arXiv:2309.06212v6 [cs.LG] 12 Jul 2024

Long-term drought prediction using deep neural networks based on geospatial weather data

Alexander Marusov *, Vsevolod Grabar *, Yury Maximov , Nazar Sotiriadi , Alexander Bulkin , Alexey Zaytsev

Abstract

The problem of high-quality drought forecasting up to a year in advance is critical for agriculture planning and insurance. Yet, it is still unsolved with reasonable accuracy due to data complexity and aridity stochasticity. We tackle drought data by introducing an end-to-end approach that adopts a spatio-temporal neural network model with accessible open monthly climate data as the input.

Our systematic research employs diverse proposed models and five distinct environmental regions as a testbed to evaluate the efficacy of the Palmer Drought Severity Index (PDSI) prediction. Key aggregated findings are the exceptional performance of a Transformer model, EarthFormer, in making accurate short-term (up to six months) forecasts. At the same time, the Convolutional LSTM excels in longer-term forecasting.

Keywords: weather, climate, drought forecasting, deep learning, long-term forecasting

1. Highlights

- We improved quality for long-term, up to 12 months drought forecasting
- We adopted modern transformers and Convolutional LSTM to solve this problem

*Authors contributed equally

- We created an extensive test bed to evaluate models consisting of 5 diverse regions
- We reduced the gap to perfect ROC-AUC by 54% and 16%, respectively

2. Software and data availability

- Software name: Long-term drought prediction
- Developer: Vsevolod Grabar [repo creator, contribution], Alexander Marusov [contribution]
- Contact information: astralex98@gmail.com
- First year available: 2023
- Program language: Python
- Cost: free
- Software and data availability: [1]
- Repository storage: 120 MB

3. CRediT author statement

Alexander Marusov: Methodology, Software, Validation, Visualization, Writing – original draft, Writing – review & editing. **Vsevolod Grabar:** Software, Validation, Visualization, Writing – original draft, Writing – review & editing. **Yury Maximov:** Conceptualization. **Nazar Sotiriadi:** Conceptualization. **Alexander Bulkin:** Conceptualization. **Alexey Zaytsev:** Conceptualization, Funding acquisition, Methodology, Project administration, Resources, Supervision, Validation, Writing - Original Draft, Writing - Review & Editing.

[1] https://github.com/Astralex98/long-term-drought-prediction/tree/main

4. Introduction

The forecasting of droughts represents a critical challenge in climate science (Mohammed et al., 2022), as these natural phenomena incur substantial losses and can significantly impact populations and various economic sectors (Adikari et al., 2021). The importance of monitoring and predicting droughts is underscored by their frequent occurrence in diverse geographical landscapes (Ghozat et al., 2023). Moreover, the likelihood of droughts is expected to increase in the context of global climate change (Xiujia et al., 2022). Their accurate forecasting, however, is a complex problem due to the inherent difficulty in predicting the onset, duration, and cessation of drought events (Mishra and Desai, 2005). Another difficulty lies in choosing a drought index suitable for the goal being targeted.

We focus on long-term decision-making, which is critical for the annual planning of agricultural and insurance companies (Zhang et al., 2019). Formally, it is desired to provide accurate forecasts of droughts that extend 12 months into the future. A particular drought severity index for prediction is also important, as various indexes take into account diverse climatic factors, including temperature and precipitation. Among the drought severity indices, the Standardized Precipitation Index (SPI) (McKee et al., 1993) and the Palmer Drought Severity Index (PDSI) (Alley, 1984) stand out as fundamental measures. Another example of a modern drought index is the Standardized Precipitation-Evapotranspiration Index (SPEI) (Vicente-Serrano et al., 2010).

Our research utilizes the monthly PDSI for several compelling reasons. Firstly, given our extended forecast horizon of 12 months, PDSI is recognized as an effective tool for long-term assessment McPherson and Richman (2022). Additionally, PDSI's extensive historical record facilitates an analysis of the impacts of global warming within the broader climate dynamics Dai (2011). It is known that drought substantially can be divided into several types: meteorological, agricultural, hydrological, and socioeconomic (Hao et al., 2017). As our study concentrates on meteorological and agricultural droughts, leaving hydrological and socioeconomic issues out of scope, PDSI emerges as the most pertinent index, aligning well with the types of droughts we are investigating. Finally, we aim for a more interpretable forecast and quantifying PDSI values into the selected bins corresponding to different drought severity levels. So, instead of predicting PDSI directly (a regression task), we treat our task as a classification problem predicting bins and estimating the

probability of severe drought, which is the most interpretable quantity for decision-makers.

There are numerous approaches already available to solve the problem at hand. As a natural climate phenomenon, drought could be evaluated with the help of various climate models. General Circulation Models (GCMs) are powerful modern methods for climate event prediction, utilizing partial differential equations to simulate Earth's systems (Jiang et al., 2021). GCMs typically model the Earth's climate with a three-dimensional grid of spatial resolution between 250 and 600 km, 10 to 20 layers in the atmosphere, and up to 30 layers in the oceans. Wang and Chen (2014) used 35 GCMs from Coupled Model Intercomparison Project Phase 5 (CMIP5) to estimate increasing drought frequency in China. Similarly, Song et al. (2022) compared SPI and SPEI drought indices in South Korea using CMIP5 and CMIP6 GCMs, revealing varied forecast reliability over different time frames. While GCMs offer comprehensive global climate insights over decades-long horizons, their inherent uncertainties, especially in predicting extreme events, and lower resolution limit their effectiveness for regional climate analysis.

An alternative is machine learning algorithms for drought forecasting tasks (Prodhan et al., 2022). Classical approaches, such as stochastic models like AutoRegressive Integrated Moving Average (ARIMA) and multiplicative Seasonal AutoRegressive Integrated Moving Average (SARIMA), have been effectively utilized for drought prediction using the SPI index Mishra and Desai (2005). In another instance, multivariate regression, incorporating historical PDSI data and other global climate indices, was employed to forecast the PDSI index in South-Central Oklahoma (McPherson and Richman, 2022). While these methods predominantly address regression problems, we focus on classification. Logistic regression has been applied for binary drought classification using the SPI index in Niaz et al. (2021), demonstrating its suitability for label prediction tasks. Also, the gradient boosting algorithm has emerged as a powerful tool in modeling geospatial data Proskura et al. (2019), Koldasbayeva et al. (2022), especially effective in handling the imbalanced datasets often encountered in drought prediction (Kozlovskaia and Zaytsev, 2017). Notably, this algorithm has also remarkably succeeded in classifying drought conditions in Turkish regions (Danandeh Mehr, 2021).

Addressing the needs of practitioners, deep learning emerged as a viable tool for drought forecasting. Mishra and Desai (2006) introduced the use of recursive and direct multistep neural networks, leveraging the SPI index. Among time-series data approaches, Recurrent Neural Networks (RNN)

(Rumelhart et al., 1985) stand out, frequently outperforming traditional methods in time-series analysis (Hewamalage et al., 2021). Specifically, Long Short-Term Memory (LSTM) networks (Schmidhuber and Hochreiter, 1997) have been shown to surpass ARIMA in long-term SPI index forecasting, although ARIMA remains competitive in short-term prediction (Poornima and Pushpalatha, 2019). To harness the strengths of both ARIMA and LSTM, Xu et al. (2022) proposed a hybrid ARIMA-LSTM model.

