Computers and Electronics in Agriculture 209 (2023) 107855

Contents lists available at ScienceDirect

Computers and Electronics in Agriculture

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

Complex systems modelling of UK winter wheat yield

R.J. Hall [a], H.-L. Wei [b], S. Pearson [c,*], Y. Ma [d], S. Fang [d], E. Hanna [a]

[a] Department of Geography and Lincoln Climate Research Group, University of Lincoln, Lincoln LN6 7FL, UK
[b] Department of Automatic Control and Systems Engineering, University of Sheffield, Sheffield S1 3JD, UK
[c] Lincoln Institute of Agri-Food Technology, University of Lincoln, Riseholme Campus, Lincoln LN2 2NG, UK
[d] Institute of Ecological and Agricultural Meteorology, Chinese Academy of Meteorological Science, No. 46, Zhongguancun South Road, Beijing 100081, China

ARTICLE INFO

Keywords:
Wheat yield prediction
Machine learning
Systems identification
WOFOST
Climate

ABSTRACT

Wheat is one of the most important global crops, understanding the drivers of wheat yield has significant societal benefits. Climate variables are particularly important in determining interannual variations in wheat yield, either as primary factors which directly influence the stages of wheat growth, or as secondary factors through their influence on pests, diseases and soil conditions. Here we present a new approach to model wheat yield; an empirical method based on nonlinear complex systems identification, known as NARMAX (Nonlinear AutoRegressive Moving Average with eXogenous inputs model). We deploy the NARMAX analytical approach for a specific site, Rothamsted, UK, where detailed meteorological variables are available, together with specific information on site conditions and crop growth stages. NARMAX yield forecasts are compared with those from the WOFOST crop model and nine state-of-the-art machine learning (ML) models; experimental results show that NARMAX outperforms all the compared methods in both prediction accuracy and model interpretability. We also develop regional wheat yield forecasts derived from a new gridded meteorological data product.

The NARMAX approach produces skillful forecasts ($r = 0.78$) of Rothamsted wheat yield for a validation period, with small errors. The NARMAX regional forecasts, based on less specific information than WOFOST, also show a high degree of skill ($r = 0.73$). In addition, the predictor terms chosen for the model are identifiable and can help to give insight into potential key processes involved in the determination of wheat yield at a specific location. This approach can be extended in principle to other crop types and locations. It is straightforward and inexpensive to implement, using a limited number of meteorological predictor variables, which can be taken from site-based observations, or from gridded meteorological datasets. The method is a new tool to understand the environmental drivers of wheat yields on an annual basis.

1. Introduction

Wheat (*Triticum aestivum* L.) is one of the most important global crops. It forms the basis of many foodstuffs, providing c.20 % of all calories consumed by people and critical for animal feed (Shiferaw et al., 2013). In the UK, there has been a sustained increase in productivity over the period 1945 to 1995, but evidence indicates that this has since flattened at c. 8 t ha^{-1} (Knight et al., 2012). Globally, wheat yields are only increasing by c. 0.9% per annum whilst c. 2.4 % growth is required to meet the food requirements of an expanding population (Ray et al., 2013). Much of this increase in yield is due to technological developments in the breeding of new strains of wheat (e.g. Foulkes et al., 2007). However, plant growth and development are affected by a number of environmental factors which can cause significant fluctuations in productivity from year to year. It has been estimated that climate variability can account for more than 60 % of interannual variability in crop yields (Ray et al., 2013), while the prevalence of pests and diseases can also be influenced by weather conditions at key points in the crop life cycle. In the light of this it is particularly important to develop a thorough understanding of the impacts of climate on wheat yields, enabling society to better respond to shocks in food production. Accurate quantification of wheat yield is important for policymaking and farm management.

There are two approaches to forecasting wheat yield: (1) the use of crop growth simulation models (CGSM), such as WOFOST (Boogaard and Kroes, 1998), APSIM (Keating et al., 2003), or now multi-model ensembles (e.g. Asseng et al., 2015), which currently encompass up to 32 different CGSM platforms (Asseng et al., 2019), and (2) empirical

* Corresponding author.
E-mail address: spearson@lincoln.ac.uk (S. Pearson).

https://doi.org/10.1016/j.compag.2023.107855
Received 9 November 2022; Received in revised form 25 March 2023; Accepted 12 April 2023
Available online 27 April 2023

approaches (e.g. Landau et al. 2000, Lobell and Field, 2007). The CGSM are mechanistic, in that they try to explain both the relationship between parameters and the simulated variables, and also the biological mechanisms involved (Challinor et al., 2009). However, given that best forecasting skill has been achieved with multi-ensemble models (Asseng et al., 2019), it is clear that no one model is accepted as a definitive description of the effects of environment on wheat growth and yield. This suggests considerable opportunity exists to improve forecast skill within solus CGSMs. Statistical models on the other hand have offered less insight into processes. Due to possible overfitting, there are considerable risks when using model for prediction beyond the observed samples (Rosenzweig et al., 2013). To date they have shown limited forecasting skill, Chmielewski and Potts (1995)- using longterm yield data (1854 to 1967) at Rothamsted, UK, showed highest wheat yields occurred in dry or cool years with a multi-regression model skill of $r = 0.33$ with temperature and precipitation as critical environmental variables. Likewise, Landau et al. (2000) used a parsimonious approach resulting in a 17-parameter model that quantified the effects of environment on UK wheat yields, showing modest skill ($r = 0.41$) across a large observation data set of 2000 yields.

Recent advances in machine learning have led to the development of new techniques for yield forecasting. Pantazi et al. (2019) use machine learning to quantify wheat yield by combining advanced sensing of soil properties with neural networks and self-organising maps; however, such a forecast is site-specific and data intensive. Similarly, Cai et al. (2019) used multiple machine learning techniques combined with climate and satellite data to forecast Australian wheat yield ($r = 0.87$). Machine learning techniques offer an array of black-box tools that demonstrate high modelling skill but can provide little insight into underpinning mechanisms.

In this study we move the state of the art forward by presenting novel statistical models based on a Nonlinear AutoRegressive Moving Average with eXogenous inputs (NARMAX) approach used for system identification (Billings, 2013). The NARMAX approach is a subcategory of machine learning, but it is different from other machine learning methods in that models resulting from this approach are transparent: that is the predictor variables are apparent, the models are easily interpretable and the relationships between the response and explanatory variables are explicitly known. In addition, NARMAX can be used with relatively small datasets such as the time series available here. We use NARMAX to predict wheat yields for a specific location, given detailed monthly meteorological data only. NARMAX has been successful in revealing linear and non-linear relationships across a range of scales, in engineering, biological and environmental sciences (Billings, 2013); however, it has not previously been applied to crop yield prediction. Here we compare the NARMAX results with those from CGSMs, in this case WOFOST, as well as other typical machine learning approaches. WOFOST requires fewer input data and parameters than APSIM, and WOFOST has more simulation tests in Europe. WOFOST (WOrld FOod STudies) is a mechanistic crop growth model for the quantitative analysis of the growth and production of annual field crops (Van Diepen et al., 1989; Boogaard and Kroes, 1998; Supit et al, 1994; de Wit et al., 2019). It uses a general model to describe the process of crop growth and development. By changing crop parameters, it can be applied to different crop species or varieties. WOFOST can be applied in two modes: potential, where crop growth is caused only by solar radiation and temperature and there are no growth limiting factors, and a water limited mode which uses water availability to limit crop growth.

A typical limitation of CGSMs is that they are developed for field-level data. In reality, this situation hardly ever occurs as site-specific information is rarely available, and CGSMs are increasingly used for regional forecasts (e.g. Boogaard et al., 2013; van der Velde et al., 2019). It is therefore important to be able to upscale, making regional generalisations based on CGSMs (e.g. Asseng et al. 2015; 2019). However, there are issues with this approach for CGSMs, as biased simulations can result since biophysical processes depend on scale (Hansen and Jones, 2000) and the models built to represent processes at a given scale may not produce valid results when scaled up. Furthermore, there may be a nonlinear aggregation error arising from changing resolution (Hoffmann et al, 2016). In addition to spatial aggregation, further considerations are crop model parameters (Iizumi et al., 2014) and climate data, for example temperature biases due to mean elevation difference between a coarse grid cell and specific site (Baron et al, 2005). However, a NARMAX approach based on meteorological parameters suffers from no such issues, so we also present a regional forecast model for selected England wheat-growing regions, using the NARMAX methodology.

To fairly evaluate the strengths of the proposed approach, we also compare the NARMAX results with those from nine state-of-the-art machine learning (ML) models, namely, least absolute shrinkage and selection operator (LASSO) (Hastie, Tibshirani and Friedman, 2008), decision tree (DT) (Breiman et al., 1984; Loh, 2002), random forest (RF) (Breiman, 2001; Hastie, Tibshirani and Friedman, 2008), generalized additive model (GAM) (Hastie, Tibshirani and Friedman, 2008; Lou et al., 2013), Gaussian process regression (GPR) (Rasmussen and Williams, 2006), support vector machine (SVM) (Fan, Chen and Lin, 2005), feedforward neural network (FFNN, shallow neural network) (Reed and Marks, 1999), Long-Short-Term Memory (LSTM, deep neural network) (Hochreiter and Schmidhuber, 1997; Graves and Schmidhuber, 2005; Bengio, 2013), and bidirectional LSTM (BiLSTM) (Schuster and Paliwal, 1997; Goodfellow, Bengio, and Courville, 2016; Siami-Namini et al., 2019). Note that whilst NARMAX models are transparent and parsimonious, most of these nine ML models (e.g., the last six) are opaque and complex, and cannot be written down and easily explained. Experimental results show that NARMAX outperforms all these compared methods in both prediction accuracy and model interpretability.

The main contributions and advantages of the work are as follows:

- It presents a new approach to wheatyield modelling based on an empirical method based on nonlinear complex system identification, known as NARMAX.
- It demonstrates the NARMAX analytical approach for a specific site, Rothamsted, UK, where detailed meteorological variables are available, together with specific information on site conditions and crop growth stages.
- It also demonstrates how regional wheatyield forecasts might be derived from a new gridded meteorological data product.
- Unlike most other machine learning methods which work in a black-box manner, the proposed model is completely explainable due to its transparent, interpretable, reproducible and parsimonious (TRIP) properties. It explicitly tells us which meteorological factors significantly affect wheat yield, and reveals the relationship between wheat yield and meteorological factors.
- The performance of the identified NARMAX models is compared with those of nine state-of-the-art machine learning models; it turns out that NARMAX outperforms all the nine compared models.

