nature food

Brief Communication

https://doi.org/10.1038/s43016-024-01029-3

High-resolution global maps of yield potential with local relevance for targeted crop production improvement

Received: 2 November 2023

Accepted: 11 July 2024

Published online: 29 July 2024

Check for updates

Fernando Aramburu-Merlos[1,2], Marloes P. van Loon[3], Martin K. van Ittersum[3] & Patricio Grassini[1]

Identifying untapped opportunities for crop production improvement in current cropland is crucial to guide food availability interventions. Here we integrated an agronomically robust bottom-up approach with machine learning to generate global maps of yield potential of high resolution (ca. 1 km^2 at the Equator) and accuracy for maize, wheat and rice. These maps serve as a robust reference to benchmark farmers' yields in the context of current cropping systems and water regimes and can help to identify areas with large room to increase crop yields.

Meeting future food demand without massive land conversion depends on the capacity of existing cropland to support higher yields[1]. Estimating yield potential (Ypot), that is, the maximum yield of a locally adapted crop cultivar, serves as a basis for identifying areas with large room to increase crop yields and provides essential input to studies assessing food security, land use and climate change from local to global levels[2,3]. Because of spatial and temporal variation in the factors governing Ypot and limitations to achieve perfection in crop and soil management, measuring Ypot via field experimentation is not feasible at large spatial scales. Alternatively, well-validated crop simulation models, coupled with high-quality weather, soil and cropping system data, can be used to estimate Ypot at local, regional and global levels[4]. Consequently, most large-scale studies rely on crop simulation models to estimate Ypot.

While there is consensus on the use of crop modelling to estimate Ypot, there is considerable debate about the proper spatial framework to use. On the one hand, simulations following 'top-down' approaches rely on global crop models lacking local validation, gridded synthetic climate and soil data, and rough assumptions concerning cropping systems and water regimes[5]. Yet, they allow estimation of Ypot at global scale and high resolution (for example, 5 arc-minutes (ref. 5)) with a modest investment of time and effort. On the other hand, so-called bottom-up approaches are based on sites strategically selected to represent the largest fraction of the harvested area and prioritize using measured weather and soil data, local agronomic data and locally calibrated and evaluated crop models[4]. Not surprisingly, this approach leads to more accurate local Ypot estimates than top-down approaches[6]. However, the better performance of bottom-up approaches is at the expense of higher data requirements and associated time investment in data collection and model calibration and evaluation, making the application of this approach challenging in data-scarce regions[7].

Over the past decade, substantial improvements in computing power, spatial information on soil and climate, and advancement in the use of machine learning (ML) for geospatial analysis have provided new tools that can help address the limitations of bottom-up and top-down approaches[8,9]. Here, we developed a method, hereafter referred to as 'metamodel', to estimate gridded Ypot globally. The metamodel was applied at a 30-arc-second resolution (approximately 1 km^2 at the Equator) to three main cereal crops (maize, wheat and rice) separately for irrigated and rainfed conditions.

The metamodel approach comprises three steps (Fig. 1). The first step consists of a bottom-up crop modelling approach developed for the Global Yield Gap Atlas (GYGA) that results in locally evaluated Ypot estimates for specific sites selected to represent the harvested area distribution. The second step involves training a ML model with these site-specific Ypot values and gridded climate, soil and cropping system data. In the last step, the ML model is used to estimate gridded Ypot and associated prediction uncertainty in areas harvested with a given crop and water regime combination. We restricted ML model predictions

[1]Department of Agronomy and Horticulture, University of Nebraska-Lincoln, Lincoln, NE, USA. [2]Instituto de Innovación para la Producción Agropecuaria y el Desarrollo Sostenible Balcarce (INTA-CONICET), Balcarce, Buenos Aires, Argentina. [3]Plant Production Systems Group, Wageningen University and Research, Wageningen, The Netherlands. e-mail: pgrassini2@unl.edu

https://doi.org/10.1038/s43016-024-01029-3

to the environmental range across the GYGA Ypot sites that were used to train the ML model.

Our high-resolution maps of yield potential shown in Fig. 2 overcome limitations of bottom-up approaches. By using state-of-the-art geospatial analytical tools[10,11], we show that our Ypot maps cover 90–95% of the land planted with these crops (Extended Data Fig. 1 and Supplementary Section 1) without losing much precision in relation to GYGA Ypot (root mean square error (RMSE) of 13% to 18%) (Extended Data Figs. 2–4 and Supplementary Sections 2 and 3), ensuring the local relevance of the gridded Ypot. Additionally, the metamodel approach is flexible enough to accommodate new data, for example, as new site-specific data on Ypot become available from bottom-up approaches like GYGA, the metamodel can be easily updated and applied to generate updated Ypot data at high spatial resolution, ultimately leading to more precise global maps of Ypot.

Our approach also has clear advantages relative to published top-down approaches. As shown in Extended Data Figs. 5–7, Supplementary Section 3 and previous studies[4,6], estimates of Ypot from top-down approaches are biased and lack local relevance. For example, Ypot estimates much lower than average farmer yields are clear evidence of Ypot underestimation. That is the case for 21% of the Ypot estimates for rainfed maize in the US Midwest from a popular top-down approach[5] (Extended Data Fig. 7), which highlights the limitations of using top-down Ypot values that have not been validated with outcomes from bottom-up approaches. In contrast, our estimations of Ypot were consistently above farmer yields.

The metamodel approach to derive gridded Ypot also has limitations. First, our uncertainty assessment is incomplete because it does not consider the uncertainty of GYGA Ypot estimations, which is larger in places with lack of measured weather data and detailed soil maps[7]. Errors in GYGA Ypot propagate to the metamodel, affecting its accuracy. Thus, more and better locally measured weather and soil data can help improve GYGA Ypot and metamodel accuracy. Second, the metamodel is weaker in reproducing Ypot at the lower and upper extremes of the Ypot range (Supplementary Section 2). These biases are common in ML algorithms such as random forest when the number of observations in extreme conditions is limited[12,13]. The GYGA Ypot sites used for model training were selected prioritizing the most important crop producing regions accounting for the largest portion of national crop area. Thus, it is not surprising that the highest metamodel uncertainty occurs in marginal lands with low and highly variable Ypot, relatively low crop area and few GYGA sites, where the gridded Ypot tends to exceed GYGA Ypot (Extended Data Figs. 2 and 4). Likewise, there are few sites with very high GYGA Ypot, which might explain the tendency of the metamodel to underestimate yield potential in those cases. Although more complex ML models could be explored, more is likely to be gained with better quality global data and more training sites derived from bottom-up approaches in such environments. In addition, we note that the metamodel cannot be applied to crop producing regions where climate and soil types differ from those used for model training (Extended Data Fig. 1). For these regions, generating Ypot using bottom-up approaches is advisable rather than using the metamodel outside the environmental range within which it was trained[9].

While our maps are a robust reference to benchmark farmers' yields in the context of current cropping systems and water regimes, we acknowledge the inherent uncertainty from the databases we used, including crop area distribution and environmental variables. Additionally, our maps represent a snapshot of average yield potential at a short period in time. Still, they serve as valuable tools for identifying regions with the greatest potential for increasing crop output through agronomic management and for studies assessing food security, land use change and climate change at local to global levels. Moreover, our method is flexible, allowing for easy updates as newer global data become available. Likewise, it remains essential to periodically update underlying point-based yield potential data to reflect changes in climate, genetics and cropping systems over time.

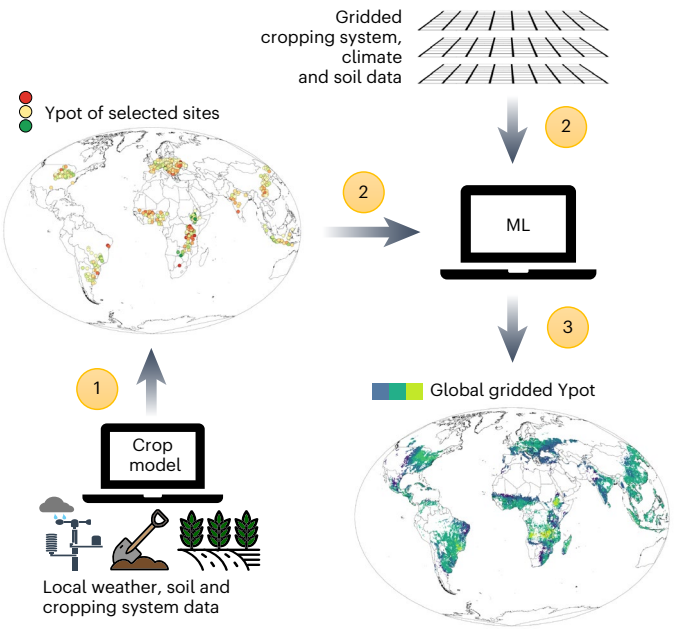


Fig. 1 | Schematic representation of the metamodel. The metamodel integrates a bottom-up approach with machine learning (ML) to estimate high-resolution global yield potential (Ypot). Metamodel steps: (1) Ypot estimation for strategically selected sites using locally calibrated and evaluated crop simulation models and the best available observed weather, soil and cropping system data. (2) Training of a ML algorithm with site-specific Ypot and relevant gridded environmental predictors. (3) Global gridded Ypot estimation for the area of applicability of the ML model and evaluation of its prediction uncertainty.

