Supplement of Earth Syst. Sci. Data, 15, 791–808, 2023
https://doi.org/10.5194/essd-15-791-2023-supplement

Open Access Earth System Science Data

Supplement of

AsiaRiceYield4km: seasonal rice yield in Asia from 1995 to 2015

Huaqing Wu et al.

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

Supplementary Methods

For machine learning, RF models were trained using the RandomForestRegressor module from the sklearn library, and XGBoost models were trained using the XGBRegressor module from the xgboost library. LSTM models were trained using the Keras library. All these libraries were utilized in Python version 3.9.

Tuning hyper-parameters is helpful to improve the ML prediction accuracy (Shahhosseini et al., 2021). Here, the grid search method was employed to determine the optimal parameter combinations (GridSearchCV from sklearn library in Python 3.9., cv=10, and scoring='neg_mean_squared_error'). Three hyper-parameters of RF were tuned: '*n_estimators*', '*max_features*', and '*min_samples_split*' and six hyper-parameters of XGBoost: '*n_estimators*', '*max_depth*', '*min_child_weight*', '*eta*', '*gamma*', '*tree_method*'. Table S4 lists the defined space of each hyper-parameter. For LSTM, the dropout rate was set to 0.2, and L2 regularization was used to reduce over-fitting. In addition, the early stopping patience was set to 50 to further reduce over-fitting. The hyper-parameter values of each optimal model are listed in Table S5.

Reference:

Shahhosseini, M., Hu, G., Huber, I., and Archontoulis, S. V.: Coupling machine learning and crop modeling improves crop yield prediction in the US Corn Belt, Sci. Rep., 11, 1606, https://doi.org/10.1038/s41598-020-80820-1, 2021.

Supplementary Tables

Table S1. Rice yield sources and the minimum yield available administrative units for countries in the study area

Country	Source	Names of the local administrative unit	Administrative divisions	Time span
Bangladesh	https://bbs.portal.gov.bd; http://www.brri.gov.bd	district	Second-level	1995-2015
Cambodia	https://www.fao.org/in-action/countrystat/national-countrystat-sites/en/	province	First-level	1995-2013
China	http://www.stats.gov.cn/	country, district	Third-level	1995-2015
India	https://data.gov.in/	division	Second-level	1997-2014
Indonesia	https://www.bps.go.id/; https://www.bps.go.id/subject/53/tanaman-pangan.html#subjekViewTab5.html	province	First-level	1995-2014
Japan	https://www.e-stat.go.jp/	prefectural division	First-level	1995-2015
Malaysia	https://www.statistics.gov.my/v1_/	state	First-level	1998-2015
Myanmar	https://www.mmsis.gov.mm/	region, state	First-level	1995-2015
Nepal	https://nepalindata.com/	district	Second-level	1995-2015
Pakistan	http://www.amis.pk/	district	Third-level	1995-2014
Philippines	https://www.fao.org/in-action/countrystat/national-countrystat-sites/en/	lalawigan	Second-level	1995-2015
Republic of Korea	https://kostat.go.kr/portal/korea/index.action	si, gu, gun	Second-level	1996-2015
Thailand	http://web.nso.go.th/	province	First-level	1996-2015
Vietnam	http://www.gso.gov.vn/default_en.aspx/	province, municipality	First-level	1996-2015

Note: names of the local administrative unit represent the specific name of the administrative divisions. These names are collected from Wikipedia (https://en.wikipedia.org/w/index.php?title=List_of_administrative_divisions_by_country&oldid=1081175578, last accessed: 7 April 2022). In addition, all the links in the table can be accessed on 8 April 2022.

Table S2. Overview of the data required in this study

Dataset	Variables	Spatial resolution	Temporal resolution	Time span	Source
Paddy rice area map	Paddy rice area map	500m	Yearly	2000-2015	Han et al. (2022)
Rice yield	Rice yield of single, double, and triple seasons	County level	Yearly	1995-2015	Table S1
Crop calendar	Planting date, heading date, and harvesting date	4km	Yearly	1995-2015	Laborte et al. (2017)
Vegetation index	LAI	0.05°	8 days	1995-2015	Xiao et al. (2016, 2013)
Climate variables	PDSI, Pre, Srad, Tmax, Tmin, Vap, Ws	4km	Monthly	1995-2015	Abatzoglou et al. (2018)
Soil properties	T_Sand, T_SILT, T_CLAY, T_BULK_DEN, T_OC, T_PH_H2O	30″	-	-	Wieder et al. (2014)
Elevation	Elevation	1km	-	-	Hastings et al., (1999)

