https://doi.org/10.5194/essd-2022-297
Preprint. Discussion started: 2 September 2022

Open Access Earth System Science Data Discussions

GlobalWheatYield4km: a global wheat yield dataset at 4-km resolution during 1982-2020 based on deep learning approaches

Yuchuan Luo[1], Zhao Zhang[1], Juan Cao[1], Liangliang Zhang[1], Jing Zhang[1], Jichong Han[1], Huimin Zhuang[1], Fei Cheng[1], Jialu Xu[1], Fulu Tao[2, 3, 4]

[1]Academy of Disaster Reduction and Emergency Management Minsitry of Emergency Management & Ministry of Education, School of National Safety and Emergency Management, Beijing Normal University, Beijing 100875, China
[2]Key Laboratory of Land Surface Pattern and Simulation, Institute of Geographical Sciences and Natural Resources Research, Chinese Academy of Sciences, Beijing 100101, China
[3]College of Resources and Environment, University of Chinese Academy of Sciences, Beijing 100049, China
[4]Natural Resources Institute Finland (Luke), FI-00790, Helsinki, Finland

Correspondence to: Zhao Zhang (zhangzhao@bnu.edu.cn)

Abstract. Accurate and spatially explicit information on crop yield over large areas is paramount for ensuring global food security and guiding policy-making. However, most public datasets are coarse resolution in both space and time. Here, we used data-driven models to develop a 4-km dataset of global wheat yield (GlobalWheatYield4km) from 1982 to 2020. First, we proposed a phenology-based approach to map spatial distribution. Then we determined the optimal grid-scale yield estimation model by comparing the performance of two data-driven models (i.e., Random Forest (RF) and Long Short-Term Memory (LSTM)), with publicly available data (i.e., satellite and climatic data from the Google Earth Engine (GEE) platform, soil properties, and subnational statistics covering ~11000 political units). The results showed that GlobalWheatYield4km captured 82% of yield variations with RMSE of 619.8 kg/ha. In addition, our dataset had a higher accuracy (R^2 ~0.73) as compared with Spatial Production Allocation Model (R^2 ~ 0.49) across all regions and years. The GlobalWheatYield4km dataset will play important roles in modelling crop system and assessing climate impact over larger areas ((DOI of the referenced dataset: https://doi.org/10.6084/m9.figshare.10025006; Luo et al., 2022b).

1 Introduction

Approximately 800 million people worldwide suffered from undernourishment in 2020 (FAO, 2021). Sustainable Development Goal (SDG) 2 is dedicated to eradicating hunger and all forms of malnutrition by 2030 and achieving food security (UN, 2017). However, the goal of eliminating hunger will remain elusive even by 2050 due to climate variability or extreme weather events and global crises such as the COVID-19 pandemic and the current Russia–Ukraine war (IFPRI, 2022). Climate change is projected to force an additional 72 million people to face hunger risks in 2050, and the COVID-19 pandemic perhaps has added 83-132 million more undernourished people in 2020 (FAO, 2020; IFPRI, 2022). In these contexts, global food production needs to increase by at least 70% to feed the unprecedented population growth up to 10 billion by 2050 (Sulser et al., 2021; van Dijk et al., 2021). To better inform a series of agricultural resource allocation and food security decisions, timely and accurate information on crop yield at global scale is of paramount significance (Folberth et al., 2020; Lobell et al., 2009; Ray et al., 2015; Rötter et al., 2018).

Process-based crop models and statistical methods are the main ways to predict crop yield (Balaghi et al., 2008; Bussay et al., 2015; Feng et al., 2021; Franch et al., 2015; Jin et al., 2017; Zhao et al., 2020). Process-based crop models can dynamically simulate crop development, growth and grain formation processes (Chen et al., 2022; Huang et al., 2019; Ines et al., 2013; Luo et al., 2021; Zhuo et al., 2021). Despite utilizing a range of fundamental mechanisms of physiological processes, crop models highly require substantial data inputs and intensive computations (Burke and Lobell, 2017). In addition, statistical models often relate crop yields to diverse predictor variables (e.g., vegetation indices and climatic variables) and calibrate the empirical relationships based on measurements (Kern et al., 2018). The main advantages of statistical models are their simplicity and less dependence on calibration data; however, they are particularly vulnerable to co-linearity problems and noise of inputs (Lobell and Burke, 2010). Fortunately, machine learning (ML) approaches provide innovative alternatives to statistical models and can address the nonlinear relationships between the predictor variables and crop yield, which have demonstrated their superior performance in many applications (Cai et al., 2019; Cao et al., 2021; Li et al., 2021; Jin et al., 2018). For instance, Kang et al. (2020) compared the performances of a set of statistical and ML methods and indicated that all ML models achieved better accuracy in predicting county-level maize yield. Emerging breakthroughs in algorithms such as deep learning (DL) approaches have accomplished more accurate crop yield estimation (Jeong et al., 2022; Zhang et al., 2021). For example, the long short-term memory (LSTM) model adopts a recurrent neural network structure that can recognize sequential information for long time periods and capture sophisticated nonlinear relationships, showing superior performance over ML models in yield prediction (Jiang et al., 2019; Tian et al., 2021).

Previous studies using ML and DL methods focused on very limited areas rather than global scales. It is well recognized that a global spatially explicit crop yield dataset has important implications for large-scale agricultural system modelling and climate change impact assessments. Although there are a few efforts to fill such data gaps, they still have some limitations. For example, a global dataset, with 10-km harvested area and yield datasets, was firstly generated for 175 crops circa 2000 (Monfreda et al., 2008), followed by the Global Agro-ecological Zones (GAEZ) datasets in 2000 and 2010 (Fischer et al., 2012), Spatial Production Allocation Model (SPAM) with 5-arcmin grid for three years (2000, 2005 and 2010) (You et al., 2014) and the latest data by Grogan et al. (2022) at 5-minute resolution for 2015. However, these four public data products only cover 1~3 years, which hamper related studies on investigating the long-term impacts of climate change on yields (Tao et al., 2006, 2009). Iizumi et al. (2020) developed a global dataset of historical yields (GDHY) for major crops at a spatial resolution of 0.5° by integrating agricultural statistics and remote sensing. GDHY covers a longer period, but its spatial resolution is relatively coarse. Moreover, these yield datasets above were established based on crop distribution maps generated by downscaling method rather than accurate satellite-derived maps, which might result in misestimated yield and inaccurate assessments of climate change impacts (Luo et al., 2022a). Therefore, there is an urgent need to acquire the global gridded yield dataset with a higher resolution and a longer time span based on the accurate spatial distribution of harvesting areas.

https://doi.org/10.5194/essd-2022-297
Preprint. Discussion started: 2 September 2022

In this study, by integrating multi-source data (e.g., remote sensing, climate, soil data and census statistics) and data-driven methods, we aim to 1) propose a phenology-based method to obtain the spatial distribution of wheat across the globe; 2) compare the performance of two ML and DL models in predicting gridded yields; 3) choose the optimal models to generate global wheat yield datasets. The resultant dataset with 4-km spatial resolution will benefit to investigate spatiotemporal patterns of crop production, assess climate change impacts and modelling crop growth processes over large spatial extents (Luo et al., 2022b).

