusion

We have used nine GNSS ground stations in order to obtain PWV and use it in order to train, test, and validate a classifier

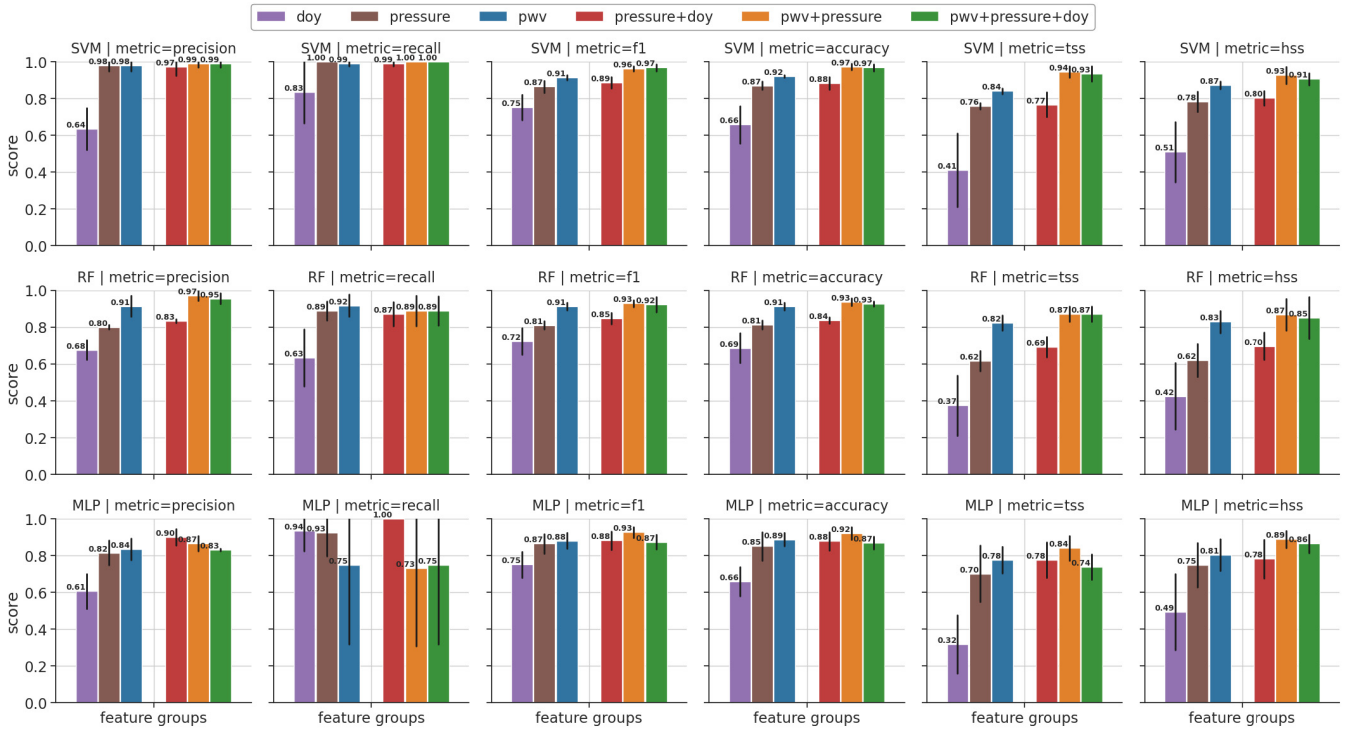

Fig. 18. Mean test scores for the SVM, RF, and MLP classifiers (row) and each metric (column). The feature groups consist of DoY (purple), surface pressure (brown), PWV (blue), PWV and surface pressure (orange), and all three together (green). The mean scores are indicated in the top left of each bar and the SD of four data splits is represented by the error bar length. The difference from this figure and Fig. 9 is that in this figure, the hyperparameters were optimized for each inner fold, and thus, there is a different set of hyperparameters for each fold (e.g., Fig. 16) as opposed to one set of hyperparameters in used in Fig. 9 (Table III).