2. Data

2.1. Rothamsted

Rothamsted is a long-established agricultural research station in Hertfordshire, north of London. Our data are taken from the Broadbalk winter wheat experiment, started in 1843, one of the oldest agronomic experiments in the world (Perryman et al., 2018; MacDonald et al., 2018). Detailed management and agronomic data are available, together with daily data from an on-site meteorological station (51.8066 N, −0.3602E, 128 m asl). We use meteorological and wheat yield data from Rothamsted Research, Hertfordshire, UK, for the period 1968–2018, as in 1968 short-stalked wheat varieties were introduced; resulting in a large increase in yield. We use winter first wheat yield, which is typically sown in October and rotated round a limited number of plots within a field. Winter wheat is always grown on Broadbalk, with

the exception of the 2015 yield, when weather conditions prevented the sowing of winter wheat and a spring wheat variety was substituted. For the CGSMs, date of sowing and harvesting is available for each harvest, together with cultivar information. Soil data are taken from Salter and Williams (1969) and the Rothamsted electronic archive (http://www.era.rothamsted.ac.uk/Broadbalk).

For the NARMAX and ML models we have selected 13 meteorological variables that may be important for influencing wheat yield, listed in Table 1. These variables may have a direct influence on yield and potentially influence secondary considerations such as pests and diseases.

Daily data are available and have been aggregated to provide monthly mean values for each of the 11 variables. We have focussed purely on wheat yield prediction based on meteorological inputs; we do not consider variations in drainage and soil type for example, although on the Rothamsted plot these are likely to be small. We assume that the impacts of cultivar changes during the period are likely to be of secondary importance. Since the wheat is sown in October, we include meteorological variables from the preceding August, to try and capture any preconditioning that may influence crop yield prior to sowing, such as soil moisture content. We do not detrend data, but the magnitudes of the explanatory meteorological variables show considerable variation. To alleviate this inbalance in magnitude, the mean value for the period is subtracted from each monthly value.

2.2. Regional forecasts

England is divided into eight administrative regions (hereafter AR to distinguish them from the wider regional forecasting) (Fig. 1).

Here we use monthly HadUK-Grid meteorological variables (Met Office 2018), which are available for each AR separately, averaged over the AR. The product interpolates data from meteorological stations onto a range of uniform grids and at different resolutions. Details of the interpolation methods can be found in Perry and Hollis (2005). A different range of variables is available from this product, compared with data available from Rothamsted (Table 1), although there is considerable overlap. Our aim here is to demonstrate that NARMAX can be used for regional forecasting without the need for gridded meteorological data. WOFOST is not used for these forecasts as it requires daily data as input.

Table 1
Meteorological variables available for use in NARMAX and the WOFOST crop model, and the source of the data.

Rothamsted		Regional	
NARMAX	**WOFOST**	**NARMAX**	**Abbreviation used**
Source: Rothamsted Met station	Source: Rothamsted Met station	Source: HadUK-Grid	
Monthly (aggregated from daily data)	daily	monthly	
Dewpoint /°C			dp
Solar radiation /J cm^{-2}	Solar radiation /J cm^{-2}		sol
Rainfall /mm	Rainfall /mm	Rainfall/mm	rain
Relative humidity		Relative humidity	rhum
Sun hours		Sun hours	sun
Max temperature /°C	Max temperature /°C	Max temperature /°C	Tmax
Min temperature /°C	Min temperature /°C	Min temperature /°C	Tmin
Wet bulb temperature /°C			Twb
Vapour pressure /mb	Vapour pressure /mb	Vapour pressure /mb	VP
Wind run /km			wrun
Average wind speed /ms^{-1}	Average wind speed /ms^{-1}	Average wind speed /ms^{-1}	wind
		Ground frost days	grf
		Atmospheric pressure /mb	AP

Fig. 1. Administrative regions of England used in the regional forecasts. (Basemap: Wikipedia.org).

Total national wheat yield is supplied from the UK DEFRA open datasets (DEFRA, 2019) available from 1885, which also provide an indication of the percentage of total wheat harvest contributed by each AR (from 1999), summarised in Table 2. While these figures are for 2018, they are relatively constant over time.

Four ARs account for 75 % of the England annual wheat yield; Yorkshire and Humber, East Midlands, East of England and the South East. Due to the proximity of the UK to the jet stream, ARs show considerable variations in weather conditions for a given season and year. For example, the north and west of England receive much more rainfall than the south and east, due to orographic influences and jet stream variability (Hall and Hanna, 2018). We therefore focus on developing a regional wheat yield forecast for the four most significant

Table 2
Percentage of total England wheat yield provided by each administrative region 2018 (DEFRA, 2019) and the contributions of the 4 main wheat growing ARs to the regional yield.

Administrative Region	% of national wheat yield	% of regional yield
North East	4	NA
North West	2	NA
Yorkshire and Humber	14	18.67
East Midlands	20	26.67
West Midlands	10	NA
East of England	28	37.33
South East and London	13	17.31
South West	9	NA

wheat-producing ARs combined, which are more homogenous in terms of meteorological variability, being located in east and south-east England, in the rain shadow of the upland regions. We use a time-constant weighted average of the meteorological variables, based on the contribution of each AR to the regional wheat yield given in Table 2. For example, for rainfall, the AR value is multiplied by the fraction of regional wheat total contributed by the AR (Table 2): in the case of East of England, all meteorological variables are multiplied by 0.3733, whereas those for Yorkshire and the Humber are multiplied by 0.1867. This weighting ensures the contribution of meteorological variables from the ARs is in direct proportion to the contribution to the regional wheat yield.

3. Methods

3.1. NARMAx

NARMAX can reveal and describe non-linear dynamic relationships between a range of input variables (predictors), and the output (response), in this case wheat yield. NARMAX will construct the most parsimonious or compact model that best describes the system; therefore if a linear model provides a good representation of the system, the method will stop at this point (Billings, 2013, p9).

The general form of the NARMAX model for a multiple input single output (MISO) case can be given by (Wei, 2019; Wei and Billings, 2022):

$$\begin{aligned} y(k) = F\big[&y(k-1), y(k-2)\cdots, y\big(k-n_y\big), u_1(k-d), u_1(k-d \\ &-1), u_1(k-d-2), \cdots, u_1(k-n_u), u_2(k-d), u_2(k-d-1), \cdots, u_2(k \\ &-n_u), \cdots, u_r(k-d), u_r(k-d-1), \cdots, u_r(k-n_u), e(k-1), e(k \\ &-2), \cdots, e(k-n_e)\big] + e(k) \end{aligned} \tag{1}$$

where $y(k)$ is the measured system response, $u_j(k)$ ($j = 1,2, \ldots, r$) are the inputs at time k, $e(k)$ is a noise sequence which is not measurable but can be estimated once a model is built; r being the number of input variables; n_y, n_u, and n_e are the maximum lags for the system output, input, and noise; $F(\bullet)$ is some non-linear function to be determined; and d is a time delay (typically $d = 0$ or $d = 1$). For an identified model, the noise $e(k)$ can be estimated as the prediction errors: $e(k) = y(k) - \hat{y}(k)$, where $\hat{y}(k)$ is the predicted value at time instant k generated by an estimated model. The noise terms are included to accommodate the effects of measurement noise, modelling errors, and/or unmeasured disturbances. Note that the NARMAX model (1) for a single-input and single-output case can easily be extended to multiple-input and multiple-output cases (Billings, 2013).

In this study, we consider a special case of model (1), with the following considerations: 1) the time delay and the maximum time lags being zero, that is, $d = n_y = n_u = n_e = 0$; 2) and the total number of input (candidate explanatory) variables used for building models is r. The NARMAX model for such a special case can be written as:

$$y(k) = f(u_1(k), u_2(k), \cdots u_r(k)) + e(k) \tag{2}$$

For example, a simple special case with two inputs u_1 and u_2, the initial full model of degree 2 would be:

$$y(k) = a_0 + a_1u_1(k) + a_2u_2(k) + a_3{u_1}^2(k) + a_4u_1(k)u_2(k) + a_5{u_2}^2(k) + e(k) \tag{3}$$

Note that although power-form polynomials are commonly used as the basic elements for model building in many real applications (e.g. Ayala-Solares et al., 2016, 2018; Aguirre, 2019; Hall et al., 2019; Gu et al., 2019), other types of functions such as fractional-power polynomials (Royston and Sauerbrei, 2008; Wei et al., 2012), and Gaussian and radial basis functions (Chen et al., 1990) can also be used as building elements for model construction.

Also note that in most applications , the identified model is much simpler than the initial full model, because those candidate model terms which are not important are removed from the full model, and only the most important model terms that make significant contributions to explaining the variation of the system output are included in the final model. NARMAX uses an orthogonal forward selection algorithm, called the Forward Regression Orthogonal Least Squares (FROLS) algorithm (Billings, 2013), to select the important terms. Taking the case of multi-input, single-output systems as an example, the idea of the FROLS algorithm is as follows. The procedure of NARMAX modelling begins by specifying a dictionary D consisting of a number (say M) of elements (possible useful model terms). The algorithm searches for and selects the first model term from D in such a way that it explains the variation of the system response (output) better than any other candidate elements. This model term is called B_1. The algorithm then searches for and selects the second model term B_2 as follows: the combination of B_1 and B_2 better explains the variation of the system response than any other combination of B_1 and B_x where B_x can be any of the remaining $M-1$ elements in D (without including B_1). The process continues until the penalized error-to-signal ratio (PESR), a metric used to measure both model performance and model complexity, reaches a minimum. The search procedure is implemented with an orthogonal least squares or one of its variants, in each search step the importance of the selected model term is measured by an index called the error reduction ratio (ERR) (Chen et al. 1990; Billings, 2013). This allows the model to be built up term by term in a manner that exposes the significance of each new term that is added. Model structure detection is a fundamental part of the NARMAX procedure because searching for the structure ensures that the model is as simple as possible and a model with good generalisation properties is obtained. In this way, the model can be written down, and the dependent relationship of the response on the predictor variables and the interactions between the predictors can be analysed. Further details may be found in Wei et al. (2004; 2008), Wei et al. (2007).

In this study, a leave-v-out cross-validation, with v being approximately 10 % of the training sample, is considered in NARMAX for searching the most important regressors (predictors and cross-product terms) by using a FROLS algorithm. A sparse and parsimonious common model is identified from subsets of the training period, and common model parameters are then estimated using all the training data.

We now present the model specifications, considering the regional model first as its terms are simpler.

3.1.1. Regional model specification

The dataset involves a total of 108 candidate predictors, all of which are considered as candidate model inputs in the initial stage. The initial model is of the form:

$$\begin{aligned} y(k) &= f(u_1(k), u_2(k), \cdots, u_r(k)) + e(k) \\ &= a_0 + a_1u_1(k) + a_2u_2(k) + \cdots + a_ru_r(k) + \sum_{i=1}^{r}\sum_{j=1}^{r} a_{ij}u_i(k)u_j(k) + e(k) \end{aligned} \tag{4}$$

where $r = 108$, $y(k)$ represents the wheat yield in year k, and $u_1(k), u_2(k), \cdots u_r(k)$ represent the meteorological input variables. Following Billings (2013), the initial full model (4) involves a total of $((r+1)\times(r+2))/2 = (109 \times 110)/2 = 5995$ candidate model terms. However, as will be shown later that the model can be greatly reduced using the FROLS algorithm.