However, these methods primarily focus on historical (temporal) data, neglecting the spatial aspects. Addressing both temporal and spatial dependencies, the ConvLSTM method has been applied to various fields, including precipitation prediction (Shi et al., 2015), earthquake prediction (Kail et al., 2021), and was notably used by Park et al. (2020) for short-term (eight-day) drought forecasting using satellite-based indices like Scaled Drought Condition Index (SDCI) and SPI. The emergence of Transformer architectures, originally developed for Natural Language Processing (NLP) (Vaswani et al., 2017), led to new models in diverse domains such as Computer Vision (CV) (Zheng et al., 2021). This makes their adoption of spatiotemporal modeling a compelling choice. Prominent examples include EarthFormer (Gao et al., 2022) and FourCastNet (Pathak et al., 2022), which excel in various spatiotemporal tasks, including regression challenges like precipitation nowcasting, while these approaches can suffer from low amount of available training data and high stochasticity of a target in a long-term forecasting problem. In our research, we have adapted these advanced Transformer architectures to enhance drought forecasting, aiming to leverage their capabilities in handling complex spatiotemporal data.

In the paper, we build spatio-temporal models for PDSI index forecasting, solving the long-term drought prediction problem and obtaining the following key contributions:

1. Comprehensive study of diverse spatio-temporal models. We adapted advanced deep-learning methods from different domains, benchmarking the most prominent options, including transformer-based models EarthFormer and FourCastNet. Also, classic approaches (logistic regression and gradient boosting) were constructed to account for both spatial and temporal dependencies. According to our knowledge, we are the first to adopt the above-mentioned spatio-temporal neural network models for long-term drought forecasting within the PDSI index with notable quality of the developed models.

2. Development of a test bed encompassing five distinct global regions for objective model evaluation with publicly accessible data. Together with a wide range of compared models, this research provides a systematic answer to the question of how one should predict droughts one to 12 months ahead and do we need deep learning to provide accurate forecasts.
3. Identification of our neural networks based on EarthFormer as optimal for medium-term forecasting up to six months and our variant ConvLSTM for long-term predictions. Thus, we make transformers work for a relatively small amount of input and training data. The input data for our models are easy to obtain and preprocess, making the model straightforward to run and more robust compared to elaborated preprocessing and feature engineering used in previous studies. While logistic regression and gradient boosting are enough in short-term forecasting, deep learning methods excel in four-month and longer time frames, which is crucial for decision-making.
4. Consistent improvement of models for horizons ranging from 1 to 12 months. This focus contrasts with previous studies, which typically addressed either very short-term predictions (up to a month) or much longer-term forecasts spanning decades.

5. Data

In this section, we consider the definition of the selected target variable PDSI, drought classification based on its values, and the characteristics of input features for model prediction.

The Palmer Drought Severity Index (PDSI) is a standardized index where absolute values above 4 indicate extreme conditions, and intermediate numbers are divided into bins and assigned to various wet or dry environments, where the latter corresponds to negative PDSI, see Table 1. It is calculated using a version of the Palmer formula, which combines reference evapotranspiration, precipitation, and a static soil water-holding capacity layer.

We have used publicly available geospatial data from Google Earth Engine (Gorelick et al., 2017). Specifically, to obtain the PDSI data, we employed the TerraClimate Monthly dataset [2]. Our PDSI data encompasses

[2] `https://developers.google.com/earth-engine/datasets/catalog/IDAHO_EPSCOR_TERRACLIMATE`

PDSI value	Drought severity class
4.00 and above	Extreme wet spell
3.00-3.99	Severe wet spell
2.00-2.99	Moderate wet spell
1.00-1.99	Mild wet spell
-1.00 to 0.99	Normal
-1.00 to -1.99	Mild dry spell
-2.00 to -2.99	Moderate dry spell
-3.00 to -3.99	Severe dry spell
-4.00 and below	Extreme dry spell

Table 1: Classification of various PDSI values, source: Liu et al. (2015)

a comprehensive range of climatic values from 1958 to 2022, covering the Earth's entirety. To test the consistency of the considered models, we examined regions across continents and climate zones: from the state of Missouri to Poland to India. The considered regions are depicted in Figure 1, and their characteristics are shown in Table 2.

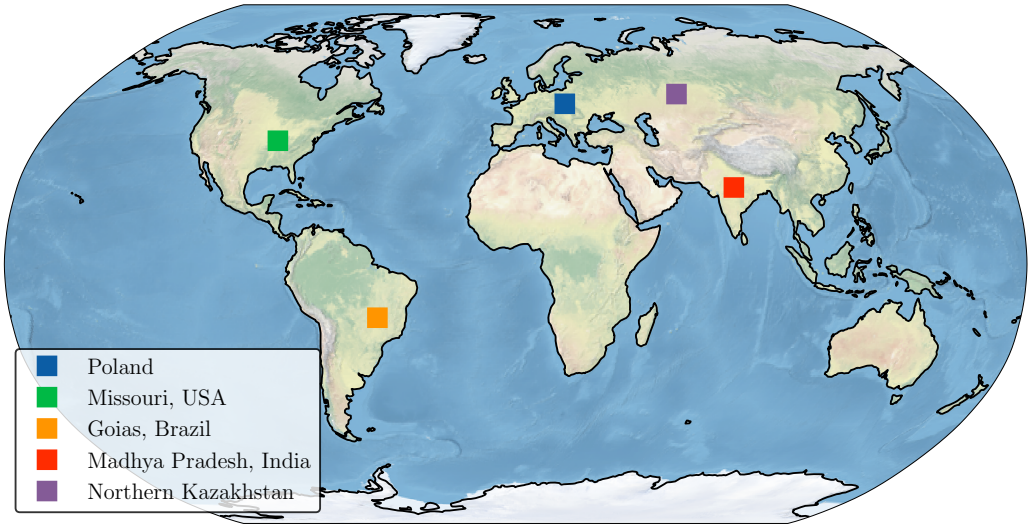


Figure 1: Diverse regions chosen for PDSI forecast

The input data, downloaded in a tif format, were transformed into a 3D tensor containing PDSI values. This tensor's structure includes one timescale dimension (monthly intervals) and two spatial dimensions (x and y

coordinates) representing the grid. The regional datasets exhibit varied resolutions, with grid dimensions ranging from 30 x 60 to 128 x 192 cells, allowing for varying degrees of granularity and spatial detailing. The spatial resolution for a cell in the TerraClimate data is approximately 5 km.