Methods

Yield potential definitions

Yield potential (Yp) is defined as the maximum yield of a locally adapted crop cultivar as determined by solar radiation, temperature, carbon dioxide and genetic traits that govern length of growing period, light interception by the crop canopy, its conversion to biomass and partitioning of biomass to the harvestable organs[4]. In the case of rainfed crops, water-limited yield potential (Yw) is also determined by precipitation patterns and soil properties influencing the crop water balance[4]. Herein, we use Ypot to refer to the Yp of irrigated crops and/or the Yw of rainfed crops. The difference between Ypot and average farmer yields is the yield gap[4].

Yield potential for selected sites from a bottom-up approach

Our framework builds on the site-specific Ypot estimates of the GYGA. This long-running project has become a reference for agronomically robust Ypot and yield gap data, providing valuable information for food security risk assessments of large regions and countries. GYGA follows a bottom-up protocol for site selection, data collection and crop modelling and makes use of best available data sources, giving priority to measured weather, fine-resolution soil maps and locally validated crop calendars. Over the past 10 years, this protocol has been applied to quantify yield gaps in more than 70 countries and for multiple crops in rainfed and irrigated conditions. This section briefly introduces this protocol; more details can be found at http://www.yieldgap.org and references therein.

First, at each country and crop included in GYGA, sites were strategically selected on the basis of GYGA climate zones (CZ)[14] and harvested area distribution. A CZ is a region with similar climatic conditions as defined by its growing degree days (that is, growing season length in

https://doi.org/10.1038/s43016-024-01029-3

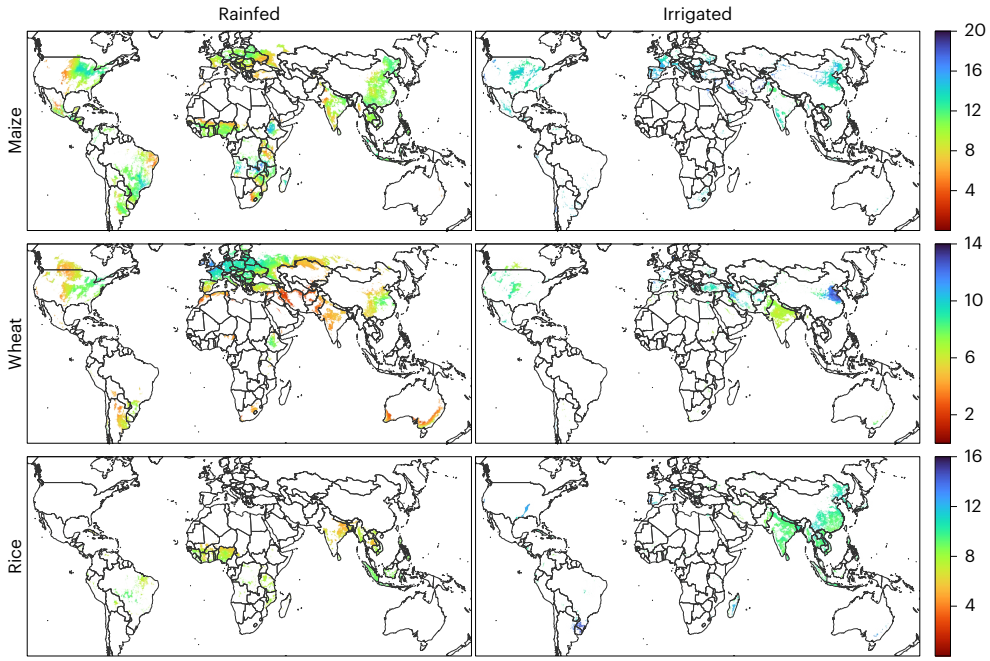


Fig. 2 | Global gridded yield potential for the three main cereal crops around year 2020. Yield potential (Ypot) of irrigated and rainfed maize, wheat and rice was estimated at a 30-arc-second spatial resolution with a machine learning metamodel trained with site-specific Ypot values from the Global Yield Gap Atlas (GYGA, www.yieldgap.org). Predictions were restricted to the metamodel area of applicability (Extended Data Fig. 1). Prediction uncertainty is shown in Extended Data Fig. 5.

thermal time), aridity index (water-stress indicator) and temperature seasonality. On average, the size of each GYGA CZ is ca. 0.3 million km^2, which is smaller than other available agroecological zone frameworks[14]. Within the most important CZs for the target crop, representative sites with long-term weather records were selected to represent the harvested area of a given crop. For this purpose, we considered the harvested area within a 100-km-radius buffer around each site constrained to its corresponding CZ borders such that the buffer does not extend to different CZs. On average, the buffers of selected sites include ca. 100,000 ha of the given crop[15]. Next, the best available observed data for these sites (and their buffers) were retrieved with help from local experts in agronomy, including weather, soil and crop management (for example, cultivar maturity, sowing and harvest dates, and planting density). Soil data are needed only for rainfed conditions, and the three to five most dominant soil types where the target crop is grown were selected for each buffer. These datasets, subject to thorough quality control, are the foundation of crop simulation model performance. Then, crop simulation models were locally calibrated and evaluated in their capacity to reproduce yields of well-managed local field experiments with adapted and commonly used cultivars where yield-limiting and reducing factors had been effectively controlled. The Ypot of selected sites was subsequently simulated with calibrated crop models for multiple years to capture the inter-annual climate variability. These steps resulted in high-quality, unbiased site-specific Ypot estimates that captured the variability in climates, soils and farmers' cropping systems. Up to our last access date (10 June 2023), GYGA provided Ypot data for 543, 573 and 325 sites (weather stations) for, respectively, maize, wheat and rice distributed across 74 countries (Extended Data Fig. 8).

Selection of gridded environmental predictors for the metamodel

A complete list of environmental predictors, descriptions, spatial resolution, sources and references is provided in Supplementary Table 1. Just as crop simulation model performance relies on good local data, selecting relevant gridded environmental predictors of spatial variation in Ypot is key for metamodel performance. For a given crop and cultivar, the spatial variation in irrigated Yp is a function of climatic conditions during its growing season, mainly solar radiation and temperature, while rainfed Yw variation also depends on soil properties and rainfall amount and distribution[4]. Considering these factors, we created crop-specific climatic Ypot predictors from crop calendar information and monthly gridded climatic data. To account for variations in climatic conditions during the cropping season that may affect the crop differently depending on its growing cycle stage, we split the crop growing season into three equal thermal-time periods (early vegetative, flowering and grain filling) and computed climatic variables for each period. Up to three crops per year can be grown in (sub)tropical environments with ample water supply from rainfall and/or irrigation. In those cases, we split each cropping season of each crop cycle into three periods but then computed average climate means for each period across the crop cycles as our goal was to estimate the average Ypot for a given grid.

In addition, we included a set of annual bioclimatic variables that might help to explain spatial variation in Ypot. To that end, we considered those variables used to define CZs (annual growing degree days, aridity index and temperature seasonality), together with other variables describing seasonal rainfall and temperature patterns (for example, precipitation of the warmest quarter of the year). These variables have demonstrated value for spatial prediction of the suitability and distribution of plant species, including cereal crops[16]. Moreover, when used for global Ypot prediction, these bioclimatic variables resulted in better ML model performance than the use of monthly values (for example, average precipitation of each month of the year) as model predictors[8].

We retrieved data on plant-available soil water holding capacity in the first and second metre of soil depth to account for the capacity of the soil to supply water during rain-free periods (note that these were used only for Yw and not for Yp predictions). For model training and validation, we considered the climatic conditions at the GYGA Ypot sites and the average soil and cropping system properties within a 100-km-radius buffer around each site.

Finally, we used harvested area maps from the Spatial Production Allocation Model (SPAM)[17] to define the target area of metamodel

predictions for each crop and water regime. We chose SPAM maps because they remain the only source that provides crop- and water-regime-specific harvested area maps for the three crops (maize, wheat and rice) with global coverage. We note that crop harvested area was not used as a predictor in the metamodel, so SPAM uncertainties have no impact on metamodel performance.

