Coupling agricultural system models with machine learning to facilitate regional predictions of management practices and crop production

Liujun Xiao, Guocheng Wang, Hangxin Zhou, Xiao Jin and Zhongkui Luo

1 College of Environmental and Resource Sciences, Zhejiang University, Hangzhou 310058, People’s Republic of China
2 State Key Laboratory of Atmospheric Boundary Layer Physics and Atmospheric Chemistry, Institute of Atmospheric Physics, Chinese Academy of Sciences, Beijing 100029, People’s Republic of China
3 Key Laboratory of Agricultural Remote Sensing and Information System, Zhejiang University, Hangzhou 310058, People’s Republic of China
4 Academy of Ecological Civilization, Zhejiang University, Hangzhou 310058, People’s Republic of China
5 Key Laboratory of Environment Remediation and Ecological Health, Ministry of Education, Zhejiang University, Hangzhou 310058, People’s Republic of China

* Author to whom any correspondence should be addressed.

E-mail: luozk@zju.edu.cn

Keywords: crop model, carbon sequestration, nitrogen fertilizers, machine learning, sustainable agriculture, yield

Supplementary material for this article is available online

Abstract

Process-based agricultural system models are a major tool for assessing climate-agriculture-management interactions. However, their application across large scales is limited by computational cost, model uncertainty, and data availability, hindering policy-making for sustainable agricultural production at the scale meaningful for land management by farmers. Using the Agricultural Production System sIMulator (APSIM) as an example model, the APSIM model was run for 101 years from 1980 to 2080 in a typical cropping region (i.e., the Huang-Huai-Hai plain) of China. Then, machine learning (ML)-based models were trained to emulate the performance of the APSIM model and used to map crop production and soil carbon (which is a key indicator of soil health and quality) dynamics under a great number of nitrogen and water management scenarios. We found that ML-based emulators can accurately and quickly reproduce APSIM predictions of crop yield and soil carbon dynamics across the region under different spatial resolutions, and capture main processes driving APSIM predictions with much less input data. In addition, the emulators can be easily and quickly applied to identify optimal nitrogen management to achieve yield potential and sequester soil carbon across the region. The approach can be used for modelling other complex systems and amplifying the usage of agricultural system models for guiding agricultural management strategies and policy-making to address global environmental challenges from agriculture intensification.

1. Introduction

Process-based mechanistic models integrate physical, chemical, and biological knowledge into the modelling of complex climate-agriculture-management interactions (Reid et al. 2010, Bonan and Doney 2018). They are a major tool to address global environmental challenges (e.g. climate change, food security, biodiversity conservation, and natural resource management). In agroecosystems, crop production involves complex interactions with climate, soil, and management practices (e.g. the application of fertilizers, irrigation, and residue management). In addition to crop production, agricultural lands face a big challenge to reduce environmental footprints (e.g. soil carbon loss, greenhouse gas emissions, and nutrient leaching), which also actively respond to those factors controlling crop production (Tilman et al. 2002, Altieri 2018). It is very difficult, if not impossible, to conduct manipulative field experiments or extrapolate plot-scale results across large regions. Process-based farming system models with detailed plant and soil processes are usually used to simulate crop productivity, nutrient
cycling and environmental impacts as influenced by environmental variability and management interventions (Bamière et al 2011, Luo et al 2017b).

However, some significant limitations are hindering the usage of process-based models. First, models are abstractions of complex systems. The consequence is that those discrepancies in model structure and parameterization can induce large uncertainties in model predictions (Bradford et al 2016, Getz et al 2018). To partly address this uncertainty, multi-model ensemble simulations are usually required (Alissar et al 2012, Bopp et al 2013, Friedlingstein et al 2014). However, such ensemble simulations can only focus on several specific scenarios, and rarely reflect the real world and potential possibilities (Luo et al 2016, Bonan and Doney 2018). Second, although we are in a data-rich era, most data are not specifically collected for modelling purposes and can be rarely directly used by process-based biophysical models (Weng and Luo 2011). When applying crop models across large scales under fine resolution, great uncertainties may be induced by limited data availability to drive the model (Luo et al 2017a, Xiong et al 2020). Third, applying high-resolution spatiotemporal process-based models over large extents presents significant computational challenges (Levin et al 1997, Nichols et al 2011, Dawson and Gerritsen 2016). Whilst a single plot-based scenario may be completed within just seconds or minutes, many hundreds even thousands of scenarios may be required over fine spatial units across large scales. This cannot be done within an acceptable time using traditional computing methods. Recent advances in computing techniques such as parallel high-performance computing (HPC, e.g. supercomputers) have reduced this computational barrier (Zhao et al 2013). However, using HPC usually needs to migrate to new hardware and software environments (Bryan 2013). To overcome these limitations, machine learning (ML) techniques are increasingly used for predicting crop yield and its intervention with climate and management (Jiang et al 2019, Paudel et al 2020). Compared with process-based models, ML has limited ability to interpret processes underlying the interactions of crops, climate, soil, and management practices. However, coupling process-based model with ML can not only gain process understanding but also improve the efficiency of large-scale simulations at finer resolutions, representing a prevailing approach for large-scale prediction of agroecosystems (Feng et al 2019, Folberth et al 2019, Paudel et al 2020).