Region	Span, months	% of normal PDSI ≥ -2	% of drought PDSI ≤ -2	Spatial sizes km×km
Missouri, USA	754	74.91	25.09	416×544
Madhya Pradesh, India	754	70.66	29.34	512×768
Goias, Brazil	754	68.97	31.03	640×640
Northern Kazakhstan	742	68.70	31.30	256×480
Poland	742	66.28	33.72	352×672

Table 2: Regions' summary statistics

6. Methods

We compare deep learning approaches, including Convolutional LSTM and novel transformer models, such as FourCastNet from Nvidia and Earth-Former from Amazon, with classic methods, including the baseline model, gradient boosting, and logistic regression.

6.1. Baseline

As a global baseline and a coherence check, we took the most prevalent class from the training data as the prediction and compared it with actual targets from the test subset. We also checked a rolling baseline (i.e., the most frequent class from recent history, from 6 to 24 months). Still, the results were almost indistinguishable from the global baseline, so we did not include them in our tables and graphs.

6.2. Basic methods: Logistic regression and Gradient boosting

Both logistic regression and gradient boosting cannot work with raw data. Therefore, we created a data module that treated each grid cell as an individual value and transformed our task into a typical time series forecasting problem. To benefit from spatial correlations, we incorporate values from neighboring cells. For example, if we consider a 3x3 cell neighborhood, this includes eight additional time series. It is important to note that for "edge" cells, some of the neighboring cells may contain all zeros due to data limitations.

Logistic Regression. Logistic regression is usually the natural choice for tasks with linear dependence or as a baseline model. The novel research Zeng et al. (2023) shows that time series linear models are a good choice.

Gradient boosting. We adopted the gradient boosting of decision trees, implemented using the well-established library XGBoost Chen and Guestrin (2016). XGBoost is renowned for its speed and efficiency across a wide range of predictive modeling tasks and has consistently been favored by data science competition winners. It operates as an ensemble of decision trees, with new trees aimed at rectifying errors made by existing trees in the model. Trees are successively added until no further improvements can be made.

6.3. Convolutional LSTM (ConvLSTM)

Our model, inspired by Kail et al. (2021), modifies the Convolutional LSTM architecture (Shi et al. (2015)), blending Recurrent Neural Networks (RNNs) with Long Short-Term Memory (LSTM) to capture temporal dependencies. This adaptation involves extending LSTM's traditional one-dimensional hidden states to two-dimensional feature maps, facilitating grid-to-grid transformations essential for spatial tasks. We process PDSI grids varying from 50×50 to 200×200 cells, integrating Convolutional Neural Networks (CNNs) for spatial analysis. This combination, depicted in Figure 2, leverages both RNN and CNN strengths, simultaneously capturing temporal and spatial information in the architecture for drought prediction.

Details of architecture. Convolutional LSTM follows the pipeline:

1. Represent data as a sequence of grids: For each cell, we specify a PDSI value for a particular month; the input grid at each time moment has dimension $grid_h \times grid_w$ (varying from 50×50 to 200×200 for different regions of interest).
2. Pass the input grid through a convolutional network to create an embedding of grid dimensionality with 16 channels. As an output of LSTM at each time moment, we have a hidden representation (short-term memory) of size $hidden \times grid_h \times grid_w$, cell (long-term memory) representation of a similar size, and the output of size $hidden \times grid_h \times grid_w$.
3. Transform the output to the size $1 \times grid_h \times grid_w$ using convolution 1×1 to receive probabilities of the drought for each cell as a final prediction or to $k \times grid_h \times grid_w$ in case of multiclass classification, where $k > 2$ is the number of classes of drought condition that we are trying to predict.

As an additional hyperparameter, we vary the forecasting horizon (to forecast PDSI for the next month or f-th month).

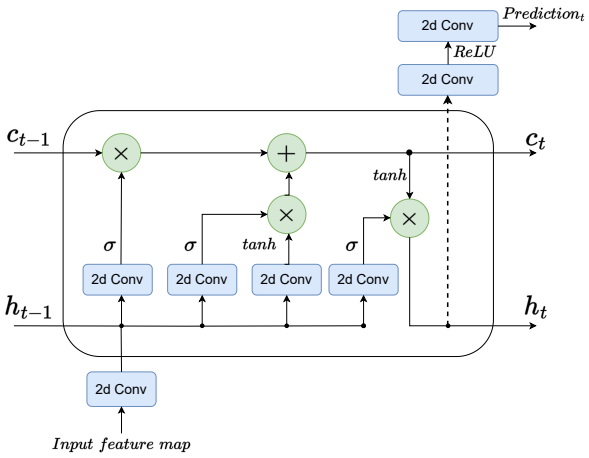


Figure 2: Our version of ConvLSTM architecture

6.4. Transformer-based methods

We consider two recent transformer-based models, FourCastNet and Earthformer, adopted for usage in spatiotemporal problems such as drought prediction.

FourCastNet. Fourier ForeCasting Neural Network (FourCastNet) was developed by Pathak et al. (2022) as a weather forecasting model and a part of NVIDIA's Modulus Sym deep learning framework for solving applied physics tasks. This model combines Adaptive Fourier Neural Operator (AFNO) from Guibas et al. (2022) with Vision Transformer. The model is computationally and memory efficient, with low spatial mixing complexity of $O(N \log N)$, where N is the sequence length. The authors produced high-resolution short-term wind-speed and precipitation forecasts in the original papers. We modified the last layer of the model, switching it from a regression to a classification task and evaluating it for the long-term forecasting problem with a significantly smaller amount of available data.

EarthFormer. Vanilla transformers have $O(N^2)$ complexity, and consequently, it is hard to apply them to spatiotemporal weather data because of their large dimensionality. Gao et al. (2022) suggest using the "divide and conquer" method: they divide the original data into non-intersecting parts (called cuboids) and use a self-attention mechanism on each cuboid separately in parallel. Such an approach allows significantly reduced complexity, bridging the gap between transformers and CNNs. The authors introduced EarthFormer for regression tasks, and we adopted it for a classification problem in a way similar to FourCastNet.

6.5. Technical details

Hyperparameters of deep learning models are mostly taken from corresponding original papers and can be found in our GitHub. This section in Table 3 presents some necessary optimization settings for all our deep learning methods.

	Optimization			
Models	Epochs num (average)	Batch	Optimizer	Learning rate scheduler
ConvLSTM	70	8	Adam	absent
EarthFormer	20	16	AdamW	Cosine
FourCastNet	30	8	Adam	Cosine

Table 3: Optimization characteristics

7. Results

Formal problem statement. We define a binary classification for drought forecasting using a PDSI threshold of -2, aligning with McPherson and Richman (2022) and PDSI bin categorizations presented in Table 1. The model's objective is to predict drought occurrences, classified as serious, when PDSI falls below this threshold.

For validation, we divided each local dataset (that corresponds to one of the five regions including Missouri, Northern Kazakhstan, Madhya Pradesh, Eastern Europe, and Goias) into training (70%) and testing (30%) subsets. The splitting uses out-of-time validation where the test set follows the training set chronologically. Model training and subsequent quality metric evaluations are conducted using these distinct data subsets.