Machine-learning algorithm used in the metamodel

We used random forest regression as ML algorithm in the metamodel. We trained the random forest regression model with site-specific Ypot from GYGA and gridded environmental predictors to generate global gridded Ypot estimates for maize, wheat and rice. This algorithm is flexible enough to capture the complex interactions between crops, climate conditions and soil properties that result in nonlinear yield responses to variation in environmental conditions and has a relatively low computational cost[8]. Such low computational cost was required due to the many iterations of the spatial cross-validation procedure (see below). We tuned the algorithm to avoid overfitting by selecting the random forest tuning parameters (number of covariables considered in each split and minimum node size for a split) that resulted in the lowest spatially cross-validated RMSE for each crop and water regime (see below).

Metamodel validation method

We evaluated metamodel performance by comparing GYGA Ypot with metamodel predictions derived using a spatial cross-validation method: the nearest-neighbour-distance-matching leave-one-out cross-validation (NNDM LOO CV)[10], which is explained below. In the present study, spatial cross-validation and NNDM LOO CV are used interchangeably. We used the following performance metrics: the RMSE derived from the spatial cross-validation, expressed as absolute value and as percentage of the average GYGA Ypot (that is, relative RMSE or normalized RMSE), concordance correlation coefficient, coefficient of determination and mean bias error. In addition, we calculated the percentage of the error due to lack of accuracy and precision. Only the relative RMSE is shown in the main text; other metrics (and their corresponding references) are shown in Supplementary Section 2.

To assess how well a model performs in new sites in the absence of an independent validation dataset, model validation is usually done by partitioning the data into training and testing subsets. When sites are randomly distributed, the data partitioning might be done randomly. For example, the model might be trained with all observations but one (the left-out), and its prediction error tested on the left-out, repeating this procedure for each observation. This method is the leave-one-out cross-validation (LOO CV). However, when observations are clustered in specific regions and spatially autocorrelated, as is the case with data of our study (Extended Data Fig. 8), this validation strategy would inflate the prediction performance metrics. In such cases, the data partitioning strategy must consider the spatial structure of the data, which can be done with the NNDM LOO CV method[10].

NNDM LOO CV is a modification of the LOO CV in which observations near testing sites are excluded for model training. Sites to be excluded are defined such that the distribution function of distances between testing sites and their nearest training sites matches the nearest-neighbour distance distribution function between the prediction area and their nearest training sites. In other words, we validated the capacity of the metamodel to reproduce each GYGA Ypot while excluding that GYGA Ypot and nearest neighbours from model training. The neighbours to be excluded for each GYGA Ypot during the NNDM LOO CV procedure were defined such that the distances between GYGA Ypot used for testing and their nearest GYGA Ypot used for model training match the distances between the prediction grid cells and their nearest GYGA Ypot (Extended Data Fig. 9). For each crop and water regime, the prediction grid was defined as those areas with more than 0.5% of crop harvested area according to SPAM at a spatial resolution of 5 arc-minutes (9.3 km at the Equator).

Delineation of metamodel area of applicability

ML models can generate predictions for any environment, so it is possible to estimate global gridded Ypot independently of the number and distribution of the bottom-up local Ypot used to train it. However, extrapolating to environments outside the environmental range captured by sites with bottom-up estimates leads to meaningless predictions with unknown uncertainty[9]. Therefore, it is imperative to constrain the predictions to the area of applicability of the metamodel. By delineating the geographical scope in which the metamodel remains representative and with known accuracy, the risk of disseminating substantial errors is mitigated, ensuring the reliability and meaningfulness of gridded Ypot predictions.

We delineated the area of applicability of the metamodel following Meyer and Pebesma[11]. First, we computed a dissimilarity index between the biophysical conditions in the training data (GYGA Ypot) and the target prediction area, defined as the global gridded harvested area of each crop and water regime as reported in SPAM at a 5-arc-minute resolution. For each grid cell with more than 0.5% of harvested area for the given crop and water regime, the dissimilarity index equalled the Euclidean distance in the environmental space to the most similar GYGA Ypot, with environmental variables (that is, gridded predictors listed in Supplementary Table 1) weighted by their relative importance in the random forest model. Therefore, the dissimilarity index indicates how different a grid cell is from its most similar GYGA site in terms of biophysical properties defining Ypot. Second, this dissimilarity index was compared with the dissimilarity between the training and testing sites in the NNDM LOO CV procedure. Those grid cells that were more dissimilar than the outlier-removed maximum dissimilarity between NNDM LOO CV training and testing sites were considered outside the area of applicability of the metamodel. Thus, the area of applicability is the geographic region where the estimated NNDM LOO CV performance holds because environmental conditions are similar enough to those in GYGA sites. We computed the area of applicability at a 5-arc-minute resolution and disaggregated it to a 30-arc-second resolution to match that of metamodel Ypot predictions. Given that SPAM crop area maps have a lower spatial resolution than our Ypot estimates grids, our maps might include some grid cells where the target crop is not grown and may miss very isolated areas.

Estimation of global yield potential and prediction uncertainty

We estimated Ypot for maize, wheat and rice for rainfed and irrigated conditions within the area of applicability of the metamodel using the random forest regression model trained with GYGA Ypot and gridded environmental variables. We applied the metamodel at a 30-arc-second resolution (ca. 1^2 km at the Equator), matching the resolution of the climatic data. We aggregated gridded soil predictors available at finer spatial resolution to match this 30-arc-second resolution.

We used the association of the dissimilarity index between training and testing sites with their spatially cross-validated errors to estimate the uncertainty of gridded Ypot predictions[11] (Extended Data Fig. 10). For each crop and water regime, we adjusted a nonlinear model to the association between each observation error and its dissimilarity index with its most similar training data. We used that model to predict the uncertainty (that is, expected RMSE) in the prediction grid at a given dissimilarity index level. The prediction uncertainty is expected to be smaller (larger) in those regions with environmental conditions that are similar (different) to those in GYGA sites, eventually approaching the GYGA Ypot standard deviation in very dissimilar regions.

Comparison of metamodel performance with climate-zone-based approaches

We compared the performance of our metamodel for global gridded Ypot predictions against that of a country-blind CZ extrapolation approach. For each crop and water regime, this approach assumes that Ypot is constant within a CZ (that is, all prediction grid cells within a

CZ have the same Ypot) and equal to the weighted average of all GYGA Ypot within the same CZ worldwide. Weights were defined on the basis of the harvested area in a 100 km buffer zone around each GYGA Ypot site, circumscribed to its corresponding CZ and country. We evaluated CZ gridded predictions with the same NNDM LOO CV method used for the metamodel (Extended Data Fig. 9). That is, each GYGA Ypot was compared against the weighted average of all other GYGA Ypots in the same CZ worldwide while excluding the nearest neighbours.

We note that this approach generates different Ypot values than those reported in GYGA at the CZ level. Whereas the country-blind CZ approach used above as baseline does not consider country borders to generate gridded Ypot predictions, Ypot estimates derived from the original GYGA bottom-up upscaling protocol via CZ are country specific. Therefore, we also tested how the metamodel performs compared with the original GYGA country-specific CZ level Ypot estimations. This assessment is of interest for agronomists and researchers who want to generate more granular Ypot maps than those provided by GYGA at the CZ level. We evaluated the metamodel and GYGA CZ level Ypot estimates with the NNDM LOO CV, considering the crop harvested area of countries included in GYGA as prediction target. For GYGA CZ level Ypot, we made use of CZs with multiple Ypot sites and compared the Ypot of each site against the weighted average Ypot of other sites within the same CZ and country. This cross-validation approach was therefore not possible for CZ countries with only one Ypot site.

Assessment of local relevance in the metamodel and a top-down approach

Yield potential estimates are locally relevant when they are calculated by leveraging local knowledge and data, such that they are agronomically sound and unbiased at subnational levels[4]. By definition, the yield gap between potential and farmer yields cannot be negative; therefore, Ypot estimates that are (much) lower than average farmer yields are unequivocal evidence of Ypot underestimation. We assessed whether Ypot estimates were lower than average farmer yields in the US Midwest rainfed maize at the county level. We chose this region and crop owing to its high yields and the quality of its official statistics, which discriminate crop yields by water regime. We retrieved average farmer yields between 2005 and 2015 from USDA-NASS Quick Stats (https://quickstats.nass.usda.gov). We only considered counties that explicitly reported non-irrigated maize yields or with less than 5% of irrigated area and at least 3 years of data. We calculated the difference between average farmers yield and average gridded Ypot at county level for 861 counties across 12 states. For comparison, we repeated the analysis using Ypot estimates from FAO Global Agroecological Zone[5]. We retrieved the 'agro-climatic potential yield' for rainfed maize from https://gaez.fao.org/ and converted it from dry weight to US harvest weight by dividing it by 0.845. We also used that information to compare spatial patterns of rainfed maize Ypot as derived from a top-down approach (FAO Global Agroecological Zone), a bottom-up approach (GYGA CZ) and the metamodel. We chose East Africa for this comparison for being a data-scarce region with high environmental variability, where the contrasts between methods become more apparent.

