Review: Improving wheat phenology and yield forecasting with a deep learning-enhanced WOFOST model under extreme weather conditions

The paper presents WOFOST-EW, an enhanced version of the WOFOST crop model that employs an LSTM to integrate extreme weather indices, aiming to improve the phenology component of WOFOST. The model is validated using data from the North China Plain, focusing on the prediction of two phenological stages (heading and maturity) and yield, particularly during extreme weather years.

While the approach of integrating machine learning into a process-based crop growth model is interesting, the implementation and evaluation raise several significant concerns, outlined below.

Use of Extreme Weather Indices in phenology prediction

- The authors propose that extreme weather events, such as excessive rainfall or temperature anomalies, can improve the prediction of phenological stages. However, it is commonly understood that phenological development in cereal crops is primarily driven by temperature (thermal time), photoperiod, and vernalization. To strengthen their argument, the authors should provide a clear rationale explaining why and how extreme weather events would directly affect phenological stages.
- It appears that these extreme weather indicators are not incorporated into other model components (e.g., growth, biomass accumulation, or stress responses), but are instead used solely to modulate phenology. This approach attributes all extreme weather effects to phenology alone, which risks conflating physiological processes and may compromise the model's interpretability.

LSTM training

- The description of the LSTM network implementation lacks detail regarding the input sequence. Given that LSTM networks are designed for time series data and the extreme weather indices appear to be annual scalar values rather than time series over a growing season, it is unclear what sequence is fed into the LSTM. Is it a sequence of these annual scalars across the training period? This approach for LSTMs seems questionable, particularly when a held out test year falls within the training period and the output of the final timestep is taken as the prediction.
- It is also unclear what the training target of the LSTM is—does it aim to predict final yield, phenological stages, or something else? Furthermore, are the parameters of WOFOST kept fixed during the LSTM optimization? If the LSTM (and thus WOFOST-EW) is trained using phenological stage data while the original WOFOST model is not, then the comparison between the two models may not be fair or meaningful.
- While WOFOST is calibrated using data from the years 1990–2000, the LSTM (and therefore WOFOST-EW) appears to be trained on a broader range of years through cross-validation. This introduces an unfair advantage in the evaluation.

- It would be interesting to see a plot of only the intermediate LSTM output in the different years and if any patterns can be noticed w.r.t. the presence of extreme weather events.

Reproducibility

- The provided codebase does not include any means to reproduce the results. For instance, the “LSTM” repository appears to be merely a clone of the Keras library.
- Section 2.2.2 states that yield data is used from the Agricultural Yearbook of the respective provinces, and is supplemented with data not disclosed by survey bureaus. Where does this supplemented data come from

Calibration WOFOST

- It is unclear which parameters are selected during the WOFOST calibration process. How many? How many iterations? Which (if any) parameter value ranges were used?

Agricultural management data

- No details are provided regarding farm management. The authors should clarify whether the winter wheat was irrigated and specify the fertilization applied.

Phenological stages

- The authors evaluate two phenological stages, namely heading and maturity. In WOFOST, phenology is represented by the Development Stage (DVS), which ranges from 0 (emergence) to 2 (maturity), with 1 corresponding to flowering. What DVS value corresponds to the heading stage?

Other

- Abstract: There is no mention of why deep learning models should be expected to help improve winter wheat simulations in this setting.
- Section 2.3.2: LDD and HDD terms are used without introduction
- Section 2.3.3: LSTM is a recurrent neural network architecture that mitigates the inherent weakness in rnns in dealing with long temporal dependencies. It is not inherently stable and high performing in long term prediction tasks.
- Section: 2.3.4: What are the parameter values used for t_base and t_max in the WOFOST temperature response functions? (F(T) in the paper). Discussion on the behavior and possible limitations could be beneficial.
- Tables S3 and S4: Would it be possible to include error bars/confidence intervals?
- The separate evaluation in years with extreme weather events is very interesting. Maybe including a statistical significance test could help in giving insight to whether two years of data is sufficient to claim an improvement.