Evaluation procedure. We use $ROC\,AUC$, $PR\,AUC$ and $F1$ scores to evaluate the model. During validation for early stopping and hyperparameter optimization, we chose the $ROC\,AUC$ score. All these scores are the medians of all spatial predictions (because we want to remove the impact of outliers), as we compute a temporal vector at every spatial prediction cell. Next, we receive a single score for each cell, so we end up with a grid of metrics. Finally, we get the median of scores at each set. Higher values for all scores correspond to better models. $ROC\,AUC$ scores have the perfect value of 1 and the value for a random prediction of 0.5.

Compared methods. For our main experiments, we explored the baseline's performance (the most frequent class from historical data), gradient boosting (XGBoost), logistic regression, and our modifications of ConvLSTM, FourCastNet, and EarthFormer.

XGBoost and logistic regression are two basic algorithms often used as strong baselines. The last three are variants of neural networks that performed strongly in various geospatial problems. They represent two dominating architectures in geospatial modeling: ConvLSTM is a combination of recurrent and convolutional neural networks; FourCastNet and EarthFormer are Transformers.

7.1. Main results

Analysis of results. The primary results are depicted in Figure 3. Our findings indicate that EarthFormer outperforms other approaches for shorter horizons. In particular, EarthFormer reached $ROC\,AUC$ score of 0.95 for a one-month prediction. But EarthFormer falls short of ConvLSTM at longer horizons of 9-12 months. The ConvLSTM model showed the second-best result (after EarthFormer), achieving an impressive $ROC\,AUC$ score of 0.9 for a one-month prediction. Notably, ConvLSTM exhibits a gradual decline in performance, reaching $0.6 - 0.65$ for longer forecasting horizons (ranging from 9 to 12 months) but nevertheless beating all other models. The standard gradient boosting approach initially yields a similar $ROC\,AUC$ score of 0.9 but sharply drops to 0.5 as the forecasting horizon is extended.

Additionally, we present the results for six-month predictions by regions in Figure 5, where we show the variation in scores for different geographies across models.

Why does transformer fail in long-term prediction? Our assumption for such behavior is the permutation-invariance of the attention mechanism. De-

spite positional encoding, transformers cannot effectively extract temporal information from long input sequences. Since a long input sequence is essential for long-term forecasting, transformers do not show good results. Similar results for different time-series tasks were observed in Zeng et al. (2023). ConvLSTM naturally extracts temporal information via LSTM.

Why is logistic regression impressive? Since gradient boosting results are almost identical to those of logistic regression, we discuss only logistic regression performance. First, the power of linear models was already shown in Zeng et al. (2023), where they beat modern transformer architectures on almost all datasets. In our experiments, linear models are worse than other models on *long-term* prediction, but on a *short-term* scale, we can see comparable results. We tried different history lengths, but our results show that taking only the element nearest to the future horizons is sufficient. Our intuition is that the nearest future predictor variables are closely (particularly linearly) related to the history element. For example, PDSI in July is close to PDSI in August but far away from December. Hence, linear models are good at *short-term* predictions but poor at *long-term* forecasting.

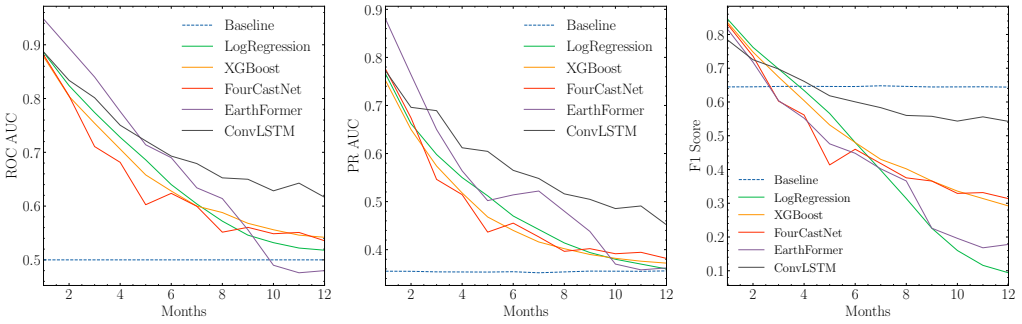


Figure 3: Quality metrics for the binary drought severity classification: median ROC AUC, PR AUC, and F1 for different forecast horizons averaged over five considered regions

7.2. Predictions and errors for a particular region

To assess the performance of the Convolutional LSTM algorithm (which proved to be the most stable and promising for long-term drought forecasting), we focused on the region of Missouri, where we ran several ablation studies. To illustrate, the spatial distribution of *ROC AUC* scores is depicted in Figure 4. Notably, we observed a non-uniform distribution of *ROC AUC* values across the cells within the region. The standard deviation of the scores

Horizon, months	1	3	6	9	12
Median ROC AUC:					
Baseline	0.5	0.5	0.5	0.5	0.5
LogRegression	0.886	0.774	0.640	0.546	0.518
XGBoost	0.878	0.754	0.628	0.568	0.542
FourCastNet	0.881	0.711	0.624	0.561	0.536
EarthFormer	**0.948**	**0.840**	0.690	0.556	0.480
ConvLSTM	0.887	0.802	**0.693**	**0.650**	**0.617**
Median PR AUC:					
Baseline	0.355	0.354	0.354	0.355	0.356
LogRegression	0.766	0.598	0.470	0.394	0.360
XGBoost	0.752	0.574	0.44	0.39	0.372
FourCastNet	0.776	0.546	0.455	0.402	0.382
EarthFormer	**0.880**	0.650	0.514	0.438	0.362
ConvLSTM	0.772	**0.689**	**0.565**	**0.505**	**0.452**
Median F1:					
Baseline	0.645	0.646	**0.646**	**0.645**	**0.644**
LogRegression	**0.846**	**0.698**	0.480	0.226	0.094
XGBoost	0.836	0.674	0.480	0.366	0.292
FourCastNet	0.831	0.603	0.460	0.366	0.314
EarthFormer	0.816	0.604	0.448	0.226	0.178
ConvLSTM	0.784	**0.698**	0.600	0.558	0.543

Table 4: Median Metrics vs. Forecast Horizon, binary classification; best values are in bold, second best are underlined

is substantial, and individual values range from close to random predictors ($ROC\ AUC = 0.6$) to near-perfect scores approaching 1.0. This variability highlights the diverse predictive capabilities of our algorithm across different spatial locations within Missouri.

7.2.1. Performance Evaluation for Cropped Region

Description of experiment. As is typical with *ROC AUC* maps, the worst predictions are found on the edges and some corners. We have observed that this behavior is consistent regardless of the region being studied. Consequently, making predictions for a larger region and cropping the desired region of interest may be advantageous. We have conducted a study to test this hypothesis for Figure 4, and the results are shown in Table 6.

