Stochastic Environmental Research and Risk Assessment (2025) 39:3333–3352
https://doi.org/10.1007/s00477-025-03020-1

ORIGINAL PAPER

Seasonal heatwave forecasting with explainable machine learning and remote sensing data

Jung-Ching Kan[1,2] · Marlon Vieira Passos[1,2] · Georgia Destouni[1,3] · Karina Barquet[2] · Carla S. S. Ferreira[4,5] · Zahra Kalantari[1]

Accepted: 10 May 2025 / Published online: 5 June 2025

Abstract
Heatwaves can greatly impact societies, underscoring the need to extend current heatwave prediction lead times. This study investigates multiple machine learning (ML) model approaches for heatwave occurrence prediction with long lead times of one to five months. Five ML classifiers, built using Google Earth Engine remote sensing datasets, are developed and tested for heatwave prediction for the national scale (case example of Sweden) over time period 1989–2019. The ML modelling is based on 13 final explanatory atmospheric and landscape features. The balanced random forest model exhibits the consistently best performance among the tested ML models, stable across all investigated lead times (from one to five months) with balanced accuracy of around 0.77, even though not overall identifying actual heatwave occurrence (decreased recall for heatwave occurrence from 0.87 to 0.81). Application of SHapley Additive exPlanations technique for model interpretation shows increasing importance of model output with increasing lead time for landscape features such as runoff and soil water. Overall, more frequent heatwave occurrence emerges for places characterized by lower values of geopotential height, evaporation, precipitation, and topographical slope, and higher values of temperature, runoff, and sea level pressure. The study also exemplifies how the developed ML modelling approach could be used to identify and warn for early signs of forthcoming heatwave occurrence, and further step-wise improve the identification and warning toward less uncertainty for shorter lead times. This can facilitate earlier warning in support of better planning of measures to mitigate adverse heatwave impacts, up to several months ahead of their possible occurrence.

Keywords Summer heatwaves · Machine-learning models · Explanatory-predictive factors · Landscape factors · Atmospheric climate factors · Geopotential height

1 Introduction

A heatwave is an extreme temperature event defined as a period of excessive heat that usually lasts for one week or at least three days (Perkins-Kirkpatrick and Lewis 2020). Research into the impacts of heatwaves on human health has consistently revealed that a disproportionately high number of older individuals are affected (Xu et al. 2016; Vu et al. 2019; Rodrigues et al. 2021). The 2003 heatwave in Europe led to 70,000 excess mortalities (Robine et al. 2008) and over 13 billion euros in economic damage (United Nations Environment Programme 2003). Heatwaves can also have far-reaching cascading and compound impacts, exacerbating drought impacts on water, soil, energy and agriculture sectors (Niggli et al. 2022; Soares et al. 2023). Global predictions indicate that climate change will lead to increased intensity, frequency, and duration of heatwaves, with Europe

✉ Jung-Ching Kan
jckan@kth.se

1 Department of Sustainable Development, Environmental Science and Engineering, Sustainability Assessment and Management, KTH Royal Institute of Technology, SE-100 44, Stockholm, Sweden

2 Stockholm Environment Institute (SEI), Stockholm, Sweden

3 Department of Physical Geography and Bolin Centre for Climate Research, Stockholm University, 106 91 Stockholm, Sweden

4 Polytechnic Institute of Coimbra, Applied Research Institute, Rua da Misericórdia, Lagar dos Cortiços– S. Martinho do Bispo, 3045-093 Coimbra, Portugal

5 Research Centre for Natural Resources Environment and Society (CERNAS), Polytechnic Institute of Coimbra, Bencanta, 3045-601 Coimbra, Portugal

identified as a future hotspot. The projected upward trend for Europe is three to four times faster than in other northern mid-latitude regions, with Europe experiencing a mean increase of 0.61 heatwave days per decade, compared to a mean increase of 0.21 heatwave days per decade in other mid-latitudes regions worldwide (Rousi et al. 2022).

Early warnings of heatwaves with a sufficient lead time can enable timely and more effective societal responses to protect vulnerable populations and socioeconomic activities (Merz et al. 2020). Extending heatwave predictions to a seasonal lead time would allow for a timelier and more effective risk reduction and longer term health system preparedness, with better resource management (Weirich-Benet et al. 2023). State-of-the-art deterministic heatwave prediction models that employ physical laws to solve the equations of atmospheric physics to forecast future states of the atmosphere can accurately predict heatwaves up to two weeks in advance (Lorenz 1963; Domeisen et al. 2023). Predictions beyond two weeks lead time using deterministic approaches are challenging due to the increase of inherent uncertainty in complex variables interactions over time, such as between atmospheric variables (e.g. wind speed and temperature) and landscape variables status and feedbacks (e.g. land cover and soil moisture) (Domeisen et al. 2023). Heatwaves are often associated with high-pressure systems, while memory in landscape states, such as moisture deficit and land–atmosphere feedbacks, plays a crucial role in atmospheric predictability on sub-seasonal time scales (Ford et al. 2018). Knowledge gaps in atmospheric and landscape model parameterisations, as well as forecast initialisation, impact the reliability of forecasts and require further research, particularly for sub-seasonal heatwave prediction (Ford et al. 2018). The challenges of detecting early signs and forecasting heatwaves weeks in advance are further compounded by a lack of historical data on extreme events, the difficulties of building reliable statistics using climate models (due to the substantial computational cost required to simulate a sufficient number of events), and biases in models that hinder accurate quantitative assessments of extremes given the scarcity of available data (Jacques-Dumas et al. 2022).

As the capabilities of current deterministic approaches for heatwave prediction beyond a 2-week lead time may be insufficient for deploying effective protection and mitigation strategies, more studies now explore the potential of machine learning (ML) models (Barriopedro et al. 2023). Use of ML-based approaches can potentially increase heatwave prediction lead times as ML models have the advantage of identifying, rather than pre-assuming, the complex interactions between various possible explanatory and predictive factors. ML-based modelling therefore is increasingly applied in studies of hydro-climatic hazards, such as floods, droughts, and heatwaves (Rahmati et al. 2020; Panahi et al. 2022). For example, a study by Khan et al. (2021) achieved accurate predictions with a one-month lead time of total summer heatwave days (HWD) in Pakistan using a support vector machine. Further, Asadollah et al. (2022) developed and compared three ML models (AdaBoost regression decision tree (ABR-DT), Random Forest, and Decision Tree) for total summer HWD prediction in Iran with a three-month lead time, and found ABR-DT to be the most accurate. Weirich-Benet et al. (2023) predicted heatwaves at weekly resolution using different approaches and concluded that ML can improve sub-seasonal heatwave prediction. Straaten et al. (2022) found that explainable ML using high-dimensional remote sensing data can complement physically-based models for heatwave prediction, providing an effective alternative for such predictions with a lead time longer than two weeks. However, the development and application of ML models for heatwave research and prediction are still in their infancy compared with work on other natural hazards, such as droughts and floods (Mosavi et al. 2018; Gyaneshwar et al. 2023). Additionally, most ML/deep learning-based heatwave studies to date have considered sub-seasonal forecast lead times (less than one month), without applying model explainability techniques, and the focus has been on relatively warm and/or dry climate regions such as the Middle East and Central Europe (Domeisen et al. 2023).

The aim of this study is to advance ML model development for heatwave prediction by creating a comprehensive framework for seasonal prediction with longer lead times, useful for practical implementation. This research supports interdisciplinary efforts to mitigate climate change risks and underscores the importance of integrating scientific advances into environmental management and policy frameworks. As a practical case study, we focus here on predicting summer HWD at a monthly resolution for the national-scale case example of Sweden. The specific objectives of this study are: (i) to evaluate and compare the performance of various ML models developed using relevant remote sensing data for the potential prediction of heatwave occurrences with lead times ranging from one to five months, and (ii) to analyse the relative importance of various explanatory-predictive features (e.g. related to the atmospheric climate, the landscape, and other relevant case aspects) in terms of their influence the on ML model outputs.

2 Materials and methods

2.1 Study area

Sweden is located in the high-latitude region of northern Europe (55–69°N; 11–24°E), which is characterised by long, cold winters (Sköld Gustafsson et al. 2023b). Although major documented natural hazards in Sweden mostly comprise flooding and wildfire events (Sköld Gustafsson et al. 2023a), heatwave exposure is of increasing concern due to rising temperatures over the past two decades, e.g. mean peak temperatures in Sweden were higher in the period 1991–2020 (Fig. 1b) than in 1961–1990 (Fig. 1a). Trend analysis suggests that heavily populated areas in the southern part of the country will experience more prolonged and frequent heatwaves in the near future (Vieira Passos et al. 2024). A persistent heatwave in Sweden in the summer of 2018 resulted in around 750 cases of excess mortality (SMHI 2020). Sweden faces a considerable risk of heatwave impacts due to its population being more acclimated to cooler climates. For example, Sweden is accustomed to building homes and organizing activities designed to withstand severe cold rather than heat (SWECO 2024). Additionally, heatwaves are generally the most common primary hazard interacting with other natural hazards (Sköld Gustafsson et al. 2023b). Heatwaves in Sweden mostly occur in the summer months (June–August) and are defined in this study as days with the maximum temperature exceeding 27 °C for at least three consecutive days, based on the threshold adopted by the Swedish Meteorological and Hydrological Institute (SMHI) for issuing high-temperature warnings (Oudin Åström et al. 2020).

2.2 Development method

The framework developed and tested in this study comprised three main steps (Fig. 2): (1) data extraction, (2) data pre-processing, (3) model training, validation, and testing.

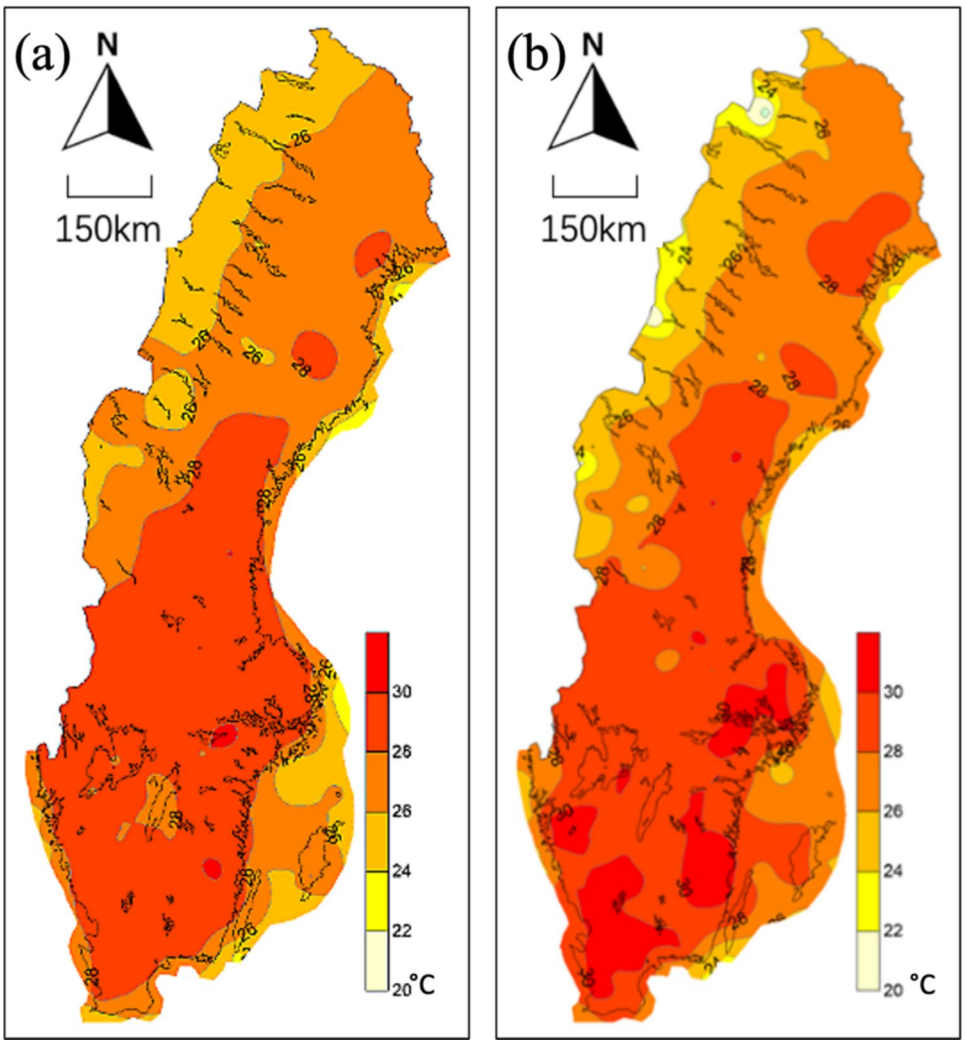

Fig. 1 Maps of Sweden showing the average value of the years' highest temperatures during the periods (**a**) 1961–1990; (**b**) 1991–2020 (adapted from SMHI 2023)

