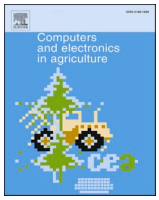
Computers and Electronics in Agriculture 239 (2025) 110965

Contents lists available at ScienceDirect

Computers and Electronics in Agriculture

journal homepage: www.elsevier.com/locate/compag



Wheat yield forecasts with seasonal climate models and long short-term memory networks

Maximilian Zachow [a], Stella Ofori-Ampofo [b], Harald Kunstmann [c,d], Rıdvan Salih Kuzu [e], Xiao Xiang Zhu [b,f], Senthold Asseng [a,*]

[a] Technical University of Munich, Department of Life Science Engineering, Digital Agriculture, HEF World Agricultural Systems Center, Freising, Germany
[b] Technical University of Munich, TUM School of Engineering and Design, Chair of Data Science in Earth Observation, Munich, Germany
[c] Institute of Geography, University of Augsburg, Augsburg, Germany
[d] Institute of Meteorology and Climate Research (IMK-IFU), Karlsruhe Institute of Technology, Campus Alpin, Garmisch-Partenkirchen, Germany
[e] Remote Sensing Technology Institute, German Aerospace Center, Wessling, Germany
[f] Munich Center for Machine Learning, Munich, Germany

ARTICLE INFO

Keywords:
Seasonal climate models
Crop yield
Wheat
Agriculture
LSTM

ABSTRACT

The potential of seasonal climate forecasts (SCFs) within machine learning models to forecast crop yields remains unexplored. We propose a workflow for integrating SCF data into a long short-term memory (LSTM) network to forecast wheat yield at the county level across the Great Plains in the United States. Each month, past predictors were filled with observations and future weather predictors were forecasted using the seasonal climate model of the European Centre for Medium-Range Weather Forecasts (SCF approach). This approach was benchmarked with the truncate approach that only used observed predictors. Using all observed predictors at harvest, the model achieved an R^2 of 0.46, an NRMSE of 0.24, and an MSE of 0.46 t/ha on the test set. The SCF approach and truncate approach performed poorly from January to March. The SCF approach outperformed the truncate approach in April and May. At the beginning of May, three months before harvest, the SCF approach achieved an MSE of 0.6 t/ha, improving the truncate approach by 10 %. In June, the SCF approach further improved but did not outperform the truncate approach. Predictor importance analysis revealed the critical role of SCF data at the beginning of May for the latter half of May. This study suggests that weather forecasts issued at the right time, when both crop development and forecast skill align, could be as short as 16 days and still significantly improve the accuracy of sub-national wheat yield forecasts over other approaches.

1. Introduction

Climate change increases interannual crop yield variabilities that challenge global food security. To navigate these uncertainties, crop yield models that provide timely estimates are crucial for stakeholders such as policymakers, food traders, food processors, NGOs and farmers (Basso and Liu, 2019). Crop yield models can be employed for both in-season forecasts and end-of-season estimations. The challenges of in-season forecasts are the unknown growth conditions for the remainder of the crop season from the forecast date to harvest. This lack of information on growth conditions hinders crop yield projections, as data about crucial development stages that affect yield formation are missing. Among the many factors influencing crop yield, weather conditions are most important, and advancements in forecasting methodologies have primarily targeted these variables (Schauberger et al., 2020). Weather conditions can be forecasted several months in advance using seasonal climate forecasts (SCFs) (Johnson et al., 2019). The application of SCFs to forecast crop yield has been assessed for process-based crop models and statistical crop models. However, for machine learning crop models, in-season crop yield forecasts rely on observed predictors from the start of the season to the forecast date (truncate approach) without using forecasted information from the forecast date until harvest (Paudel et al., 2022). This approach demonstrates the ability of machine learning models to leverage information from key crop development stages. Typically, the model's performance is evaluated by using data from the entire season and is then re-evaluated while successively excluding the latest time steps. When time steps around critical development stages are removed, a drop in performance suggests that the model has learned

* Corresponding author at: Liesel-Beckmann-Str. 2, 85354 Freising, Germany.
E-mail address: senthold.asseng@tum.de (S. Asseng).

https://doi.org/10.1016/j.compag.2025.110965
Received 26 March 2025; Received in revised form 22 August 2025; Accepted 3 September 2025
Available online 15 September 2025

to link these stages to crop yield outcomes. While this is valuable for making machine learning models more interpretable, in-season crop yield forecasts that rely solely on observed predictors are not suited for early-season forecasts because they miss most of the key phenological stages and they may overlook potential improvements that could be achieved using SCFs. Without SCFs, the implicit assumption is that weather conditions between the forecast date and harvest will not impact final crop yields. This simplification may limit the crop yield's forecast accuracy. Our study proposes a machine learning approach to forecast administrative unit-level wheat yield before harvest by combining forecasted weather predictors from a seasonal climate model and observed static and sequential predictors of soil properties, vegetation indices and weather variables. This approach directly addresses the limitation of the conventional truncate method, which disregards the impact of future weather conditions. We selected wheat as our model crop due to its high global significance for food security and substantial nutritional value. Our study targeted the U.S. Great Plains. The Great Plains is a major breadbasket region, and winter wheat yield in the Great Plains has been affected by severe droughts (Rippey, Dec. 2015) and compound hot-dry-windy events (Zhao et al., 2022), suggesting the potential usefulness of SCF data. We relied on a long short-term memory (LSTM) network (Hochreiter and Schmidhuber, Nov. 1997), which is designed to capture sequential input data interactions and has previously been used for regional crop yield models (Cunha et al., 2018). Yield data as well as static and sequential observed predictor data were obtained from the "Crop Yield Benchmark Dataset" (Paudel et al., 2024), while forecasted weather information was obtained from the European Centre of Medium-Range Weather Forecast's (ECMWF) fifth-generation seasonal forecast system (SEAS5) (Johnson et al., 2019). By incorporating forecasted weather conditions into a machine learning model, our study advances in-season crop yield forecasting for mitigating risks to food security, particularly in the face of climate change. This newly proposed SCF approach of making in-season crop yield forecasts was compared to the more popular truncate approach and benchmarked with a 3-year average model whose yield forecasts consisted of the mean yield from the previous three years.

2. Material and methods

A high-level overview of each forecasting approach is given in Fig. 1. Yield and predictor data were retrieved from the "Crop Yield Benchmark Dataset" (CY-Bench) (Paudel et al., 2024), a subnational crop yield forecasting database, which offers open-source access to analysis-ready data covering a range of meteorological and environmental variables as well as crop yield statistics. The various data sources included in CY-Bench are publicly available and were gathered based on the proven suitability of the variables for crop yield modeling.

2.1. Dataset and predictors

2.1.1. Yield data

Winter wheat (hereafter called wheat) is grown in various states across the United States. However, the Great Plains, comprising South Dakota, Nebraska, Colorado, Kansas, Oklahoma, and Texas, are often considered the heartland of U.S. wheat cultivation. This region contributes approximately 50 % of the 40 million tons of wheat produced annually in the United States (USDA/NASS, xxxx). In addition, the average wheat-harvested area constitutes a significant part of each county's size in the Great Plains (Fig. 2a), highlighting the importance of wheat farming in this region. County-level wheat yield data for the Great Plains was extracted from CY-Bench alongside the corresponding set of predictors, as shown in Table 1. The data spans a 21-year period (2003–2023) for each of the 540 counties in the study region. The wheat season begins with planting in the fall of the previous year and concludes with harvesting in the summer of the following year. Therefore, the first harvest year for which predictors are available for the entire season from planting until harvest is the harvest of 2004. In addition, there were counties with missing records for some years. Eventually, there were 6,789 yield records available for the final yield dataset with harvests from the 20-year period of 2004–2023. The average yield at the county level ranges from 1 t/ha in some parts of Texas to 4 t/ha in South Dakota (Fig. 2b). We forecasted absolute wheat yield at the county level, treating each season independently.