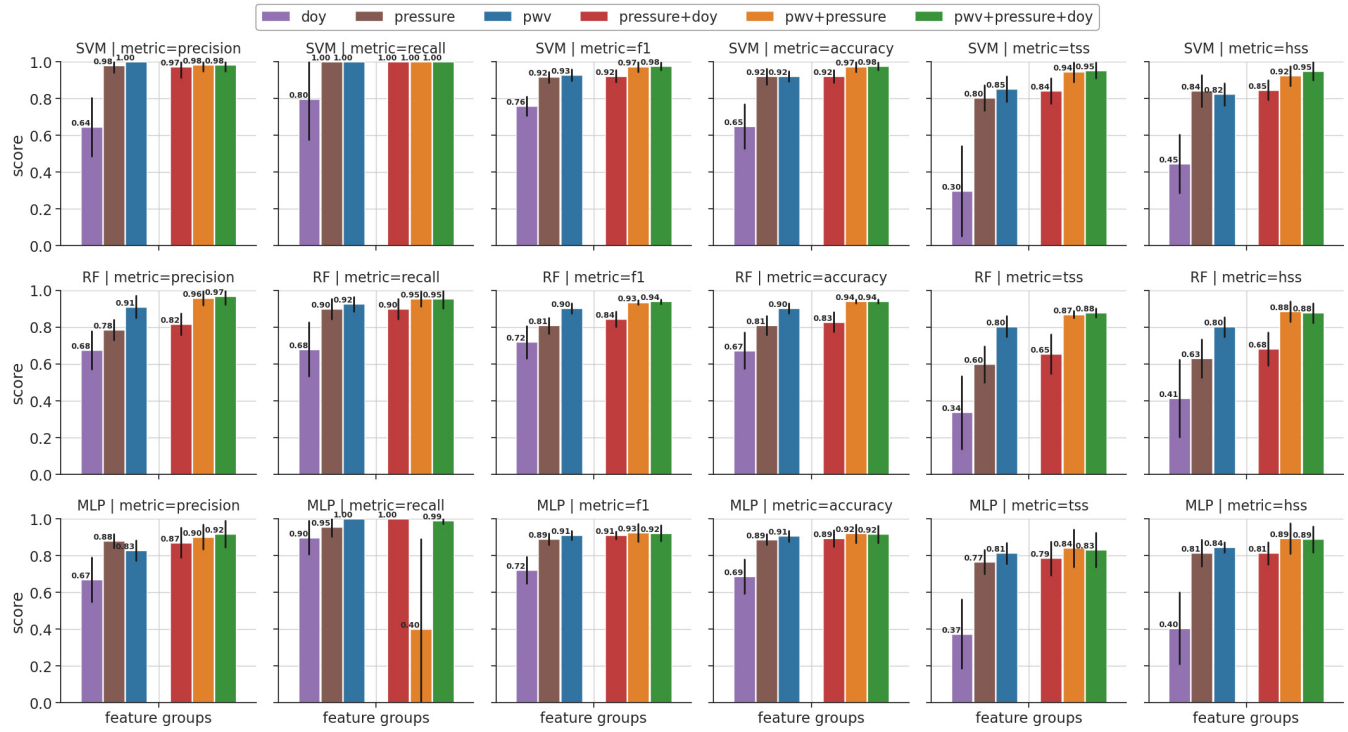

Fig. 19. Mean test scores for the SVM, RF, and MLP classifiers (row) and each metric (column). The feature groups consist of DoY (purple), surface pressure (brown), PWV (blue), PWV and surface pressure (orange), and all three together (green). The mean scores are indicated in the top left of each bar and the SD of five data splits is represented by the error bar length. The difference from this figure and Fig. 9 is that in this figure, the hyperparameters were optimized for five inner folds' CV strategy, and thus, there is a different set of hyperparameters for each fold (e.g., Fig. 16) as opposed to one set of hyperparameters in used in Fig. 9 (Table III).

