

## Supplementary information

We utilized satellite remote sensing data focusing on climate and phenological variables relevant to grain crop yields, including evapotranspiration (ET), average surface temperature (AT), precipitation (Precip), leaf area index (LAI), vegetation index (NDVI), and tropospheric NO<sub>2</sub> monitoring data. ET and Ts were sourced from NASA's MODIS sensor datasets, specifically MOD16A2 and MOD11A2. MOD16A2, an 8-day evapotranspiration product, is produced by the Global Modeling and Assimilation Office and widely used in ecosystem and water resource management. MOD11A2 provides 8-day surface temperature data with a 1 km resolution, suitable for climate research, agriculture, and environmental monitoring. LAI data came from the MOD15A2H dataset, offering 8-day composite LAI and FPAR data at a 500 m resolution, broadly applied in vegetation research and ecosystem modeling. NDVI data were obtained from the MOD13Q1 dataset, which provides global 16-day vegetation indices, ideal for long-term vegetation change analysis. Precipitation data came from CHIRPS, a product developed by the Climate Hazards Group at the University of California, Santa Barbara, combining satellite imagery and ground station data with 0.05° spatial resolution, widely used in climate research and agricultural monitoring.

NO<sub>2</sub> emissions data were sourced from the Sentinel-5P satellite's TROPOMI module, provided by the European Space Agency. We used Level 3 data processed on the GEE platform, offering daily estimates since June 2018 with a resolution of 0.01° (~1 km). The TROPOMI algorithm separates stratospheric and tropospheric NO<sub>2</sub>, ensuring accuracy by removing stratospheric NO<sub>2</sub> influenced by solar cycles, leaving primarily anthropogenic emissions in the troposphere. N<sub>2</sub>O data were obtained from Emissions Database for Global Atmospheric Research (EDGAR), developed by RIVM and the EU's Joint Research Centre, aggregates emissions data from various international sources, including FAO, UNSD, IEA, and the World Bank. EDGAR's high spatial resolution and global coverage, combined with satellite validation by ESA, make it a reliable resource for

emissions assessments. We used agricultural sector N<sub>2</sub>O data from EDGAR v8.0 (updated to 2022).

Soil characteristics, including total nitrogen content (TNC), organic carbon stock (OCS), bulk density (Bdod), cation exchange capacity (CEC), soil organic carbon (SOC), clay, sand, and silt, were sourced from the SoilGrids dataset, a global soil information system developed by ISRIC. SoilGrids uses advanced machine learning to map soil properties across six depth intervals, with a spatial resolution of 250 m. It integrates over 230,000 soil profiles and 400 environmental covariates, producing global soil property maps with quantified uncertainties. These soil characteristics are provided in raster format across six depth intervals (0–5 cm, 5–15 cm, 15–30 cm, 30–60 cm, 60–100 cm, and 100–200 cm). The latest SoilGrids data have been updated to 2020.

Crop yield data for rice, maize, wheat, and soybean were derived from the SPAM dataset, developed by IFPRI. SPAM provides high-resolution global crop yield data at 5 arc-minute (~10 km) resolution, including detailed crop area, yield, and production statistics for 42 crops. SPAM's inputs include biophysical crop suitability assessments, land use data, and population density, processed through a cross-entropy method for precise estimation of crop distributions. The latest SPAM 2020 v1.0 data were released in April 2024.

Future climate scenarios were based on IPCC projections under the A1AIM and B1IMAGE scenarios for 2030 and 2050, using 2020 as the baseline. These projections come from the IPCC's Special Report on Emissions Scenarios (SRES), which explores different socio-economic development pathways and their associated emissions trajectories. We primarily focused on NO<sub>2</sub> emissions growth rates under these scenarios. Future climate scenarios were based on IPCC projections under the A1 and B1 scenarios. The SRES framework includes four "marker" scenarios: A1B-AIM, A2-ASF, B1-IMAGE, and B2-MESSAGE, along with two additional scenarios from the A1 family: A1G-MiniCAM (A1FI) and A1T-MESSAGE. The A1 family assumes rapid economic growth, a peak in global population around the mid-21st century, and swift technological