2.1.2. Observed predictor data

An initial assessment across the sequential predictors available in CY-Bench was performed to obtain our final set of predictors, indicated in Table 1. Namely three variables available in CY-Bench were not considered in our study. The normalized difference vegetation index was removed due to its high correlation with the fraction of absorbed photosynthetically active radiation, which has been found to provide better information for crop yield modeling (Kolotii, 2015). Surface soil moisture was removed as it is highly correlated with root zone soil moisture, which we considered the more important variable for our study (Deines et al., Sep. 2024). Lastly, potential evapotranspiration was discarded, which is the difference between precipitation and climatic water balance (both kept). The non-linear data-driven model used here could easily derive such a difference if it is relevant for wheat yield estimation. In addition to sequential predictors, static predictors are also presented. Sequential predictors were obtained in different temporal resolutions for each year from 2003 to 2023. Predictors from 2003 were needed because the earliest available yield data was for harvest in 2004,

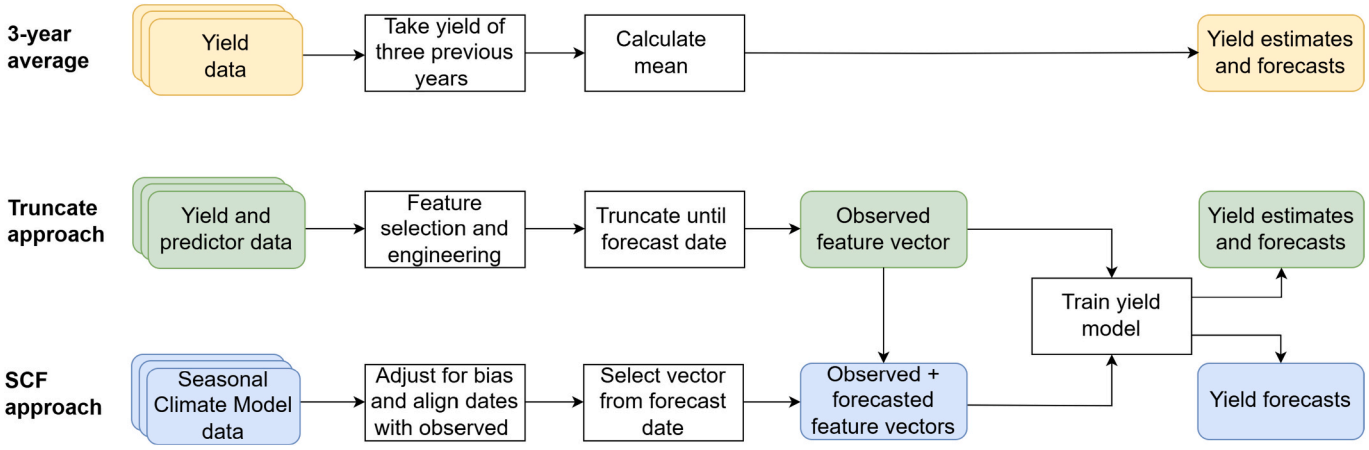


Fig. 1. Flow chart. The chart consists of three horizontal streams, each representing a different approach. It displays the input data and steps for forecasting yield. Actions are in white squares, and data is in colored rounded rectangles. For details, see section 2.1 and 2.2.

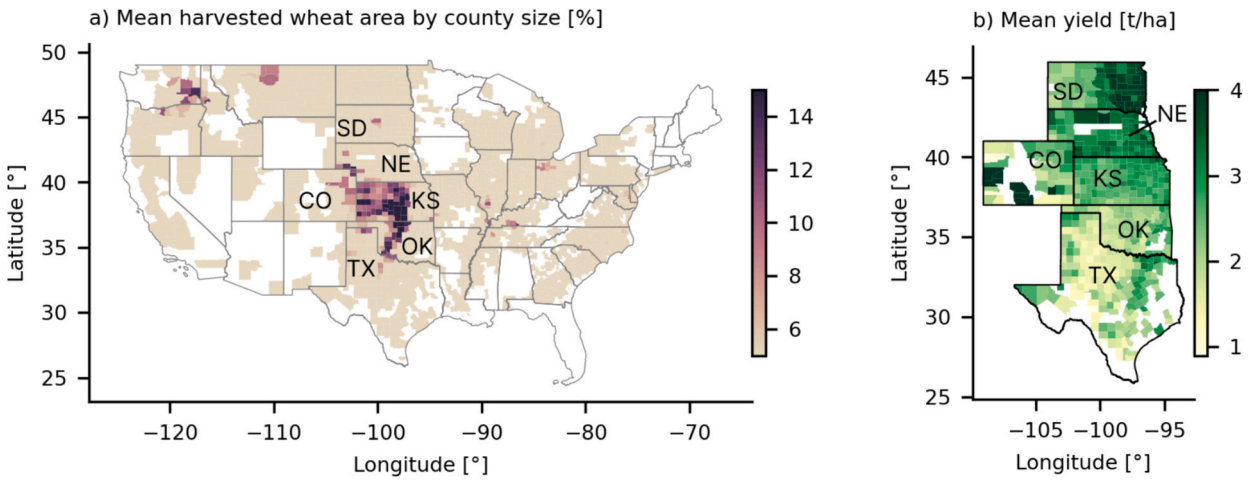


Fig. 2. Study region. (a) Average relative harvested area for wheat per county from wheat harvests 2004–2023 for the United States. (b) Average absolute wheat yield per county from wheat harvests 2004–2023 for the main wheat-producing states. South Dakota (SD), Nebraska (NE), Colorado (CO), Kansas (KS), Oklahoma (OK) and Texas (TX).

Table 1
Details of the data retrieved from CY-Bench. Crop calendar information guides our selection of the start and end of the wheat season. Data sources are indicated in the CY-Bench repository (Paudel et al., 2024).

Category	Name	Temporal resolution	Variable (unit)
Weather	temperature	daily	tmean, tmax, tmin (°C)
Weather	precipitation	daily	prec (*mm*)
Weather	solar radiation	daily	rad ($J\ m^{-2}$)
Soil	root zone soil moisture	daily	rsm (*kg m-2*)
Soil	available water capacity	constant	awc (*cm m-1*)
Soil	drainage class	constant	drainage class (*category*)
Soil	bulk density	constant	bulk density (*kg dm-3*)
Crop	fraction of absorbed photosynthetically active radiation	10 days	fpar (%)
Crop	crop calendar	constant	eos (*day of year*)
Crop	crop yield	yearly	yield (*t/ha*)
Crop	climatic water balance	daily	cwb (*mm*)
Geometry	latitude/longitude	Constant	lat, lon (°)

with the season starting in the fall of 2003. To harmonize the different temporal resolutions (daily, 10-day) of the predictors indicated in Table 1, an alignment was required. We chose 16-day time steps as a balance between capturing temporal dynamics and avoiding too much noise from very short time intervals. Daily data was averaged into 16-day time steps, while the FPAR data, originally available at 10-day intervals, was linearly interpolated to match the 16-day time steps. The final set of predictors included 8 sequential variables (listed under the variable column in Table 1), each aligned to 16-day intervals, resulting in 23 time steps per year. In addition, we designed 10 static predictors based on the original data sources summarized in Table 1. These static predictors comprised available water capacity, bulk density, two soil drainage classes represented as dummy variables (well-drained and excessively drained), end-of-season day, centroid latitude and longitude of each county and the yield values from the previous three years. This combination of sequential and static features was chosen to capture both temporal crop development patterns and location-specific agro-environmental characteristics relevant for yield prediction.

2.1.3. Seasonal climate forecast data

We studied the application of forecasted weather predictors from seasonal climate models to forecast wheat yield in-season. We used the fifth-generation seasonal forecast system (SEAS5) (Johnson et al., 2019) from the European Centre for Medium-Range Weather Forecasts (ECMWF). The data is readily accessible through the Copernicus Climate Change Service (dataset, , 2018). The SEAS5 forecasts are initialized on the first of each month and made available shortly after, on the fifth. To train and test the in-season performance of the SCF approach, we collected historical forecasts from 2004 to 2023. The features retrieved were daily mean, maximum and minimum temperature as well as daily accumulated precipitation. Forecasted weather conditions for 7 months ahead were collected at each initialization date. To account for uncertainties in initial conditions, seasonal climate models generate ensembles of forecasts at each initialization date. SEAS5 provided 25 ensemble members for the hindcast period from 2004 to 2016 and 50 members for the forecast period from 2017 to 2023. To ensure consistency across the full 2004–2023 period, we truncated the ensemble size of the forecast period to the first 25 members. While truncating from 50 to 25 members may alter the original statistical properties of the 50-member ensemble, 25 members are still considered sufficient to provide reliable estimates of both mean and forecast uncertainty. This choice follows the approach of the ECMWF to standardize their hindcast archive at 25 members, a size that balances computational cost while still providing assessments that are sufficiently robust to calibrate forecasts from the 50-member ensemble. The daily outputs of the 25 ensemble members for each of the four weather variables were averaged into 16-day time steps, as outlined in section 2.1.2. The raw SCF outputs were then bias-adjusted to align with observational data using normal quantile mapping for all four variables (Jakob, pp., 2011), in accordance with recommendations for SEAS5 usage within climate services (Crespi, 2021). The bias was learned using the training years to avoid information leakage, and then the bias was adjusted for all years in the training, validation and test split. The bias-adjusted SCF data were then assigned to the counties. The SCF data is available at a spatial resolution of 1°, and the assignment of the grid cells to the counties was done by matching the centroid of each county's polygon shape to the closest SCF grid cell, potentially assigning two counties to the same SCF grid. The SCF data consists of outlooks up to 7 months into the future and can offer valuable insights into variables such as temperature and precipitation anomalies in certain regions. However, the skill of seasonal forecasts depends on several factors, including initialization date, forecast variable, region, spatial and temporal resolution, and lead time. Based on a review of the literature (Ardilouze et al., 2019; Klemm and McPherson, Sep. 2018; Peings et al., 2025), we determined that incorporating forecasts beyond a 2-month lead time would not provide useful information for our study

area. Hence, for each of the 25 ensemble members, we only kept the first 4 time steps for each initialization date, which roughly corresponds to the first 2 months (4 periods of 16 days each).