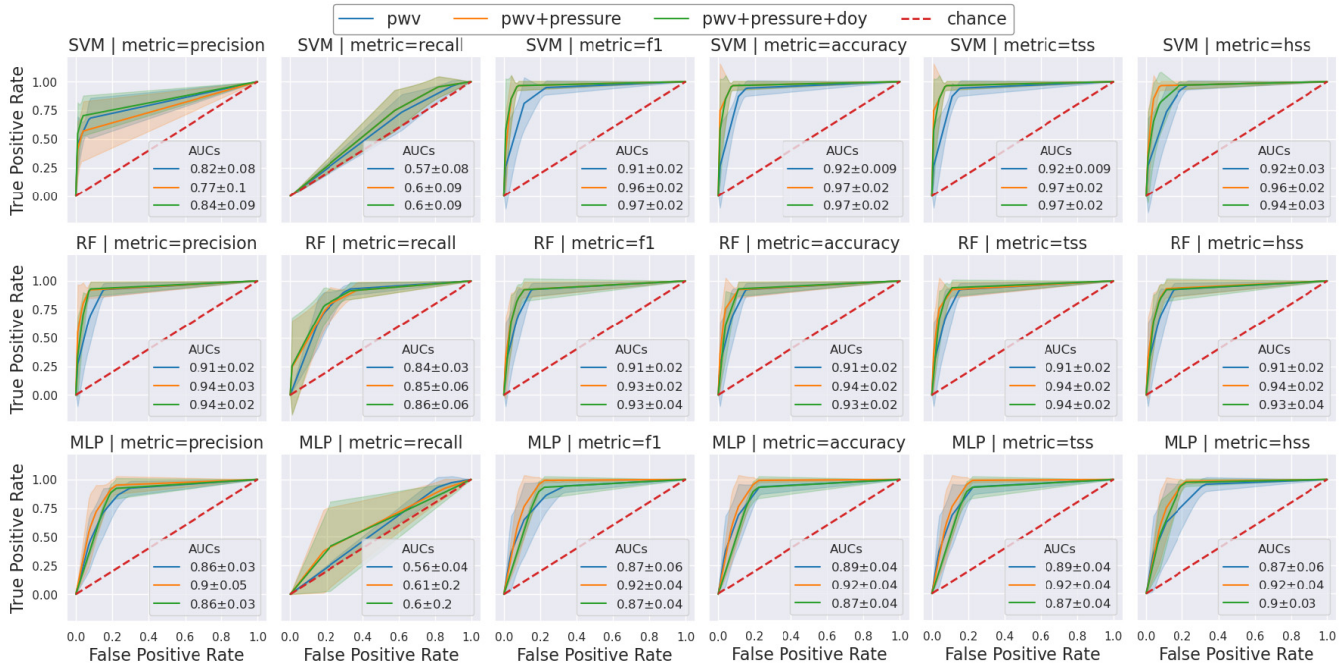

Fig. 20. Mean ROC curves for the SVM, RF, and MLP classifiers (row) with optimized hyperparameters for each fold individually where the positive and negative classes are balanced. Each set of hyperparameters was selected using a different metric (column). The feature groups consist of PWV (blue), PWV and surface pressure (orange), and PWV + surface pressure + DoY (green), where the chance curve indicates a no skill curve (red dashed line). The shaded area is the SD of four folds for each point in the ROC curve. The area under the curve is denoted in the legend with the SD of four fold as variability.