Note: LAI, Leaf area index; PDSI, Palmer Drought Severity Index; Pre, precipitation accumulated; Srad, downward surface shortwave radiation; Tmax, maximum temperature; Tmin, minimum temperature; Vap, vapor pressure; Ws, wind speed; T_Sand, Topsoil Sand Fraction; T_SILT, Topsoil Silt Fraction; T_CLAY, Topsoil Clay Fraction; T_BULK_DEN, Topsoil Reference Bulk Density; T_OC, Topsoil Organic Carbon; T_PH_H2O, Topsoil pH (H2O).

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, 1–12, https://doi.org/10.1038/sdata.2017.191, 2018.

Han, J., Zhang, Z., Luo, Y., Cao, J., Zhang, L., Zhuang, H., Cheng, F., Zhang, J., and Tao, F.: Annual paddy rice planting area and cropping intensity datasets and their dynamics in the Asian monsoon region from 2000 to 2020, Agric. Syst., 200, 103437, https://doi.org/10.1016/j.agsy.2022.103437, 2022.

Hastings, D. A., Dunbar, P. K., Elphingstone, G. M., Bootz, M., Murakami, H., Maruyama, H., Masaharu, H., Holland, P., Payne, J., and Bryant, N. A.: The global land one-kilometer base elevation (GLOBE) digital elevation model, version 1.0 National Oceanic and Atmospheric Administration, National Geophysical Data Center [data set], 325, 80305–3328, 1999.

Laborte, A. G., Gutierrez, M. A., Balanza, J. G., Saito, K., Zwart, S. J., Boschetti, M., Murty, M. V. R., Villano, L., Aunario, J. K., Reinke, R., Koo, J., Hijmans, R. J., and Nelson, A.: RiceAtlas, a spatial database of global rice calendars and production, Sci. Data, 4, 170074, https://doi.org/10.1038/sdata.2017.74, 2017.

Wieder, W. R., Boehnert, J., Bonan, G. B., and Langseth, M.: Regridded harmonized world soil database v1. 2 ORNL DAAC [data set], https://doi.org/10.3334/ORNLDAAC/1247, 2014.

Xiao, Z., Liang, S., Wang, J., Chen, P., Yin, X., Zhang, L., and Song, J.: Use of general regression neural networks for generating the GLASS leaf area index product from time-series MODIS surface reflectance, IEEE Trans. Geosci. Remote Sens., 52, 209–223, https://doi.org/10.1109/TGRS.2013.2237780, 2013.

Xiao, Z., Liang, S., Wang, J., Xiang, Y., Zhao, X., and Song, J.: Long-time-series global land surface satellite leaf area index product derived from MODIS and AVHRR surface reflectance, IEEE Trans. Geosci. Remote Sens., 54, 5301–5318, https://doi.org/10.1109/TGRS.2016.2560522, 2016.

Table S3. All predicted variables in four categories

Categories	Predicted variables	Abbreviation
TI	Year	-
CEC	Longitude	Lon
	Latitude	Lat
	Elevation	Ele
	Topsoil Sand Fraction	T_Sand
	Topsoil Silt Fraction	T_SILT
	Topsoil Clay Fraction	T_CLAY
	Topsoil Reference Bulk Density	T_BULK_DEN
	Topsoil Organic Carbon	T_OC
	Topsoil pH (H2O)	T_PH_H2O
CGP	Sum of LAI for whole growing period	Sum_LAI_WGP
	Sum of LAI for vegetative period	Sum_LAI_VEP
	Sum of LAI for reproductive period	Sum_LAI_REP
	Sum of Palmer Drought Severity Index for whole growing period	Sum_PDSI_WGP
	Sum of Palmer Drought Severity Index for vegetative period	Sum_PDSI_VEP
	Sum of Palmer Drought Severity Index for reproductive period	Sum_PDSI_REP
	Sum of Precipitation accumulated for whole growing period	Sum_Pre_WGP
	Sum of Precipitation accumulated for vegetative period	Sum_Pre_VEP
	Sum of Precipitation accumulated for reproductive period	Sum_Pre_REP
	Sum of srad for whole growing period	Sum_Srad_WGP
	Sum of srad for vegetative period	Sum_Srad_VEP
	Sum of srad for reproductive period	Sum_Srad_REP
	Sum of maximum temperature for whole growing period	Sum_Tmax_WGP
	Sum of maximum temperature for vegetative period	Sum_Tmax_VEP
	Sum of maximum temperature for reproductive period	Sum_Tamx_REP
	Sum of minimum temperature for whole growing period	Sum_Tmin_WGP
	Sum of minimum temperature for vegetative period	Sum_Tmin_VEP
	Sum of minimum temperature for reproductive period	Sum_Tmin_REP
	Sum of vapor pressure for whole growing period	Sum_Vap_WGP
	Sum of vapor pressure for vegetative period	Sum_Vap_VEP
	Sum of vapor pressure for reproductive period	Sum_Vap_REP