Software

All the analysis was done in R[18]. A list of R packages used in the analysis is provided in Supplementary Section S5.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

The high-resolution global maps of yield potential have been deposited in Zenodo (https://doi.org/10.5281/zenodo.12209708) (ref. 19). Data on yield potential are available on the GYGA (https://www.yieldgap.org/). Global climatic data are available on WorldClim (https://www.worldclim.org/). Global gridded soil data are available on ISRIC (https://data.isric.org/). Global crop calendar data are available on SAGE, UW-Madison (https://sage.nelson.wisc.edu/data-and-models/datasets/crop-calendar-dataset/), RiceAtlas (https://www.nature.com/articles/sdata201774) and CropMonitor (https://cropmonitor.org/index.php/eodatatools/baseline-data/). Crop distribution maps are available on SPAM (https://mapspam.info). Source data are provided with this paper.

Code availability

The R code used in the current study is publicly available on GitHub (https://github.com/AramburuMerlos/gGYGA).

References

1. Cassman, K. G. & Grassini, P. A global perspective on sustainable intensification research. *Nat. Sustain.* **3**, 262–268 (2020).
2. van Ittersum, M. K. et al. Can sub-Saharan Africa feed itself? *Proc. Natl Acad. Sci. USA* **113**, 14964–14969 (2016).
3. Marin, F. R. et al. Protecting the Amazon forest and reducing global warming via agricultural intensification. *Nat. Sustain.* **5**, 1018–1026 (2022).
4. van Ittersum, M. K. et al. Yield gap analysis with local to global relevance—a review. *Field Crops Res.* **143**, 4–17 (2013).
5. FAO and IIASA. Global Agro Ecological Zones version 4 (GAEZ v4). http://www.fao.org/gaez/ Accessed 29 Sep 2023.
6. Rattalino Edreira, J. I. et al. Spatial frameworks for robust estimation of yield gaps. *Nat. Food* **2**, 773–779 (2021).
7. Grassini, P. et al. How good is good enough? Data requirements for reliable crop yield simulations and yield-gap analysis. *Field Crops Res.* **177**, 49–63 (2015).
8. Cedrez, C. B. & Hijmans, R. J. Methods for spatial prediction of crop yield potential. *Agron. J.* **110**, 2322–2330 (2018).
9. Meyer, H. & Pebesma, E. Machine learning-based global maps of ecological variables and the challenge of assessing them. *Nat. Commun.* **13**, 2208 (2022).
10. Milà, C., Mateu, J., Pebesma, E. & Meyer, H. Nearest neighbour distance matching leave-one-out cross-validation for map validation. *Methods Ecol. Evol.* **13**, 1304–1316 (2022).
11. Meyer, H. & Pebesma, E. Predicting into unknown space? Estimating the area of applicability of spatial prediction models. *Methods Ecol. Evol.* **12**, 1620–1633 (2021).
12. Breiman, L. Random forests. *Mach. Learn.* **45**, 5–32 (2001).
13. Jeong, J. H. et al. Random forests for global and regional crop yield predictions. *PLoS ONE* **11**, e0156571 (2016).
14. van Wart, J. et al. Use of agro-climatic zones to upscale simulated crop yield potential. *Field Crops Res.* **143**, 44–55 (2013).
15. van Bussel, L. G. J. et al. From field to atlas: upscaling of location-specific yield gap estimates. *Field Crops Res.* **177**, 98–108 (2015).
16. Aramburu Merlos, F. & Hijmans, R. J. Potential, attainable, and current levels of global crop diversity. *Environ. Res. Lett.* **17**, 044071 (2022).
17. Global spatially-disaggregated crop production statistics data for 2010 Version 2.0. Harvard Dataverse. *International Policy Research Institute* https://doi.org/10.7910/DVN/PRFF8V (2019).
18. R Core Team. *R: A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, 2020).
19. Aramburu-Merlos, F., van Loon, M., van Ittersum, M. & Grassini, P. Global gridded maps of yield potential of the Global Yield Gap Atlas (GYGA). *Zenodo* https://doi.org/10.5281/zenodo.12209708 (2024).

Acknowledgements

This study was supported by the National Institute of Food and Agriculture of the United States Department of Agriculture (grants Hatch NEB-22-399 to P.G.) and the National Science Foundation (NSF

#2214604 to P.G.) We thank M. Alimagham (Wageningen University and Research) for his feedback on an early version of this manuscript.

Author contributions

F.A.-M., M.P.v.L, M.K.v.I. and P.G. conceived the research. F.A.-M. performed data acquisition, data processing, modelling and data analysis. F.A.-M., M.P.v.L, M.K.v.I. and P.G. wrote the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

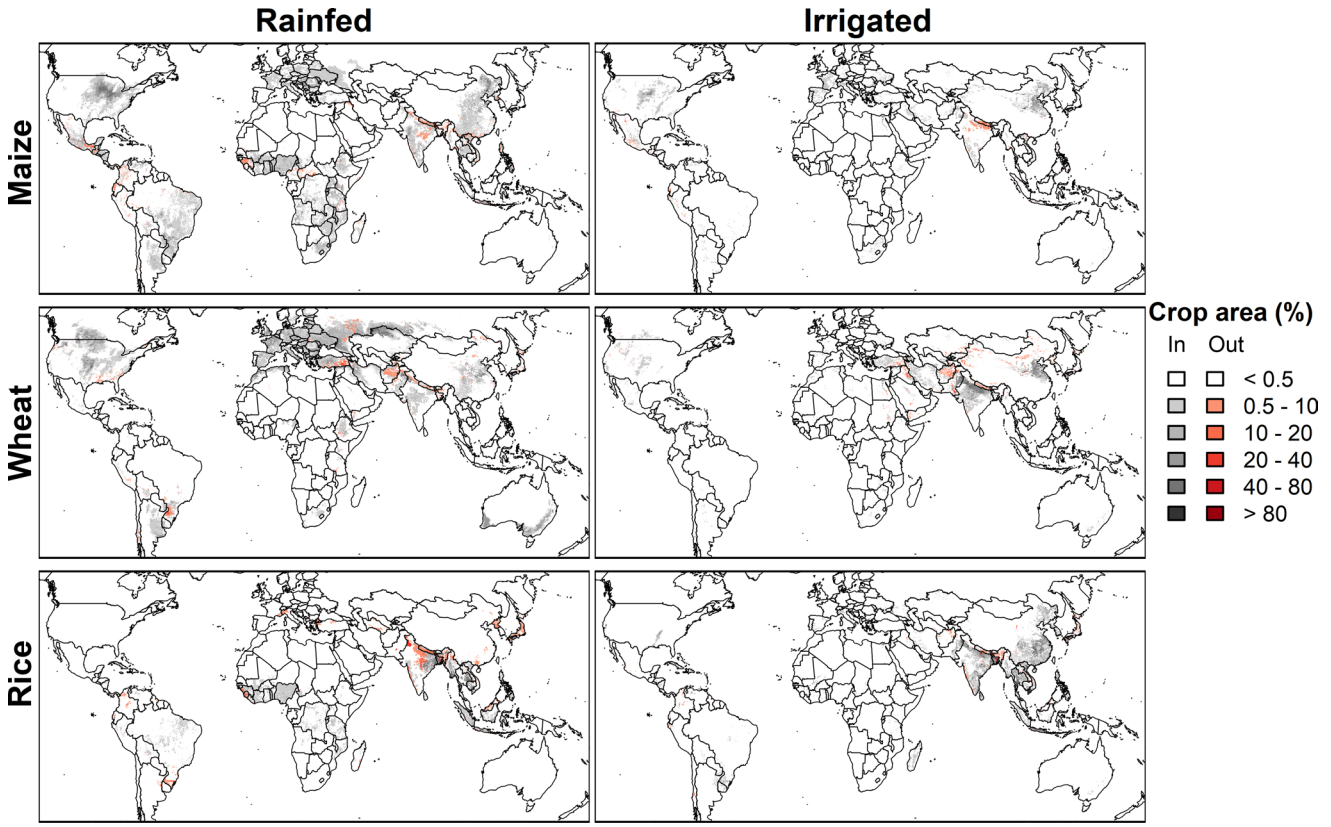
Extended data is available for this paper at https://doi.org/10.1038/s43016-024-01029-3.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s43016-024-01029-3.

Correspondence and requests for materials should be addressed to Patricio Grassini.