As growing systems and genetics change, we limit the data to 1984–2017 (34 data points). Earlier data are not used due to inconsistencies in changes in wheat yield during this earlier period. Following a conventional practice, about 70 % of the data are used for model training and the remaining 30 % for model test. For the regional modelling case here, the first 24 years of the data (1984–2007) are used for model estimation, and the remaining 10 years are used as the validation dataset. For convenience of model explanation, data are not pre-processed, as pre-processing would normally result in the loss of the

original physical meanings of the model input variables.

The final identified models are shown in Table 3, where it can be noticed that the FROLS algorithm can generate very simple but accurate nonlinear models for wheat yield prediction.

3.1.2. Rothamsted model specification

While for the regional data case, ordinary power-form polynomial models supply the best fit, our numerical experiments show that such models cannot well capture the varying trend of the response variable (wheat yield) for the Rothamsted case study. This motivated us to consider fractional polynomials. More specifically, for this case we use the following 5 groups of variables to build NARMAX models:

$$\underbrace{x_1,\ldots,x_{132}}_{\text{Group 1}},\ \underbrace{|x_1|^{1/2},\ldots,|x_{132}|^{1/2}}_{\text{Group 2}},\ \underbrace{|x_1|^{1/3},\ldots,|x_{132}|^{1/3}}_{\text{Group 3}},\ \underbrace{|x_1|^{1/4},\ldots,|x_{132}|^{1/4}}_{\text{Group 4}},\ \underbrace{|x_1|^{1/5},\ldots,|x_{132}|^{1/5}}_{\text{Group 5}}$$

where Group (1) contains the original basic 132 explanatory variables, Group (2) contains 132 new variables derived from the 132 basic variables, and so on for Groups (3)–(5). Therefore, there are a total of 132 × 5 = 660 variables in the 5 groups. For convenience, denote by $x_1, x_2, \ldots, x_{660}$ the newly derived variables, and denote by y wheat yield. The following dictionaries are defined:

$$D_0 = \{1\}\ \text{(constant)}$$
$$D_1 = \{x_1, x_2, \ldots, x_{660}\}$$
$$D_2 = \{x_1x_1, x_1x_2, \ldots, x_1x_{660},\ x_2x_2, \ldots, x_2x_{660},\ x_{659}x_{660},\ x_{660}x_{660}\}$$

Note that the dictionary D_2 contains a total of $661 \times 660/2 = 218130$ elements (candidate model terms), which is equivalent to the number of all possible selections of 2 elements taken from 660 objects with repetitions being allowed. The NARMAX model is constructed using the 1 + 660 + 218130 = 218791 elements in $D_0 + D_1 + D_2$ as the building blocks.

Experimental results (see below) show that fractional polynomials are preferred in this model as a better fit between predictor and predictand is obtained, compared with the use of standard polynomials. Fractional powers offer a greater range of curved fits than lower-order polynomials, which are the alternative given the short dataset (Royston and Sauerbrei, 2008). Another advantage of using fractional polynomials is that the initial full model is still a linear-in-the-parameters (i.e., the response variable linearly depends on all the model parameters) form and the resulting models are transparent and interpretable as shown in equations (7) and (8).

The NARMAX approach will build a model based on a training set of the data, with reserved years not included in the model construction being used as a test dataset (out-of-sample). Here, again we use around 70 % of the samples (i.e. the 35 data values of the period 1969–2003) for model estimation, and the remaining 30 % (i.e. the 15 data values of the period 2004–2018) is used for model performance examination. It is important to emphasise that the 2004–2018 data are not used in any stage of the model development.

Table 3
Parameters and coefficients for the best three regional NARMAX models, with verification statistics, including those for the average model.

Variable	Parameter of 9-term model	Parameter of 10-term model	Parameter of 11-term model	P-value
[aug Tmax]* [sep rhum]	0.0023	0.0014	0.0012	$3.24\text{x}10^{-4}$
[jun Tmin]* [jul Tmax]	0.0146	0.0130	0.0097	$1.57\text{x}10^{-6}$
[jan wind]* [dec grf]	0.0146	0.0153	0.0151	$8.28\text{x}10^{-10}$
[sep Tmin]* [oct sun]	0.0003	0.0005	0.0004	0.01
[jul rain]* [jul wind]	−0.0014	−0.0013	−0.0018	$1.29\text{x}10^{-6}$
[mar VP]* [may wind]	0.0662	0.0685	0.0617	$3.69\text{x}10^{-8}$
[apr wind]* [jun VP]	−0.0464	−0.0518	−0.0545	$1.58\text{x}10^{-8}$
[may Tmin]* [dec wind]	0.0293	0.0316	0.0334	$6.50\text{x}10^{-7}$
[apr VP]* [sep wind]	−0.0315	−0.0291	−0.0282	$8.36\text{x}10^{-5}$
[jul VP]* [aug Tmax]		0.0053	0.0063	$8.42\text{x}10^{-5}$
[jun Tmax]* [jul rhum]			0.0009	$2.84\text{x}10^{-3}$
	RMSE_train = 0.11 RMSE_test = 0.60 MAE_train = 0.09 MAE_test = 0.43 Corr_train = 0.98 Corr_test = 0.74	RMSE_train = 0.08 RMSE_test = 0.60 MAE_train = 0.06 MAE_test = 0.47 Corr_train = 0.99 Corr_test = 0.71	RMSE_train = 0.06 RMSE_test = 0.61 MAE_train = 0.05 MAE_test = 0.46 Corr_train = 0.99 Corr_test = 0.73	
Model Averaging:	RMSE_train = 0.06, RMSE_test = 0.60 MAE_train = 0.05, MAE_test = 0.46 Corr_train = 0.99, Corr_test = 0.73			

3.1.3. Model averaging

Given the relatively short timeseries available for constructing the models, the number of observations is small and much smaller than the number of regressors that are available (the explanatory variables and their cross-product interactions). This means that the models can be very sensitive to adding or removing a particular explanatory term. To reduce the risk of using a single model, with its associated uncertainties, we use a model-averaging approach of either two or three models. These models are the best ones selected from a total of 661x662/2 = 218791 candidate models. The models used in model averaging differ only in the number of terms used and have most of their predictors in common. We use a weighted average of the selected best models, to produce a model average. The variance of the prediction errors over the training period are calculated for each model, $\widehat{\sigma}_1{}^2$ and $\widehat{\sigma}_2{}^2$. The weighted average prediction of two models is defined as:

$$\widehat{y} = w_1\widehat{y}_1 + w_2\widehat{y}_2 \tag{5}$$

where $\widehat{y}_1$ and $\widehat{y}_2$ are model predictions from the two models, respectively, and w_1 and w_2 are weight coefficients, defined as

$$w_1 = \frac{1/\widehat{\sigma}_1^2}{1/\widehat{\sigma}_1^2 + 1/\widehat{\sigma}_2^2} \text{and} w_2 = \frac{1/\widehat{\sigma}_2^2}{1/\widehat{\sigma}_1^2 + 1/\widehat{\sigma}_2^2} \tag{6}$$

By extension, a similar procedure can be applied to three models, as

is the case with the regional forecasting models.

3.2. Other machine learning methods

Apart from NARMAX, nine state-of-the-art ML methods, mentioned in Section 1 are also used to build models for the same datasets – both for the Reginal case and Rothamsted case. The nine ML methods are very briefly introduced as follows. **LASSO** (least absolute shrinkage and selection operator) is used for estimating sparse linear regression models; the coefficient estimates can be used to determine which predictors should be included in a model. **DT** (decision tree) is a non-parametric algorithm and graphical representation of all the possible outcomes of a decision based on the explanatory variables; it has a tree-like structure where each node represents a decision, each branch represents the outcome of that decision, and each leaf node represents the final decision. **RF** (random forest) uses a combination of many decision trees; it effectively makes use of trees via ensemble learning approach. **GAM** (generalized additive model is a non-parametric method for dealing with generalized linear models where the response is represented as a linear combination of a set of unknown smooth functions determined by a number of the candidate variables. **GPR** (Gaussian process regression) is a non-parametric approach which uses Gaussian kernel functions of all candidate predictors as basis functions to build regression models to approximate variation of the response variable. **SVM** (support vector machine) is a non-parametric method for building regression models with kernel functions (e.g. radial basis functions, polynomial functions); kernel parameters are determined defining a hyperplane which controls the egression model performance over the training data; model hyperparameters are estimated by using some optimization algorithms. **FFNN** (feedforward neural network) is a type of shallow neural networks which usually only have a single hidden layer. **LSTM** (Long-Short-Term Memory) is a special type of recurrent neural networks (RNN); it is a sequential neural network with deep learning, suitable for time series prediction. **BiLSTM** (bidirectional LSTM) is a variant of LSTM; it allows the input signals flows in both forward and backword directions.

3.3. WOFOST parameterisation for Rothamsted

The WOFOST model was parameterized by using the data of sowing date, maturity date (using harvest date as a proxy) and yield of winter wheat from 1969 to 2003 at Rothamsted. In this study, the development process of WOFOST is changed from the temperature sum model (TSUM) (Supit et al., 1994) to a total heat model (THU) (Matthews and Hunt, 1994). The development parameters included heat units accumulated during sowing to emergence (THUEM), emergence to flowering (THU1), and flowering to maturity (THU2). Using the observation data, THU from sowing to maturity in each year was calculated, and the three cardinal temperatures were 0, 26 and 34 ℃, respectively. The THUEM was determined by assuming that winter wheat emerged when THU accumulated to 75 degree days after sowing. Among the other THU, THU1 accounts for 0.64 and THU2 for 0.36. The initial development value (sowing date) of the model was taken from the measured values from Rothamsted for each year.

It is assumed that the growth differences of winter wheat in different regions are determined by the parameters of some key varieties. According to sensitivity and constraints analysis, they mainly include leaf maximum photosynthetic capacity (AMAX), specific leaf area (SLA) and life span of leaves growing at 35 Celsius (SPAN). These parameters were optimized by using the winter wheat yield data of Rothamsted from 1969 to 2003. The values of the three parameters were 37.5 $\mu mol \bullet m^{-2} \bullet s^{-1}$, 0.0016 $ha \bullet kg^{-1}$ and 38 days respectively.

Soil parameters are obtained from Salter and Williams (1969), including soil moisture content at field capacity, wilting point and saturation. In addition, due to the lack of observational data it is assumed that the initial value of soil moisture in the surface layer is 50 % of the field holding capacity and 60 % of the root layer is the field holding capacity. In practice, in places like the UK where there is sufficient rainfall, the effect of this parameter is eliminated. The wheat yield data of 2004–2017 were used for the independent simulation test of WOFOST. The development parameters and initial values (sowing date) of WOFOST were used for the average values from 1969 to 2003.

