DeepOryza: A knowledge-guided machine learning model for rice growth simulation

Jingye Han1,2, Liangsheng Shi1,*, Christos Pylianidis2, Qi Yang1, Ioannis N. Athanasiadis2
1State Key Laboratory of Water Resources and Hydropower Engineering Sciences, Wuhan University
2Wageningen University and Research, Wageningen, the Netherlands

*Corresponding author, E-mail: liangshs@whu.edu.cn

Abstract
Even though process-based crop models are widely used to simulate crop growth, the challenge of parameter calibration makes it difficult to use them in practice. Also, it is labor-consuming to improve the model by adjusting the model description. To address these issues, we propose a knowledge-guided machine learning model (DeepOryza) to directly learn the crop growth pattern from data. A synthetic dataset generated by a process-based model (ORYZA2000) was used to pre-train the DeepOryza. An observation dataset was used to finetune and evaluate the DeepOryza. The preliminary results showed that DeepOryza can perform equally or better than the well-calibrated ORYZA2000. To investigate the effect of the proposed knowledge-guided structure, we designed two DeepOryza models with different structures. Results showed that the knowledge-guided structure can improve the performance of DeepOryza when the synthetic dataset was generated by the uncalibrated ORYZA2000. This finding indicates that the knowledge-guided structure could potentially reduce the calibration requirement of the process-based model.

Introduction
In agriculture, crop models with the ability to describe interactions between crops and the environment, are widely used as physics-based methods to simulate crop growth and optimize crop production (Ewert et al., 2015; Keating et al., 2003). Studies over the past decades have developed crop models such as WOFOST (van Diepen et al., 1989), SWAP (van Dam, J. C., Huygen, J., Wesseling, J. G., Feddes, R. A., Kabat, P., van Walsum and P. E. V., Groenendijk, P., & van Diepen, 1997), and ORYZA (Bouman et al., 2001) based on different physical mechanisms. Despite their extensive use, these crop models have several limitations due to simplified representations of the physical processes (Feng et al., 2019; Karpatne et al., 2017) or challenges in parameter calibration (Seidel et al., 2018; Wallach et al., 2021). The limitations of physics-based models cut across discipline boundaries and are well-known in the scientific community (Lall, 2014).

To address these issues, various methods have focused on improving the accuracy of parameter estimation as well as reducing the model structure error. For parameter estimation, a model-independent software (e.g., PEST) (Doherty, 1994) or package (e.g., SOPTPY) (Houska et al., 2015) could be employed to calibrate models, but applications are limited by the difficulty of coupling the model and the software/package (Christian et al., 2020). As a result, most researchers used trial-and-error procedures to search for the best-fit parameters (Seidel et al., 2018) even though many other algorithms (least squares, generalized likelihood uncertainty estimation, Bayesian parameter estimation) could be used. However, the optimal parameters could change as the evaluation criteria change, which means that there will never be an optimal set of parameters that allow all simulated variables to be optimal at the same time. Regarding model structure error, it always arises from the imperfect description of crop growth mechanism. Thus, the only way to decrease the structure error of a process-based model is to adjust the model description, such as the model function, the model structure or involve additional inputs into the model. However, this procedure is labor-consuming and may increase the model complexity, making it more difficult to calibrate (Yin et al., 2021).

Data-driven methods, especially machine learning (ML) methods, have shown superior performance in various disciplines for their feature extraction capabilities and capturing nonlinear dynamics (Kashinath et al., 2021). For agricultural models, the application of purely data-driven methods is restricted by the lack of data, and also from the learned dynamics which may violate the laws of crop growth and physics. In recent years, research is being directed towards combining process-based crop models and data-driven methods (Everingham et al., 2016; Feng et al., 2020; Guzmán et al., 2018; Pagani et al., 2017; Shahhosseini et al., 2021). Prior studies consider crop models as tools for feature
engineering to discover valuable information from raw data, but they do not truly integrate the two modeling approaches.

To fundamentally integrate scientific knowledge from process-based models and ML models, a new paradigm termed knowledge-guided machine learning (KGML) was proposed. There have been attempts to apply this new paradigm to several disciplines by transforming physical knowledge into different forms for integration into machine learning models, such as logic rules, constraints, simulated datasets, etc. A detailed review of these forms can be found in (Karpatne et al., 2017; Von Rueden et al., 2019; Willard et al., 2020). A study conducted by Read et al. (2019) used a neural network pre-trained on a synthetic dataset generated by a physical model to accurately predict lake water temperature for various conditions. Another research developed a model to estimate \( \text{N}_2\text{O} \) emissions using a knowledge-guided model structure that was designed according to the relationship of variables defined in equations of process-based models (Liu et al., 2021). However, to the best of our knowledge, there is no research exploring the potential of the KGML for crop growth simulation, which involves the interactions between crops and environment. As agricultural data acquisition becomes easier, it is necessary to build a crop growth model that can learn from both real observed data and knowledge embedded in crop model.