2 Data and methods

2.1 Study area

The study area contains 54 countries across the globe, covering ~92% of the total harvested areas and ~93% of the productions (FAOSTAT, 2020) (Fig. 1). These countries possess abundant subnational-level statistics, with diverse climatic conditions and cropping systems. Winter wheat dominates the majority (>75%) of the total wheat harvested areas while spring wheat covers <25% (primarily in Northern Hemisphere high latitude areas such as the United States, Russian Federation and Canada) (Ren et al., 2019; USDA, 1994).

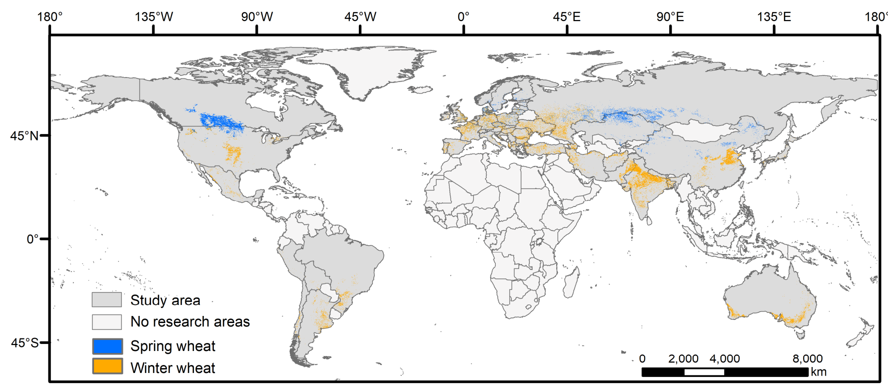

Figure 1: The spatial distribution of spring and winter wheat across the study areas covering 54 countries globally.

2.2 Data

2.2.1 Remote sensing data

We acquired the global daily 0.05° Normalized Difference Vegetation Index (NDVI) data during 1981-2021 derived from the Advanced Very High-Resolution Radiometer (AVHRR) sensor on the Google Earth Engine (GEE) platform (https://

https://doi.org/10.5194/essd-2022-297
Preprint. Discussion started: 2 September 2022

developers.google.com/earth-engine/datasets/). The main strength of AVHRR NDVI lies in its longest time coverage which can be used to derive predictors for yield prediction (Vermote et al., 2014). In addition, the 8 d composite Global Land Surface Satellite (GLASS) Leaf Area Index (LAI) at 1-km spatial resolution and Global Food Security-support Analysis Data (GFSAD) 1 km Crop Mask product (GFSAD1KCM) were used to map spatial distributions of wheat. GLASS LAI was retrieved using general regression neural networks with multiple inputs (http://glass-product.bnu.edu.cn/?pid=3&c=1), with the specific advantages of being spatiotemporally continuous without gaps and having higher accuracy than other datasets (Xiao et al., 2014, 2016). GFSAD1KCM provides global cropland extent for the nominal year 2010 and is produced based on four inputs with the highest accuracy of 85% (Teluguntla et al., 2016). Moreover, the annual dataset of 1 km wheat harvested area (named ChinaCropArea1km) in China during 2000-2015 were used (Luo et al., 2020).

2.2.2 Wheat harvested area and yield

We collected subnational-level census statistics on harvested area (unit: ha), production (unit: ton), and yield (unit: kg/ha) from ~11000 administrative units for the 54 countries, with the longest time coverage spanning from 1981-2020. Yield is calculated as production divided by harvested area. Overall, 97% of statistical data came from administrative unit level 2 (ADM2) and 3 (ADM3). The temporal coverage of crop statistics differs across the study area (Table S1). We eliminated outliers of census data with values +/− 2 standard deviation from the average.

2.2.3 Environmental Data

Meteorological information was obtained from high-spatial resolution (1/24°, ~4-km) monthly TerraClimate datasets (Abatzoglou et al., 2018). The climate variables used for this analysis were maximum temperature (T_{min}), minimum temperatures (T_{max}), precipitation (Pre), vapor pressure (Vap), vapor pressure deficit (Vpd), evapotranspiration (Pet), Soil moisture (Soil), palmer drought severity index (Pdsi), and downward surface shortwave radiation (Srad) from 1981 to 2021. In addition, soil properties were derived from Harmonized World Soil Database (HWSD) at 0.00833° (~1 km), involving bulk density, organic carbon, pH, gravel, clay, sand and silt fraction for the topsoil (0-30cm) (Nachtergaele et al., 2012).

2.3 Methods

We applied the framework, Global Wheat Production Mapping System (GWPMS), developed by Luo et al. (2022a) for mapping global wheat harvesting area and estimating gridded-yield.

2.3.1 Identifying the spatial distribution of wheat

Phenology information plays a paramount role in large-scale crop mapping (Dong et al., 2015; Luo et al., 2020; Song et al., 2017). The occurrence of phenological stages of wheat differs from other crops, especially for winter wheat. More specifically, winter wheat is often sown in autumn, reaches heading in late spring and ultimately matures in early summer. That is, the phenological dates of winter wheat are earlier than other crops (e.g., summer crops and spring wheat). In addition, the duration

https://doi.org/10.5194/essd-2022-297
Preprint. Discussion started: 2 September 2022

of growth period of winter wheat is generally longer. Spring wheat can also be differentiated from other summer crops as its phenological phases occur earlier. Therefore, we developed a wheat detecting algorithm that formalized these features in rules to automatically detect the harvested areas of spring and winter wheat (Luo et al., 2022a). In addition, we modified the algorithm when applied it in some regions where winter wheat was not a dominant crop or grown in rotation with other crops. For example, the rule for the senescence phase was loosened or even eliminated when the signal was weak due to the mixed pixel issues or the short duration of the interval between the maturity date of winter wheat and the planting date of the second crop.

First, we compared the cropland map derived from the GFSAD1KCM with statistics to determine whether to use it as a cropland mask; that is, the mask was utilized only when the GFSAD1KCM-derived areas matched with (or were larger than) statistics. Then, we combined all the available GLASS LAI images during the wheat growing season together and obtained LAI time series for each cropland pixel. The commonly used Savitzky-Golay (S-G) filter method was used to remove the noise from the data, which had shown good performance for smoothing time series (Geng et al., 2014; Savitzky and Golay, 1964; Wang et al., 2018). Finally, we applied the algorithm to extract annual spatial distribution of spring and winter wheat during 2006-2014.

2.3.2 Estimating gridded-yield using data-driven models

We first compared the predictive performance of two commonly used ML and DL approaches, i.e., Random Forest (RF) and LSTM. RF combines a set of decision trees that are constructed from a random subset of data (Breiman, 2001). Each tree is trained separately on these samples, and the remaining data are called out of bag (OOB) sample and can be used to validate the RF model. In this study, we used Python scikit-learn module to develop the RF regression model. The number of decision trees, the maximum depth of the tree and the number of features were selected for tuning. The LSTM network performs a framework of recurrent neural network (RNN) and memory gate structure, demonstrating superior performance in coping with sequential data and capturing the nonlinear and cumulative relationships between crop yield and meteorological factors (Hochreiter and Schmidhuber, 1997; Jiang et al., 2019). The model consists of an input layer, one or more LSTM layers and an output layer. The LSTM layers are composed of LSTM cells, in which information is forgotten or outputted decided by three gates. Batch normalization were firstly implemented for all the input data. The transient data (i.e., NDVI and climate data) were dealt with two LSTM layers that has 200 hidden units, whereas the non-sequential data (i.e., soil properties) were appended to the final LSTM layer and then fully connected to the output layer. In addition, a rectified linear unit (ReLU) activation function was used for all the layers. Model were run for 2000 maximum iterations with a mini-batch size of 500 and RMSprop was used to optimize hyperparameters with a learning rate of 0.001. The LSTM network for estimating gridded-yield was performed on TensorFlow (GPU version 2.0).