4. Results

4.1. Regional forecasts

4.1.1. Regional NARMAX forecasts

Table 3 presents the three best models identified using the NARMAX method and selected from a total of 5595 candidate models, with 9, 10 and 11 parameters respectively. Note that the predictors in the table are the standardized versions of their corresponding original predictors as follows:

$x_{new} = (x_{old} - m_{old})/s_{old}$, where m_{old} and s_{old} are the mean and standard deviation of the associated variable. Data standardization does not change the model structure but the model parameters, with and without standardization, will be different.

The averaged regional model also provides a skilful forecast of wheat yield for the validation period (r = 0.73), with RMSE 0.60. The regional forecast is comparable in skill to the Rothamsted version (see below), with a slightly lower correlation (0.73 c.f. 0.78) and an improved RMSE (0.60c.f. 0.82). Fig. 2 shows the time series of actual and predicted regional yield, using the averaged NARMAX model. Years with large prediction errors are 2008, 2013 and 2015.

Note that the models shown in Table 3 are completely explainable due to their transparent, interpretable, reproducible and parsimonious (TRIP) properties. They explicitly show which factors significantly affect the annual wheat yield, and reveal the quantitative relationship between wheat yield and the most important meteorological factors.

4.1.2. Regional forecasts using other machine learning models

To further evaluate the performance of the proposed NARMAX approach for UK wheat yield modelling and prediction, we also applied nine state-of-the-art machine learning (ML) models, mentioned in Section 3.2, to the same dataset: 1984–2017 (34 samples) in total, of which the first 24 samples (70 %) were used for model training and the remaining 10 samples (30 %) for model testing.

Prediction results from the nine ML models and comparison with actual wheat yield are shown in Fig. 3. For the last three models (artificial neural networks, ANN), their structures are as follows. **FFNN**: 3 layers, hidden layer with 7 neurons; **LSTM**: 5 layers, 1 fully connected

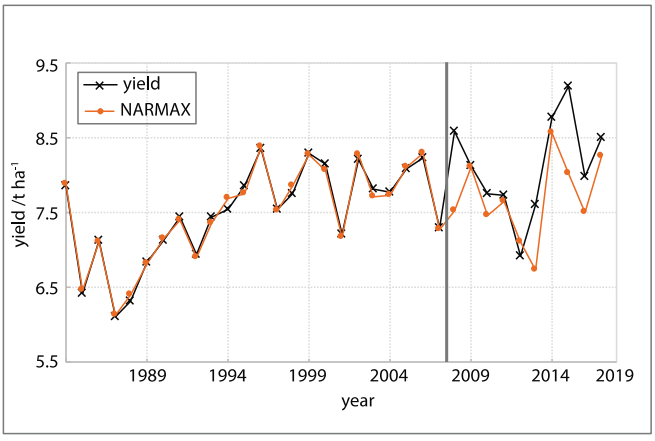

Fig. 2. Observed regional wheat yield and predicted values from the averaged NARMAX model. Black vertical line separates training and testing data (out-of-samples), which are not used in any stage of the model development. The number of candidate predictors used is 108.

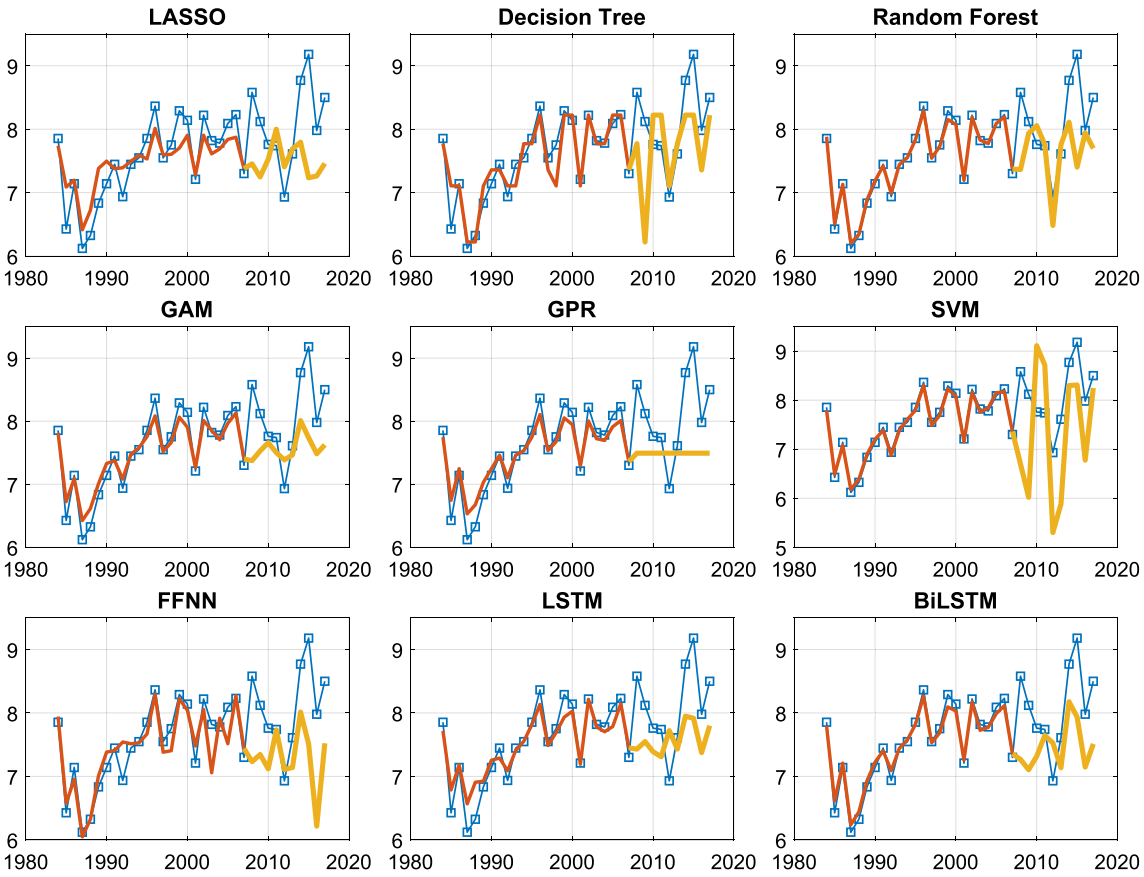

Fig. 3. Prediction results from nine ML models and comparison with actual regional wheat yield. Thin blue line (with squares): actual yield; thick red line: prediction on training dataset; thick yellow line: predictions on test dataset (out-of-samples). The values of the 108 candidate predictors used for these ML models are the same as that of for the NARMAX model listed in Table 3. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 4
The values of the three metrics, RMSE, MAE and correlation coefficient calculated based on the model predictions of regional wheat yield (1984–2017). The values of the 108 candidate predictors used for these ML models are the same as that of for the NARMAX model listed in Table 3.

Methods	Metrics					
	Training Dataset			Test dataset (out-of-sample)		
	RMSE	MAE	Corr	RMSE	MAE	Corr
LASSO	0.3129	0.2546	0.9308	0.9341	0.7728	−0.1869
DT	0.2364	0.1618	0.9249	0.8025	0.6393	0.3695
RF	0.0503	0.0379	0.9985	0.7786	0.5589	0.3642
GAM	0.1655	0.1366	0.9925	0.7600	0.6314	0.5966
GPR	0.1859	0.1518	1.0000	0.8818	0.7372	0
SVM	0.0617	0.0617	0.9964	1.3738	1.2466	0.4745
FFNN	0.2729	0.1949	0.9004	1.0252	0.8588	0.3893
LSTM	0.2135	0.1528	0.9748	0.7586	0.6881	0.5108
BiLSTM	0.0955	0.0812	0.9973	0.8435	0.7598	0.4759
NARMAX	0.06	0.05	0.99	0.60	0.46	0.73

layer, LSTM layer with 16 hidden units, dropout rate being 20 %; **BiLSTM**: 4 layers, 1 fully connected layer, BiLSTM layer with 12 hidden units, dropout rate being 20 %. For all the three neural networks, the input layer has a total of 108 variables. These models are implemented in Matlab (R2022a), and these settings are optimized through cross-validations on the training datasets. Our experiments showed that increasing the numbers of hidden neurons (nodes) or adding hidden layers did not help improve the network performance.

The values of the three metrics, RMSE, MAE and correlation coefficient, on the training and test datasets, calculated from the model predictions are shown in Table 4.

From Fig. 3 and Table 4, it can be observed that the performances of these ML models are not as good as that of the NARMAX models. The reason may be that the small sample size of the datasets makes most of these ML methods ineffective at finding a good model with satisfactory generalization properties (prediction ability). Our previous experience, together with the results here, shows that NARMAX is suitable for solving not only large data modelling problems but also small data modelling tasks; it can usually perform well for solving 'small n, large p' problems with small datasets, where the number of observations (n) is far smaller than the number of input variables (p).

4.2. Rothamsted wheat yield forecasts

4.2.1. Rothamsted yield forecasts using NARMAX models

Using the NARMAX and the FROLS algorithm, we have identified two of the best models below, which will be used in the model averaging:

$$\begin{aligned}\text{Model 1}: \; yield = {} & 16.97\ |\text{may rain}|^{1/5} \\ & +0.37(\text{feb rain} \times \text{jul Tmax}) \\ & -6.97\left(|\text{may rain}|^{1/3} \times |\text{may rain}|^{1/4}\right) \\ & -3.35\left(|\text{jan sun}|^{1/4} \times |\text{jun Twb}|^{1/5}\right)\end{aligned} \tag{7}$$

Table 5
Root mean square error (RMSE) correlation and adjusted R^2 for NARMAX models and the model average. Values are given for both training and testing data. The number of candidate predictors used is 132.

	RMSE		Correlation		Adjusted R-squared	
	Training data	Test data	Training data	Test data	Training data	Test data
Model 1	0.92	0.98	0.76	0.80	0.58	0.49
Model 2	0.78	0.91	0.84	0.78	0.70	0.54
Model average	0.80	0.90	0.83	0.78	Overall R-squared = 0.69 Adjusted R-squared not available (it depends on No of model terms)	

$$\begin{aligned}\text{Model 2}: \quad yield = {} & 17.96\ |\text{may rain}|^{1/5} \\ & +0.31(\text{feb rain} \times \text{jul Tmax}) \\ & -7.57\left(|\text{may rain}|^{1/3} \times |\text{may rain}|^{1/4}\right) \\ & -3.72\left(|\text{jan sun}|^{1/4} \times |\text{jun Twb}|^{1/5}\right) \\ & +0.60(\text{may rain} \times \text{jun Twb})\end{aligned} \tag{8}$$

Note that the predictors in (1) and (2) are the mean-removed versions of their corresponding original predictors as follows: $x_{new} = x_{old} - m_{old}$, where m_{old} is the mean of the associated variable.

A limited range of explanatory variables were selected within the models: February and May rainfall, July Tmax, January sun and June Twb. May rainfall has been converted to absolute anomalies as part of the pre-processing, and so indicates the magnitude of the May rainfall anomaly, not its sign. The variances of the prediction errors, over the training data, from the two models are $\hat{\sigma}_1^2 = 0.8721$ and $\hat{\sigma}_2^2 = 0.6211$, respectively. Some basic model statistics are shown in Table 5.

