Improved drought forecasting in Kazakhstan using machine and deep learning a non-contiguous drought analysis approach

Sadrtdinova, Renata; Perez, Gerald Augusto Corzo; Solomatine, Dimitri P.

DOI
10.2166/nh.2024.154

Publication date
2024

Document Version
Final published version

Published in
Hydrology Research

Citation (APA)
Sadrtdinova, R., Perez, G. A. C., & Solomatine, D. P. (2024). Improved drought forecasting in Kazakhstan using machine and deep learning: a non-contiguous drought analysis approach. *Hydrology Research*, *55*(2), 237-261. https://doi.org/10.2166/nh.2024.154

Hydrology Research

Hydrology Research Vol 55 No 2, 237 doi: 10.2166/nh.2024.154

Improved drought forecasting in Kazakhstan using machine and deep learning: a non-contiguous drought analysis approach

Renata Sadrtdinova [a], Gerald Augusto Corzo Perez [a,*] and Dimitri P. Solomatine [a,b,c]

[a] Department of Hydroinformatics and Socio-Technical Innovation, IHE Delft Institute for Water Education, 2611 AX Delft, The Netherlands
[b] Water Problems Institute of RAS, Gubkina 3, 119333 Moscow, Russia
[c] Water Resources Section, Delft University of Technology, 2628 CD Delft, The Netherlands
*Corresponding author. E-mail: g.corzo@un-ihe.org

RS, 0000-0001-9849-8350; GACP, 0000-0002-2773-7817; DPS, 0000-0003-2031-9871

ABSTRACT

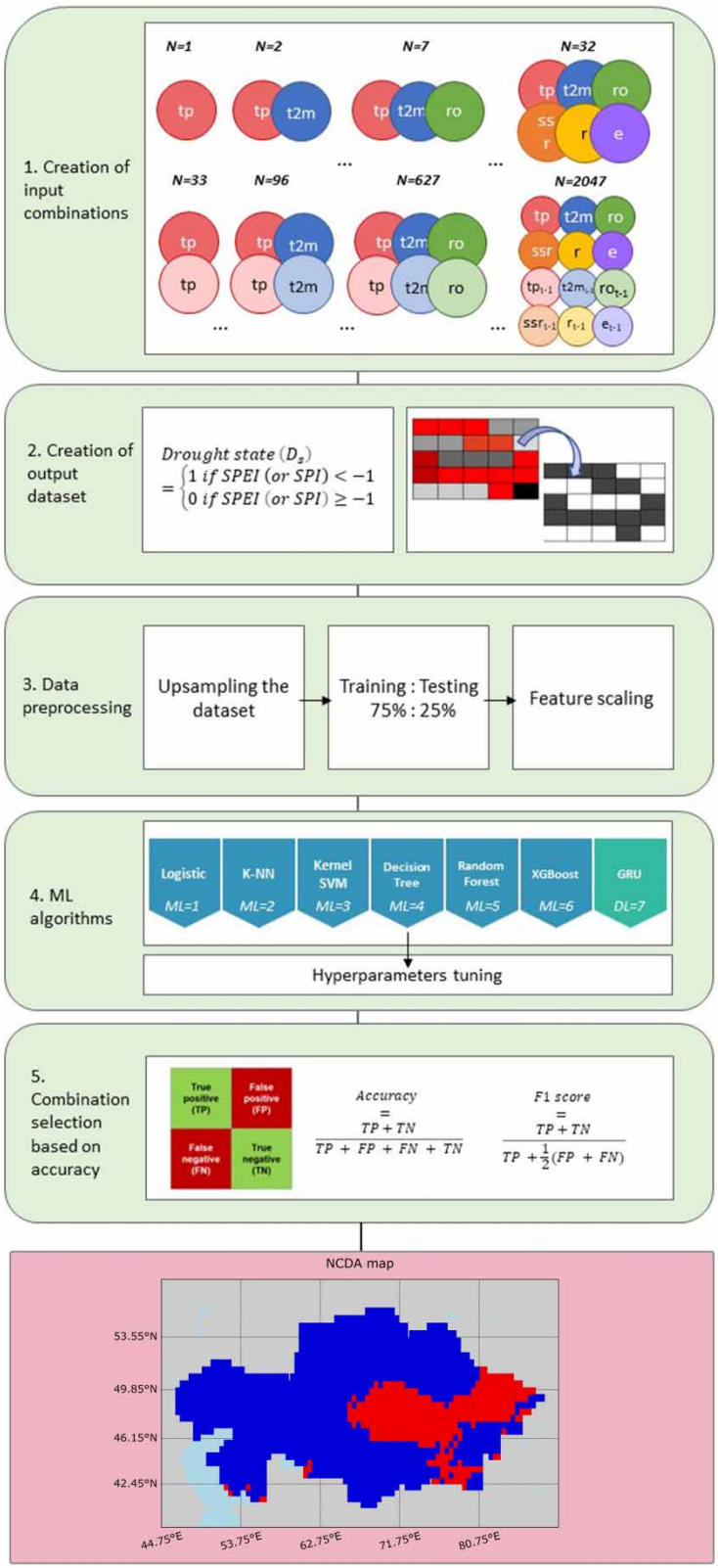
Kazakhstan is recently experiencing an increase in drought trends. However, low-capacity probabilistic drought forecasts and poor dissemination have led to a drought crisis in 2021 that resulted in the loss of thousands of livestock. To improve drought forecasting accuracy, this study applies Machine Learning and Deep Learning (ML and DL) algorithms to capture the sequences of drought events using a non-contiguous drought analysis (NCDA). Precipitation, 2-m temperature, runoff, solar radiation, relative humidity, and evaporation were collected from the ERA5 database as input variables. Combinations of inputs were used to build ML models, including seven classifiers (Logistic, K-NN, Kernel SVM, Decision Tree, Random Forest, XGBoost, and GRU). The output events were defined by standardized precipitation index (SPI) and SPEI indicators as binary classes. Weekly time series from 1991 to 2021 for each cell were used to forecast a lead time from 1 week to 6 months. GRU provided 97–99% accuracy in more volatile regions while Random Forest and XGBoost showed 94–99% accuracy at a lead time of 6 months. The accuracy evaluation was based on the confusion matrix and F1 score to analyze the stage change capture. This study demonstrates the effectiveness of using ML and DL algorithms for drought forecasting, with potential applications for other regions.

Key words: deep learning, machine learning, NCDA, spatiotemporal drought forecasting, SPEI, SPI

HIGHLIGHTS

- Advanced Forecasting: ML and DL algorithms, including non-contiguous drought analysis, were implemented.
- Data Diversity: ERA5 data on precipitation, temperature, and more is used for model construction.
- High Accuracy: GRU achieves 97-99% accuracy, and Random Forest/XGBoost show 94-99% accuracy at a 6-month lead time.
- Global Relevance: Study highlights ML/DL effectiveness in drought forecasting, applicable to similar regions.

GRAPHICAL ABSTRACT



1. INTRODUCTION

Droughts are drawing the attention of experts in various fields increasingly from year to year as an emerging environmental disaster. Droughts occur in all climatic zones, including both high- and low-rainfall areas. Hot temperatures, low relative humidity, and the timing and characteristics of precipitation – including the distribution of wet days during agricultural growing seasons, the strength and duration of the rain, as well as its onset and termination – all aggravate them (Mubenga-Tshitaka *et al.* 2021). A drought, unlike aridity, which is a permanent component of climate and is limited to low-rainfall areas, has a temporary occurrence. Drought can strike everywhere on the planet, having a devastating impact on water supplies and economic activities. According to FAO (2017), in the last 40 years, the percentage of land affected by drought has doubled. For developing countries, 80% of losses are associated with droughts since agriculture is influenced the most. Although drought impacts are severest for the developing world, developed countries are under the same risk. For example, the United States experienced the harshest flash droughts in 2007, 2012, and 2017 (USDM 2021). The World Meteorological Organization (WMO) emphasizes the importance of developing and implementing national policies based on the best definition and characterization of drought to improve drought impact mitigation. Droughts are produced by a distinct combination of environmental and economic elements in meteorological, ecological, agricultural, hydrologic, and socioeconomic droughts, making it complicated to construct a single holistic definition of it. The recent appearance of flash drought has added to the complexity. Flash droughts are defined by their sudden onset, which is usually caused by abnormally elevated temperatures, high evapotranspiration, little precipitation, and low soil moisture. The current monitoring is mostly accounting for a lack of water (Otkin *et al.* 2018).

To be prepared for drought, it is essential to know its spatial distribution by having a coherent procedure for forecasting and dependable models. Recent technologies such as machine learning (ML) have been studied to provide reliable long-term forecasts months in advance. ML, called data-driven models, is significantly less complex than physical-based models, uses far fewer computational resources, and can reach greater accuracy. The capacity to uncover subtle or hidden patterns in complicated geographical data with a limited prior understanding of how these factors interact is a benefit of ML (Brust *et al.* 2021).

ML has significantly advanced drought research, enhancing our ability to predict, monitor, and mitigate drought conditions. These applications include early warning systems for drought prediction, the use of remote sensing and satellite data for real-time monitoring, crop yield prediction to aid farming decisions, optimization of water resource management during droughts, forecasting droughts using meteorological and environmental data, assessing risks related to drought-related disasters, integrating diverse data sources for a comprehensive understanding of drought conditions, developing climate change adaptation strategies, creating decision support systems for policy and response guidance, and building public awareness tools to educate and engage the public. These developments rely on high-quality data and ongoing model refinement, making ML crucial for addressing drought challenges (Prodhan *et al.* 2022; Ghobadi & Kang 2023). The accuracy of the results however varies and depends on the data quality and quantity, the ability of modelers to incorporate all important factors of the problem in question, and the specifics (physics) of the modeled processes. An emerging trend is the 'explainable AI' (see Molnar 2020) aiming at building ML models which would not be black boxes but allow for interpretation by domain experts; for example, M5 model trees (being equivalent to piece-wise linear models) belong to this class of models. A similar trend is named 'physics-aware AI', aiming at a combination of process-based (which encapsulate the process physics) and data-driven (ML) models (e.g., Jiang Zheng & Solomatine 2020). In this paper, a somewhat simpler (and more often used) approach to the incorporation of physical knowledge is employed: its essence is in the choice of the most relevant set of variables (features) to be used in ML models as inputs, aiming either at optimizing the resulting model performance, or via incorporation of expert knowledge, or both. This approach is typically referred to as 'feature engineering', or 'input variables selection' (see e.g., Moreido *et al.* 2021).

In terms of the ML techniques used, there are various techniques that have been used in modeling hydrometeorological processes. Lately, a lot of attention has been given to the so-called deep learning (recurrent multi-layered neural networks), e.g., Long Short-Term Memory (LSTM) and Gated Recurrent Units (GRUs) (a version of LSTM), which also demonstrated high accuracy in drought forecasting (Brust *et al.* 2021; Dikshit & Pradhan 2021).

Consequently, the goal of the study is to develop and examine the methodology for forecasting the meteorological drought events and spatial extent using the standardized precipitation index (SPI) and SPEI (classifying subregions to be in a state of drought, or not). As a case study, this work will focus on Kazakhstan. Although most of the land in Kazakhstan is already arid

and semi-arid, seasonal droughts create even more hostile conditions for agriculture and normal human life activities. This research employs a methodology based on the input variable selection and comparative analysis of several ML models, in which forecasting uncertainty is also analyzed. Finally, conclusions are drawn, and further recommendations are made.

2. STUDY AREA AND DATA

2.1. Case study

Kazakhstan is the ninth largest country in the world and the largest land-locked country, located in Central Asia. Although the large area (2.7 million sq. km), the population does not exceed 19 million people. The climate is 'extreme' continental, with hot summers and very cold winters. The weather is a mix of arid and semi-arid, with particularly dry winters (UNESCAP 2021). Nonetheless, the territory experiences significant climate variations, encompassing five distinct types based on the Köppen climate classification. These include continental climates characterized by cold winters and hot (Dfa) and warm (Dfb) summers, cold dry semi-arid regions (BSk), desert (BWk) climates in the southern and western areas, and a Dsa climate in the Turkestan region marked by dry, hot summers (Peel *et al.* 2007). The yearly precipitation varies mostly by latitude and height, resulting in regions with a mean annual sum of less than 100 mm in the southern semidesert plains and 1,000 mm and more, though usually as snow, in the high mountain ranges. In January, the range of average temperatures is from −5 °C in the south to −20 °C in the north. On the contrary, in July, the range of temperatures is from 18 °C (in the north) to 29 °C (south) (Dubovyk *et al.* 2019), which can be observed in Figure 1 which shows the study area and Walter-Lieth climate graphs for various locations across the country.