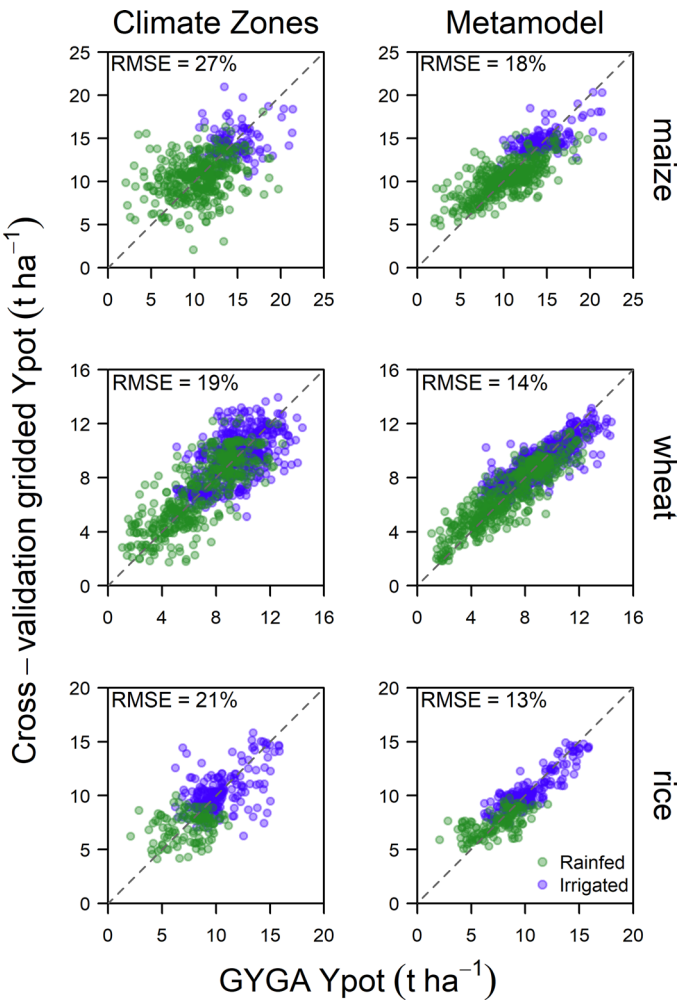
Peer review information *Nature Food* thanks Nimai Senapati, Francisco Villalobos, Bingfang Wu and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

https://doi.org/10.1038/s43016-024-01029-3



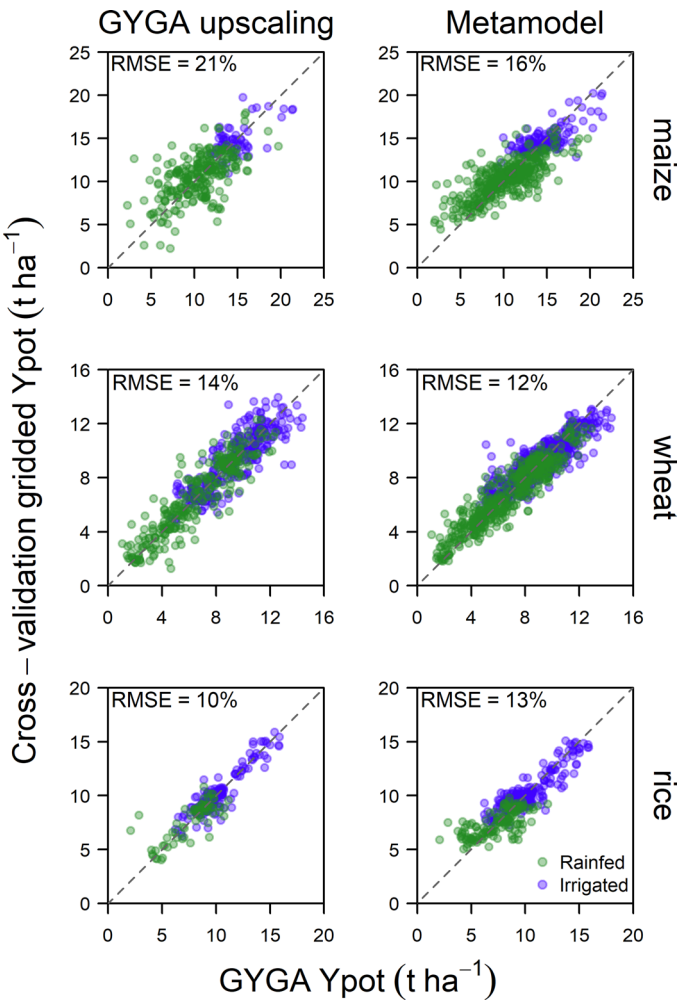
Extended Data Fig. 1 | Area of applicability of metamodel. The metamodel area of applicability is shown in shades of gray. Red areas were excluded from metamodel predictions due to their high dissimilarity with the environments explored by data used to train the metamodel. Color darkness represents the crop harvested area of each crop and water regime combination as percentage of the total area of each pixel. White areas have < 0.5% of given crop and water regime and were not considered. Crop harvested areas were retrieved from SPAM[17].

https://doi.org/10.1038/s43016-024-01029-3



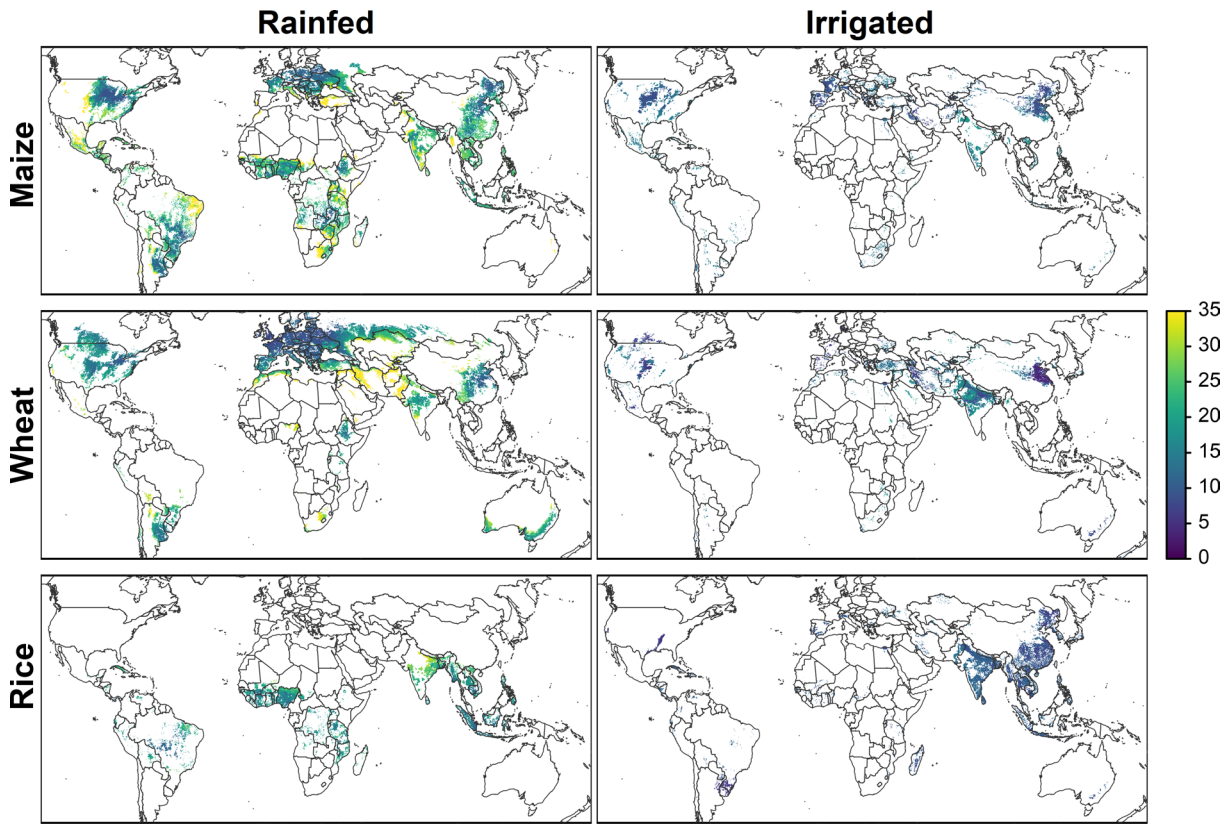
Extended Data Fig. 2 | Global gridded yield potential (Ypot) comparison of two approaches. Comparison of gridded Ypot predictions based on country-blind climate zones extrapolation (left panels) and metamodel (right panels) versus site-level Ypot from the Global Yield Gap Atlas (GYGA Ypot) for three crops and two water regimes. Each point represents a simulation site (reference weather station). Predictions were derived following nearest-neighbor-distance-matching leave-one-out cross-validation method. The root mean square error relative to GYGA Ypot average (RSME %, also known as normalized RMSE) is shown for each method and crop combination. Other model performance metrics are shown in Supplementary Table 2.

