Smart Agricultural Technology 12 (2025) 101540

Contents lists available at ScienceDirect

Smart Agricultural Technology

journal homepage: www.journals.elsevier.com/smart-agricultural-technology

On the strategy of exploring spatio-temporal information from Earth observation data for crop yield prediction

Stella Ofori-Ampofo [a,*], Ridvan Salih Kuzu [b], Peter Schauer [c], Martin Willberg [c], Adrian Höhl [a], Xiao Xiang Zhu [a,d,*]

[a] Technical University of Munich, Germany; TUM School of Engineering and Design, Department of Aerospace and Geodesy, Germany
[b] Remote Sensing Technology Institute, German Aerospace Center, Weßling, Germany
[c] Industrieanlagen-Betriebsgesellschaft mbH, Ottobrunn, Germany
[d] Munich Center for Machine Learning, Munich, Germany

ARTICLE INFO

Dataset link: https://github.com/ellaampy/SpatioTemporalYield

Keywords:
Crop yield
Remote sensing
Machine learning
Forecasting
Time series

ABSTRACT

Crop yield information plays a pivotal role in ensuring food security. Advances in Earth Observation technology and the availability of historical yield records have promoted the use of machine learning for yield prediction. Significant research efforts have been made in this direction, encompassing varying choices of yield determinants and particularly how spatial and temporal information are encoded. However, these efforts are often conducted under diverse experimental setups, complicating their inter-comparisons. In this paper, we present our findings on multiple strategies for encoding spatial-spectral information at the county level through averaging pixel values, pixel sampling, and image histograms alongside approaches for encoding temporal information, including recurrent neural networks, temporal convolutions, and attention mechanisms. Our numerical results indicate that predicting crop yield solely using time series data can be effective, even without spatial information, and classical machine learning methods remain competitive in this application. Surface reflectance information emerges as a critical predictor in the absence of weather and spectral indices. While machine learning models typically require an extensive sample size, our findings suggest that reliance on long-term historical data may hinder models' ability to accurately reflect current conditions. This study provides valuable insights into feature and model selection for county-level yield prediction, highlighting the interplay between data structure, model complexity, and predictive performance.

1. Introduction

Agriculture production is expected to more than double to meet the pressure from growing food demand [1]. While interventions like genetic modification, cropping intensification and mechanization have transformed the agriculture sector, the desire to satisfy a food-insecure population remains challenging under the threats of erratic climate conditions, economic slowdowns and conflicts [2]. Climate change, particularly rising temperatures, adversely affects crop growth [3–5]. Extreme weather conditions have, in some cases, accelerated insect populations and activity, leading to increased crop consumption and yield loss [6]. Conflicts in various regions further exacerbate these challenges by restricting access to farmlands and disrupting harvesting activities, which severely impacts agricultural revenues [7,8]. These factors collectively reduce crop production and revenue, further constraining investments in subsequent planting seasons [9].

Researchers have applied various environmental and weather factors from satellite data to understand crop conditions and to estimate crop yield. On the one hand, a generalization of the widely used approaches includes process-based models which simulate crop growth under various environmental and management conditions [10,11]. On the other hand, data-driven or empirical approaches, such as statistical and machine learning (ML) models, approximate a function that correlates productivity drivers to yield. The former requires expert knowledge of crop biophysical processes and can present a high level of uncertainty due to excessive parameterization [12]. As such, data-driven approaches

* Corresponding authors at: Technical University of Munich, Germany.
E-mail addresses: stella.ofori-ampofo@tum.de (S. Ofori-Ampofo), ridvan.kuzu@dlr.de (R.S. Kuzu), schauer@iabg.de (P. Schauer), willberg@iabg.de (M. Willberg), adrian.hoehl@tum.de (A. Höhl), xiaoxiang.zhu@tum.de (X.X. Zhu).

https://doi.org/10.1016/j.atech.2025.101540
Received 21 December 2024; Received in revised form 9 September 2025; Accepted 5 October 2025

Available online 16 October 2025

have become more attractive, and the advances in ML methods coupled with the availability of satellite data facilitate the exploitation of spatial and temporal information.

The capacity of ML to predict crop yields has been demonstrated in several studies and for varying prediction units [13–16]. A prediction unit refers to the geographic boundary, such as a pixel, farm, or administrative region, at which crop yield data is collected or aggregated. At large prediction units like county level or districts, encoding spatial information presents unique challenges due to variations in crop conditions across different ecological zones and the irregular nature of administrative boundaries. Moreover, the inherent phenological characteristics of crops incite the use of ML models designed to efficiently exploit temporal information. Despite methodological advances in ML for yield prediction, there is limited insight into the trade-offs between the numerous ways of handling spatial and temporal information and their respective impact on prediction performance. Addressing this gap is essential for informing data collation efforts and guiding model selection, particularly when developing large-scale prediction systems in resource-constrained environments [17]. Relying only on temporal information to predict yield is prevalent [13,18,19], but the approaches used are far from exhaustive, leaving room for improvement. Inter-study comparisons are also challenging, as each operates under different experimental setups, including feature combinations and geographical and temporal scopes.

In this paper, we present a first comprehensive comparison of existing spatial and temporal encoding techniques for corn yield prediction in the USA. We build upon current methods by introducing innovative ideas from crop classification studies, particularly using pixel-set encoding [20], to complement the widely adopted mean-averaged and histogram-based transformations prevalent in yield prediction research [16,21,22]. For temporal feature extraction, we explore state-of-the-art temporal encoding techniques, including multi-scale residual networks (MSResNet [23] and InceptionTime [24]) and a multi-headed temporal attention encoder [25] in addition to standard recurrent neural networks (RNN) and single-kernel convolution [26,27]. Our work not only provides a rigorous evaluation of spatial and temporal encoding strategies but also delivers practical insights into their trade-offs and contributions to yield prediction accuracy. Furthermore, leveraging the MSResNet architecture, we conduct an in-depth analysis of key factors affecting model performance, such as sample size, feature combinations, and in-season yield forecast. In this context, we also examine the role of prior-year observations: an often overlooked yet informative addition [28]. These immediate preceding values may implicitly capture recent agronomic practices, management conditions, or persistent environmental effects, which in turn can enhance model generalization and predictive accuracy. By addressing these aspects, we aim to inform future research and support the development of robust, scalable yield prediction models applicable to both data-rich and resource-constrained environments. To support this analysis, we compiled a multi-source dataset consisting of yearly county-level corn yield statistics, gridded 8-day MODIS surface reflectance and derived spectral indices, as well as gridded daily weather variables from Daymet.

The remainder of the paper is organized as follows: Section 2 reviews previous work on machine learning techniques for crop yield prediction. Section 3 describes the methods used to encode spatial and temporal information. Sections 4 and 5 introduce the dataset and outline the experimental setup respectively. Finally, Section 6 presents and discusses the results of our study.

2. Related work

Crop yield is influenced by a complex interplay of biotic and abiotic factors that vary across agro-ecological zones. Yield prediction models predominantly relied on meteorological variables such as precipitation and temperature, which are known to significantly influence yield variability [29,3,30]. These climatic variables are often supplemented with vegetation indices, which serve as proxies for crop health and biomass accumulation throughout the growing season [31–33]. In some studies, additional inputs such as soil characteristics and farm management practices are incorporated as they play a critical role in determining crop productivity by influencing root development, nutrient availability, and water retention capacity [13,18]. Relative to these more commonly used features, surface reflectance (SR) despite being a more direct measurement of land surface condition remains relatively underexplored. SR retains fine-grained spectral information that can better capture subtle variations in crop health, growth stages, and stress responses [34]. There is growing evidence that ML models relying on such data can yield promising results in predicting crop yield or mapping crops [15,16,21,27,35].

In assessing the relevance of crop yield predictors or features, existing approaches employ a stage-specific feature selection strategy—choosing predictors aligned to known crop phenological phases [14, 16,13]. However, this approach risks missing interactions across time and may suffer from temporal misalignment due to variability in planting dates or crop progress. In contrast, our work adopts a season-long perspective, leveraging full time series via feature grouping to capture evolving conditions and interactions throughout the crop lifecycle. Moreover, utilizing the entire time series instead of relying on certain features at specific states is more practical.

Machine learning methods applied in yield prediction range from traditional ML techniques to deep learning models that efficiently capture temporal and spatio-temporal information and have consistently outperformed statistical and tree-based methods, as shown in Table 1. Early approaches often relied on traditional ML models such as random forests, lasso regression, or multilayer perceptrons (MLPs) [13,18]. However, these models are limited in their ability to capture temporal dependencies. To address this, recurrent neural networks, and one-dimensional (1D) convolutions have been employed and shown improved performances than standard multi-layer perceptrons [13,18] and statistical or traditional ML techniques. Hybrid models combining 1D convolutional neural networks (CNNs) with LSTMs have also been proposed to jointly capture local patterns and longer-term dependencies [18,13,22]. Attention-based mechanisms further improve model performance by dynamically weighting relevant time steps, allowing the model to focus on critical growth periods [18,37]. While there has been a growing adoption of novel deep learning architectures, several state-of-the-art approaches remain underexplored. For example, multi-scale residual networks (MSResNet) [23] and InceptionTime [24] have achieved state-of-the-art results in a range of time series classification and regression tasks, yet their application to yield prediction has so far been limited.

Beyond temporal information, spatial variability plays a crucial role in yield prediction. Geographic heterogeneity in soil properties, weather patterns, and management practices can significantly influence crop outcomes. To encode spatial structure, it is common practice to aggregate features over predefined prediction units, which obscures intra-region variability. Meanwhile, the availability of satellite image time series (SITS) has resulted in a gradual adoption of spatial convolution architectures. 2D convolutional neural networks (2D CNNs) have been used to extract spatial patterns from individual image frames and are often coupled with RNNs to model temporal dependencies. In other cases, 3D CNNs or ConvLSTMs are used to directly encode spatio-temporal information [21,36]. The introduction of transformer-based models, such as the Multi-Modal Spatial-Temporal Vision Transformer (MMST-ViT), has also pushed the boundaries of yield prediction by incorporating both spatial and temporal contexts in a unified architecture [36]. Spatial modeling at large prediction units (e.g., counties) presents unique challenges. Counties are irregularly sized and large, thus directly feeding SITS is impractical. A feasible option is sampling uniformly sized mini-SITS per county however, the success of SITS-based models are questionable in how SITS data is prepared for training. Crop masking is omitted, which will otherwise result in an SITS with reduced pixel contiguity, [21,38,39,36] and limiting the ability to extract meaningful

Table 1
Inter-comparison of ML models for corn and soybean yield prediction in the USA. RMSE units are in bushels per acre (bu/acre). Performances are drawn from different studies and are not directly comparable, as they are based on varying experimental setups and input data sources. ^{DA} denotes the application of temporal data augmentation, while * indicates self-supervised pretraining. SR, FM and M represent surface reflectance, farm management and meteorological features.