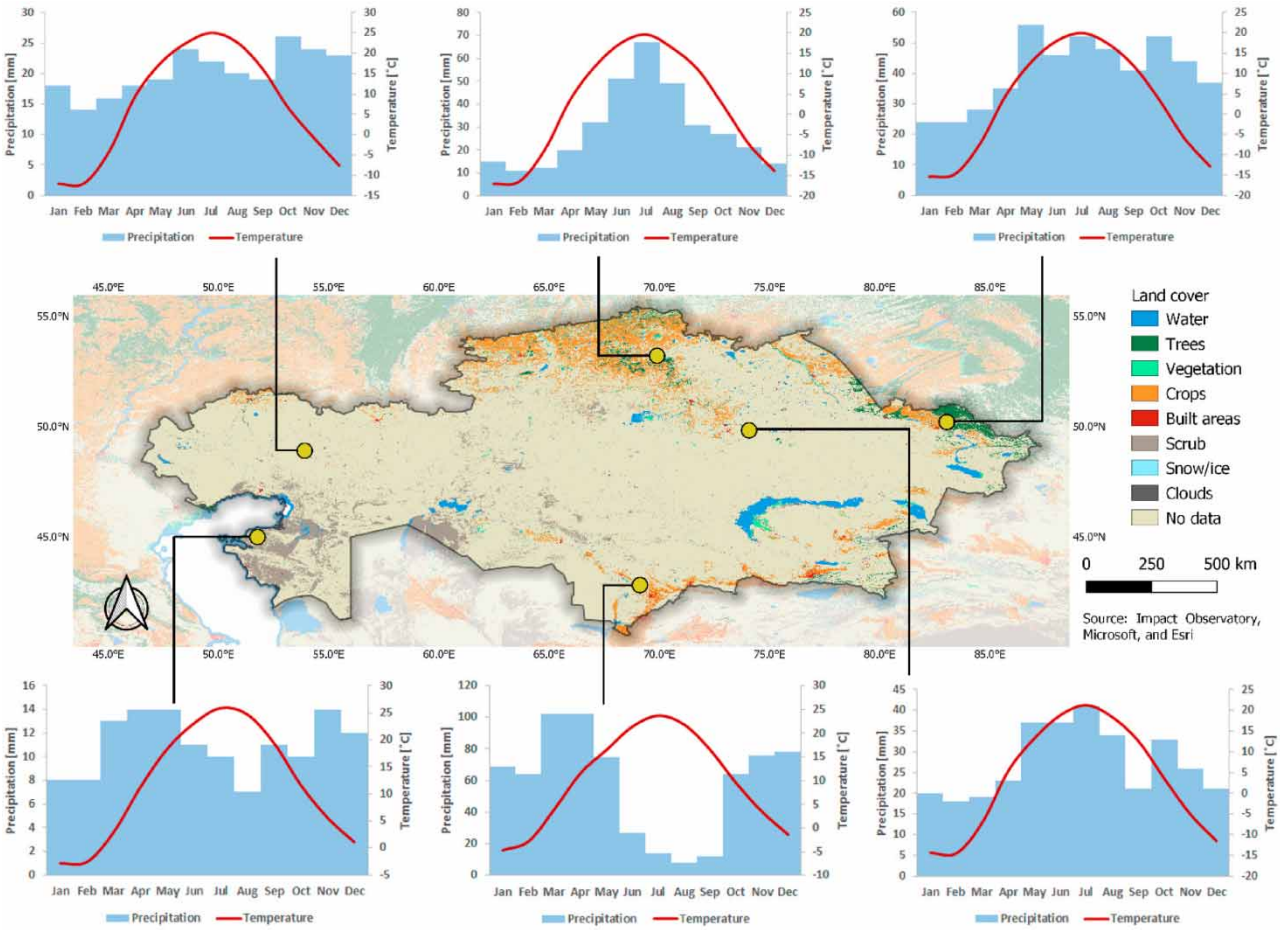


Figure 1 | The study area and Walter-Lieth climate graphs for various locations across the country.

The utilization of water resources in Kazakhstan represents a complex set of interconnected challenges encompassing social, political, and economic aspects. Inadequate management of water resources serves as a hindrance to their effective utilization. As the population in the region continues to grow, the issue of water scarcity in Central Asia becomes increasingly pressing as the population is increasing. Given the scarcity of water resources, it is imperative to view water security as an integral component of national security in the Republic of Kazakhstan. The WMO has defined four levels of stress linked to water scarcity. In Kazakhstan, the highest levels of stress are observed in five out of eight water economic basins (WEBs), with the Shu-Talas and Nura-Sarysu WEBs registering indices of 0.98 and 1, indicating full utilization of river runoff. Anticipated reductions in river runoff in Kazakhstan may lead to significant shifts in both the quantities and patterns of water consumption (Tursunova *et al.* 2022).

In a land-locked country like Kazakhstan, evidence of climate change such as warming oceans, decreasing ice sheets, sea-level rise, and ocean acidification is barely discernible, yet temperature rise and the increased likelihood of occurrence of extreme events (floods and droughts) are of great concern. The ICCP RCP8.5 projection of temperature and precipitation changes is going to extremes by the end of the century. If the temperature is rising all over the territory, the precipitation pattern is different depending on the region: it is increasing in the North and East while decreasing in the South and West. West Kazakhstan is already experiencing severe drought conditions for the last 3 years, having the severest in history last summer, in 2021 (Pannett 2021). It is forecast that Kazakhstan will be affected by all types of droughts (meteorological, hydrological, and agricultural). Currently, an annual mean probability of a meteorological drought is not exceeding 5%. Nevertheless, even with the most fortunate scenario (RCP2.6), the probability of severe drought is increasing up to 40%. The greatest impact is on the West and South (Mangystau and Kyzylorda), having a probability of over 80% by the end of the century (ADB 2021). Mangystau region is already experiencing drought conditions and drastic loss of livestock.

An early drought warning is the responsibility of the National Hydro-Meteorological Service (NMHS). NMHS develops forecasts and predictions of extreme weather events and risks based on the network of hydrometeorological stations for prompt communication of all key stakeholders, including state authorities, economic sectors, and the public. To be able to provide improved operational monitoring and drought event forecasting, the present system needs to advance its technological capabilities. Responsible authorities should consider tools like ML forecasting to provide accurate and timely information for knowledge-driven decisions related to droughts in the country's various social and economic sectors considering the rising risks associated with the impacts of climate variability and climate change in the region (Dubovyk *et al.* 2019).

2.2. Data

2.2.1. Data for input variables

One of the most common hydrometeorological datasets used in research is the ERA5 data, from the European Centre for Medium-Range Weather Forecasts (ECMWF). Initially, we aimed to consider the hydrometeorological variables typically used in drought forecasting, e.g., in the study by Brust *et al.* (2021), namely 13 variables: precipitation, surface soil moisture, vapor pressure deficit, rootzone soil moisture, wind speed, minimum temperature, solar radiation, maximum relative humidity, maximum temperature, minimum relative humidity, evapotranspiration, gross primary product, and runoff. However, not all variables were present in the database for the chosen period (1991–2021). Soil moisture is an essential indicator of the water content available in the soil, and it was a desired input parameter to consider. Unfortunately, none of the available ECMWF datasets could provide soil moisture values within the historical period chosen (1991–2021) and for the whole area. We were also not able to acquire additional datasets during the duration of this study. Finally, data for the six variables are presented in Table 1. Final input parameters were used. The monthly averaged data were retrieved from Muñoz Sabater (2019) databases. Determining whether the parameters are enough for accurate forecasting is a part of the research. As an important reference, Liu *et al.* (2021) also used a (limited) set of 5 variables for similar research (precipitation, air temperature, relative humidity, sunshine duration, and wind), obtaining successful results.

Data on input parameters were collected from the ERA5 database as '.nc' files having dimensions of a rectangle, ideally covering the area of Kazakhstan from the upper left corner at 55 °N 46 °E to the lower right corner at 40 °N 88 °E. The data were further processed using R to break down the area into cells $0.5\,^{\circ} \times 0.5\,^{\circ}$ following the shape of Kazakhstan (total of 1,480 cells) for further detailed analysis, as shown in Figure 2.

The aim of this research is to demonstrate that precise forecasting can be achieved using readily accessible data, such as ERA5, even in regions where a diverse range of data is typically unavailable. Consequently, data from weather stations in Kazakhstan was neither utilized as a data source nor employed for verification purposes.

Table 1 | Final input parameters

Index	Meteorological parameter	Units
1	Precipitation	m
2	2 m temperature	K
4	Solar radiation	Jm^{-2}
5	Relative humidity	%
6	Evaporation	m of water equivalent
	Hydrological parameter	
3	Runoff	M

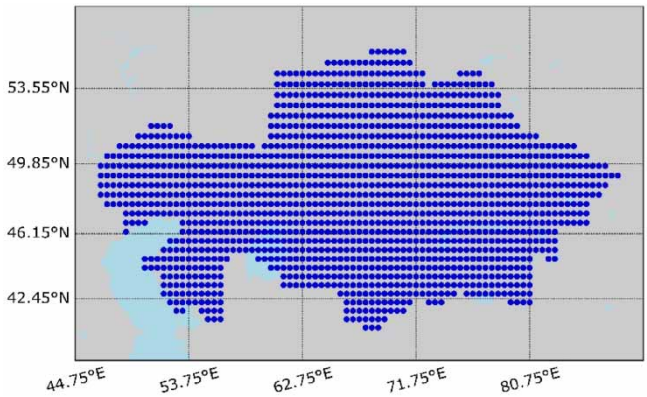


Figure 2 | The division of Kazakhstan into 1,480 cells with a size of 0.5° × 0.5°.

2.2.2. Data for output variables

2.2.2.1. Standardized precipitation index. The SPI from precipitation data was calculated by using Python code, by defining the *spi* function as the inverse of CDF (gamma distribution). Since the precipitation data were retrieved from the ECMWF database, it was possible to obtain the SPI for all pixels (0.5° × 0.5°, 1,480 pixels in total following the shape of Kazakhstan) for a monthly interval from January 1991 to December 2021. The average monthly data were converted to a weekly one by assuming that 1 month = 4 weeks using the formulae below.

The formulae for converting monthly to weekly data for 2-m temperature and relative humidity are as follows:

$$\text{Week 1} = f(\text{value of the Month 1})$$

$$\text{Week 2} = \frac{3}{4}[\text{Month 1}] + \frac{1}{4}[\text{Month 2}]$$

$$\text{Week 3} = \frac{1}{2}[\text{Month 1}] + \frac{1}{2}[\text{Month 2}]$$

$$\text{Week 4} = \frac{1}{4}[\text{Month 1}] + \frac{3}{4}[\text{Month 2}]$$

For the precipitation, runoff, solar radiation, and evaporation the formulae are as follows:

$$\text{Week 1} = f\left(\frac{\text{Month 1}}{4}\right)$$

$$\text{Week 2} = \frac{1}{4}\left(\frac{3}{4}[\text{Month 1}] + \frac{1}{4}[\text{Month 2}]\right)$$

$$\text{Week } 3 = \frac{1}{4}\left(\frac{1}{2}[\text{Month } 1] + \frac{1}{2}[\text{Month } 2]\right)$$

$$\text{Week } 4 = \frac{1}{4}\left(\frac{1}{4}[\text{Month } 1] + \frac{3}{4}[\text{Month } 2]\right)$$

2.2.2.2. Standardized Precipitation Evaporation Index. The SPEI-1 indices were collected from the Global SPEI database (Beguería *et al.* 2014). The collected data for Kazakhstan will contain only pixels related to Kazakhstan (similarly to SPI, $0.5° \times 0.5°$, 1,480 pixels in total). Since the database contains time series up to December 2018, it was decided not to infuse the available data with the data from other datasets. Therefore, the historical period of analysis was also modified for SPEI-based model analysis to be January 1991–December 2018.

3. METHODOLOGY

This study mainly follows the spatiotemporal drought analysis methodology, and non-contiguous drought analysis (NCDA), developed by Corzo Perez *et al.* (2011). NCDA is used to identify the hydrological drought on a larger scale. Although the reference study analyzes hydrological droughts, the same methodology can be applied to meteorological droughts since all droughts are caused by a lack of precipitation. NCDA focuses on Kazakhstan as a whole. First, the drought index calculation was conducted where the definition of the threshold of the drought anomaly was chosen based on the literature review. The majority of the studies follow McKee Doesken & Kleist (1993), choosing -1 as a threshold. This indicates that for time scale i, a drought event is defined as a period during which the SPI is consistently negative and achieves a value of -1 or below. The drought begins when the SPI drops below zero for the first time and ends when the SPI returns to a positive value after a value of -1 or less. Therefore, -1 was chosen as a threshold for both SPI and SPEI as Liu *et al.* (2021) used it.

3.1. ML model setup

There were two sets of modeling experiments, called stages, as shown in Figure 3. The first stage (called 'regionalization') consisted of building models for only six reference points (representative of a specific region), spread across the area. In the second stage, more advanced analysis was undertaken for all the 1,480 points.