In this study, we used ML as a partner of a biophysical model—the Agricultural Production Systems sIMulator (APSIM) (Holzworth et al 2014)—to conduct regional simulations under a great number of management scenarios. The APSIM model allows flexible specification of management options such as crop and rotation types, tillage, residue management, and fertilizer application, and has the capacity to simulate the interactions of crop growth and soil processes with soil, climate and management practices. The APSIM model was first applied to simulate soil organic carbon (SOC, which is a key indicator of soil health and quality) dynamics and yield during the period 1980–2080 under various agricultural management practices (i.e. scenarios) at 3409 grid cells across a typical cropping region, i.e. the Huang-Huai-Hai plain (HHHp) in China. Based on the APSIM simulations, then we developed machine learning-based statistical models to mimic APSIM predictions. The ML approach enables us to cover all possible scenarios considering the whole range of environmental variability and management options, otherwise impossible using APSIM alone because of huge computational demands and limited data availability. Specifically, the major aim of this study is to demonstrate a novel and powerful approach combining process-based biophysical models and ML to facilitate regional simulations, and to evaluate the implication of this approach focusing on crop yields and SOC as impacted by management practices.

2. Materials and methods

2.1. Study region and data sources

The selected study region is the HHHp of China (figure 1). This region constitutes around 17% of China's total cultivated area, and produces 35%–40% and 60%–80% of China's wheat and summer maize, respectively (Kong et al 2014). Winter-wheat summer--maize rotation is the most common rotation system in the region. We obtained soil data layers at the resolution of 10 km from the China Soil Scientific Database (Yu et al 2007, Shangguan et al 2013), which was generated based on more than 7000 soil profile measurements in the late 1970s/early 1980s across China. Major soil properties include SOC, nitrogen, bulk density, clay fraction, pH, saturated water content, drained upper limit, lower limit of crop water extraction in different soil layers down to 2.5 m, which are required to initialize and parameterize the APSIM model.

Daily climatic variables from 1980 to 2010 at around 670 meteorological stations across China were firstly sourced from the China Meteorological Administration. Then, we generated the gridded climatic data in each 10 km grid cell across the HHHp by applying an interpolation algorithm (Thornton et al 1997). Figures S1(a) and (b) show the average annual precipitation (which ranges from 508 to 1155 mm) and temperature (which ranges from 8.34 °C to 16.4 °C) during the period 1980–2010 in each grid cell across the whole study region, respectively.

In this study, we conducted a 101 year simulation from 1980 to 2080. The climate data from 1980 to 2010 (the period of climate record) were used for the first 31-year simulations. For climate data in the
following 70 years (i.e. 2011–2080), we used 28 Global Circulation Models (GCMs) from Coupled Model Intercomparison Project (CMIP5) dataset (Liu and Zuo 2012). In practice, GCM projected climate data is difficult to be directly used to conduct high resolution simulations as GCM climate products have to be downscaled to the scale of interests due to coarse spatial resolution of the GCM data (Ham et al 2018, Jiang et al 2018, Zhang et al 2019). We used a commonly used bias-correction approach to generate climate data during the period 2011–2080, i.e. linearly shift observed temperature and precipitation with the rate projected by the GCM models, following the approach of Liu et al (2017) and Liu and Zuo (2012). Although this approach is less effective of projecting climate extremes (because of no temporal downscaling processes) (Hawkins et al 2013, Lange 2019), it is based on real observed data. This approach has also been widely used for generating daily climate inputs for the application of APSIM model under future climate change scenarios (Liu et al 2014, 2017, Wang et al 2016, 2020, Xiao et al 2020). The additive shift in temperature and multiplicative shift in precipitation from 2011–2080 were shown in table S1. Here, we also would like to note that the temperature and precipitation shifts were predicted under the emission scenario of Representative Concentration Pathways 4.5 (RCP4.5). RCP4.5 predicts that annual greenhouse gas emissions reach to a peak in around 2040 and then declines throughout the latter 21st century, representing a stabilization scenario (Jones et al 2013). The APSIM model also simulates the effect of CO$_2$ concentration on plant growth (Vanuytrecht and Thorburn 2017). Future CO$_2$ concentration was estimated using an empirical equation to fit the CO$_2$ concentration of RCP4.5 from CMIP5 (Liu et al 2017).