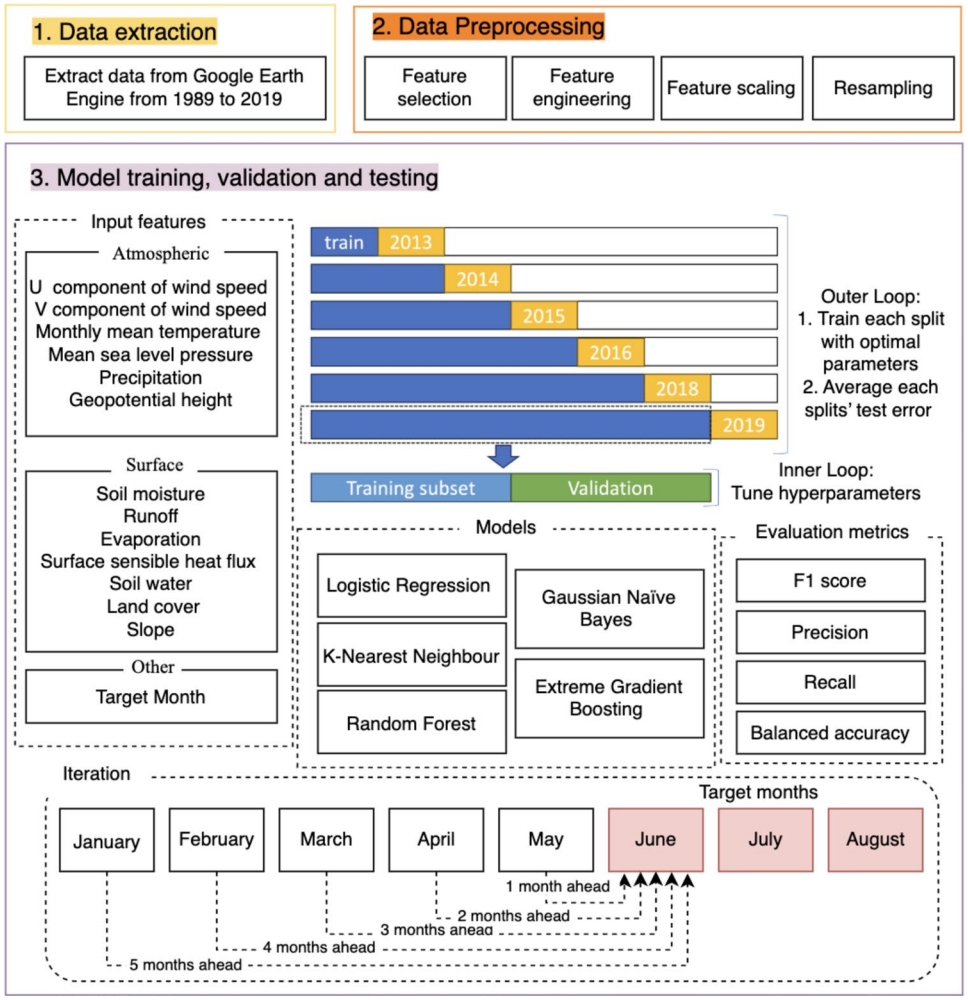

Fig. 2 Flowchart of the framework developed and used for heatwave prediction in Sweden

2.2.1 Data extraction

A total of 19 possible explanatory and predictive physical features, representing various factors that may influence the heatwave dynamics, were selected for ML heatwave modelling. These features were initially selected based on a comprehensive literature review (Table 1) (Khan et al. 2021; Asadollah et al. 2022; Weirich-Benet et al. 2023). Atmospheric features, including the u component of wind speed (u wind), v component of wind speed (v wind), monthly mean temperature, mean sea level pressure (MSLP), specific humidity, precipitation, and geopotential height, play important roles in atmospheric circulation and have been used for the synoptic analysis of heatwave phenomena in the Northern Hemisphere (Drouard et al. 2019; Fonseca-Rodríguez et al. 2023). In addition to atmospheric processes, regional controls of heatwaves occurrence at sub-seasonal time scale may also include landscape features, such as evaporation, soil moisture, runoff, land cover/ use properties, and topographical slope, influencing the evaporative cooling through latent heat flux and surface sensible heat flux (SSHF) (Olsson et al. 2016; Domeisen et al. 2023; Tak et al. 2024). Studies of the 2018 heatwave in northern Europe have shown that soil water content deficits suppressed cloud formation through land–atmosphere feedbacks, contributing to a prolonged period of high temperatures (Tak et al. 2024). Additionally, the target month is included as a feature, indicating whether the heatwave prediction is for June, July, or August. This inclusion helps capture monthly variability, as certain conditions may consistently precede heatwaves in July but not in June.

The atmospheric and landscape features were retrieved from the remote sensing dataset using Google Earth Engine (GEE). GEE is a cloud computing platform that allows the acquisition of data from various sources for any specific study area worldwide, and is increasingly being used in heatwave-related studies. The study period in GEE was set to 1989–2019, with a resolution of 27,830 m based on data availability and quality. GEE utilises the image pyramid technique to manage data from multiple sources and each source is ingested at its native resolution (Gorelick et al. 2017). The pyramid structure allows GEE to efficiently

Table 1 Features considered in this study, data source for each feature and brief description

Feature	Category	Data source	Description
u wind	Atmospheric	ERA 5	Monthly average 10 m u-component of wind
v wind	Atmospheric	ERA 5	Monthly average 10 m v-component of wind
Mean temperature	Atmospheric	ERA 5	Monthly average air temperature at 2 m height
Mean sea level pressure	Atmospheric	ERA 5	Monthly average sea level pressure
Specific humidity	Atmospheric	FLDAS	Monthly average specific humidity
Precipitation	Atmospheric	ERA 5	Monthly sum of total precipitation
Geopotential height	Atmospheric	NCEP	Monthly average surface geopotential height
Soil moisture	Landscape	FEWS NET	Monthly average soil moisture 0–10 cm underground
Runoff	Landscape	ERA 5 land	Monthly sum of surface runoff and subsurface runoff
Evaporation	Landscape	ERA 5 land	Monthly sum of total evaporation
Latent heat flux	Landscape	ERA 5 land	Monthly average latent heat exchange with surface
Sensible heat flux	Landscape	FLDAS	Monthly average heat transfer between surface and atmosphere
Soil water 1	Landscape	ERA 5 land	Monthly average volume of water in 0–7 cm soil layer
Soil water 2	Landscape	ERA 5 land	Monthly average volume of water in 7–28 cm soil layer
Soil water 3	Landscape	ERA 5 land	Monthly average volume of water in 28–100 cm soil layer
Soil water 4	Landscape	ERA 5 land	Monthly average volume of water in 100–289 cm soil layer
Land cover	Landscape	ESA	22 land cover classes defined with the United Nations Land Cover Classification System
Slope	Landscape	USGS	
Target month	Spatial–temporal		6, 7, 8 correspond to targeted months as in June, July, and August

handle large datasets, while also allowing users to select data based on spatial and temporal specifications (Gorelick et al. 2017). For continuous value images, pixel values at the upper levels of the pyramid represent the mean of the pixels from the lower levels. In contrast, for the discrete-value land cover image, pixel values correspond to the most frequently occurring values. The built-in ee.Terrain.slope function in Google Earth Engine is used to calculate terrain slope from a digital elevation model (DEM). This function determines the steepness or gradient of the terrain at each point by analysing the rate of elevation change between adjacent pixels in the DEM. Figure 3a illustrates the Pearson correlation coefficient (R) between some considered features and the target number of heatwave days, suggesting that there are no linear relationships between the selected indicators and heatwave occurrence. Hence, advanced ML models could be deployed to address the nonlinear relationship. Heatwave days were identified based on daily maximum temperature. Since almost all heatwaves in the case of Sweden occur in the summer months (June, July, or August) (Sjulgård et al. 2023), only these were set as target months. Monthly heatwave occurrence was further defined using binary classification, with a value of 1 assigned for one or more HWD occurring during a month (Class 1) and a value of 0 assigned for no HWD occurrence (Class 0). Figure 3b illustrates the density distribution of the two classes for selected continuous features.

2.2.2 Data pre-processing

The data obtained from GEE were subjected to four data pre-processing procedures, feature selection, engineering, scaling and data resampling to ensure high-quality, consistent, and suitable data for subsequent modelling.

2.2.3 Feature selection

The feature selection consisted of identifying an effective subset of features from the full set by removing redundant features (Tang et al. 2014). The feature selection method applied in this study is a correlation-based filter that searches feature subsets according to the degree of redundancy and finds features with low intercorrelation (Khalid et al. 2014). Features with strong linear correlation contain