Six parameters including precipitation, 2-m temperature, runoff, solar radiation, relative humidity, and evaporation were used as inputs (at time $= t$ and $t - 1$) for modeling the drought conditions using six ML models such as Logistic, K-Nearest Neighbor (K-NN), Kernel Support Vector Machine (Kernel SVM), Decision Tree, Random Forest, and XGBoost, and 1 deep learning model such as GRU. The models were trained and assessed with 75 and 25% of the dataset, respectively. The dataset consisted of the weekly target value and the prediction factors from January 1991 to December 2021 obtained from the ECMWF database (ERA5). To identify the best combination of input variables, a comprehensive feature selection analysis was carried out using the algorithm. Since the database was imbalanced with drought events, the upsampling was performed, which will be discussed later. The hyperparameters of ML models were tuned to obtain the highest possible accuracy. The applied methodology is depicted in Figure 4.

Preventing overtraining (overfitting) is a critical aspect of ML, and there are various measures taken to ensure that the model generalizes effectively to the new, unseen data. It can be done in different ways, e.g. introducing a cross-validation

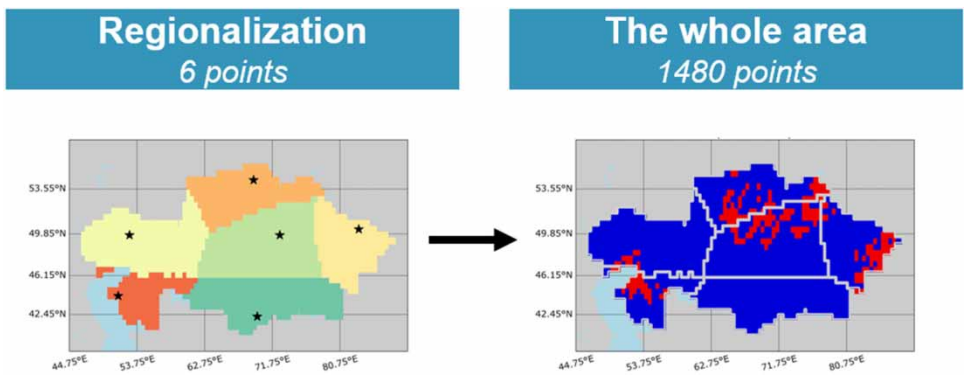


Figure 3 | Hierarchy of the analysis.

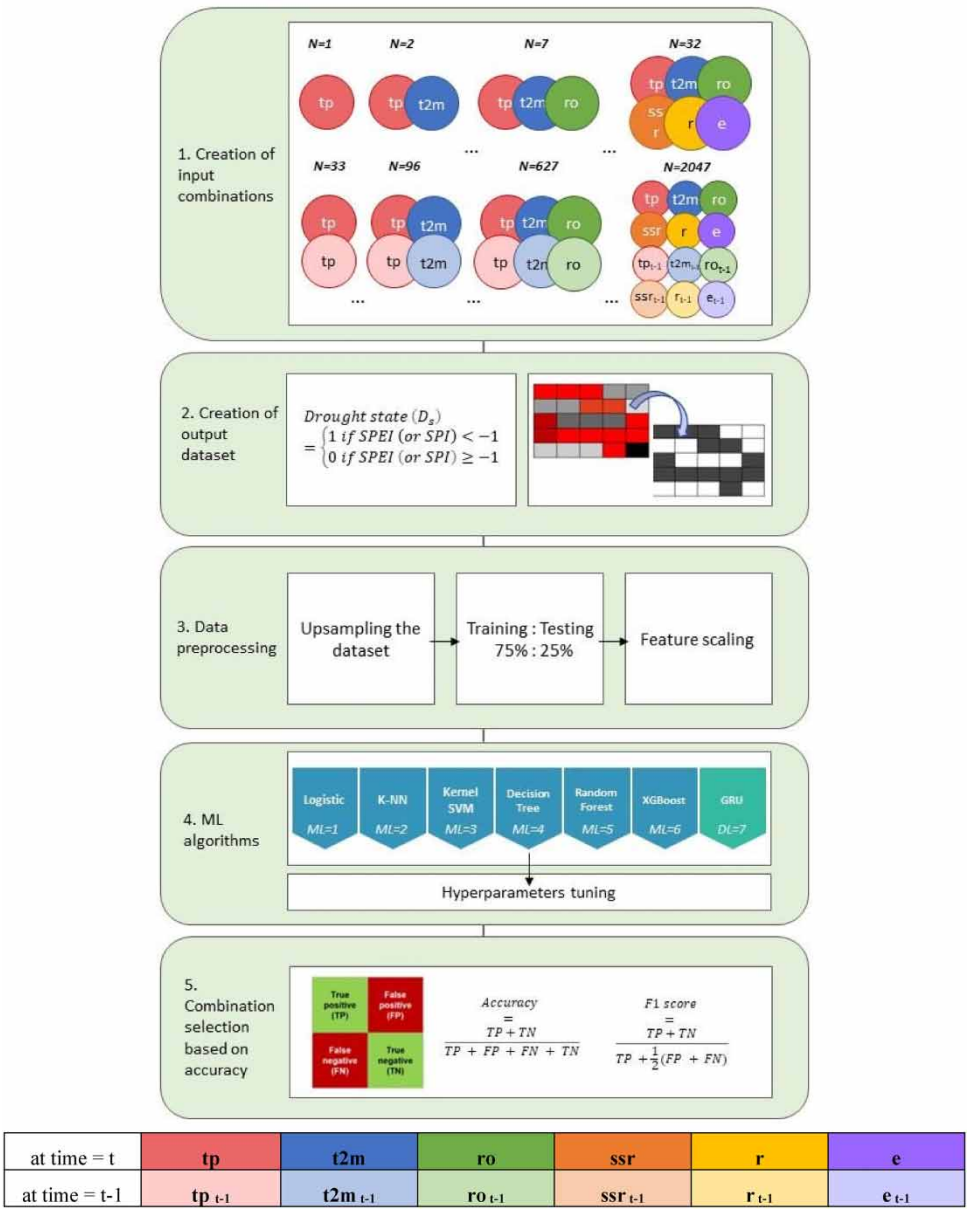


Figure 4 | Methodology flow chart.

set. We have not done that, but undertook several other ways to prevent overfitting, mainly by ensuring that the model is not more complex than needed (Occam's razor principle). First, at various stages we have employed the 'leave-one-out' method to check validity of the model on unseen data. Second, we conducted feature selection to eliminate irrelevant features, and hence building simpler models. Thirdly, we experimented with models of different structures (complexities) to determine the most suitable level. Fourth, we applied data augmentation to increase the dataset's size and introduce data variations. Fifth, we fine-tuned the model's hyperparameters to discover the best settings for the ML models. These strategies are discussed in more detail in the following sections.

3.1.1. Creation of input combinations

Determining the best input combination for the ML model is essential to achieve the highest model efficiency and accuracy. Therefore, to recognize the best input combination, a feature selection analysis was carried out. Having 12 potential input variables (six at time $= t$ and six at time $= t - 1$), the total number of possible combinations with precipitation

(tp at time $= t$) is 2,047. The formation is partially shown in Figure 5, as combinations of input variables. Statistically, there are more combinations of variables without repetition. However, we are interested only in combinations containing precipitation (tp) since both SPI and SPEI are derived from them. Therefore, we excluded combinations that do not contain tp. In conclusion, the best input variable combination was chosen based on the highest accuracy.

3.1.2. Creation of output dataset

Unlike input variables, the output is at time $= t + k$ where k is a lead time. The formula for a drought state is represented below:

$$\text{Drought state } (D_S) = F[tp_{t+k},\ t2m_{t+k}, ro_{t+k},\ ssr_{t+k},\ r_{t+k},\ e_{t+k}]$$

*tp = precipitation, $t2m$ = 2 m temperature, ro = runoff, ssr = solar radiation, r = relative humidity, e = evaporation, and k is a lead time and varies from 1 week to 6 months.

The spatial analysis is based on the occurrence of an event. The analysis was performed for two learning problem types: classification and regression.

3.1.2.1. Classification. The output encoding for classification is presented below:

$$\text{Drought state } (D_s) = \{1 \text{ if SPEI (or SPI)} < -1\ 0 \text{ if SPEI (or } SPI) \geq -1$$

Then, the mask of 0 ('no drought') and 1 ('drought') values is created representing the non-drought and drought events, as shown in Figure 6.

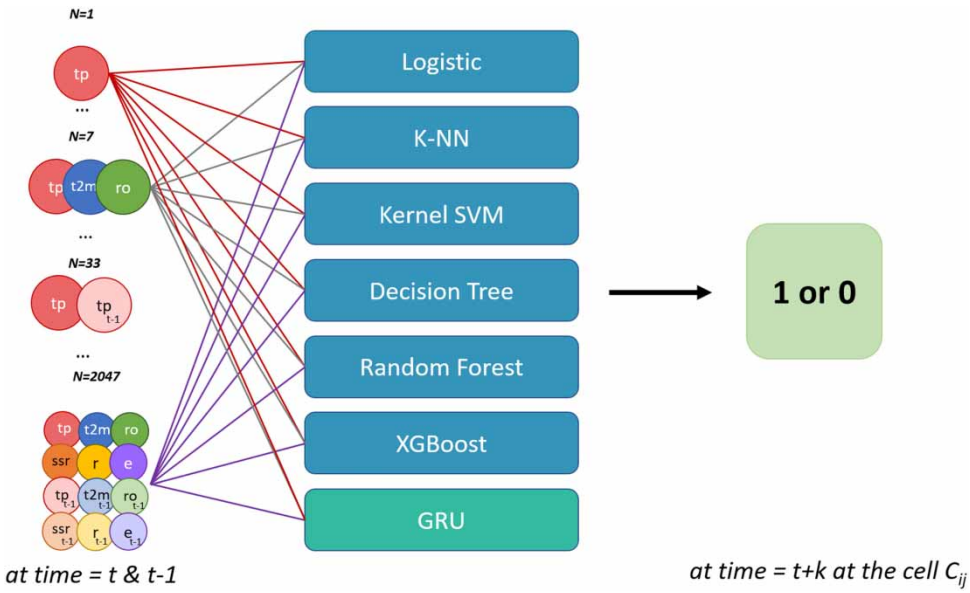


Figure 5 | Combinations of input variables.

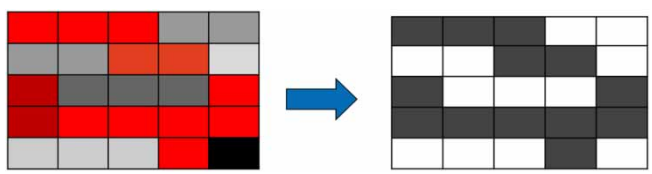


Figure 6 | SPEI distribution from −3 to +3 (left); a mask of 0 and 1: non-drought and drought states (right).

Another important parameter is the percentage of drought area (PDA). PDA can be estimated using the following equation (Corzo Perez *et al.* 2011):

$$\mathrm{PDA}(t) = \frac{100}{A_{\mathrm{tot}}} \times \sum_{c=1}^{N} (D_s(t)A)$$

where A_{tot} is the total land area, A is the area of each cell, and N is the number of cells. The equation can be simpler transformed into the expression below since every input and output parameter are transformed into 0.5° × 0.5° cells, which will be discussed later:

$$\mathrm{PDA}(t) = \frac{\text{Number of drought cells}}{\text{Total number of cells}} \times 100\%$$

3.1.2.2. Regression. The difference between the regression analysis from the classification is that the mask was not applied. The output was SPI or SPEI values themselves, being from −3 to +3. The regression was evaluated only on six ML models, excluding GRU.

3.1.3. Data preprocessing

The training and test sets were built by using *sklearn.model_selection train_test_split* as follows:

$$\text{Dataset} = \{\text{training} \rightarrow 75\% \text{ testing} \rightarrow 25\%\}$$

Input variables for both training and test sets were scaled by *sklearn.preprocessing StandardScaler* since the range of the variables varied significantly (from the negative values of evaporation to 10^6 for the solar radiation).

3.1.3.1. Upsampling to balance the dataset. The analyzed area was divided into regions due to the inhomogeneity of climate and terrestrial ecosystem within the research area. Therefore, the number of drought events also varies from region to region. This led to a significant imbalance of drought events in Regions 1 and 2, having only 15% of drought events in the dataset. In this case, the accuracy is mostly dependent on 'true positives' while the number of 'true negatives' is significantly lower, which questions the reliability of such accuracy. Consequently, it was decided to balance each dataset with drought cases (the drought events (1 s) were duplicated to meet the criteria so that drought and no-drought cases are quantitatively equal) using *resample from sklearn.utils*, obtaining a 50:50 proportion of drought and not drought events, as shown in Figure 7.

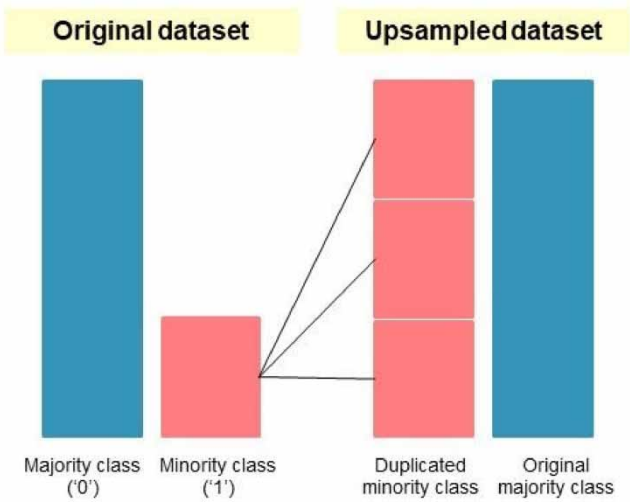


Figure 7 | Visual representation of upsampling.