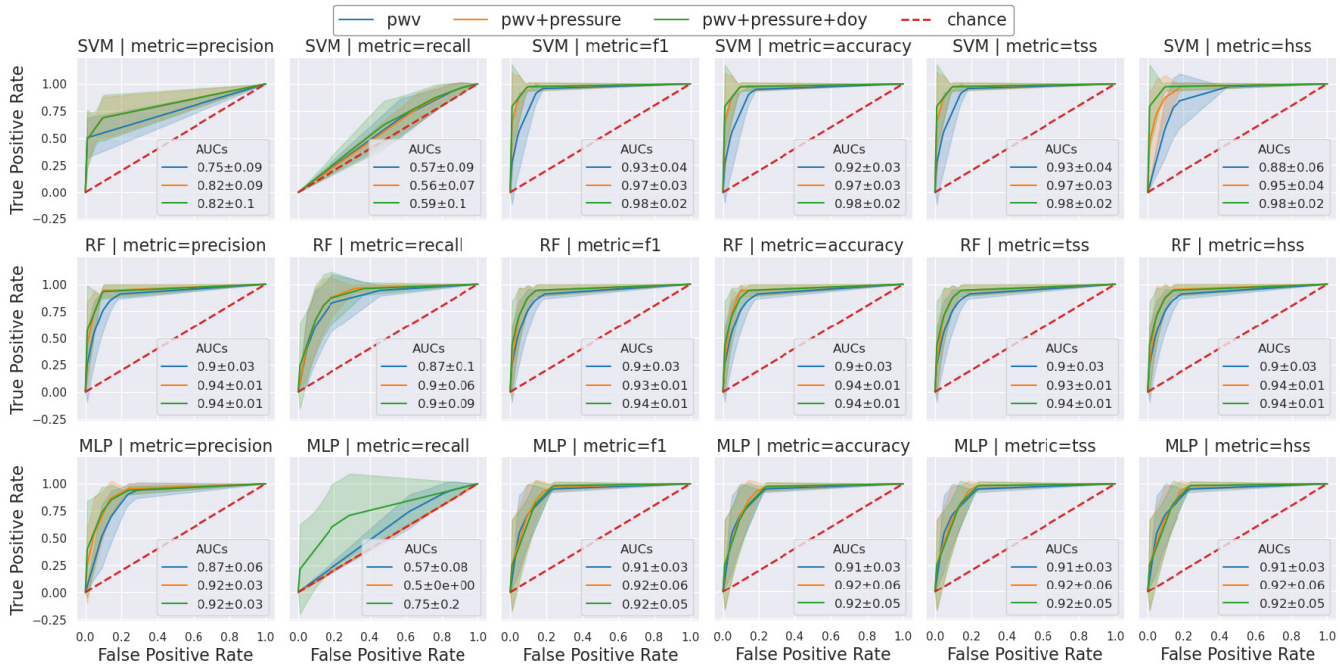

Fig. 21. Mean ROC curves for the SVM, RF, and MLP classifiers (row) with optimized hyperparameters for five inner folds' CV strategy where the positive and negative classes are balanced. Each set of hyperparameters was selected using a different metric (column). The feature groups consist of PWV (blue), PWV and surface pressure (orange), and PWV + surface pressure + DoY (green), where the chance curve indicates a no skill curve (red dashed line). The shaded area is the SD of five folds for each point in the ROC curve. The area under the curve is denoted in the legend with the SD of five fold as variability.

for predicting flash floods in the arid part of the EM region. The conclusions are given as follows.

1) Forning them together shows only a slight improvement.
2) The ROC curves showed that the SVM model achieved the highest mean AUC and the lowest AUC variability compared to the RF and MLP models.
3) The feature importance plots from the RF model showed that the PWV predictor is the most important one (72%), followed by surface pressure (27%) and DoY (<1%). An hourly breakdown of the PWV predictor shows a major peak from 2 to 6 h prior to a flood, with two smaller peaks on 14 and 19 h prior to a flood.

4) The nested CV technique is very informative and can quantify the model's performance variability due to data split selection. Furthermore, we show that for this dataset that is composed of 214 samples (balanced classes), a CV of five or five folds is either acceptable and produces similar results.
5) From all the score metrics that were used to find the optimal hyperparameters in this analysis, only recall was found unstable and resulted in poor ROC curves.
6) The permutation tests showed a clear class structure for the RF and SVM models, and however, the MLP achieved less than desirable results in these series of tests.
7) All the models have been tested with a highly imbalanced dataset, which simulates a more realistic flash flood occurrence scenario. The models show a drop in the false alarm rate (precision) with the hit rate (recall) remaining high.
8) A possible improvement to the flash flood prediction approach is to solve a multiclass classification task where the peak discharge can be used as a threshold parameter, i.e., predict whether the flood will be large, medium, or small.
9) The flash floods prediction approach as demonstrated in this work can be used to develop a near real-time flash floods early warning system.

ACKNOWLEDGMENT

The authors would like to thank the Israeli Water Authority for the floods data and the Israeli Meteorological Service for the surface temperature and pressure data.