Reference	Features				Crop masking	Model	Encoder		RMSE (bu/acre)	
	SR	FM	M	Soil			Spatial	Temporal	Corn	Soybean
[21]	✓		✓			ridge				8.10
						decision tree				7.64
						MLP				7.19
						2D CNN (histogram)	✓	✓		6.60
						3D CNN	✓	✓		5.22
[13]		✓	✓	✓		lasso			31.30	9.49
						random forest			26.02	12.78
						MLP			21.37	5.89
						1D CNN-LSTM		✓	17.64	4.91
[16]	✓		✓		✓	ridge				7.96
						MLP				8.32
						decision tree				7.73
						LSTM (histogram)		✓		5.83
						2D CNN (histogram)	✓	✓		5.55
[18]		✓	✓	✓		random forest			28.76	8.21
						lasso			31.23	9.24
						SVM			29.84	8.37
						DNN			28.69	8.35
						1D CNN-LSTM		✓	24.33	7.42
						LSTM		✓	29.62	9.77
						LSTM-Attention		✓	17.49	5.90
[19]			✓	✓		MLP-LSTM		✓	20.71	5.27
[36]	✓		✓			ConvLSTM	✓	✓	18.6	7.2
						CNN-RNN		✓	14.6	5.8
						MMST-ViT	✓	✓	13.2	5.1
						MMST-ViT*	✓	✓	**10.5**	**3.9**
[22]	✓		✓	✓		1D CNN	✓	✓	27.00	9.68
						1D CNN-LSTM		✓	25.48	10.96
						1D CNN-LSTM (histogram)	✓	✓	22.46	7.96
						1D CNN-LSTM^{DA}(histogram)	✓	✓	18.94	-

spatial patterns. This raises concerns about whether SITS-based models are actually learning from valid crop-specific pixels. An alternative strategy that bypasses these shortfalls at the expense of natural spatial order is the transformation of SITS into image histograms [16]. Pixel-set encoders, such as PSE-TAE [20], offer another promising solution by modeling sets of crop-specific pixels as unordered collections, enabling flexible spatial summarization without relying on contiguity or regular grids. This approach while dominant for crop application at the farm level has been unexplored for large prediction units. As summarized in Table 1, incorporating spatial information results in improved performance; however, comparison is limited to a single temporal model. Existing studies are limited by fragmented experimental setups, inconsistent feature choices, and incomplete spatial preprocessing. Our work fills this gap by systematically benchmarking a range of temporal models under a unified dataset and experimental setup.

3. Encoding spatial and temporal information

Accurately estimating crop yield from remote sensing data depends on the model's ability to learn spatial patterns and temporal dynamics. Spatial patterns reflect the variability within agricultural regions as a result of crop management, or local environmental conditions. Temporal patterns capture phenological stages over the growing season from emergence to senescence. Altogether, these patterns can provide essential signals related to spatial and temporal variability and ultimately crop productivity. Satellite image time series especially from multi-spectral satellites, present a rich source of information for capturing these insights. Their high-dimensionality requires summarizing (encoding) spatial and temporal structure into meaningful features. As highlighted in Section 2, numerous methods have been proposed to

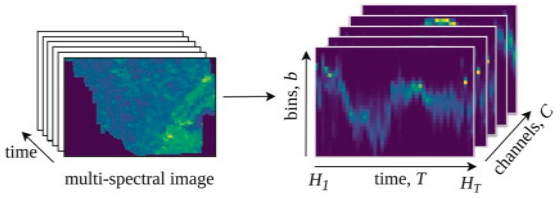

Fig. 1. Transformation of a multi-spectral SITS data into 3D histograms.

encode spatial and temporal information for yield prediction using satellite data. In this section, we detail these strategies, introducing the application of pixel-set encoders to effectively capture spatio-spectral information at large spatial scales (e.g., county level) and residual and attention-based networks to encode temporal information. We begin by discussing spatial encoding methods that transform pixel-level imagery into compact representations, followed by temporal models that learn patterns and dependencies across time. For these discussions, we target yield prediction at county level, where for a given county and year, we have a multi-variate SITS (X) in the format $X \in \mathbb{R}^{T\times C\times H\times W}$. H and W are the height and width of the bounding box of an irregularly shaped county and vary due to the different geometrical sizes of each county. T and C represents the number of observations per year and the number of channels/features respectively. In Section 3.1.1 and Section 3.2.1, we present baseline approaches to spatial and temporal encoding.

3.1. Encoding spatial information

3.1.1. Pixel averages

Most studies addressing yield prediction at broader prediction units (e.g., counties) simplify spatial representation by averaging all pixels

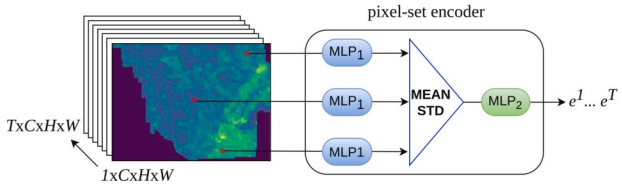

Fig. 2. Illustration of pixel-set encoder (adapted from [20]). Random pixels are sampled within a county and at the same location across the temporal dimension.

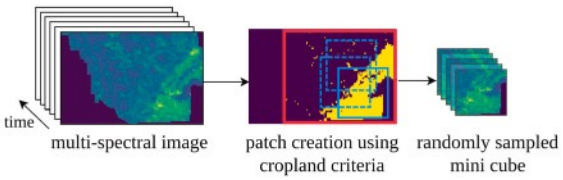

Fig. 3. Illustration of random cropping of uniform patches across the prediction unit (adapted from [21]). Red and blue outlines denote cropland bounds and candidate patches, respectively.

within a given region [14,13,40]. While this approach reduces data dimensionality, it significantly limits the model's capacity to learn meaningful spatial features. Considering a SITS, $X \in \mathbb{R}^{T\times C\times H\times W}$, the height, H and width W dimensions of a satellite image are reduced into a scalar value by computing the mean of all pixels $[p_1, ..., p_N]$ within a prediction unit, where N is the number of pixels. This results in a compact data structure of dimension, time (T) × channel (C), and permits the learning of spectro-temporal features only.

3.1.2. Histograms

To retain some information about the pixel distribution while still discarding spatial arrangement, another approach encodes each image channel into a histogram. Assuming that the spatial arrangement of crop pixels merely impacts yield, [16] suggested the mapping of a SITS into a histogram of pixel counts. Pixels within each image channel ($c \in [1, .., C]$) are discretized into a user-defined number of bins, b to produce a histogram h_c. A compact representation of the multi-spectral image is obtained by concatenating each histogram $h_1, ..., h_C$ to form H. Considering the multi-temporal nature of SITS, we obtain $H_1, ..., H_T$ as depicted in Fig. 1. The resulting histogram images can be operated by spatial CNNs.

3.1.3. Pixel-set encoders

Spatial CNNs may be less effective on coarse-resolution satellite data due to limited texture. Compared to images with higher spatial resolution, coarse-resolution images may exhibit minimal texture to extract expressive spatial features. [20] proposed the pixel-set encoder (Fig. 2) to learn statistical descriptors of the spectral distribution of pixels in a prediction unit which is invariant to the permutation of the pixels' location. To achieve this, a subset of pixels, $S \subset [1, ...N]$ is randomly selected from the total number of pixels, N within a prediction unit (as in Equation (1)) and across the time dimension, T. Every element $s \in S$ is passed through an MLP consisting of a succession of fully connected layers, batch normalization and rectified linear unit activation function (Equation (2)).

$$S = subset(S, N) \tag{1}$$

$$\hat{e}_s^t = MLP_1(X_s^t), \forall s \in S \tag{2}$$

$$e^t = MLP_2\left(pooling(\hat{e}_s^t)\right) \tag{3}$$

Since prediction units vary in spatial extent, the total number of available pixels N can differ significantly. To obtain a consistent spatio-temporal embedding e^t, a fixed-sized subset S is sampled. For smaller counties where $N < S$, pixels are sampled with replacement (i.e., repeated sampling) to meet the required input size. The resulting layer is pooled along the subset dimension and further processed by a configuration of MLPs (Equation (3)) to obtain a spatial representation. Generalization in this approach is enforced by randomly selecting new subsets during each training step. The encoding of spatio-spectral information as set-encoders or histograms overcomes the strict requirement of having fixed-sized input, which is impractical for administrative units.

3.1.4. Image patches

While the PSE and histogram methods retain some degree of spatial variability, they disregard the positional relationships among pixels. As [21] argued, omitting or randomly permuting the spatial layout leads to a loss of local contextual information, particularly the variations between neighboring pixels, which may be critical for capturing spatial structure. Such variations can form distinctive textural patterns, even in coarse-resolution satellite imagery, that are informative for CNNs trained to recognize spatial structure. Following this rationale, [21] designed input data as randomly sampled mini-images (patches) of a prediction unit (Fig. 3). Spatial CNNs can then be used to encode spatial information by leveraging convolutional layers that apply localized filters across the input, generating feature maps that capture hierarchical spatial features such as textures and patterns. These feature maps serve as progressively abstract representations of the spatial structure in the data. The shortfalls of this approach in the context of crop yield prediction where crop masking is essential is described in Section 2.

3.2. Encoding temporal information

Following the application of spatial encoding techniques, the resulting SITS representations retain their temporal structure, which encapsulates crop phenological progression over the growing season. To effectively leverage this temporal dimension, it is necessary to employ models capable of capturing time-dependent patterns and sequential dynamics. In this section, we present a range of temporal encoding approaches, beginning with traditional machine learning methods that operate on temporally unordered features, and progressing to advanced deep learning architectures specifically designed to model temporal dependencies.