Here, we first resampled the gridded input data (i.e., NDVI, climate, and soil data) into 4 km and unified NDVI and climate data into monthly time steps by the maximum value synthesis and monthly mean method, individually. Then, we derived an integrated wheat map to represent reliable spatial distribution over a long-term period on the basis of the grids with cultivation

https://doi.org/10.5194/essd-2022-297
Preprint. Discussion started: 2 September 2022

for at least 5 years during 2006-2014. Finally, all input data were averaged on the subnational scale after being masked by wheat cultivated pixels. These processes were performed on the Google Earth Engine (GEE) platform.

We implemented "leave-one-year-out" method to examine the practical performance of the two ML and DL models, that is, one-year data was used for testing and the records of the remaining years for training. More specifically, each model was first trained separately by excluding one year in the data. The best hyperparameters were determined with the ten-fold cross-validated coefficient of determination (R^2). Then, the optimized models were used to estimate gridded-yield for the excluded year. Finally, the resultant yield maps were aggregated to the corresponding ADM level and were compared with census data for the excluded year. R^2 and root mean square error (RMSE) were calculated to validate the accuracy of yield estimation. The whole process was repeated 20 times and the mean R^2 and RMSE were used to compare the performance of the two data-driven models.

To improve the accuracy of the yield dataset and lengthen its time coverage, we combined the census data of some countries together to train the models and estimate gridded yields. For example, we only collected statistics of Kazakhstan for the years 2014-2020. Since the growing season of spring wheat was identical in Russian Federation and Kazakhstan, their statistics were integrated to feed into the model and the yield maps were ultimately generated from 1995 to 2020. The above treatment was repeated for all European countries, as well as Afghanistan and Iran. In addition, we applied the pre-trained model to other years where statistics are unavailable, aiming at generating a spatiotemporally continuous yield dataset.

2.3.3 Comparison with other global yield datasets

We compared our gridded yield estimates with a prevalent product (i.e., SPAM) using statistical data to demonstrate the reliability of our dataset. These two datasets could be directly compared as they were both generated using statistics. More specifically, we calculated the R^2 and RMSE between the statistical yield and the estimates of SPAM or GlobalWheatYield4km in 2000, 2005, and 2010.

3 Results

3.1 Accuracy assessment of wheat distribution maps

To illustrate the reliability of the wheat distribution maps, we validated them with the statistical area at the subnational level. The estimated areas generally matched well with the statistics, with R^2 ranging from 0.65 to 0.89 (average R^2 of 0.8) and RMSE ranging from 5.9 Kha to 355.6 Kha (Fig. 2, Table S2). The mapped areas were overestimated in Russian Federation, Kazakhstan, Australia, Canada and the United States, while they were underestimated in South America. The possible reason for the overestimation was that spring wheat was dominantly grown in these countries and was prone to be confused with other spring cereals such as spring barley. In addition, the wheat distribution maps showed the lowest accuracy in South America with R^2 ranging from 0.65 to 0.82, which was ascribed to the mixed pixels and the larger uncertainties from remote sensing

https://doi.org/10.5194/essd-2022-297
Preprint. Discussion started: 2 September 2022

products. Overall, the comparisons showed the high consistency between the resultant maps and the census data, demonstrating that the derived maps were reliable for further yield prediction.

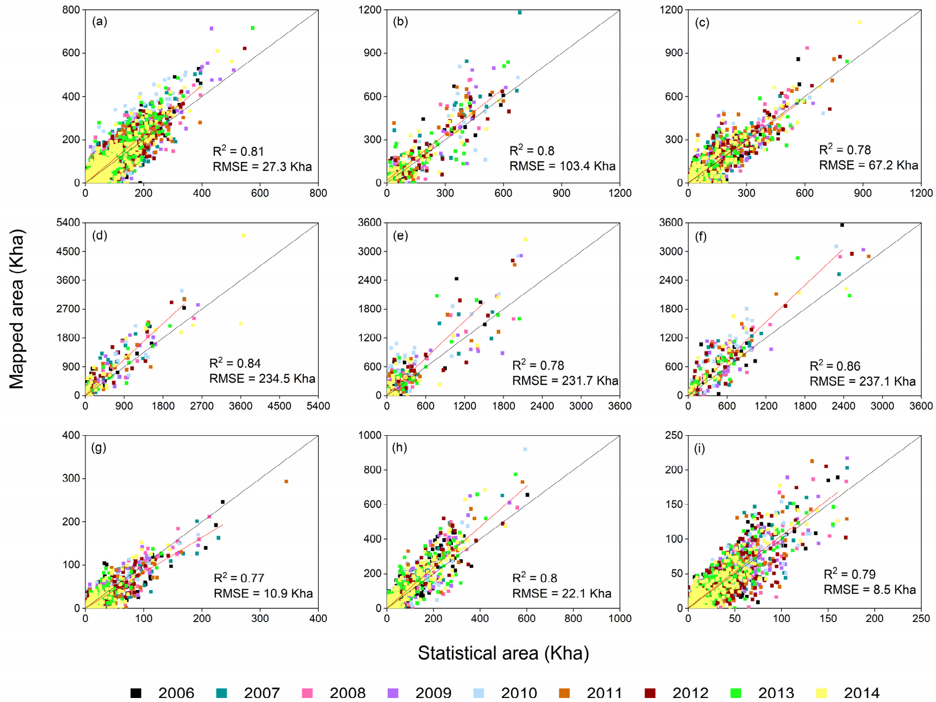

Figure 2: Comparisons between mapped area by the phenology-based method and subnational-level statistics during 2006-2014. (a) South and East Asia, (b) Central Asia, (c) Europe, (d) spring wheat in Russian Federation and Kazakhstan, (e) winter wheat in Russian Federation, (f) Australia, (g) South America, (h) spring wheat in North America, (i) winter wheat in North America.