3.1.3.2. Hyperparameter tuning. It was decided to tune the hyperparameters of ML models to achieve the maximum accuracy possible, which was applied to all classifiers except for Logistic. *RandomizedSearchCV* from *sklearn.model_selection* was used to perform an exhaustive search of the most suitable hyperparameters.

3.1.4. ML algorithms

For this paper, we are applying several ML techniques, one of which is a deep learning model (gated recurrent unit, GRU, a version of a recurrent neural network), and all others we may call 'shallow' ML models.

3.1.4.1. Logistic classification or regression. It is a linear regression model extension for classification problems. Although a linear regression model can be effective for some regression problems, it is too simple to perform well for classification (Molnar 2020). Therefore, it was not expected to have good results on logistic classification, but it was of the research interest to observe the behavior of such a classifier on a complicated correlation between six input variables and a binary outcome.

3.1.4.2. K-Nearest Neighbor. The K-NN approach classifies an unseen sample based on the majority of the k neighbors' (output) classes. Distances to these neighbors can be weighted by the inverse distance to the unseen instance (or using a kernel function of it) (Mucherino Papajorgji & Pardalos 2009).

3.1.4.3. Kernel Support Vector Machine. The SVM model, based on the ideas developed by Vladimir Vapnik in the 1970s (see Vapnik 1999), constructs an N-dimensional surface with margins, which separates samples belonging to different classes, ensuring effective generalization ability even without using a cross-validation set (so-called 'big margin classifier').

3.1.4.4. Decision tree. A decision tree is one of the oldest classification models that uses a series of tests stated at each branch (or node) in the tree to recursively segment a dataset into smaller subdivisions. A root node (made from all the data), a collection of internal nodes (splits), and a set of terminal nodes make up the tree (leaves). The dataset is classed in this framework by systematically subdividing it according to a certain criterion (typically, minimizing entropy in each resulting subset), and a class label is issued to each observation based on which leaf node it falls into (Friedl & Brodley 1997).

3.1.4.5. Random forest. Random forest is a model proposed by Breiman & Cutler (2001) which is a set (ensemble) of classification trees (typically, Breiman's regression trees), which are basic models that predict outcomes using binary splits on predictor variables. Many classification trees are built in the random forest scenario utilizing randomly selected training datasets and random subsets of predictor variables for modeling outcomes. As a result, as compared to a single decision tree model, random forest frequently gives superior accuracy while retaining some of the tree model's advantages. The capacity to manage datasets with many predictor variables is one of the key advantages of utilizing random forests in a wide variety of applications. Regarding variable selection methods for random forests, see, e.g., Speiser *et al.* (2019).

3.1.4.6. XGBoost. Boosting is an approach leading also to an ensemble of decision or regression trees, but it is a sequential model, where each subsequent tree is dependent on the outcome of the previous. Boosting assigns weak learners to a weighted subset of the original dataset. Weak learners have little predictive ability and perform just marginally better than random guessing. Subsets that were previously misclassified are given more weight and hence the probability to be selected for the subsequent learner. As a result, the ensemble has a good generalizing ability. The two widely used versions of boosting are adaptive boosting AdaBoost (see e.g., Shrestha & Solomatine 2006), and gradient boosting (Friedman 2001). A popular implementation of the latter is in XGBoost (extreme gradient boosting), a C++ library with APIs for several languages (XGBoost 2023), and it was used in this study.

3.1.4.7. Gated Recurrent Unit. Cho *et al.* (2014) proposed GRU, a deep learning model, which is comparable to LSTM but easier to compute and apply. The reset gate r and the update gate z make up a typical GRU cell. The update gate selects how to use the previously stored information to generate the new state, whereas the reset gate chooses how to mix the new input

with the previously stored data. Utilizing the hidden state at time t-1 and the input time series value at time t, the hidden state output at time t is calculated. Details can be found in Cho *et al.* (2014) and Lynn *et al.* (2019).

The ML model configuration is presented in Figure 5. The forecasting was performed for both S2S (1–4 weeks) and seasonal periods (2–6 months). Consequently, the total number of models created:

$$2,047 \text{ combinations of input variables} \times 7 \text{ ML \& DL classifiers} \times 2 \text{ set of outputs (SPI \& SPEI)} \times 9 \text{ lead times} \times 6 \text{ regions} = 1,547,532 \text{ models}$$

3.2. Stage 1: regionalization

The research area was divided into regions based on the highest SPI (or SPEI) correlation of every point to the chosen reference points. The assumption is that the area within the region is homogeneous. The points were chosen as they would be in various locations as well as in different terrestrial ecosystems and climate zones, as it is shown in Figure 1. Initially, the points were chosen near the main administrative centers of 5 common divisions of Kazakhstan (North, East, West, Center, and South), being close to Petropavl, Oskemen, Oral, Karaganda, and Shymkent cities, respectively. The optimal number of regions was identified by achieving a correlation of SPI (or SPEI) not lower than 50% of every 1,480 cells to the reference points. Such correlation was still low in the Magystau region, which has a very distinct climate and land type from the rest of the area and where the drought disaster happened in 2021, which was discussed previously. Therefore, it was decided to add one more point near Aktau City, creating Region 0. Thus, the correlation requirement was satisfied, and the regions were cohesively divided. Region 0 is special to this research because it was used as a lumped catchment at which the ML models were evaluated first. The reason behind this is the fact that this is the region where the recent drought disaster occurred as well as the driest region in the country. Therefore, the probability of drought occurrence is quite high. The final regional division (Regions 0–5) is shown in Figure 8(a) and 8(b). For this stage, the reference points were responsible for the forecasting in their whole region. The description of the regions is shown in Table 2.

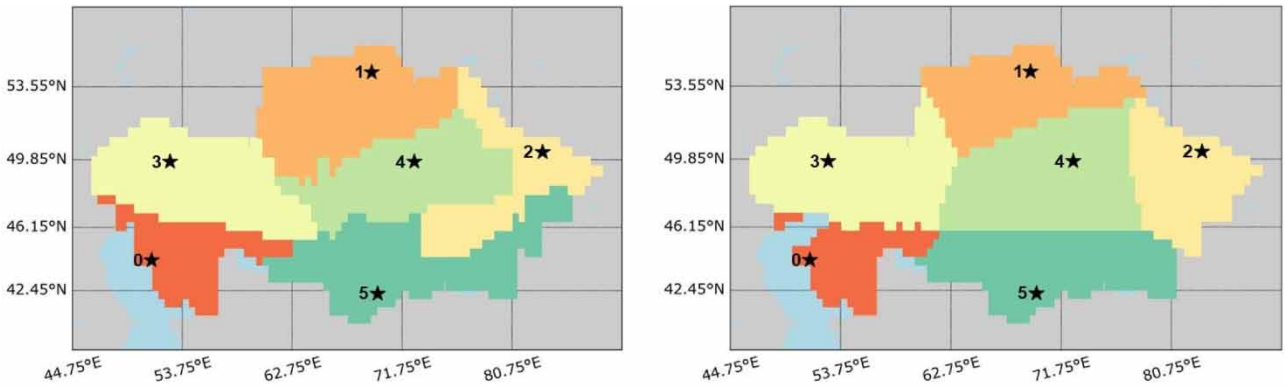


Figure 8 | Regional division based on (a) SPI and (b) SPEI.

Table 2 | Summary of the regions

N	Location of the reference point	Terrestrial ecosystem	Köppen climate class	N of wet days/year	N of cells (SPI based)	N of cells (SPEI based)
0	44.25 ° N 51.25 ° E	Desert	BWk	60	157	132
1	54.25 ° N 69.25 ° E	Forest & cropland	Dfb	147	280	204
2	50.25 ° N 83.25 ° E	Grassland (mountain)	Dfb	141	173	187
3	49.75 ° N 52.75 ° E	Arable land	BSk	94	290	310
4	49.75 ° N 72.75 ° E	Grassland (steppe)	Dfa	129	264	342
5	42.25 ° N 69.75 ° E	Desert	Dsa	86	316	305
					1,480 cells	

3.3. Stage 2: The whole area

At this stage, the multivariate time series of all points (1,480 cells) within the regions are engaged in drought forecasting instead of only six reference points as in Stage 1. The division of the research area into cells was discussed in Section 2.3.2. The difference from Stage 1 is that instead of only 1 point per region, the forecasting is performed for every point within a region. The number of points per region is shown in Table 2. This gives the highest possible precision of the forecasting since the exact location of drought cells is identified. This gives us a clear picture of where exactly drought spatial location is expected. However, Stage 1 was needed to identify the best combinations of input variables and the best performing ML classifier for every lead time and SPI and SPEI-based drought indices, to be used in this stage. Therefore, the exhaustive search is not performed here which saves computational capacity and time.

3.4. Model performance metrics

3.4.1. Accuracy for binary classification

Classifier performance is evaluated using a 2×2 confusion matrix (CM), which includes the numbers of true positive (TP), false negative (FN), true negative (TN), and false positive (FP) predictions, and is used to present the results of a binary classification problem, as shown in Figure 9. CM provides all required information to assess the performance (accuracy) of the results, as shown in the following equation (WCRP 2017).

$$\text{Accuracy} = \frac{\Sigma\text{TP} + \text{TN}}{\Sigma\text{TP} + \text{FP} + \text{FN} + \text{TN}}$$

3.4.2. Accuracy for the state change

Although the method described above is classical for the determination of a classifier's accuracy, it is important to identify whether it works at critical moments such as state change. This means it is essential for the classifier to catch not only drought/no-drought states but also correctly identify the switch from no-drought to drought or vice versa in the initial dataset before the upsampling was introduced. Therefore, an alternative accuracy estimation method is applied called the F1 score. The advantage of the F1 score is that it evaluates both precisions and recall for a fairer summary of model effectiveness (Lipton Elkan & Naryanaswamy 2014). The equation is presented below:

$$\text{F1 Score} = \frac{\text{TP}}{\text{TP} + (1/2)(\text{FP} + \text{FN})}$$

3.4.3. Accuracy for regression

When continuous variables are forecasted (regression problem), the difference between the forecasted and observed values is measured (Hu Palta & Shao 2006). R^2 statistics, also known as coefficients of determination, are one approach to evaluate a model's efficiency because they represent the strength of the regression relationship. The mean of the observed data is estimated as follows:

$$\underline{p} = \frac{\sum_{i=1}^{n} y_i}{n}$$

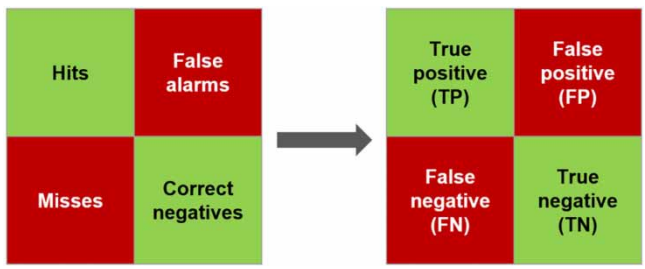


Figure 9 | Confusion matrix for binary classification.

Therefore, two sums of squares formulas can be used to calculate the dataset's variability, using the sums-of-squares type of R^2. Total and residual sums are defined as follows:

$$\mathrm{SST} = \sum_{i=1}^{n} (y_i - \underline{p})^2$$

$$\mathrm{SSE} = \sum_{i=1}^{n} (y_i - \hat{p}_i)^2$$

Consequently, R^2 equals to

$$R^2 = 1 - \frac{\mathrm{SSE}}{\mathrm{SST}}$$

4. RESULTS AND DISCUSSIONS

4.1. Stage 1. Regionalization

4.1.1. Preparing data

4.1.1.1. Choice of the reference points (regions) number. The location of the chosen reference points and their characteristics are presented in Figure 8 and Table 2. As mentioned before, the division was performed based on the highest linear correlation of SPI or SPEI between every 1,480 cells to the reference point. The area is divided into six regions since it is the optimal number of regions to achieve at least a minimum (50%) correlation. Table 3 shows the minimum correlation and the corresponding number of regions. As can be seen from Table 3, the optimal number is 6.

As it was discussed in Section 2.3.1.3.1, upsampling of the training set was performed to balance the dataset with drought events as well as hyperparameter tuning to improve the forecasting accuracy. As discussed in Section 2.3.1.1, the number of input parameters varies from 2 to 6 forming all combinations with precipitation. For easier representation, the hydrological and meteorological parameters in combinations are labeled with numbers.