REFERENCES

[1] M. Borga, E. N. Anagnostou, G. Blöschl, and J.-D. Creutin, "Flash flood forecasting, warning and risk management: The HYDRATE project," *Environ. Sci. Policy*, vol. 14, no. 7, pp. 834–844, Nov. 2011.
[2] V. Andréassian, A. Oddos, C. Michel, F. Anctil, C. Perrin, and C. Loumagne, "Impact of spatial aggregation of inputs and parameters on the efficiency of rainfall-runoff models: A theoretical study using chimera watersheds," *Water Resour. Res.*, vol. 40, no. 5, pp. 1–9, May 2004.
[3] S. Rozalis, E. Morin, Y. Yair, and C. Price, "Flash flood prediction using an uncalibrated hydrological model and radar rainfall data in a Mediterranean watershed under changing hydrological conditions," *J. Hydrol.*, vol. 394, nos. 1–2, pp. 245–255, Nov. 2010.
[4] D. Zoccatelli, M. Borga, F. Zanon, B. Antonescu, and G. Stancalie, "Which rainfall spatial information for flash flood response modelling? A numerical investigation based on data from the Carpathian range, Romania," *J. Hydrol.*, vol. 394, nos. 1–2, pp. 148–161, Nov. 2010.
[5] H. Yakir and E. Morin, "Hydrologic response of a semi-arid watershed to spatial and temporal characteristics of convective rain cells," *Hydrol. Earth Syst. Sci.*, vol. 15, no. 1, pp. 393–404, Jan. 2011.
[6] D. C. Goodrich, J. M. Faurès, D. A. Woolhiser, L. J. Lane, and S. Sorooshian, "Measurement and analysis of small-scale convective storm rainfall variability," *J. Hydrol.*, vol. 173, nos. 1–4, pp. 283–308, Dec. 1995.
[7] K. H. Syed, D. C. Goodrich, D. E. Myers, and S. Sorooshian, "Spatial characteristics of thunderstorm rainfall fields and their relation to runoff," *J. Hydrol.*, vol. 271, nos. 1–4, pp. 1–21, Feb. 2003.
[8] M.-L. Segond, H. S. Wheater, and C. Onof, "The significance of spatial rainfall representation for flood runoff estimation: A numerical evaluation based on the Lee catchment, UK," *J. Hydrol.*, vol. 347, nos. 1–2, pp. 116–131, Dec. 2007.
[9] M. Karklinsky and E. Morin, "Spatial characteristics of radar-derived convective rain cells over Southern Israel," *Meteorol. Zeitschrift*, vol. 15, no. 5, pp. 513–520, 2006.
[10] E. Morin, Y. Jacoby, S. Navon, and E. Bet-Halachmi, "Towards flash-flood prediction in the dry Dead Sea region utilizing radar rainfall information," *Adv. Water Resour.*, vol. 32, no. 7, pp. 1066–1076, 2009.
[11] N. Peleg and E. Morin, "Convective rain cells: Radar-derived spatiotemporal characteristics and synoptic patterns over the Eastern Mediterranean," *J. Geophys. Res., Atmos.*, vol. 117, no. D15, pp. 1–17, Aug. 2012.
[12] M. Bevis, S. Businger, T. A. Herring, C. Rocken, R. A. Anthes, and R. H. Ware, "GPS meteorology: Remote sensing of atmospheric water vapor using the global positioning system," *J. Geophys. Res.*, vol. 97, no. D14, pp. 15787–15801, Oct. 1992.
[13] M. Bevis *et al.*, "GPS meteorology: Mapping zenith wet delays onto precipitable water," *J. Appl. Meteorol.*, vol. 33, no. 3, pp. 379–386, Mar. 1994.
[14] Y. Reuveni, S. Kedar, S. E. Owen, A. W. Moore, and F. H. Webb, "Improving sub-daily strain estimates using GPS measurements," *Geophys. Res. Lett.*, vol. 39, no. 11, pp. 1–7, Jun. 2012.