https://doi.org/10.1038/s43016-024-01029-3



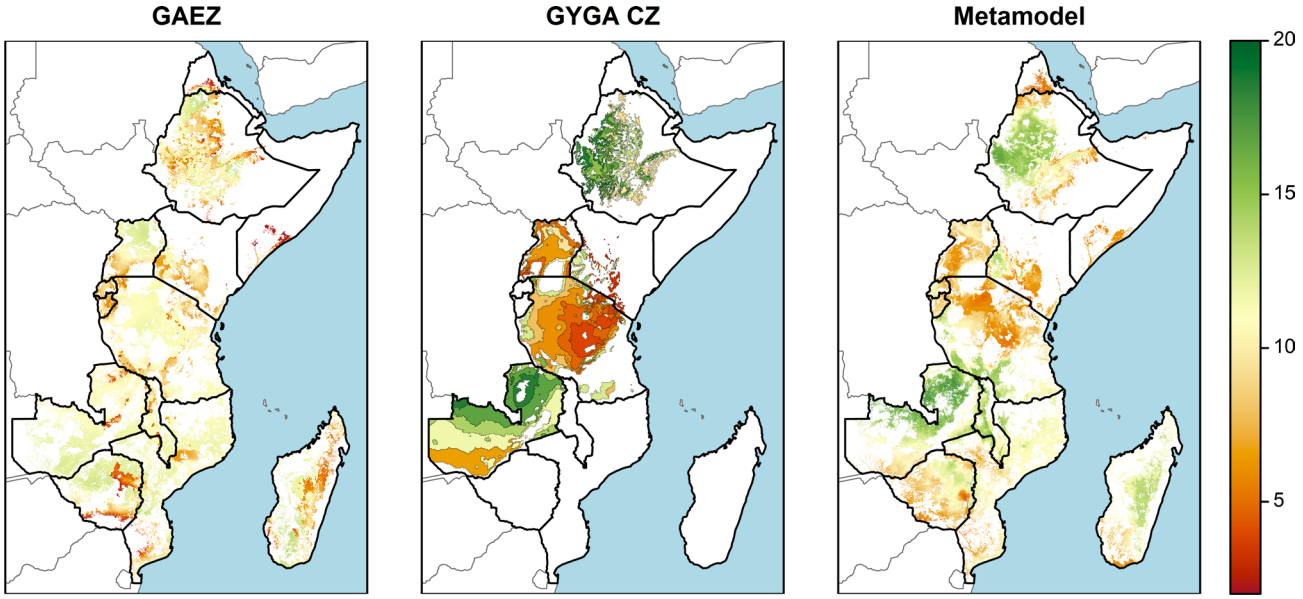
Extended Data Fig. 3 | National gridded yield potential (Ypot) comparison of two approaches. Comparison of Ypot predictions based on GYGA upscaling approach (left panels) and the metamodel (right panels) versus site-level Ypot from the Global Yield Gap Atlas (GYGA Ypot) for three crops and two water regimes. Each point represents a site (reference weather station). Predictions were derived following the Nearest-Neighbor-Distance-Matching Leave-One-Out Cross-Validation method, with crop harvested area of countries included in GYGA as target prediction area. The root mean square error relative to GYGA Ypot average (RSME, %) is shown for each method and crop combination.

https://doi.org/10.1038/s43016-024-01029-3



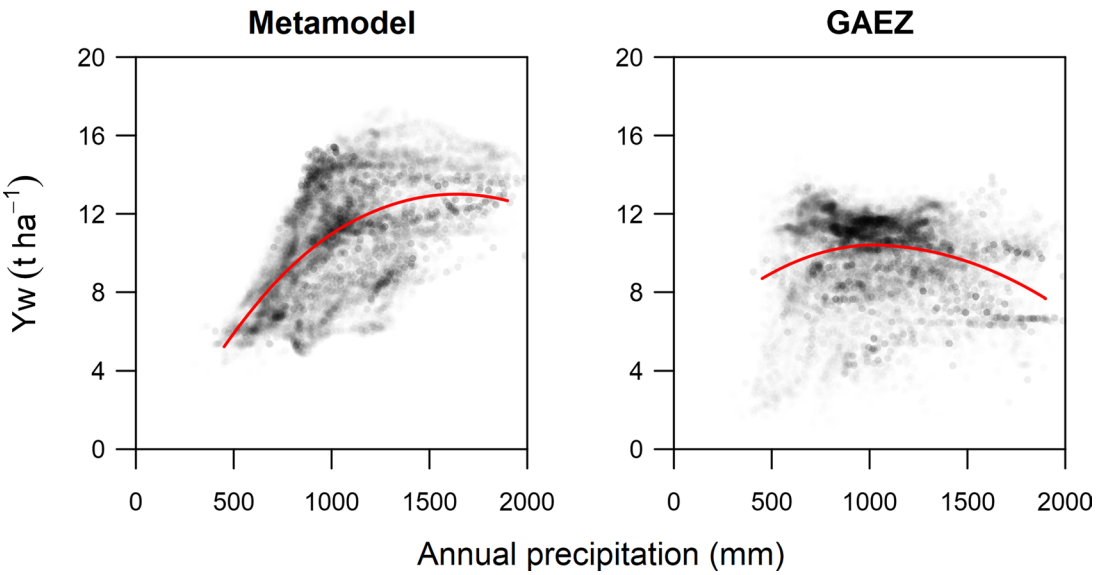
Extended Data Fig. 4 | Metamodel prediction uncertainty. Expected normalized root mean square error (NRMSE) of global gridded yield potential estimates derived from the metamodel, expressed as percentage of the predicted yield potential.

https://doi.org/10.1038/s43016-024-01029-3



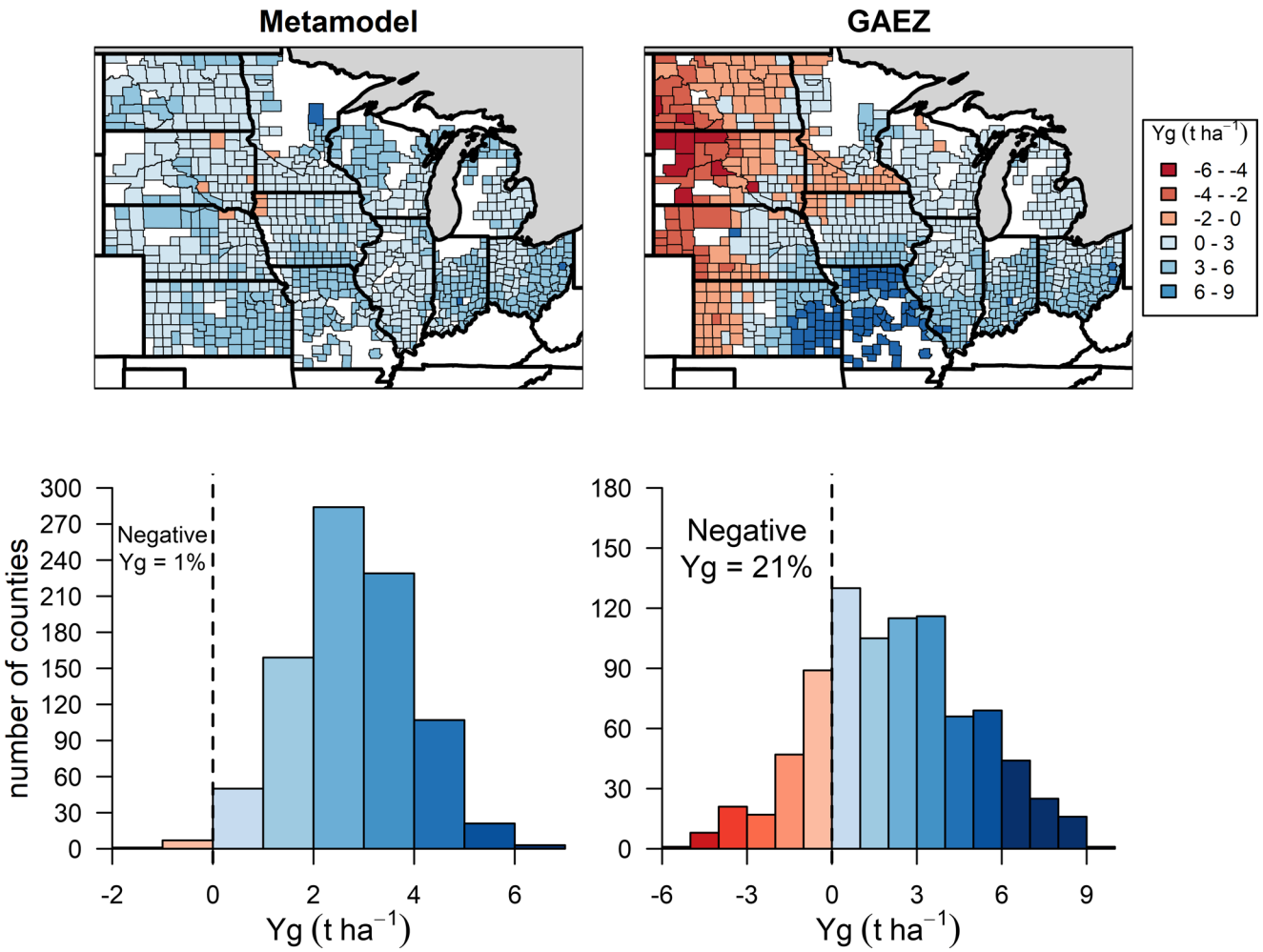
Extended Data Fig. 5 | Yield potential derived from different approaches. Comparison of maize water-limited yield potential derived from a top-down approach (GAEZ, gaez.fao.org)[5], a bottom-up approach (GYGA CZ, www.yieldgap.org), and a metamodel that integrates a bottom-up approach with machine learning (Metamodel) in East Africa.