	Sum of wind speed for whole growing period	Sum_Ws_WGP
	Sum of wind speed for vegetative period	Sum_Ws_VEP
	Sum of wind speed for reproductive period	Sum_Ws_REP
EGP	Minimum LAI	Min_LAI
	Maximum LAI	Max_LAI
	Minimum Palmer Drought Severity Index	Min_PDSI
	Maximum Palmer Drought Severity Index	Max_PDSI
	Minimum Precipitation accumulated	Min_Pre
	Maximum Precipitation accumulated	Max_Pre
	Minimum downward surface shortwave radiation	Min_Srad
	Maximum downward surface shortwave radiation	Max_Srad
	Minimum of maximum temperature	Min_Tmax
	Maximum of maximum temperature	Max_Tmax
	Minimum minimum temperature	Min_Tmin
	Maximum minimum temperature	Max_Tmin
	Minimum vapor pressure	Min_Vap
	Maximum vapor pressure	Max_Vap
	Minimum wind speed	Min_Ws
	Maximum wind speed	Max_Ws

Table S4. Detailed information about the hyper-parameters of ML

Hyperparameter	Hyperparameter space	Algorithm
max_features	[2, 6, 8, 12]	RF
min_samples_split	[2, 3, 4]	
n_estimators	[1, 10, 100, 200, 500]	
eta	[0.1,0.2,0.3,0.4,0.5,0.6]	XGBoost
gamma	[0.1,0.2,0.3,0.4,0.5]	
max_depth	[int(x) for x in np.linspace(2, 20, 10)]	
min_child_weight	[int(x) for x in np.linspace(1, 10, 10)]	
n_estimators	[int(x) for x in np.linspace(200, 2000, 10)]	
objective	['reg:squarederror', 'reg:squaredlogerror']	
tree_method	['auto', 'exact', 'approx', 'hist']	

Table S5. Values of hyper-parameters for each optimal estimated model

Country	Case number	Optimal model	Hyper-parameters
Cambodia	1	RF	max_features=12, min_samples_split=2, n_estimators=200
China	2	RF	max_features=12, min_samples_split=2, n_estimators=500
India	3	XGBoost	eta=0.1, gamma=0.4, max_depth=16, min_child_weight=3, n_estimators=1200, objective=reg:squarederror, tree_method=auto
Indonesia	4	XGBoost	eta=0.3, gamma=0.4, max_depth=18, min_child_weight=4, n_estimators=1400, objective=reg:squarederror, tree_method=auto
Japan	5	XGBoost	eta=0.1, gamma=0.4, max_depth=16, min_child_weight=3, n_estimators=1200, objective=reg:squarederror, tree_method=auto
Malaysia	6	XGBoost	eta=0.3, gamma=0.5, max_depth=4, min_child_weight=5, n_estimators=600, objective=reg:squarederror, tree_method=approx
Myanmar	7	XGBoost	eta=0.1, gamma=0.4, max_depth=16, min_child_weight=3, n_estimators=1200, objective=reg:squarederror, tree_method, auto
Nepal	8	RF	max_features=12, min_samples_split=4, n_estimators=200
Pakistan	9	XGBoost	eta=0.1, gamma=0.4, max_depth=16, min_child_weight=3, n_estimators=1200, objective=reg:squarederror, tree_method=auto
Republic of Korea	10	XGBoost	eta=0.3, gamma=0.5, max_depth=4, min_child_weight=5, n_estimators=600, objective=reg:squarederror, tree_method=approx
Thailand	11	XGBoost	eta=0.3, gamma=0.4, max_depth=18, min_child_weight=4, n_estimators, 1400, objective=reg:squarederror, tree_method=auto
China	12	XGBoost	eta=0.3, gamma=0.4, max_depth=18, min_child_weight=4, n_estimators=1400, objective=reg:squarederror, tree_method=auto
	13	RF	max_features=12, min_samples_split=3, n_estimators=500
India	14	XGBoost	eta=0.1, gamma=0.4, max_depth=16, min_child_weight=3, n_estimators, 1200, objective=reg:squarederror, tree_method=auto
	15	RF	max_features=12, min_samples_split=2, n_estimators=500
Philippines	16	RF	max_features=12, min_samples_split=2, n_estimators=500
	17	RF	max_features=12, min_samples_split=2, n_estimators=200
Thailand	18	XGBoost	eta=0.3, gamma=0.5, max_depth, 4, min_child_weight= 5, n_estimators= 600, objective=reg:squarederror, tree_method=approx