In addition to the regional data, data from three long-term experimental sites (figure 1 and table 1) as a part of the National Long-term Fertilization Experimental Network in China (Xu et al 2015, Bai et al 2018) were collected to validate the performance of the APSIM model on simulating crop yield and SOC changes under various
management practices (see section 2.2). These three sites (i.e. Fengqiu, Xuzhou, Yucheng) were located in the HHHp (figure 1). The double cropping system of winter-wheat and summer-maize rotation was adopted at the three sites. The long-term experiment at Yucheng began in 1986 with three N fertilizer treatments (0 kg N ha\(^{-1}\) yr\(^{-1}\), 400 kg N ha\(^{-1}\) yr\(^{-1}\), 800 kg N ha\(^{-1}\) yr\(^{-1}\)). The field experiment at Fengqiu was firstly carried out in 1989 with three fertilizer treatments (0 kg N ha\(^{-1}\) yr\(^{-1}\), 300 kg N ha\(^{-1}\) yr\(^{-1}\), 150 kg N ha\(^{-1}\) yr\(^{-1}\) plus 4500 kg manure ha\(^{-1}\) yr\(^{-1}\)). The long-term experiment at Xuzhou was carried out in 1980 with three fertilizer treatments (0 kg N ha\(^{-1}\) yr\(^{-1}\), 300 kg N ha\(^{-1}\) yr\(^{-1}\), 800 kg N ha\(^{-1}\) yr\(^{-1}\) plus 1.85–3.75 Mg manure ha\(^{-1}\) yr\(^{-1}\)). Across the experimental sites, the multi-year average crop yields varied from 0.32 to 8.4 Mg ha\(^{-1}\) for wheat and from 0.23 to 11.02 Mg ha\(^{-1}\) for maize under variable fertilizer treatments. The initial SOC stock varied from 20 to 11.02 Mg ha\(^{-1}\) (0–30 cm soil profile). Table 1 summarizes the climate and site information at each study site. Detailed information on the experimental design and data can be found in Xie et al (2009), Jiang et al (2014), Xu et al (2015), Zhang et al (2016) and Bai et al (2018).

2.2. The APSIM model
APSIM is a process-based biophysical model simulating crop growth and soil processes in response to climatic and edaphic status and agricultural management practices on a daily time-step (Holzworth et al 2014). The model has the capability to well simulate crop growth, soil water, carbon and nutrient dynamics under a wide range of environmental and management conditions across spatiotemporal scales including in China. A full list of publications testing and verifying the APSIM model can be found at www.apsim.info/download-apsim/downloads (accessed on 7 July 2022). This history of model verification gives a sound basis for the use of the model in this study. Here we used the APSIM model's capability to systematically simulate biophysical processes that control crop production and a selected environmental footprint [i.e. SOC changes in this study]. The simulation outputs were used to derive ML models (see sections below).

The APSIM model has been comprehensively calibrated for simulating wheat and maize in the study area (e.g. Zhao et al 2014, 2015, 2017, Wang et al 2018). In this study, we have adopted the relevant crop cultivars and SOC parameters calibrated by those studies. Specifically, crop cultivar parameters for wheat and maize were adopted from Zhao et al (2017), parameters relating to SOC dynamics were adopted from Wang et al (2018). In addition, to explicitly test the performance of the APSIM model in simulating crop yields and SOC dynamics, we have collected long-term field experiment data at three sites (Fengqiu, Xuzhou and Yucheng) to further validate the APSIM model (figures 2 and 3).

2.3. Scenario simulations
In this study, we used the APSIM version 7.8 to conduct simulations under different scenarios (figure 2). In theory, there are limitless combinations of management practices, and we intended to cover all potential possibilities. Thus, we generated a scenario combination taking into account N management (from 0 to 300 kg N ha\(^{-1}\) yr\(^{-1}\) with the increment of 1 kg), irrigation [no, one, two, three, and four times (sowing, floral initiation, flowering, start of grain filling, respectively) of irrigation (irrigated to field capacity in the top 50 cm soil), and auto-irrigation (i.e. irrigating once soil water content drops to drained lower limit)], crop residue management (0%, 10%, …, 100% removal), manure application (from 0 to 2000 kg farmyard manure per year with the increment of 10 kg). Considering that we had 28 climate change scenarios and two crops, at last, there were a total of \(1.1 \times 10^{15}\) scenario combinations. To complete these simulations for each of the 3409 grid cells across the HHHp, millions of years were required even using parallel computing of 2000 modern computer cores. It was impossible to run all scenarios. To relax this challenge of computational cost, we randomly selected 2000 scenarios from the \(1.1 \times 10^{15}\) scenario combinations to run the model continuously from 1980 to 2080 in each grid cell. Crop yield and SOC stock in the top 30 cm soil were output each year.

2.4. Reproducing APSIM model outputs using ML at grid cell scale
Using the APSIM outputs under the 2000 randomly selected scenarios, we trained ML models to predict APSIM model outputs (figure 2). The ML models have the advantage of not requiring detailed information to initialize the process-based APSIM model and saving computing time. As examples to demonstrate the ML approach, we focused on the simulated SOC change (\(\Delta SOC\), the difference between SOC at the start and end of the 101 year simulations, which is a key indicator of soil health and has important implications for carbon sequestration in agricultural soils to mitigate climate change) and average yield (which is the central element of interest of agricultural production) in each grid cell. We assumed that both \(\Delta SOC\) and yield were associated with the specific management in each scenario, including N fertilizer amount for each group (N), residue retention fraction (R), irrigation times (I), farmyard manure input (F), and climate projection models (GCM). At last, at each grid, we obtained a database of \(\Delta SOC\) as well as yield with its covariates of N, R, I, F, and GCM based on the APSIM simulation results from the 2000 scenarios.