2.2. Modeling workflow

2.2.1. Three-year average model

Our baseline approach was a 3-year average model, where forecasts were based on historical yield. For each county, the average yield from the three previous years was used as the forecast for the current year. The three-year period gives a realistic estimate of the current yield trend level. The mean yield of periods longer than three years would have likely been too low, given that yield levels consistently increase due to technology improvements (USDA/NASS, xxxx), while extreme years could have affected periods shorter than three years. The 3-year average model forecasted yield as soon as the new season began and did not adjust it throughout the season. A model that does not outperform the 3-year average model is considered unskillful.

2.2.2. Truncate approach

The interaction of predictors and their impact on crop yield can be captured by a long short-term memory network (LSTM) (Hochreiter and Schmidhuber, Nov. 1997). The primary advantage of LSTMs lies in their ability to understand sequential relationships in the feature space. In the context of crop yield modeling, this capability allows for the analysis of how current growth conditions affect yield while also taking into account the influence of previous conditions. By doing so, LSTMs are able to capture important temporal dependencies in yield development. For instance, the impact of a dry spell on crop yield depends, among other factors, on the amount of precipitation received prior to that dry spell. Numerous studies conducted across various crops and regions have demonstrated the potential of LSTMs in crop yield modeling (Cunha et al., 2018).

Our machine learning workflow for the truncate approach began by aligning the observed feature set described in section 2.1.2 with the historical yield data from section 2.1.1. For our study region, the end of the wheat season typically falls around the end of July (Fig. 3a). Therefore, predictor time steps were reorganized such that time step one was at the beginning of August of the previous year and time step 23 at the end of July of the harvest year. For each yield record, the predictors from August of the previous year until July of the harvest year were provided to estimate yield. The samples were then divided into the train (2004–2017), validation (2018–2020) and test set (2021–2023) with 5222, 793 and 774 samples, respectively. The distribution of states within each split was relatively constant (Fig. 3b). Wheat harvests of 2022 and 2023 have been especially challenging in the study region due to drought events (USDA/NASS), suggesting a realistic assessment for years with extreme conditions when crop yield forecasts are most needed. An alternative strategy that involves a (nested) leave-one-year-out evaluation has been used in previous studies (Meroni et al., 2021). This involves evaluating model performance in a year whose previous and subsequent data is already present in the training set. Our setup was selected because it closely mirrors an operational scenario, where a model that is trained on historical data is fine-tuned with data from recent years and then applied to make predictions for future years (Paudel et al., 2022).

The input to an LSTM was provided as a three-dimensional array ($N \times T \times C$) corresponding to the number of samples N, time steps T, and predictors C. Here, we provided an array containing both sequential and static predictors, with static predictors repeated for each time step (Leontjeva and Kuzovkin, 2016). The hyperparameters of the LSTM were determined using the validation set and fully observed predictors for a theoretical end-of-season estimation model. We found that the best results for the end-of-season model were achieved with information from time step 3 to time step 23, corresponding to the period from mid-September to the end of July. The model configuration was characterized by a hidden dimension of 64, a stack of 4 LSTM layers and a dropout rate of 0.5. For training, we found optimal performance using the Adam optimizer (Kingma et al., 2017) with a batch size of 64, a learning rate of 1×10^{-4} and an early stopping patience of 20 epochs. The hyperparameters found in the end-of-season experiments were fixed for the subsequent truncate and SCF approach experiments. To investigate in-season wheat yield forecast performance of the truncate approach, time steps starting from the season's end were successively removed by trimming the input array, and a new model was trained and evaluated on the test set. This procedure was repeated until the beginning of January, corresponding to a temporal truncation from time step 3 to time step 9.

2.2.3. SCF approach

Some studies have suggested training an LSTM model using weather observations and switching to weather forecasts during testing without retraining the model (Cunha et al., 2018). Alternatively, an LSTM can be trained with forecasted weather data (Zhao et al., 2024). We chose the latter approach and provided SCF data during training to allow the LSTM to emphasize SCF data based on its reliability.

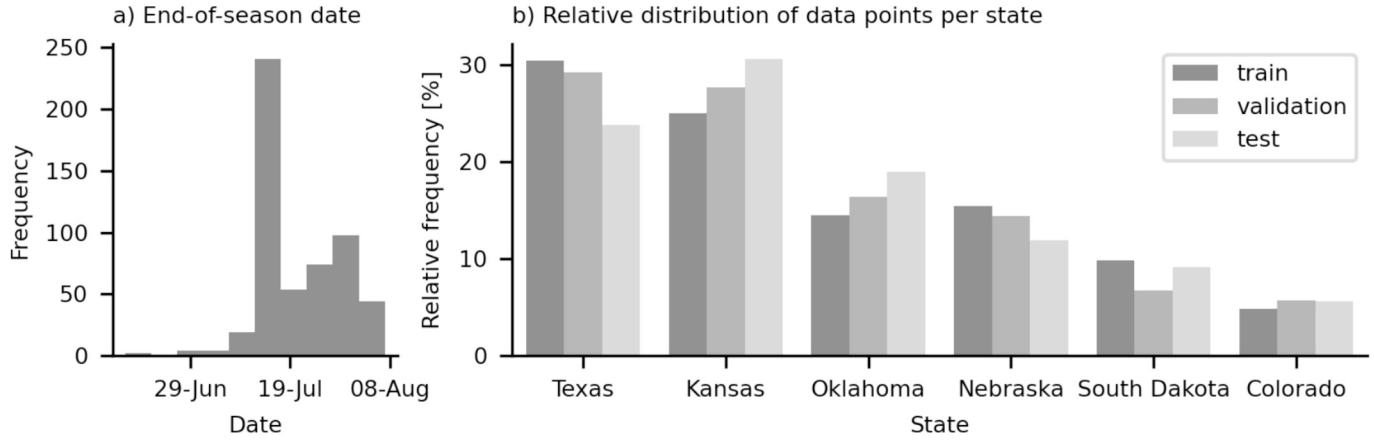


Fig. 3. Crop calendar and distribution of available data points. (a) The histogram of the end-of-season dates for the counties within the six states of our study region. (b) The number of data points per state is shown separately for the training, validation, and test splits, with values normalized within each split. That is, for each split, the bars across all states sum to 100%, allowing for direct comparison of the relative distribution across states, regardless of the overall size of the split. Although the training set covers a longer period (2004–2017) than the validation (2018–2020) and test (2021–2023) sets, the bar heights for states such as Texas are still comparable across splits. This indicates that there is no major distribution shift between them, and that the regional distribution of data remains consistent across the splits.

When newly forecasted daily weather data became available at the beginning of each month, we trained and tested new SCF-based wheat yield forecasts using an updated three-dimensional input array. The modification of the three-dimensional input array to the LSTM was as follows (Fig. 4). First, we determined the last time step for which observed features are used and the first time step for which SCF data is available by looking at the time step in which the SCF initialization date falls. All time steps before remained unchanged and used observed predictor data. Four future time steps were provided from SCF data, and all time steps beyond that were dropped from the array. For example, a new SCF on 1 April falls into time step 16, which spans from 30 March to 14 April. Until time step 15, the three-dimensional array contained observed features. For time steps 16–19, we replaced the observed tavg, tmax, tmin and prec with SCF data. In addition, we set all other sequential features for the time steps 16–19 to 0 (we assumed there were no observations for these predictors) and kept the static features unchanged (Fig. 4). All days within the first SCF time step up to the SCF initialization were ignored (30 March − 1 April) when calculating the time step average such that only SCF data from 2 April to 14 April was considered. This procedure was applied to all other months, from January to June.

The SCF approach consists of two sub-experiments. In the first sub-experiment, we calculated the mean of the ensemble from the 25 SCF predictors, which resulted in one new input array for the LSTM, along with one yield forecast. In the second sub-experiment, we created 25 modified input arrays, one for each ensemble member, and ran 25 yield forecasts. In the second experiment, we calculated the mean and standard deviation of the resulting wheat yield forecasts to analyze forecast accuracy and uncertainty.