Region	Northern Kazakhstan	Poland	Madhya Pradesh	Goias	Missouri
Median ROC AUC:					
Baseline	0.5	0.5	0.5	0.5	0.5
LogReg	<u>0.60</u>	0.63	0.61	<u>0.61</u>	**0.75**
XGBoost	0.59	0.64	0.59	<u>0.61</u>	0.71
FourCastNet	<u>0.60</u>	0.55	0.65	<u>0.61</u>	0.69
EarthFormer	0.54	**0.69**	**0.84**	**0.65**	<u>0.73</u>
ConvLSTM	**0.71**	<u>0.68</u>	<u>0.71</u>	0.60	**0.75**
Median PR AUC:					
Baseline	0.37	0.43	0.35	0.39	0.23
LogReg	0.46	**0.67**	0.24	**0.55**	0.43
XGBoost	0.46	0.63	0.20	<u>0.54</u>	0.37
FourCastNet	0.46	0.55	0.36	<u>0.54</u>	0.37
EarthFormer	<u>0.47</u>	0.22	**0.77**	0.52	**0.59**
ConvLSTM	**0.55**	<u>0.65</u>	<u>0.57</u>	<u>0.54</u>	<u>0.50</u>
Median F1:					
Baseline	<u>0.63</u>	0.57	<u>0.65</u>	**0.61**	**0.77**
LogReg	0.42	<u>0.62</u>	0.35	0.41	0.60
XGBoost	0.45	0.58	0.30	0.54	0.53
FourCastNet	0.42	0.39	0.46	0.51	0.52
EarthFormer	0.03	0.18	**0.80**	<u>0.55</u>	<u>0.65</u>
ConvLSTM	**0.65**	**0.64**	0.64	0.52	0.56

Table 5: Median Metrics vs. Region, binary classification, six-month horizon; best values are in bold, second best are underlined

Percent of map cropped	0	10	20	30	40
median ROC AUC	0.7525	0.7592	0.7665	0.7749	0.7834
Percent of map cropped	50	60	70	80	90
median ROC AUC	0.7886	0.7899	0.7880	0.7838	0.7825

Table 6: ROC AUC score vs. crop percentage, six-month forecast, ConvLSTM model for Missouri

Analysis of results. Based on this experiment's findings, we deduce that cropping approximately 40-50% of the initially selected region maximizes our score. In other words, choosing a region that is initially 1.6-2 times larger than our target region is advisable. However, the precise amount of zoom required for optimal results must be determined through further experiments.

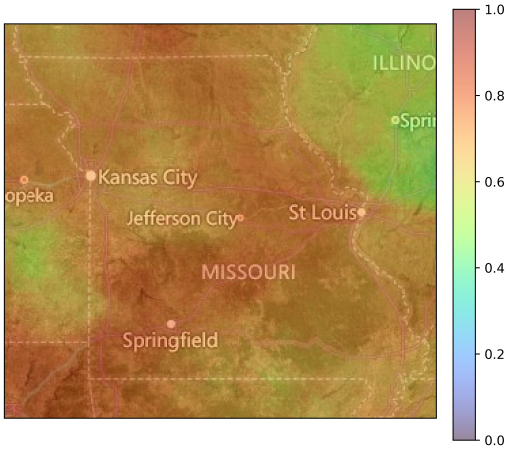


Figure 4: Spatial distribution of ROC AUC for 6 month forecast, Missouri, ConvLSTM

Next, we conducted several similar experiments to investigate how the model predictions change with the decrease in the total squared area. We took the same geographic region, Missouri, and examined various combinations of history length and forecast horizon. We trained a new model for each history length variant, forecast horizon, and region area (varying from the entire state to about a quarter of it).

The results are summarized in Table 7, and they are presented in more detail in Tables 9, 10, 11, and 12. We observed that predicting a smaller area with a pretrained model from a larger area generally works better. However, the degree of improvement is marginal, usually not exceeding 0.5-1%. Figure 5 presents the evolution of spatial maps.

Region area	100%	75%	53%	27%
$h = 6,\ f = 1$	0.9292	0.9466	0.9432	0.9326
$h = 12,\ f = 3$	0.8746	0.8525	0.8595	0.8710
$h = 9,\ f = 6$	0.7525	0.6901	0.7597	0.7544
$h = 24,\ f = 12$	0.6117	0.5890	0.6126	0.6264

Table 7: ROC AUC score vs zoomed region area (h - length of input history, f - forecast horizon)

7.2.2. Standard deviations and ensembling

Description of experiment. To assess the consistency of our findings, we repeated the experiment of PDSI binary classification for Missouri with five random seeds for each configuration of history length and forecast horizon. In addition, we tested an averaged ensemble of these five trained models.

Analysis of results. The results for horizons of 1, 3, 5, 6, 9, and 12 months can be found in Figure 6. Ensemble scores are denoted by the red crosses on these plots. Overall, the numbers vary, and there is no definitive optimal choice of history length for any horizon. However, we observed that extreme values yield better performance, such as the shortest or longest history lengths (e.g., 6, 9, 21, and 24 months). Notably, the averaged ensemble of models outperforms the individual underlying models in most cases. We note that this effect is more sound for neural networks, as they provide more diverse predictions, even if there are no differences in the architecture, and the only difference is a starting point for training (Fort et al., 2019). On the contrary, the logistic regression ensemble is similar to a single model, as the optimization problem is convex (Bishop and Nasrabadi, 2006), and gradient boosting is an ensemble per se (Danandeh Mehr, 2021). The obtained ensembles can also be used to access the uncertainty of predictions by machine learning models, improving the decision-making process (Jain et al., 2020).

8. Conclusion

Droughts, increasingly severe and frequent due to climate change, pose significant threats to agriculture and public health. The summer of 2022 in the Northern Hemisphere highlighted the urgency of these issues. Our research focuses on improving drought forecasting, a crucial step in mitigating the adverse impacts of these natural disasters. To tackle the task, we employed various models (from classic models to modern transformers) and many distinct regions to test their performance.

We succeeded in providing a better model suitable for agricultural decision-making and insurance applications. Our variant of EarthFormer shows the best result in *short-term* forecasting. In one-month prediction *ROC AUC* score is 0.948. Our variant of ConvLSTM is much better than other models in *long-term* forecasting, achieving an impressive *ROC AUC* score of 0.617 in twelve-month prediction. The metric values above are much higher than classic approaches: we significantly reduced the gap between perfect *ROC AUC*