Using the \(\Delta SOC\) and yield database, we trained three ML-based statistical models, including
Figure 2. Framework for modelling procedures used in this study.

Figure 3. Observed winter wheat yield (A), summer maize yield (B), and SOC (C) compared with APSIM-modelled values, respectively. Observed data are from three sites, i.e. Fengqiu (FQ), Xuzhou (XZ), and Yucheng (YC), in the study region.

Multivariate adaptive regression splines (MARS), random forest (RF), and boosted regression trees (BRT) to predict ΔSOC and yield. MARS is an advanced algorithm designed for multivariate non-linear regression problems and provides a convenient approach to capture the nonlinearity aspect of polynomial regression by assessing cut-points (knots) similar to step functions. BRT and RF are two popular tree-based ensemble learning approaches by combining multiple simple trees (weak learners) to produce better outcomes and decrease generalization errors (Yang et al. 2016). The RF model is a bagging-based model that grows multiple decision trees which are merged together for a more accurate prediction (Hastie et al. 2009). The BRT model is an additive regression model which combines the strengths of regression trees and boosting (a numerical optimization technique for minimizing the loss function by adding a new tree that best reduces the loss function) (Elith et al. 2008). These three widely used ML have advantages of handling non-linear relationships (Yang et al. 2016, Garosi et al. 2019). The three models also allow us to assess potential uncertainties induced by different algorithms.
Fundamentally, the three ML models for predicting APSIM-predicted variables \( y \), such as \( \Delta \text{SOC} \) and yield) in each grid can be written as:

\[
y = f(N, R, F, GCM)
\]  

(1)

where \( N, R, F \) are the fertilizer amount (kg ha\(^{-1}\)), residue retention fraction (0–1), farmyard manure input (kg ha\(^{-1}\)), respectively; \( It \) represents the frequency of irrigation (no irrigation, one, two, three, four times of irrigation and auto irrigation) and was treated as a factor variable with six levels in the ML model; GCM was the 28 GCMs indicating future climate change predicted by different GCMs. A grid cell was randomly selected as an example to demonstrate the ML performance at the grid cell scale (figure 1).

The three ML models were evaluated using a common 10-fold cross-validation strategy using the \textit{caret} package (Kuhn, 2008 in R 3.3.1 (R Core Team 2016). Specifically, BRT, MARS and RF were trained using the algorithms implemented in the \textit{R} packages \textit{gbm}, \textit{ranger} and \textit{earth} respectively. Although ML models are generally robust to overfitting, to further confirm that the model is not over-fitted, we randomly selected 1800 data points to train the model, while the left 200 data points were used to validate the model. Then, the coefficient of determination of model training and validation were compared and used to indicate the model predictive power. After this procedure, we obtained three ML models for \( \Delta \text{SOC} \) and yield predictions, respectively, for each 10 km grid.

2.5. Reproducing APSIM model outputs using ML at regional scale

Besides specific ML models for each grid cell, we also developed ML models which can be applied across the whole study region (i.e. regional ML models). These regional models took into account some additional primary variables representing soil and climate variability across the region (equation (1)). Due to the spatial variability of climatic and edaphic conditions, some primary variables relating to soil and climate variability across the region have to be taken into account when deriving regional grid-specific ML models. Specifically, we considered three regional covariates: initial SOC stock at the start of the simulation \( (C_0) \), mean annual precipitation (MAP) and temperature (MAT), which are the primary controls underpinning the prediction of SOC changes and crop yields. As we focused on SOC changes and average yields during the whole simulation period, for MAP and MAT, we used the value during the period 1980–2010 rather than during the simulation period 1980–2080. As such, by revising the generic ML model for grid-specific prediction (i.e. equation (1)), the generic regional ML model can be written as:

\[
y = f(C_0, \text{MAT}, \text{MAP}, N, R, F, GCM). 
\]  

(2)

To train this model, we randomly selected 2000 grid-scenario combinations. The distribution of selected grids and the management scenarios were presented in figure S2. All model fitting followed the same procedure to that for fitting grid-specific ML models.

For both grid-specific and regional ML models, all the three ML models allow estimation of the relative importance of each individual variables, i.e. the contribution of variables in the model. We derived the average relative importance of each individual variables by weighting the relative importance predicted by the three ML models, with weights estimated based on the performance of each model (i.e. the determinant coefficient \( R^2 \)). This assessment enables us not only to identify the most important factor/management influencing \( \Delta \text{SOC} \) and yield but only to judge whether the importance identified by the ML models correctly reflects the processes driving APSIM predictions.