4.1.1.2. Effect of upsampling. The upsampling was essential since some of the areas had a comparably low number of drought events to be able to train and test the model (by splitting the dataset, the already small number of events was decreasing even more). The effect of upsampling is represented in Figure 10 on a span of 32 combinations (not including $t-1$) as an example, which compares the accuracy before and after the dataset balancing on an example of Region 1 for SPEI-based models: a region has distinct seasons with cold snowy winters and warm rainy summers. As can be seen, the strong variation between the combinations got smoothened and the overall accuracy increased on average at 0.2, while the highest from 0.75 to 0.95. Therefore, it was decided to proceed with balancing the training datasets to allow the machine to learn and be trained better.

4.1.2. Classification by 'shallow' ML models: drought or no-drought

4.1.2.1. Tuning hyperparameters based on the model's performance. One of the major accuracy improvements was to tune the hyperparameters of ML models to achieve the maximum accuracy possible, which was applied to K-NN, Kernel SVM,

Table 3 | The number of regions and minimum achieved correlation

	Minimum correlation [%]	
Number of regions	Based on SPI	Based on SPEI
2	18.0	19.7
3	23.7	24.2
4	32.3	35.6
5	47.6	48.0
6	50.0	51.1

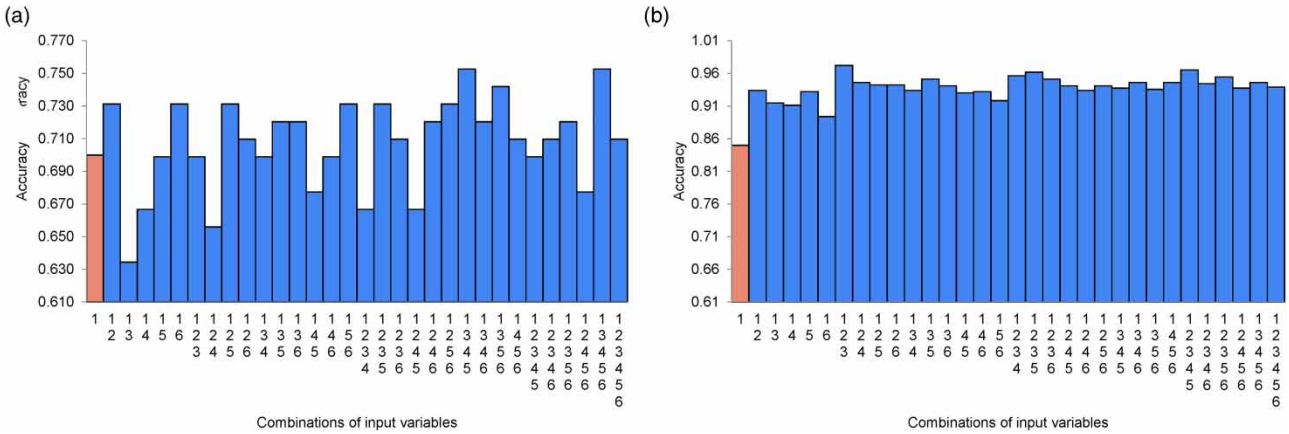


Figure 10 | Accuracy distribution for Region 1 SPI-based drought at lead time = 4 weeks indices (a) before upsampling and (b) after upsampling.

Decision Tree, Random Forest, and XGBoost. Random Forest and XGBoost were already providing high accuracy results (>86%) and the boost of the accuracy was not as significant as for other ML models. Table 4 shows an example of the parameters of the ML output to achieve the highest accuracy at lead time = 4 weeks: region number, SPI or SPEI based, a combination of the best input parameters, the highest accuracy itself, CM (true positive, false positive, false negative, true negative), tuned parameters obtained through a randomized search of the hyperparameters' combinations.

As can be observed from Table 4, before tuning, XGBoost was showing the highest accuracy with the combination of input variables. Then, when the parameters were tuned (for every ML model except Logistic), Random Forest showed better accuracy with another combination of inputs by adjusting the number of estimators, and minimum sample splits. Therefore, the boost of the accuracy from 0.86 to 0.94 was obtained for the SPEI-based model of Region 0 and from 0.94 to 0.99 for the SPI-based model of Region 1.

Figure 11 shows the highest accuracy among six ML models for every combination of input parameters for every region at lead time = 4 weeks as an example.

We can see that Region 0 shows significantly lower results than the rest of the regions. Possible reasons why Region 0 performs worse than other regions will be discussed later. Figure 12 represents the dynamics of the implication of every one of the input variables (tp, t2m, ro, ssr, r, and e) in the best combination while seeing whether they were at time = t, $t-1$, or the same parameter at t and $(t-1)$.

The evolution of the variables' best combination as well as their feature importance is shown in Figure 13 as an example of lead time = 1 week and 6 months (minimum and maximum lead times). Graphs for other lead times are presented in

Table 4 | A comparison of the ML output when the parameters are untuned and tuned

Region ID	SPI/SPEI	Combination	Accuracy	ML model	Confusion matrix		Untuned/Tuned parameters
0	SPEI	1 2 3 5 6 7 9 12	0.8627	XGBoost	236 31	15 53	Untuned
0	SPEI	1 2 3 6 9 10 11 12	0.9413	Random Forest	228 34	23 50	{'n_estimators': 90, 'min_samples_split': 2, 'min_samples_leaf': 2, 'max_features': 'auto', 'max_depth': 30, 'bootstrap': False}
1	SPI	1 2 3 4 5 6 7 10 11 12	0.9407	XGBoost	325 19	3 24	Untuned
1	SPI	1 2 3 4 7 8 10 12	0.9923	Random Forest	324 28	4 15	{'n_estimators': 50, 'min_samples_split': 2, 'min_samples_leaf': 1, 'max_features': 'auto', 'max_depth': 50, 'bootstrap': False}

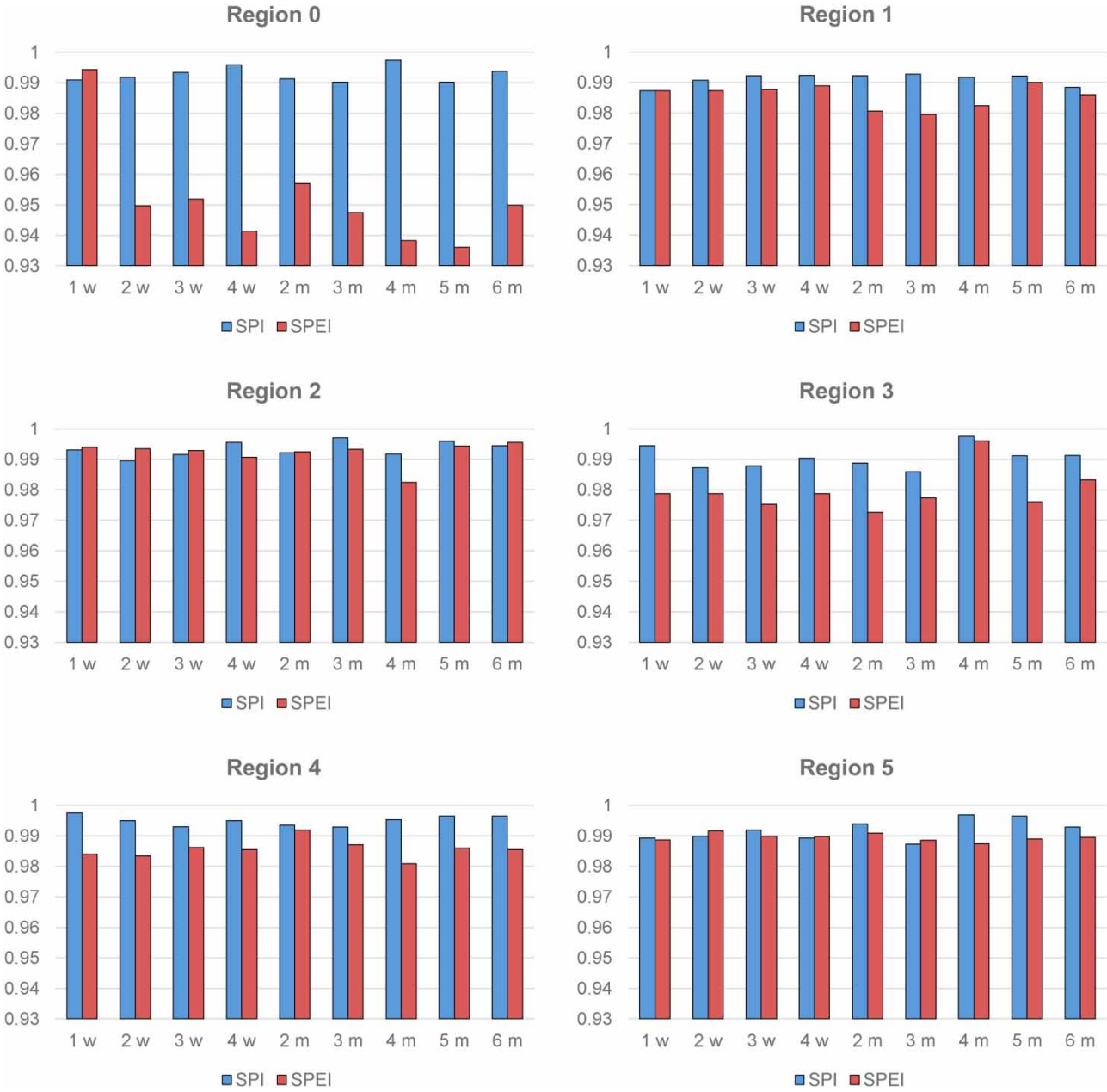


Figure 11 | The highest accuracy distribution of every input parameters combination among six ML classification models for SPI- and SPEI-based drought indices.

Supplementary material, Appendix, Figure A1. However, it is difficult to see any pattern. ML is not deterministic; it is stochastic. There are no physical constraints. As you can see from the figure, even though several best combinations for different regions, lead times, and SPI/SPEI base have runoff (at time t or/and $t-1$) as one of the input variables, it has weight zero. However, when the same combination exists without a runoff, the accuracy is lower. This means that the addition of runoff, disregarding whether it is hollow or not, readjusts the weights of input parameters, thereby increasing the accuracy of the forecasting.

As can be seen from Figure 13, for SPI-based models at lead time = 1 week, precipitation at $(t-1)$ has the highest importance having a much higher value than the rest of the variables, except for Region 5. For SPEI-based models, the importance of the variables is more distributed than for the SPI-based. Regarding the lead time = 6 months, it is seen that the

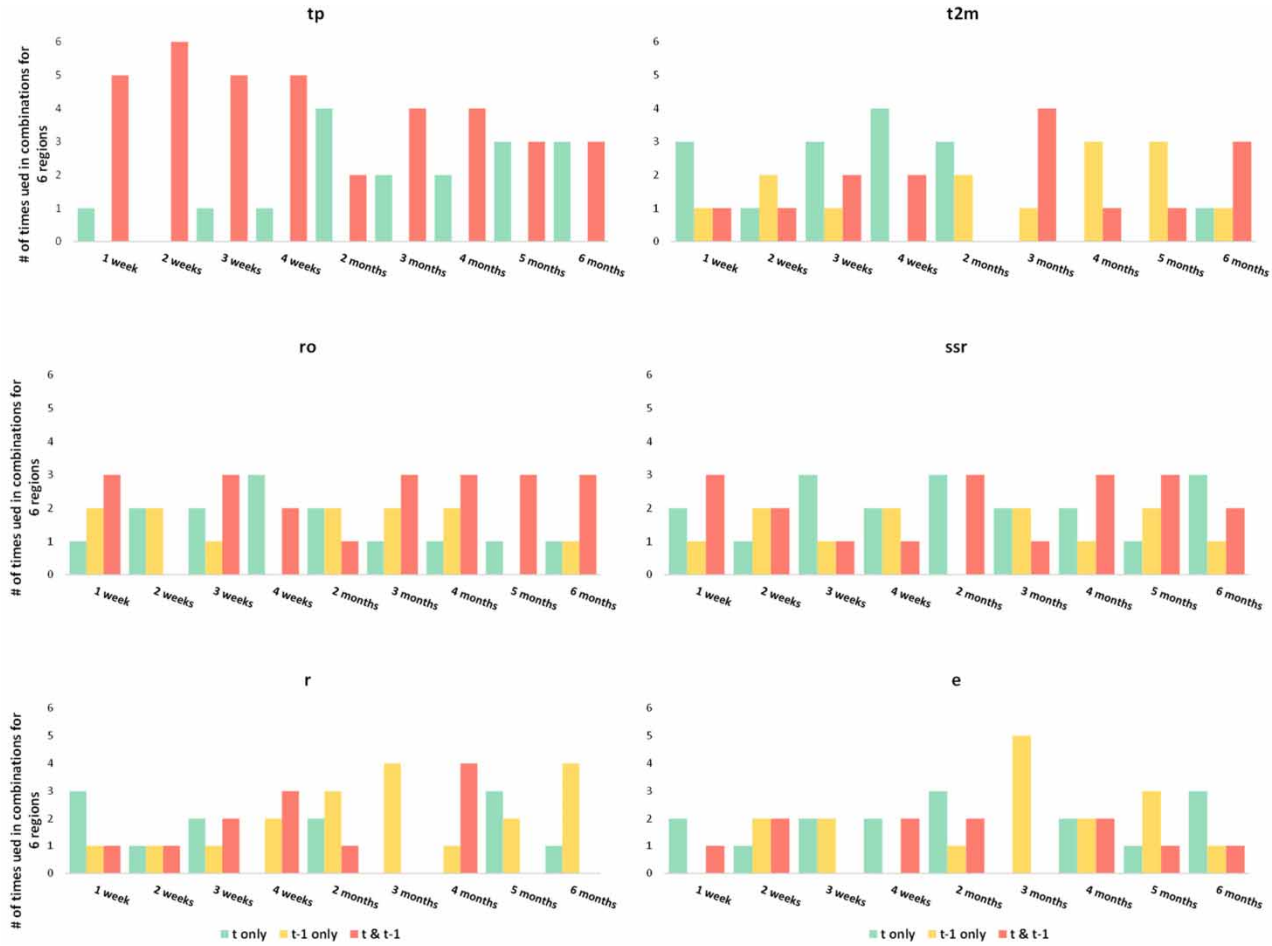


Figure 12 | Number of times the input variable appeared in the best combination for all six regions through lead times from 1 week to 6 months.

combinations change significantly for SPI-based rather than for SPEI-based. What is observed for SPEI-based is that evaporation at $(t-1)$ becomes an important parameter to identify the drought conditions.