Fig. 4 shows the inter-annual variability of the averaged NARMAX model compared with the Rothamsted annual wheat yields. It is notable that the local maxima and minima of interannual wheat yield are well captured by the NARMAX models over the testing period, with the exception of the period 2010–2013, where errors are largest. The greatest error occurs in 2013. This coincides with a very late sowing due to wet autumn and winter, so the model may not capture the yield of this particular year, predicting a lower yield influenced by the wet winter. The correlation of training and testing period data are comparable for the average model, (0.83c.f. 0.78), and the RMSE for the testing period is only slightly greater than that obtained for the training period (0.90c.f. 0.80).

4.2.2. Rothamsted yield forecasts using other machine learning models

In this section, the nine machine learning methods used in Section 4.1.2 are employed to model Rothamsted wheat yield. The samples considered are limited to the period 1969–2008. There are a total of 50 samples, of which around 70 % of the samples (i.e. the 35 data values of the period 1969–2003) is used for model estimation, and the remaining 30 % (i.e. the 15 data values of the period 2004–2018) is used for model performance testing. Again, it is important to emphasise that the 2004–2018 data are not used in any stage of the model development. Prediction results from the nine ML models, and the comparison with actual wheat yield are shown in Fig. 5. The values of the three metrics, RMSE, MAE and correlation coefficient, on the training and test datasets, calculated from the model predictions are shown in Table 6.

Note that the results shown in Fig. 5 and Table 6 were produced from the nine ML models trained using the 132 input variables, rather than the 660 candidate predictors mentioned in Section 3.1.2. To fairly compare these ML models with the proposed method, further experiments have been conducted by using all the 660 predictors. All the other experimental settings are the same as that for the 132 predictors.

Prediction results from the nine ML models are shown in Fig. 6. The values of the three metrics, RMSE, MAE and correlation coefficient, on the training and test datasets, calculated from the model predictions are shown in Table 7.

Comparing Fig. 6 and Fig. 5 (Table 7 and Table 6, respectively), it can be observed that the inclusion of the 528 extended candidate predictors into the ML model inputs does not help improve their prediction performances.

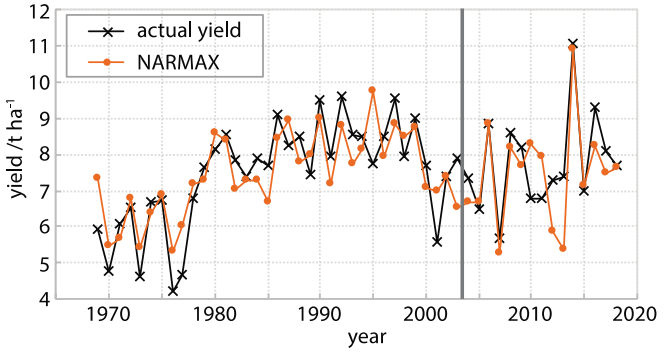

Fig. 4. Predictions from averaged Rothamsted NARMAX model and the actual wheat yield. Bold vertical line separates the training and testing periods. The number of candidate predictors used is 132.

4.3. WOFOST model for Rothamsted

The WOFOST model, albeit with minimal local parameterisation, shows an overall correlation of 0.31 with the actual crop yield. For the model-building period to 2003, the correlation is 0.39. While some periods are modelled well (e.g. 1974–1984; Fig. 7 other periods show limited correlation.

Indeed, the pattern of 11-year running correlations is of concern, showing that after initial success, the WOFOST model correlation deteriorates (Fig. 8).

Fig. 9 compares WOFOST predicted yield and that from NARMAX, for the NARMAX validation period and the WOFOST independent simulation period (2004–2018). Over this period, while WOFOST has a correlation of 0.09 and RMSE of 4.11, the respective values for NARMAX are 0.80 and 0.90. Certain years, such as 2005, 2014 and 2017 are well-predicted by both NARMAX and WOFOST, while 2012 is relatively poorly predicted by both. NARMAX performs particularly well for the periods 2005–2009 and 2014–2018.

4.4. Analysis of Rothamsted NARMAX model terms

The NARMAX model terms selected for Rothamsted are few, and being site-specific, may offer some insight into physical processes that influence wheat yield, when many of the variables such as soil and drainage are controlled. However, for the regional forecasts, while the skill levels are high, many more terms are selected, which are less easily explained, as a result of increased uncertainty in the data and confounding effects of different variables and physical interactions in the different regions. However, the model still provides highly skilful forecasts and should perhaps be better regarded as a black box approach more akin to other machine learning methods.

Tables 8 organises the selected predictors from the Rothamsted NARMAX models according to wheat growth stage (AHDB, 2018).

5. Discussion

Using the NARMAX methodology, we have made skilful out-of-sample predictions of wheat yield for Rothamsted (r = 0.80) and for regional wheat production (r = 0.73). These improve on the skill obtained from other empirical studies (e.g. Chmielewski and Potts (1995), r = 0.33; Landau et al., 2000 (r = 0.41), while the more advanced

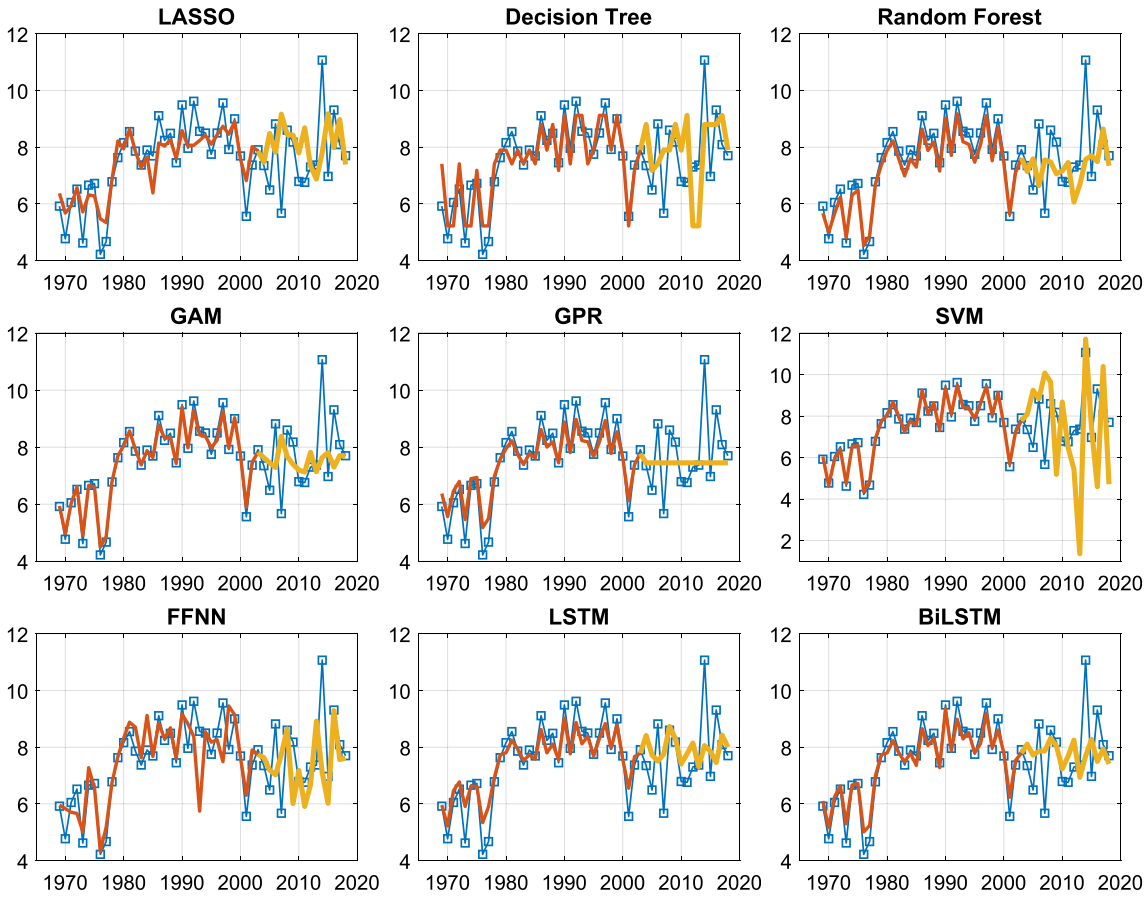

Fig. 5. Machine learning model prediction results and comparison with the actual wheat yield (Rothamsted). Thin blue line (with squares): actual yield; thick red line: prediction on training dataset; thick yellow line: predictions on test dataset (out-of-samples). The values of the 132 candidate predictors used for these ML models are the same as that of for the NARMAX models given by (7) and (8). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 6
The values of the three metrics, RMSE, MAE and correlation coefficient calculated from the model predictions of Rothamsted wheat yield (1969–2018). The values of the 132 candidate predictors used for these ML models are the same as that of for the NARMAX models given by (7) and (8).

Methods	Metrics					
	Training Dataset			Test dataset (out-of-sample)		
	RMSE	MAE	Corr	RMSE	MAE	Corr
LASSO	0.6389	0.4687	0.9107	1.6410	1.2307	−0.2656
DT	0.6130	0.4793	0.9017	1.5492	1.3754	0.2372
RF	0.3244	0.3111	0.9965	1.4423	1.1706	0.0612
GAM	0.1618	0.1222	0.9966	1.4012	1.0258	−0.2063
GPR	0.4211	0.3332	1.0000	1.3174	0.9680	0.0284
SVM	0.1351	0.1351	0.9955	2.8033	2.2069	0.1606
FFNN	0.8186	0.5389	0.8271	1.4940	1.0685	0.2128
LSTM	0.4938	0.3577	0.9789	1.2739	0.9905	0.1937
BiLSTM	0.3300	0.2559	0.9920	1.3313	1.0172	0.0122
NARMAX	0.80	0.67	0.83	0.90	0.69	0.78

approach of Cai et al., (2019), which combines machine learning approaches with satellite data, achieves a skill of 0.87. However, the simplicity of the NARMAX approach, with its use of monthly meteorological variables only, yet high level of skill, and its usefulness for both individual location and regional predictions, make it a valuable contribution to wheat yield forecasting, which can be extended to other regions and crop types, being inexpensive to implement and only requires relatively small amounts of data. First in this section we consider how the above results obtained using NARMAX may be interpreted, although within the constraints of the available data there are naturally some remaining uncertainties and limitations. Further experimental investigations would be required to confirm the specific drivers of seasonal yield; however, the NARMAX approach has highlighted the potential significance of key parameters, such as high sensitivity to May rainfall (Rothamsted dataset). We then consider the wider implications of our findings.