3.2 Accuracy of GlobalWheatYield4km

The performance of RF and LSTM models in gridded yield prediction during 2006-2014 for each region/country were shown in Fig. 3. Generally, the LSTM model outperformed RF with average R^2 (nRMSE) of 0.72 (13.1) and 0.64 (16.2), respectively. More specifically, LSTM achieved the highest accuracy in the United States, Europe, China, India and Pakistan ($R^2 > 0.8$) while RF showed comparable performance (R^2 of 0.7 ~ 0.82), perhaps due to the abundant training samples. Moreover, the LSTM models were improved R^2 by around 15% as compared with RF, especially in Russian Federation, Kazakhstan, Ukraine, Iran, Bangladesh, Japan, Brazil, Peru and Bolivia with more improvements ranging 15%~48%. The superior performance of LSTM was attributed to its powerful temporal learning capabilities that can capture nonlinear and cumulative relationships between yield and meteorological factors over long time periods.

https://doi.org/10.5194/essd-2022-297
Preprint. Discussion started: 2 September 2022

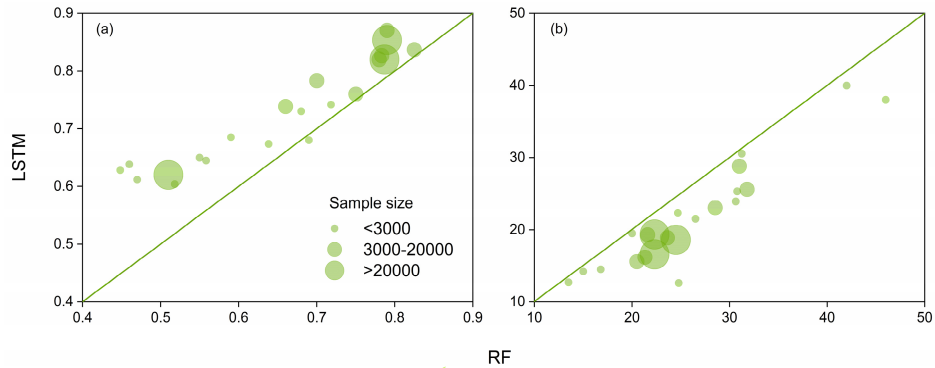

Figure 3: Performance of RF and LSTM in yield estimation during 2006-2014 across all regions: (a) R^2, (b) nRMSE.

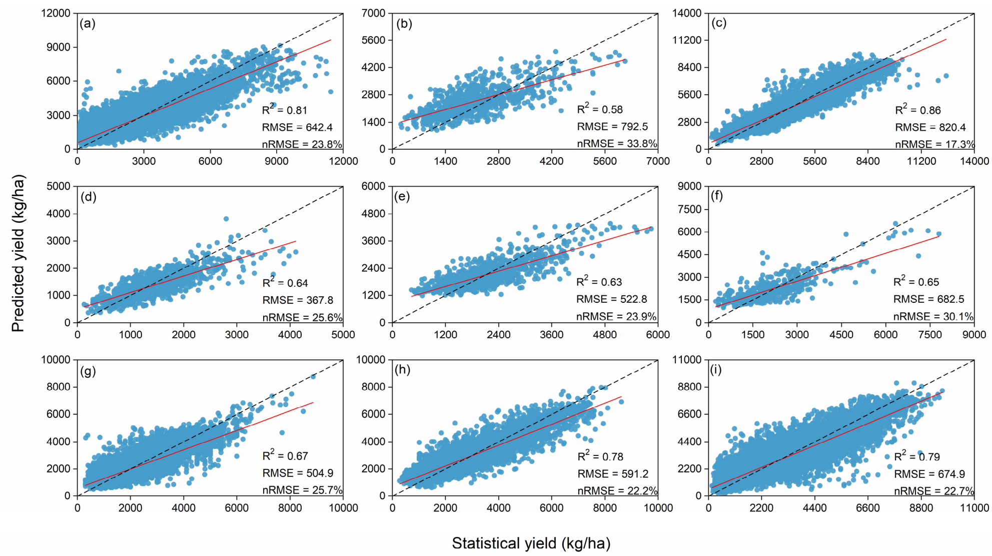

Figure 4: Comparisons between the predicted yields of GlobalWheatYield4km and statistics. (a) South and East Asia, (b) Central Asia, (c) Europe, (d) spring wheat in Russian Federation and Kazakhstan, (e) winter wheat in Russian Federation, (f) Australia, (g) South America, (h) spring wheat in North America, (i) winter wheat in North America.

Therefore, the optimal LSTM model was implemented to predict global wheat yield at grid scale. Overall, the predicted yield agreed well with the census data as they were closely and consistently distributed around the 1:1 line, with R^2 (0.56~0.86), RMSE (123.2~911.3 kg/ha) and nRMSE (13.8~33.8%) (Figs. 4, S1). The highest R^2 was found in Bangladesh (R^2 = 0.86, nRMSE = 14.9%) and Europe (R^2 = 0.86, nRMSE = 17.3%), followed by China, Chile, Pakistan, India, Canada and the United

https://doi.org/10.5194/essd-2022-297
Preprint. Discussion started: 2 September 2022

States (R^2 of 0.77~0.82). By contrast, the lowest R^2 was found in Japan (R^2 = 0.56, nRMSE = 20.6%), Afghanistan and Iran (R^2 = 0.58, nRMSE = 33.8%), which might be caused by the less wheat cultivation or insufficient yield records.

Fig. 5 showed the spatial distributions of GlobalWheatYield4km and census data in 2010. Generally, the spatial patterns of predicted yields were consistent with the statistical yields, with a large variability from 130–11546 kg/ha. We further summarized the gridded yield by countries. The averages of yield were highest in Europe (e.g., Belgium: 8457 kg/ha; Netherlands: 8011 kg/ha), followed by Chile (5201 kg/ha) and China (4658 kg/ha). By contrast, Kazakhstan, Bangladesh and Bolivia achieved the lowest average yield (< 1000 kg/ha).

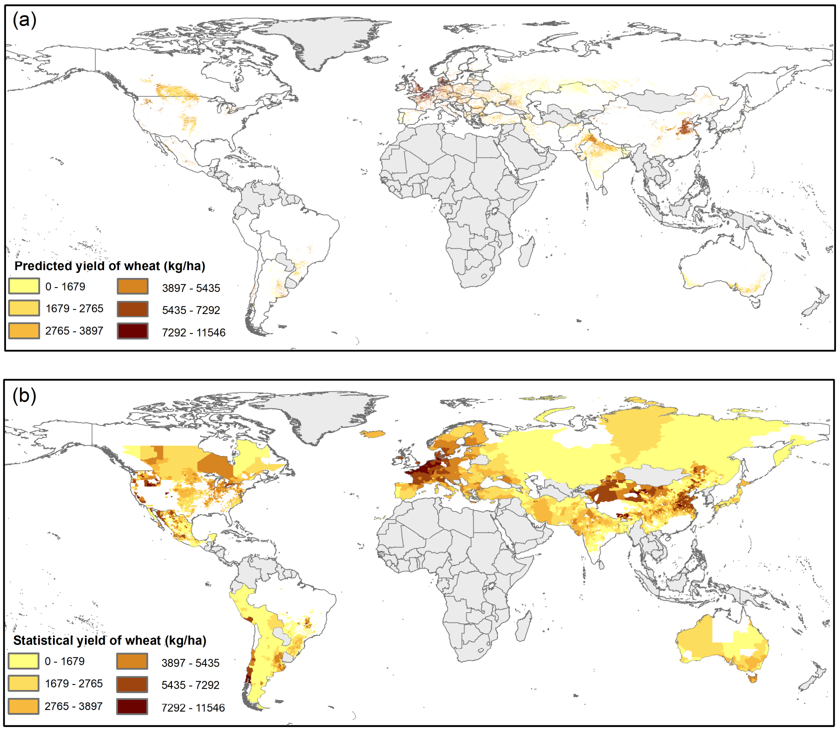

Figure 5: Spatial distribution of the predicted yield (a) and the observed yields (b) in 2010.