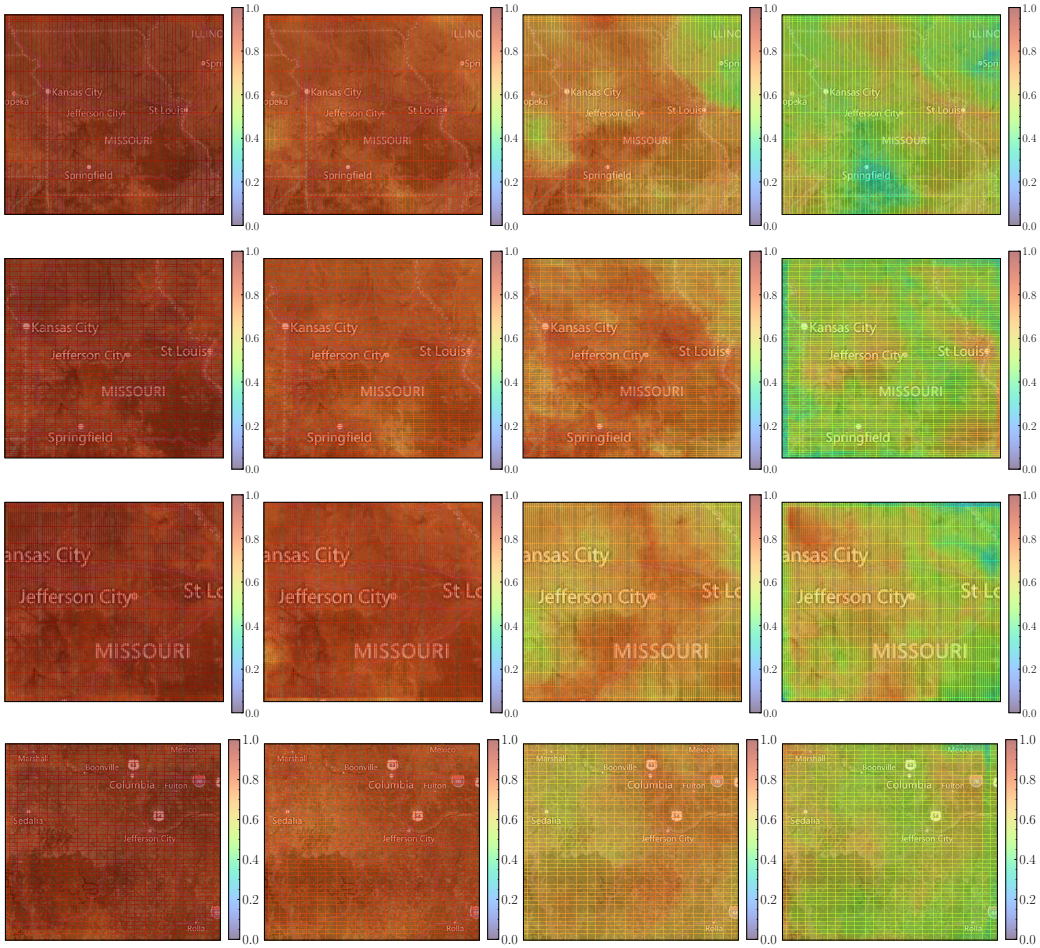


Figure 5: Evolution of spatial maps with zooming in: top to bottom area decreases from 100% of the region to 27%, left to right forecast horizons are 1, 3, 6, and 12 months correspondingly.

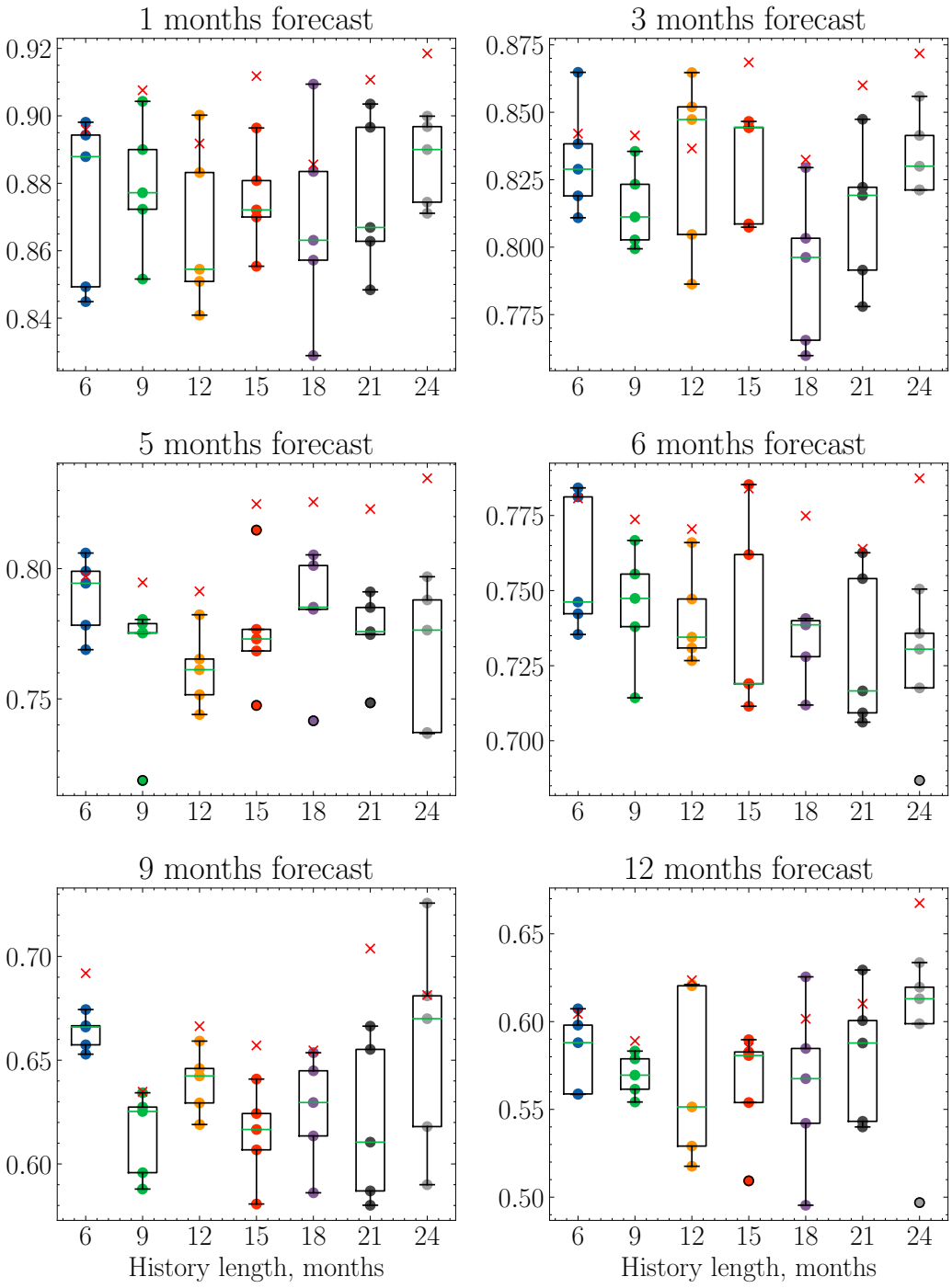


Figure 6: Variation of ROC AUC scores for five different random seeds; ConvLSTM binary classification for Missouri state. Metrics for an ensemble of five models are marked as red crosses.

score and ours on 54% and 16%, for short and long-term predictions, respectively. Before our study, 12-months ahead prediction gave close to random results, which is no longer the case. Additional improvement can be obtained by using an ensemble of deep learning models and increasing the amount of used data.

So we recommend to use EarthFormer for *short-term* predictions and ConvLSTM for *long-term*. For better predictions, one should use an ensemble for such models. Such a combination leads to a good model for the considered time horizons.

References

Adikari, K.E., Shrestha, S., Ratnayake, D.T., Budhathoki, A., Mohanasundaram, S., Dailey, M.N., 2021. Evaluation of artificial intelligence models for flood and drought forecasting in arid and tropical regions. Environmental Modelling & Software 144.