Term 1 ($|\text{may rain}|^{1/5}$) in equations (7) and (8) emphasises the importance of May rainfall in the construction stage of wheat growth. Subtraction of term 3 ($|\text{may rain}|^{1/3} \times |\text{may rain}|^{1/4}$) has the effect of reducing the input value for particularly high or low years of term 1 ($|\text{may rain}|^{1/5}$). It should be remembered that May rainfall has been converted to absolute values as part of the fractional polynomial NARMAX methods (see Section 3.1.2), so |may rain| indicates the magnitude of the May rainfall anomaly, not its sign. The fractional power terms for May rain effectively damp the signal. Interestingly, when mapping back to the absolute rainfall anomalies, it is the years with very small or vary large anomalies, of either sign, that would be predicted to have a lower wheat yield, other thing being equal, based on terms 1 and 3 alone. Thus significant departures from the average May rainfall in either direction, have negative impacts on predicted wheat yield. This makes sense physiologically. May is an important time for the development of the wheat crop: too dry and the canopy will not expand and assimilation will be reduced, whereas too wet means the crop may suffer detrimentally from increased fungal infections. It is both interesting and significant that NARMAX has identified this variable as being important.

Terms 2 (feb_rainxjul_Tmax) and $4(|\text{jan\_sun}|^{1/4}\text{x}|\text{jun\_Twb}|^{1/5})$ (equations (7) and (8) consist of meteorological variables some months

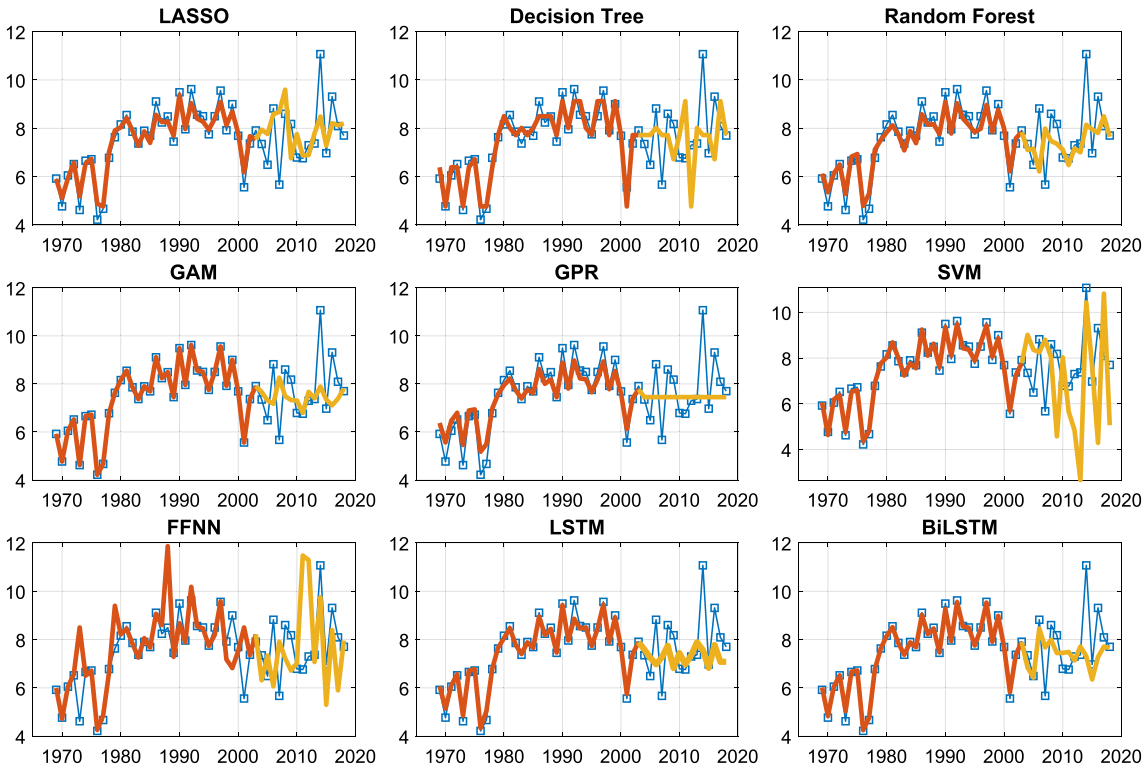

Fig. 6. Machine learning model prediction results and comparison with the actual wheat yield (Rothamsted). Thin blue line (with squares): actual yield; thick red line: prediction on training dataset; thick yellow line: predictions on test dataset (out-of-samples). The 660 candidate predictors used for these ML models are the elements of the dictionary D_1 defined in Section 3.1.2; the values of these 660 variables are the same as that of for the NARMAX models given by (7) and (8). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 7
The values of the three metrics, RMSE, MAE and correlation coefficient calculated from the model predictions of Rothamsted wheat yield (1969–2018). The 660 candidate predictors used for these ML models are the elements of the dictionary D_1 defined in Section 3.1.2; the values of these 660 variables are the same as that of for the NARMAX models given by (7) and (8).

Methods	Metrics					
	Training Dataset			Test dataset (out-of-sample)		
	RMSE	MAE	Corr	RMSE	MAE	Corr
LASSO	0.2957	0.2292	0.9867	0.9012	0.9310	0.3315
DT	0.3690	0.3013	0.9655	1.2546	1.4004	−0.0702
RF	0.3659	0.3075	0.9889	1.6780	0.9823	0.2172
GAM	0.0102	0.0106	0.9998	1.3383	0.9906	−0.0270
GPR	0.4211	0.3332	0.9997	1.3704	0.9680	0.0284
SVM	0.1351	0.1351	0.9957	1.3174	2.1734	0.1333
FFNN	1.1336	0.5421	0.7127	2.5962	1.6860	0.0799
LSTM	0.1766	0.1026	0.9962	2.1119	0.9926	0.2546
BiLSTM	0.2957	0.2292	0.9987	1.3155	0.9310	0.2744
NARMAX	0.80	0.67	0.83	0.90	0.69	0.78

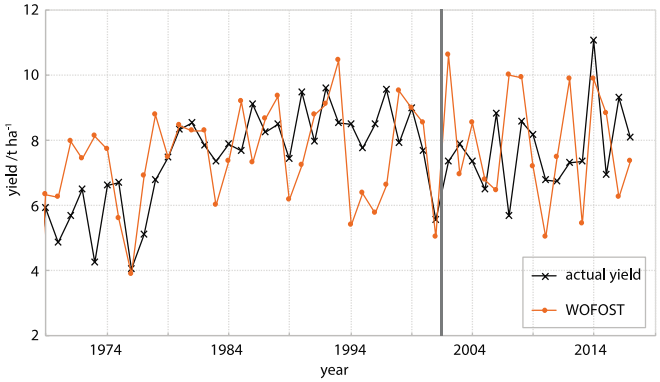

Fig. 7. WOFOST modelled and actual crop yields, Rothamsted first wheat, 1969–2017. Bold vertical line separates training and testing periods.

apart. Term 2 is based on straightforward anomalies; the sign and magnitude are both important. However, term 4 is again based on absolute anomalies, with fractional powers, therefore it is the magnitude and not sign of these anomalies in term 4 that are important. The second term (feb_rainxjul_Tmax) is added, therefore high values contribute to higher wheat yields. As values are determined by multiplication of positive and negative anomalies, positive contributions to wheat yield arise when both components of the term have the same sign: positive/negative July Tmax anomalies combined with positive/negative February rain anomalies. On the other hand, the lowest values of term 2, reducing predicted wheat yield, occur when February rain and July Tmax are of opposite signs. Largest contributions occur in 2014, 1995, 1993, 1990 and 1998, whereas the greatest reductions to predicted wheat yield from this term occur in 1974, 1976, 2007,1977 and 2006. 1995 is a "bust forecast", where this term may have had undue influence on predicting a yield that was quite a bit higher than the actual yield. A drier than normal February followed by a warmer than average July, or a wet February followed by a cool July, both have a negative impact on predicted wheat yield. This may reflect a priming process whereby environmental conditions in February prime the wheat to respond more sensitively to July Tmax.

The fourth term is subtracted and hence always has a negative influence, possibly related to diseases linked to humidity variability. High values for this term have the biggest negative impact on wheat yield. Both large above and below average anomalies of each component contribute to this negative impact on wheat yield. This is the hardest term to interpret and patterns are less clear. High humidity may have an

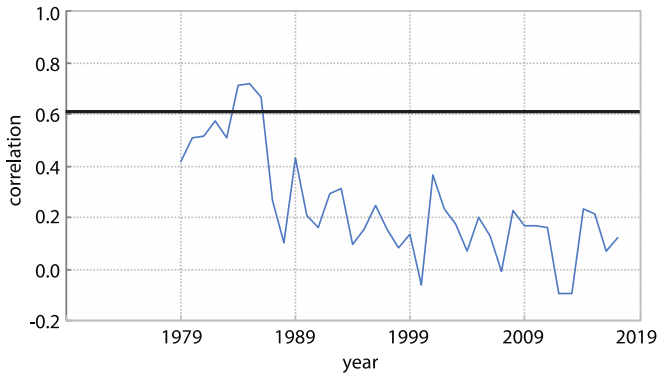

Fig. 8. 11-year running correlations between actual and WOFOST modelled first wheat yield, Rothamsted, 1969–2017. Bold horizontal line denotes 95 % significance (p = 0.05) correlation value.

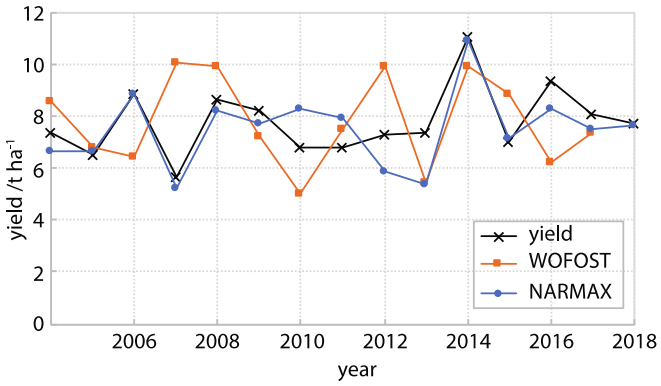

Fig. 9. Comparison of NARMAX and WOFOST predicted yields and actual yields from Rothamsted, 2004–2018.

Table 8
Predictor variables from the NARMAX Rothamsted models, organised by wheat development phase.

Variable	Development Phase	Physiological processes	Notes
May rain	Construction GS39	Start of ear formation/ root growth, stem storage & elongation	Main N uptake time. Soil N release stimulated by warm moist soils N availability controls canopy expansion
Feb rain	foundation	tillering/leaf emergence/root growth	Ear initiation
July Tmax	production	Grain filling	Disease, water availability Hot weather reduces grain growth by shortening growing period
June wet bulb T	Construction/ production GS59	Ear formation/ flowering/grain filling	Potential impact of pests and diseases
January sun	FoundationGS21 (main shoot, 1 tiller, >9 leaves)	Tillering/leaf emergence/root growth	Ear initiation

adverse impact through diseases such as mildews, while low humidity may adversely affect other processes such as pollination. Again, particularly large January sun anomalies, either positive or negative, may serve to precondition the wheat to susceptibility to humidity variations in June.