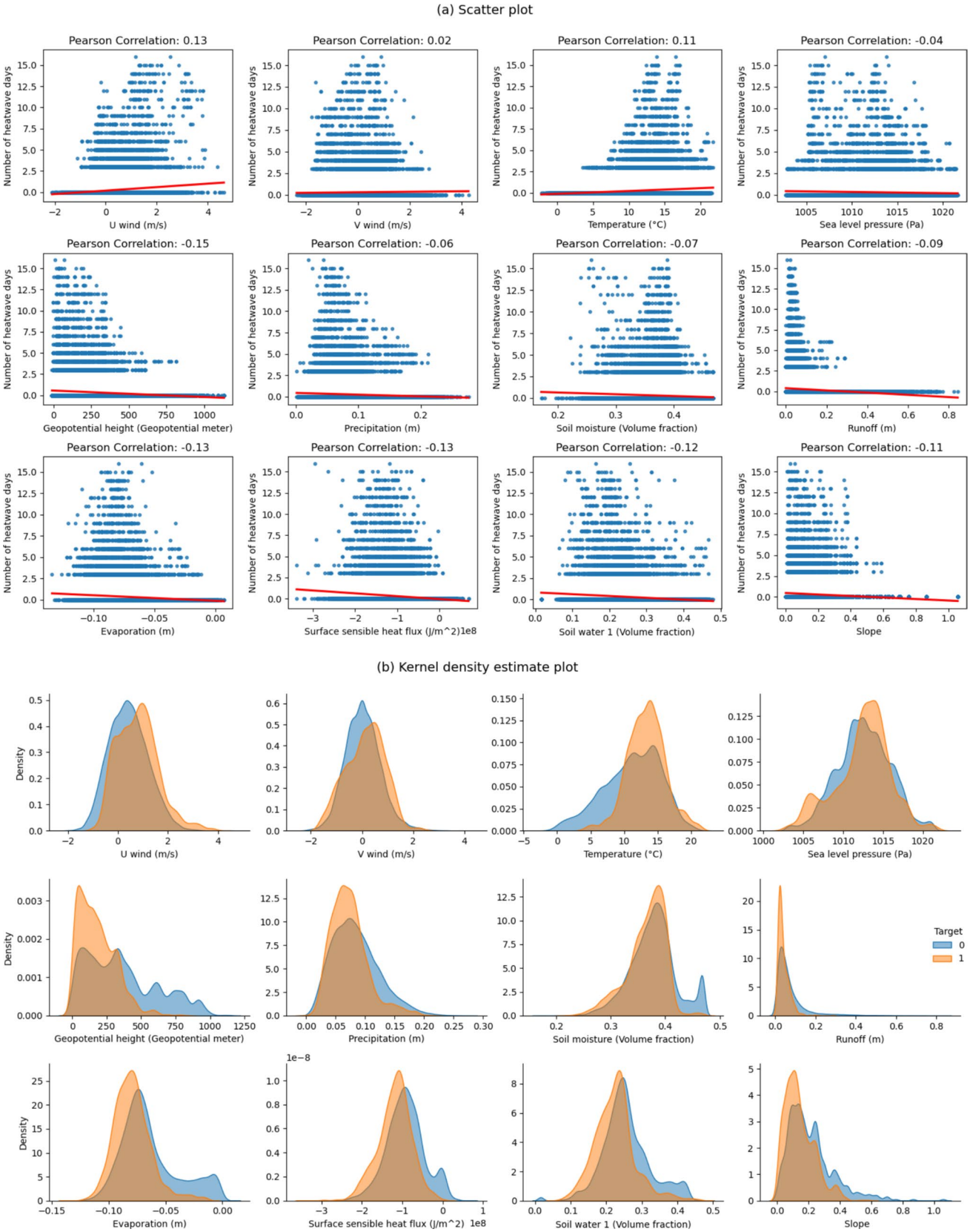

Fig. 3 (**a**) Scatterplots of selected continuous features and the number of heatwave days in the target months (June, July, and August) between 1989 and 2019. Low R values suggest that there is no linear relationship between the features and target. The red colour represents the regression line. (**b**) Kernel density estimate plot of selected continuous features and target class

highly similar information, and Pearson correlation is often used for redundancy analysis (Khalid et al. 2014). A Pearson correlation matrix of the 19 features was plotted with a heatmap to examine mutually correlated feature groups. Only one feature from each highly correlated feature group was kept and the others were excluded in the further ML modelling.

2.2.4 Feature engineering

Feature engineering is the process of designing features to represent combined actions or underlying patterns based on domain knowledge (Qu et al. 2023). To better utilise possible patterns and relationships over time, extracting them from the time series data into features that can better represent them (Verdonck et al. 2024). The means, standard deviations, and linear regression trends of values in previous months were calculated to capture central tendencies, variability, and directional movements over time. The values of precipitation, evaporation, soil moisture, and soil water in previous months were aggregated to assess the overall hydrological status of the pixel. Wind speed is calculated by combining the U and V components of wind and constructed as new features to measure the integrated effect of wind direction and intensity.

2.2.5 Feature scaling

Feature scaling was applied to transform selected features to a common range, minimising the bias of feature value distribution (Singh and Singh 2020). This procedure is required since different features have distinct value ranges, and greater numerical feature values may misleadingly dominate over smaller numerical ones (Singh and Singh 2020). The minimum–maximum normalisation method was applied, setting the maximum value equal to 1 and the minimum equal to 0. BRF and XGBoost models were excluded from the feature scaling process due to tree-based models don't require feature scaling.

2.2.6 Data resampling

Resampling is a state-of-the-art solution to address imbalanced data (Alam et al. 2020). Based on historical data, months in which a heatwave occurred were relatively infrequent compared with months without a heatwave. This caused a class imbalance problem, leading to bias towards the majority class (Class 0) (Lin et al. 2017). Data imbalance is a common problem when dealing with real-world applications, e.g. natural hazards (Oommen et al. 2011). To balance the data without losing essential information, cluster centroids, a clustering-based under-sampling technique (Alam et al. 2020), was applied in this study. Cluster centroids replace the majority class samples with cluster centres, reducing the amount of majority class samples while maintaining the representative variations (Tsai et al. 2019). This resampling technique was only applied to the training dataset for the ML models.

2.2.7 Model training, validation, and testing

A nested cross-validation scheme was implemented in this study using time series split cross-validaton and day-forward strategy, tailored to the time series nature of the data. The day-forward strategy ensures the models only use past data to predict future events, thereby avoiding look-ahead bias. Initially, each model was trained and validated in the inner loop using inner loop data that was time series split into two splits for hyperparameters optimisation. Subsequently, the models were refitted to the entire inner loop data and tested against the outer loop's test years (2013, 2014, 2015, 2016, 2018, and 2019), constituting approximately 20% of the collected data. Feature scaling and resampling are applied exclusively to the training data in each fold. This is achieved by integrating these processes within each model's pipeline, ensuring that the features for scaling and the resampling adjustments are learned only from the training dataset and then applied to the corresponding validation or test sets to prevent data leakage. The results in terms of average performance for these test years ensure a robust representation of the overall model performance. The year 2017 was excluded from testing as no recorded heatwave events occurred in that year. With class 1 missing, model performance metrics (F1 score, precision, and recall) cannot be calculated for 2017. As a result, only partial insights into model performance can be obtained from model testing for that year. Lead times ranging from one to five months were tested, and the overall setup described above was repeated for each lead time. In total, 25 ML models were developed and evaluated in this study. Following a literature review, five popular ML classifiers (Extreme Gradient Boosting, Gaussian Naïve Bayes, K-Nearest Neighbour, Balanced random Forest, and logistic regression) were selected and used in the study based on their relevant characteristics. Brief summaries of each model class and their applications are provided below.

2.2.8 Extreme gradient boosting

Extreme Gradient Boosting (XGBoost) is a highly effective and scalable algorithm derived from the ensemble of decision trees (Chen and Guestrin 2016). It has a built-in regularisation function to reduce overfitting, and parallel processing to speed up the model training process (Asadollah et al. 2022). A previous study used an XGBoost model

to predict the number of heat-related ambulance calls and achieved high accuracy (Ke et al. 2023).

2.2.9 Gaussian Naïve bayes

Gaussian Naïve Bayes (NB) is an efficient classifier identified as one of the top ten algorithms in data mining (Wu et al. 2008). Gaussian NB is based on Bayes' theorem, assuming the data are Gaussian-distributed and that all the features are independent (Wu et al. 2008). It has given satisfactory results in urban flood depth prediction (Wang et al. 2021).

2.2.10 K-Nearest neighbour

K-Nearest Neighbour (KNN) is a generic non-parametric algorithm that does not make assumptions on data distribution (Wu et al. 2008). KNN predicts outcomes by identifying the k nearest data points in the feature space, using measures such as Euclidean distance. For classification tasks, it assigns the class of a data point based on the majority class among its k nearest neighbours (Guo et al. 2003). This model has been extensively applied in drought prediction, with good results (Raja and Gopikrishnan 2022).

2.2.11 Balanced random forest

Balanced Random Forest (BRF) is an ensemble model that modifies the traditional random forest algorithm to better handle imbalanced data. Classic random forest consists of decision trees built using bootstrap samples drawn from the entire training dataset. In contrast, BRF draws equal-sized samples from each class ensuring that the trees in the forest are trained on a balanced set of data (More and Rana 2017). Random Forest is one of the most widely applied models in environmental research, and has been widely tested in previous heatwave-related studies (Asadollah et al. 2022; Weirich-Benet et al. 2023). BRF has been successfully applied in wildfire risk assessment compared to standard random forest and XGBoost, primarily because wildfire datasets are typically imbalanced and BRF effectively identifies the minority class (Wang et al. 2022). This characteristic of imbalance is also present in heatwave datasets. Although BRF has not yet been explicitly tested in heat-related studies, its proven capability to manage class imbalance suggests it could be effective for heatwave prediction.

2.2.12 Logistic regression

Logistic regression (LR) is a robust and flexible classification model for predicting a binary outcome, such as yes/no, when the features are continuous (Bartosik and Whittingham 2021). Logistic regression calculates a linear combination of the input features and then transforms them with a sigmoid function (Bartosik and Whittingham 2021). It is widely applied in the clinical field (Kannan et al. 2023) and hazard research, e.g. mapping of flood susceptibility and drought spatial patterns (Al-Juaidi et al. 2018; Niaz et al. 2021).

2.2.13 Baseline model based on probability

A baseline model is developed using historical data on heatwave occurrence to predict future events. This model assesses the probability of a heatwave in a given year based on historically observed heatwaves from the year 1989 to the given year at each pixel. It calculates the frequency of heatwaves during summer months and uses the ratio as a probability estimate for future predictions. A class is randomly generated depending on the calculated probability. This model primarily served as an initial benchmark and a point of comparison for evaluating the performance of more complex models with additional features.

2.2.14 Hyperparameter tuning

Hyperparameters are external configuration variables of algorithms. Hyperparameter tuning consists of finding the optimal settings of hyperparameters to classify imbalanced data at algorithm level (Rosales-Pérez et al. 2023), possibly achieving significant ML model improvements (Kong et al. 2019). Random search, one of the two most widely used strategies for hyperparameter tuning, was selected for use in this study due to its advantages in terms of efficiency and flexibility (Bergstra and Bengio 2012). The tuning process was integrated into the nested cross-validation framework, specifically within the inner loop. Random search was performed 50 iterations per loop, enabling a thorough exploration of the parameter space while managing computational costs. The hyperparameter tuning process concluded in the selection of the best hyperparameters based on the performance metric of balanced accuracy (as detailed in Sect. 2.2.4). The hyperparameter grids for the ML models are listed in Table 2.

2.2.15 Model evaluation

The minority data class (with a heatwave in a month) was defined as positive (Class 1), while the majority class (with no heatwave in a month) was defined as negative (Class 0). Seven evaluation metrics, F1 score, precision, and recall for both classes, and balanced accuracy (all defined in Eqs. (1)–(4) below), were selected to evaluate model performance. The F1 score symmetrically combines and represents in one metric both precision and recall (considered here also

Table 2 Hyperparameter search space for different classifiers

Classifier	Hyperparameter	Value range	References
Logistic regression	Inverse of regularization strength (C)	Loguniform (1e−5, 100)	Sun et al. (2020)
	Solver	lbfgs, newton-cg, liblinear, sag, saga	
	Max iterations	(100, 5000)	
Balanced random forest	Number of estimators	(50, 200)	Sun et al. (2020)
	Max features	sqrt, log2	
	Max depth	(3, 30)	
	Min samples split	(2, 10)	
	Min samples leaf	(1, 10)	
	Bootstrap	True, False	
	Criterion	Gini, Entropy	
	Max leaf nodes	(50, 200)	
Extreme gradient boosting	Number of estimators	(50, 200)	Pan et al. (2022)
	Learning rate	0.001, 0.01, 0.001, 0.3	
	Max depth	(2, 20)	
	Gamma	0, 0.1, 0.2, 0.3, 0.4, 0.5	
	Colsample bytree	1, 0.3, 0.5,0.7	
	Alpha	0, 0.01, 0.1, 1	
	Lambda	0, 0.01, 0.1, 1	
	Sub sample	0.5, 0.7, 1	
K nearest neighbor	Number of neighbors	(3, 20)	Halder et al. (2024)
	Weights	Uniform, distance	
	Metric	Minkowski, Euclidean, Manhattan	
	Algorithm	Auto, Ball_tree, Kd_tree, Brute	
Naïve Bayes	Smoothing	np.logspace(−10,0, num = 50)	Raza et al. (2024)

separately and defined in the following), and is a popular metric in imbalanced classification, especially when accurate prediction of heatwave occurrence (Class 1) is important (Zhang et al. 2017). Precision measures the accuracy of predictions, indicating the proportion of correctly predicted positive examples out of all predicted positives. Recall, on the other hand, assesses the model's ability to identify relevant class examples, measuring the proportion of true positive examples correctly predicted out of all actual positives. These metrics are less sensitive to changes in data distributions compared to accuracy (Bekkar and Djemaa 2013). Balanced accuracy is the average of the true positive rate (recall) and the true negative rate, providing a balanced representation of correctness across all considered classes (Bekkar and Djemaa 2013).