2.3. Permutation-based predictor importance

Machine learning models have achieved remarkable performance across various tasks, including image recognition, natural language processing and predictive analysis. Despite their success, these models are often criticized for their lack of transparency. This black-box nature challenges the understanding of how models arrive at specific predictions raising concerns about trust. As a result, research has increasingly focused on developing explainability techniques to gain insight into a model's decision-making process. Permutation feature importance (PFI) is a simple yet widely used feature attribution technique that evaluates the sensitivity of the model to changes in its predictors. While there are many ways to conduct this kind of sensitivity analysis, PFI focusses on shuffling predictor values. This process disrupts their original relationship with a target variable. A predictor is deemed important by measuring the deterioration in model performance due to the perturbation. Table 2 describes how we applied PFI in our study.

2.4. Evaluation metrics

The performance of the models presented in section 2.2 was evaluated by comparing mean squared error (MSE), normalized root mean square error (NRMSE), and the coefficient of determination (R^2). In the equations below, y_i and $\hat{y}_i$ represent the observed and predicted yield, respectively. Further, $\bar{y}$ is the mean of the observed wheat yield an n is the number of data points.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \quad (1)$$

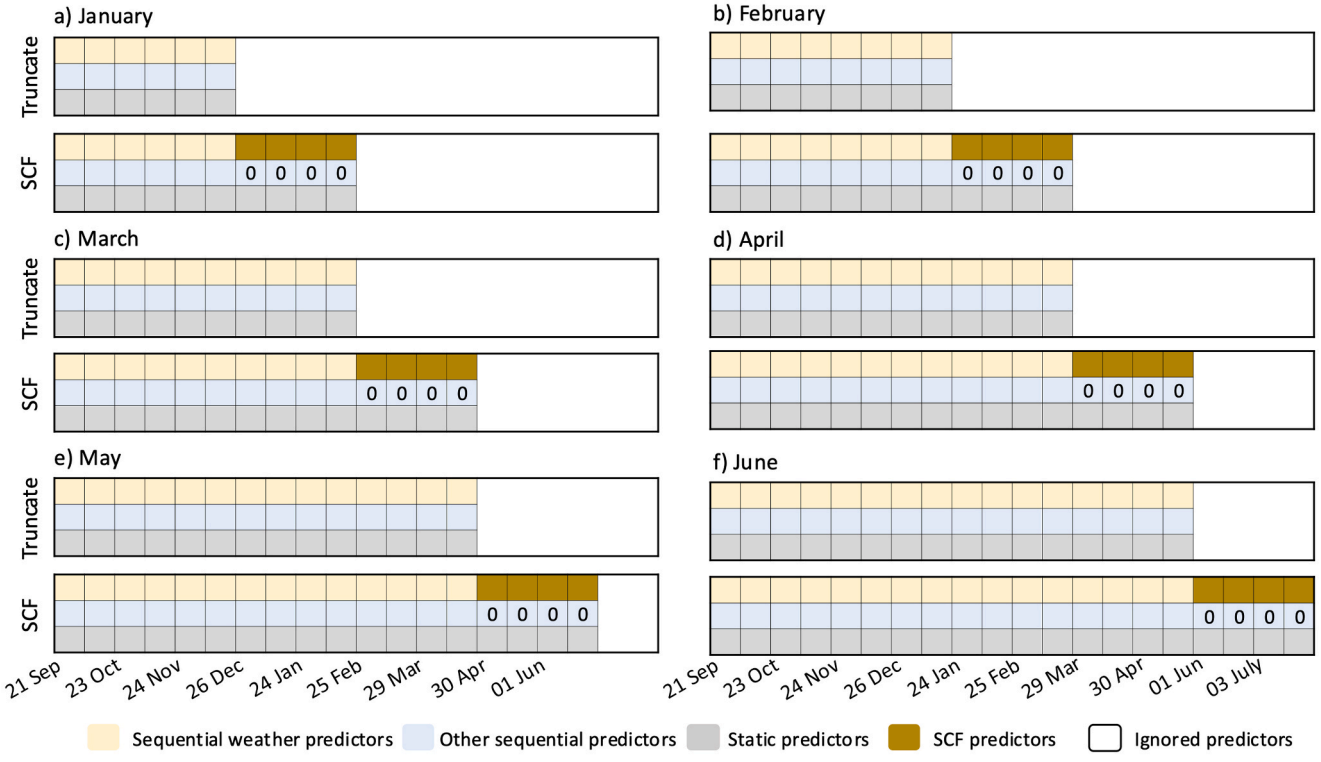


Fig. 4. Input array by approach and month of initialization. From January a) to June f), the difference in the input arrays between the truncate and seasonal climate forecast (SCF) approaches is shown. Each cell represents one 16-day time step. The months in the plot titles correspond to the month the SCF is initialized. For example, in a), the SCF is initialized on the 1st of January, which falls into the 16-day time step that starts on the 26th of December. Four time steps of weather data (copper color) are calculated using SCF data until the end of February. All other sequential data for these four time steps is set to 0 (blue color), while the four static time steps remain unmodified (grey color). Time steps beyond the 25th of February are ignored. The truncate approach uses time steps with information until the 26th of December. All future information is ignored. This procedure is applied equally to all other months of SCF initialization.

Table 2
Grouped permutation predictor importance (PFI).

Demonstrating a PFI procedure for a predictor group (a subset of predictors available at all time steps). For this analysis, our predictors are categorized into static features, sequential weather features and other sequential features. The following terms are pre-defined: trained model f, predictor vector X and target vector (ground truth) y

timesteps ← number of timesteps
idx ← indices corresponding to a group of predictors
importance_matrix (*imp*) ← *zeros* (timesteps) *// Initialize an empty array to store PFI values*
$\hat{y} \leftarrow f(X)$ *// Generate baseline score on initial predictors*
$score_{baseline} \leftarrow MSE(y, \hat{y})$
for timestep (t) = 1 to timesteps **do**
  $l \leftarrow []$
  for iteration in number of repeats **do**
    $X_{permuted} \leftarrow PFI(X_{t,idx})$ *// Shuffle values at time step and predictor indices only*
    $\hat{y}_{permuted} \leftarrow f(X_{permuted})$
    $score_{permuted} \leftarrow MSE(y, \hat{y}_{permuted})$
    l. insert($score_{permuted}$)
  end for
  $imp_t \leftarrow score_{baseline} - mean(l)$ *// Calculate change in MSE*
end for
return *imp* *// Return time-wise scores*

$$NRMSE = \frac{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}}{\bar{y}} \quad (2)$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \quad (3)$$

3. Results

With predictors available as observations for all time steps until the end of the season, the LSTM estimates absolute wheat yield across all states with a mean R^2 of 0.46, mean NRMSE of 0.24 and MSE of 0.46 t/ha (Fig. 5a-f). The performance was assessed on county-level yield data from 2021 to 2023, which covers 774 samples that were previously not seen during training or validation. When plotting estimated yield versus actual yield for each state (Fig. 5a-f), it can be seen that the model generally tends to underestimate the magnitude of high- and low-yield samples, except for Colorado (CO, Fig. 5b). This is indicated by the colored dashed regression lines being less steep than the black identity lines, representing the theoretical best fit. No spatial performance gradient exists when the NRSME is plotted per county (Fig. 5g). However, there is a difference in average NRMSE per state. South Dakota (SD, Fig. 5a) and Kansas (KS, Fig. 5c) have the best average performance across their counties, with 0.19 and 0.21 NRMSE, respectively. Texas (TX, Fig. 5e) has the highest NRMSE, with 0.3, while Nebraska (NE, Fig. 5f), Colorado (CO, Fig. 5b) and Oklahoma (OK, Fig. 5d) have good to