	19	XGBoost	eta=0.3, gamma=0.4, max_depth=18, min_child_weight=4, n_estimators=1400, objective=reg:squarederror, tree_method=auto
Vietnam	20	RF	max_features=12, min_samples_split=4, n_estimators=500
	21	XGBoost	eta=0.1, gamma=0.4, max_depth=16, min_child_weight=3, n_estimators=1200, objective=reg:squarederror, tree_method=auto
Bangladesh	22	RF	max_features=12, min_samples_split=2, n_estimators=500
	23	RF	max_features=12, min_samples_split=2, n_estimators=200
	24	RF	max_features=12, min_samples_split=3, n_estimators=500
Vietnam	25	XGBoost	eta=0.1, gamma=0.4, max_depth=16, min_child_weight=3, n_estimators=1200, objective=reg:squarederror, tree_method=auto
	26	XGBoost	eta=0.1, gamma=0.4, max_depth=16, min_child_weight=3, n_estimators=1200, objective=reg:squarederror, tree_method=auto
	27	RF	max_features=12, min_samples_split=3, n_estimators=500

Supplementary Figure

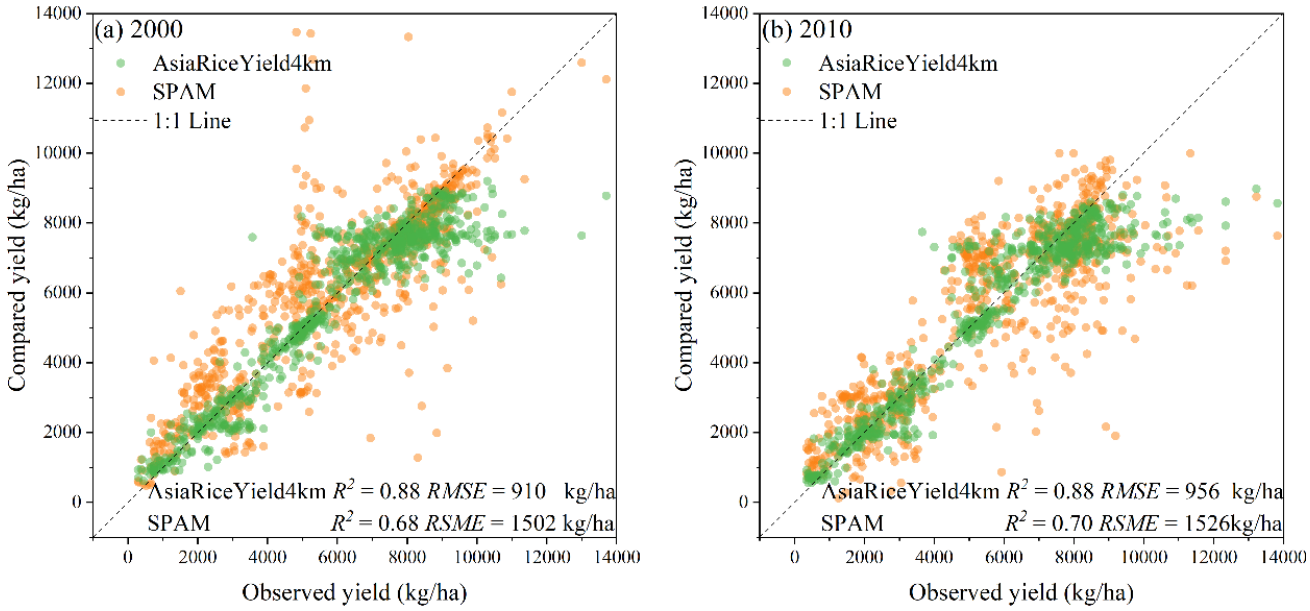

Figure S1. The accuracy of AsiaRiceYield4km and SPAM in 2000 and 2010.