2.6. The application of the derived ML models

In croplands, a central element of interests is the \( N \) management and its relationship with crop yield. As an example to showcase the usage of the derived ML models, we used the regional ML models to predict wheat and maize yield across the HHHp. First, we randomly chose 100 000 scenarios from the scenario combinations described above. Then, we run the regional ML models under these scenarios in each grid cell across the HHHp (i.e. 100 000 predictions of wheat and maize yield under different management practices specified in the scenarios in each grid cell). Next, we targeted the simulated maximum wheat and maize yield, and identified scenarios under which both wheat and maize yield reached to >90% of the simulated maximum. Last, we screened those identified scenarios to find the lowest \( N \) input values (the sum of \( N \) input in wheat and maize growing seasons) and the separate \( N \) input for each crop. In essence, the identified \( N \) input represents the optimal \( N \) requirement under otherwise best management in order to at least reach the 90% of yield potential. During the prediction, we derived ensemble estimates (i.e. yield) as the weighted average between the three ML models, with weights estimated based on model predictive power (i.e. \( R^2 \)). Maps were generated to show the spatial variation of yield and \( N \) requirement for both crops.

3. Results

Model validation indicated that the APSIM model can explain 83% of the variance of winter wheat yield, and 78% of summer maize yield across the three sites (figure 3). The dynamics of SOC could be also well captured by the model (figure 3). Figure 4 shows an example demonstrating the performance of ML models in reproducing APSIM-predicted \( \Delta \text{SOC} \), wheat
Figure 4. An example demonstrating the performance of three machine learning models in predicting APSIM-predicted SOC change (up panels), wheat (middle panels) and maize yield (bottom panels) in a specific grid cell.

and maize yield in a specific grid cell (i.e. grid-ML). The results indicated that three ML models (MARS—figure 4(A), RF—figure 4(B), and stochastic gradient boosting—figure 4(C)) can all explain >96% of the variance in APSIM-predicted ΔSOC. Applying the ML modelling approach to each grid cell across the whole study region, we found that in all grid cells, the three ML models on average can explain >92%, >95% and >93% of variance in APSIM-predicted ΔSOC (figure S3(a)), wheat (figure S3(b)) and maize yield (figure S3(c)), respectively. Similar to the performance of the grid-ML, the regional ML also could explain >92% of the variance in APSIM model outputs across the whole HHHP (figure 5). This was true for APSIM-predicted ΔSOC, wheat and maize yield for all three ML models. These results demonstrate the strong predictive power of ML models across scales regardless of which variable we are interested.

The importance of variables of the ML model can be estimated by using percentage increases in MSE (mean squared error) as scores via the Out-of-Bag approach. The ML model results enabled us to assess the relative importance of individual management practices and climate change (figure 6). At the local scale (i.e. the 10 km grid cells in this study), as expected, the most important factor affecting ΔSOC was residue management (figure 6(A)). In the studied winter wheat-summer maize double cropping system, the management of maize residues alone contributed to ~60% of the variance in ΔSOC explained by three ML models. At the regional scale, management of maize and wheat residues showed similar importance for ΔSOC taking the first and second position respectively, while initial SOC (C₀) was the third most important variable (figure 6(D)). For the yield of both wheat and maize, at the local scale, nitrogen input played the leading role, followed by irrigation (figures 6(B) and (C)). At the regional scale, MAP was the third most important factor after nitrogen input and irrigation (figures 6(E) and (F)). However,
we note that wheat was more strongly affected by irrigation (>20%) than maize (<10%). Overall, the variable importance identified by the ML models correctly captured the major processes/controls over SOC dynamics and crop growth simulated by their antecedent process-based APSIM model.

Applying the regional ML models to estimate yield potential and the relevant N requirement, we found that the sum of wheat and maize yield ranges from 14.25 to 18.34 Mg ha⁻¹ yr⁻¹ (figure 7(A)). In general, the yield was increased from the north to the south, reflecting the precipitation and temperature patterns, while the most productivity area was in the central east of the HHHp in Shandong province (figure 7(A)). The relevant nitrogen requirement shown the similar spatial pattern to the potential yield, ranging from 235 to 273 kg N ha⁻¹ yr⁻¹ (figure 7(D)). The potential yield of maize ranged from 8.33 to 11.18 Mg ha⁻¹ yr⁻¹, and the central-eastern regions had the highest yield (figure 7(E)). The N requirement generally followed the same spatial pattern of the potential yield, but the amount (which ranges from 63 to 172 kg N ha⁻¹ yr⁻¹) was much lower than the requirement by wheat and also showed larger spatial variability (figure 7(D) vs figure 7(F)).

4. Discussion

We created ML-based statistical models to accurately reproduce the behavior of a complex agroecosystem model—APSIM—in simulating SOC dynamics and crop yield under different environmental conditions.
and agricultural management. The ML models were driven by a common set of easily obtainable soil and climate variables including initial SOC content, temperature and rainfall, and/or variables indicating management practices. These variables correctly captured the major forces of SOC dynamics and crop production in the APSIM model. For example, initial SOC storage determines the carbon pool size and couples with soil nutrient reservoir, and temperature and rainfall determine the actual decay rate of carbon pools and plant photosynthesis. Another set of management variables reflects the effect of management practices on carbon input into the soil, water and nitrogen inputs. These results demonstrate the credibility of ML models to mimic the outcomes of complex process-based models without sacrificing fundamental underlying processes.