These evaluation metrics are defined as follows (Luo et al. 2019):

$$F1-\text{score}=\frac{2*\frac{TP}{TP+FP}*\frac{TP}{TP+FN}}{\frac{TP}{TP+FP}+\frac{TP}{TP+FN}} \quad (1)$$

$$\text{Precision}=\frac{TP}{TP+FP} \quad (2)$$

$$\text{Recall}=\frac{TP}{TP+FN} \quad (3)$$

$$\text{Balanced accuracy}=\frac{1}{2}\left(\frac{TP}{TP+FN}+\frac{TN}{FP+TN}\right) \quad (4)$$

where TP (true positives) represents the number of positive predictions that are correctly predicted, TN (true negatives) is the number of negative predictions that are correctly predicted, FP is the number of false positive predictions, and FN is the number of false negative predictions.

2.2.16 Model interpretation

Machine-learning models are often considered black boxes and are increasingly being used for high-stakes prediction applications to support decision-making processes (Rudin 2019). Therefore, efforts and approaches to interpreting model outputs physically are as crucial as improving predictability in order to provide scientific insights, enable model improvements and improve scientific understanding of the modelling process (Lundberg and Lee 2017). To this end, SHapley Additive exPlanations (SHAP), a state-of-the-art ML model interpretation technique based on cooperative game theory (Angelov et al. 2021), was employed. SHAP calculates the model output impact of each explanatory-predictive feature as if the different features were players in a coalition game, and the payoff of the features, referred to as the Shapley value, is a measure of their importance (Angelov et al. 2021). SHAP is a strong and insightful way to physically interpret the results of complex ML algorithms (e.g. Althoff and Destouni 2023) and offers valuable perspectives on underlying dynamics that govern heatwave patterns.

3 Results

3.1 Feature selection

A correlation matrix heatmap illustrating the associations between the 19 possible explanatory-predictive features considered initially in this study (as outlined in Table 1) is shown in Fig. 4. Three highly correlated pairs of features were identified ($R>0.8$): evaporation and surface latent heat flux, specific humidity and monthly mean temperature,

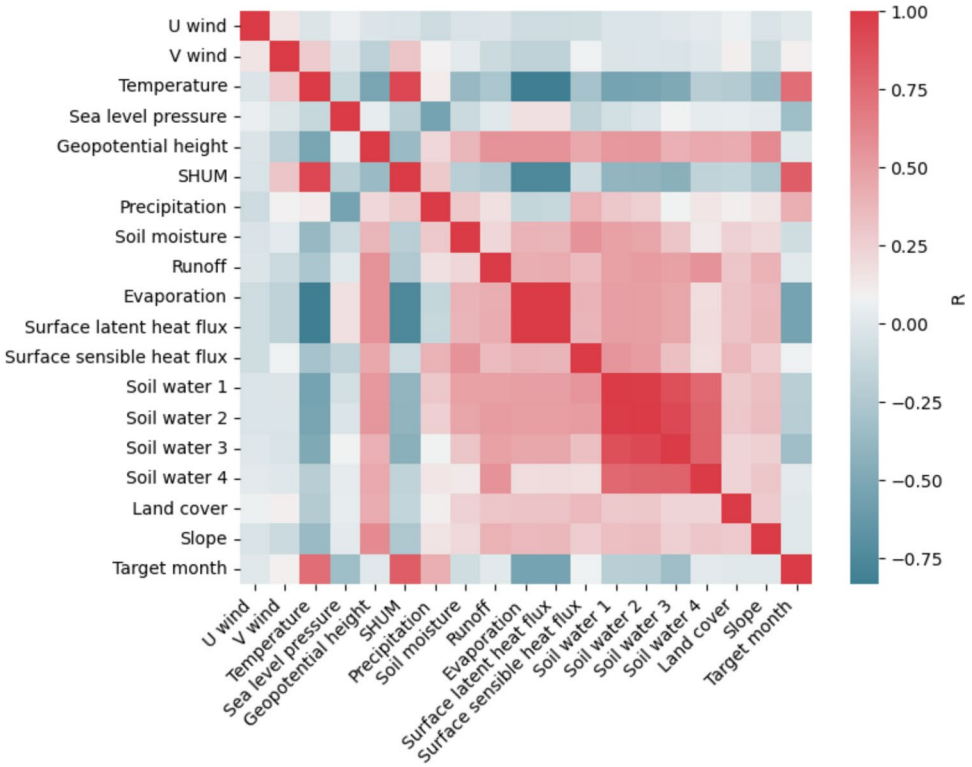

Fig. 4 Correlation matrix heatmap of possible relevant explanatory-predictive features for heatwave prediction investigated in this study. Red indicates features with strong positive linear correlations, white indicates no linear correlation, and blue indicates negative linear correlations. See Table 1 for feature descriptions

and soil water content at different depths (1, 2, 3 4). These pairs showed a similar pattern, therefore only one of each pair was retained (evaporation, monthly mean temperature, and soil water 1). The selection of features for further ML modelling was guided by domain knowledge. Evaporation was retained over surface latent heat flux due to their strong correlation and the more prevalent application of evaporation in hydrological models (Devia et al. 2015). While both evaporation and surface latent heat flux represent the transition of water from liquid to vapor in terms of mass and energy, respectively (Wild and Liepert 2010), evaporation is typically prioritized in modelling efforts because of its greater data accessibility and interpretability, particularly concerning processes such as heat accumulation (Hunt et al. 2002; Miralles et al. 2014, 2019).Temperature was chosen over humidity as it serves as a direct indicator of heatwaves, which are central to this analysis. Soil water at the surface level was selected from the various deeper levels because of its more direct interaction with the atmosphere. The remaining 13 features were then utilised in the modelling phase of the study.

3.2 Model performance

The performance of the ML models in the testing years is shown in Fig. 5. All tested models outperform the baseline probability model in F1 score and recall for heatwave occurrence (class 1) and precision for non-occurrence of heatwaves (class 0). The BRF model consistently demonstrates the highest and most stable performance in terms of balanced accuracy and F1score across all lead times. It also exhibits a declining trend in recall of heatwave occurrence (class 1), going from 0.87 for the shortest (one month) to 0.81 for the longest (five months) lead time. The standard deviations of the evaluation metrics (error bars) increase notably with longer lead times for BRF models (i.e., balanced accuracy goes from 0.06 to 0.09, F1 for class 1 from 0.23 to 0.27, and F1 for class 0 from 0.03 to 0.05 with increased lead time).

The XGBoost model performs second-best for the lead time of one month and exhibits relatively high class 1 recall performance, which improves as the lead time increases. However, its class 0 recall performance declines significantly with longer lead times, indicating a substantial failure to predict the non-occurrence of heatwaves at extended lead times. The performance of the KNN, LR, and NB models varies across different lead times. The LR model has its best performance at one-month lead time in terms of balanced accuracy (0.72) and class 1 F1 score (0.27), with its performance generally declining as the lead time increases (i.e., at a five-month lead time balanced accuracy and class 1 F1 score are 0.66 and 0.24, respectively). LR shows comparatively low performance in class 1 recall and class 0 precision. The KNN model ranks second-best for the lead times of two, four, and five months, particularly in terms of balanced accuracy, although its class 1 recall exhibits

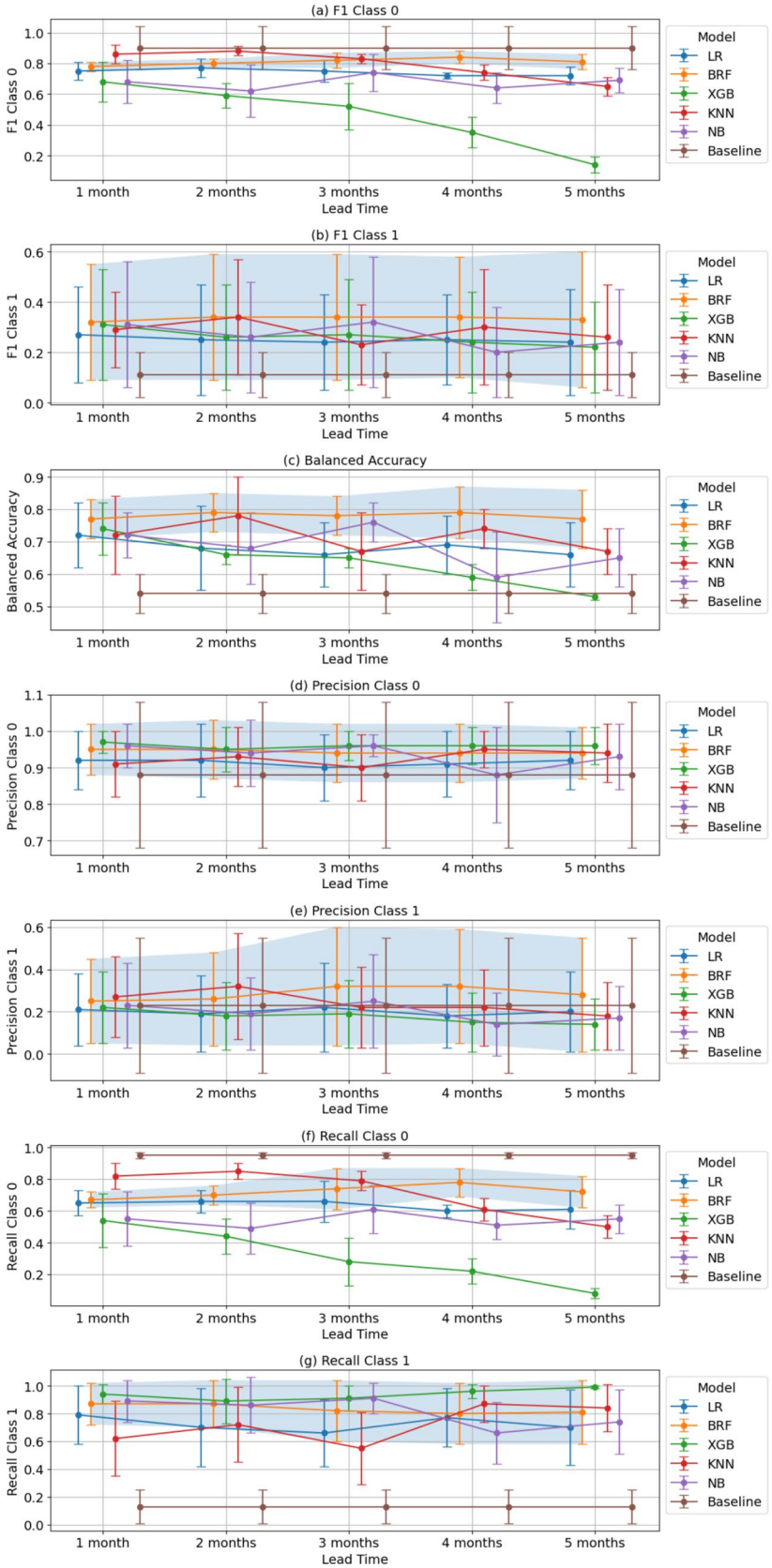

Fig. 5 Error bar plot of evaluation metrics (**a**) F1 score for Class 0; (**b**) F1 score for Class 1; (**c**) balanced accuracy; (**d**) precision for Class 0; (**e**) precision for Class 1; (**f**) recall for Class 0 (**g**) recall for Class 1 of the five machine-learning models in the prediction of heatwave in Sweden with lead times of one to five months. The error bar indicates the estimated forecast uncertainty via the standard deviation. Shadow area shows the one standard deviation range for BRF model

significant variability. The NB model performs second-best at three months lead time but shows a general decline trend in performance as the lead time increases.