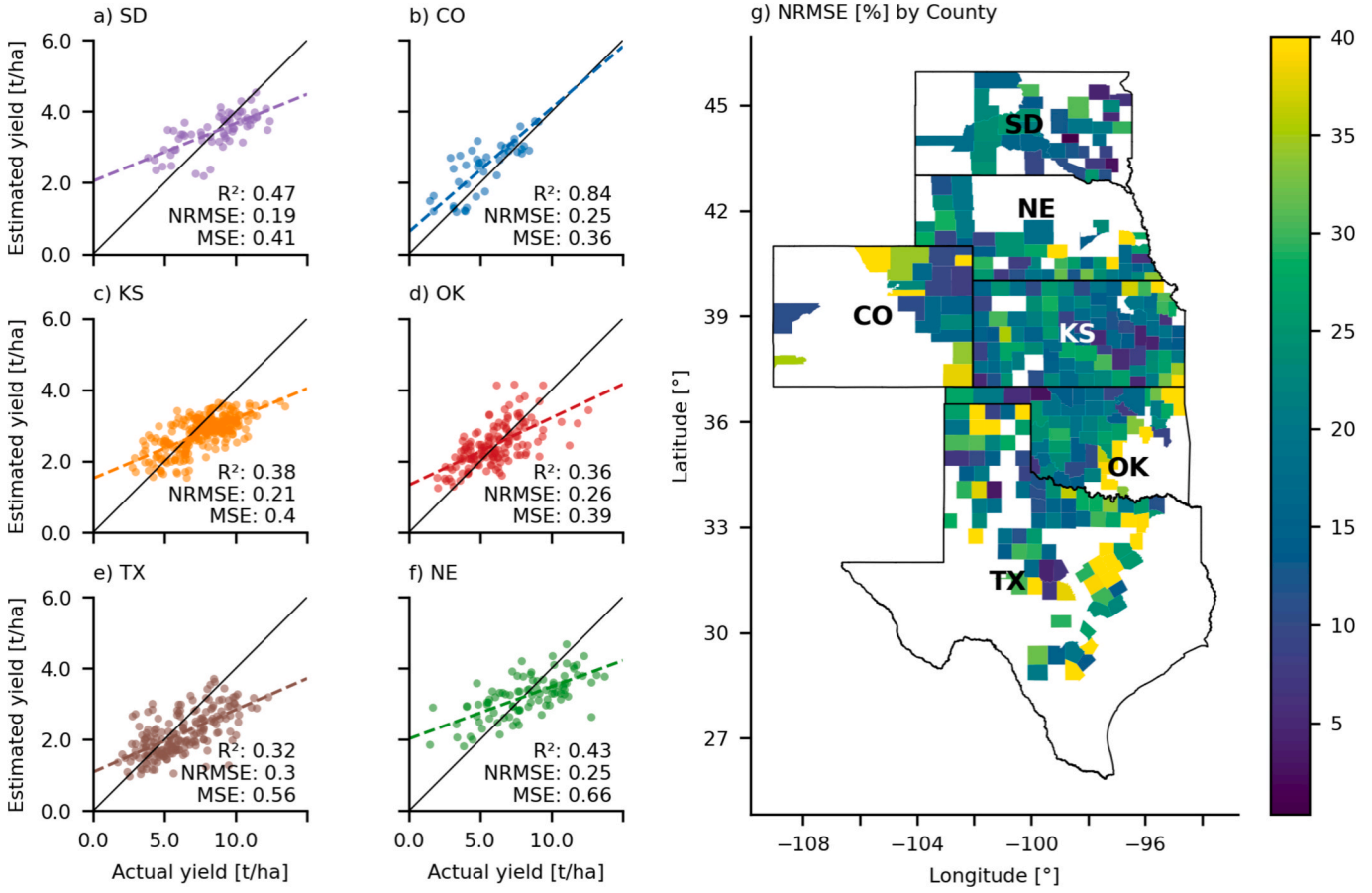


Fig. 5. Analysis of end-of-season model on test years. a-f) Actual county-level yield versus estimated yield on the test data from 2021 to 2023 per state. Yield estimates come from the LSTM model using fully observed data for all time steps until the end of the season (no truncation). Each plot shows the point cloud's regression line (colored dashed line) and the theoretical best fit's 1–1 identity line (black line). g) Spatial analysis of the normalized root mean square error (NRMSE) at the county level. The abbreviations of the states are given within each polygon. South Dakota (SD), Nebraska (NE), Kansas (KS), Colorado (CO), Oklahoma (OK), and Texas (TX).

medium performance with 0.25, 0.25, and 0.26 NRMSE, respectively. The end-of-season wheat yield estimation model was compared to the 3-year average model for each year from the test set (Table 3). In 2021, the 3-year average model and the LSTM have similar performance. In 2022 and 2023, the LSTM noticeably outperforms the 3-year average model, improving the NRMSE by 21–29 % and the R^2 from −0.02 to 0.39 in 2021 and from 0.17 to 0.27 in 2023.

The model of the end-of-season yield estimation was re-trained with all 5222 samples of the training data set to forecast wheat yield before harvest with the modified input arrays from the truncate or SCF approach (Fig. 6). Performance was again assessed on the unseen test data from 2021 to 2023, containing 774 samples. The in-season forecast approaches were compared to the 3-year average model. The 3-year average model did not update its forecasts as the season progressed (as per the definition of the 3-year average model), indicated through a constant performance with a grey horizontal line at 0.67 t/ha MSE. The truncate approach does not consistently outperform the 3-year average model early in the season and remains close to it until the beginning of May. As predictor information from May and June becomes available, the truncate approach improves by 28 % from 0.65 t/ha MSE with information until 16 May to 0.47 t/ha MSE with information until 03 July. Further improvements to the truncate approach with information from July are minimal. The two sub-experiments of the SCF approach (orange and green line, Fig. 6) show a similar result with a consistent decline in MSE from January until June, before the SCF approach becomes equivalent to the truncate approach in July when no SCF were used anymore. Early in the season, from January until the end of February, the SCF approach does not perform better than the 3-year average model or the truncate approach. From 29 March (SCF initialized on 01 April) to 30 April (SCF initialized on 01 May), the SCF approach outperforms both the 3-year average model and the truncate approach. The SCF approach (yield forecast mean, green line) in May achieves 0.6 t/ha MSE, 10 % better than the truncate approach. The SCF approach falls slightly behind the truncate approach in June. The sub-experiment where each SCF ensemble member ran its wheat yield forecast was used to plot the performance variance (green shaded area, Fig. 6). It can be seen that the variance changes with the forecast month. For example, 29 March (April forecast) has less variance than 01 June (June forecast). In addition, some SCF members outperform or remain similar to the truncate approach when the SCF ensemble mean (green line) fails to do so (e.g., 01 June).

Since both SCF approach experiments perform similarly, we analyzed predictor importance using the predictor mean SCF approach with one input array to the LSTM per instance. The importance of predictor groups of the SCF approach was estimated per forecast month and time step (Fig. 7) using permutation importance analysis. From January to March (Fig. 7a-c), the analysis indicated that static predictors are unimportant for the model. Furthermore, during these first three months, sequential weather and other sequential predictors, such as soil moisture and vegetation indices, are important across most time steps. Lastly, from January to March, the model relies on forecasted weather predictors from the SCF. In April (Fig. 7d), other sequential predictors from early time steps are most important, followed by static predictors, while forecasted sequential weather predictors are unimportant. In May (Fig. 7e), all feature groups are important. Static predictors are mostly relied on from early time steps. In June (Fig. 7f), no feature group or time step has high importance as their magnitude is distributed across the time steps.

The skill of the SCF for the four time steps of weather features was analyzed in Fig. 8. Skill was calculated as R^2 with the squared error of the SCF predictors over the test set as numerator and the squared error of using mean historical weather (2004–2017) as a forecast method over the test set as denominator. Temperature forecasts in January and February are only skillful in South Dakota (SD) and generally worsen from January to April. In March and April, no state has skillful temperature forecasts. In May and June, temperature forecasts improve, with almost all states showing skill with a R^2 greater than 0, except for Kansas (KS) and Nebraska (NE) in June. Precipitation forecast skill is given in the second row. Skill is poor in January, except for Oklahoma (OK). For the remainder of the season, no trend is observable, and there is a lot of variance regarding skill from one month to the other. For example, Texas (TX) has skillful precipitation forecasts in February (h), March (i) and May (k) and unskillful precipitation forecasts in January (a), April (j) and June (l). Most skillful states for both temperature and precipitation forecasts are found with SCFs made at the beginning of May (Fig. 8 e, k) except for precipitation forecasts in Colorado (CO) and Nebraska (NE), all quantities were skillfully forecasted.

Table 3
Annual performance compared to the benchmark model. The normalized root mean square error (NRMSE), mean squared error (MSE) and coefficient of determination (R^2) are shown for the end-of-season long short-term memory (LSTM) approach that uses observed predictors for all time steps until harvest and for the 3-year average model (3YA) that always estimates subsequent yield to be the mean yield from the previous three years. Performance is separated row-wise for each year in the test set from 2021 to 2023. The NRMSE is normalized by state using mean state-wise yield of that year and then averaged across all states to obtain the yearly NRMSE. The R^2 is calculated using the variance of each state per year in the denominator. Since R^2 can be infinitely negative but has an upper limit of 1, we take the median R^2 across all states to obtain the final yearly metric.