Our application of the ML models focusing on $\Delta$SOC suggests that maize residue management is the most importance factor affecting $\Delta$SOC (figure 6(A)), followed by wheat residue management. The larger influence of maize residue management is reasonable as maize usually have much higher stubble amount than wheat. In addition, we found that wheat N input contributes >10% of the explained variances in $\Delta$SOC, while maize N input is much less important. This may attribute to the factor that winter wheat strongly depends on N input for growth because of much less N mineralization in winter, evidenced by the much higher N requirement for wheat than for maize (figure 7). Another intriguing finding is that GCM (which represents different temperature and precipitation changes in the future predicted by different GCMs) is the least important factor in the simulated system, highlighting that future climate change (i.e. temperature and precipitation changes considered in this study) may have limited effect on SOC change or the change can be effectively offset by a suitable management shift that enhances SOC sequestration.

The application focuses on yield potential generated spatially-explicit maps of wheat and maize yield and the relevant N fertilizer requirement. On average, annual N fertilizer requirement is 387 kg N ha$^{-1}$, with the potential yield of 17.1 Mg ha$^{-1}$ (wheat + maize, figure 4). This estimation of N requirement is consistent with current farmers’ application rate in the HHHp (Wang et al 2018). However, farmer’s yield (∼5.5 Mg ha$^{-1}$ for winter wheat, and ∼6.5 Mg ha$^{-1}$ for summer maize, http://data.stats.gov.cn) is much more.
lower than our estimation. A survey on Chinese farmers’ practices during the period from 2005 to 2014 showed that, on average, 178 kg N ha$^{-1}$ and 208 kg N ha$^{-1}$ of N fertilizer were applied to produce the yield of 5.0 Mg ha$^{-1}$ for wheat and 7.2 Mg ha$^{-1}$ for maize, respectively (Cui et al 2018, Quan et al 2021). This discrepancy in crop yield between our simulations and farmers’ practice may attribute to several reasons. First, environmental constraints such as frost and pest, which are not considered by the APSIM model, may reduce farmers’ crop yield. Second, smallholder farming is dominant in the HHHp and management adopted by farmers may be suboptimal, resulting in high-input but low-efficiency systems, while the estimation in this study is locally (i.e. in each 10 km grid cell) optimized by screening simulations under various management options. Third, the APSIM model and thus the ML models may be not accurate in simulating N response of crop yields across the whole region, albeit model validation at three sites in the study region demonstrated that the APSIM model could generally capture yield and SOC dynamics. We acknowledge that additional data will be particularly important and useful for the calibration and verification of antecedent process-based models in order to provide robust regional predictions.

In the HHHp, a series of studies had been conducted using the APSIM model and demonstrated the suitability of APSIM in simulating yield (Liu et al 2012, Wang et al 2018, Zhao et al 2018). After all, the results suggested that yield gaps (i.e. the difference between the potential yield and actual farmer’s yield) probably exist in the region, highlighting the potential of increasing yield or decreasing N fertilizer input. Indeed, a recent study demonstrated that both maize and wheat yield in the region can be increased by $\sim$1 Mg ha$^{-1}$ with less N fertilizer application than farmers’ application rates, if adopting enhanced management practices such as 195 kg N ha$^{-1}$ for

Figure 7. The spatial pattern of yield potential (left panels) and the required nitrogen input (right panels) across the Huang-Huai-Hai plain of China. Top panels: the sum of wheat and maize; middle panels: wheat; bottom: maize.
maize and 190 kg N ha$^{-1}$ for wheat (Cui et al 2018). Improving management taking into account the balance between water and nitrogen supply has been suggested having the potential to reduce N input without compromising yield (Michalczyk et al 2014). With optimum N management, Ju and Zhang (2017) indeed found that N input can be reduced by 54% for wheat and 26% for maize in the North China Plain. Compared with the USA, Ju and Gu (2014) found that, to produce the similar amount of crop yield, 181 kg ha$^{-1}$ more N fertilizers were used in China. In Rothamsted Broadbalk Wheat Experiment in UK, the application rate of N fertilizers was at a stable rate of 144 kg N ha$^{-1}$ from 1969 to 2014, and the yield showed a slightly increasing trend of average yield (Quan et al 2021). Overall, other management factors (e.g. irrigation, tillage and residue retention) should be coordinated with nitrogen fertilizer management. Our digital mapping of the crop yield potential and the corresponding N fertilizer input provides a reference to identify yield gap and develop management practices that can potentially close this gap or can produce similar crop yield with less N fertilizer application at both the local and regional scales in the HHHp. However, it should be noted that this study focused on the average crop yield during the simulation period, and climate extremes have also not been considered in the projection of future climate. Annual crop yield and the required N management would vary year by year, requiring additional work on how to efficiently predict such interannual variabilities. This is particularly important in the context of directional climate change with more frequent and intense climate extremes.