The fifth term is only used in model 2 (equation (8); (may_rainxJun_Twb). The components of the term are not fractional powers. If both May rain and June Twb anomalies are of the same sign, the effect is a positive one on wheat yield, while if they are of opposite sign, the effect is negative. This term plays a minor role in the model averaging procedure.

It is interesting that the meteorological predictors selected for the regional models are almost entirely different to those identified for Rothamsted and are much more complex. While initially surprising, further reflection indicates that this is likely to be the case. The correlation between Rothamsted and regional wheat yields is 0.08, indicating that different processes are influencing yield variability. A forecast model for an individual location will be heavily influenced by local factors, which even out over the regional forecasts, which will be more impacted by larger scale weather variability. For example, the loamy soil at Rothamsted has a high field capacity and will result in a negative impact of increased rainfall on yield (Chmielewski and Potts,1995). The pooling of different ARs in the regional forecast increases uncertainty and the model will be unable to explain secondary impacts of meteorological variables on wheat yield due to the confounding influences of conditions in different regions. However, this regional model still proves highly skilful when predicting yields for a validation dataset, and can be used as a “black box” forecasting tool, akin to other machine learning techniques.

While the criticism of empirical models focuses on how they approximate the underlying processes that determine crop yield (Landau et al., 2000), here the emphasis is on providing a simple, easy-to-implement predictive tool. However, the predictors selected, while complex, show some agreement with those identified in previous studies and may help to shed some light on important links between crop physiology, meteorological variables and secondary impacts such as pests and diseases.

A limitation of crop models such as WOFOST is that they are unable to replicate these secondary effects, whereas in reality crop yield is heavily influenced by the interaction of meteorological variables with pests, disease, soil conditions and the effectiveness of management responses to these (Jamieson et al., 1999). While May rainfall maybe particularly relevant to Rothamsted due to soil types, in other locations where the soil is different this may not be a factor. Baier and Robertson (1968) find soil moisture to be more important in determining yield than rainfall. In particular, soil moisture before anthesis is identified as critical, which may relate to the May rainfall selected for the Rothamsted model. As they use a number of sites, this could be a reflection of the interaction of rainfall with soil type. At a specific site, soil does not vary, and therefore rainfall is likely to be a useful variable, although the nature of the impact of rainfall in different months will vary between sites according to soil type. Hence a model developed for a specific location such as Rothamsted cannot be used directly for another location where very different local conditions will interact with the meteorological variables. Similarly, interactions with local conditions may favour or discourage pests and diseases at different times of the year in different locations.

February rainfall is identified for Rothamsted. Higher rainfall alone would indicate an increased yield, at odds with other studies which find a negative relationship with February rainfall (e.g. Landau et al., 2000), who suggested it might be a possible consequence of leached fertiliser which is often applied at this time. Here, as noted above, the February rainfall is an interaction term with July Tmax. There are two possible explanations here. It is possible for a connection to exist between these meteorological variables several months apart. Previous studies (Wedgbrow et al., 2002; Kettlewell et al., 2003; Qian and Saunders 2003) find a link between the winter North Atlantic Oscillation (NAO), a semi-permanent dipole of high and low pressure over the Atlantic (Hurrell, 1995; Hanna and Cropper, 2017) and weather conditions in the following summer. For Rothamsted, a positive winter NAO would be

associated with mild, yet dry conditions (Hall and Hanna, 2018), followed by a more positive NAO in summer, with warmer drier weather, which has been shown to impact on yield (Atkinson et al., 2005). However, the signs of relationship are not correct, as a wetter winter followed by a warmer summer would enhance wheat yields and vice versa. This is in agreement with Brown (2013) who finds that the winter NAO relationship with wheat yield does not hold for Scotland. Hall and Hanna (2018) show that the relationship of the NAO with regional temperature and rainfall is not consistent across the country: the usual association of a positive winter NAO being wet only holds for the western side of the UK. It is more likely that the earlier meteorological variables (February rain, term 2; January sun, term 4) have an impact on wheat physiology which can precondition wheat to the influence of the second variable later in the season (July maximum temperature, term 2; June wet bulb temperature, term 4). For example, ear initiation takes place in the foundation stage. This may affect certain attributes of the ear which are later susceptible to either July maximum temperatures (term 2) or June wet bulb temperatures (term 4).

NARMAX or other machine learning tools will never substitute for CGSMs and vice versa. NARMAX is a complementary tool that can provide forecasts with significant skill, whilst also showing some insight into key environmental drivers of crop responses (e.g. sensitivity to May rainfall). NARMAX models cannot be extended beyond the extent of the data within which they are parameterised. However, CGSMs benefit from describing the state-of-the-art in mechanistic understanding of the key bio-physical driving crop growth, and as such should provide an excellent tool for forecasting impacts of long-term climate change on yield.

NARMAX is capable of generating skilful wheat yield forecasts, at both site and regional levels. It provides an easy-to-implement prediction tool, circumventing the problem of limited data availability for specific sites. The HadUK-Grid dataset can be used for sites where there is no meteorological data. However, a next step is to make the forecasting operational; HadUK-Grid is not available in near-real time and so would be unable to provide an up-to-date prediction. It will be necessary to find or develop other meteorological products to meet this need, or to work closely with operational forecasting centres. It would also be possible to develop a spatio-temporal version of NARMAX, which takes into account meteorological variables and wheat yields at a range of locations, together with other local information such as soil and crop-husbandry factors, and furthermore the method can easily be applied to other crops.

Declaration of Competing Interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Simon Pearson reports financial support was provided by Science and Technology Facilities Council. Shibo Fang reports financial support was provided by National Natural Science Foundation of China.

Data availability

Data will be made available on request.

Acknowledgements

This work was supported by the STFC Newton Fund programme, project ST/N006836/1, the Projects of International Cooperation and Exchanges NSFC (NSFC-RCUK-STFC), project 61661136005, NERC (Ref. NE/V001787/1), and EPSRC (Ref. EP/H00453X/1). We thank the Lawes Agricultural Trust and Rothamsted Research for data from the e-RA database. The Rothamsted Long-term Experiments National Capability (LTE-NC) is supported by the UK BBSRC (Biotechnology and Biological Sciences Research Council, BBS/E/C/000 J0300) and the Lawes Agricultural Trust.

References

Aguirre, L.A., 2019. Bird's Eye View of Nonlinear System Identification. arXiv: 1907.06803.
AHDB, 2019. Wheat Growth Guide. Downloaded from http://cereals.ahdb.org.uk November 1st 2018.
Asseng, S., Martre, P., Maiorano, A., Rötter, R.P., O'leary, G.J., Fitzgerald, G.J., Girousse, C., Motzo, R., Giunta, F., Babar, M.A. and Reynolds, M.P., 2019. Climate change impact and adaptation for wheat protein. Global Change Biol. 25, 155-173.
Asseng, S., Ewert, F., Martre, P., Rötter, R.P., Lobell, D.B., Cammarano, D., Kimball, B.A., Ottman, M.J., Wall, G.W., White, J.W., Reynolds, M.P., 2015. Rising temperatures reduce global wheat production. Nat. Clim. Chang. 5, 143.
Atkinson, M.D., Kettlewell, P.S., Hollins, P.D., Stephenson, D.B., Hardwick, N.V., 2005. Summer climate mediates UK wheat quality response to winter North Atlantic Oscillation. Agric. For. Meteorol. 130, 27–37.
Ayala-Solares, J.R., Wei, H.L., Boynton, R.J., Walker, S.N., Billings, S.A., 2016. Modeling and prediction of global magnetic disturbance in near-Earth space: A case study for Kp index using NARX models. Space Weather 14, 899–916.
Ayala-Solares, J.R., Wei, H.-L., Bigg, G.R., 2018. The variability of the Atlantic meridional circulation since 1980, as hindcast by a data-driven nonlinear systems model. Acta Geophys. 66, 683–695.
Baier, W., Robertson, G.W., 1968. The performance of soil moisture estimates as compared with the direct use of climatological data for estimating crop yields. Agric. Meteorol. 5, 17–31.
Baron, C., Sultan, B., Balme, M., Sarr, B., Traore, S., Lebel, T., Janicot, S., Dingkuhn, M., 2005. From GCM grid cell to agricultural plot: scale issues affecting modelling of climate impact. Philos. Trans. R. Soc., B 360, 2095–2108.
Bengio, Y., 2013. Practical recommendations for gradient-based training of deep architectures. In K.-R. M üller, G. Montavon, and G. B. Orr, editors, Neural Networks: Tricks of the Trade. Springer.
Billings, S.A., 2013. Non-Linear System Identification: NARMAX Methods in the Time, Frequency, and Spatio-Temporal Domains. Wiley, London.
Boogaard, H., Kroes, J., 1998. Leaching of nitrogen and phosphorus from rural areas to surface waters in the Netherlands. Nutr. Cycl. Agroecosyst. 50, 321–324.
Boogaard, H., Wolf, J., Supit, I., Niemeyer, S., van Ittersum, M.A., 2013. Regional implementation of WOFOST for calculating yield gaps of autumn-sown wheat across the European Union. Field Crop Res 143, 130–142.
Breiman, L., 2001. Random forests. Mach. Learn. 45, 5–32.
Breiman, L., Friedman, J., Olshen, R., Stone, C., 1984. Classification and Regression Trees. CRC Press, Boca Raton, FL.
Brown, I., 2013. Influence of seasonal weather and climate variability on crop yields in Scotland. Int. J. Biometeorol. 57, 605–614.
Cai, Y., Guan, K., Lobell, D., Potgieter, A.B., Wang, S., Peng, J., Xu, T., Asseng, S., Zhang, Y., You, L., Peng, B., 2019. Integrating satellite and climate data to predict wheat yield in Australia using machine learning approaches. Agric. For. Meteorol. 274, 144–159.
Challinor, A.J., Wheeler, T.R., Hemming, D., Upadhyaya, H.D., 2009. Crop yield simulations using a perturbed crop and climate parameter ensemble: sensitivity to temperature and potential for genotypic adaptation to climate change. Climate Res. 38, 117–127.
Chen, S., Billings, S.A., Cowan, C.F.N., Grant, P.M., 1990. Practical identification of NARMAX models using radial basis functions. Int. J. Control 52, 1327–1350.
Chmielewski, F.M., Potts, J.M., 1995. The relationship between crop yields from an experiment in southern England and long-term climate variations. Agric. For. Meteorol. 73, 43–66.
de Wit, A., Boogaard, H., Fumagalli, D., Janssen, S., Knapen, R., van Kraalingen, D., Supit, I., van der Wijngaart, R., van Diepen, K., 2019. 25 years of the WOFOST cropping systems model. Agr. Syst. 168, 154–167.
DEFRA 2019. Cereal and Oil Seed Production Survey. Downloaded from http://bit.lyDefraStats on May 15th 2019.
Fan, R.-E., Chen, P.-H., Lin, C.-J., 2005. Working set selection using second order information for training support vector machines. J. Mach. Learn. Res. 6, 1889–1918.
Foulkes, M.J., Snape, J.W., Shearman, V.J., Reynolds, M.P., Gaju, O., Sylvester-Bradley, R., 2007. Genetic progress in yield potential in wheat: recent advances and future prospects. J. Agric. Sci. 145, 17–29.
Goodfellow, I., Bengio, Y., Courville, A., 2016. Deep Learning. MIT Press.
Graves, A., Schmidhuber, J., 2005. Framewise phoneme classification with bidirectional LSTM and other neural network architectures. Neural Netw. 18, 602–610.
Gu, Y., Wei, H.-L., Boynton, R.J., Walker, S.N., Balikhin, M.A., 2019. System identification and data-driven forecasting of AE index and prediction uncertainty analysis using a new cloud-NARX model. J. Geophys. Res.: Space Physics 124, 248–263.
Hall, R.J., Hanna, E., 2018. North Atlantic circulation indices: links with summer and winter UK temperature and precipitation and implications for seasonal forecasting. Int. J. Climatol. https://doi.org/10.1002/joc.5398.
Hall, R.J., Wei, H.-L., Hanna, E., 2019. Complex systems modelling for statistical forecasting of winter North Atlantic atmospheric variability: a new approach to North Atlantic seasonal forecasting. Q. J. R. Meteorolog. Soc. 145, 2568–2585.
Hanna, E., Cropper, T.E., 2017. North Atlantic Oscillation. Oxford Research Encyclopedia of Climate Science. https://doi.org/10.1093/9780190228620.013.22.
Hansen, J.W., Jones, J.W., 2000. Scaling-up crop models for climate variability applications. Agr. Syst. 65, 43–72.
Hastie, T., Tibshirani, R., Friedman, J., 2008. The Elements of Statistical Learning, 2nd edition. Springer, New York.