3.2.1. Unordered temporal sequences

Random Forests (RF), Gradient Boosted Trees (GBT), and Support Vector Machines (SVM) are widely used ML models for satellite-based classification and regression tasks [16,21,35]. Due to their simplicity, low computational cost, and historical success, they are often used as strong baseline models. RF is an ensemble method that builds multiple decision trees from different data subsets and averages their predictions to overfitting and improve generalization. GBT sequentially constructs trees, each correcting the errors of the previous ones. SVM aims to find an optimal hyperplane that separates the data while maximizing the margin between support vectors. This makes them particularly effective in high-dimensional feature spaces. MLPs are fundamental building blocks in deep learning architectures. They are a class of feed-forward networks composed of multiple layers of interconnected neurons. Input features are forward-passed to a function that computes their weighted sum as a linear combination of their weights. An activation function adds non-linear transformations to the inputs to learn complex relations.

All these models require 2D inputs (samples × features). For SITS, this means the multi-temporal and multi-spectral dimensions (i.e., time steps and channels) must be flattened into a single feature vector per prediction unit. This flattening operation eliminates any notion of temporal order or sequence structure, preventing the models from capturing dynamics over time. Despite this limitation, these models are included as baselines in our study to benchmark the performance of more advanced, sequence-aware architectures described in later sections.

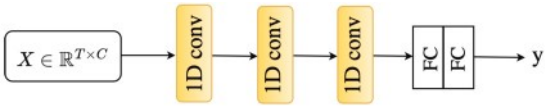

Fig. 4. Schematic diagram of TempCNN. 1D convolutions and the first dense layer are succeeded by batch-normalization, ReLU and dropout layers. FC denotes a fully connected layer.

3.2.2. Temporal convolution neural networks

CNNs are popularly used for image recognition tasks but have been applied in several remote sensing applications [41]. Although they are often used in the extraction of spatio-spectral features, they have been applied to embed temporal information for tasks involving sequences or time series data [13,27,23]. In the temporal domain, CNNs use 1D convolutional layers that slide filters over sequences of observations (e.g., multi-spectral reflance values across time. These filters capture short-term temporal patterns such as growth trends or phenological shifts. One notable architecture in this space is TempCNN [27] which relies on temporal convolutions for time series classification using satellite data. The architecture, as shown in Fig. 4, consists of sequences of 1D convolution layers succeeded by batch normalization, ReLU and dropout layers. A dense layer operates on the output of the convolution blocks to map the extracted representations to a target (e.g., yield).

Beyond simple temporal convolutions, more advanced architectures that rely on residual networks and uses multiple convolution branches with different kernel sizes to enhance temporal feature extraction have also been applied to time series applications. We detail the configuration of a specific architecture called the Multi-Scale Residual Network (MSResNet) [23].

The architecture consists of three independent streams of 1D convolutions operated by different kernel sizes (Fig. 5). These parallel streams allow the model to learn short and long-term patterns at different temporal resolutions simultaneously. Skip connections are introduced to succeeding convolution blocks to learn earlier levels of abstraction. Then, the separate representations derived by multiple scales (kernel sizes) are average-pooled and concatenated to combine features learned at multiple scales. A fully connected layer processes the concatenated features to predict a target. MSResNet requires an input temporal size of 512; hence, an initial interpolation is required for shorter sequences.

Several other models exist for benchmarking non-satellite image time series tasks, such as the UCR time series archive [42]. A notable mention is the InceptionTime model, which achieved state-of-the-art performance on a significant number of classification tasks in the UCR archive. InceptionTime is an ensemble of inception networks succeeded by a global average pooling layer and a dense layer. The central building block of the InceptionTime architecture is the inception module. Each module includes a bottleneck layer for dimensionality reduction, followed by three parallel 1D convolutional layers with different kernel sizes and a max-pooling path. The outputs from these convolutional layers are concatenated, enhancing the model's ability to learn both short and long-term dependencies efficiently. Akin to MSResNet, InceptionTime introduces residual connections within its network layers and has been explored for crop type mapping [35].

3.2.3. Recurrent neural networks

Sequence-based neural networks, such as Recurrent Neural Networks (RNNs), are specifically designed to model temporal dependencies in sequential data. Unlike traditional feedforward architectures, RNNs incorporate recurrent connections, allowing them to retain information from previous time steps. This makes them well-suited for tasks where past context is essential, such as speech recognition and language modeling. In the context of satellite time series, RNNs can model crop dynamics by processing spectral observations over time. However, standard RNNs struggle to learn long-term dependencies due to issues such as vanishing or exploding gradients. To address this limitation, Long Short-Term Memory (LSTM) networks [26] were developed. LSTMs enhance the RNN architecture by introducing memory cells that can maintain information over longer sequences. Each LSTM cell contains three gates namely the forget, input, and output gates, that regulate the flow of information. The forget gate determines what information to discard from the previous cell state, the input gate controls which new information to add, and the output gate decides what part of the cell state to expose as output. These gates are controlled by learned weights and nonlinearities, enabling the network to selectively preserve or update memory across time steps. This gated mechanism allows LSTMs to mitigate the vanishing gradient problem and effectively capture both short- and long-term dependencies. As shown in Fig. 6, the LSTM cell processes both the current input and the previous hidden state to update its internal memory and compute the output for the current time step. This structure makes LSTMs particularly suitable for modeling crop phenological stages, which unfold gradually over time and require memory of past growth conditions. Compared to temporal convolutions, LSTMs operate sequentially.

3.2.4. Attention-based networks

Attention mechanisms have emerged as a powerful alternative to RNNs for sequence modeling, offering improved parallelism and flexibility. The primary idea of attention is to allow the model to focus dynamically on the most informative parts of an input sequence when making predictions. In crop classification tasks, attention patterns have been shown to has been shown to capture narrow, distinct temporal instances that obtain classification-relevant features, while suppressing observational noise such as cloud contamination in raw satellite time series [43,25]. One such model is the Lightweight Temporal Attention Encoder (LTAE), originally proposed for satellite-based crop type mapping [25]. LTAE employs self-attention mechanism to encode satellite series into feature embeddings. It incorporates multi-headed attention [44] by splitting the input channels into groups (heads), allowing each head to capture distinct temporal patterns in parallel. For each head, a master query vector is defined, and attention scores are computed as the scaled softmax of the dot-product between the keys and the master query (Fig. 7). These scores are used to generate temporally weighted representations of the input. The outputs of each head are the temporally weighted sums of the inputs, effectively summarizing the most informative temporal segments. They are further concatenated and processed by a multi-layer perceptron to generate the final feature embedding. In empirical evaluations, LTAE has outperformed traditional RNNs (including GRUs and LSTMs) and temporal CNNs in both prediction accuracy and computational efficiency for crop type classification tasks [25].

4. Study area and data

The United States of America (USA) is the world's largest producer of corn, accounting for approximately one-third of global production. A momentous reduction in corn yield has severe implications for domestic availability, market pricing, and export quantities. 2012 was one such year where an extreme drought precipitated poor production. The majority of the crop fields entered this season with below-average moisture [3], resulting in a high variance in average yield compared to the previous year's estimates (Fig. 8). We conduct our study in the USA's top five corn-producing states: Iowa, Illinois, Indiana, Nebraska, and Minnesota. Altogether, they accounted for over one-half of the USA's corn (grain) production in 2021 [45]. We rely on gridded surface reflectance, spectral indices and weather data from our ongoing work [46] that seeks to produce a large-scale, multi-resolution spatio-temporal dataset for crop monitoring in the USA. Average farm sizes in the selected states are over 250 acres [47]. As a result, coarser resolution satellite images such as MODIS suffice for understudying satellite-based crop yield prediction and have been used extensively in this region [16,21,48].

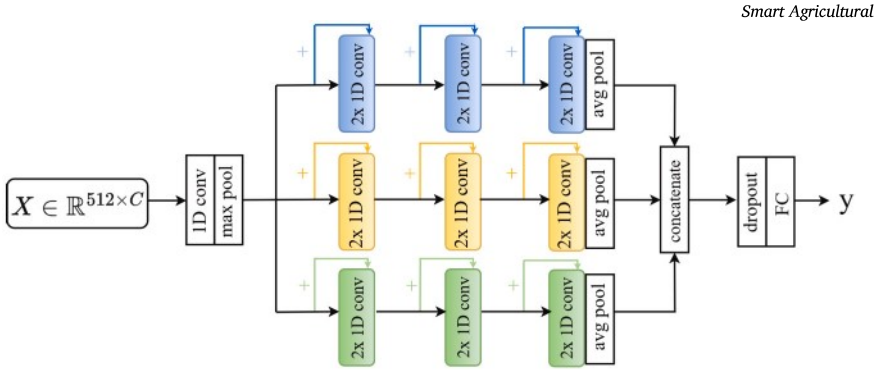

Fig. 5. Schematic diagram of MSResNet. Multiple branches of 1D convolutions succeeded by batch normalization and ReLu activation are applied to the input channel using different kernel sizes.

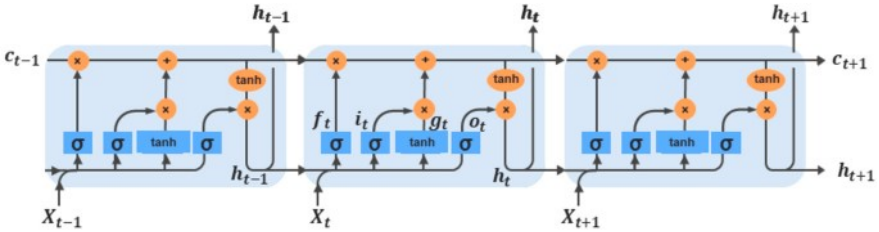

Fig. 6. Structure of an LSTM cell. This unrolled LSTM diagram illustrates how information flows across time steps through both the cell state (c_t) and hidden state (h_t). At each time step, the previous hidden state (h_{t-1}) and current input (X_t) are used to compute four gates. The forget gate (f_t) determines which parts of the previous cell state (c_{t-1}) are retained. The input gate (i_t) and candidate state (g_t) update the cell state. The cell state is revised by combining the outputs of the forget and input gates. Finally, the output gate (o_t) controls how much of the updated cell state contributes to the new hidden state (h_t).