Our results demonstrate that using ML as a partner of complex biophysical models enables us to cost-efficiently (in terms of both computational cost and data requirement) use process-based models across large scales at high resolution. The ML models require much less input information than its antecedent process-based model. For example, the required inputs for the ML models to predict SOC change were reduced to three, while the APSIM model needs tens of different inputs to initialize and parameterize the model (Luo et al 2013). With high-resolution data of those easily available variables driving the ML models, the ML models can be readily applied to estimate variables of interest (e.g. SOC and yield in this study) in the domain that the model is developed, avoiding the potential uncertainties because of limited data availability for running process-based models directly. In our study, using our approach, the time required to generate the regional map of yield and N requirement under the enormous combinations of management options considered was shortened to several minutes on a desktop, otherwise, it requires years if running the APSIM model at each 10 km grid cell for all possible scenarios using our computing resources. Although we used the total SOC change and average yield during the whole 101 year simulation period to example and demonstrate the approach, the ML modelling approach can be easily expanded to reproduce time-series outputs by introducing time into the model or other variables of interest.

5. Conclusions

We used a hybrid approach coupling APSIM with ML to spatially-explicitly predict productivity and SOC changes in HHHp of China under all possible agricultural management. The results showed that the ML-based models can accurately reproduce APSIM predictions across the region, explain >90% of the spatial variance of APSIM predictions in terms of both yield and SOC changes, capture main processes driving APSIM predictions, and require much fewer input data. The potential yield and relevant N requirement were identified across the region and showed clear spatial pattern primarily controlled by climate and soil properties, with the sum of wheat and maize yield ranging from 14.25 to 18.34 Mg ha$^{-1}$ yr$^{-1}$ (which is much higher than yield obtained by farmers) and the optimal nitrogen requirement ranging from 329 to 408 kg N ha$^{-1}$ yr$^{-1}$ (which is generally comparable to farmers’ input). Interestingly, lower yield regions required higher nitrogen input. In conclusion, coupling process-based biophysical models with ML provides a robust approach facilitating large-scale simulations, and amplifies the usage of those models for guiding agricultural management strategies and policy-making to ensure food security as well as address global environmental challenges.

Data availability statement

The data that support the findings of this study are available upon reasonable request from the authors.

Acknowledgments

This study is financially supported by National Key Research Program of Ministry of Science and Technology of China (2021YFE0114500), National Natural Science Foundation of China (42001105), the Postdoctoral Science Foundation of China (2020M681872), and State Key Laboratory of Atmospheric Boundary Layer Physics and Atmospheric Chemistry (LAPC-KF-2022-10).