Year	NRMSE (LSTM)	NRMSE (3YA)	MSE (LSTM)	MSE (3YA)	R^2 (LSTM)	R^2 (3YA)
2021	0.23	0.22	0.45	0.44	0.07	0.01
2022	0.27	0.38	0.42	0.79	0.39	−0.02
2023	0.26	0.33	0.54	0.92	0.27	0.17

4. Discussion

We proposed an approach to forecast county-level wheat yield across the U.S. Great Plains before harvest using an LSTM with forecasted weather features from the seasonal climate model of the ECMWF. This is the first machine learning approach that integrates data from a SCF to make timely forecasts while also relying on observed static and dynamic features about weather, soil and vegetation conditions. Traditionally, machine learning models forecast crop yields before harvest by only relying on observed features until the forecast date (Paudel et al., 2022). At the end of the season, with all predictors being available as observation, our model achieved an R^2 of 0.46, an NRMSE of 0.24, and an MSE of 0.46 t/ha. At the beginning of May, three months before harvest, wheat yield forecasts with the SCF approach were possible with an MSE of 0.6 t/ha, improving the truncate approach, which only relies on observed data, by 10 %. Predictor permutation analysis showed that forecasted predictors from the SCF data were important for yield forecasts throughout the season. When the SCF approach most notably outperformed the truncate approach in May, the LSTM demonstrated the strongest dependence on SCF data. This dependence on SCF data in May was especially high for the first time step (16 days), suggesting that long-range climate forecast could be replaced by other weather forecast products that target medium or sub-seasonal time scales up to 4-6 weeks.

At the beginning of the season, from January to March, the SCF approach did not outperform the truncate approach or the 3-year average model. In April, the SCF approach performed marginally better, and in May, the SCF approach showed the largest improvement compared to the other approaches before falling slightly behind the truncate approach in June. The development of the performance of the SCF approach can be explained in several ways. First, the model cannot provide skillful forecasts if the available information on observed and forecasted time steps does not cover the most crucial crop development stages. Winter wheat transitions from booting to the heading stage in May and reaches maturity in June. Forecasts from January to March did not include information from May or June, explaining why no skillful

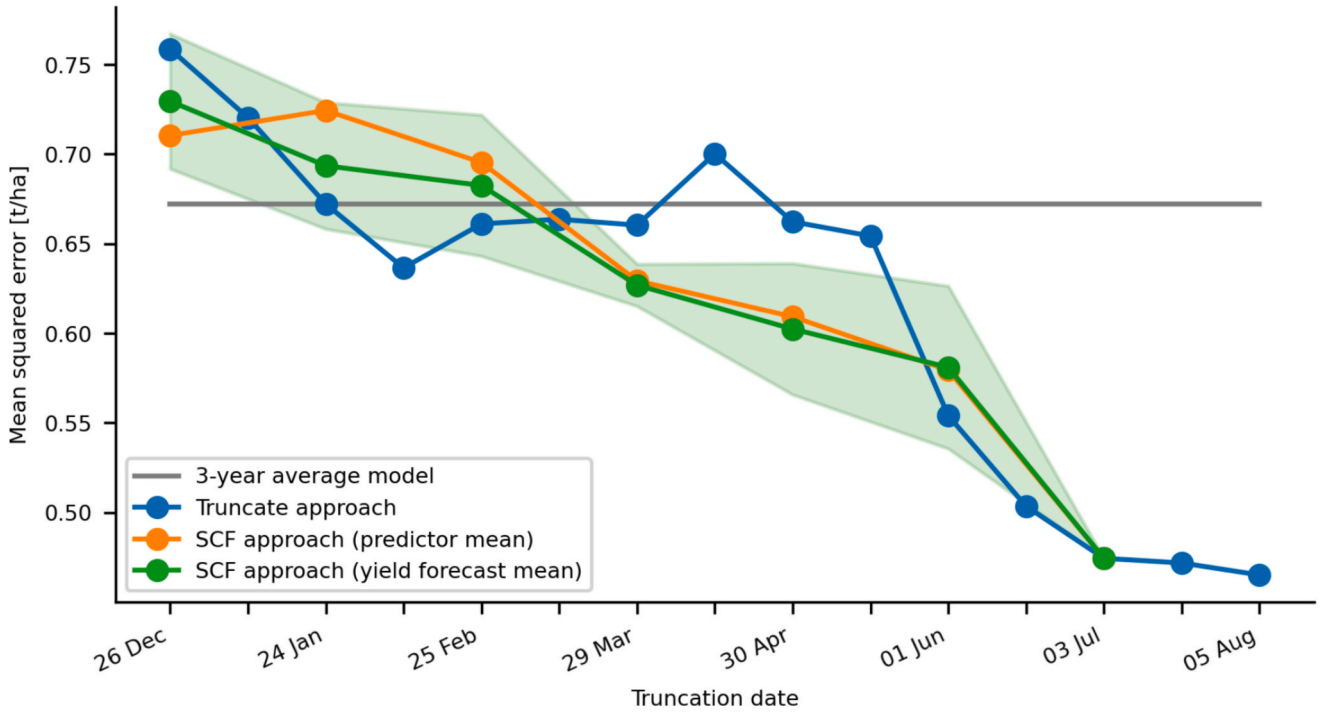


Fig. 6. In-season yield forecast performance. The wheat yield forecast performance for the 3-year average model (grey horizontal line), the truncate approach (blue line) and the two SCF approach experiments are shown. The truncate approach drops future features at a given truncation date (blue line). The 3-year average model always estimated the average yield of the last 3 years. The first SCF approach experiment consists of taking the mean of the predictors of its 25 ensemble members and running one wheat yield forecast per instance (predictor mean, orange line). The other experiment involved running a separate wheat yield forecast for each ensemble member. Then, the mean (yield forecast mean, green line) and the variance (one standard deviation) were plotted (green shaded area). Both experiments of the SCF approach use observed features for time steps before the truncation date (x-axis) and forecasted precipitation and temperature features for four time steps starting from the truncation date into the future. Performance is measured as the mean squared error in t/ha across all counties and test years from 2021 to 2023.

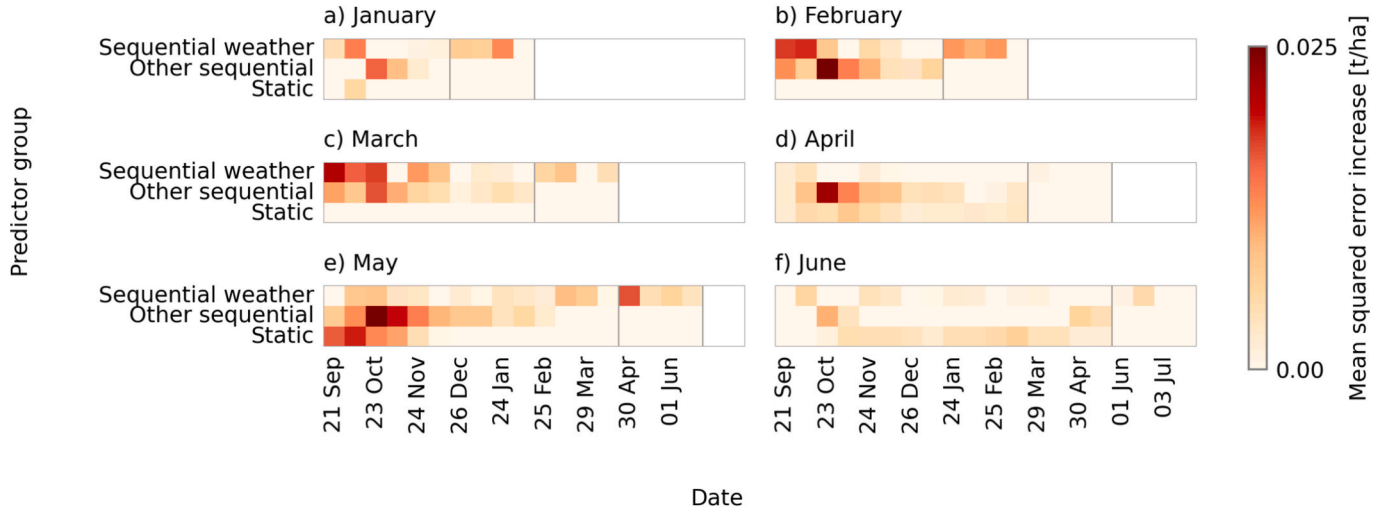


Fig. 7. Permutation predictor group importance for seasonal climate forecast approach. The importance of each predictor group (y-axis) used in the seasonal climate forecast (SCF) approach is shown for each time step (x-axis) and forecast month (a-f), where larger mean squared error (MSE) increases (color bar) represent higher importance. Predictors are grouped into sequential weather predictors (mean, maximum and minimum temperature, precipitation), other sequential predictors (fraction of absorbed photosynthetically active radiation, solar radiation, root zone soil moisture, climatic water balance) and static predictors. The importance of predictor groups is calculated using a permutation importance analysis, where values are permuted across samples. If the MSE increases due to the permutation, the predictor group at the given time step is important because changing its values affects the model output. In contrast, changing irrelevant predictors will not lead to a change in the MSE of the wheat yield forecast. The SCF date is indicated with a vertical grey line for each month from January (a) to June (f). Sequential weather time steps beyond that date are coming from SCFs, while other sequential feature time steps are set to zero and static features are kept equal. Another vertical grey line is marked four time steps further to the right to indicate the time when future information is ignored.