https://doi.org/10.1038/s43016-024-01029-3



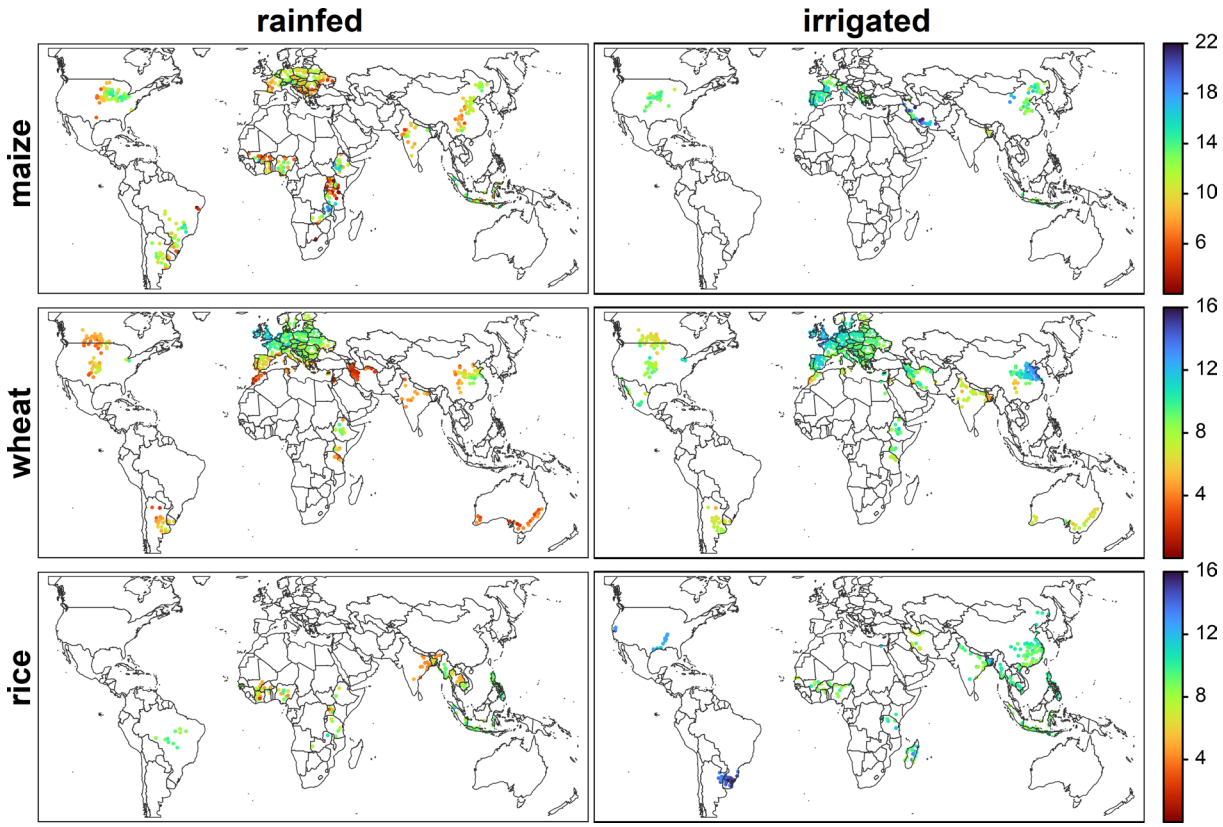
Extended Data Fig. 6 | Relation between annual precipitation and the yield potential derived from two approaches. Water-limited yield potential (Yw) of rainfed maize as a function of annual precipitation in East Africa for two yield potential prediction approaches: a metamodel that integrates a bottom-up approach with machine learning (Metamodel) and a top-down approach (GAEZ, gaez.fao.org)[5]. Each point represents a 5-arc-minute resolution grid with rainfed maize in East Africa. The red lines are local regression lines. Annual precipitation data was extracted from WorldClim (worldclim.org).

https://doi.org/10.1038/s43016-024-01029-3



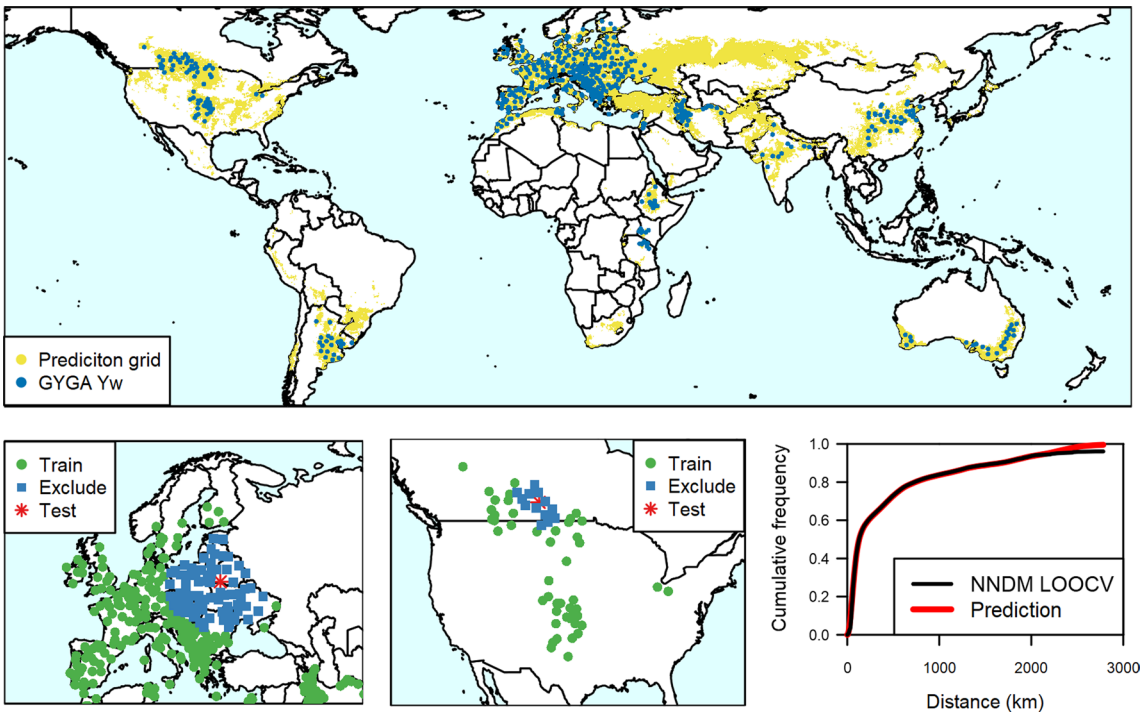
Extended Data Fig. 7 | Negative yield gap assessment. Yield gaps (Yg) between water-limited yield potential and farmers' actual yield for rainfed maize in the US Midwest at county level for two yield potential prediction approaches: a metamodel that integrates a bottom-up approach with machine learning (Metamodel) and a top-down approach (GAEZ, gaez.fao.org)[5]. Average county-level farmers' yield for rainfed maize between 2005 and 2015 was retrieved from USDA-NASS Quick Stats (quickstats.nass.usda.gov/). Only counties with less than 5% of irrigated area or reporting non-irrigated yields in 3 or more years were considered. In the histograms, the dashed vertical line indicates Yg = 0, that is, no difference between yield potential and actual yield, and the percentage of counties presenting negative Yg for each method is shown.