https://doi.org/10.5194/essd-2022-297
Preprint. Discussion started: 2 September 2022

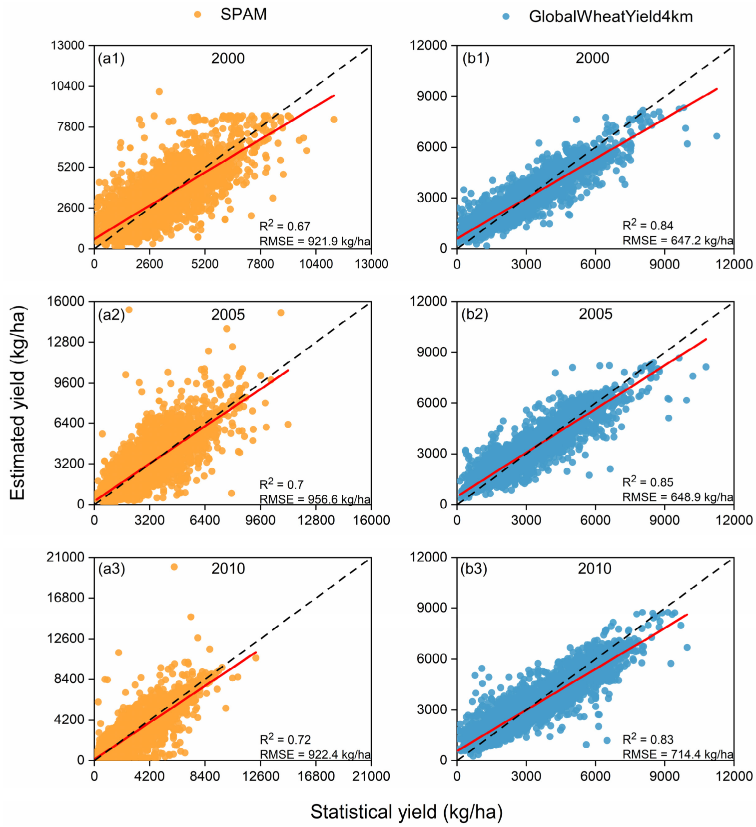

Figure 6: Comparisons between census statistics and estimated yields of SPAM (a1-a3) or GlobalWheatYield4km (b1-b3) for 2000 (a1, b1), 2005 (a2, b2) and 2010 (a3, b3).

3.3 Comparing GlobalWheatYield4km with SPAM

We aggregated gridded-yield estimates of GlobalWheatYield4km and SPAM in 2000, 2005 and 2010 to administrative units and then compared them with census yields, respectively. Overall, the yield estimates of GlobalWheatYield4km showed higher consistencies with census yields as they were closer to 1:1 line than SPAM, with average R^2 (RMSE) of 0.84 (670.2 kg/ha) and 0.7 (933.6 kg/ha), respectively (Fig. 6). In addition, GlobalWheatYield4km exhibited higher and more robust accuracies than SPAM in all years and regions (Fig. 7, Table S3). The R^2 (RMSE) of GlobalWheatYield4km was improved (reduced) by an average of 50.8% (25.1%) as comparing with SPAM, especially in Australia, Canada, Iran, Pakistan and the United States (improvements over 40% and 34% for R^2 and RMSE). We ascribed such improvement into more accurate wheat distribution

https://doi.org/10.5194/essd-2022-297
Preprint. Discussion started: 2 September 2022

maps and the consequent high-quality input data at more consistent and finer resolution. In contrast, the methodology and input data of SPAM were improved stepwise. We are sure more accurate yield datasets would be expected with higher resolution remote sensing products available in the world.

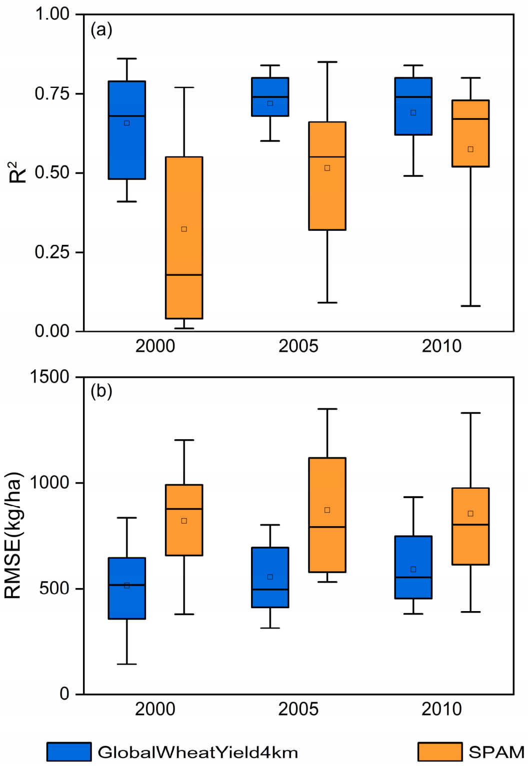

Figure 7: The comparisons of GlobalWheatYield4km-derived yield maps with the SPAM dataset across all regions and three years: (a) R^2, (b) RMSE.

4 Discussions

4.1 Advantages of GlobalWheatYield4km

Compared with other global crop yield products (e.g., SPAM, GAEZ), GlobalWheatYield4km had the following advantages: 1) the highest spatial resolution of 4km among all yield datasets presently available; 2) a higher and more stable accuracy than SPAM as comparing with census data; 3) more accurate spatial distribution and clear subdivision of spring and winter wheat, highly contributing to agricultural system modeling over larger areas; 4) clearly charactering the temporal dynamics of wheat

https://doi.org/10.5194/essd-2022-297
Preprint. Discussion started: 2 September 2022

yields over 40 years. Moreover, we compared two typical ML and DL models that were commonly used for yield prediction and determined the optimal model to generate gridded-yield estimates, which could partly improve the accuracy of our dataset. We found that LSTM consistently outperformed RF despite of years and regions, which was well supported by many previous studies (Jeong et al., 2022; Luo et al., 2022a; Schwalbert et al., 2020; Tian et al., 2021). The strengths of the LSTM model are its recurrent neural network structure, which had been proved to successfully capture cumulative and complex nonlinear relationships between crop yields and climatic factors (Jiang et al., 2019; Zhang et al., 2021).

4.2 Uncertainties

Despite the higher accuracy of GlobalWheatYield4km, there were still some uncertainties. First, the largest constraint was the uncertainties of remote sensing data. For example, cloud and snow contaminations could cause noise in GLASS LAI products and consequently dampen the wheat detection signal (Xiao et al., 2014). Moreover, the coarse spatial resolution of 1 km could result in mixed pixel issues, thereby reducing the accuracy of our dataset, especially in areas where wheat was sparsely and less cultivated such as South America. To partly avoid these uncertainties, we integrated annual maps during 2006-2014 to generate a single map of wheat harvested area on the basis of the grids with cultivation for at least 5 years. However, it could lead to errors in aggregated features as wheat growing areas changed over time. In future studies, we will attempt to map the spatial distribution of wheat using remote sensing images with finer spatial resolutions (Nie et al., 2022; Wang et al., 2020). In addition, another limitation lay in the accessibility of statistical data. On the one hand, the performance of LSTM was dependent on the quantity and quality of statistics. It was particularly difficult to collect finer-scale census data with longer time coverage in some countries such as Kazakhstan and Afghanistan. On the other hand, reliable statistical data were not available in Africa, leading to data gaps in GlobalWheatYield4km.