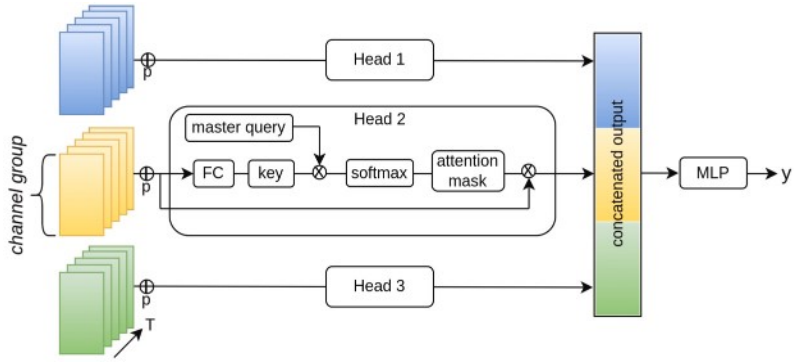

Fig. 7. Schematic diagram of LTAE (adapted from [25]) demonstrating multi-head operation for an input time series ($T \times C$). The initial channel dimension, C, undergoes a linear projection before a multi-head learning process.

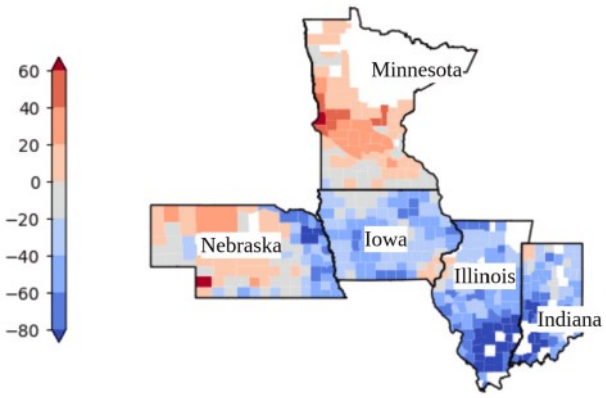

Fig. 8. Map of the five states showing the difference in average corn yield (bu/acre) for the drought year (2012) and the pre-drought year (2011). Extreme reduction in yield losses was observed in Iowa, Illinois, Indiana and eastern Nebraska.

4.1. County-level yield records

County-level corn yield (bu/acre) data is obtained from the National Agriculture Statistics Office of the United States Department of Agriculture (USDA-NASS). The data is available yearly for several commodities. We retrieved 19 years of corn records in tabular form, from 2003 to 2021 and for the 473 counties in this region. Throughout these years, some yield records were missing; for example, in 2020 and 2021, only 94% and 77% of the counties, respectively, had yield records.

4.2. Crop type maps

The USDA-NASS provides annual gridded crop type maps at 30-meter resolution. The maps are produced using decision-tree methods to provide acreage estimates for major crops. For earlier years, e.g., 2003, no crop map was available for Minnesota; hence, satellite time series for this location were discarded. Overall, the reported mapping accuracy for corn exceeds 90% in the selected states [49]. The crop-type layer is used to mask out the SITS per year.

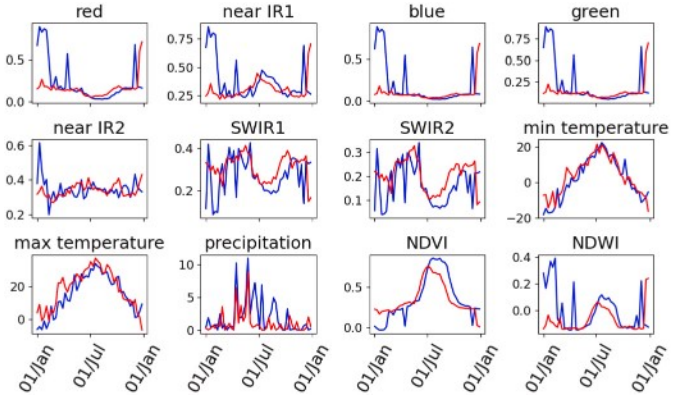

Fig. 9. Comparing the temporal variability in spectral reflectance, weather variables, and spectral indices for a pre-drought year (2011 - blue) and drought year (2012 - red) for a selected county in Nebraska. Surface reflectance features are unitless and range between 0 and 1. Temperature and precipatation plots are in °C and *mm* respectively. The temporal variations are prominent in short-wave infrared (SWIR), NDVI, NDWI, and precipitation deficits from July onwards. During these years, the average yield for the selected county was 164.9 and 99.1 bu/acres, respectively, for 2011 and 2012.

4.3. Surface reflectance and spectral indices

The moderate-resolution imaging spectroradiometer instrument (MODIS) is one of the longest-standing remote sensing missions. Its global coverage and high temporal frequency allow for near real-time monitoring. The collection MOD9A1.061 [50] estimates land surface reflectance along the visible and infrared (IR) range (see Table A.1 in Appendix A) at a spatial resolution of 500 meters and a revisit of 8 days (resulting in ideally 46 timesteps every year). Each pixel in this product represents the highest-quality observation identified over an 8-day interval and chosen based on criteria such as extensive observation coverage, minimal viewing angle, clear atmospheric conditions (absence of clouds and cloud shadows), and low aerosol presence. Two spectral indices are computed from MODIS bands to supplement the reflectance data: normalized difference vegetation index (NDVI) and normalized difference water index (NDWI) to incorporate domain-knowledge features. The two-band NDWI is derived from a near IR band and a second IR band as in [51]. NDVI is computed as the normalized difference between the near IR and red band. NDVI and NDWI approximate vegetation greenness and canopy water content, respectively.

4.4. Weather data

We consider precipitation and temperature information from Daymet [52] as indicators of weather conditions during crop growth. Daymet interpolates several weather variables from ground-station measurements using statistical techniques at a daily temporal resolution and at 1 km spatial resolution. For this study, only precipitation and minimum and maximum temperature are used. Fig. 9 shows the temporal variation of the selected weather variables, surface reflectance and spectral indices for a selected area across different crop seasons (years), highlighting their potential to capture yield dynamics.

As our task involves yield prediction at the county level, the dataset is structured such that for each county (i) at a given year, we construct a multi-variate time series (X) in the format $X_i \in \mathbb{R}^{T\times C\times H\times W}$. H and W are the height and width of the bounding box of an irregularly shaped county and vary due to the different geometrical size of each county. T and C represents the number of observations per year (46 timesteps) and the number of channels (12 features) respectively, the harmonization of features with varying spatial and temporal resolutions is detailed in Section 5.

5. Experiment

In this section, we describe how the multi-source data presented in Section 4 is harmonized before training the yield prediction models. In addition, the yield prediction task is formulated, and the experiment setup is described.

5.1. Data processing

Daymet data is resampled to the resolution of MODIS. Given the coarse nature of MODIS and Daymet data and the relatively higher spatial resolution of the cropland layer, a remapping scheme is adopted to downsample the cropland layer to 500 meters. First, the cropland mask is overlaid over the 500-meter MODIS grid; then a threshold is defined to remap each grid as corn if it constitutes 60% or more coverage. A year-by-year crop masking is performed on the SITS to filter out non-corn pixels. Then the start and end period of the yearly time series is truncated to focus on the usual corn season (April to October) [53], reducing the 46 timesteps to 27. To harmonize the different temporal frequencies of MODIS and Daymet, daily Daymet variables are reduced to 8-day temporal resolution by averaging. The resulting data is normalized channelwise using each channel's 2^{nd} and 98^{th} percentile calculated over the whole data. Our final SITS is of size $X \in \mathbb{R}^{27\times 12\times H\times W}$ and form the basis for transforming the input data into the various formats described in Section 3. For the period considered (2003-2021), the total number of samples available for training, validation and testing was 4940, 2116 and 795 respectively. Fig. 10 shows our data preparation workflow.

5.2. Experiment setup

We formulate the task by predicting the average yield of a county (i) at a given year (y) using its corresponding multi-variate time series X such that $X_i \in \mathbb{R}^{T\times C\times H\times W} \rightarrow y_i \in \mathbb{R}$. Our baseline models are tree-based methods (RF, XGBoost), SVM and MLP. Here, spatial information is summarized as pixel averages, and temporal sequences are merely treated as features. 1D temporal convolution-based models (TempCNN, MSResNet, InceptionTime), LSTM and LTAE are similarly applied to pixel averages to investigate how the different ways of handling temporal information compare to models that do not consider temporal order. Under this category of experiments, we also augment an LSTM with an attention mechanism. Histogram images are processed with 2D convolutions (Histogram-2D) similar to the architecture used in [16]. In another experiment, the sequential part of the histogram is handled using LSTMs (Histogram-LSTM) and temporal convolutions (Histogram-TempCNN) and the bin and channel dimensions are flattened to expand the feature space. Pixel-set encoders are applied to sampled pixels, and the temporal dimension is handled with LTAE (PSE-LTAE) as per the original architecture [25]. In this strategy, we first select 80% of pixels within each county (defined by the distance to the county's mean NDVI) to reduce the influence of extreme values and then extract the values of 500 of these pixels. The amount of pixels sampled was determined by first inspecting the pixel count for all and experimenting across a 100 to 1000 sampling range. Where necessary, repeated sampling is performed to obtain a uniform number of pixels across all counties and years. For all experiments, we find the optimal parameters that will minimize the sum of the squared difference in observed (y_i) and predicted ($\hat{y}_i$) yield as seen in Equation (4). For deep learning models, we use the Adam optimizer [54] initialized with momentum parameters β_1, β_2 = 0.9, 0.99. Our validation set serves as the reference for hyperparameter optimization and consists of 30% of the counties within each state, with the remaining 70% used for training. This split is based on the number of counties and not individual samples to ensure spatial separation between training and validation data. The model configuration that results in the lowest validation loss is then used to predict outcomes for our independent test set (2020 and 2021), which are the recent years in our dataset.