Figure 6 presents the spatial performance of the one-month lead time BRF model in terms of the evaluation metrics considered during the testing years. In southern Sweden, the model exhibits lower F1 score and recall for class 0 and balanced accuracy, while showing higher F1 score, precision, and recall for class 1. Conversely, northern Sweden displays relatively lower performance for class 1 metrics (particularly F1 score and recall) and higher performance for class 0. The dark blue areas on the F1 score, precision, and recall maps indicate regions where one class is absent in either the true observed or predicted data, resulting in zero-valued metrics.

3.3 Feature importance

Figure 7 shows the feature importance ranking and impact value obtained using the SHAP approach for the best-performing BRF models across the studied lead times of one to five months. Geopotential height emerges as a key explanatory-predictive feature for HWD occurrence, consistently ranking in the top two for the BRF model at different lead times. Temperature, evaporation, and related variables have notable high impacts on model output at lead times of one and two months. For longer lead times of three, four, and five months, the importance of these variables slightly decreases but they remain influential. In contrast, the importance of landscape explanatory variables such as runoff and soil water, generally increases with increasing lead time. Precipitation has an overall moderate influence without a clear trend with changing lead time. Different physical drivers influence the formation of heatwaves at various timescales (Domeisen et al. 2023). At shorter lead times, atmospheric features associated with synoptic systems dominate, as they are highly sensitive to changes in radiation and temperature (Perkins 2015; Miralles et al. 2019; Mascolo et al. 2025). In contrast, landscape features play a key role in initiating heatwaves at sub-seasonal to seasonal lead times, driven by slow-evolving land surface processes (Perkins 2015; Dirmeyer et al. 2021). Empirical studies have shown that dry winters and springs can induce summer heatwaves in Europe (Quesada et al. 2012), while a precipitation deficit generally contributes to summer heatwaves globally (Mueller and Seneviratne 2012). However, precipitation anomalies are more influential in southern Europe compared to northern Europe (Quesada et al. 2012), which aligns with the relatively lower importance observed in the study."

Figure 7 also shows the variability and distribution of SHAP values for the features in the BRF model. The impact ranges for the features narrow and become more concentrated around zero, implying weaker influence on model outputs, for longer lead times. Overall, heatwave formation tends to be higher for lower values of geopotential height, precipitation, and slope, and for higher values of temperature, mean sea level pressure, and runoff.

3.4 Example of application for heatwave in July 2019

Heatwave occurrence prediction maps were generated for July 2019 across Sweden utilising the best-performing BRF model for each lead time in a series from longer to shorter lead times (Fig. 8). A large-scale heatwave occurred in Sweden in July 2019, and the present predictions were made as one specific test example of the potential of the ML-based methodology framework developed in this paper to support early-warning systems. The maps in Fig. 7 present the observed heatwave occurrence (Fig. 8a) alongside the best ML model predictions at lead times from five months to one month before the occurrence (Fig. 8b–f). The number of correctly classified pixels for the various lead times, from five months to one month, are 807, 911, 913, 919, and 930,

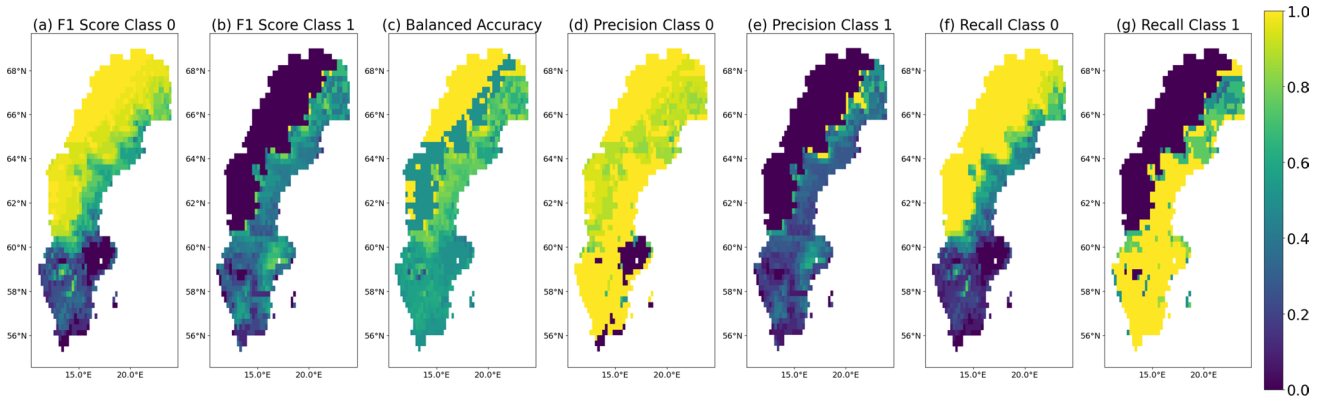

Fig. 6 Spatial plot of evaluation metrics: (**a**) F1 score for Class 0; (**b**) F1 score for Class 1; (**c**) balanced accuracy; (**d**) precision for Class 0; (**e**) precision for Class 1; (**f**) recall for Class 0; (**g**) recall for Class 1 of the one-month lead time balanced random forest model in predicting heatwaves in Sweden

respectively, out of a total of 1294 pixels across Sweden. These stable results for the BRF model are consistent with its reliable performance in terms of balanced accuracy.

4 Discussion

This study presents a ML methodology framework to comparatively develop, identify, and extract insights from the suitable models for predicting heatwave occurrences with seasonal lead times, leveraging global remote sensing datasets. The methodology framework could be adapted to other regions using global datasets on GEE platforms and open-source ML libraries. This study's results demonstrate that ML models can provide early signs of forthcoming heatwaves up to five months in advance. The BRF consistently emerged as the top-performing model due to its superior balanced accuracy of around 0.77, demonstrating its ability to handle imbalanced datasets. Most tested models outperformed the comparative baseline model that predicted heatwaves by estimating the probability of occurrence based on approximately 30 years of historical data. However, it is difficult to compare the prediction performance of the present study with previous studies due to different definitions of heatwaves, used data, and target variables (e.g., annual total number of heatwave days, binary weekly heatwave occurrence, daily maximum temperature). Ultimately, the present study results may add to such contributions toward further strategic deployment of ML models in climate adaptation efforts, offering a methodical approach to enhancing predictive accuracy and operational readiness for heatwaves, thereby reducing their socio-economic impacts.

With overall consistency, the BRF model outperformed all other tested models across the different studied lead times (of one to five months) with stable balanced accuracy. The associated increase in standard deviations for longer lead times implies that the heatwave predictions of this model might occasionally still be rather inaccurate at longer lead times; to some degree, this may also be due to uncertainties in the underlying data time series. The strong performance of the BRF model is likely attributed to its random forest foundation, which has been well-validated and widely applied in various heat (Giamalaki et al. 2022; Suthar et al. 2023) and other natural hazard-related studies (Zennaro et al. 2021). The consistent superior performance of this model should encourage further explorations of its utility for prediction of natural hazards and particularly extreme events that commonly involve issues of imbalanced data. Spatial evaluation metrics maps (Fig. 6) indicate that the BRF model demonstrates a greater accuracy in predicting non-heatwave conditions in northern Sweden; however, it encounters difficulties in forecasting heatwave occurrences in this region, potentially attributable to the infrequent occurrence of heatwave events. In southern Sweden, the BRF model achieves higher Class 1 F1 scores and recall, although certain areas, such as Stockholm and Skåne County, exhibit a tendency toward more frequent heatwave events. The BRF model is constructed using data from all counties in Sweden. Developing separate models for the northern and southern regions could provide valuable insights in relation to the findings of this study and potentially further enhance the performance of heatwave predictions.

While baseline probability model effectively predicted non-heatwave conditions (demonstrating high recall for class 0), it exhibited notably low recall for heatwave occurrence (class 1). This highlights the limitation of this simple extrapolation model, which relies solely on historical frequencies and does not account for additional explanatory features. All other tested models outperformed the baseline model predictions, except for XGBoost at a lead time of five months. The XGBoost model exhibits high performance at one-month lead time, capturing most heatwave occurrences with high class 1 recall and balanced accuracy. However, its performance decreases notably with increased lead time, as seen by its low F1 score for class 0 and balanced accuracy (Fig. 5a,c). This declining performance trend suggests that XGBoost is biased to class 1, likely due to the data resampling process. The KNN model demonstrates high flexibility in handling various data distributions but is also more sensitive to these distributions. The variability in heatwave class distributions over the years makes it difficult for KNN to consistently determine the optimal number of neighbours (Zhang 2022), leading to its fluctuating performance.

The intuitive land–atmosphere feedback mechanism, where dry land reduces land evaporation, leading to drier air conditions, suggests an associated decrease in precipitation that could potentially trigger meteorological drought. As evaporation decreases, the latent heat flux also diminishes, while the sensible heat flux increases, potentially driving the onset of a heatwave (Miralles et al. 2019). The current SHAP value analysis identifies geopotential height as a consistent primary contributor to heatwave formation, with temperature and evaporation also emerging as significant atmospheric and landscape features, respectively. Other landscape features, such as runoff and soil water, gain importance in model outputs at longer lead times, likely reflecting the relatively stable cumulative effects of landscape conditions over time. Overall, feature importance values decrease for longer lead times, reflecting the complexity of predicting heatwave occurrence over extended time periods. For models like XGBoost and KNN, showing varying performance for different lead times, improved understanding and accounting for the physical relationship complexity could guide effective adjustments in model configuration to