5 Data availability

The 4-km global dataset of wheat yield from 1982 to 2020 is available at https://doi.org/10.6084/m9.figshare.10025006 (Luo et al., 2022b).

6 Conclusions

We generated a long-term global wheat yield dataset at a spatial resolution of 4-km using data-driven models. First, we mapped the spatial distribution of wheat harvested area using a phenology-based method. Then, we compared the predictive performance of two commonly used ML and DL models and finally developed the optimal model to estimate yield at grid scale. The wheat distribution map had a high accuracy with an average R^2 of 0.8. The LSTM model outperformed RF in predicting gridded yields, with R^2 (nRMSE) ~0.72 (13.1) and 0.64 (16.2), respectively. The GlobalWheatYield4km dataset were highly consistent with statistical data, indicated by R^2 (RMSE) of 0.82 (619.8 kg/ha) and 51% higher accuracy than those

https://doi.org/10.5194/essd-2022-297
Preprint. Discussion started: 2 September 2022

of SPAM (R^2 ~0.49) across all regions and three years. Our GlobalWheatYield4km can be applied for many purposes, including large-scale agricultural system modeling and climate change impact assessments.

Author contribution

Z. Z, F. T and Y. L designed the research. J. C, J. Z and F. C collected and processed datasets. J. H and H. Z validated the resultant crop maps. Y. L implemented the research and wrote the paper; Z. Z, L. Z, J. X and F. T revised the manuscript.

Competing interests

The authors declare that they have no conflict of interest.

Acknowledgements

This study was funded by National Natural Science Foundation of China (42061144003, 41977405).

References

Abatzoglou, J. T., Dobrowski, S. Z., Parks, S. A., and Hegewisch, K. C.: TerraClimate, a high-resolution global dataset of monthly climate and climatic water balance from 1958-2015, Sci. Data, 5, 170191, 2018.

Balaghi, R., Tychon, B., Eerens, H., and Jlibene, M.: Empirical regression models using NDVI, rainfall and temperature data for the early prediction of wheat grain yields in Morocco, Int. J. Appl. Earth Obs., 10, 438-452, https://doi.org/10.1016/j.jag.2006.12.001, 2008.

Breiman, L.: Random forests, Mach. Learn., 45, 5-32, https://doi.org/10.1023/A:1010933404324, 2001.

Burke, M. and Lobell, D. B.: Satellite-based assessment of yield variation and its determinants in smallholder African systems, Proc. Natl. Acad. Sci. U. S. A., 114, 2189-2194, https://doi.org/10.1073/pnas.1616919114, 2017.

Bussay, A., van der Velde, M., Fumagalli, D., and Seguini, L.: Improving operational maize yield forecasting in Hungary, Agric. Syst., 141, 94-106, https://doi.org/10.1016/j.agsy.2015.10.001, 2015.

Cai, Y. P., Guan, K. Y., Lobell, D., Potgieter, A. B., Wang, S. W., Peng, J., Xu, T. F., Asseng, S., Zhang, Y. G., You, L. Z., and Peng, B.: Integrating satellite and climate data to predict wheat yield in Australia using machine learning approaches, Agric. For. Meteorol., 274, 144-159, https://doi.org/10.1016/j.agrformet.2019.03.010, 2019.

Cao, J., Zhang, Z., Tao, F. L., Zhang, L. L., Luo, Y. C., Zhang, J., Han, J. C., and Xie, J.: Integrating Multi-Source Data for Rice Yield Prediction across China using Machine Learning and Deep Learning Approaches, Agric. For. Meteorol., 297, 108275, https://doi.org/10.1016/j.agrformet.2020.108275, 2021.

https://doi.org/10.5194/essd-2022-297
Preprint. Discussion started: 2 September 2022

Chen, Y. and Tao, F. L.: Potential of remote sensing data-crop model assimilation and seasonal weather forecasts for early-season crop yield forecasting over a large area, Field Crop. Res., 276, 108398, https://doi.org/10.1016/j.fcr.2021.108398, 2022.

Dong, J. W., Xiao, X. M., Kou, W. L., Qin, Y. W., Zhang, G. L., Li, L., Jin, C., Zhou, Y. T., Wang, J., Biradar, C., Liu, J. Y., and Moore, B.: Tracking the dynamics of paddy rice planting area in 1986-2010 through time series Landsat images and phenology-based algorithms, Remote Sens. Environ., 160, 99-113, https://doi.org/10.1016/j.rse.2015.01.004, 2015.

FAO, IFAD, UNICEF, WFP, and WHO: The State of Food Security and Nutrition in the World 2020. Transforming food systems for affordable healthy diets, FAO, Rome, Italy, https://doi.org/10.4060/ca9692en, 2020.

FAO, IFAD, UNICEF, WFP and WHO.: The State of Food Security and Nutrition in the World 2021. Transforming food systems for food security, improved nutrition and affordable healthy diets for all. Rome, FAO. https://doi.org/10.4060/cb4474en, 2021.

FAOSTAT: Food and Agriculture Organization of the United Nations Statistics Division, available at: http://www.fao.org/faostat/ en/#data (last access: August 2022), 2020

Feng, L. W., Wang, Y. M., Zhang, Z., and Du, Q. Y.: Geographically and temporally weighted neural network for winter wheat yield prediction, Remote Sens. Environ., 262, https://doi.org/10.1016/j.rse.2021.112514, 2021.

Fischer, G., Nachtergaele, F. O., Prieler, S., Teixeira, E., Tóth, G., Velthuizen, H. V., Verelst, L., and Wiberg, D.: Global Agro- ecological Zones (GAEZ v3.0), IIASA, Laxenburg, Austria and FAO, Rome, Italy, 2012.

Folberth, C., Khabarov, N., Balkovic, J., Skalsky, R., Visconti, P., Ciais, P., Janssens, I. A., Penuelas, J., and Obersteiner, M.: The global cropland-sparing potential of high-yield farming, Nat. Sustain., 3, 281-289, https://doi.org/10.1038/s41893-020-0505-x, 2020.

Geng, L. Y., Ma, M. G., Wang, X. F., Yu, W. P., Jia, S. Z., and Wang, H. B.: Comparison of Eight Techniques for Reconstructing Multi-Satellite Sensor Time-Series NDVI Data Sets in the Heihe River Basin, China, Remote Sens., 6, 2024-2049, https://doi.org/10.3390/rs6032024, 2014.

Grogan, D., Frolking, S., Wisser, D., Prusevich, A., and Glidden, S.: Global gridded crop harvested area, production, yield, and monthly physical area data circa 2015, Sci. Data, 9, 15, https://doi.org/10.1038/s41597-021-01115-2, 2022.

Hochreiter, S. and Schmidhuber, J.: Long short-term memory, Neural Comput., 9, 1735-1780, https://doi.org/10.1162/neco.1997.9.8.1735, 1997.