[15] Y. Reuveni, S. Kedar, A. Moore, and F. Webb, "Analyzing slip events along the Cascadia margin using an improved subdaily GPS analysis strategy," *Geophys. J. Int.*, vol. 198, no. 3, pp. 1269–1278, Sep. 2014.
[16] Y. Reuveni, Y. Bock, X. Tong, and A. W. Moore, "Calibrating interferometric synthetic aperture radar (InSAR) images with regional GPS network atmosphere models," *Geophys. J. Int.*, vol. 202, no. 3, pp. 2106–2119, Sep. 2015.
[17] H.-J. Eueler and C. C. Goad, "On optimal filtering of GPS dual frequency observations without using orbit information," *Bull. Géodésique*, vol. 65, no. 2, pp. 130–143, Jun. 1991.
[18] J. L. Davis, T. A. Herring, I. I. Shapiro, A. E. E. Rogers, and G. Elgered, "Geodesy by radio interferometry: Effects of atmospheric modeling errors on estimates of baseline length," *Radio Sci.*, vol. 20, no. 6, pp. 1593–1607, 1985.
[19] G. D. Thayer, "An improved equation for the radio refractive index of air," *Radio Sci.*, vol. 9, no. 10, pp. 803–807, 1974.
[20] W. Bertiger *et al.*, "GipsyX/RTGx, a new tool set for space geodetic operations and research," *Adv. Space Res.*, vol. 66, no. 3, pp. 469–489, Aug. 2020.
[21] F. Webb and J. Zumberge, "An introduction to GIPSY/OASIS-II. JPL publication D-11088," Jet Propuls. Lab., Pasadena, CA, USA, Tech. Rep. D-11088, 1993.
[22] J. Duan *et al.*, "GPS meteorology: Direct estimation of the absolute value of precipitable water," *J. Appl. Meteorol.*, vol. 35, pp. 830–838, Jun. 1996.
[23] J. Boehm, B. Werl, and H. Schuh, "Troposphere mapping functions for GPS and very long baseline interferometry from European centre for medium-range weather forecasts operational analysis data," *J. Geophys. Res., Solid Earth*, vol. 111, no. B2, pp. 1–9, Feb. 2006.
[24] J. Boehm, A. Niell, P. Tregoning, and H. Schuh, "Global mapping function (GMF): A new empirical mapping function based on numerical weather model data," *Geophys. Res. Lett.*, vol. 33, no. 7, pp. 1–4, 2006.
[25] Z. Li, "Comparison of precipitable water vapor derived from radiosonde, GPS, and moderate-resolution imaging spectroradiometer measurements," *J. Geophys. Res.*, vol. 108, no. D20, pp. 1–20, 2003.
[26] J. Van Baelen, J.-P. Aubagnac, and A. Dabas, "Comparison of near–real time estimates of integrated water vapor derived with GPS, radiosondes, and microwave radiometer," *J. Atmos. Ocean. Technol.*, vol. 22, no. 2, pp. 201–210, Feb. 2005.
[27] J. Wang, L. Zhang, A. Dai, T. Van Hove, and J. Van Baelen, "A near-global, 2-hourly data set of atmospheric precipitable water from ground-based GPS measurements," *J. Geophys. Res.*, vol. 112, no. D11, pp. 1–17, 2007.
[28] D. Pérez-Ramírez *et al.*, "Evaluation of AERONET precipitable water vapor versus microwave radiometry, GPS, and radiosondes at ARM sites," *J. Geophys. Res., Atmos.*, vol. 119, no. 15, pp. 9596–9613, 2014.
[29] A. Leontiev and Y. Reuveni, "Combining Meteosat-10 satellite image data with GPS tropospheric path delays to estimate regional integrated water vapor (IWV) distribution," *Atmos. Meas. Tech.*, vol. 10, no. 2, pp. 537–548, Feb. 2017.
[30] Y. Wang *et al.*, "Evaluation of precipitable water vapor from four satellite products and four reanalysis datasets against GPS measurements on the Southern Tibetan Plateau," *J. Climate*, vol. 30, no. 15, pp. 5699–5713, Aug. 2017.