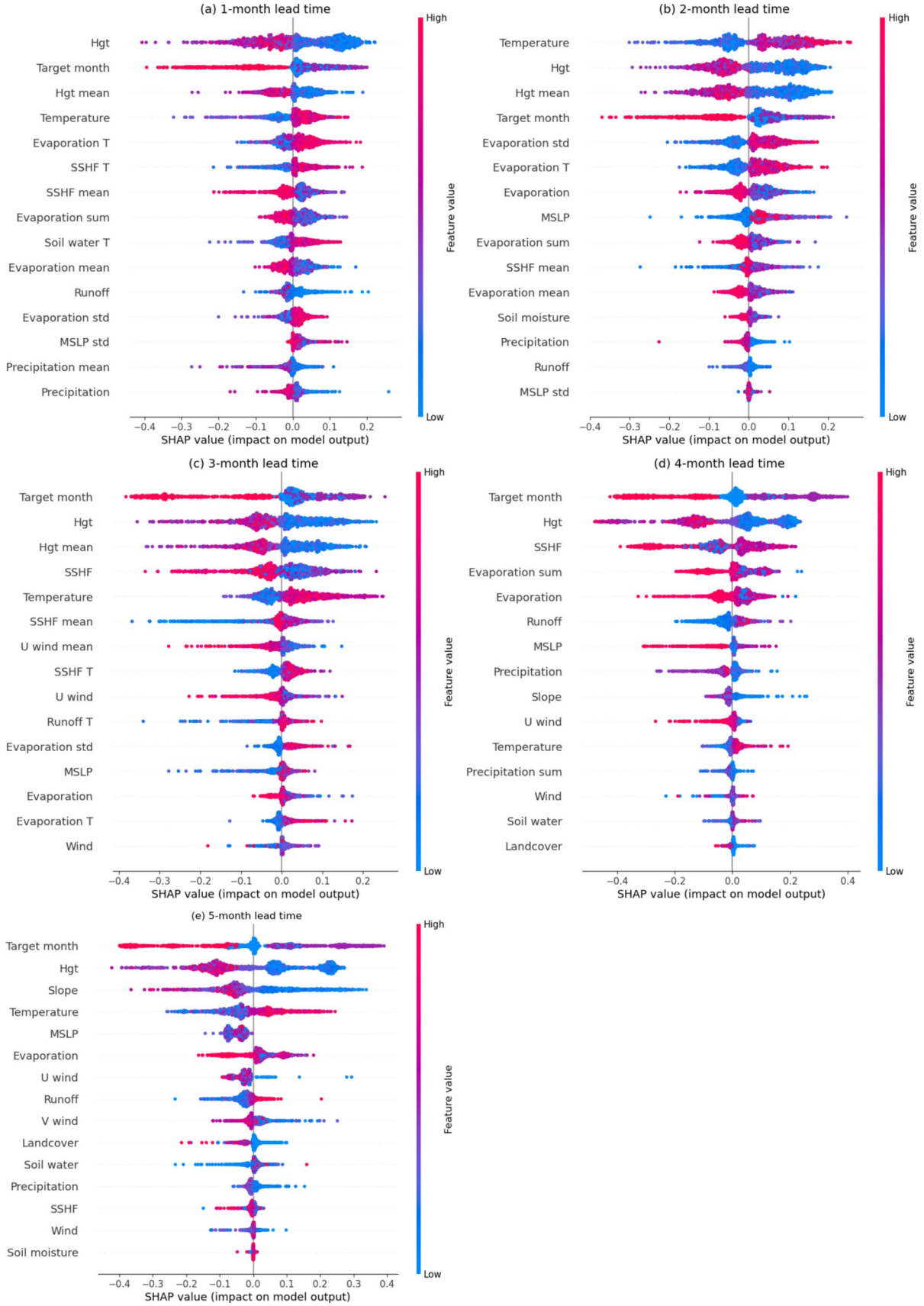
(a) 1-month lead time
Hgt
Target month
Hgt mean
Temperature
Evaporation T
SSHF T
SSHF mean
Evaporation sum
Soil water T
Evaporation mean
Runoff
Evaporation std
MSLP std
Precipitation mean
Precipitation
SHAP value (impact on model output)
Feature value
High
Low
(b) 2-month lead time
Temperature
Hgt
Hgt mean
Target month
Evaporation std
Evaporation T
Evaporation
MSLP
Evaporation sum
SSHF mean
Evaporation mean
Soil moisture
Precipitation
Runoff
MSLP std
SHAP value (impact on model output)
Feature value
High
Low
(c) 3-month lead time
Target month
Hgt
Hgt mean
SSHF
Temperature
SSHF mean
U wind mean
SSHF T
U wind
Runoff T
Evaporation std
MSLP
Evaporation
Evaporation T
Wind
SHAP value (impact on model output)
Feature value
High
Low
(d) 4-month lead time
Target month
Hgt
SSHF
Evaporation sum
Evaporation
Runoff
MSLP
Precipitation
Slope
U wind
Temperature
Precipitation sum
Wind
Soil water
Landcover
SHAP value (impact on model output)
Feature value
High
Low
(e) 5-month lead time
Target month
Hgt
Slope
Temperature
MSLP
Evaporation
U wind
Runoff
V wind
Landcover
Soil water
Precipitation
SSHF
Wind
Soil moisture
SHAP value (impact on model output)
Feature value
High
Low

Fig. 7 SHapley Additive exPlanations (SHAP) values for each feature in the balanced random forest model in prediction with lead times of one to five months, where blue represents lower and red represents higher feature values, and positive and negative SHAP values (x-axis) imply a higher and lower probability of heatwave occurrence, respectively. Hgt: geopotential height; SSHF: Sensible heat flux; MSLP: mean sea level pressure; T: trend

better capture the occurrence dynamics of heatwaves. For instance, adjusting the weight given to geopotential height and/or landscape features might improve the performances of these models for longer lead times.

From a practical point of view, these insights not only improve understanding of heatwave formation, but also facilitate the structured preparations for such events, guiding interventions well before the heatwave occurs. By identifying key predictive features, such as geopotential height, along with the significance of landscape characteristics, these models will enable disaster management agencies to develop a tiered, proactive response strategy. For instance, with the example shown in Fig. 8, agencies can predict and monitor a large-scale heatwave four to five months ahead. Between one and three months ahead, agencies can launch broad-scale awareness campaigns with increased certainty, optimise resource allocation, and implement early-stage mitigation strategies for the spatial coverage of the heatwave. These strategies could include adjustments in public transportation schedules to reduce heat exposure and the establishment of specific water-management protocols to handle increased demand during peak heat periods. In the long term, these results can inform future spatial planning, facilitating the design of houses and other buildings that are better equipped to withstand both heat and cold. For example, nature-based solutions (NbS) such as incorporating shaded areas in the summer while allowing sunlight during the winter can be effectively integrated into residential layouts (Pan et al. 2021; Barnett and Bouw 2022). Additionally, including NbS elements like pathways, water bodies, and various types of vegetated areas can further enhance resilience (Sahani et al. 2023; Ibsen et al. 2024). As the lead time shortens, the focus can shift towards fine-tuning emergency response plans, mobilising community support mechanisms, and deploying targeted health services to vulnerable populations.

Machine-learning methodologies are progressively employed to enhance the accuracy of heatwave prediction, extending their ability beyond current predictability limits (Domeisen et al. 2023). However, it is also important to acknowledge the inherent ML limitations. These models rely heavily on the quality and availability of training data and are unable to generate predictions beyond the patterns captured in the provided data (L'Heureux et al. 2017). This makes it particularly challenging to predict extreme events due to the common spatiotemporal gaps in data and the imbalanced distribution of heatwave occurrence versus non-occurrence data. Such imbalances often bias models toward the majority class (non-occurrence of heatwaves), necessitating adjustments to address this issue. Careful model calibration and evaluation are thus crucial for good model predictions, which are essential for relevant issuing of heatwave warnings and for the efficacy of heat-health action plans (Kotharkar and Ghosh 2022).

Heatwave warning systems aim to mitigate the adverse impacts of heatwaves and enhance communication among stakeholders. According to the World Health Organization's heat-health action plan guidelines, the adverse impacts of heatwaves can be significantly reduced through well-coordinated actions at multiple levels (Matthies et al. 2008). These measures include accurate and timely alert systems to provide timely public advisories and a heat-related health information plan about what, with whom, and when to communicate (Matthies et al. 2008). In this study, the models were developed to predict the occurrence of heatwaves. For such model to be utilised practically, e.g., in a multi-hazard early-warning system, follow-up research is further needed, e.g., on: (i) an impact-based prediction modelling and (ii) action road maps.

In an impact-based prediction model, the consequences of the heatwaves are not only related to the hazard itself, but also to the characteristics of the population in the affected area. For example, people with pre-existing health problems, socially isolated elderly people with fragile health conditions, young children, people suffering from obesity, and cardiovascular diseases (Liu et al. 2022) are particularly vulnerable to heatwaves. As an increase in the aging population is expected (SCB 2022), heatwaves are projected to have more severe impacts in the future, not only due to climate change but also as a result of rising urbanization (Hu et al. 2024). It is crucial for authorities to identify and locate vulnerable populations and areas in need of medical intervention during a heatwave within urban environments. By classifying the impact level of heatwaves—such as high-impact or low-impact—and integrating this information into the model used in this study, the accuracy of predicted risks can be significantly enhanced.

In action road maps, communication science is extremely important. Roadmaps can be created to improve societal and organisational preparedness, prevent negative consequences, and ensure the timely response to heatwave-related events and crises in Sweden. The roadmaps could be structured based on systemic mapping of best management practices in the EU and other parts of the world, their adaptability to the Swedish context, and relevant policies and regulations to support societal security in Sweden.

This study has limitations that require improvements and need to be addressed in future research. First, to align with

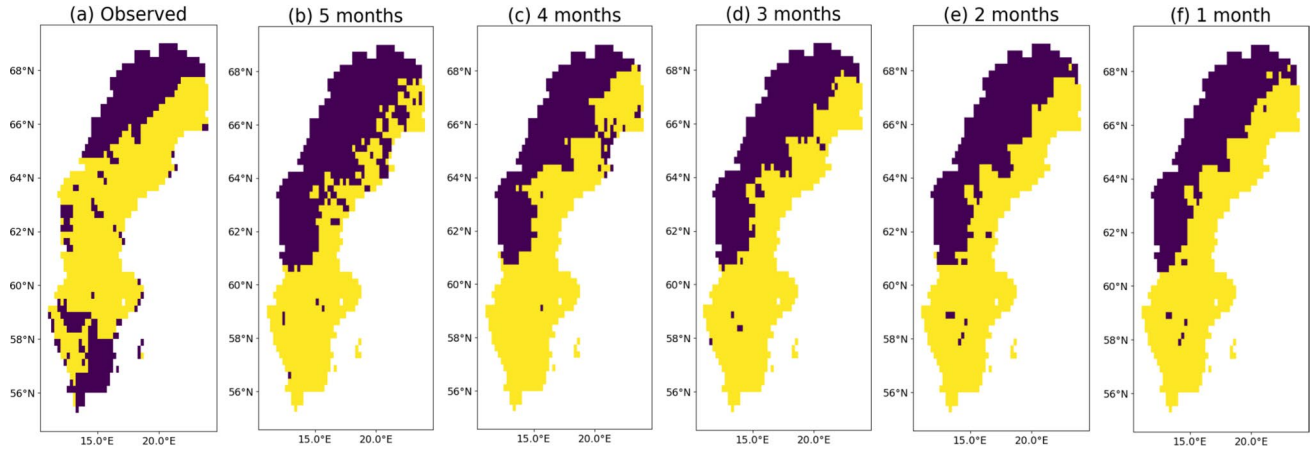

Fig. 8 Example of heatwave occurrence observed in July 2019 (**a**) and associated predictions at one to five months (**b** to **f**) lead time based on BRF model. The yellow area indicates the presence of a heatwave, while the purple area indicates regions without a heatwave

the current national heatwave warning system and simplify the modelling process, the threshold heatwave temperature in this study was set to the same value (27 °C) throughout the whole Sweden. However, a single threshold value may not be suitable across all parts of Sweden, as heatwave impacts may differ considerably between northern and southern regions. Second, besides temperature as a heatwave indicator, a heat index that combines temperature and humidity could also better represent heatwave impacts and should be further considered in future modelling efforts (Xu et al. 2016). Third, the selected explanatory-predictive features investigated in this study were based on a literature review on ML-based heatwave prediction approaches and were limited by data availability in GEE. Future research could integrate other relevant features and test for importance in seasonal heatwave prediction, e.g. cloud cover and the amount of snowfall in winter (Hansen et al. 2014; Dirmeyer et al. 2021). Fourth, the present ML modelling was built on reanalysis data, which are not real-time accessible. Future work should attempt to integrate real-time data and evaluate the operational model readiness for early warning applications. To adapt the modelling for operational forecasting purposes, real-time data– such as those from local weather stations and satellite images—should supplement or replace reanalysis and/or other combined model-observation data combinations. This is particularly important because global reanalysis products have been shown to considerably diverge from other relevant datasets, potentially misrepresenting many water-related landscape variables (Zarei and Destouni 2024) that this study has identified as essential features for heatwave prediction. The real-time data should be for the same or analogous explanatory features as the reanalysis data used in this study, to enable transfer learning techniques, such as translations of the distributions of features and transformations of feature representations (Segev et al. 2017). Shifting to real-time data introduces new challenges, such as increased noise, but it has the potential to improve model performance with more precise data, provided the model is retrained and recalibrated with adjusted hyperparameters to account for the data updates.