Hochreiter, S., Schmidhuber, J., 1997. Long short-term memory. Neural Comput. 9, 1735–1780.

Hoffmann, H., Zhao, G., Asseng, S., Bindi, M., Biernath, C., Constantin, J., Coucheney, E., Dechow, R., Doro, L., Eckersten, H., Gaiser, T., Grosz, B., Heinlein, F., Kassie, B.T., Kersebaum, K.-C., Klein, C., Kuhnert, M., Lewan, E., Moriondo, M., Nendel, C., Priesack, E., Raynal, L., Rogerro, P.P., Rötter, R.P., Siebert, S., Specka, X., Tao, F., Teixera, E., Trombi, G., Wallach, D., Weihermüller, L., Yelurapati, J., Ewert, F., 2016. Impact of spatial soil and climate input data agrgregation on regional yield simulations. PLOS-one. https://doi.org/10.1371/journal.pone.0151782.

Hurrell, J.W., 1995. Decadal trends in the North Atlantic Oscillation: regional temperature and precipitation. Science 269, 676–679.

Iizumi, T., Tanaka, Y., Sakurai, G., Ishigooka, Y., Yokozawa, M., 2014. Dependency of parameter values of a crop model on the spatial scale of simulation. J. Adv. Model. Earth Syst. 6, 527–540.

Jamieson, P.D., Porter, J.R., Semenov, M.A., Brooks, R.J., Ewert, F., Ritchie, J.T. 1999. Comments on "Testing winter wheat simulation models predictions against observed UK grain yield" by Landau et al. (1998). Agric. Forest Meteorol. 96, 157-161.

Keating, B.A., Carberry, P.S., Hammer, G.L., Probert, M.E., Robertson, M.J., Holzworth, D., Huth, N.I., Hargreaves, J.N., Meinke, H., Hochman, Z., McLean, G., 2003. An overview of APSIM, a model designed for farming systems simulation. Eur. J. Agron. 18, 267–288.

Kettlewell, P.S., Stephenson, D.B., Atkinson, M.D., Hollins, P.D., 2003. Summer rainfall and wheat grain quality: relationships with the North Atlantic Oscillation. Weather 58, 155–163.

Knight, S.; Kightley, S.; Bingham, I.; Hoad, S.; Lang, B.; Philpott, H.; Stobart, R.; Thomas, J.; Barnes, A.; Ball, B. Desk Study to Evaluate Contributory Causes of the Current 'Yield Plateau' in Wheat and Oilseed Rape; HGCA Project Report No. 502; AHDB: Kenilworth, UK, 2012.

Landau, S., Mitchell, R.A.C., Barnett, V., Colls, J.J., Craigon, J., Payne, R.W., 2000. A parsimonious, multiple-regression model of wheat yield response to environment. Agric. For. Meteorol. 101, 151–166.

Lobell, D.B., Field, C.B., 2007. Global scale climate–crop yield relationships and the impacts of recent warming. Environ. Res. Lett. 2, 014002.

Loh, W.Y., 2002. Regression trees with unbiased variable selection and interaction detection. Stat. Sin. 12, 361–386.

Lou, L., Caruana, R., Gehrke, J. and Hooker, G. 2013. Accurate intelligible models with pairwise interactions, in: ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, 2013, pp. 623–631.

Macdonald, A. , Poulton, P. , Clark, I. , Scott, T. , Glendining, M. , Perryman, S. , Storkey, J. , Bell, J. , Shield, I. , McMillan, V. and Hawkins, J., 2018. Guide to the Classical and other Long-term experiments, Datasets and Sample Archive. Rothamsted Research, Harpenden, UK. doi: 10.23637/ROTHAMSTED-LONG-TERM-EXPERIMENTS-GUIDE-2018.

Matthews, R.B., Hunt, L.A., 1994. GUMCAS: A model describing the growth of cassava (*Manihot esculenta* L. Crantz). Field Crop Res 36, 69–84.

Pantazi, X.E., Moshou, D., Alexandris, T., Whetton, R.L., Mouazen, A.M., 2019. Wheat yield prediction using machine learning and advanced sensing techniques. Comput. Electron. Agric. 121, 57–65.

Perry, M., Hollis, D., 2005. The development of a new set of long-term climate averages for the UK. Int. J. Climatol. 25, 1023–1039.

Perryman, S.A., Castells-Brooke, N.I., Glendining, M.J., Goulding, K.W., Hawkesford, M. J., Macdonald, A.J., Ostler, R.J., Poulton, P.R., Rawlings, C.J., Scott, T., Verrier, P.J., 2018. The electronic Rothamsted Archive (e-RA), an online resource for data from the Rothamsted long-term experiments. Sci. Data 5, 180072.

Qian, B., Saunders, M.A., 2003. Summer U.K. Temperature and its links to preceding Eurasian Snow Cover, North Atlantic SSTs, and the NAO. J. Clim. 16, 4108–4120.

Rasmussen, C.E., Williams, C.K.I., 2006. Gaussian Processes for Machine Learning. MIT Press, Cambridge, MA, USA.

Ray, D.K., Mueller, N.D., West, P.C., Foley, J.A., 2013. Yield trends are insufficient to double global crop production by 2050. PLoS One 8 (6), e66428.

Reed, R., Marks, R.J., 1999. Neural Smithing: Supervised Learning in Feedforward Artificial Neural Networks. MIT Press, Cambridge, MA, USA.

Rosenzweig, C., Jones, J.W., Hatfield, J.L., Ruane, A.C., Boote, K.J., Thorburn, P., Antle, J.M., Nelson, G.C., Porter, C., Janssen, S., Asseng, S., Basso, B., Ewert, F., Wallach, D., Baigorria, G., Winter, J.M., 2013. The Agricultural Model Intercomparison and Improvement Project (AgMIP): protocols and pilot studies. Agric. For. Meteorol. 170, 166–182.

Royston, P., Sauerbrei, W., 2008. Multivariable Model-building: A Pragmatic Approach to Regression Analysis based on Fractional Polynomials for Modelling Continuous Variables. Wiley, Chichester.

Salter, P.J., Williams, J.B., 1969. The moisture characteristics of some Rothamsted, Woburn and Saxmundham soils. J. Agric. Sci. Cambridge 73, 155–158.

Schuster, M., Paliwal, K.K., 1997. Bidirectional recurrent neural networks. IEEE Trans. Signal Process. 45 (11), 2673–2681.

Shiferaw, B., Smale, M., Braun, H.J., Duveiller, E., Reynolds, M., Muricho, G., 2013. Crops that feed the world 10. Past successes and future challenges to the role played by wheat in global food security. Food Security 5, 291–317.

Siami-Namini, S., Tavakoli, N. and Namin, A.S. 2019. The Performance of LSTM and BiLSTM in Forecasting Time Series. In Proceedings of the 2019 IEEE International Conference on Big Data (Big Data), pp.3285–3292.

Supit, I., Hooyer, A.A., van Diepen, C.A., 1994. System description of the WOFOST 6.0 crop simulation model implemented in CGMS. Theory and Algorithms 1, 146 pp.

van der Velde, M., van Diepen, C.A., Baruth, B., 2019. The European crop monitoring and yield forecasting system: Celebrating 25 years of JRC MARS Bulletins. Agr. Syst. 168, 56–57.

Van Diepen, C.V., Wolf, J., Van Keulen, H., Rappoldt, C., 1989. WOFOST: a simulation model of crop production. Soil Use Manag. 5, 16–24.

Wedgbrow, C.S., Wilby, R.L., Fox, H.R., O'Hare, G., 2002. Prospects for seasonal forecasting of summer drought and low river flow anomalies in England and Wales. Int. J. Climatol. 22, 219–236.

Wei, H.-L. and Billings, S.A. 2022. Modelling COVID-19 pandemic dynamics using transparent, interpretable, parsimonious and simulatable (TIPS) machine learning models: A case study from systems thinking and system identification perspectives. In Jiang, R., Crookes, D., Wei, H. L., Zhang, L., Chazot, P. (Editors): Recent Advances in AI-enabled Automated Medical Diagnosis, pp.13-27.

Wei, H.-L., Billings, S.A., Liu, J., 2004. Term and variable selection for non-linear system identification. Int. Control J. 77, 86–110.

Wei, H.-L., Billings, S.A., 2008. Model structure selection using an integrated forward orthogonal search algorithm assisted by squared correlation and mutual information. Int. J. Model. Ident. Control 3, 341–356.

Wei, H.-L., Zhao, Y., Billings, S.A. and Zhao, J. 2012. Fractional power NARX model identification using a harmony search algorithm. In: IEEE International Conference on Computational Intelligence for Measurement Systems and Applications (CIMSA), Tianjin, China, 2–4 July 2012, pp.102–107.

Wei, H.-L., Zhu, D., Billings, S.A., Balikhin, M.A., 2007. Forecasting the geomagnetic activity of the Dst index using multiscale radial basis function networks. Adv. Space Res. 40, 1863–1870.

Wei, H.-L. 2019. Sparse, interpretable and transparent predictive model identification for healthcare data analysis. In: Advances in Computational Intelligence. IWANN 2019, pp. 103–114.