[31] A. Leontiev and Y. Reuveni, "Augmenting GPS IWV estimations using spatio-temporal cloud distribution extracted from satellite data," *Sci. Rep.*, vol. 8, no. 1, pp. 1–10, Dec. 2018.
[32] A. Leontiev, D. Rostkier-Edelstein, and Y. Reuveni, "On the potential of improving WRF model forecasts by assimilation of high-resolution GPS-derived water-vapor maps augmented with METEOSAT-11 data," *Remote Sens.*, vol. 13, no. 1, p. 96, Dec. 2020.
[33] S. Bonafoni, R. Biondi, H. Brenot, and R. Anthes, "Radio occultation and ground-based GNSS products for observing, understanding and predicting extreme events: A review," *Atmos. Res.*, vol. 230, Dec. 2019, Art. no. 104624.
[34] J. Van Baelen and G. Penide, "Study of water vapor vertical variability and possible cloud formation with a small network of GPS stations," *Geophys. Res. Lett.*, vol. 36, no. 2, pp. 1–4, Jan. 2009.
[35] E. Priego, J. Jones, M. J. Porres, and A. Seco, "Monitoring water vapour with GNSS during a heavy rainfall event in the Spanish Mediterranean area," *Geomatics, Natural Hazards Risk*, vol. 8, no. 2, pp. 282–294, Dec. 2017.
[36] A. W. Moore *et al.*, "National weather service forecasters use GPS precipitable water vapor for enhanced situational awareness during the Southern California summer monsoon," *Bull. Amer. Meteorolog. Soc.*, vol. 96, no. 11, pp. 1867–1877, Nov. 2015.
[37] H. K. Huelsing, J. Wang, C. Mears, and J. J. Braun, "Precipitable water characteristics during the 2013 Colorado flood using ground-based GPS measurements," *Atmos. Meas. Techn.*, vol. 10, no. 11, pp. 4055–4066, Nov. 2017.
[38] B. Lynn, Y. Yair, Y. Levi, S. Z. Ziv, Y. Reuveni, and A. Khain, "Impacts of non-local versus local moisture sources on a heavy (and deadly) rain event in Israel," *Atmosphere*, vol. 12, no. 7, p. 855, Jun. 2021.
[39] W. Sun *et al.*, "Forecasting of ionospheric vertical total electron content (TEC) using LSTM networks," in *Proc. Int. Conf. Mach. Learn. Cybern. (ICMLC)*, vol. 2, Jul. 2017, pp. 340–344.
[40] L. Liu, S. Zou, Y. Yao, and Z. Wang, "Forecasting global ionospheric total electron content (TEC) using deep learning," in *Proc. AGU Fall Meeting Abstr.*, 2020, pp. 4–17.
[41] S. Asaly, L.-A. Gottlieb, and Y. Reuveni, "Using support vector machine (SVM) and ionospheric total electron content (TEC) data for solar flare predictions," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 1469–1481, 2021.
[42] S. Asaly, L.-A. Gottlieb, N. Inbar, and Y. Reuveni, "Using support vector machine (SVM) with GPS ionospheric TEC estimations to potentially predict earthquake events," *Remote Sens.*, vol. 14, no. 12, p. 2822, Jun. 2022.
[43] L.-T. Hsu, "GNSS multipath detection using a machine learning approach," in *Proc. IEEE 20th Int. Conf. Intell. Transp. Syst. (ITSC)*, Oct. 2017, pp. 1–6.
[44] N. Linty, A. Farasin, A. Favenza, and F. Dovis, "Detection of GNSS ionospheric scintillations based on machine learning decision tree," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 55, no. 1, pp. 303–317, Feb. 2019.
[45] S. Ziskin Ziv, Y. Yair, P. Alpert, L. Uzan, and Y. Reuveni, "The diurnal variability of precipitable water vapor derived from GPS tropospheric path delays over the Eastern Mediterranean," *Atmos. Res.*, vol. 249, Feb. 2021, Art. no. 105307.
[46] Y. E. Bar-Sever, "Real-time GNSS positioning with JPL's new GIPSYx software," in *Proc. AGU Fall Meeting Abstr.*, 2016, p. 4.
[47] S. Z. Ziv, P. Alpert, and Y. Reuveni, "Long-term variability and trends of precipitable water vapour derived from GPS tropospheric path delays over the Eastern Mediterranean," *Int. J. Climatol.*, vol. 41, no. 5, pp. 6433–6454, 2021.
[48] H. Saaroni and B. Ziv, "Summer rain episodes in a Mediterranean climate, the case of Israel: Climatological-dynamical analysis," *Int. J. Climatol.*, vol. 20, no. 2, pp. 191–209, Feb. 2000.
[49] F. Pedregosa *et al.*, "Scikit-learn: Machine learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, Jan. 2011.
[50] B. E. Boser, I. M. Guyon, and V. N. Vapnik, "A training algorithm for optimal margin classifiers," in *Proc. 5th Annu. Workshop Comput. Learn. Theory*, 1992, pp. 144–152.
[51] C. Cortes and V. Vapnik, "Support vector machines," *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, Sep. 1995.
[52] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.
[53] G. E. Hinton, "Connectionist learning procedures," in *Machine Learning*. Amsterdam, The Netherlands: Elsevier, 1990, pp. 555–610.
[54] G. B. Orr and K.-R. Müller, *Neural Networks: Tricks of the Trade*. Berlin, Germany: Springer, 2003.
[55] V. Landa and Y. Reuveni, "Low-dimensional convolutional neural network for solar flares GOES time-series classification," *Astrophys. J. Suppl. Ser.*, vol. 258, no. 1, p. 12, Jan. 2022.
[56] A. Hanssen and W. Kuipers, *On the Relationship Between the Frequency of Rain and Various Meteorological Parameters. (With Reference to the Problem of Objective Forecasting)*. De Bilt, The Netherlands: Koninklijk Nederlands Meteorologisch Instituut, 1965.
[57] D. S. Bloomfield, P. A. Higgins, R. T. J. McAteer, and P. T. Gallagher, "Toward reliable benchmarking of solar flare forecasting methods," *Astrophys. J.*, vol. 747, no. 2, p. L41, Mar. 2012.
[58] J. Friedman *et al.*, *The Elements of Statistical Learning* (Springer Series in Statistics), vol. 1. New York, NY, USA: Springer, 2001.
[59] G. C. Cawley and N. L. Talbot, "On over-fitting in model selection and subsequent selection bias in performance evaluation," *J. Mach. Learn. Res.*, vol. 11, pp. 2079–2107, Jul. 2010.
[60] T. Hastie, R. Tibshirani, and J. Friedman, "Model assessment and selection," in *The Elements of Statistical Learning*. New York, NY, USA: Springer, 2009, pp. 219–259.
[61] R. Kohavi *et al.*, "A study of cross-validation and bootstrap for accuracy estimation and model selection," in *Proc. Int. Joint Conf. AI*, Montreal, QC, Canada, Aug. 1995, vol. 14, no. 2, pp. 1137–1145.
[62] M. Ojala and G. C. Garriga, "Permutation tests for studying classifier performance," *J. Mach. Learn. Res.*, vol. 11, no. 6, pp. 1833–1863, 2010.
[63] H. Ishwaran, "Variable importance in binary regression trees and forests," *Electron. J. Stat.*, vol. 1, pp. 519–537, Nov. 2007.
[64] C. Strobl, A.-L. Boulesteix, A. Zeileis, and T. Hothorn, "Bias in random forest variable importance measures: Illustrations, sources and a solution," *BMC Bioinf.*, vol. 8, no. 1, pp. 1–21, Dec. 2007.
[65] L. S. Shapley, *17. A Value for N-Person Games*. Princeton, NJ, USA: Princeton Univ. Press, 2016.
[66] S. M. Lundberg *et al.*, "From local explanations to global understanding with explainable AI for trees," *Nature Mach. Intell.*, vol. 2, no. 1, pp. 56–67, 2020.

Shlomi Ziskin Ziv received the B.Sc., M.Sc., and Ph.D. degrees from Hebrew University, Jerusalem, Israel, in 2007, 2009, and 2018, respectively.

He joined the Remote Sensing Laboratory, Physics Department, Ariel University, Ariel, Israel, in 2019. His research focuses primarily on data sciences and geophysics, but he is also engaged in using machine learning (ML) in other fields (e.g., social sciences) in order to better understand various phenomena.

Yuval Reuveni received the B.Sc., M.Sc., and Ph.D. degrees from Tel Aviv University, Tel Aviv, Israel, in 2002, 2005, and 2011, respectively.

He joined the Eastern Research and Development Center, Ariel, Israel, as the Head of the Department of Geophysics and Space Sciences in 2014. He joined the Physics Department, Ariel University, Ariel, in 2015. His research focuses primarily on combining various data analysis techniques and remote sensing measurements from ground and space-based instruments, to study electromagnetic wave propagation, ionospheric physics, space weather, and space geodesy phenomena.