In this paper, we present a knowledge-guided machine learning model for crop growth. A baseline model (DeepOryza_rizya_baseline) and a knowledge-guided machine learning model (DeepOryza_kgml) are proposed and compared with a process-based crop growth model (ORYZA2000). DeepOryza was developed as an LSTM for multivariate time series forecasting of seven crop states. The baseline model does not consider any physical relationship between the outputs. In contrast, the knowledge-guided machine learning DeepOryza enforces known relationships between them. DeepOryza is demonstrated for a case study in China with field observations from two years. The DeepOryza model was trained with synthetic data generated using ORYZA2000.

The aim of this study is to investigate: 1) how a ML model performs against a well-calibrated process-based crop growth model; 2) the effect of knowledge-guided structures in ML models; 3) the necessity of calibration for the process-based model before generating a synthetic dataset.

### Methodology and data

To develop DeepOryza and evaluate its performance, we design the following framework (Fig. 1):

1. **Synthetic dataset generation**: The ORYZA2000 model was calibrated on observed data from YEAR-A. The calibrated and uncalibrated ORYZA2000 models were then used to generate synthetic datasets, respectively.

2. **Model pretraining**: The DeepOryza model was trained on the synthetic dataset to learn basic patterns of rice growth.

3. **Model finetuning**: The pre-trained DeepOryza model was finetuned on the observation dataset from Year-A and the finetuned DeepOryza model was tested on the observation dataset from both Year-A and Year-B.

4. **Model evaluation**: The results of the uncalibrated ORYZA2000 (with default parameters), the calibrated ORYZA2000, and the finetuned DeepOryza were compared to evaluate model performance, the necessity for calibration, and the effect of knowledge-guided structures.

![Fig. 1 The framework for building and evaluating the DeepOryza model.](image)

**The synthetic dataset**

The ORYZA2000 model was selected for generating synthetic datasets due to its verification through a broad range of studies (Cao et al., 2017; Han et al., 2022, 2020; Li et al., 2013). After determining the model parameters, the date of seeding and transplanting, ORYZA2000 can simulate rice growth in daily steps for the rate of dry matter production and phenological development driven by the daily inputs (weather, irrigation, and fertilization) (Bouman et al., 2001). To reduce the complexity of the model, the irrigation and model parameters were not involved in the synthetic dataset. Also, only seven model outputs (development stage, plant area index, leaf/stem/grain biomass, above-ground biomass, final yield) which had the corresponding observations in the real datasets were selected for the synthetic dataset. Thus, the synthetic datasets consist of weather, management (fertilization, dates of seeding and transplanting), and the seven corresponding outputs of the ORYZA2000. Cligen (Nicks et al. 1995), a software that can generate random synthetic time-series weather data that match the pattern of historical weather of a specific site, was used to generate weather data. The management was randomly generated according to the schedule of local farmers. To calibrate the ORYZA2000 model, 14 parameters that are sensitive to the seven output variables were selected. The PEST (Doherty, 1994) software was then used to calibrate these parameters with the real dataset.

Finally, two synthetic datasets were created with the uncalibrated ORYZA2000 with default parameters (only adjusting the DVS-related parameters to make the growth timetable roughly consistent with the real data) and calibrated ORYZA2000. Each synthetic dataset included 250k
samples, and was further divided into five sub-datasets to pre-train five models in order to reduce the effect of random factors. The results of DeepOryza were derived by averaging the results of the models in the five sub-datasets.

### The observation dataset

The observation data were collected in a late-season rice experiment in a small region (23°5′52″–23°7′23″ N, 108°57′7″–108°58′34″ E) located in Binyang County of Guangxi in China. The meteorological data, including solar radiation, air temperature, relative humidity, and precipitation acquired from https://power.larc.nasa.gov. There were 65 plots in 2018 and 40 plots in 2019 were used in this study. The management recording was obtained by performing surveys during the growing season. The observation values of DVS (development stage), PAI (plant area index), WLV (leaf biomass), WST (stem biomass), WSO (storage organ biomass), and AGB (above-ground biomass) were obtained by destructive sampling. The final yield was obtained by combine-harvesters. The details of the dataset can be found in Han et al., (2022).