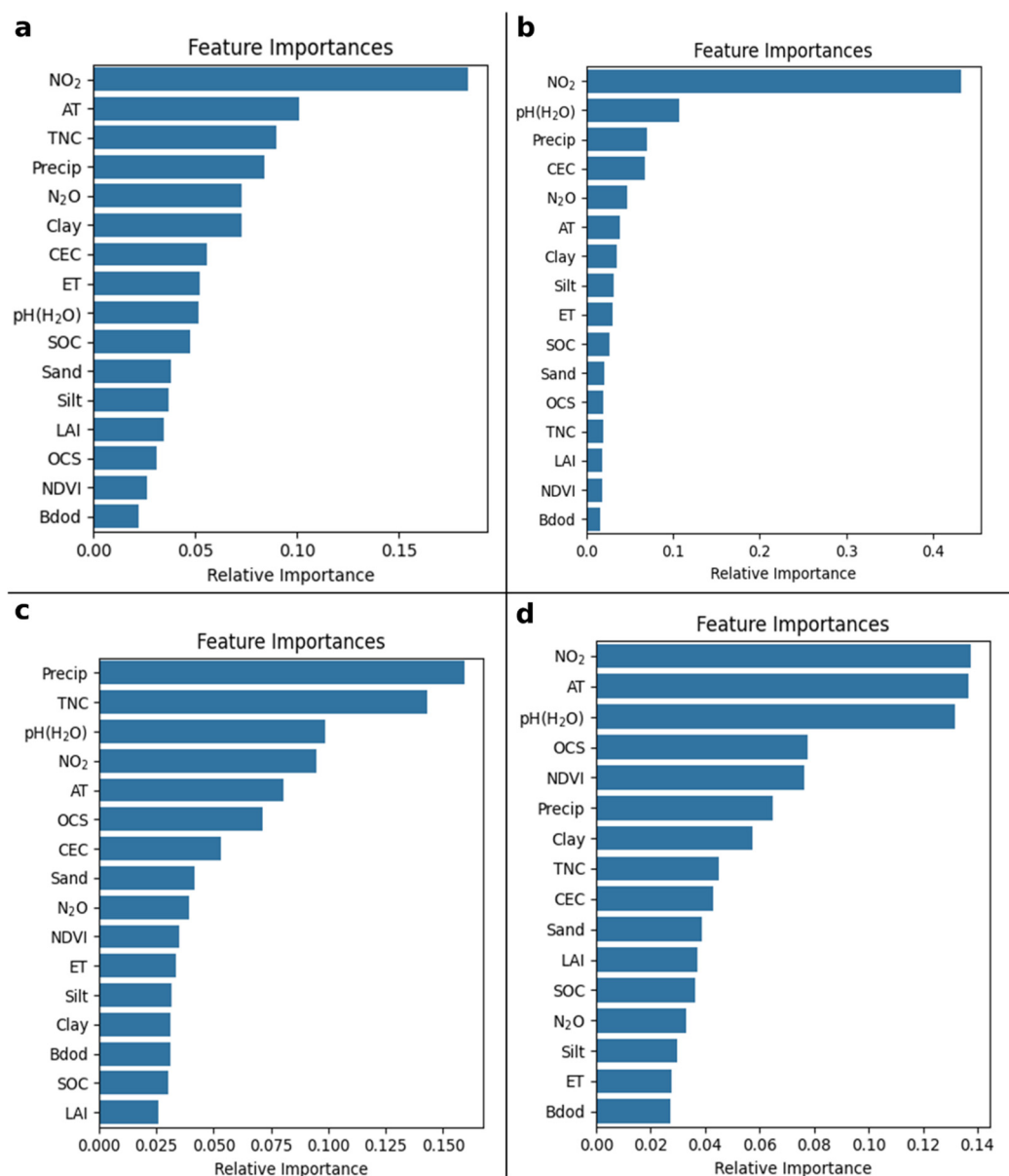
advancement. The A1 scenarios are further subdivided based on energy use: A1B-AIM assumes a balanced approach using both fossil and non-fossil energy sources; A1G-MiniCAM (A1FI) is fossil fuel-intensive; A1T-MESSAGE prioritizes non-fossil energy technologies. The B1 family, on the other hand, assumes a world with an emphasis on clean and resource-efficient technologies, along with a peak in global population by mid-century.

Both XGBoost and Random Forest Regression (RFR) are widely used ensemble learning techniques for regression and classification tasks. However, they differ substantially in their underlying mechanisms and application. Random Forest Regression is a classical supervised algorithm that builds multiple decision trees by sampling subsets of the data. The method introduces randomness in two ways: first by bootstrapping data samples, and second by randomly selecting a subset of features at each node to split. Each decision tree is trained independently, and the final prediction is obtained by averaging the outputs of all trees. This method reduces the risk of overfitting while enhancing model generalization. Additionally, RFR can handle missing data naturally and provides insights into feature importance by evaluating the contribution of each feature during the splitting process. In contrast, XGBoost follows a boosting approach, where each new tree corrects the residual errors made by the previous ones. XGBoost improves prediction accuracy by sequentially adding trees that target the mistakes of prior models. The algorithm employs regularization techniques (L1 and L2) to control model complexity and mitigate overfitting, making it particularly suited for large-scale datasets. However, it requires careful tuning of hyperparameters, and the training process is more computationally intensive since each tree depends on the residuals of its predecessor.

For this study, we applied both RFR and XGBoost to model crop yields, as these algorithms are well-suited for analyzing complex interactions between environmental factors and agricultural outputs. RFR allowed us to assess feature importance, while XGBoost provided high predictive accuracy through iterative model refinement. These

models enabled a comprehensive evaluation of how TNC and NO<sub>2</sub> emissions influence the yields of major grain crops in China.

### Supplementary figure



**Supplementary Figure S1. Importance of variables in crop yield models based on nitrogen cycling.** **a**, Importance analysis of the RFR crop yield model for maize. **b**, Importance analysis of the RFR crop yield model for rice. **c**, Importance analysis of the RFR crop yield model for soybean. **d**, Importance analysis of the RFR crop yield model for wheat. We constructed the RFR model using the sklearn library in Python (Version 3.12.3); relevant data are detailed in Data availability. The number of valid random sample points was Maize = 662, Rice = 890, Soybean = 1063, Wheat = 843.