4.1.3. Classification by a deep learning algorithm (GRU)

Although tuned Random Forest and XGboost ML models already presented high results (up to 0.99), the results for Region 0 (the region where the drought event occurred in 2021) were still lower. Therefore, it was decided to test the deep learning model such as GRU the same as Brust *et al.* (2021) did. The combined accuracy results of tuned classification models and GRU are presented in Figure 14.

What can be observed from Figure 14 is that the accuracy increased from 0.97 to 0.99 across six regions, the lowest for Region 0, as it was with ML models. However, it is still higher than what Logistic, K-NN, Kernel SVM, Decision Tree, Random Forest, and XGBoost could provide. For the rest, there were occasional improvements in accuracy, but not for all regions. Therefore, GRU can provide better results for more volatile regions than traditional ML models.

Is deep learning (GRU) better? Indeed, we can observe an improvement in accuracy, most probably since for each forecast made DL can automatically include many more data instances from the past, and because the model is much more complex (has many more weights) than other considered models. However, this 'deepness' may mean that a DL model automatically makes physically irrelevant (lagged) inputs part of a model (on this, see e.g., Moreido *et al.* 2021). The disadvantage of GRU compared to other (non-deep) ML models is the significantly greater computational time and load. Other ML classification models have much simpler configurations and more transparent and physically explainable sets of inputs.

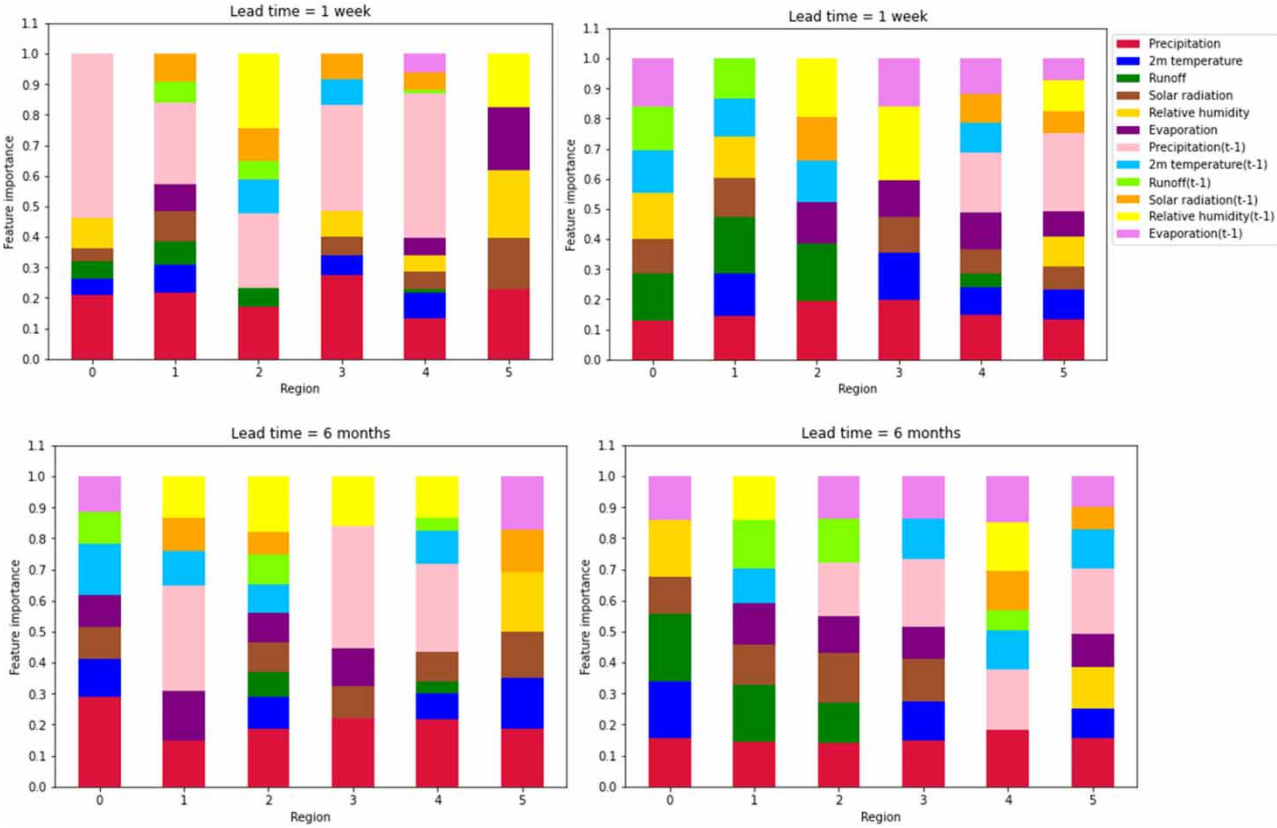


Figure 13 | The combinations of input variables for every region at lead time 1 week (top) and 6 months (bottom) and their feature importance for (a) SPI-based drought indices models; and (b) SPEI-based drought indices models.

4.1.4. Regression: forecast of SPI or SPEI

Since the previous analysis focused on drought or no-drought states, it is also an interest of the research to identify how effective ML would be to forecast SPI or SPEI values directly, instead of binary classes. The same family of ML algorithms can be used (since by design, all of them can be set up to solve regression problems). The results for the SPI/SPEI forecasting are shown in Figure 15.

As can be seen from Figure 15, the accuracy is quite high (around 0.8) for the lead time within 2 weeks for Regions 0–4 for SPI-based models (Region 5 is an outlier) while dramatically decreasing as the lead time increases and reaches months. This is no surprise since a regression problem is much more difficult than a binary classification problem. Interestingly, the accuracy of SPI-based models decreases more significantly than that of SPEI-based models. Better tuning may improve accuracy at a longer lead time. It would be also reasonable to test deep learning models, which are reportedly accurate for time series forecasting problems.

4.1.5. Forecasting the state change between drought and no-drought

It is critical to correctly differentiate the drought and no-drought conditions at the borderline case where SPI (or SPEI) is around the threshold (-1) to prove the reliability of the ML for drought forecasting. The probability of hitting the 'state change' case during testing was 6% for Regions 1–6 and 8% for Region 0. The comparison of the classification accuracy and F1 score results are shown in Figure 16 for the lead time being 1 week and 6 months for SPEI-based drought indices, as an example. As can be observed from the figure, the accuracy at state change (F1 score) is still relatively high (minimum 87%), although it is lower than the regular accuracy. It can be concluded that ML forecasting provides not only reliable results to classify 'drought' and 'no-drought' conditions but also shows decent outcomes for borderline cases.

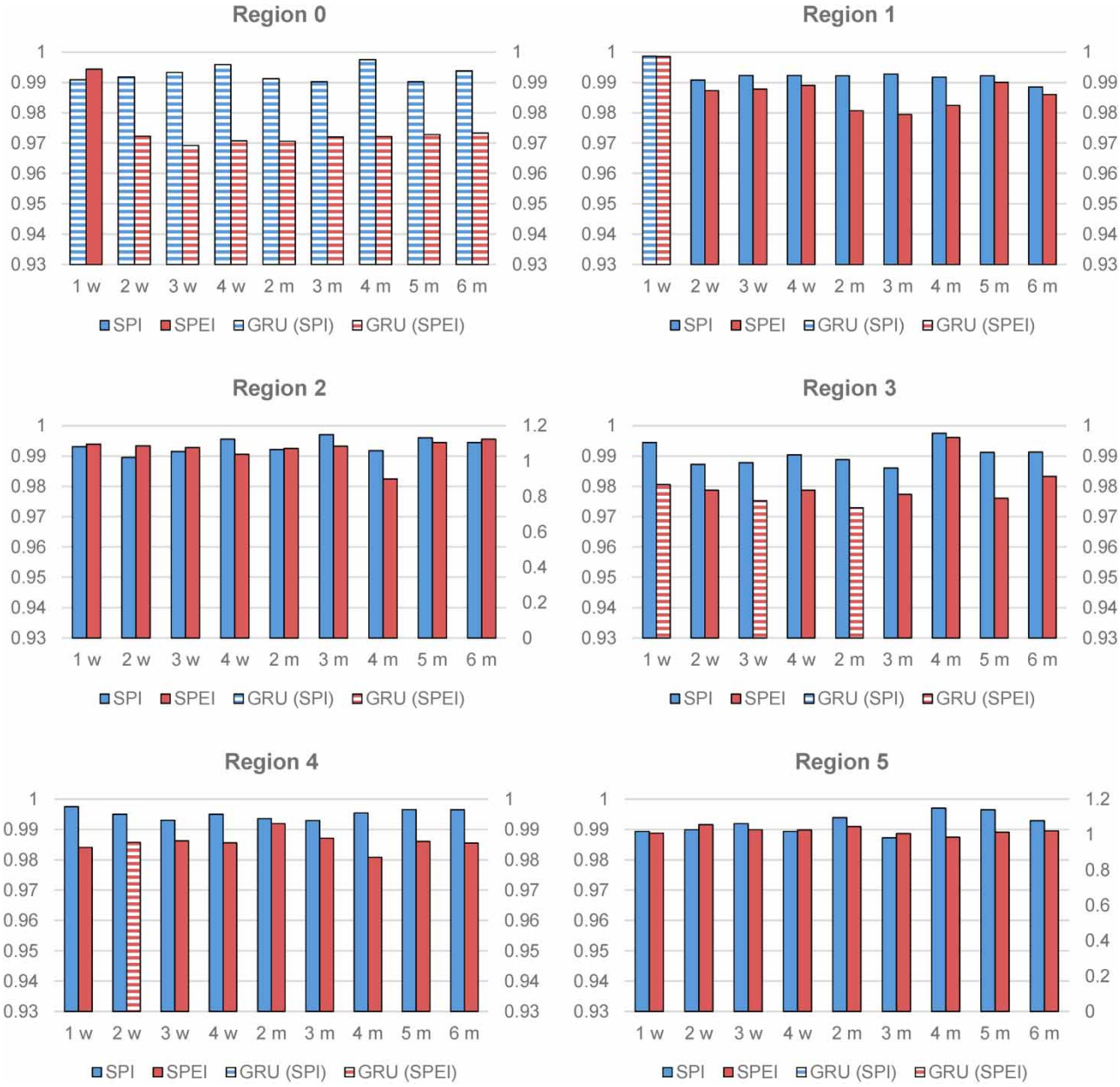


Figure 14 | The accuracy of the GRU model for SPI- and SPEI-based drought indices.

4.2. Stage 2. Considering the whole area

Although for combinations the accuracy was high (minimum 86% among six regions and two types of bases for drought indices for ML models and 97%+ for GRU models), there is still a question: how representative is 1 point for the relatively large area to characterize a state of drought? How homogeneous is the region? So, as was explained earlier, Stage 2 was undertaken, where all points of the area are considered. This section will discuss the PDA provided by the forecasting in every cell and the comparison of the forecasted results and observed drought cells to identify the level of accuracy.

4.2.1. Percentage of drought area

PDA for each time step and region was explored using the equation discussed in Section 2.3.1.2.1. Figure 17 shows the PDA evolution of the weekly period of January 1991–December 2021 or December 2018 for each region for SPI and SPEI-based drought indices, respectively. PDA gives us the severity of the drought at a particular time in a particular region.