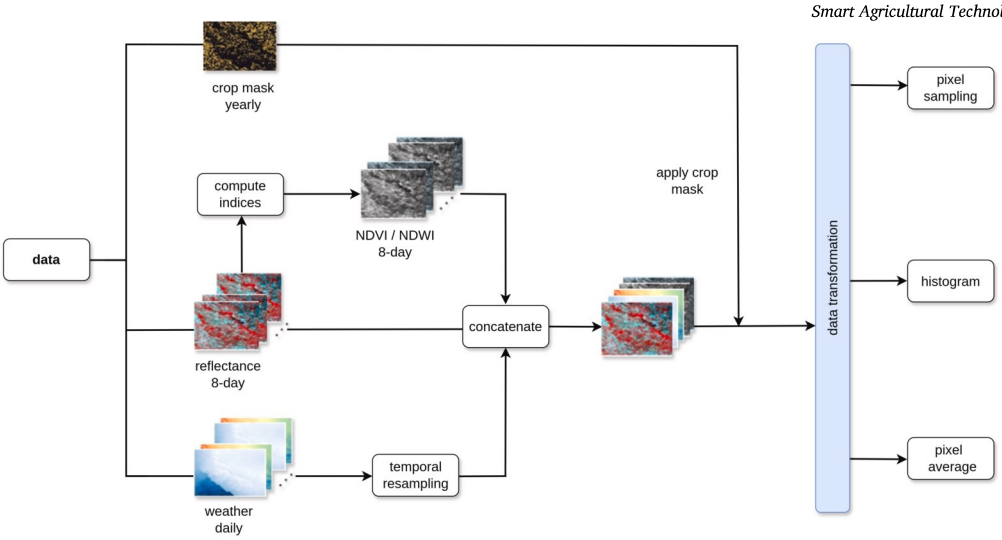

Fig. 10. Overview of the data preparation workflow, demonstrating the pipeline for a single year spatio-temporal time series data.

$$\mathcal{J}(y_i, \hat{y}_i) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{4}$$

5.3. Evaluation metrics

The experiments are quantitatively evaluated using standard statistical metrics, namely mean absolute percentage error (MAPE), root mean squared error (RMSE), and coefficient of determination (R^2). $\bar{y}$ is the mean of the observed corn yield and n is the number of data points.

$$MAPE = \frac{100}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \tag{5}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{6}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{7}$$

6. Results and discussion

In this section, we report the performances of the spatio-temporal models and further investigate, using MSResNet, the influence of the time window selection, feature combination, in-season forecasting and the impact of the previous year's information on the model's performance. The MSResNet is selected for further analysis due to its unexplored potential in this use case compared to single-kernel temporal convolution.

6.1. Comparing spatial-temporal encoding techniques

Table 3 summarizes model performance across three metrics: MAPE, RMSE, and R^2. The average MAPE on the test sets is generally below 10%, and the RMSE is higher in the test year 2021 compared to 2020, likely due to excluding 2020 samples from the training data, a point discussed further in later sections. Among models using time-series pixel averages, SVM outperformed other baselines that do not account for temporal order, as well as the LSTM and Histogram-2D models. This contrasts with Table 1, which suggests that classical methods generally underperform relative to deep learning models. When metrics are averaged across both test years, TempCNN achieves the best overall performance, as further evidenced by Fig. 11, where it reduces high percentage difference errors. However, its improvement over MSResNet and SVM is modest. We observed that the performance of LTAE declines when the pixel-set encoder (PSE) is introduced, possibly due to PSE's limitations in handling sampling within large prediction units where there is high spectral variation, unlike the more consistent conditions in farmlands.

Table 2
Model complexity by number of learnable parameters. Each model is optimized separately using Optuna [55], an automatic hyperparameter optimization framework.

Model	Number of parameters
RF/XGBoost/SVM	-
MLP	297 K
TempCNN	488 K
LSTM	466 K
LSTM-Attn	52 K
MSResNet	8M
InceptionTime	208 K
LTAE	65 K
Histogram-LSTM	54M
Histogram-2D CNN	23M
Histogram-TempCNN	3M
PSE-LTAE	3M

Flattening the histogram and modeling the temporal component with 1D temporal convolution proves more efficient than using an LSTM or applying a spatial CNN to histogram images (Histogram-2D), as it results in a less complex model (see Table 2). Introducing attention into LSTMs can result in a less complex model with performance comparable to standard LSTMs, provided that hyperparameters are carefully tuned. Meanwhile, prior studies have reported that attention-based LSTMs can significantly reduce RMSE up to 40% in yield prediction tasks [18].

6.2. Assessing feature importance via grouped feature analysis

Our initial feature selection was informed by existing studies [16,21] and the impact of weather on crop yield in our study area [3]. As illustrated in Fig. 9, the effects of one of the most prominent droughts can be observed in our choice of features, suggesting that they can capture yield variations. To further isolate feature contributions, we explored how different subsets of features performed relative to the entire set of features. Specifically, we conducted independent experiments using spectral features, satellite-derived spectral indices, weather features, or their combinations. From Table 4, spectral features (surface reflectance) consistently perform well independently, particularly in terms of R^2, indicating that it captures a substantial amount of variability in yield as

Table 3
Performance of the different machine learning models. RMSE is reported in bushels per acre (bu/acre) and bold values reflect the lowest RMSE column-wise.

Model	Test(2020)			Test(2021)		
	MAPE	RMSE	R^2	MAPE	RMSE	R^2
RF	$7.11^{\pm 0.07}$	$15.74^{\pm 0.18}$	$0.48^{\pm 0.02}$	$8.12^{\pm 0.28}$	$18.60^{\pm 0.53}$	$0.57^{\pm 0.03}$
XGBoost	$6.93^{\pm 0.04}$	$15.24^{\pm 0.07}$	$0.51^{\pm 0.01}$	$7.73^{\pm 0.23}$	$17.80^{\pm 0.44}$	$0.60^{\pm 0.02}$
SVM	$6.46^{\pm 0.37}$	$14.21^{\pm 0.80}$	$0.58^{\pm 0.05}$	$7.48^{\pm 0.23}$	$16.52^{\pm 0.50}$	$0.66^{\pm 0.02}$
MLP	$7.40^{\pm 0.18}$	$15.61^{\pm 0.34}$	$0.49^{\pm 0.02}$	$7.20^{\pm 0.14}$	$16.44^{\pm 0.36}$	$0.66^{\pm 0.01}$
LSTM	$6.76^{\pm 0.26}$	$14.59^{\pm 0.58}$	$0.56^{\pm 0.04}$	$8.56^{\pm 0.49}$	$19.07^{\pm 0.93}$	$0.54^{\pm 0.04}$
LSTM (attention)	$6.85^{\pm 0.52}$	$15.01^{\pm 1.09}$	$0.53^{\pm 0.07}$	$8.38^{\pm 0.34}$	$18.67^{\pm 0.77}$	$0.56^{\pm 0.04}$
TempCNN	$6.48^{\pm 0.52}$	$13.89^{\pm 0.95}$	$0.60^{\pm 0.05}$	$\mathbf{7.00^{\pm 0.32}}$	$\mathbf{16.05^{\pm 0.87}}$	$\mathbf{0.68^{\pm 0.03}}$
MSResNet	$6.23^{\pm 0.19}$	$13.77^{\pm 0.37}$	$0.60^{\pm 0.02}$	$7.27^{\pm 0.38}$	$16.93^{\pm 1.17}$	$0.64^{\pm 0.05}$
InceptionTime	$6.99^{\pm 0.55}$	$15.14^{\pm 0.57}$	$0.52^{\pm 0.04}$	$7.13^{\pm 0.55}$	$16.36^{\pm 1.32}$	$0.66^{\pm 0.05}$
LTAE	$8.07^{\pm 0.67}$	$17.37^{\pm 1.39}$	$0.37^{\pm 0.10}$	$7.55^{\pm 0.68}$	$17.42^{\pm 1.90}$	$0.62^{\pm 0.09}$
Histogram-LSTM	$6.57^{\pm 0.26}$	$14.28^{\pm 0.56}$	$0.57^{\pm 0.03}$	$7.92^{\pm 0.66}$	$18.13^{\pm 0.98}$	$0.59^{\pm 0.05}$
Histogram-TempCNN	$\mathbf{6.21^{\pm 0.16}}$	$\mathbf{13.65^{\pm 0.30}}$	$\mathbf{0.61^{\pm 0.02}}$	$7.53^{\pm 0.57}$	$17.37^{\pm 1.22}$	$0.62^{\pm 0.05}$
Histogram-2D CNN	$6.81^{\pm 0.44}$	$14.99^{\pm 0.65}$	$0.53^{\pm 0.04}$	$7.78^{\pm 1.47}$	$17.78^{\pm 2.96}$	$0.60^{\pm 0.14}$
PSE-LTAE	$8.93^{\pm 0.37}$	$19.08^{\pm 0.59}$	$0.24^{\pm 0.05}$	$12.63^{\pm 2.0}$	$27.18^{\pm 3.54}$	$0.07^{\pm 0.24}$

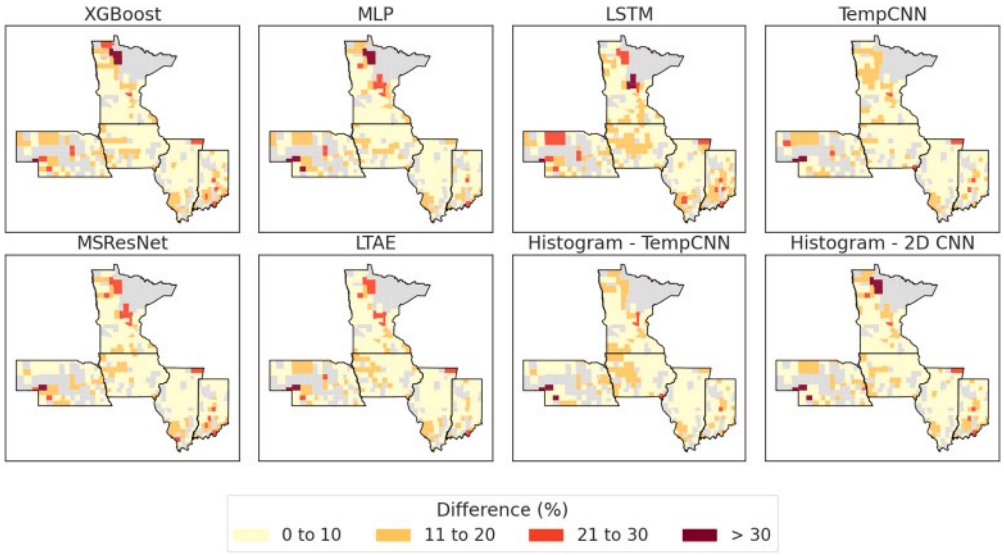

Fig. 11. Maps showing the percentage difference in observed and predicted corn yield for 2021. Areas with no data are depicted in gray.

reported in related studies [15,16]. Fig. 9 also illustrates that the influence of weather is evident in the surface reflectance bands, and the correlation between weather and spectral features or indices has been substantiated in the literature [56,57]. Models incorporating surface reflectance and either weather or spectral indices achieved the best overall performance and stability across different test years. Weather features alone show moderate performance on the validation set but exhibit significant errors during generalization, indicating that this combination is far worse than using the mean of the target values. Including all features does not necessarily equate to improved performance [16,27]. The performance gain observed with SR, or SR combined with spectral indices or weather variables, compared to using all features, may be model-specific, as our findings are based on a single model demonstration.