5 Conclusions

This study bridges a research gap between physically based models that commonly enable relatively short-term heatwave prediction, and ML models for possible considerably longer-term prediction and associated earlier warning. The task of early heatwave prediction, i.e., with a seasonal lead time, is challenging due to the complex interactions between different types of variables (atmospheric, landscape, and others). This study demonstrates the potential ability of the present multi-model ML methodology framework, using open-source remote sensing data from GEE, to enhance the prediction of heatwaves with seasonal lead times and be implemented in, e.g., a multi-hazard early-warning system. Among the five ML models investigated, BRF emerged as best suited for this purpose with persistent high performance and ability to handle imbalanced data. Further research should aim at integrating real-time data and additionally also impact perspectives in the ML modelling for practical early-warning and risk assessment purposes.

The novelty of the present study lies in its development of and comparison framework for five different ML models, and the concrete application example of how best-performing models could potentially facilitate better heatwave preparedness and mitigation strategies. The results further show the possibility of good long-term predictive power for the best-performing ML model at lead times from one to five months, while the feature importance assessment and interpretation using the SHAP approach show the explanatory basis for this performance. Overall, the present results

demonstrate that machine learning approaches have the potential to provide early signs and warnings of impending heatwave occurrences. These could eventually be integrated into an early warning system with impact assessments to address the associated societal and environmental challenges, aiding future response plans and actions.

Author contribution J.C.K.: Formal analysis, Methodology, Writing– original draft. M.V.P.: Methodology, Writing– review & editing. G.D.: Writing– review & editing. K.B.: Writing– review & editing. C. S.S.F.: Validation, Writing– review & editing. Z.K.: Conceptualization, Supervision, Funding acquisition, Writing– review & editing. All authors reviewed the manuscript.

Funding Open access funding provided by Royal Institute of Technology.
We gratefully acknowledge funding from the Swedish Research Council (VR) through the projects: Science for a secure society: Hydroclimatic hazard, risk, and crisis management in Sweden (CrisAct) (grant 2021-06309); and Coupled freshwater system variations, trends and their drivers around the world (grant 2022-04672). Carla Ferreira was supported by the Portuguese Foundation for Science and Technology, through the institutional scientific employment program-contract (CEECINST/00077/2021). This study arose from a collaborative Master of Science dissertation by Li (2023) entitled 'Monthly Heatwave Prediction in Sweden Based on Machine Learning Techniques with Remote Sensing Data'.

Data availability All data were accessed from Google Earth Engine, a publicly available cloud-based platform, via (https://developers.google.com/earth-engine/datasets/).

Declarations

Conflict of interest The authors declare no competing interests.

References

Alam TM, Shaukat K, Hameed IA et al (2020) An investigation of credit card default prediction in the imbalanced datasets. IEEE Access 8:201173–201198. https://doi.org/10.1109/ACCESS.2020.3033784

Al-Juaidi AEM, Nassar AM, Al-Juaidi OEM (2018) Evaluation of flood susceptibility mapping using logistic regression and GIS conditioning factors. Arab J Geosci 11:765. https://doi.org/10.1007/s12517-018-4095-0

Althoff D, Destouni G (2023) Global patterns in water flux partitioning: irrigated and rainfed agriculture drives asymmetrical flux to vegetation over runoff. One Earth 6:1246–1257. https://doi.org/10.1016/j.oneear.2023.08.002

Angelov PP, Soares EA, Jiang R et al (2021) Explainable artificial intelligence: an analytical review. Wires Data Min Knowl Discov 11:e1424. https://doi.org/10.1002/widm.1424

Asadollah SBHS, Khan N, Sharafati A et al (2022) Prediction of heat waves using meteorological variables in diverse regions of Iran with advanced machine learning models. Stoch Environ Res Risk Assess 36:1959–1974. https://doi.org/10.1007/s00477-021-02103-z

Barnett J, Bouw M (2022) Managing the climate crisis: designing and building for floods, heat, drought, and wildfire. Island Press

Barriopedro D, García-Herrera R, Ordóñez C et al (2023) Heat waves: physical understanding and scientific challenges. Rev Geophys 61:e2022RG000780. https://doi.org/10.1029/2022RG000780

Bartosik A, Whittingham H (2021) Chapter 7—Evaluating safety and toxicity. In: Ashenden SK (ed) The Era of artificial intelligence, machine learning, and data science in the pharmaceutical industry. Academic Press, pp 119–137

Bekkar M, Djemaa DHK (2013) Evaluation measures for models assessment over imbalanced data sets

Bergstra J, Bengio Y (2012) Random search for hyper-parameter optimization

Chen T, Guestrin C (2016) XGBoost: a scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Association for Computing Machinery, New York, NY, USA, pp 785–794

Devia GK, Ganasri BP, Dwarakish GS (2015) A review on hydrological models. Aquat Procedia 4:1001–1007. https://doi.org/10.1016/j.aqpro.2015.02.126

Dirmeyer PA, Balsamo G, Blyth EM et al (2021) Land-atmosphere interactions exacerbated the drought and heatwave over northern europe during summer 2018. AGU Adv 2:e2020AV000283. https://doi.org/10.1029/2020AV000283

Domeisen DIV, Eltahir EAB, Fischer EM et al (2023) Prediction and projection of heatwaves. Nat Rev Earth Environ 4:36–50. https://doi.org/10.1038/s43017-022-00371-z

Drouard M, Kornhuber K, Woollings T (2019) Disentangling dynamic contributions to summer 2018 anomalous weather over Europe. Geophys Res Lett 46:12537–12546. https://doi.org/10.1029/2019GL084601

Fonseca-Rodríguez O, Adams RE, Sheridan SC, Schumann B (2023) Projection of extreme heat- and cold-related mortality in Sweden based on the spatial synoptic classification. Environ Res 239:117359. https://doi.org/10.1016/j.envres.2023.117359

Ford T, Dirmeyer P, Benson D (2018) Evaluation of heat wave forecasts seamlessly across subseasonal timescales. Npj Clim Atmospheric Sci. https://doi.org/10.1038/s41612-018-0027-7

Giamalaki K, Beaulieu C, Prochaska JX (2022) Assessing predictability of marine heatwaves with random forests. Geophys Res Lett 49:e2022GL099069. https://doi.org/10.1029/2022GL099069

Gorelick N, Hancher M, Dixon M et al (2017) Google earth engine: planetary-scale geospatial analysis for everyone. Remote Sens Environ 202:18–27. https://doi.org/10.1016/j.rse.2017.06.031

Guo G, Wang H, Bell D et al (2003) KNN model-based approach in classification. In: Meersman R, Tari Z, Schmidt DC (eds) On the move to meaningful internet systems 2003: CoopIS, DOA, and ODBASE. Springer, Berlin, Heidelberg, pp 986–996

Gyaneshwar A, Mishra A, Chadha U et al (2023) A contemporary review on deep learning models for drought prediction. Sustainability 15:6160. https://doi.org/10.3390/su15076160

Halder RK, Uddin MN, Uddin MdA et al (2024) Enhancing K-nearest neighbor algorithm: a comprehensive review and performance

analysis of modifications. J Big Data 11:113. https://doi.org/10.1186/s40537-024-00973-y
Hansen BB, Isaksen K, Benestad RE et al (2014) Warmer and wetter winters: characteristics and implications of an extreme weather event in the High Arctic. Environ Res Lett 9:114021. https://doi.org/10.1088/1748-9326/9/11/114021
Hu X, Cao J, Qian Y et al (2024) Extreme heat events in mainland China from 1981 to 2015: spatial patterns, temporal trends, and urbanization impacts. Sustain Cities Soc 100:104999. https://doi.org/10.1016/j.scs.2023.104999
Hunt JE, Kelliher FM, McSeveny TM, Byers JN (2002) Evaporation and carbon dioxide exchange between the atmosphere and a tussock grassland during a summer drought. Agric for Meteorol 111:65–82. https://doi.org/10.1016/S0168-1923(02)00006-0
Ibsen PC, Crawford BR, Corro LM et al (2024) Urban tree cover provides consistent mitigation of extreme heat in arid but not humid cities. Sustain Cities Soc 113:105677. https://doi.org/10.1016/j.scs.2024.105677
Jacques-Dumas V, Ragone F, Borgnat P et al (2022) Deep learning-based extreme heatwave forecast. Front Clim. https://doi.org/10.3389/fclim.2022.789641
Kannan S, Subbaram K, Faiyazuddin Md (2023) Chapter 17—Artificial intelligence in vaccine development: significance and challenges ahead. In: Philip A, Shahiwala A, Rashid M, Faiyazuddin Md (eds) A Handbook of Artificial Intelligence in Drug Delivery. Academic Press, pp 467–486
Ke D, Takahashi K, Takakura J et al (2023) Effects of heatwave features on machine-learning-based heat-related ambulance calls prediction models in Japan. Sci Total Environ 873:162283. https://doi.org/10.1016/j.scitotenv.2023.162283
Khalid S, Khalil T, Nasreen S (2014) A survey of feature selection and feature extraction techniques in machine learning. In: 2014 Science and Information Conference. pp 372–378
Khan N, Shahid S, Ismail TB, Behlil F (2021) Prediction of heat waves over Pakistan using support vector machine algorithm in the context of climate change. Stoch Environ Res Risk Assess 35:1335–1353. https://doi.org/10.1007/s00477-020-01963-1
Kong J, Kowalczyk W, Nguyen DA, et al (2019) Hyperparameter optimisation for improving classification under class imbalance. In: 2019 IEEE Symposium Series on Computational Intelligence (SSCI). pp 3072–3078
Kotharkar R, Ghosh A (2022) Progress in extreme heat management and warning systems: a systematic review of heat-health action plans (1995–2020). Sustain Cities Soc 76:103487. https://doi.org/10.1016/j.scs.2021.103487
L'Heureux A, Grolinger K, Elyamany HF, Capretz MAM (2017) Machine learning with big data: challenges and approaches. IEEE Access 5:7776–7797. https://doi.org/10.1109/ACCESS.2017.2696365
Li Z (2023) Monthly heatwave prediction in Sweden based on Machine Learning techniques with remote sensing data
Lin W-C, Tsai C-F, Hu Y-H, Jhang J-S (2017) Clustering-based undersampling in class-imbalanced data. Inf Sci 409–410:17–26. https://doi.org/10.1016/j.ins.2017.05.008
Liu J, Varghese BM, Hansen A et al (2022) Heat exposure and cardiovascular health outcomes: a systematic review and meta-analysis. Lancet Planet Health 6:e484–e495. https://doi.org/10.1016/S2542-5196(22)00117-6
Lorenz EN (1963) Deterministic nonperiodic flow
Lowe R, García-Díez M, Ballester J, Creswick J, Robine J-M, Herrmann F, Rodó X (2016) Evaluation of an early-warning system for heat wave-related mortality in europe: implications for sub-seasonal to seasonal forecasting and climate services. Int J Environ Res Public Health 13(2):206. https://doi.org/10.3390/ijerph13020206
Lundberg SM, Lee S-I (2017) A unified approach to interpreting model predictions. In: Advances in Neural Information Processing Systems. Curran Associates, Inc.
Luo H, Pan X, Wang Q, et al (2019) Logistic regression and random forest for effective imbalanced classification. In: 2019 IEEE 43rd Annual Computer Software and Applications Conference (COMPSAC). pp 916–917
Mascolo V, Le Priol C, D'Andrea F, Bouchet F (2025) Comparing the influence of Atlantic multidecadal variability and spring soil moisture on European summer heat waves. Oxf Open Clim Change 5:kgae023. https://doi.org/10.1093/oxfclm/kgae023
Matthies F, Bickler G, Marín NC, Hales S (2008) Heat-health Action Plans: Guidance. World Health Organization
Merz B, Kuhlicke C, Kunz M et al (2020) Impact forecasting to support emergency management of natural hazards. Rev Geophys 58:e2020RG000704. https://doi.org/10.1029/2020RG000704
Miralles DG, Teuling AJ, van Heerwaarden CC, Vilà-Guerau de Arellano J (2014) Mega-heatwave temperatures due to combined soil desiccation and atmospheric heat accumulation. Nat Geosci 7:345–349. https://doi.org/10.1038/ngeo2141
Miralles DG, Gentine P, Seneviratne SI, Teuling AJ (2019) Land–atmospheric feedbacks during droughts and heatwaves: state of the science and current challenges. Ann N Y Acad Sci 1436:19–35. https://doi.org/10.1111/nyas.13912
More AS, Rana DP (2017) Review of random forest classification techniques to resolve data imbalance. In: 2017 1st International Conference on Intelligent Systems and Information Management (ICISIM). pp 72–78
Mosavi A, Ozturk P, Chau K (2018) Flood prediction using machine learning models: literature review. Water 10:1536. https://doi.org/10.3390/w10111536
Mueller B, Seneviratne SI (2012) Hot days induced by precipitation deficits at the global scale. Proc Natl Acad Sci 109:12398–12403. https://doi.org/10.1073/pnas.1204330109
Niaz R, Zhang X, Iqbal N et al (2021) Logistic regression analysis for spatial patterns of drought persistence. Complexity 2021:e3724919. https://doi.org/10.1155/2021/3724919
Niggli L, Huggel C, Muccione V et al (2022) Towards improved understanding of cascading and interconnected risks from concurrent weather extremes: analysis of historical heat and drought extreme events. PLOS Clim 1:e0000057. https://doi.org/10.1371/journal.pclm.0000057
Olsson J, Arheimer B, Borris M et al (2016) Hydrological climate change impact assessment at small and large scales: key messages from recent progress in Sweden. Climate 4:39. https://doi.org/10.3390/cli4030039
Oommen T, Baise LG, Vogel RM (2011) Sampling bias and class imbalance in maximum-likelihood logistic regression. Math Geosci 43:99–120. https://doi.org/10.1007/s11004-010-9311-8
Oudin Åström D, Åström C, Forsberg B et al (2020) Heat wave–related mortality in Sweden: a case-crossover study investigating effect modification by neighbourhood deprivation. Scand J Public Health 48:428–435. https://doi.org/10.1177/1403494818801615
Pan H, Page J, Cong C et al (2021) How ecosystems services drive urban growth: integrating nature-based solutions. Anthropocene 35:100297. https://doi.org/10.1016/j.ancene.2021.100297
Pan S, Zheng Z, Guo Z, Luo H (2022) An optimized XGBoost method for predicting reservoir porosity using petrophysical logs. J Pet Sci Eng 208:109520. https://doi.org/10.1016/j.petrol.2021.109520
Panahi M, Rahmati O, Kalantari Z et al (2022) Large-scale dynamic flood monitoring in an arid-zone floodplain using SAR data and hybrid machine-learning models. J Hydrol 611:128001. https://doi.org/10.1016/j.jhydrol.2022.128001
Perkins SE (2015) A review on the scientific understanding of heatwaves—Their measurement, driving mechanisms, and changes at