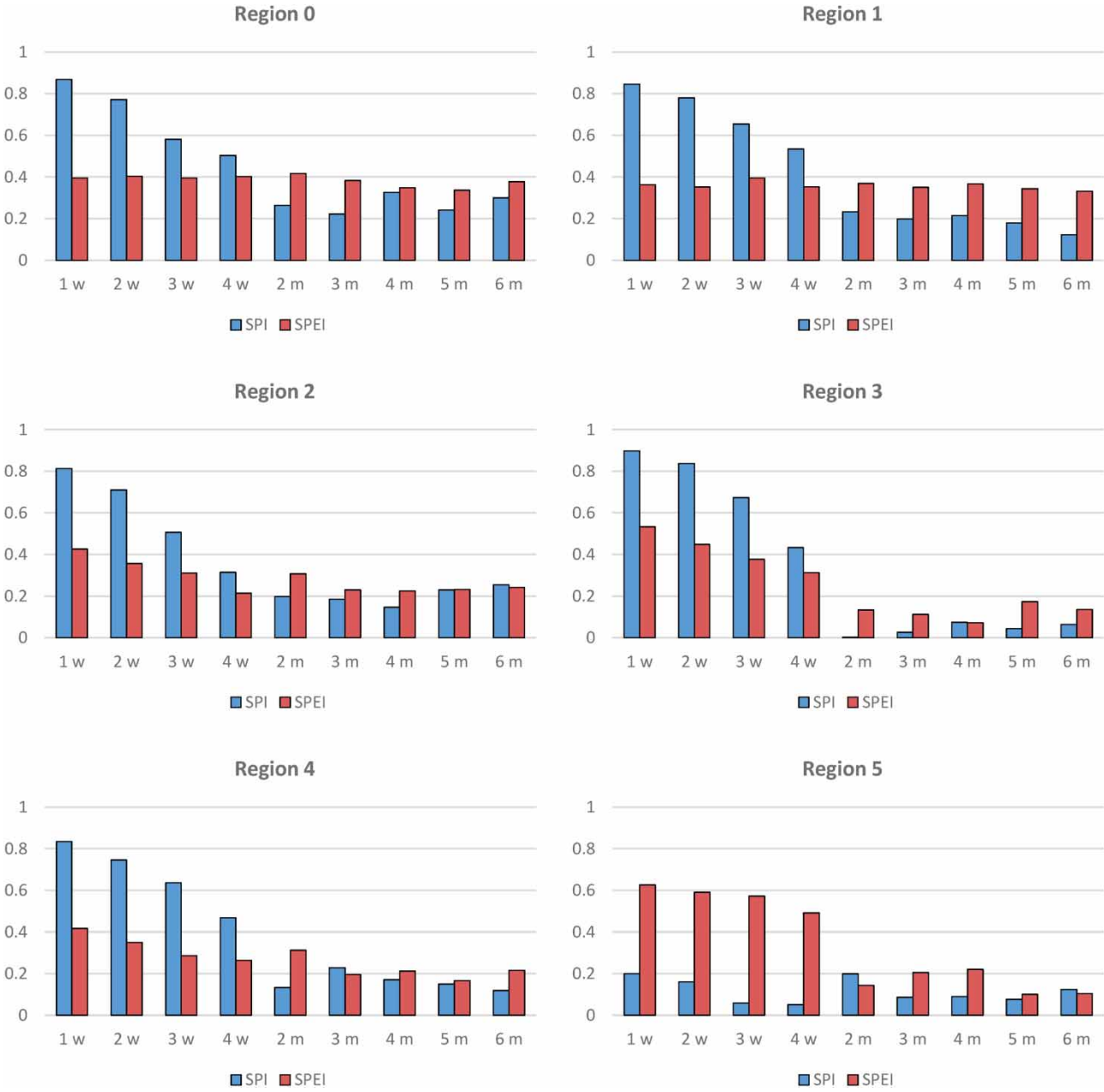


Figure 15 | The highest accuracy distribution of every input parameters combination among six ML regression models for SPI- and SPEI-based drought indices.

The PDA time series indicate the most drought-susceptible seasons for every region. As can be seen from Figure 17, SPI-based PDA has a pattern for every region except for Region 0 while it is chaotic for SPEI-based PDA. As discussed before, forecasting for SPEI is a more complicated problem due to a more entangled correlation between input and output parameters. Droughts covering up to 100% happened in different years at different months making the region drought susceptible at any time of the year. This is threatening since regional agriculture is concentrated around farming. Consequently, the water supply must be abundant for the entire year. On the contrary, as seen from the SPI-based PDA, Region 1 is most susceptible to drought during the end of the summer-fall seasons and has low PDA for the rest of the year with few outliers. This also creates a problem since this is the main cereal-producing region, and summer is the crop-producing season. The region historically relies on natural precipitation, which is getting less dependable as the climate is changing.

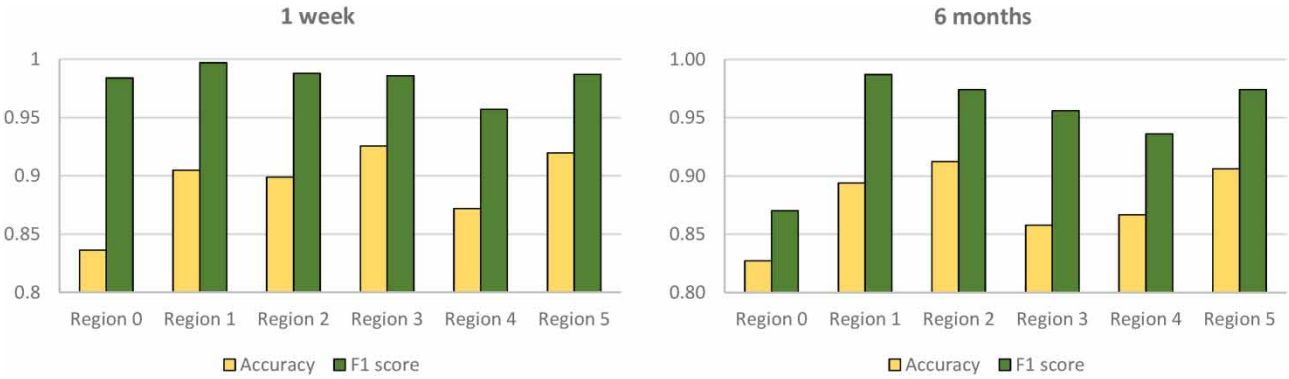


Figure 16 | Comparison of classification accuracy and F1 score for the borderline (state change) cases for SPEI-based models.

Region 2 is most susceptible during wintertime, which affects crop production indirectly such as not enough snow layer to protect the soil. However, this does not generate the same level of an issue as for Region 1. Region 3 shows a similar pattern as Region 0 (both are in the West part of Kazakhstan) with occasional drought area peaks at any time of the year. However, most of the time, summer is the most drought-susceptible season, which affects agriculture since the vast area is arable land. Another problem is the drying of the available freshwater resources: The Ural River and the Caspian Sea. This is a major concern of the region for the last few years. Region 4 has a similar pattern. The difference between Region 1 and 4 is that this region has irrigation (Karatal irrigation massif) and is more protected from drought. Region 5 is mostly affected during the winter and fall seasons.

4.2.2. Comparison to the reference and determination of the forecasting accuracy

The whole area analysis uses the best input parameters and ML classification model as per the results of Stage 1. The resulting overlaps of the forecasting output and observed reference for one date and lead time = 1 week, 4 weeks, and 6 months for SPI-based drought cells are shown in Figure 18(a)–18(c). As can be seen from Figure 18, the forecasting exaggerated the extent of the drought area (black cells) disregarding the duration of lead time. The accuracy of the forecast is more than 95% since the number of missed or false alarm cells (black cells) is relatively low. Therefore, involving all cells as input time series of the forecasting skyrockets the forecasting spatial reliability. However, the main limitation of this stage is that it requires more computational power. Another general limitation is the latency of the ERA5 dataset: the latency varies from 2 to 3 months. This issue can be resolved by choosing another database for the analysis.

4.3. Discussion

In the 21st century, the active utilization of machine and deep learning for drought forecasting has become increasingly prominent. Despite the rapid advancements in technology, our research focuses on crafting a finely tuned model by synthesizing existing models prevalent at the time. A key innovation in our approach lies in meticulous feature selection, where we explored 2,047 combinations of input variables, tested seven machine and deep learning classifiers, and two types of output variables. This exhaustive exploration resulted in the evaluation of over 1.5 million models to identify the most suitable configuration for six distinct regions across nine lead times.

Contrary to the common belief that maximizing input variables enhances model performance, our study reveals the potential risk of overtraining and its adverse impact on forecasting capability. To refine model accuracy, we employed hyperparameter tuning, an often overlooked facet in model development. Notably, we achieved an 87% accuracy in forecasting "borderline events," which are pivotal for identifying transitions from a non-drought to a drought state and determining drought duration.

While the ensemble forecasting approach, specifically utilizing the stacking method, significantly improved accuracy, detailed discussion on this aspect is beyond the scope of this work. Our ensemble strategy demonstrated heightened accuracy, particularly over a lead time of up to 6 months. The study recognizes the potential of ensembles to enhance accuracy and outlines avenues for future research, emphasizing the importance of nuanced model development strategies to optimize drought forecasting precision. These insights have implications for understanding and mitigating the impact of drought events.

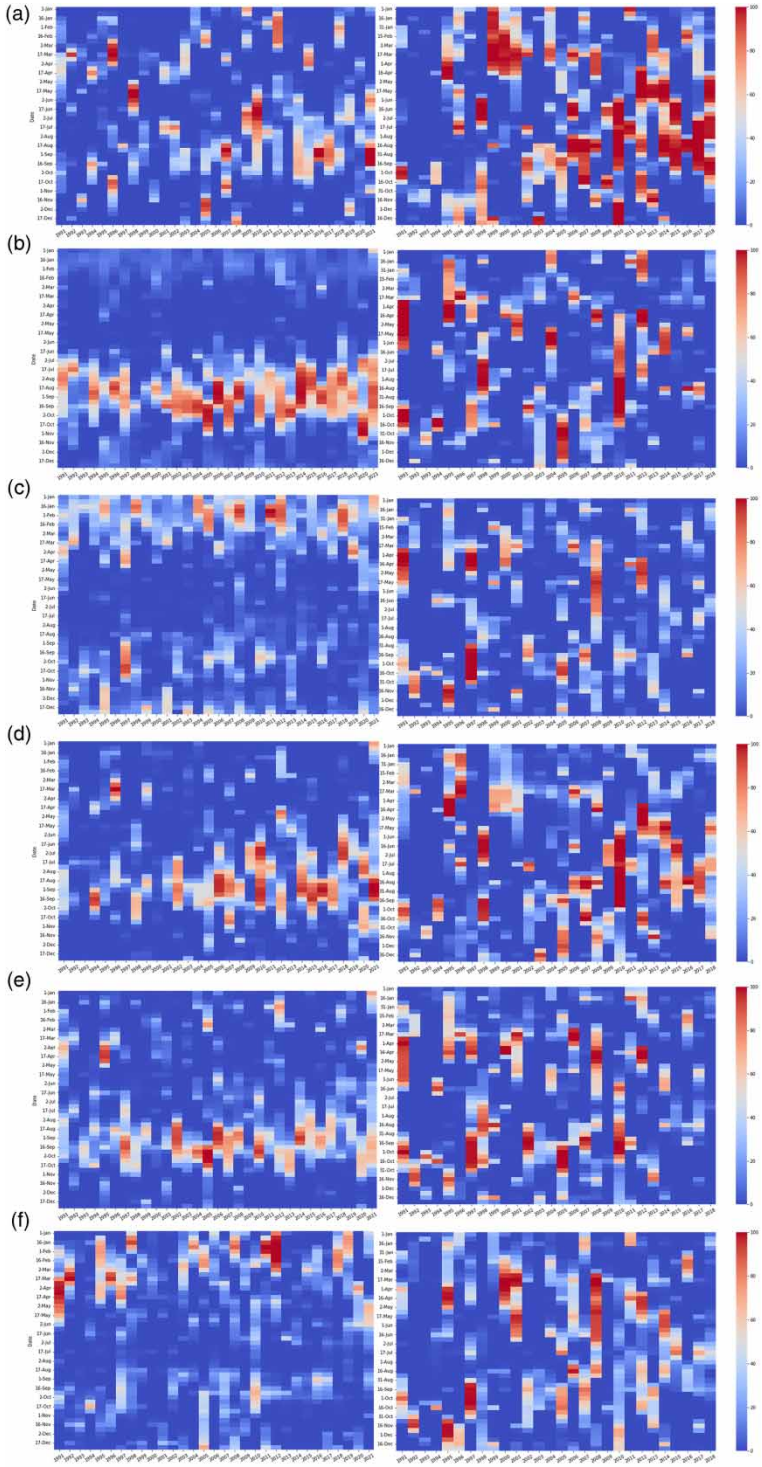


Figure 17 | PDA time series for 1991–2021 for SPI-based drought indices (left) and SPEI-based drought indices (right) for (a) Region 0; (b) Region 1; (c) Region 2; (d) Region 3; (e) Region 4; and (f) Region 5.