Conflict of 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.
References

Alissar C et al 2012 Climate change impacts on tree ranges: model intercomparison facilitates understanding and quantification of uncertainty Ecol. Lett. 15 533–44
Altieri M A 2018 Agroecology: The Science of Sustainable Agriculture (Boca Raton, FL: CRC Press)
Bai Z et al 2018 Effects of agricultural management practices on soil quality: A review of long-term experiments for Europe and China Agric. Ecosystems. Environ. 265 1–7
Barriere J, Havlík P, Jacquet F, Lherm M, Millet G and Bretagnolle V 2011 Farming system modelling for agri-environmental policy design: the case of a spatially non-aggregated allocation of conservation measures Ecol. Econ. 70 891–9
Bonan G B and Doney S C 2018 Climate, ecosystems, and planetary futures: the challenge to predict life in Earth system models Science 359 eaam8328
Bopp L et al 2013 Multiple stressors of ocean ecosystems in the 21st century: projections with CMIP5 models Biogeosciences 10 6223–42
Bradford M A, Wiener W R, Bonan G B, Fierer N, Raymond P A and Crowther T W 2016 Managing uncertainty in soil carbon feedbacks to climate change Nat. Clim. Change 6 751
Bryan B A 2013 High-performance computing tools for the integrated assessment and modelling of social–ecological systems Environ. Modelling Softw. 39 295–303
Cai Z et al 2018 Pursuing sustainable productivity with millions of smallholder farmers Nature 555 363
Dawson C and Gerritsen M G 2016 Computational Challenges in the Geosciences (Berlin: Springer)
Elith J, Leathwick J R and Hastie T 2008 A working guide to boosted regression trees J. Anim. Ecol. 77 802–13
Feng P, Wang B, Liu D L, Waters C and Yu Q 2019 Incorporating machine learning with biophysical model can improve the evaluation of climate extremes impacts on wheat yield in south-eastern Australia Agric. For. Meteorol. 275 100–13
Folberth C, Baklanov A, Balković J, Skalský R, Khabarow N and Obersteiner M 2019 Spatio-temporal downscaling of gridded crop model yield estimates based on machine learning Agric. For. Meteorol. 264 1–15
Friedlingstein P, Meinshausen M, Arora V K, Bopp L,Complete text here.
Quan Z, Zhang X, Fang Y and Davidson E A 2021 Different quantification approaches for nitrogen use efficiency lead to divergent estimates with varying advantages Nat. Food. 2 241–5
R Core Team 2016 R: a language and environment for statistical computing, version 3.3.1 (Vienna: R Foundation for Statistical Computing)
Reid W V et al 2010 Earth system science for global sustainability: grand challenges Science 330 916–7
Shangguan W et al 2013 A China data set of soil properties for land surface modeling J. Adv. Modelling Earth Syst. 5 212–24
Thornton P E, Running S W and White M A 1997 Generating surfaces of daily meteorological variables over large regions of complex terrain J. Hydrol. 190 214–51
Tilman D, Cassman K G, Matson P A, Naylor R and Polasky S 2002 Agricultural sustainability and intensive production practices Nature 418 671
Vanuytrecht E and Thorburn P J 2017 Responses to atmospheric CO2 concentrations in crop simulation models: a review of current simple and semicomplex representations and options for model development Glob. Change Biol. 23 1806–20
Wang B et al 2020 Sources of uncertainty for wheat yield projections under future climate are site-specific Nat. Food. 1 720–8
Wang B, Liu D L, Macadam I, Alexander L V, Abramowitz G and Yu Q 2016 Multi-model ensemble projections of future extreme temperature change using a statistical downscaling method in south eastern Australia Clim. Change 138 85–98
Wang G, Luo Z, Wang E and Zhang W 2018 Reducing greenhouse gas emissions while maintaining yield in the croplands of Huang-Huai-Hai Plain, China J. Agric. Meteorol. 260–261 80–94
Weng E and Luo Y 2011 Relative information contributions of model vs. data to short- and long-term forecasts of forest carbon dynamics Ecol. Appl. 21 1490–505
Xiao D, Liu D L, Wang B, Feng P, Bai H and Tang J 2020 Climate change impact on yields and water use of wheat and maize in the North China Plain under future climate change scenarios Agric. Water Manage. 238 106238
Xie W, Zhou J, Wang H, Chen X, Lu Z, Yu J and Chen X 2009 Short-term effects of copper, cadmium and cypermethrin on dehydrogenase activity and microbial functional diversity in soils after long-term mineral or organic fertilization Agric. Ecosyst. Environ. 129 450–6
Xiong W, Asseng S, Hoogenboom G, Hernandez-Ochoa I, Robertson R, Sonder K, Pequeno D, Reynolds M and Gerard B 2020 Different uncertainty distribution between high and low latitudes in modelling warming impacts on wheat Nat. Food 1 63–69
Xu M G, Zhang W J and Shuang S W 2015 China Soil Fertility Evolution 2nd edn (Beijing: China Agriculture and Science Technology Publishing House)
Yang B-M, Zhang G-L, Liu F, Lu Y-Y, Yang F, Yang F, Yang M, Zhao Y-G and Li D-C 2016 Comparison of boosted regression tree and random forest models for mapping soil organic carbon concentration in an alpine ecosystem Ecol. Indic. 60 870–8
Yu D S, Shi X Z, Wang H J, Sun W X, Chen J M, Liu Q H and Zhao Y C 2007 Regional patterns of soil organic carbon stocks in China J. Environ. Manage. 85 680–9
Zhang Q, Shen Z, Xu C-Y, Sun P, Hu P and He C 2019 A new statistical downscaling approach for global evaluation of the CMIP5 precipitation outputs: model development and application Sci. Total Environ. 690 1048–67
Zheng X, Sun N, Wu L, Xu M, Bingham I J and Li Z 2016 Effects of enhancing soil organic carbon sequestration in the topsoil by fertilization on crop productivity and stability: Evidence from long-term experiments with wheat-maize cropping systems in China Sci. Total Environ. 562 247–59
Zhao G, Bryan B A, King D, Luo Z, Wang E, Bende-Michl U, Song X and Yu Q 2013 Large-scale, high-resolution agricultural systems modeling using a hybrid approach combining grid computing and parallel processing Environ. Modelling Softw. 41 231–8
Zhao Z, Qin X, Wang E, Carberry P, Zhang Y, Zhou S, Zhang X, Hu C and Wang Z 2015 Modelling to increase the eco-efficiency of a wheat–maize double cropping system Agric. Ecosyst. Environ. 210 36–46
Zhao Z, Qin X, Wang Z and Wang E 2018 Performance of different cropping systems across precipitation gradient in North China Plain J. Agric. Meteorol. 259 162–72
Zhou Z, Qin X, Zang H, Chen C, Zhang Y and Wang Z 2017 Value of groundwater used for producing extra grain in North China Plain Field Crops Res. 210 47–51
Zhao Z, Wang E, Xue L, Wu Y, Zang H, Qin X, Zang J and Wang Z 2014 Accuracy of root modelling and its impact on simulated wheat yield and carbon cycling in soil Field Crops Res. 165 99–110