Alley, W.M., 1984. The Palmer drought severity index: limitations and assumptions. Journal of Applied Meteorology and Climatology 23, 1100–1109.

Bishop, C.M., Nasrabadi, N.M., 2006. Pattern recognition and machine learning. volume 4. Springer.

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, pp. 785–794.

Dai, A., 2011. Characteristics and trends in various forms of the palmer drought severity index during 1900–2008. Journal of Geophysical Research: Atmospheres 116.

Danandeh Mehr, A., 2021. Drought classification using gradient boosting decision tree. Acta Geophysica 69, 909–918.

Fort, S., Hu, H., Lakshminarayanan, B., 2019. Deep ensembles: A loss landscape perspective. arXiv preprint arXiv:1912.02757 .

Gao, Z., Shi, X., Wang, H., Zhu, Y., Wang, Y.B., Li, M., Yeung, D.Y., 2022. Earthformer: Exploring space-time transformers for earth system forecasting. Advances in Neural Information Processing Systems 35.

Ghozat, A., Sharafati, A., Asadollah, S.B.H.S., Motta, D., 2023. A novel intelligent approach for predicting meteorological drought based on satellite-based precipitation product: Application of an emd-dfa-dbn hybrid model. Computers and Electronics in Agriculture 211.

Gorelick, N., Hancher, M., Dixon, M., Ilyushchenko, S., Thau, D., Moore, R., 2017. Google earth engine: Planetary-scale geospatial analysis for everyone. Remote sensing of Environment 202, 18–27.

Guibas, J., Mardani, M., Li, Z., Tao, A., Anandkumar, A., Catanzaro, B., 2022. Adaptive fourier neural operators: Efficient token mixers for transformers. `arXiv:2111.13587`.

Hao, Z., Hao, F., Singh, V.P., Ouyang, W., Cheng, H., 2017. An integrated package for drought monitoring, prediction and analysis to aid drought modeling and assessment. Environmental Modelling & Software 91.

Hewamalage, H., Bergmeir, C., Bandara, K., 2021. Recurrent neural networks for time series forecasting: current status and future directions. IJF .

Huynh, T.D., Nguyen, T.H., Truong, C., 2023. Climate risk: The price of drought. Journal of Corporate Finance 65.

Jain, S., Liu, G., Mueller, J., Gifford, D., 2020. Maximizing overall diversity for improved uncertainty estimates in deep ensembles, in: Proceedings of the AAAI conference on artificial intelligence, pp. 4264–4271.

Jiang, Z., Rashid, M.M., Johnson, F., Sharma, A., 2021. A wavelet-based tool to modulate variance in predictors: An application to predicting drought anomalies. Environmental Modelling & Software 135.

Kail, R., Burnaev, E., Zaytsev, A., 2021. Recurrent convolutional neural networks help to predict the location of earthquakes. IEEE Geoscience and Remote Sensing Letters 19, 1–5.

Koldasbayeva, D., Tregubova, P., Shadrin, D., Gasanov, M., Pukalchik, M., 2022. Large-scale forecasting of heracleum sosnowskyi habitat suitability under the climate change on publicly available data. Scientific reports 12, 6128.

Kozlovskaia, N., Zaytsev, A., 2017. Deep ensembles for imbalanced classification, in: 2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA), IEEE. pp. 908–913.

Liu, Y., Ren, L., Ma, M., Yang, X., Yuan, F., Jiang, S., 2015. An insight into the palmer drought mechanism based indices: comprehensive comparison of their strengths and limitations. Stochastic Environmental Research and Risk Assessment 30. doi:10.1007/s00477-015-1042-4.

Marusov, A., Zaytsev, A., 2023. Non-contrastive representation learning for intervals from well logs. IEEE Geoscience and Remote Sensing Letters .

McKee, T.B., Doesken, N.J., Kleist, J., 1993. The relation of drought frequency and duration to time scales. Proceedings of the Eighth Conference on Applied Climatology , 179–184.

McPherson, Renee A., I.L.C., Richman, M.B., 2022. A place-based approach to drought forecasting in south-central oklahoma. Earth and Space Science 9.

Mishra, A.K., Desai, V.R., 2005. Drought forecasting using stochastic models. Stochastic environmental research and risk assessment 19, 202–216.

Mishra, A.K., Desai, V.R., 2006. Drought forecasting using feed-forward recursive neural network. Ecological Modelling 198, 127–138.

Mohammed, S., Elbeltagi, A., Bashir, B., Alsafadi, K., Alsilibe, F., Alsalman, A., Harsányi, E., 2022. A comparative analysis of data mining techniques for agricultural and hydrological drought prediction in the eastern mediterranean. Computers and Electronics in Agriculture 197, 1–19.

Niaz, R., Zhang, X., Iqbal, N., Almazah, M.M., Hussain, T., Hussain, I., 2021. Logistic regression analysis for spatial patterns of drought persistence. Complexity , 1–13.

Park, S., Im, J., Han, D., Rhee, J., 2020. Short-term forecasting of satellite-based drought indices using their temporal patterns and numerical model output. Remote Sensing 12, 3499.

Pathak, J., Subramanian, S., Harrington, P., Raja, S., Chattopadhyay, A., Mardani, M., Kurth, T., Hall, D., Li, Z., Azizzadenesheli, K., Hassanzadeh, P., Kashinath, K., Anandkumar, A., 2022. Fourcastnet: A global data-driven high-resolution weather model using adaptive fourier neural operators. `arXiv:2202.11214`.

Poornima, S., Pushpalatha, M., 2019. Drought prediction based on spi and spei with varying timescales using lstm recurrent neural network. Soft Computing , 8399–8412.

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.

Proskura, P., Zaytsev, A., Braslavsky, I., Egorov, E., Burnaev, E., 2019. Usage of multiple rtl features for earthquakes prediction, in: International Conference on Computational Science and Its Applications, Springer. pp. 556–565.

Rumelhart, D.E., Hinton, G.E., Williams, R.J., 1985. Learning internal representations by error propagation. California Univ San Diego La Jolla Inst for Cognitive Science .

Schmidhuber, J., Hochreiter, S., 1997. Long short-term memory. Neural Comput , 1735–1780.

Shi, X., Chen, Z., Wang, H., Yeung, D.Y., Wong, W.K., Woo, W.c., 2015. Convolutional lstm network: A machine learning approach for precipitation nowcasting. Advances in neural information processing systems 28.

Song, Y.H., Shahid, S., Chung, E.S., 2022. Differences in multi-model ensembles of cmip5 and cmip6 projections for future droughts in south korea. International Journal of Climatology 42, 2688–2716.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Polosukhin, I., 2017. Attention is all you need. Advances in Neural Information Processing Systems 30.

Vicente-Serrano, S., Santiago, B., Juan, L., 2010. A multiscalar drought index sensitive to global warming: The standardized precipitation evapotranspiration index. Journal of climate 23.