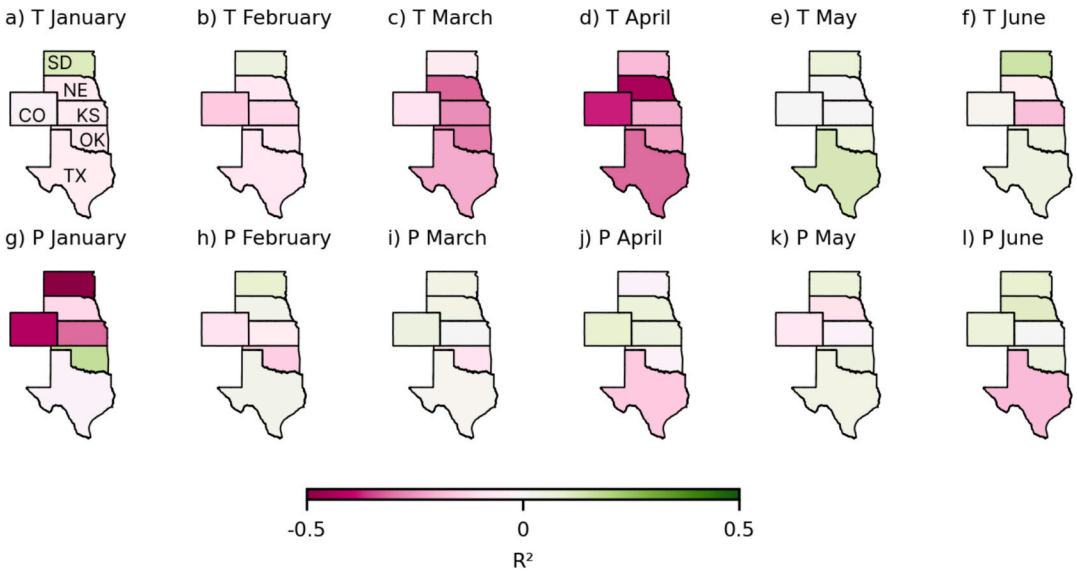


Fig. 8. Spatial analysis of seasonal climate forecast skill. Skill is measured as R^2, where higher values are better and negative values indicate a lack of skill. The top row (a-f) corresponds to the average skill of forecasting temperature features (mean, maximum and minimum). The lower row (g-l) shows precipitation forecast skill. Each column represents a forecast initialization month from January to June. Each month, forecasts for four time steps (64 days) ahead are evaluated. The abbreviations of the states are given in the first plot within each polygon: South Dakota (SD), Nebraska (NE), Kansas (KS), Colorado (CO), Oklahoma (OK), and Texas (TX).

wheat yield forecasts were possible. During May and June, rainfall is more relevant than during the beginning of the year and yield formation processes are most vulnerable to extreme weather conditions (Zhao et al., 2022) that can lead to irreversible damages. In contrast, suboptimal weather conditions in months like January or February do not have the same impact, as plants can still recover during the remainder of the season. Second, the skill of the SCF for sequential weather predictors influenced how the SCF approach compared to the truncate approach. We calculated SCF skill in Fig. 8, by dividing the squared errors of the SCF with the squared errors of using mean historical weather (2004–2017) as a forecast method. This approach was chosen because the truncate approach of ignoring future weather conditions implicitly assumes that average historical conditions will apply. In Fig. 8, we show a skill score that directly relates the SCF to the truncate approach. The resulting skill score was highest in May (Fig. 8). In May, the permutation predictor importance analysis also revealed that SCF predictors are important (Fig. 7), and the SCF approach most notably outperformed the truncate approach (Fig. 6). In some months (e.g., April), the predictor importance analysis (Fig. 7) suggests that other sequential variables, such as FPAR or soil moisture, are more important than sequential weather predictors. This does not mean that weather conditions are not relevant at this stage but that the other sequential predictors already capture much of the weather signal and thus act as indirect proxies.

Previous studies using statistical and machine learning methods for end-of-season wheat yield forecasting in the U.S. have reported a range of results. While direct comparisons to our approach are challenging due to differences in geographic scope, predictor variables, and evaluation approaches, these studies provide useful context. For instance, one machine learning model forecasting winter wheat yields across all U.S. states reported an R^2 of 0.75 for the 2020 and 2021 test years (Joshi et al., 2023). Another statistical model focused on Kansas achieved a normalized RMSE of 7 % using cross-validation for the 2000–2008 period (Becker-Reshef et al., Jun. 2010). We also explored analog-year benchmarks as an alternative to the 3-year average or truncate approaches. However, their performance was limited by the small number of available historical years (8–17 per county) combined with the high dimensionality of the predictor space (Suppl. Fig. S1).

While the results are promising and encourage an application to other crops and regions, there is potential to further improve our approach. The predictors used here were obtained from the CY-Bench project, in which a static crop land mask from 2021 (Tricht, 2023) was utilized to aggregate grid cell data to the county-level. Dynamic in-season crop land masks, obtained from high-resolution satellite data, such as from Copernicus Sentinel-2, could help to make more targeted predictors and better yield forecasts (Ben, et al., 2025). The weather across the U.S. Great Plains is difficult to forecast (Peings et al., 2025) and SCF-based machine learning yield forecast could be more useful in regions with higher SCF skills. Furthermore, our approach suffered from the spring predictability barrier (Duan and Wei, Apr. 2013), where SCF skill beyond spring (e.g., April) is poor but suddenly recovers for forecasts made after spring (e.g., May). Our approach could be further improved for crops with crucial development stages that are better aligned with SCF skills. In addition, it has been shown that averaging the output of SCFs from different climate centers, called the Multi-Model Ensembles technique, can provide more reliable forecasted weather predictors than using the output from a single SCF, like we and others have demonstrated elsewhere (Knutti et al., 2010). Future studies on machine learning crop yield forecasts could leverage the strength of multi-model ensembles of SCFs to have forecasted weather predictors available that are more skillful than those coming from a single SCF. The spatial resolution of the SCF of the ECMWF used here is 1° by 1°, which led to some counties sharing the same SCF grid cell and forecasted weather conditions. Forecasted weather predictors in higher spatial resolution can be validated for better in-season wheat yield forecasts on the county level. Furthermore, SCFs are generally initialized once a month, while crop yield forecasts before harvest could benefit from real-time forecasts. A possibility would be the application of extended-range forecasts, which are initialized daily, but with shorter forecast horizons. Here, we applied the standard bias adjustment method quantile mapping and let the LSTM find meaningful information within the SCF data by itself. Alternatively, one could analyze the skill of SCF data during data preprocessing and only provide features that are considered skillful at providing a shortcut in the learning process of the LSTM. Our study lacks detailed information on the spatial distribution of wheat growth stages and on how forecasted meteorological factors in May relate to actual yield outcomes on the county-level. Addressing this gap in the future by linking forecast data, weather conditions and yield impacts might improve forecasts. A general statement about the usability of SCF

data to machine learning-based crop yield models is still challenging to make. In this study, we only evaluated our approach using an LSTM, while other model types may hold other inherent challenges. In addition, the approach's performance in other breadbasket regions has to be explored.

5. Conclusion

We have enhanced in-season wheat yield forecasts by integrating seasonal climate forecasts (SCFs) into an LSTM-based prediction system. We demonstrated for wheat across the U.S. Great Plains that forecasted weather variables, can improve yield forecasts, especially during May, leading up to a 10 % reduction in MSE compared to relying solely on observed predictors. This mid-season skill is particularly beneficial given that wheat transitions through critical developmental stages in late spring. By providing more reliable forecasts two to three months before harvest, our approach could help farmers, agribusinesses, and policymakers make timely and informed decisions regarding resource allocation, risk mitigation, and market planning.

However, SCF skills remain low from January to March. Further gains in improving forecast skills might also come from using multi-model ensembles instead of single-models or incorporating higher-resolution seasonal products. Future research should evaluate this framework across other major breadbasket regions and for different crops to verify its adaptability and robustness. As weather forecasts are expected to become increasingly accurate, our findings suggest the potential benefits of more integrated climate–crop modeling to strengthen food security and foster sustainable agricultural practices.