5. CONCLUSION

The main objectives of the research were to examine the performance of ML-based forecasting of spatiotemporal meteorological drought events in Kazakhstan employing extensive input variable selection. To analyze the spatiotemporal drought

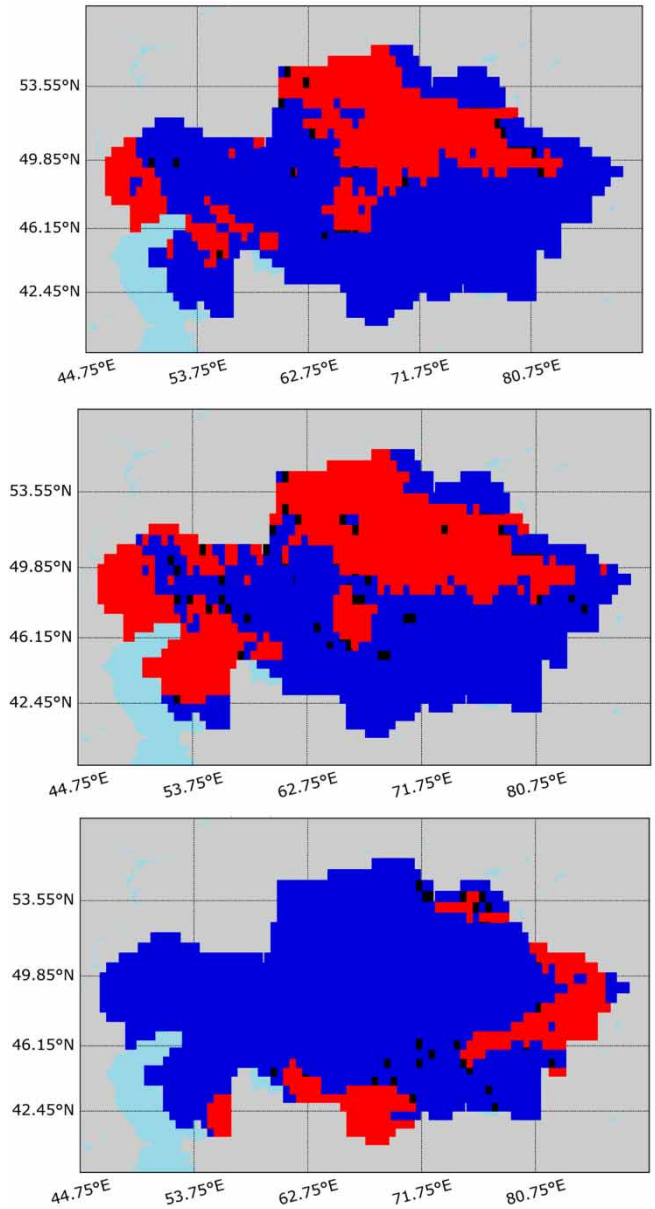


Figure 18 | Comparison of forecasted results and its reference at 13.07.2017 at the lead time (a) 1 week; (b) 4 weeks; and (c) 6 months. The extra drought cells forecasted are shown in black.

development and its characteristics utilizing large-scale gridded time series of hydrometeorological data, the NCDA methodology was utilized. This allowed for the identification of spatial extension and drought event occurrence in time. The work was performed by gradually increasing the complexity of the model. Therefore, two stages were used, from analyzing one reference point per region to the whole area of Kazakhstan. From analyzing the results, some key findings should be formulated:

- Although SPI was derived from precipitation, the combinations of input variables lead to a better result than when the precipitation was the only input. Therefore, it is recommended to use combinations of meteorological and hydrological variables instead of only precipitation for drought forecasting.
- The inclusion of lagged input variables for the previous time step (along with those for the current step) did not only increase the selection of the input combinations but also provided more accurate forecasting, increasing the accuracy of tuned models up to 99% for SPI-based and 94% for SPEI-based.

- GRU (a deep learning technique) performs somewhat better than 'shallow' ML models having an accuracy of up to 97% for SPEI-based models and 99% for SPI-based for a more volatile region (Region 0 where the severe drought occurred in 2021). This can be explained by the fact that for each forecast made DL is including data from the (deeper) past, and because the model is much more complex than 'shallow' learning models. However, this also means that a DL model takes physically irrelevant (far-lagged) inputs, which makes it less explainable than the 'shallow' ML models. DL models are also more complicated, and computationally demanding during training.
- ML techniques provided not only good results to classify 'drought' or 'no-drought' conditions but also delivered adequate results for the 'borderline events' at the state change (change from drought to no-drought and vice versa). This was obtained using an F1 score, having a minimum of 87% accuracy.
- Regardless of the complexity of the analysis, it was observed that SPI-based drought indices models performed better at shorter lead times while worse with increasing lead time, compared to SPEI-based. The core of the reason also lies in the different derivations of the indices. SPEI is better suited for monitoring agricultural and hydrological drought, which happens over a longer period than meteorological drought for which SPI is suited better.
- Generally, accuracy for both SPI and SPEI does not fall below 94%. It was found that all models typically overestimate – the drought area is exaggerated.
- Regression (numerical forecasting of SPI or SPEI indices) showed relatively high results for shorter lead times (1–2 weeks) but failed to achieve good results when lead time increased to months in advance.

For further research interest, it is suggested to explore the use of other deep learning techniques, committee models (weighted ensembles), multiclass drought classification, and explore dynamics of the drought cells and clusters.

DATA AVAILABILITY STATEMENT

All relevant data are included in the paper or its Supplementary Information.

CONFLICT OF INTEREST

The authors declare there is no conflict.

REFERENCES

ADB. 2021 Climate risk country profile: Kazakhstan. https://www.adb.org/sites/default/files/publication/722246/climate-risk-country-profile-kazakhstan.pdf (accessed 3 February 2022).

Beguería, S., Vicente-Serrano, S. M., Reig, F. & Latorre, B. 2014 Standardized precipitation evapotranspiration index (SPEI) revisited: parameter fitting, evapotranspiration models, tools, datasets and drought monitoring. *International Journal of Climatology* **34** (10), 3001–3023.

Breiman, L. & Cutler, R. A. 2001 Random forests machine learning [J]. *Journal of Clinical Microbiology* **2**, 199–228.

Brust, C., Kimball, J. S., Maneta, M. P., Jencso, K. & Reichle, R. H. 2021 Droughtcast: a machine learning forecast of the United States drought monitor. *Frontiers in big Data* **4**, 773478.

Cho, K., van Merrienboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H. & Bengio, Y. 2014 Learning phrase representations using RNN encoder–decoder for statistical machine translation. In: *('Encoder-Decoder Recurrent Neural Network Models for Neural Machine …') Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1724–1734.

Corzo Perez, G. A., van Huijgevoort, M. H. J., Voß, F. & van Lanen, H. A. J. 2011 On the spatio-temporal analysis of hydrological droughts from global hydrological models. *Hydrology And Earth System Sciences* **15** (9), 2963–2978.

Dikshit, A. & Pradhan, B. 2021 Explainable AI in drought forecasting. *Machine Learning with Applications* **6**, 100192.

Dubovyk, O., Ghazaryan, G., González, J., Graw, V., Löw, F. & Schreier, J. 2019 Drought hazard in Kazakhstan in 2000–2016: a remote sensing perspective. *Environmental Monitoring and Assessment* **191**, 1–17.

FAO. 2017 Infographic: Drought & Agriculture. https://www.fao.org/3/i7378e/i7378e.pdf (accessed 5 February 2022).

Friedl, M. A. & Brodley, C. E. 1997 Decision tree classification of land cover from remotely sensed data. *Remote Sensing of Environment* **61** (3), 399–409. doi: 10.1016/S0034-4257(97)00049-7.

Friedman, J. H. 2001 Greedy function approximation: a gradient boosting machine. *Annals of statistics* **29** (5), 1189–1232.

Ghobadi, F. & Kang, D. 2023 Application of machine learning in water resources management: a systematic literature review. *Water* **15** (4), 620.

Hu, B., Palta, M. & Shao, J. 2006 Properties of R2 statistics for logistic regression. *Statistics in Medicine* **25** (8), 1383–1395.

Jiang, S., Zheng, Y. & Solomatine, D. 2020 Improving AI system awareness of geoscience knowledge: symbiotic integration of physical approaches and deep learning. *Geophysical Research Letters* **47** (13), e2020GL088229.

Lipton, Z. C., Elkan, C. & Naryanaswamy, B. 2014 Optimal Thresholding of Classifiers to Maximize F1 Measure. In: Machine Learning and Knowledge Discovery in Databases. ECML PKDD 2014. Lecture Notes in Computer Science, vol 8725 (Calders, T., Esposito, F., Hüllermeier, E. & Meo, R., eds.). Springer, Berlin, Heidelberg, pp. 225–239.

Liu, C., Yang, C., Yang, Q. & Wang, J. 2021 Spatiotemporal drought analysis by the standardized precipitation index (SPI) and standardized precipitation evapotranspiration index (SPEI) in Sichuan Province, China. *Scientific Reports* **11** (1), 1–14.

Lynn, H. M., Pan, S. B. & Kim, P. 2019 A deep bidirectional GRU network model for biometric electrocardiogram classification based on recurrent neural networks. *IEEE Access* **7**, 145395–145405.

McKee, T. B., Doesken, N. J. & Kleist, J. 1993 The relationship of drought frequency and duration to time scales. In *Proceedings of the 8th Conference on Applied Climatology (Vol. 17, No. 22, pp. 179–183)*.

Molnar, C. 2020 *Interpretable Machine Learning. A Guide for Making Black Box Models Explainable*.

Moreido, V., Gartsman, B., Solomatine, D. P. & Suchilina, Z. 2021 How well can machine learning models perform without hydrologists? application of rational feature selection to improve hydrological forecasting. *Water* **13** (12), 1696.

Mubenga-Tshitaka, J. L., Muteba Mwamba, J. W., Dikgang, J. & Gelo, D. 2021 Risk spillover between climate variables and the agricultural commodity market in East Africa.

Mucherino, A., Papajorgji, P. J. & Pardalos, P. M. 2009 k-nearest neighbor classification. In: Data Mining in Agriculture. Springer. *Optimization and Its Applications, vol 34* (Mucherino, A., Papajorgji, P.J., Pardalos, P.M., eds.). Springer, New York, pp. 83–106.

Muñoz Sabater, J. 2019 ERA5-Land monthly averaged data from 1981 to present. Copernicus Climate Change Service (C3S) Climate Data Store (CDS). doi: 10.24381/cds.68d2bb3.

Otkin, J. A., Svoboda, M., Hunt, E. D., Ford, T. W., Anderson, M. C., Hain, C. & Basara, J. B. 2018 Flash droughts: a review and assessment of the challenges imposed by rapid-onset droughts in the United States. *Bulletin of the American Meteorological Society* **99** (5), 911–919.

Pannett, R. 2021 Horse graves on the steppes as Kazakhstan is battered by one of the worst droughts in living memory. *The Washington Post* https://www.washingtonpost.com/world/2021/08/09/horses-kazakhstan-heatwave-grave/ (accessed 10 January 2022).

Peel, M. C., Finlayson, B. L. & McMahon, T. A. 2007 Updated world map of the köppen-Geiger climate classification. *Hydrology and Earth System Sciences* **11** (5), 1633–1644.

Prodhan, F. A., Zhang, J., Hasan, S. S., Sharma, T. P. P. & Mohana, H. P. 2022 A review of machine learning methods for drought hazard monitoring and forecasting: current research trends, challenges, and future research directions. *Environmental Modelling & Software* **149**, 105327.

Shrestha, D. L. & Solomatine, D. P. 2006 Experiments with adaBoost. RT, an improved boosting scheme for regression. *Neural Computation* **18** (7), 1678–1710.

Speiser, J. L., Miller, M. E., Tooze, J. & Ip, E. 2019 A comparison of random forest variable selection methods for classification prediction modeling. *Expert Systems with Applications* **134**, 93–101.

Tursunova, A., Medeu, A., Alimkulov, S., Saparova, A. & Baspakova, G. 2022 Water resources of Kazakhstan in conditions of uncertainty. *Journal of Water and Land Development* **5**, 54–149.

UNESCAP. 2021 Kazakhstan – Climate Change and Disaster Risk Profile. https://www.unescap.org/sites/default/d8files/event-documents/Kazakhstan%20-%20Climate%20Change%20and%20Disaster%20Risk%20Profile.pdf (accessed 10 February 2022).

USDM. 2021 Map Archive | U.S. Drought Monitor. Droughtmonitor.Unl.Edu. https://droughtmonitor.unl.edu/Maps/MapArchive.aspx (accessed 16 February 2022).

Vapnik, V. N. 1999 An overview of statistical learning theory. *IEEE Transactions on Neural Networks* **10** (5), 988–999.

WCRP. 2017 Forecast Verification methods Across Time and Space Scales. https://www.cawcr.gov.au/projects/verification/ (accessed 14 April 2022).

XGBoost. 2023 https://github.com/dmlc/xgboost (accessed on March 4, 2023).

First received 5 September 2023; accepted in revised form 9 January 2024. Available online 24 January 2024