Wang, L., Chen, W., 2014. A cmip5 multimodel projection of future temperature, precipitation, and climatological drought in china. International Journal of Climatology 34, 2059–2078.

Xiujia, C., Guanghua, Y., Jian, G., Ningning, M., Zihao, W., 2022. Application of wnn-pso model in drought prediction at crop growth stages: A case study of spring maize in semi-arid regions of northern china. Computers and Electronics in Agriculture 199.

Xu, D., Zhang, Q., Ding, Y., Zhang, D., 2022. Application of a hybrid arima-lstm model based on the spei for drought forecasting. Environmental Science and Pollution Research , 4128–4144.

Zeng, A., Chen, M., Zhang, L., Xu, Q., 2023. Are transformers effective for time series forecasting? Proceedings of the AAAI conference on artificial intelligence , 11121–11128.

Zhang, S., Wu, Y., Sivakumar, B., Mu, X., Zhao, F., Sun, P., Han, J., 2019. Climate change-induced drought evolution over the past 50 years in the southern chinese loess plateau. Environmental Modelling & Software 122.

Zheng, C., Zhu, S., Mendieta, M., Yang, T., Chen, C., Ding, Z., 2021. 3d human pose estimation with spatial and temporal transformers. Proceedings of the IEEE/CVF International Conference on Computer Vision , 11656–11665.

9. Appendix

9.1. Multiclass problem

Description of experiment. As an additional experiment and study of models' limits, we looked at the multiclass classification problem for Missouri. For the 3 class problem, we arbitrarily set up thresholds of PDSI values as $-1,\ 1$ and for the 5 class as $-3,\ -1,\ 1,\ 3$. As an evaluation metric, we use median accuracy over all celled predictions. For gradient boosting and logistic regression, we use default implementations for multiclass predictions. For the neural networks-based model, we replace two possible output cells with the number of cells equal to the number of classes.

Analysis of results. Results, provided in Figure 7 and Table 8, are similar to a binary problem. Logistic regression and gradient boosting hold their superiority longer, the Convolutional LSTM score is relatively stable, and only the transformers' prediction disappoints, falling to the level of the historical baseline.

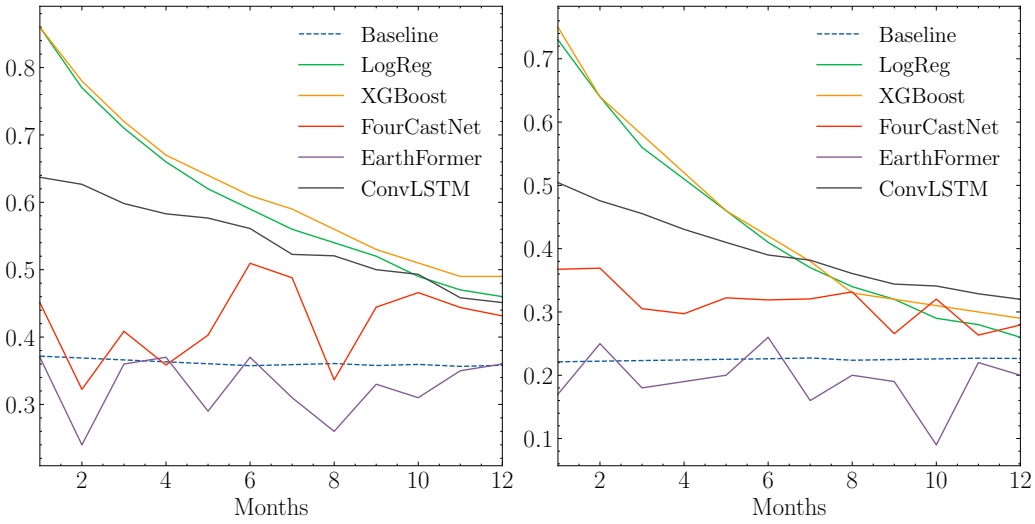


Figure 7: Accuracy vs. forecast horizon, 3 and 5 class task, Missouri

Horizon, months	1	3	6	9	12
median Accuracy, 3 class:					
Baseline	0.372	0.366	0.358	0.358	0.358
LogReg	**0.86**	<u>0.71</u>	<u>0.59</u>	<u>0.52</u>	<u>0.46</u>
XGBoost	**0.86**	**0.72**	**0.61**	**0.53**	**0.49**
FourCastNet	0.451	0.409	0.509	0.444	0.431
EarthFormer	0.37	0.36	0.37	0.33	0.36
ConvLSTM	<u>0.637</u>	0.598	0.561	0.500	0.451
median Accuracy, 5 class:					
Baseline	0.221	0.223	0.226	0.225	0.223
LogReg	<u>0.73</u>	<u>0.56</u>	<u>0.41</u>	<u>0.32</u>	0.26
XGBoost	**0.75**	**0.58**	**0.42**	<u>0.32</u>	<u>0.29</u>
FourCastNet	0.367	0.305	0.319	0.266	0.279
EarthFormer	0.17	0.18	0.26	0.19	0.20
ConvLSTM	0.504	0.455	0.389	**0.344**	**0.312**

Table 8: Median Accuracy vs. Forecast Horizon, 3 and 5 possible class problems, Missouri; best values are in bold, second best are underlined

Region area	104x136 (100%)	88x120 (75%)	72x104 (53%)	48x80 (27%)
104x136 (100%)	0.9292	0.9302	0.9327	0.9356
88x120 (75%)	-	0.9466	0.9487	0.9515
72x104 (53%)	-	-	0.9432	0.9438
48x80 (27%)	-	-	-	0.9326

Table 9: ROC AUC score trained on a subset (left) and evaluated on a different region (top) (length of input history = 6, forecast horizon = 1)

Region area	104x136 (100%)	88x120 (75%)	72x104 (53%)	48x80 (27%)
104x136 (100%)	0.8746	0.8775	0.8816	0.8829
88x120 (75%)	-	0.8525	0.8536	0.8508
72x104 (53%)	-	-	0.8595	0.8655
48x80 (27%)	-	-	-	0.8710

Table 10: ROC AUC score trained on a subset (left) and evaluated on a different region (top) (length of input history = 12, forecast horizon = 3)

Region area	104x136 (100%)	88x120 (75%)	72x104 (53%)	48x80 (27%)
104x136 (100%)	0.7525	0.7624	0.7722	0.7897
88x120 (75%)	-	0.6901	0.6974	0.7081
72x104 (53%)	-	-	0.7597	0.7794
48x80 (27%)	-	-	-	0.7544

Table 11: ROC AUC score trained on a subset (left) and evaluated on a different region (top) (length of input history = 9, forecast horizon = 6)

Region area	104x136 (100%)	88x120 (75%)	72x104 (53%)	48x80 (27%)
104x136 (100%)	0.6117	0.6154	0.6142	0.6210
88x120 (75%)	-	0.5890	0.5980	0.6033
72x104 (53%)	-	-	0.6126	0.6184
48x80 (27%)	-	-	-	0.6264

Table 12: ROC AUC score trained on a subset (left) and evaluated on a different region (top) (length of input history = 24, forecast horizon = 12)