Huang, J. X., Gomez-Dans, J. L., Huang, H., Ma, H. Y., Wu, Q. L., Lewis, P. E., Liang, S. L., Chen, Z. X., Xue, J. H., Wu, Y. T., Zhao, F., Wang, J., and Xie, X. H.: Assimilation of remote sensing into crop growth models: Current status and perspectives, Agric. For. Meteorol., 276, 107609, https://doi.org/10.1016/j.agrformet.2019.06.008, 2019.

Iizumi, T. and Sakai, T.: The global dataset of historical yields for major crops 1981-2016, Sci Data, 7, 97, https://doi.org/10.1038/s41597-020-0433-7, 2020.

Ines, A. V. M., Das, N. N., Hansen, J. W., and Njoku, E. G.: Assimilation of remotely sensed soil moisture and vegetation with a crop simulation model for maize yield prediction, Remote Sens. Environ., 138, 149-164, https://doi.org/10.1016/j.rse.2013.07.018, 2013.

https://doi.org/10.5194/essd-2022-297
Preprint. Discussion started: 2 September 2022

International Food Policy Research Institute.: 2022 Global Food Policy Report: Climate Change and Food Systems. Washington, DC: International Food Policy Research Institute. https://doi.org/10.2499/9780896294257, 2022.

Jeong, S., Ko, J., and Yeom, J. M.: Predicting rice yield at pixel scale through synthetic use of crop and deep learning models with satellite data in South and North Korea, Sci. Total Environ., 802, 149726, https://doi.org/10.1016/j.scitotenv.2021.149726, 2022.

Jiang, H., Hu, H., Zhong, R. H., Xu, J. F., Xu, J. L., Huang, J. F., Wang, S. W., Ying, Y. B., and Lin, T.: A deep learning approach to conflating heterogeneous geospatial data for corn yield estimation: A case study of the US Corn Belt at the county level, Glob. Change Biol., 26, 1754-1766, https://doi.org/10.1111/gcb.14885, 2020.

Jin, S. C., Su, Y., Gao, S., Hu, T. Y., Liu, J., Guo, Q. H.: The Transferability of Random Forest in Canopy Height Estimation from Multi-Source Remote Sensing Data, Remote Sens., 10, 1183, https://doi.org/10.3390/rs10081183, 2018.

Jin, Z. N., Azzari, G., and Lobell, D. B.: Improving the accuracy of satellite-based high-resolution yield estimation: A test of multiple scalable approaches, Agric. For. Meteorol., 247, 207-220, https://doi.org/10.1016/j.agrformet.2017.08.001, 2017.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., and Liu, T.-Y.: LightGBM: a highly effcient gradient boosting decision tree. In: Proceedings of the Thirty-First Conference on Neural Information Processing System. Long Beach, CA, USA, 4, 2017.

Kern, A., Barcza, Z., Marjanovic, H., Arendas, T., Fodor, N., Bonis, P., Bognar, P., and Lichtenberger, J.: Statistical modelling of crop yield in Central Europe using climate data and remote sensing vegetation indices, Agric. For. Meteorol., 260, 300-320, https://doi.org/10.1016/j.agrformet.2018.06.009, 2018.

Lobell, D. B. and Burke, M. B.: On the use of statistical models to predict crop yield responses to climate change, Agric. For. Meteorol., 150, 1443-1452, https://doi.org/10.1016/j.agrformet.2010.07.008, 2010.

Lobell, D. B., Cassman, K. G., and Field, C. B.: Crop Yield Gaps: Their Importance, Magnitudes, and Causes, Annu. Rev. Environ. Resour., 34, 179-204, https://doi.org/10.1146/annurev.environ.041008.093740, 2009.

Luo, Y. C., Zhang, Z., Zhang, L. L., and Cao, J.: Spatiotemporal patterns of winter wheat phenology and its climatic drivers based on an improved pDSSAT model, Sci. China Earth Sci., 64, 2144-2160, https://doi.org/10.1007/s11430-020-9821-0, 2021.

Luo, Y. C., Zhang, Z., Li, Z. Y., Chen, Y., Zhang, L. L., Cao, J., and Tao, F. L.: Identifying the spatiotemporal changes of annual harvesting areas for three staple crops in China by integrating multi-data sources, Environ Res Lett, 15, 074003, https://doi.org/10.1088/1748-9326/ab80f0, 2020.

Luo, Y. C., Zhang, Z., Cao, J., Zhang, L. L., Zhang, J., Han, J. C., Zhuang, H. M., Cheng, F., and Tao, F. L.: Accurately mapping global wheat production system using deep learning algorithms, Int. J. Appl. Earth Obs., 110, 102823, https://doi.org/10.1016/j.jag.2022.102823, 2022a.

Luo, Y. C., Zhang, Z., Cao, J., Zhang, L. L., Zhang, J., Han, J. C., Zhuang, H. M., Cheng, F., Xu J. L., and Tao, F. L.: GlobalWheatYield4km: a global wheat yield dataset at 4-km resolution during 1982-2020 based on deep learning approaches, https://doi.org/10.6084/m9.figshare.10025006, 2022b.

https://doi.org/10.5194/essd-2022-297
Preprint. Discussion started: 2 September 2022

Monfreda, C., Ramankutty, N., and Foley, J. A.: Farming the planet: 2. Geographic distribution of crop areas, yields, physiological types, and net primary production in the year 2000, Glob. Biogeochem. Cy., 22, GB1022, https://doi.org/10.1029/2007GB002947, 2008.

Nachtergaele, F. O., van Velthuizen, H., Verelst, L., Wiberg, D., Batjes, N. H., Dijkshoorn, J. A., van Engelen, V. W. P., Fischer, G., Jones, A., Montanarella, L., Petri, M., Prieler, S., Teixeira, E., and Shi, X.: Harmonized World Soil Database (version 1.2). Food and Agriculture Organization of the UN, International Institute for Applied Systems Analysis, ISRIC - World Soil Information, Institute of Soil Science - Chinese Academy of Sciences, Joint Research Centre of the EC. http://www.iiasa.ac.at/Research/LUC/External-World-soil-database/HWSD_Documentation.pdf, 2012.

Niu, Q. D., Li, X. C., Huang, J. X., Huang, H., Huang, X. D., Su, W., and Yuan, W. P.: A 30 m annual maize phenology dataset from 1985 to 2020 in China, Earth Syst. Sci. Data, 14, 2851–2864, https://doi.org/10.5194/essd-14-2851-2022, 2022.

Ray, D. K., Gerber, J. S., MacDonald, G. K., and West, P. C.: Climate variation explains a third of global crop yield variability, Nat. Commun., 6, 5989, https://doi.org/10.1038/ncomms6989, 2015.

Ren, S. L., Qin, Q. M., and Ren, H. Z.: Contrasting wheat phenological responses to climate change in global scale, Sci. Total Environ., 665, 620-631, https://doi.org/10.1016/j.scitotenv.2019.01.394, 2019.

Rötter, R. P., Hoffmann, M. P., Koch, M., and Müller, C.: Progress in modelling agricultural impacts of and adaptations to climate change, Curr. Opin. Plant Biol., 45, 255-261, https://doi.org/10.1016/j.pbi.2018.05.009, 2018.