the global scale. Atmospheric Res 164–165:242–267. https://doi.org/10.1016/j.atmosres.2015.05.014
Perkins-Kirkpatrick SE, Lewis SC (2020) Increasing trends in regional heatwaves. Nat Commun 11:3357. https://doi.org/10.1038/s41467-020-16970-7
Qu Q, Xu J, Kang W et al (2023) Ensemble learning model identifies adaptation classification and turning points of river microbial communities in response to heatwaves. Glob Change Biol 29:6988–7000. https://doi.org/10.1111/gcb.16985
Quesada B, Vautard R, Yiou P et al (2012) Asymmetric European summer heat predictability from wet and dry southern winters and springs. Nat Clim Change 2:736–741. https://doi.org/10.1038/nclimate1536
Rahmati O, Falah F, Dayal KS et al (2020) Machine learning approaches for spatial modeling of agricultural droughts in the south-east region of Queensland Australia. Sci Total Environ 699:134230. https://doi.org/10.1016/j.scitotenv.2019.134230
Raja A, Gopikrishnan T (2022) Drought prediction and validation for desert region using machine learning methods. Int J Adv Comput Sci Appl 13(7):47–53. https://doi.org/10.14569/IJACSA.2022.0130707
Raza A, Phan T-L, Li H-C et al (2024) A comparative study of machine learning classifiers for enhancing knee osteoarthritis diagnosis. Information 15:183. https://doi.org/10.3390/info15040183
Robine J-M, Cheung SLK, Le Roy S et al (2008) Death toll exceeded 70,000 in Europe during the summer of 2003. C R Biol 331:171–178. https://doi.org/10.1016/j.crvi.2007.12.001
Rodrigues M, Santana P, Rocha A (2021) Modelling of temperature-attributable mortality among the elderly in lisbon metropolitan area, Portugal: a contribution to local strategy for effective prevention plans. J Urban Health Bull N Y Acad Med 98:516–531. https://doi.org/10.1007/s11524-021-00536-z
Rosales-Pérez A, García S, Herrera F (2023) Handling imbalanced classification problems with support vector machines via evolutionary bilevel optimization. IEEE Trans Cybern 53:4735–4747. https://doi.org/10.1109/TCYB.2022.3163974
Rousi E, Kornhuber K, Beobide-Arsuaga G et al (2022) Accelerated western European heatwave trends linked to more-persistent double jets over Eurasia. Nat Commun 13:3851. https://doi.org/10.1038/s41467-022-31432-y
Rudin C (2019) Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat Mach Intell 1:206–215. https://doi.org/10.1038/s42256-019-0048-x
Sahani J, Kumar P, Debele SE (2023) Efficacy assessment of green-blue nature-based solutions against environmental heat mitigation. Environ Int 179:108187. https://doi.org/10.1016/j.envint.2023.108187
SCB (2022) After age 60. A description of older people in Sweden
Segev N, Harel M, Mannor S et al (2017) Learn on source, refine on target: a model transfer learning framework with random forests. IEEE Trans Pattern Anal Mach Intell 39:1811–1824. https://doi.org/10.1109/TPAMI.2016.2618118
Singh D, Singh B (2020) Investigating the impact of data normalization on classification performance. Appl Soft Comput 97:105524. https://doi.org/10.1016/j.asoc.2019.105524
Sjulgård H, Keller T, Garland G, Colombi T (2023) Relationships between weather and yield anomalies vary with crop type and latitude in Sweden. Agric Syst 211:103757. https://doi.org/10.1016/j.agsy.2023.103757
Sköld Gustafsson V, Andersson Granberg T, Pilemalm S, Waldemarsson M (2023a) Identifying decision support needs for emergency response to multiple natural hazards: an activity theory approach. Nat Hazards. https://doi.org/10.1007/s11069-023-06305-2
Sköld Gustafsson V, Hjerpe M, Strandberg G (2023b) Construction of a national natural hazard interaction framework: the case of Sweden. iScience 26:106501. https://doi.org/10.1016/j.isci.2023.106501
SMHI (2020) Health effects of heat stress. https://www.smhi.se/en/research/research-departments/meteorology/heat-and-air-quality-in-cities/health-effects-of-heat-stress-1.167391. Accessed 15 Feb 2024
SMHI (2023) Värmebölja. https://www.smhi.se/kunskapsbanken/meteorologi/temperatur/varmebolja-1.22372. Accessed 15 Feb 2024
Soares PR, Harrison MT, Kalantari Z et al (2023) Drought effects on soil organic carbon under different agricultural systems. Environ Res Commun 5:112001. https://doi.org/10.1088/2515-7620/ad04f5
Sun D, Wen H, Wang D, Xu J (2020) A random forest model of landslide susceptibility mapping based on hyperparameter optimization using Bayes algorithm. Geomorphology 362:107201. https://doi.org/10.1016/j.geomorph.2020.107201
Suthar G, Singh S, Kaul N et al (2023) Prediction of maximum air temperature for defining heat wave in Rajasthan and Karnataka states of India using machine learning approach. Remote Sens Appl Soc Environ 32:101048. https://doi.org/10.1016/j.rsase.2023.101048
SWECO (2024) Building heatwave resilience in European cities https://www.sweco.se/wp-content/uploads/sites/3/2024/06/Building-heatwave-resilience-in-European-cities_webb_240620.pdf
Tak S, Seo E, Dirmeyer PA, Lee M-I (2024) The role of soil moisture-temperature coupling for the 2018 Northern European heatwave in a subseasonal forecast. Weather Clim Extrem 44:100670. https://doi.org/10.1016/j.wace.2024.100670
Tang J, Alelyani S, Liu H (2014) Feature selection for classification: a review
Tsai C-F, Lin W-C, Hu Y-H, Yao G-T (2019) Under-sampling class imbalanced datasets by combining clustering analysis and instance selection. Inf Sci 477:47–54. https://doi.org/10.1016/j.ins.2018.10.029
United Nations Environment Programme (2003) Impacts of Summer 2003 Heat Wave in Europe - Environment Alert Bulletin 2. https://wedocs.unep.org/xmlui/handle/20.500.11822/40942. Accessed 15 Feb 2024
van Straaten C, Whan K, Coumou D et al (2022) Using explainable machine learning forecasts to discover subseasonal drivers of high summer temperatures in western and central europe. Mon Weather Rev 150:1115–1134. https://doi.org/10.1175/MWR-D-21-0201.1
Verdonck T, Baesens B, Óskarsdóttir M, vanden Broucke S (2024) Special issue on feature engineering editorial. Mach Learn 113:3917–3928. https://doi.org/10.1007/s10994-021-06042-2
Vieira Passos M, Kan J-C, Destouni G et al (2024) Identifying regional hotspots of heatwaves, droughts, floods, and their co-occurrences. Stoch Environ Res Risk Assess. https://doi.org/10.1007/s00477-024-02783-3
Vu A, Rutherford S, Phung D (2019) Heat health prevention measures and adaptation in older populations—a systematic review. Int J Environ Res Public Health 16:4370. https://doi.org/10.3390/ijerph16224370
Wang H, Wang H, Wu Z, Zhou Y (2021) Using multi-factor analysis to predict urban flood depth based on naive bayes. Water 13:432. https://doi.org/10.3390/w13040432
Wang Z, He B, Lai X (2022) Balanced random forest model is more suitable for wildfire risk assessment. In: IGARSS 2022–2022 IEEE International Geoscience and Remote Sensing Symposium. pp 3596–3599
Weirich-Benet E, Pyrina M, Jiménez-Esteve B et al (2023) Subseasonal prediction of central european summer heatwaves with linear and random forest machine learning models. Artif Intell Earth Syst. https://doi.org/10.1175/AIES-D-22-0038.1

Wild M, Liepert B (2010) The earth radiation balance as driver of the global hydrological cycle. Environ Res Lett 5:025203. https://doi.org/10.1088/1748-9326/5/2/025203

Wu X, Kumar V, Ross Quinlan J et al (2008) Top 10 algorithms in data mining. Knowl Inf Syst 14:1–37. https://doi.org/10.1007/s10115-007-0114-2

Xu Z, FitzGerald G, Guo Y et al (2016) Impact of heatwave on mortality under different heatwave definitions: a systematic review and meta-analysis. Environ Int 89–90:193–203. https://doi.org/10.1016/j.envint.2016.02.007

Zarei M, Destouni G (2024) A global multi catchment and multi dataset synthesis for water fluxes and storage changes on land. Sci Data 11:1333. https://doi.org/10.1038/s41597-024-04203-1

Zennaro F, Furlan E, Simeoni C et al (2021) Exploring machine learning potential for climate change risk assessment. Earth-Sci Rev 220:103752. https://doi.org/10.1016/j.earscirev.2021.103752

Zhang S (2022) Challenges in KNN Classification. IEEE Trans Knowl Data Eng 34:4663–4675. https://doi.org/10.1109/TKDE.2021.3049250

Zhang C, Wang G, Zhou Y, Jiang J (2017) A new approach for imbalanced data classification based on minimize loss learning. In: 2017 IEEE Second International Conference on Data Science in Cyberspace (DSC). pp 82–87