CRediT authorship contribution statement

Maximilian Zachow: Writing – original draft, Visualization, Validation, Software, Resources, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Stella Ofori-Ampofo:** Writing – original draft, Visualization, Validation, Software, Resources, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Harald Kunstmann:** Writing – review & editing, Validation, Supervision, Methodology, Investigation. **Rıdvan Salih Kuzu:** Writing – review & editing, Validation, Supervision, Methodology, Investigation. **Xiao Xiang Zhu:** Writing – review & editing, Validation, Supervision, Methodology, Investigation. **Senthold Asseng:** Writing – original draft, Validation, Supervision, Methodology, Investigation.

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.

Acknowledgments

The work of S. Ofori-Ampofo was funded by the Munich Aerospace e. V. scholarship. The results of our study contain modified Copernicus Climate Change Service information 2024. Neither the European Commission nor ECMWF is responsible for any use that may be made of the Copernicus information or data it contains.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.compag.2025.110965.

Data availability

Data used is publicly available and indicated in the manuscript. The repository with code to reproduce our study is indicated in the manuscript.

References

Ardilouze, C., Batté, L., Decharme, B., Déqué, M., 2019. "On the link between summer dry bias over the U.S. Great Plains and seasonal temperature prediction skill in a dynamical forecast system. DOI: 10.1175/WAF-D-19-0023.1.

Basso, B., Liu, L., 2019. Chapter Four - Seasonal crop yield forecast: Methods, applications, and accuracies. In: Advances in Agronomy, D. L. Sparks, Ed., vol. 154, Academic Press, pp. 201–255.

Becker-Reshef, I., Vermote, E., Lindeman, M., Justice, C., Jun. 2010. A generalized regression-based model for forecasting winter wheat yields in Kansas and Ukraine using MODIS data. Remote Sens. Environ. 114 (6), 1312–1323. https://doi.org/10.1016/j.rse.2010.01.010. ISSN: 0034-4257.

Ben, A.W. et al., 2025. "JRC MARS Bulletin - Global outlook - Crop monitoring European neighbourhood - Ukraine - June 2025," JRC Publications Repository. Accessed: Jul. 24, 2025.

A. Crespi, M. Petitta, P. Marson, C. Viel, and L. Grigis, "Verification and Bias Adjustment of ECMWF SEAS5 Seasonal Forecasts over Europe for Climate Service Applications," Climate, vol. 9, no. 12, Art. no. 12, Dec. 2021, DOI: 10.3390/cli9120181.

R. L. F. Cunha, B. Silva, and M. A. S. Netto, "A Scalable Machine Learning System for Pre-Season Agriculture Yield Forecast," en, in 2018 IEEE 14th International Conference on e-Science (e-Science), Amsterdam: IEEE, Oct. 2018, pp. 423–430, ISBN: 978-1-5386-9156-4. DOI: 10.1109/eScience.2018.00131.

[dataset] Copernicus Climate Change Service, Seasonal forecast daily and subdaily data on single levels, dataset, 2018. DOI: 10.24381/ CDS.181D637E.

Deines, J.M., Archontoulis, S.V., Huber, I., Lobell, D.B., 2024. Observational evidence for groundwater influence on crop yields in the United States. Proc. Natl. Acad. Sci. 121 (36), e2400085121. https://doi.org/10.1073/pnas.2400085121.

Duan, W., Wei, C., 2013. The 'spring predictability barrier' for ENSO predictions and its possible mechanism: results from a fully coupled model. Int. J. Climatol. 33, 1280–1292. https://doi.org/10.1002/joc.3513.

S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," Neural Comput., vol. 9, no. 8, pp. 1735–1780, Nov. 1997. DOI: 10 . 1162 / neco . 1997 . 9 . 8 . 1735. ISSN: 0899-7667.

M. Jakob Themeßl, A. Gobiet, and A. Leuprecht, "Empirical-statistical downscaling and error correction of daily precipitation from regional climate models," en, International Journal of Climatology, vol. 31, no. 10, pp. 1530–1544, 2011,_eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/j. DOI: 10.1002/joc.2168. ISSN: 1097-0088.

S. J. Johnson, T. N. Stockdale, L. Ferranti, et al., "SEAS5: The new ECMWF seasonal forecast system," English, Geoscientific Model Development, vol. 12, no. 3, pp. 1087–1117, Mar. 2019, Publisher: Copernicus GmbH. DOI: 10.5194/gmd-12-1087-2019. ISSN: 1991-959X.

A. Joshi, B. Pradhan, S. Chakraborty, and M. D. Behera, "Winter wheat yield prediction in the conterminous United States using solar-induced chlorophyll fluorescence data and XGBoost and random forest algorithm," Ecological Informatics, vol. 77, p. 102 194, Nov. 2023. DOI: 10.1016/j.ecoinf.2023.102194. ISSN: 1574-9541.

D. P. Kingma and J. Ba, Adam: A Method for Stochastic Optimization, arXiv:1412.6980 [cs], Jan. 2017. DOI: 10.48550/arXiv.1412.6980.

Klemm, T., McPherson, R., Sep. 2018. Assessing Decision timing and Seasonal climate Forecast needs of Winter Wheat Producers in the South-Central United States. J. Appl. Meteorol. Climatol. 57, 2129–2140. https://doi.org/10.1175/JAMC-D-17-0246.1.

R. Knutti, R. Furrer, C. Tebaldi, J. Cermak, and G. A. Meehl, "Challenges in Combining Projections from Multiple Climate Models," EN, Journal of Climate, vol. 23, no. 10, pp. 2739–2758, May 2010, Publisher: American Meteorological Society Section: Journal of Climate, 1520-0442. DOI: 10.1175/2009JCLI3361.1. ISSN: 0894-8755.

A. Kolotii et al., "Comparison of biophysical and satellite predictors for wheat yield forecasting in Ukraine," The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, vol. XL-7-W3, pp. 39–44, Apr. 2015, DOI: 10.5194/isprsarchives-XL-7-W3-39-2015.

A. Leontjeva and I. Kuzovkin, "Combining Static and Dynamic Features for Multivariate Sequence Classification," in 2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA), arXiv:1712.08160 [cs], Oct. 2016, pp. 21–30. DOI: 10.1109/ DSAA.2016.10.

M. Meroni, F. Waldner, L. Seguini, H. Kerdiles, and F. Rembold, "Yield forecasting with machine learning and small data: What gains for grains?" Agricultural and Forest Meteorology, vol. 308-309, p. 108 555, Oct. 2021. DOI: 10.1016/j.agrformet.2021.108555. ISSN: 0168-1923.

D. Paudel, H. Boogaard, A. de Wit, et al., "Machine learning for regional crop yield forecasting in Europe," Field Crops Research, vol. 276, p. 108 377, Feb. 2022. DOI: 10.1016/j.fcr.2021.108377. ISSN: 0378-4290.

D. Paudel, H. Baja, R. van Bree, et al., CY-Bench: A comprehensive benchmark dataset for subnational crop yield forecasting, en, dataset, Sep. 2024. DOI: 10.5281/ZENODO.13839134.

Peings, Y., Dong, C., Magnusdottir, G., 2025. Seasonality in the predictive skill of southwest united states precipitation in the seasonal forecast systems. Geophysical Research Letters, vol. 52, no. 12, p. e2025GL115972, 2025, DOI: 10.1029/2025GL115972.

Rippey, B.R., Dec. 2015. The U.S. drought of 2012. Weather and Climate Extremes, USDA Research and Programs on Extreme Events 10, 57–64. https://doi.org/10.1016/j.wace.2015.10.004. ISSN: 2212-0947.

B. Schauberger, J. Jägermeyr, and C. Gornott, "A systematic review of local to regional yield forecasting approaches and frequently used data resources," en, European

Journal of Agronomy, vol. 120, p. 126 153, Oct. 2020. DOI: 10.1016/j.eja.2020.126153. ISSN: 1161-0301.

Tricht, K. V. et al., 2023. "WorldCereal: a dynamic open-source system for global-scale, seasonal, and reproducible crop and irrigation mapping," May 2023, DOI: 10.5194/essd-2023-184.

USDA/NASS QuickStats Ad-hoc Query Tool, dataset.

H. Zhao, L. Zhang, M. B. Kirkham, et al., "U.S. winter wheat yield loss attributed to compound hot-dry-windy events," en, Nature Communications, vol. 13, no. 1, p. 7233, Nov. 2022, Publisher: Nature Publishing Group. DOI: 10 . 1038 / s41467 - 022-34947-6. ISSN: 2041-1723.

Zhao, J., Guo, Y., Lin, Y., Zhao, Z., Guo, Z., 2024. "A novel dynamic ensemble of numerical weather prediction for multi-step wind speed forecasting with deep reinforcement learning and error sequence modelling. Energy, vol. 302, p. 131 787, Sep. 2024. DOI: 10 . 1016 / j . energy. 131787. ISSN: 0360-5442.