While it is generally expected that combining diverse data sources enhances model performance [28], our results showing that SR alone provides the best configuration can be attributed to several factors. First, SR bands directly capture key vegetation properties (e.g., chlorophyll and biomass), while spectral indices are derived from these same bands, often adding redundant than complementary information. Second, the early fusion strategy (feature concatenation) used in our setup may not adequately handle heterogeneity between data sources, limiting the model's ability to extract additional value from non-SR inputs. Finally, increasing the number of input features expands the dimensionality of the input space, which may lead to overfitting, especially when the added features are not sufficiently informative. In combination, these factors can dilute the model's learning signal and result in reduced generalization performance. Sophisticated fusion strategies can be explored to better integrate heterogeneous data sources [58,22,59], while feature selection methods may help to more effectively leverage complementary information [28].

6.3. *Influence of time window selection on model performance*

Considering that the factors influencing yield and their patterns over a long-term window can change, we study the impact of a reduced time window (consequently training data size) on the generalization capability of the MSResNet model. For this experiment, we truncate the year range to generate samples for building a model by considering a 4-year (2016-2019) and an 8-year window (2012-2019) before 2020. These specific time windows were selected based on two criteria (i) analysis of drought patterns from the US drought monitor [60] which highlighted significant interannual variability during this period, and (ii) the need to balance recency with data sufficiency. The 4-year window captures only the most recent trends, which may reflect current climatic and management conditions. However, it results in a smaller number of training samples. In contrast, the 8-year window offers a broader view of his-

Table 4
The performance of MSResNet on the different feature groups. RMSE is reported in bushels per acre (bu/acre), and bold values reflect the lowest RMSE column-wise. SR denotes surface reflectance.

Feature	Test(2020)			Test(2021)		
	MAPE	RMSE	R^2	MAPE	RMSE	R^2
SR	**$5.56^{\pm 0.14}$**	**$12.70^{\pm 0.15}$**	**$0.66^{\pm 0.01}$**	**$6.33^{\pm 0.36}$**	**$15.30^{\pm 1.37}$**	**$0.71^{\pm 0.05}$**
Weather	$12.77^{\pm 1.29}$	$27.34^{\pm 2.57}$	$-0.57^{\pm 0.30}$	$14.62^{\pm 1.74}$	$31.48^{\pm 3.38}$	$-0.25^{\pm 0.27}$
Indices	$6.96^{\pm 0.67}$	$16.40^{\pm 1.48}$	$0.44^{\pm 0.1}$	$7.74^{\pm 0.47}$	$18.08^{\pm 0.97}$	$0.59^{\pm 0.04}$
Indices+Weather	$6.62^{\pm 0.25}$	$14.69^{\pm 0.27}$	$0.55^{\pm 0.02}$	$7.25^{\pm 0.24}$	$17.08^{\pm 0.62}$	$0.63^{\pm 0.03}$
SR+Weather	$5.96^{\pm 0.24}$	$13.16^{\pm 0.57}$	$0.64^{\pm 0.03}$	$6.80^{\pm 0.17}$	$15.91^{\pm 0.67}$	$0.68^{\pm 0.03}$
SR+Indices	$5.81^{\pm 0.37}$	$12.85^{\pm 0.66}$	$0.66^{\pm 0.04}$	$6.55^{\pm 0.6}$	$15.85^{\pm 1.33}$	$0.60^{\pm 0.07}$
All features	$6.23^{\pm 0.19}$	$13.77^{\pm 0.37}$	$0.60^{\pm 0.02}$	$7.27^{\pm 0.38}$	$16.93^{\pm 1.17}$	$0.64^{\pm 0.05}$

Table 5
The performance of MSResNet on different sizes of the training data. RMSE is reported in bushels per acre (bu/acre), and bold values reflect the lowest RMSE column-wise.

Window	Test(2020)			Test(2021)		
	MAPE	RMSE	R^2	MAPE	RMSE	R^2
MSResNet (4 years)	$6.67^{\pm 0.64}$	$15.16^{\pm 1.32}$	$0.52^{\pm 0.09}$	$6.89^{\pm 0.32}$	$15.46^{\pm 0.64}$	$0.70^{\pm 0.02}$
MSResNet (8 years)	$5.92^{\pm 0.58}$	**$13.42^{\pm 1.08}$**	**$0.62^{\pm 0.06}$**	**$6.39^{\pm 0.41}$**	**$14.89^{\pm 0.85}$**	**$0.72^{\pm 0.03}$**
MSResNet (17 years)	$6.23^{\pm 0.19}$	$13.77^{\pm 0.37}$	$0.60^{\pm 0.02}$	$7.27^{\pm 0.38}$	$16.93^{\pm 1.17}$	$0.64^{\pm 0.05}$

Table 6
In-season yield prediction performance of MSResNet. Experiments rely on data from the start of the season until mid-season and with progressive increments of two timesteps (16 days). RMSE is reported in bushels per acre (bu/acre), and bold values reflect the lowest RMSE column-wise.

Date ending	Test(2020)			Test(2021)		
	MAPE	RMSE	R^2	MAPE	RMSE	R^2
July 28	$6.81^{\pm 0.29}$	$15.35^{\pm 0.55}$	$0.51^{\pm 0.03}$	$7.59^{\pm 0.26}$	$18.09^{\pm 0.42}$	$0.59^{\pm 0.02}$
August 13	$6.41^{\pm 0.06}$	$14.44^{\pm 0.22}$	$0.57^{\pm 0.01}$	$7.05^{\pm 0.58}$	$17.33^{\pm 1.29}$	$0.62^{\pm 0.06}$
August 28	**$5.60^{\pm 0.32}$**	**$12.78^{\pm 0.61}$**	**$0.66^{\pm 0.03}$**	$6.94^{\pm 1.05}$	$16.73^{\pm 2.20}$	$0.65^{\pm 0.10}$
September 14	$5.92^{\pm 0.45}$	$13.46^{\pm 0.91}$	$0.62^{\pm 0.05}$	$7.04^{\pm 0.67}$	$16.97^{\pm 1.14}$	$0.64^{\pm 0.05}$
September 30	$5.93^{\pm 0.33}$	$13.12^{\pm 0.89}$	$0.64^{\pm 0.05}$	$6.34^{\pm 0.35}$	**$15.47^{\pm 0.79}$**	**$0.70^{\pm 0.03}$**
October 24	$6.23^{\pm 0.19}$	$13.77^{\pm 0.37}$	$0.60^{\pm 0.02}$	$7.27^{\pm 0.38}$	$16.93^{\pm 1.17}$	$0.64^{\pm 0.05}$

torical variability and includes more extreme events, such as the severe drought of 2012 [3], while avoiding earlier years that may no longer be relevant due to shifts in production practices or climate. Within this 8-year window, the average yield across most states showed improvement compared to earlier years (pre-2012), apart from the notable dip during the 2012 drought. This setup allows us to evaluate whether recency or volume of training data has a greater impact on model generalization. Compared to the baseline results, where 17 years of data are considered, the 8-year window generalized better to both test years (Table 5). The model's improved performance using recent data suggests that concept drift [61] may be present in the data. An adaptive training approach, such as weighting recent data more, may be appropriate when considering longer historical data to ensure the model's reliance on recent trends. The poorer performance observed for the 4-year window may be attributed to the limited amount of training data, which restricts the model's ability to capture the diversity and variability in yield-influencing factors over time.

6.4. In-season forecasting

Predicting yield both in-season and at the end of the season is crucial for effective agricultural management. These predictions enable timely interventions to prevent disruptions in the food supply chain and provide valuable insights for crop marketing and distribution planning. Numerous studies on in-season yield forecasting have shown that models become increasingly reliable as more temporal information becomes available [15,16,21]. In this context, forecasting involves progressively increasing the sequence of time steps leading up to harvest. In contrast, [19] relied on long-term climate forecasts (combined with static soil parameters) to achieve a much longer lead time. However, long-term forecasts may be subject to high uncertainties that make them unreliable [62]. We address forecasting using the former approach with the end of July (mid-season) as our baseline and gradually adding 2 time steps (16 days) until the usual harvest dates. As shown in Table 6, longer time spans improve yield prediction accuracy in-season, but the performance gain is not linear. Optimal performance is observed around late August. After this period, extending the time steps provides diminishing returns. The variability in yield estimation across years points to the inherent challenge of generalizing machine learning models for time series yield prediction. Our results emphasize the importance of capturing seasonal patterns accurately while managing the complexity introduced by extending the prediction timeline.

6.5. Including the previous year's data in training

Our experiment setup mimics a scenario where both test years are treated independently. Meanwhile, the previous year's season may bear the closest resemblance to the current season since adjustments to management practices or the effects of extreme events can linger into the current season. It has been established that the inclusion of the previous year's yield as a feature can enhance prediction accuracy [28]. In our case, we make a broader assumption that the absence of the training samples (including features and yield) from the previous year can reduce the predictability of a present season's yield. To evaluate the validity of this assumption, we redesigned our experiment by incorporating samples from 2020 into the training and validation sets while reserving 2021

exclusively as the test set. The MSResNet model under this configuration achieved an MAPE = $6.36^{\pm 0.38}$, RMSE = $14.89^{\pm 1.03}$ and $R^2 = 0.72^{\pm 0.04}$. Compared to the scenario where 2020 was omitted, the RMSE is reduced by 2 units, and the R^2 improved by 12%. We compare our results to existing studies using ML for corn yield prediction in the USA. From Table 1, models for corn yield prediction based on 1D CNN-LSTM and LSTM (with attention) reported an RMSE of around 17 bu/acre on the test sets. Our MSResNet model outperforms these references, achieving 13.77 in 2020 and 14.89 in 2021 (when 2020 is included in its training set), albeit under different experiment setups and input data. Our result based on light-weight time series data is comparable to the performances of ConvLSTM and MMST-ViT [36], and better than multi-modal Histogram-LSTM models [22], which are trained on a massive amount of high-resolution spatio-temporal satellite data (Sentinel-2). However, the MMST-ViT becomes advantageous when pre-trained via self-supervised learning.

7. Conclusion

Machine learning models have seen significant advancements, leading to their growing application in yield prediction. These models encompass a wide range of architectural designs, requiring simple to complex input data structures. Although existing research highlights progress in ML-based yield prediction, the uniqueness of each study in terms of data inputs and experiment setup challenges their intercomparison. This study provides a comprehensive comparison of various ML techniques for corn yield prediction, highlighting different ways to encode spatial and temporal information. We underscore the predictive strength of time series data without spatial features and the benefits of surface reflectance data. Moreover, excluding previous-year data in training decreases prediction accuracy for the current year, reinforcing the value of temporal continuity in training. Our time-controlled experiments confirm the effectiveness of ML models for in-season predictions. While our experiments focused on commonly used deep learning models, future work could explore more complex hybrid architectures, such as CNN-Transformer combinations, which combines local and global feature extraction and have shown promise in related applications [63]. Similarly, pretrained models that leverage large-scale remote sensing datasets may offer improved generalization and reduced training requirements. These directions, alongside adaptive learning strategies that prioritize recent data, could enhance the robustness of crop yield prediction models under changing environmental and management conditions.

CRediT authorship contribution statement

Stella Ofori-Ampofo: Writing – original draft, Visualization, Methodology, Formal analysis, Data curation, Conceptualization. **Ridvan Salih Kuzu:** Writing – review & editing, Validation, Supervision, Methodology, Conceptualization. **Peter Schauer:** Writing – review & editing, Supervision, Conceptualization. **Martin Willberg:** Writing – review & editing, Supervision, Conceptualization. **Adrian Höhl:** Writing – review & editing, Data curation. **Xiao Xiang Zhu:** Writing – review & editing, Supervision, Resources, Methodology, Conceptualization.

Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work the authors used ChatGPT v2 and Grammarly in order to improve the readability and diction of the manuscript. After using this tool/service, the authors reviewed and edited the content as needed and take full responsibility for the content of the published article.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

The work of S. Ofori-Ampofo was funded by the Munich Aerospace e.V. scholarship. The work of A. Höhl was funded by the project ML4Earth by the German Federal Ministry for Economic Affairs and Climate Action under grant number 50EE2201C. The authors are responsible for the content of this publication.

Appendix A

Table A.1
MODIS band designation with corresponding wavelength range. IR refers to infrared.

Band	Description	Wavelength (nm)
band 1	red	620 - 670
band 2	near IR	841 - 876
band 3	blue	459 - 479
band 4	green	545 - 565
band 5	near IR	1230 - 1250
band 6	short wave IR	1628 - 1652
band 7	short wave IR	2105 - 2155

Data availability

The dataset used is the first version of a multi-sensor US-wide dataset currently being collated by the Chair of Data Science in Earth Observation at the Technical University of Munich (under the MONITOR and ML4Earth project) to facilitate methodological advances in remote sensing for crop monitoring and climate-related applications. The dataset and the code supporting this study are available in our GitHub repository: https://github.com/ellaampy/SpatioTemporalYield.

References

[1] FAO, The future of food and agriculture – trends and challenges, Tech. Rep., FAO, 2017, http://www.fao.org/3/a-i6583e.pdf.
[2] FAO, IFAD, UNICEF, WFP, WHO, The state of food security and nutrition in the world 2021. Transforming food systems for food security, improved nutrition and affordable healthy diets for all, Tech. Rep., FAO and IFAD and UNICEF and WFP and WHO, 2021, https://doi.org/10.1016/j.jag.2021.102436.
[3] B.R. Rippey, The U.S. drought of 2012, Weather and Climate Extremes 10, 2015, pp. 57–64, uSDA Research and Programs on Extreme Events, https://doi.org/10.1016/j.wace.2015.10.004.
[4] B. Sultan, D. Defrance, T. Iizumi, Evidence of crop production losses in West Africa due to historical global warming in two crop models, Sci. Rep. 9 (2019), https://doi.org/10.1038/s41598-019-49167-0.
[5] S. Chen, X. Chen, J. Xu, Impacts of climate change on agriculture: evidence from China, J. Environ. Econ. Manag. 76 (2015), https://doi.org/10.1016/j.jeem.2015.01.005.
[6] C. Deutsch, J. Tewksbury, M. Tigchelaar, D. Battisti, S. Merrill, R. Huey, R. Naylor, Increase in crop losses to insect pests in a warming climate, Science 361 (2018) 916–919, https://doi.org/10.1126/science.aat3466.
[7] M.A. Arias, A.M. Ibáñez Londoño, J.A. Zambrano Riveros, et al., Agricultural Production amid Conflict: the Effects of Shocks, Uncertainty, and Governance of Non-state Armed Actors, 2014.
[8] FAO, The state of food security and nutrition in the world 2021. Transforming food systems for food security, improved nutrition and affordable healthy diets for all, Tech. Rep., FAO, July 2022, https://doi.org/10.4060/cc1025en.
[9] R. Miao, M. Khanna, H. Huang, Responsiveness of crop yield and acreage to prices and climate, Am. J. Agric. Econ. 98 (2015), https://doi.org/10.1093/ajae/aav025.
[10] C.v. Van Diepen, J.v. Wolf, H. Van Keulen, C. Rappoldt, Wofost: a simulation model of crop production, Soil Use Manag. 5 (1) (1989) 16–24.
[11] R. McCown, G. Hammer, J. Hargreaves, D. Holzworth, D. Freebairn, Apsim: a novel software system for model development, model testing and simulation in agricultural systems research, Agric. Syst. 50 (3) (1996) 255–271, https://doi.org/10.1016/0308-521X(94)00055-V.

[12] G. Leng, J. Hall, Predicting spatial and temporal variability in crop yields: an inter-comparison of machine learning, regression and process-based models, Environ. Res. Lett. 15 (2020), https://doi.org/10.1088/1748-9326/ab7b24.

[13] S. Khaki, L. Wang, S.V. Archontoulis, A cnn-rnn framework for crop yield prediction, Front. Plant Sci. 10 (2020), https://doi.org/10.3389/fpls.2019.01750.

[14] D. Paudel, H. Boogaard, A. Wit, S. Janssen, S. Osinga, C. Pylianidis, I. Athanasiadis, Machine learning for large-scale crop yield forecasting, Agric. Syst. 187 (2020) 103016, https://doi.org/10.1016/j.agsy.2020.103016.

[15] M. Qiao, X. He, X. Cheng, P. Li, H. Luo, L. Zhang, Z. Tian, Crop yield prediction from multi-spectral, multi-temporal remotely sensed imagery using recurrent 3d convolutional neural networks, Int. J. Appl. Earth Obs. Geoinf. 102 (2021) 102436, https://doi.org/10.1016/j.jag.2021.102436.

[16] J. You, X. Li, M. Low, D. Lobell, S. Ermon, Deep Gaussian process for crop yield prediction based on remote sensing data, in: AAAI, 2017, pp. 4559–4566.

[17] M. Bernardi, J. Delince, W. Durand, N. Zhang, Crop Yield Forecasting: Methodological and Institutional Aspects, 2016.

[18] A. Kaur, P. Goyal, R. Rajhans, L. Agarwal, N. Goyal, Fusion of multivariate time series meteorological and static soil data for multistage crop yield prediction using multi-head self attention network, Expert Syst. Appl. 226 (2023) 120098, https://doi.org/10.1016/j.eswa.2023.120098.

[19] I. Oliveira, R.L. de Freitas Cunha, B.L.B. Silva, M.A.S. Netto, A scalable machine learning system for pre-season agriculture yield forecast, in: 2018 IEEE 14th International Conference on e-Science (e-Science), 2018, pp. 423–430.

[20] V.S.F. Garnot, L. Landrieu, S. Giordano, N. Chehata, Satellite image time series classification with pixel-set encoders and temporal self-attention, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 12325–12334.

[21] H. Russello, W. Shang, Convolutional Neural Networks for Crop Yield Prediction Using Satellite Images, 2018.

[22] A. Kaur, P. Goyal, K. Sharma, L. Sharma, N. Goyal, A generalized multimodal deep learning model for early crop yield prediction, in: 2022 IEEE International Conference on Big Data (Big Data), 2022, pp. 1272–1279.

[23] Z. Wang, W. Yan, T. Oates, Time series classification from scratch with deep neural networks: a strong baseline, in: 2017 International Joint Conference on Neural Networks (IJCNN), 2017, pp. 1578–1585.

[24] H. Ismail Fawaz, B. Lucas, G. Forestier, C. Pelletier, D.F. Schmidt, J. Weber, G.I. Webb, L. Idoumghar, P.-A. Muller, F. Petitjean, Inceptiontime: finding alexnet for time series classification, Data Min. Knowl. Discov. 34 (6) (2020) 1936–1962.

[25] V.S.F. Garnot, L. Landrieu, Lightweight temporal self-attention for classifying satellite images time series, in: V. Lemaire, S. Malinowski, A. Bagnall, T. Guyet, R. Tavenard, G. Ifrim (Eds.), Advanced Analytics and Learning on Temporal Data, Springer International Publishing, Cham, 2020, pp. 171–181.