---

### The structure, training, and evaluation

The LSTM model, proposed by Hochreiter and Schmidhuber (1997), was employed to build the DeepOryza because it can learn longer dependencies between variables on long time series data. As shown in Fig. 2, the hidden states of the LSTM cell were initialized from the values of the initial crop state, which could reduce the effect of random initialization. The time series of weather and management were used as input because they are known drivers of rice growth. The DeepOryza_baseline estimates seven crop states directly without considering their relationship. The structure of DeepOryza_kgml was designed by considering the relationship between the crop states: 1) AGB should be the sum of WLV, WST, and WSO; 2) the yield comes from the partition biomass of the storage organ and the translocation biomass of the stem over the stage after the flowering stage (Laza et al., 2003). These two relationships were also described by functions in ORYZA2000. Thus, the AGB was derived from WLV, WST, and WSO; the YIELD was derived from DVS, WST, and WSO in DeepOryza_kgml.

The input size was 8 (temperature_min, temperature_max, irradiance, vapor pressure, wind speed, precipitation, fertilization amount, if transplanting). The hidden state size of the LSTM cell was 64, and the other layer setting was shown in Fig. 2. ADAM was used to optimize the model parameters. The loss function was defined as:

\[
    \text{Loss} = \sum_{\text{state}} \text{Loss}_{\text{state}} 
\]

\[
    \text{Loss}_{\text{state}} = \frac{\sum (\text{prediction}_{\text{state}} - \text{observation}_{\text{state}})^2}{N} 
\]

where state includes DVS, PAI, WLV, WST, WSO, AGB and YIELD; N is the number of observations for the corresponding state.

The model performance was evaluated by RMSE. The pretraining process was repeated 25 times (5 synthetic datasets × 5 random seeds) to make the results more robust. The RMSE was the average value of the 25 repeats.

### Results

#### The results of Calibrated-ORYZA2000, DeepOryza_baseline and DeepOryza_kgml

Since the ORYZA2000 was calibrated with the PEST software, its performance improved significantly (Case 5 and Case 6 in Table 1). As shown in Table 1, the DeepOryza models had better performance than calibrated-ORYZA2000 in Case 1 and Case 3. But the performances of the three models were similar in Case 2 and Case 4. A reason for that may be that ORYZA2000 model was well
calibrated in the dataset of 2019, making it difficult to further improve the model performance with DeepOryza. Comparing the accuracy of WLV, WST, WSO and YIELD (Case 1-4 in Table 1), there was no obvious winner between the two DeepOryza models because the synthetic dataset was generated by the calibrated ORYZA2000 model, which had embedded the prior knowledge between the variables of the real dataset. Thus, the knowledge-guided structure did not improve the model performance.

The effect of knowledge-guided structure and the necessity of calibration for the process-based model

In Table 1, the DeepOryza models in Case 5 and Case 6 were pre-trained on the synthetic dataset generated by the uncalibrated ORYZA2000. The performance of DeepOryza_baseline of Case 5 was worse than that of Case 3. This is due to the lack of restriction of the knowledge-guided structure, which made the DeepOryza_baseline easier to overfit and violate crop growth pattern when fine-tuning. The DeepOryza_kgml of Case 3 had similar accuracy to that of Case 5, which meant that a reasonable structure, although not as detailed as the process model, could also have a positive effect on model training. Furthermore, the accuracy of biomass-related variables (except the YIELD in Case 6) of DeepOryza_kgml was better than that of DeepOryza_baseline, which indicated that the knowledge-guided structure could potentially reduce the requirement of calibration.

Conclusions and Limitations

In this study, we proposed a knowledge-guided machine model named DeepOryza for rice growth simulation. The LSTM model was used as the basic model. Its initialization and structure design were guided by the knowledge of the ORYZA2000 crop growth model. This is the first attempt of simulating multivariate time-series for crop growth. The preliminary results showed that the DeepOryza could perform equally or better than a well-calibrated process-based model. The knowledge-guided structure could improve the model performance and potentially reduce the calibration requirement of the process-based model.

However, the dataset is relatively small and requires more case studies and a variety of datasets to verify the described methods. Also, fundamental principles (e.g., mass conservation) are not embedded in the structure or loss function. The naive LSTM cell makes the iterative process of crop growth a black box and weakens the interpretability. Therefore, a better structure and interpretation should also be further investigated in the future.
Acknowledgments

This study was supported by the National Natural Science Foundation of China Grant 51979200; the Open Research Fund of Guangxi Key Laboratory of Water Engineering Materials and Structures, Guangxi Institute of Water Resources Research, under grant NO: GXHRI-WEMS-2020-06 and the China Scholarship Council for 1.5 years’ study at the Wageningen university & research.

References