Savitzky, A. and Golay, M. J. E.: Smoothing and Differentiation Of Data by Simplified Least Squares Procedures, Anal Chem, 36, 1627-1639, https://doi.org/10.1021/ac60214a047, 1964.

Schwalbert, R. A., Amado, T., Corassa, G., Pott, L. P., Prasad, P. V. V., and Ciampitti, I. A.: Satellite-based soybean yield forecast: Integrating machine learning and weather data for improving crop yield prediction in southern Brazil, Agric. For. Meteorol., 284, 107886, https://doi.org/10.1016/j.agrformet.2019.107886, 2020.

Song, X. P., Potapov, P. V., Krylov, A., King, L., Di Bella, C. M., Hudson, A., Khan, A., Adusei, B., Stehman, S. V., and Hansen, M. C.: National-scale soybean mapping and area estimation in the United States using medium resolution satellite imagery and field survey, Remote Sens. Environ., 190, 383-395, https://doi.org/10.1016/j.rse.2017.01.008, 2017.

Sulser, T., Wiebe, K. D., Dunston, S., Cenacchi, N., Nin-Pratt, A., Mason-D'Croz, D., Robertson, R. D., Willenbockel, D., and Rosegrant, M. W.: Climate change and hunger: Estimating costs of adaptation in the agrifood system. Food policy report June 2021. Washington, DC: International Food Policy Research Institute (IFPRI). https://doi.org/10.2499/9780896294165, 2021.

Tao, F., Yokozawa, M., and Zhang, Z.: Modelling the impacts of weather and climate variability on crop productivity over a large area: A new process-based model development, optimization, and uncertainties analysis, Agric. For. Meteorol., 149, 831-850, https://doi.org/10.1016/j.agrformet.2008.11.004, 2009.

Tao, F. L., Yokozawa, M., Xu, Y. L., Hayashi, Y., and Zhang, Z.: Climate changes and trends in phenology and yields of field crops in China, 1981-2000, Agric. For. Meteorol., 138, 82-92, https://doi.org/10.1016/j.agrformet.2006.03.014, 2006.

https://doi.org/10.5194/essd-2022-297
Preprint. Discussion started: 2 September 2022

Teluguntla, P., Thenkabail, P., Xiong, J., Gumma, M., Giri, C., Milesi, C., Ozdogan, M., Congalton, R., Tilton, J., Sankey, T., Massey, R., Phalke, A., Yadav, K.: NASA Making Earth System Data Records for Use in Research Environments (MEaSUREs) Global Food Security Support Analysis Data (GFSAD) Crop Mask 2010 Global 1 km V001 [Data set]. NASA EOSDIS Land Processes DAAC. https://doi.org/10.5067/MEaSUREs/GFSAD/GFSAD1KCM.001, 2016.

Thenkabail, P. S., Biradar, C. M., Noojipady, P., Dheeravath, V., Li, Y. J., Velpuri, M., Gumma, M., Gangalakunta, O. R. P., Turral, H., Cai, X. L., Vithanage, J., Schull, M. A., and Dutta, R.: Global irrigated area map (GIAM), derived from remote sensing, for the end of the last millennium, Int. J. Remote Sens., 30, 3679-3733, https://doi.org/10.1080/01431160802698919, 2009.

Tian, H. R., Wang, P. X., Tansey, K., Zhang, J. Q., Zhang, S. Y., and Li, H. M.: An LSTM neural network for improving wheat yield estimates by integrating remote sensing data and meteorological data in the Guanzhong Plain, PR China, Agric. For. Meteorol., 310, 108629, https://doi.org/10.1016/j.agrformet.2021.108629, 2021.

UN.: Sustainable Development Goal 2. Sustainable Development Knowledge Platform. https://sustainabledevelopment.un.org/sdg2, 2017 (last access: August 2022).

USDA: Major world crop areas and climatic profiles. Agricultural Handbook. 664, p. 279, 1994.

van Dijk, M., Morley, T., Rau, M. L., and Saghai, Y.: A meta-analysis of projected global food demand and population at risk of hunger for the period 2010-2050, Nat. Food, 2, 494-501, https://doi.org/10.1038/s43016-021-00322-9, 2021.

Vermote, E., Justice, C., Csiszar, I., Eidenshink, J., Myneni, R., Baret, F., Masuoka, E., Wolfe, R., Claverie M., and NOAA CDR Program: NOAA Climate Data Record (CDR) of Normalized Difference Vegetation Index (NDVI), Version 5. NOAA National Climatic Data Center. doi:10.7289/V5PZ56R6, 2014.

Wang, C. Z., Zhang, Z., Chen, Y., Tao, F. L., Zhang, J., and Zhang, W.: Comparing different smoothing methods to detect double-cropping rice phenology based on LAI products - a case study in the Hunan province of China, Int. J. Remote Sens., 39, 6405-6428, https://doi.org/10.1080/01431161.2018.1460504, 2018.

Wang, S., Di Tommaso, S., Deines, J. M., and Lobell, D. B.: Mapping twenty years of corn and soybean across the US Midwest using the Landsat archive, Sci. Data, 7, 307, https://doi.org/10.1038/s41597-020-00646-4, 2020.

Xiao, Z. Q., Liang, S. L., Wang, J. D., Xiang, Y., Zhao, X., and Song, J. L.: Long-Time-Series Global Land Surface Satellite Leaf Area Index Product Derived From MODIS and AVHRR Surface Reflectance, IEEE Trans. Geosci. Remote Sensing, 54, 5301-5318, https://doi.org/10.1109/TGRS.2016.2560522, 2016.

Xiao, Z. Q., Liang, S. L., Wang, J. D., Chen, P., Yin, X. J., Zhang, L. Q., and Song, J. L.: Use of General Regression Neural Networks for Generating the GLASS Leaf Area Index Product From Time-Series MODIS Surface Reflectance, IEEE Trans. Geosci. Remote Sensing, 52, 209-223, https://doi.org/10.1109/TGRS.2013.2237780, 2014.

You, L. Z., Wood, S., Wood-Sichra, U., and Wu, W. B.: Generating global crop distribution maps: From census to grid, Agr Syst, 127, 53-60, https://doi.org/10.1016/j.agsy.2014.01.002, 2014.

Zhang, L. L., Zhang, Z., Luo, Y. C., Cao, J., Xie, R. Z., and Li, S. K.: Integrating satellite-derived climatic and vegetation indices to predict smallholder maize yield using deep learning, Agric. For. Meteorol., 311, 108666, https://doi.org/10.1016/j.agrformet.2021.108666, 2021.

Zhao, Y., Potgieter, A. B., Zhang, M., Wu, B. F., and Hammer, G. L.: Predicting Wheat Yield at the Field Scale by Combining High-Resolution Sentinel-2 Satellite Imagery and Crop Modelling, Remote Sens., 12, 1024, https://doi.org/10.3390/rs12061024, 2020.

Zhuo, W., Fang, S. B., Gao, X. R., Wang, L., Wu, D., Fu, S. L., Wu, Q. L., and Huang, J. X.: Crop yield prediction using MODIS LAI, TIGGE weather forecasts and WOFOST model: A case study for winter wheat in Hebei, China during 2009-2013, Int. J. Appl. Earth Obs., 106, 102668, https://doi.org/10.1016/j.jag.2021.102668, 2022.