[26] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural Comput. 9 (8) (1997) 1735–1780, https://doi.org/10.1162/neco.1997.9.8.1735.

[27] C. Pelletier, G. Webb, F. Petitjean, Temporal convolutional neural network for the classification of satellite image time series, Remote Sens. 11 (2019) 523, https://doi.org/10.3390/rs11050523.

[28] Y. Wang, Z. Zhang, L. Feng, Q. Du, T. Runge, Combining multi-source data and machine learning approaches to predict winter wheat yield in the conterminous United States, Remote Sens. 12 (8) (2020), https://doi.org/10.3390/rs12081232, https://www.mdpi.com/2072-4292/12/8/1232.

[29] T. van Klompenburg, A. Kassahun, C. Catal, Crop yield prediction using machine learning: a systematic literature review, Comput. Electron. Agric. 177 (2020) 105709, https://doi.org/10.1016/j.compag.2020.105709.

[30] C. Zhao, B. Liu, S. Piao, X. Wang, D.B. Lobell, Y. Huang, M. Huang, Y. Yao, S. Bassu, P. Ciais, J.-L. Durand, J. Elliott, F. Ewert, I.A. Janssens, T. Li, E. Lin, Q. Liu, P. Martre, C. Müller, S. Peng, J. Peñuelas, A.C. Ruane, D. Wallach, T. Wang, D. Wu, Z. Liu, Y. Zhu, Z. Zhu, S. Asseng, Temperature increase reduces global yields of major crops in four independent estimates, Proc. Natl. Acad. Sci. USA 114 (35) (2017) 9326–9331, https://doi.org/10.1073/pnas.1701762114, https://www.pnas.org/doi/pdf/10.1073/pnas.1701762114.

[31] R. Schwalbert, T. Amado, G. Corassa, L. Pott, P.V.V. Prasad, I. Ciampitti, Satellite-based soybean yield forecast: integrating machine learning and weather data for improving crop yield prediction in southern Brazil, Agric. For. Meteorol. 284 (2020), https://doi.org/10.1016/j.agrformet.2019.107886.

[32] D. Jiang, X. Yang, N. Clinton, N. Wang, An artificial network model for estimating crop yields using remotely sensed information, Int. J. Remote Sens. 25 (2004) 1723–1732, https://doi.org/10.1080/0143116031000150068.

[33] A. Prasad, L. Chai, R. Singh, M. Kafatos, Crop yield estimation model for Iowa using remote sensing and surface parameters, Int. J. Appl. Earth Obs. Geoinf. 8 (2006) 26–33, https://doi.org/10.1016/j.jag.2005.06.002.

[34] C. Piedallu, V. Cheret, J.-P. Denux, V. Perez, J.S. Azcona, I. Seynave, J.-C. Gegout, Soil and climate differently impact ndvi patterns according to the season and the stand type, Sci. Total Environ. 651 (2019) 2874–2885.

[35] M. Rußwurm, C. Pelletier, M. Zollner, S. Lefèvre, M. Körner, Breizhcrops: a time series dataset for crop type mapping, arXiv preprint, arXiv:1905.11893, 2019.

[36] F. Lin, S. Crawford, K. Guillot, Y. Zhang, Y. Chen, X. Yuan, L. Chen, S. Williams, R. Minvielle, X. Xiao, et al., Mmst-vit: climate change-aware crop yield prediction via multi-modal spatial-temporal vision transformer, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 5774–5784.

[37] H. Tian, P. Wang, K. Tansey, D. Han, J. Zhang, S. Zhang, H. Li, A deep learning framework under attention mechanism for wheat yield estimation using remotely sensed indices in the guanzhong plain, PR China, Int. J. Appl. Earth Obs. Geoinf. 102 (2021) 102375, https://doi.org/10.1016/j.jag.2021.102375.

[38] F. Liu, X. Jiang, Z. Wu, Attention mechanism-combined lstm for grain yield prediction in China using multi-source satellite imagery, Sustainability 15 (2023) 9210, https://doi.org/10.3390/su15129210.

[39] K. Gavahi, P. Abbaszadeh, H. Moradkhani, Deepyield: a combined convolutional neural network with long short-term memory for crop yield forecasting, Expert Syst. Appl. 184 (2021) 115511, https://doi.org/10.1016/j.eswa.2021.115511.

[40] M. Shahhosseini, G. Hu, I. Huber, S. Archontoulis, Coupling machine learning and crop modeling improves crop yield prediction in the US corn belt, Sci. Rep. 11 (2021), https://doi.org/10.1038/s41598-020-80820-1.

[41] X.X. Zhu, D. Tuia, L. Mou, G.-S. Xia, L. Zhang, F. Xu, F. Fraundorfer, Deep learning in remote sensing: a comprehensive review and list of resources, IEEE Geosci. Remote Sens. Mag. 5 (4) (2017) 8–36, https://doi.org/10.1109/MGRS.2017.2762307.

[42] H.A. Dau, A.J. Bagnall, K. Kamgar, C.M. Yeh, Y. Zhu, S. Gharghabi, C.A. Ratanamahatana, E.J. Keogh, The UCR time series archive, CoRR, arXiv:1810.07758 [abs], 2018, arXiv:1810.07758.

[43] M. Rußwurm, M. Körner, Self-attention for raw optical satellite time series classification, ISPRS J. Photogramm. Remote Sens. 169 (2020) 421–435, https://doi.org/10.1016/j.isprsjprs.2020.06.006, https://www.sciencedirect.com/science/article/pii/S0924271620301647.

[44] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, arXiv:1706.03762, 2017.

[45] U. S. D. o. A. National Agricultural Statistics Service, Crop Production 2021 Summary, ISSN 1057-7823, 2022.

[46] A. Höhl, S. Ofori-Ampofo, I. Obadic, M.-Á. Fernández-Torres, R. Salih Kuzu, X. Zhu, Uscc: a benchmark dataset for crop yield prediction under climate extremes, in: EGU General Assembly Conference Abstracts, 2023, pp. EGU–15540.

[47] U. S. D. o. A. National National Agricultural Statistics Service, Crop Production 2021 Summary, ISSN: 1995-2004 2022.

[48] N. Kim, K.-J. Ha, N.-W. Park, J. Cho, S. Hong, Y.-W. Lee, A comparison between major artificial intelligence models for crop yield prediction: case study of the midwestern United States, 2006–2015, ISPRS Int. J. Geo-Inf. 8 (5) (2019), https://doi.org/10.3390/ijgi8050240.

[49] U. S. D. o. A. national agricultural statistics service, cropland data layer metadata, https://www.nass.usda.gov/Research_and_Science/Cropland/metadata/meta.php, 2022. (Accessed 1 January 2022).

[50] NASA LP DAAC, Mod09a1 version 6: modis surface reflectance 8-day l3 global 500 m sin grid, https://doi.org/10.5067/MODIS/MOD09A1.006, 2015. (Accessed 5 June 2022).

[51] B. cai Gao, Ndwi—a normalized difference water index for remote sensing of vegetation liquid water from space, Remote Sens. Environ. 58 (3) (1996) 257–266, https://doi.org/10.1016/S0034-4257(96)00067-3.

[52] P. Thornton, R. Shrestha, T. Michele, S.-C. Kao, Y. Wei, B. Wilson, Gridded daily weather data for North America with comprehensive uncertainty quantification, Sci. Data 8 (2021), https://doi.org/10.1038/s41597-021-00973-0.

[53] U. S. D. o. A. national agricultural statistics service, field crops: usual planting and harvesting dates, https://usda.library.cornell.edu/concern/publications/vm40xr56k, October 2010.

[54] D.P. Kingma, Adam: a method for stochastic optimization, arXiv preprint, arXiv:1412.6980, 2014.

[55] T. Akiba, S. Sano, T. Yanase, T. Ohta, M. Koyama, Optuna: a next-generation hyperparameter optimization framework, in: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2019, pp. 2623–2631.

[56] P. Schultz, M. Halpert, Global correlation of temperature, ndvi and precipitation, Adv. Space Res. 13 (5) (1993) 277–280.

[57] D. Sun, M. Kafatos, Note on the ndvi-lst relationship and the use of temperature-related drought indices over North America, Geophys. Res. Lett. 34 (24) (2007).

[58] C. Pohl, J.L.V. Genderen, Review article multisensor image fusion in remote sensing: concepts, methods and applications, Int. J. Remote Sens. 19 (5) (1998) 823–854, https://doi.org/10.1080/014311698215748, https://doi.org/10.1080/014311698215748.

[59] S. Ofori-Ampofo, C. Pelletier, S. Lang, Crop type mapping from optical and radar time series using attention-based deep learning, Remote Sens. 13 (2021) 4668, https://doi.org/10.3390/rs13224668.

[60] N. D. M. Center, U. D. of Agriculture, N. Oceanic, A. Administration, United States drought monitor, https://droughtmonitor.unl.edu/, 2023. (Accessed 17 September 2023).

[61] I. Žliobaitė, M. Pechenizkiy, J. Gama, An overview of concept drift applications, in: Big Data Analysis: New Algorithms for a New Society, 2016, pp. 91–114.

[62] S.J. Johnson, T.N. Stockdale, L. Ferranti, M.A. Balmaseda, F. Molteni, L. Magnusson, S. Tietsche, D. Decremer, A. Weisheimer, G. Balsamo, S.P.E. Keeley, K. Mogensen, H. Zuo, B.M. Monge-Sanz, SEAS5: the new ECMWF seasonal forecast system, Geosci. Model Dev. 12 (3) (2019), https://doi.org/10.5194/gmd-12-1087-2019 1087–1117, publisher: Copernicus GmbH, https://gmd.copernicus.org/articles/12/1087/2019/.

[63] Y. Wang, L. Feng, W. Sun, L. Wang, G. Yang, B. Chen, A lightweight cnn-transformer network for pixel-based crop mapping using time-series sentinel-2 imagery, Comput. Electron. Agric. 226 (2024) 109370, https://doi.org/10.1016/j.compag.2024.109370, https://www.sciencedirect.com/science/article/pii/S0168169924007610.