https://doi.org/10.1038/s43016-024-01029-3



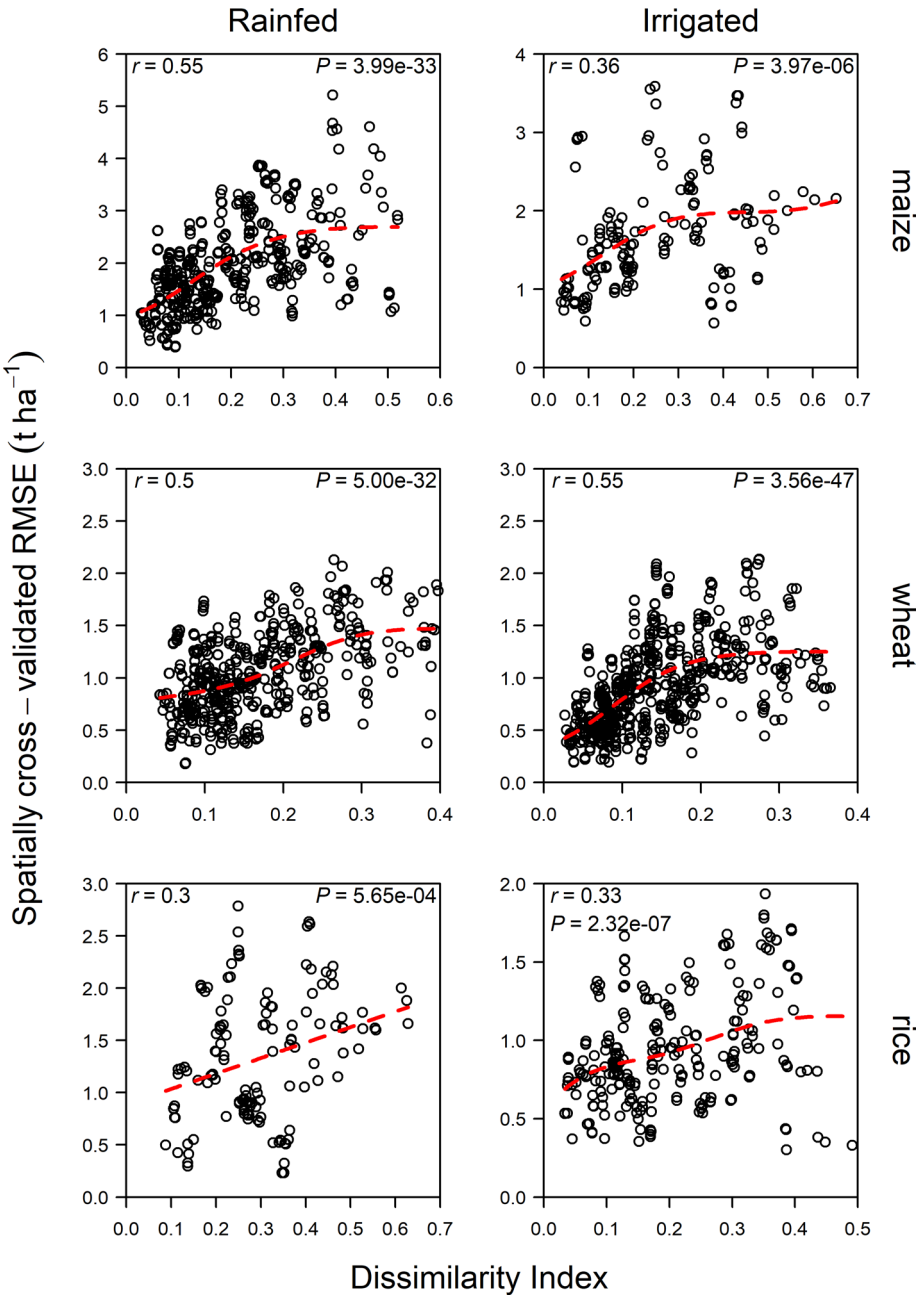
Extended Data Fig. 8 | Site-specific yield potential. Yield potential of irrigated crops and water-limited yield potential of rainfed crops reported in the Global Yield Gap Atlas (GYGA, www.yieldgap.org) at reference weather station level for the three main cereal crops. *Last access: July 10th, 2023.*

https://doi.org/10.1038/s43016-024-01029-3



Extended Data Fig. 9 | Nearest-neighbor-distance-matching leave-one-out cross-validation (NNDM LOO CV) examples for rainfed wheat. The top panel shows the distribution of site-specific water-limited yield potential (Yw) of rainfed wheat from the Global Yield Gap Atlas (GYGA) and the prediction grid (lands harvested with rainfed wheat as reported by SPAM[17]). The lower left and middle panels show the GYGA Yw sites used for model testing and training and excluded sites due to their proximity to the testing site for two iterations of the NNDM LOO CV[10]. The neighbors to be excluded are defined so that the cumulative frequency of distances between testing sites and their nearest training site in the NNDM LOO CV procedure matches the cumulative frequency of distances between the prediction grid cells and their nearest GYGA Yw, as shown in the lower right panel.

https://doi.org/10.1038/s43016-024-01029-3



Extended Data Fig. 10 | Relation between yield potential uncertainty and environmental dissimilarity. Relationship between spatially cross-validated root mean square errors (RMSE) and dissimilarity indexes between testing and training sites for each crop and water regime. Values were derived following the nearest-neighbor-distance-matching leave-one-out cross-validation (NNDM LOO CV) method[10] and the dissimilarity index used to estimate the area of applicability of the metamodel[11]. This association was used to estimate the expected RMSE of yield potential predictions from the dissimilarity index between the prediction area and the training sites. Pearson correlation coefficients (r) and their *P* values are shown.

nature portfolio

Corresponding author(s): Patricio Grassini

Last updated by author(s): Jun 26, 2024

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐	A description of all covariates tested
☒	☐	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Data on yield potential were downloaded from the Global Yield Gap Atlas website API with R and the packages httr (v1.4.7) and jsonlite (v1.8.7). Climatic data were downloaded with R using the geodata package (v0.5.9). Other datasets were downloaded from their web repositories.
Data analysis	All the analysis was done in R. We used the packages terra (v1.7.55) for spatial data analysis, data.table (v1.14.8) for data manipulation, ranger (v0.15.1) for Random Forest Regression (i.e., the machine learning algorithm), and caret (v6.0.94) and CAST (v0.8.1) for spatial cross-validation and area of applicability estimation. The R code used in the study is publicly available on GitHub (https://github.com/AramburuMerlos/gGYGA).

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

The high-resolution global maps of yield potential have been deposited in Zenodo (https://zenodo.org/doi/10.5281/zenodo.12209708)
All data used in this study is open data. Data on yield potential was downloaded from the Global Yield Gap Atlas website API (https://www.yieldgap.org/). Global climatic variables were downloaded from WorldClim (https://www.worldclim.org/). Soil data was downloaded from ISRIC (https://data.isric.org/). Crop calendar data was downloaded from SAGE, UW-Madison (https://sage.nelson.wisc.edu/data-and-models/datasets/crop-calendar-dataset/), RiceAtlas (https://www.nature.com/articles/sdata201774), and CropMonitor (chttps://cropmonitor.org/index.php/eodatatools/baseline-data/). Crop distribution maps were downloaded from SPAM (https://mapspam.info/)

Human research participants

Policy information about studies involving human research participants and Sex and Gender in Research.

Reporting on sex and gender	N/A
Population characteristics	N/A
Recruitment	N/A
Ethics oversight	N/A

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☐ Life sciences ☐ Behavioural & social sciences ☒ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Ecological, evolutionary & environmental sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	We integrated an agronomically robust bottom-up approach with machine learning to generate high-resolution global maps of yield potential for maize, wheat, and rice. Our machine learning metamodel leveraged site-specific yield potential derived from locally evaluated crop growth simulations and gridded climate, soil, and cropping system global databases.
Research sample	We trained our machine learning model with existing site-specific data on yield potential of maize, wheat, and rice from the Global Yield Gap Atlas. This dataset includes 543 (maize), 573 (wheat), and 325 (rice) sites distributed across 74 countries. Those sites were strategically selected to represent the harvested area distribution of each crop based on predefined climate zones and weather data availability.
Sampling strategy	To assess whether these data were sufficient to predict yield potential globally, we calculated the area of applicability of the machine learning model by considering the dissimilarity in biophysical properties between sites with data and global crop areas.
Data collection	We used existing databases of yield potential. Details about the protocols used to derive these yield potential estimates can be found in the Global Yield Gap Atlas (www.yieldgap.org) and references therein.
Timing and spatial scale	The Global Yield Gap Atlas provides yield potential estimates circa 2010 to 2020, depending on the country. We used machine learning to estimate yield potential globally, but we evaluated the accuracy of our results at local level.
Data exclusions	No yield potential data were excluded from the study.
Reproducibility	We provide access to the R code used for the analysis and raw data inputs.
Randomization	Not applicable as we did not conduct field experiments.

Blinding | Not applicable.

Did the study involve field work? ☐ Yes ☒ No

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging