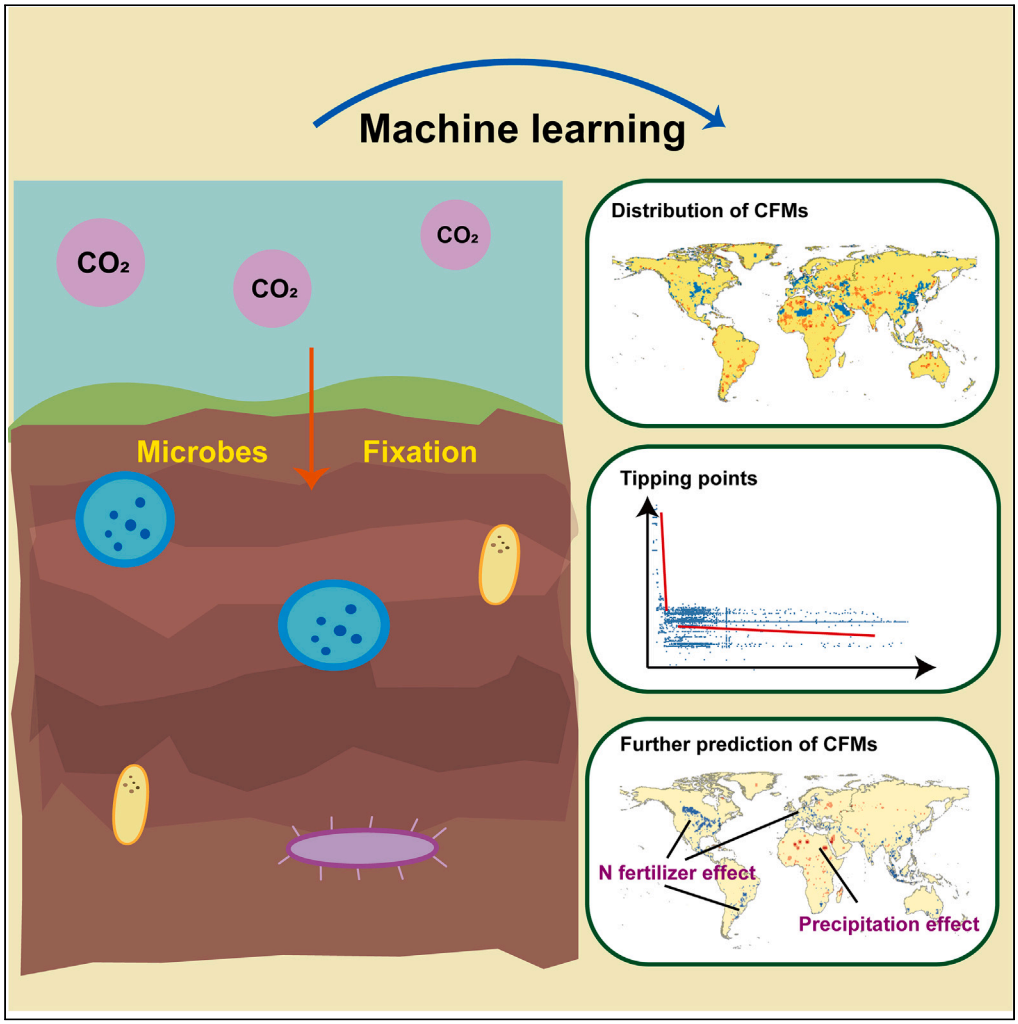
iScience

CellPress
OPEN ACCESS

Article

# Environmental tipping points for global soil carbon fixation microorganisms




Yueqi Hao, Hao Liu, Jiawei Li, Li Mu, Xiangang Hu

muli@caas.cn (L.M.)
huxiangang@nankai.edu.cn (X.H.)

## Highlights

Total C, N fertilizer and precipitation are the key factors for C-fixing microbes

The N fertilizer tipping point for C-fixing microbes is 9.45 kg $ha^{-1} \cdot y^{-1}$

The precipitation tipping point for C-fixing microbes is 22.38 mm

N fertilizer causes 46% of cropland to have decline in C-fixing microbes in 2100

Hao et al., iScience 26, 108251
November 17, 2023 
https://doi.org/10.1016/j.isci.2023.108251

iScience

CellPress
OPEN ACCESS

Article

# Environmental tipping points for global soil carbon fixation microorganisms

Yueqi Hao,[1,3] Hao Liu,[1,3] Jiawei Li,[1] Li Mu,[2,4,*] and Xiangang Hu[1,*]

## SUMMARY

**Carbon fixation microorganisms (CFMs) are important components of the soil carbon cycle. However, the global distribution of CFMs and whether they will exceed the environmental tipping points remain unclear. According to the machine learning models, total carbon content, nitrogen fertilizer, and precipitation play dominant roles in CFM abundance. Obvious stimulation and inhibition effects on CFM abundance only happened at low levels of total carbon and precipitation, where the tipping points were 6.1 $g \cdot kg^{-1}$ and 22.38 mm, respectively. The abundance of CFMs in response to nitrogen fertilizer changed from positive to negative (tipping point at 9.45 kg $ha^{-1} \cdot y^{-1}$). Approximately 46% of CFM abundance decline happened in cropland at 2100. Our work presents the distribution of carbon-fixing microorganisms on a global scale and then points out the sensitive areas with significant abundance changes. The previously described information will provide references for future soil quality prediction and policy decision-making.**

## INTRODUCTION

Carbon fixation microorganisms (CFMs) in soils are critical contributors to ecosystem function since they convert atmospheric $CO_2$ into soil carbon.[1] Microbe-associated autotrophic carbon fixation provides approximately 2–3 $\times 10^{15}$ g $C \cdot y^{-1}$ in terrestrial areas of the world.[2] CFMs are directly associated with the soil fixation process and further affect soil carbon storage and the fertility of terrestrial ecosystems; moreover, they can offset global warming.[3] Autotrophic microorganisms have six pathways capable of fixing atmospheric $CO_2$ in soil, where the Calvin cycle and rTCA cycle are the most common pathways.[4,5] $CO_2$ fixation pathways in grassland soils are mainly influenced by precipitation.[6] A low-nutrient environment stimulates microbial carbon fixation capacity in soil.[7] Other factors, such as pH and total nitrogen, have also been proven to be related to microbial carbon fixation.[4,7,8] Additionally, the results and conclusions of previous studies may have uncertainty in real environments on a global scale.[3,9,10] The isotopic tracer technique is expensive and time-consuming, which creates problems for global research on microbial carbon fixation. In addition, the nonlinear relationship between biotic carbon fixation and multiple environmental factors makes the existence of tipping points possible.[11] Small changes in environmental or climatic factors can lead to abrupt changes or relationship reversals in ecosystems.[12] Whether and how CFMs will reach any tipping points remain unclear.

Machine learning is a data-driven modeling approach with a high capability for learning complicated patterns[13,14] that can be used to identify crucial environmental factors, organize complex relationships, and predict further trends.[15,16] Then, the conditions and locations of tipping points may be identified. The identification of tipping points for CFM abundance is urgently needed to maintain soil carbon storage and to mitigate climate change.

To solve these problems, we built a database containing 1726 observations of soil microorganism samples from 640 locations worldwide (Figure S1). The relative abundance of CFMs in the soil was taken as the research object. Fourteen environmental factors that may threaten soil CFMs were collected based on previous articles.[8,17] The factors included those related to climate (e.g., mean annual temperature [MAT] and mean annual precipitation [MAP]), soil properties (e.g., pH, soil total carbon [TC], soil organic matter [OM], soil total N [TN], soil total P [TP], soil water [SW], organic carbon [OC], and texture), and agricultural management (e.g., use of N/P/K fertilizers). A workflow for predicting the global distribution pattern of soil CFMs is presented in Figure 1A. Based on the built machine learning model, we screened out the critical factors with tipping points. Tipping points are determined as the points above or below which abrupt changes or relationship reversals happen, in this case, changes in soil CFM abundance. To explore the mechanisms underlying the occurrence of tipping points, we proposed microbial tolerance on a global scale (Figure 1C). Microbial tolerance denotes the proportion of tolerant species in the microbial community under extreme disturbance (e.g., excessive temperature or pH).[18] The higher microbial tolerance is, the less suitable the local environment is for microbes to survive.[19] This work predicted the

[1]Key Laboratory of Pollution Processes and Environmental Criteria (Ministry of Education)/Tianjin Key Laboratory of Environmental Remediation and Pollution Control, College of Environmental Science and Engineering, Nankai University, Tianjin 300080, China
[2]Key Laboratory for Environmental Factors Control of Agro-product Quality Safety (Ministry of Agriculture and Rural Affairs), Tianjin Key Laboratory of Agro-environment and Safe-product, Institute of Agro-environmental Protection, Ministry of Agriculture and Rural Affairs, Tianjin 300191, China
[3]These authors contributed equally
[4]Lead contact
*Correspondence: muli@caas.cn (L.M.), huxiangang@nankai.edu.cn (X.H.)
https://doi.org/10.1016/j.isci.2023.108251

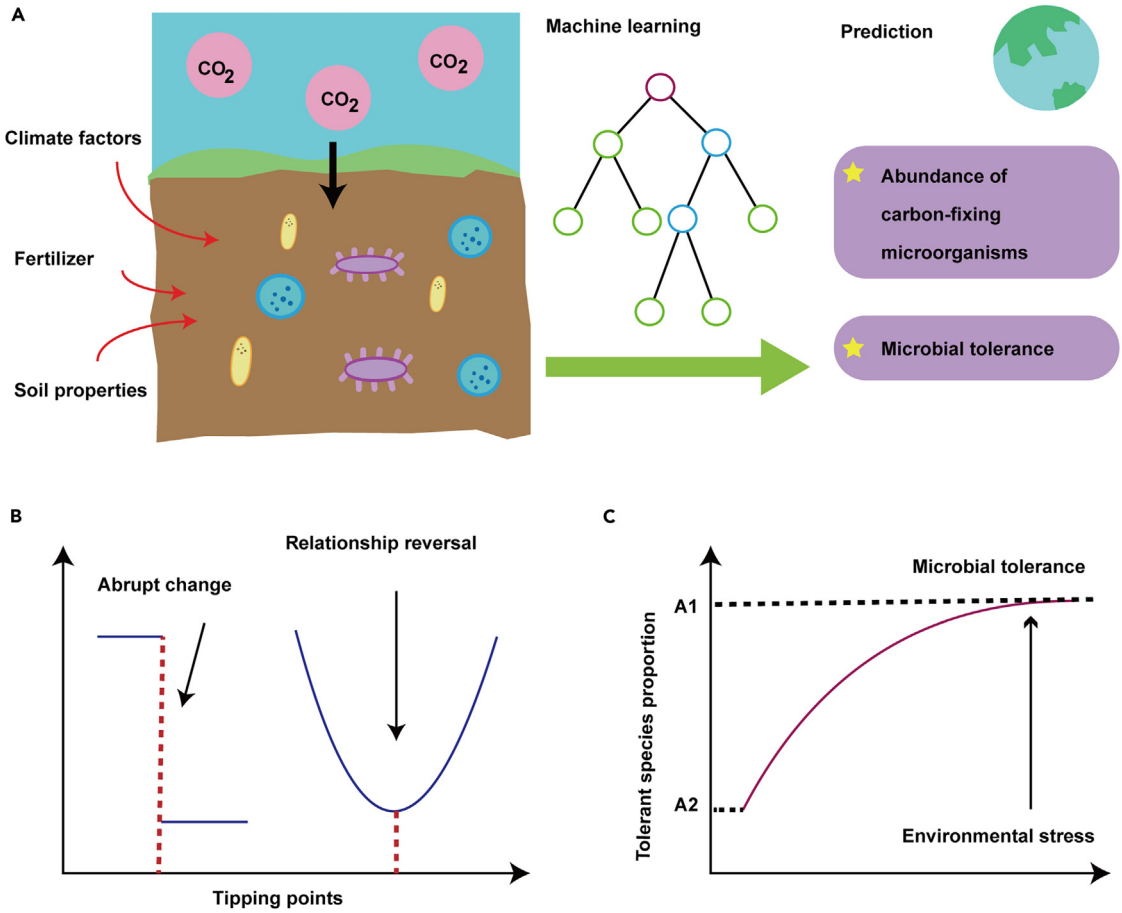



**Figure 1. Workflow and prediction of carbon fixation microorganisms (CFMs) and microbial tolerance**

(A) Workflow for predicting the global abundance distribution pattern of soil CFM abundance on a global scale.

(B) Diagram of tipping point types representing abrupt changes and relationship reversals.

(C) Microbial tolerance is the proportion of tolerant species in the microbial community under extreme conditions (e.g., excessive temperature or extreme pH). The details for the calculation are provided in the STAR Methods section.

current and future distribution and development of microbial carbon fixation in soil and identified the crucial factors and ecological tipping points affecting CFM abundance. Our work provides quantitative information for understanding soil carbon sinks and mitigating climate change driven by microbes.

## RESULTS

### Distribution patterns of global soil carbon fixation microorganisms

CFMs are key components of the soil carbon cycle.[7] The global distribution pattern of CFM abundance provides information for predicting future soil carbon storage and climate change.[1,20] A dataset containing 1726 sets of soil microbial data from 640 locations worldwide was mined and analyzed (Figure S1). Fourteen environmental factors related to CFM abundance[21–23] were collected, including MAT, MAP, pH, TC, OM, TN, TP, SW, OC, texture, and soil management (i.e., use of N/P/K fertilizers). Moreover, the microbial capacity to tolerate extreme environmental conditions was analyzed to explain the changes in the soil microbial communities (Figures 1C and 1D).

Different machine learning modeling methods were tested (Figure S2A). The training set $R^2$ values of the K-nearest neighbors, random forest, decision tree, linear regression, and ordinary least-squares regression methods were 0.94, 0.93, 0.86, 0.21, and 0.02, respectively. XGBoost performed the best among the five tested machine learning models. The training set coefficient ($R^2$) of XGBoost was 0.94. Here, we chose XGBoost as the machine learning model for the subsequent investigations. The test set $R^2$ for the XGBoost model was 0.61, and the normalized root-mean-square error was 0.1132. Model robustness evaluation was performed using new datasets together with adversarial samples to attack the model. The model was still able to maintain an $R^2$ of 0.61. The percentage of data points within the 95% confidence interval was 82.14 $\pm$ 5.90%. The mean absolute percentage error was 19.05 $\pm$ 1.96% (Figures S2B–S2E). Given the complexities of large-scale ecological modeling, our model performance is acceptable compared to other ecological models, which have $R^2$ values less than

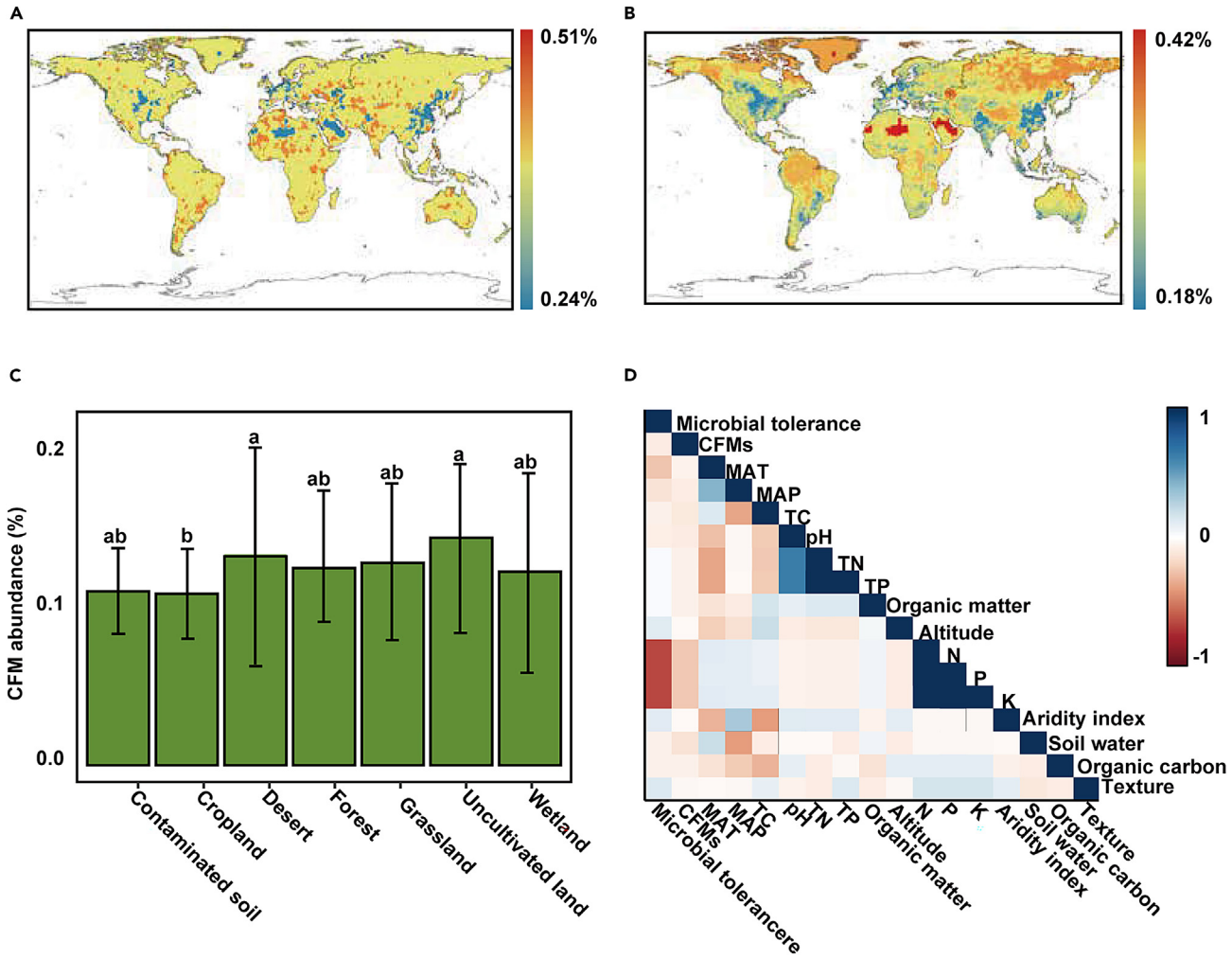



**Figure 2. Prediction of global soil carbon fixation microorganisms (CFMs) and microbial tolerance in 2021**

(A) Global distribution of CFM abundance in 2021; (B) Global distribution of microbial tolerance in 2021; (C) CFM abundance in different habitats,different letters indicate significant differences between columns ($p<0.05$); (D) Pearson correlation coefficients between the independent and dependent variables of the model.

0.6 and uncertainty values of more than 20%.[24,25] The percentage of CFM and tolerance points within the 95% confidence interval was higher than 80%. Moreover, the predictive performance for different regions (e.g., different continents) was replicable at the global scale, indicating that our model is stable.

Based on the previously established XGBoost model, the relative abundance of global CFMs was determined, as shown in Figure 2A. The relative abundance range of global CFMs was 0.24%–0.51%. The distribution of CFM abundance was dispersed (Figure 2A), although 66.71% and 46.35% of the high (top 10%) and low (bottom 10%) CFM abundance areas occurred in cropland, respectively. Some studies have proposed that habitat is an important factor influencing soil carbon fixation.[7] Figure 2B also shows that the abundance of CFMs in desert and uncultivated lands was significantly higher than that in croplands ($p < 0.05$). This pattern is related to the response of CFMs to nitrogen fertilizer and carbon content. The detailed relationships of nitrogen fertilizer and TC with CFMs are analyzed in the following paragraphs. Microbial tolerance is the abundance of tolerant species that can survive in extreme environments and is positively correlated with the level of severity of the environmental conditions.[26] Hot and cold spots of microbial tolerance were found in the temperate zone and cropland areas (Figure 2B). In addition, microbial tolerance had obvious hot spots in North Africa and West Asia (Figure 1H).

## Tipping points of critical factors for carbon fixation microorganisms

To assess the important factors influencing the CFMs community and microbial tolerance, we calculated SHAP values for 14 factors. The SHAP values estimate the contribution of each factor to the model output.[27] The heatmap of mean SHAP values (Figure S3A) shows that soil TC, nitrogen fertilizer application (N), and MAP are the most important factors affecting soil CFMs. The contributions of these three factors to CFMs on a global scale were all more than 50% (Figure 3A). In addition to TC, N, and MAP, MAT was a crucial factor affecting

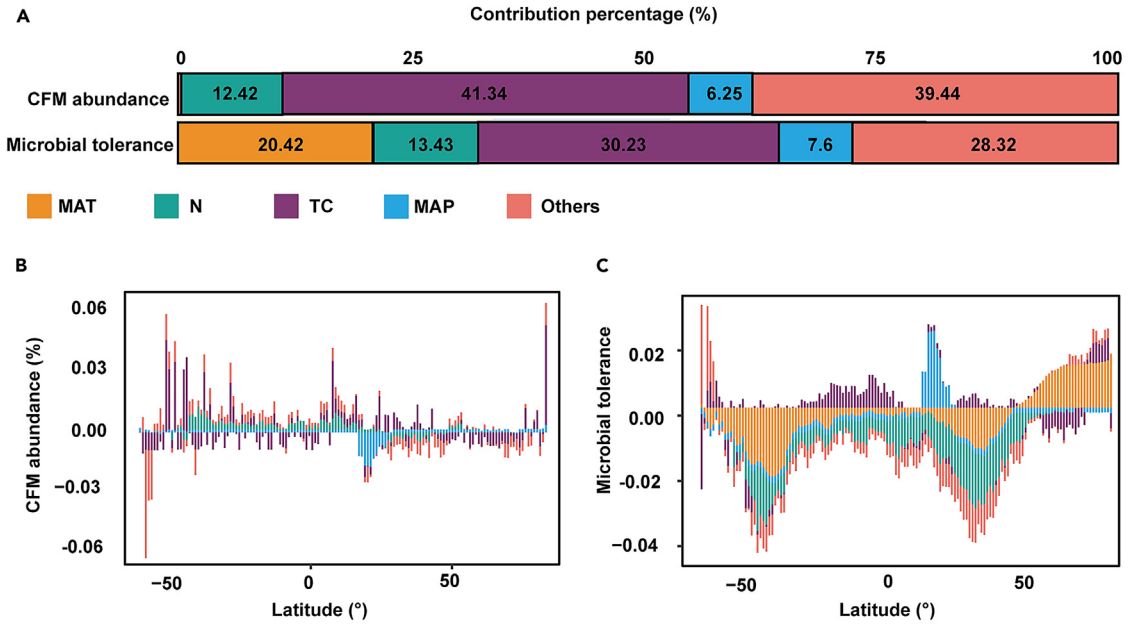



**Figure 3. Quantitative contribution of critical environmental factors to soil carbon fixation microorganisms (CFMs) and environmental tolerance in 2021**

(A) Quantitative contribution of each critical factor. SHAP value distribution of critical factors for CFMs (B) and environmental tolerance (C) with latitude.

microbial tolerance (Figures 3A–3C). Our work considered the minor (0.36%) contribution of temperature to soil CFMs on a global scale based on data from field studies. The response pattern also showed that the abundance of CFMs does not change with temperature (Figure S5A).

In our study, the screening of critical factors laid the foundation for tipping point analysis. An ecological tipping point is an abrupt change and relationship reversal in one ecological feature caused by a minor change in critical factors under special conditions[12] that becomes the premise of subsequent management. We found that a significant stimulation effect of soil TC content on CFM abundance only occurred when TC was low and that the tipping point was 6.1 g $kg^{-1}$ (Figure 4A). Low carbon content soil (<10 g $kg^{-1}$) accounted for 11.7% of the terrestrial region (Figure S6A). The areas of low TC in Africa, Oceania, and Asia accounted for 27.8%, 17.2%, and 9.6%, respectively.

The response of CFMs to nitrogen fertilizer also exhibited a trend of first increasing and then declining. The tipping point of CFM abundance in response to N fertilizer was at 9.45 kg $ha^{-1}\cdot y^{-1}$, and the CFM abundance remained stable after crossing the tipping point at 103.65 kg $ha^{-1}\cdot y^{-1}$ (Figure 3D). Areas with N fertilizer had CFM abundances lower than average (Figure 2A; the global N fertilizer application is shown in Figure S6B). Excessive application of nitrogen fertilizer may lead to a decline in soil carbon fixation rate. The response of microbial tolerance to nitrogen fertilizer consistently showed a downward trend (Figure 3F), suggesting that high nitrogen fertilization weakens adaptability to environmental changes. The tipping point of microbial tolerance to nitrogen fertilization was 10.06 kg $ha^{-1}\cdot y^{-1}$ at the global scale.

MAP made important contributions to CFM abundance in the areas from 0 to 10 N° (Figure S5E). In extreme areas with annual precipitation less than 22.38 mm, biotic carbon fixation was reduced. The combined effect of low TC (10 < g·$kg^{-1}$) and low nitrogen fertilizer (<50 kg $ha^{-1}\cdot y^{-1}$) resulted in the highest CFM abundance (Figure 4C). MAP under 20 mm limited the effect of TC and N fertilizer on CFM abundance (Figures 4E and 4F). Precipitation is an important limiting factor of autotrophic carbon fixation in barren areas.

## Prediction of global carbon fixation microorganisms in future climate scenarios

Figure 5 shows the changing proportion of CFMs from 2021 to 2100. The increases in temperature and nitrogen fertilizer were determined according to the high emissions and intensive agriculture scenarios in Climate Model Intercomparison Project Phase 6.[28,29] Precipitation and nitrogen fertilizer use will increase by approximately 1.1 and 1.6 times by 2100, respectively. Future precipitation changes produce hot spots only in North Africa and West Asia, accounting for 0.5% of global terrestrial soil (Figure 5A). Global climate change has little impact on soil microbial carbon fixation, and the abundance of CFMs increases after the precipitation limit is crossed in barren areas in North Africa and West Asia. Nitrogen fertilizer is the dominant factor determining further CFM development. Under the scenario with the combined effects of precipitation and nitrogen fertilizer, the relative change in CFM abundance was similar to that with the effect of nitrogen fertilizer alone (Figures 5B and 5C). Croplands in the Americas, Europe, and Southeast Asia appear to be cold spots for CFM abundance changes. The cropland CFM abundance in North America, South America, and Oceania decreases by 2.85%, 1.98%, and 2.58%, respectively. The countries with CFM abundance increases higher than 5% are presented in Figure 5D.

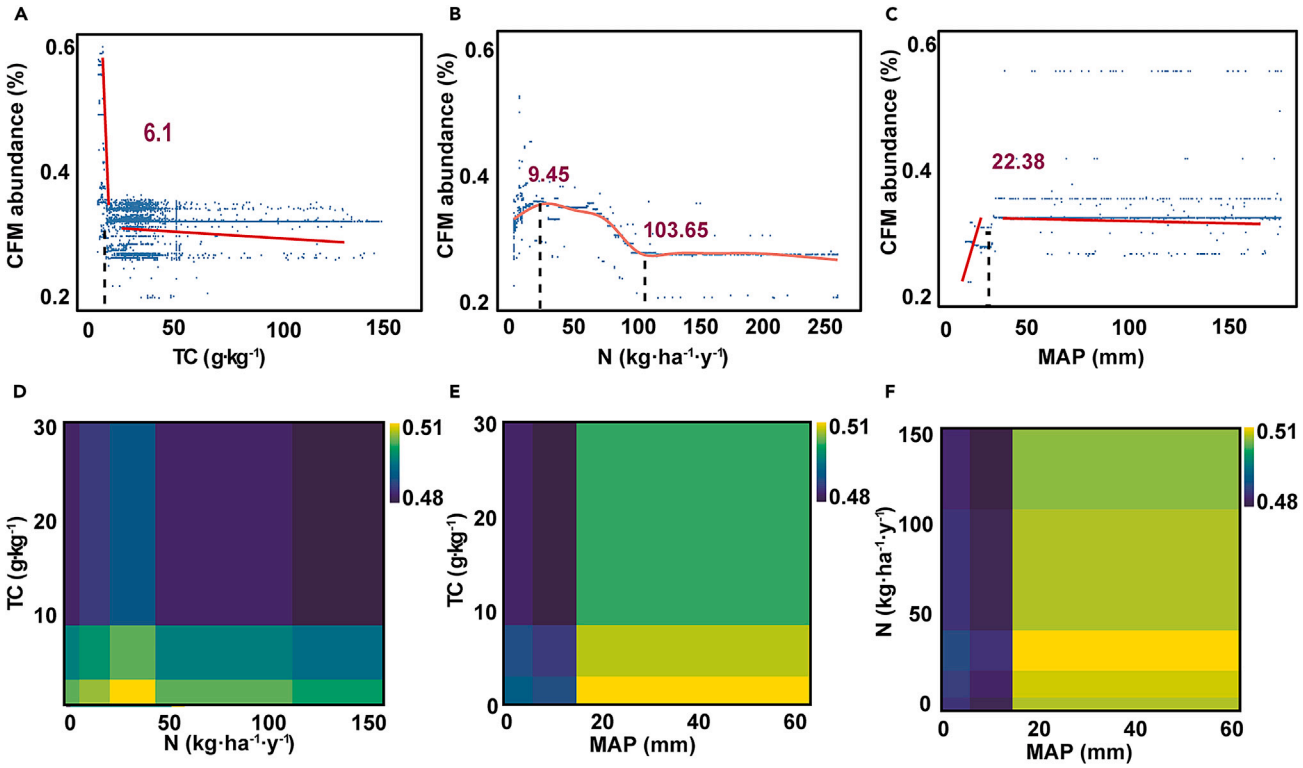



**Figure 4. Identification of tipping points and double-variable partial dependence of carbon fixation microorganisms (CFMs)**

(A) The response of CFM abundance to TC; (B) the response of CFM abundance to N; (C) the response of CFM abundance to MAP; (D) double-variable partial dependence of CFM abundance on TC and N; (E) double-variable partial dependence of CFM abundance on TC and MAP; (F) double-variable partial dependence of CFM abundance on N and MAP. Black dashed lines and numbers in purple font represent the identified tipping points; the pink line represents the smoothed trend fitted by the generalized additive model (GAM), and the red lines are the fitted lines obtained from the segmented linear regression (SLR) model.

## DISCUSSION

XGBoost provides a parallel tree boosting to achieve efficiency, flexibility, and portability.[30] The uncertainty assessment also proves the stability of the model. Based on the proposed XGBoost model, present and further CFMs abundances can be efficiently attained. Autotrophic microorganisms explained approximately 56% of the variation in $CO_2$ fixation in soil.[4] Prediction of global CFM distribution will provide a foundation for identification of the biochemical cycle of global carbon and climate change. The distribution of CFMs demonstrates the imbalance of $CO_2$ fixation rates over the world, where cropland, arid, and barren regions are the places we should pay attention to.

Approximately half of the hot and cold spots of CFM abundance occurred in cropland, which is related to the influence of N fertilizer on CFMs. A small amount of nitrogen addition can stimulate the growth of autotrophic microorganisms,[31] while a large amount of nitrogen addition can play a negative role in the microbial community due to the inhibition of nondiazotrophs and to community complexity.[32,33] Considering the stimulatory effect of nutrient addition to soil, the overuse of N fertilizer increases the sensitivity of soil carbon and produces a positive feedback effect on global warming.[34] Effective N fertilizer management ensures the optimal trade-off between yield and fertilizer usage. It was proposed that the combined application of half inorganic N plus half organic N might have the potential to enhance soil C sequestration in cropland.[35] Manure fertilization increased the abundance of functional genes involved in the rTCA cycle.[36] Nitrogen fertilizer has also become a decisive factor in the future global distribution of CFMs (Figures S5B and S5C). Nitrogen fertilizer in the Americas, Europe, and Southeast Asia is applied with concentrations of 10–110 kg $ha^{-1}\cdot y^{-1}$; in this range, CFM abundance decreases with increasing nitrogen fertilizer. Considering that more than 60% of global cropland is supplemented with nitrogen fertilizer at rates higher than 10 kg $ha^{-1}\cdot y^{-1}$, the addition of nitrogen fertilizer has a negative feedback effect on global CFMs. A negative impact will occur on global soil carbon storage and atmospheric $CO_2$ content.

The impact of global warming on future CFMs is relatively small due to the weak response of CFMs to temperature. The important influence of temperature on soil microbes is widely known[9,22] but is different from data from previous local studies or laboratory experiments.[10] It was proposed that temperature is not the dominant factor in the carbon fixation process and that precipitation plays an important role in controlling $CO_2$ fixation.[4] Temperature affects heterotrophic processes more strongly than it affects autotrophic processes.[37] A reduction in autotrophic CFMs may neutralize the positive effect of temperature on carbon fixation.[37]

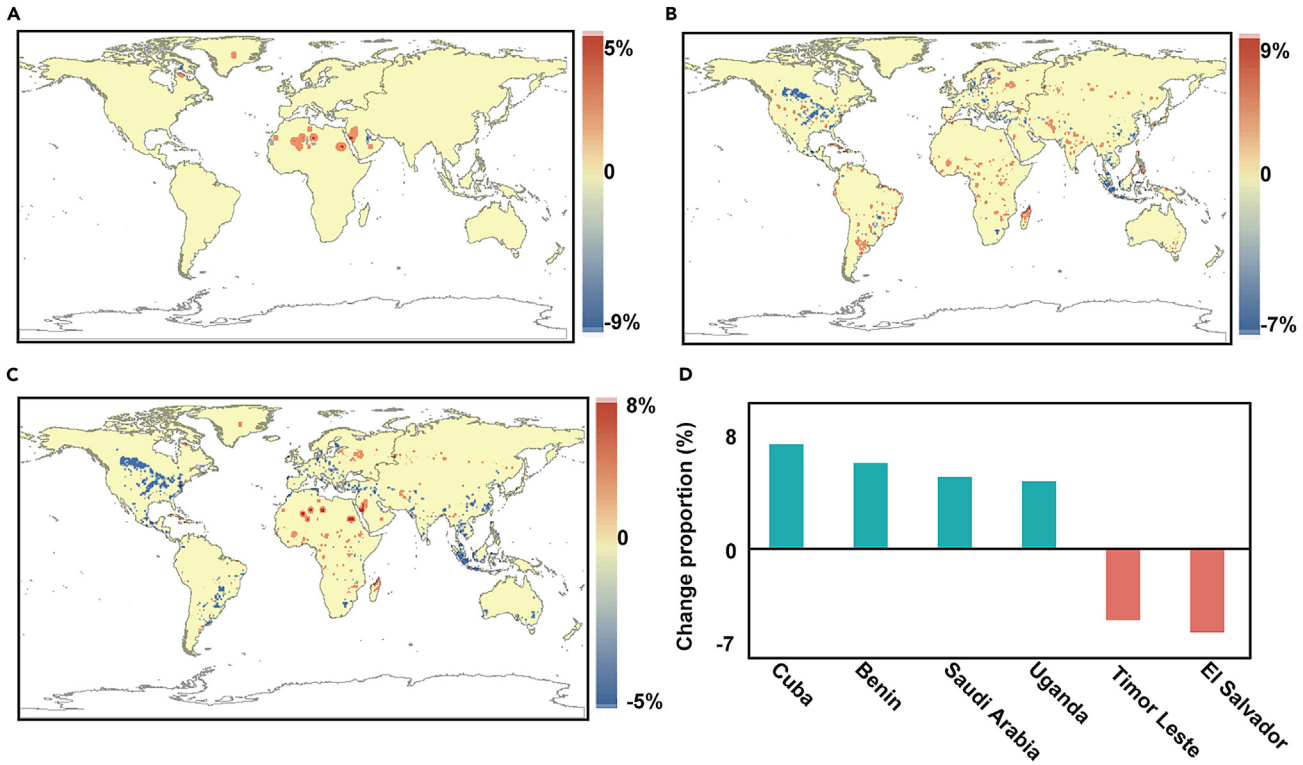



**Figure 5. Relative change in carbon fixation microorganisms (CFMs) from 2021 to 2100**

(A) Relative change in carbon fixation microorganism (CFM) abundance due to precipitation from 2021 to 2100; (B) relative change in CFM abundance due to N fertilizer from 2021 to 2100; (C) relative change in CFM abundance due to precipitation and N fertilizer from 2021 to 2100; (D) relative change in CFM abundance from 2021 to 2100 in different countries.

The existence of tipping points in TC and MAP makes the CFM abundance in barren and arid regions worth noting. It has been proposed that the microbial $CO_2$-fixation efficiency can be higher than that of plants under barren conditions,[7] but the likelihood of such a relationship and quantitative information at global or regional scales remain unclear. When the soil carbon content exceeds the tipping point, autotrophic microbes no longer play a prevalent role and exhibit a decreasing trend relative to the whole microbial community.[2] A higher adaptability and growth rate of microbes critically contribute to carbon fixation in barren areas.[7] Thus, TC deserves much attention in the aforementioned regions as a soil microbial carbon sink. In low precipitation regions, the abundance of autotrophic microbes was also lower.[38] The $CO_2$ fixation rates by CFMs in wetland, arid grasslands, and desert soils were 85, 22, and 6.4 mg C $m^{-2}$ $d^{-1}$, respectively,[6] suggesting that precipitation is the limiting factor of soil carbon fixation efficiency.

## Conclusion

Elucidating the potential and trend of microbial-mediated carbon fixation in soil is important for understanding ecosystem function and global climate change. Through machine learning, we screened out TC, nitrogen fertilizer, and precipitation as the key factors triggering the tipping points of soil CFM abundance. According to our research, cropland presents a wide range of CFM abundance. Appropriate application of nitrogen fertilizer (near tipping points) may maximize the carbon fixation ability of soil CFMs. Then, a positive feedback effect on soil carbon storage and greenhouse gas reductions may be achieved. Precipitation is the limiting factor for CFM abundance in barren land. Precipitation increases will relieve the carbon fixation limitation in North Africa and West Asia, and vice versa. Our study provides quantitative information on a global scale to assist in soil management.

## Limitations of the study

We collected soil microbial data from related articles based on 16S rRNA sequencing technology; more soil metadata can be pulled from soil metagenomes in further research.

## STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

## SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.isci.2023.108251.

## ACKNOWLEDGMENTS

This work was financially supported by the National Key Research and Development Program of China (grant no. 2019YFC1804603), National Natural Science Foundation of China (grant nos. 22176103 and U22A20615) and the Ministry of Education of China (grant no. T2017002) , and the Central Public Research Institutes Basic Funds for Research and Development (Institute of Agro-Environmental Protection, Ministry of Agriculture and Rural Affairs, China).

## AUTHOR CONTRIBUTIONS

Conceptualization, L.M. and X.H.; methodology, H.L. and Y.H.; investigation, Y.H. and J.L.; writing – original draft, Y.H. and H.L.; writing – review and editing, L.M. and X.H.

## DECLARATION OF INTERESTS

The authors declare no competing financial interests.

Received: March 15, 2023
Revised: July 18, 2023
Accepted: October 16, 2023
Published: October 27, 2023

## REFERENCES


1. Hartley, I.P., Hill, T.C., Chadburn, S.E., and Hugelius, G. (2021). Temperature effects on carbon storage are controlled by soil stabilisation capacities. Nat. Commun. *12*, 6713–6717.
2. Lynn, T.M., Ge, T., Yuan, H., Wei, X., Wu, X., Xiao, K., Kumaresan, D., Yu, S.S., Wu, J., and Whiteley, A.S. (2017). Soil Carbon-Fixation Rates and Associated Bacterial Diversity and Abundance in Three Natural Ecosystems. Microb. Ecol. *73*, 645–657.
3. Crowther, T.W., van den Hoogen, J., Wan, J., Mayes, M.A., Keiser, A.D., Mo, L., Averill, C., and Maynard, D.S. (2019). The global soil community and its influence on biogeochemistry. Science *365*, eaav0550.
4. Liao, H., Hao, X., Qin, F., Delgado-Baquerizo, M., Liu, Y., Zhou, J., Cai, P., Chen, W., and Huang, Q. (2023). Microbial autotrophy explains large-scale soil $CO_2$ fixation. GCB *29*, 231–242.
5. Zheng, Z., Liu, B., Fang, X., Fa, K., and Liu, Z. (2022). Dryland farm soil may fix atmospheric carbon through autotrophic microbial pathways. Catena *214*, 106299.
6. Huang, Q., Huang, Y., Wang, B., Dippold, M.A., Li, H., Li, N., Jia, P., Zhang, H., An, S., and Kuzyakov, Y. (2022). Metabolic pathways of $CO_2$ fixing microorganisms determined C-fixation rates in grassland soils along the precipitation gradient. Soil Biol. Biochem. *172*, 108764.
7. Chen, H., Wang, F., Kong, W., Jia, H., Zhou, T., Xu, R., Wu, G., Wang, J., and Wu, J. (2021). Soil microbial $CO_2$ fixation plays a significant role in terrestrial carbon sink in a dryland ecosystem: A four-year small-scale field-plot observation on the Tibetan Plateau. Sci. Total Environ. *761*, 143282–143287.
8. Buchanan, P.J., Chase, Z., Matear, R.J., Phipps, S.J., and Bindoff, N.L. (2019). Marine nitrogen fixers mediate a low latitude pathway for atmospheric $CO_2$ drawdown. Nat. Commun. *10*, 4611–4710.
9. Canfield, D.E., Glazer, A.N., and Falkowski, P.G. (2010). The Evolution and Future of Earth's Nitrogen Cycle. Science *330*, 192–196.
10. Jiang, J., Li, Z., Xiao, H., Wang, D., Liu, C., Zhang, X., Peng, H., and Zeng, G. (2018). Labile organic matter plays a more important role than the autotrophic bacterial community in regulating microbial $CO_2$ fixation in an eroded watershed. Land Degrad. Dev. *29*, 4415–4423.
11. Thakur, M.P., Reich, P.B., Hobbie, S.E., Stefanski, A., Rich, R., Rice, K.E., Eddy, W.C., and Eisenhauer, N. (2018). Reduced feeding activity of soil detritivores under warmer and drier conditions. Nat. Clim. Chang. *8*, 75–78.
12. Berdugo, M., Delgado-Baquerizo, M., Soliveres, S., Hernández-Clemente, R., Zhao, Y., Gaitán, J.J., Gross, N., Saiz, H., Maire, V., Lehmann, A., et al. (2020). Global ecosystem thresholds driven by aridity. Science *367*, 787–790.
13. Ban, Z., Yuan, P., Yu, F., Peng, T., Zhou, Q., and Hu, X. (2020). Machine learning predicts the functional composition of the protein corona and the cellular recognition of nanoparticles. Proc. Natl. Acad. Sci. USA *117*, 10492–10499.
14. Yu, F., Wei, C., Deng, P., Peng, T., and Hu, X. (2021). Deep exploration of random forest model boosts the interpretability of machine learning studies of complicated immune responses and lung burden of nanoparticles. Sci. Adv. *7*. eabf4130–14.

15. Ban, Z., Hu, X., and Li, J. (2022). Tipping points of marine phytoplankton to multiple environmental stressors. Nat. Clim. Chang. *12*, 1045–1051.
16. Karimi, B., Terrat, S., Dequiedt, S., Saby, N.P.A., Horrigue, W., Lelièvre, M., Nowak, V., Jolivet, C., Arrouays, D., Wincker, P., et al. (2018). Biogeography of soil bacteria and archaea across France. Sci. Adv. *4*, eaat1808.
17. Liang, C., and Balser, T.C. (2012). Warming and nitrogen deposition lessen microbial residue contribution to soil carbon pool. Nat. Commun. *3*, 1222–1224.
18. D Vinebrooke, R., L Cottingham, K., Norberg Marten Scheffer, J., I Dodson, S., C Maberly, S., Sommer, U., and Sommer, U. (2004). Impacts of multiple stressors on biodiversity and ecosystem functioning: the role of species co-tolerance. OIKOS *104*, 451–457.
19. Antão, L.H., Weigel, B., Strona, G., Hällfors, M., Kaarlejärvi, E., Dallas, T., Opedal, Ø.H., Heliölä, J., Henttonen, H., Huitu, O., et al. (2022). Climate change reshuffles northern species within their niches. Nat. Clim. Chang. *12*, 587–592.
20. Patoine, G., Eisenhauer, N., Cesarz, S., Phillips, H.R.P., Xu, X., Zhang, L., and Guerra, C.A. (2022). Drivers and trends of global soil microbial carbon over two decades. Nat. Commun. *13*, 4195–4210.
21. Bahram, M., Hildebrand, F., Forslund, S.K., Anderson, J.L., Soudzilovskaia, N.A., Bodegom, P.M., Bengtsson-Palme, J., Anslan, S., Coelho, L.P., Harend, H., et al. (2018). Structure and function of the global topsoil microbiome. Nature *560*, 233–237.
22. Li, D., Zhang, Q., Xiao, K., Wang, Z., and Wang, K. (2018). Divergent responses of biological nitrogen fixation in soil, litter and moss to temperature and moisture in a karst forest, southwest China. Soil Biol. Biochem. *118*, 1–7.
23. Rillig, M.C., Ryo, M., Lehmann, A., Aguilar-Trigueros, C.A., Buchert, S., Wulf, A., Iwasaki, A., Roy, J., and Yang, G. (2019). The role of multiple global change factors in driving soil functions and microbial biodiversity. Science *366*, 886–890.
24. Hong, C., Burney, J.A., Pongratz, J., Nabel, J.E.M.S., Mueller, N.D., Jackson, R.B., and Davis, S.J. (2021). Global and regional drivers of land-use emissions in 1961–2017. Nature *589*, 554–561.
25. Li, Y., Brando, P.M., Morton, D.C., Lawrence, D.M., Yang, H., and Randerson, J.T. (2022). Deforestation-induced climate change reduces carbon storage in remaining tropical forests. Nat. Commun. *13*, 1964–2013.
26. Allison, S.D., and Martiny, J.B.H. (2008). Resistance, resilience, and redundancy in microbial communities. Proc. Natl. Acad. Sci. USA *105*, 11512–11519.
27. Lundberg, S.M., and Lee, S.I. (2017). A unified approach to interpreting model predictions. In Proceedings of the 31st International Conference on Neural Information Processing Systems, *17*, pp. 4768–4777.
28. O'Neill, B.C., Tebaldi, C., van Vuuren, D.P., Eyring, V., Friedlingstein, P., Hurtt, G., Knutti, R., Kriegler, E., Lamarque, J.F., Lowe, J., et al. (2016). The Scenario Model Intercomparison Project (ScenarioMIP) for CMIP6. Geosci. Model Dev. (GMD) *9*, 3461–3482.
29. Hurtt, G.C., Chini, L., Sahajpal, R., Frolking, S., Bodirsky, B.L., Calvin, K., Doelman, J.C., Fisk, J., Fujimori, S., Klein Goldewijk, K., et al. (2020). Harmonization of global land use change and management for the period 850–2100 (LUH2) for CMIP6. Geosci. Model Dev. (GMD) *13*, 5425–5464.
30. Chen, T., and Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In Proceedings of the 22nd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining, pp. 785–794.
31. Wang, C., Liu, D., and Bai, E. (2018). Decreasing soil microbial diversity is associated with decreasing microbial biomass under nitrogen addition. Soil Biol. Biochem. *120*, 126–133.
32. Fan, K., Delgado-Baquerizo, M., Guo, X., Wang, D., Wu, Y., Zhu, M., Yu, W., Yao, H., Zhu, Y.-G., and Chu, H. (2019). Suppressed N fixation and diazotrophs after four decades of fertilization. Microbiome *7*, 143.
33. Zhou, Z., Wang, C., and Luo, Y. (2020). Meta-analysis of the impacts of global change factors on soil microbial diversity and functionality. Nat. Commun. *11*, 3072–3110.
34. Luo, R., Kuzyakov, Y., Liu, D., Fan, J., Luo, J., Lindsey, S., He, J.S., and Ding, W. (2020). Nutrient addition reduces carbon sequestration in a Tibetan grassland soil: Disentangling microbial and physical controls. Soil Biol. Biochem. *144*, 107764–107814.
35. Chen, Z., Xu, Y., Fan, J., Yu, H., and Ding, W. (2017). Soil autotrophic and heterotrophic respiration in response to different N fertilization and environmental conditions from a cropland in Northeast China. Soil Biol. Biochem. *110*, 103–115.
36. Hu, A., Choi, M., Tanentzap, A.J., Liu, J., Jang, K.S., Lennon, J.T., Liu, Y., Soininen, J., Lu, X., Zhang, Y., et al. (2022). Ecological networks of dissolved organic matter and microorganisms under global change. Nat. Commun. *13*, 3600.
37. Akinyede, R., Taubert, M., Schrumpf, M., Trumbore, S., and Küsel, K. (2022). Temperature sensitivity of dark $CO_2$ fixation in temperate forest soils. Biogeosciences *19*, 4011–4028.
38. Wilken, S., Huisman, J., Naus-Wiezer, S., and Van Donk, E. (2013). Mixotrophic organisms become more heterotrophic with rising temperature. Ecol. Lett. *16*, 225–233.
39. Clarke, A.C., Prost, S., Stanton, J.-A.L., White, W.T.J., Kaplan, M.E., and Matisoo-Smith, E.A.; Genographic Consortium (2014). From cheek swabs to consensus sequences: an A to Z protocol for high-throughput DNA sequencing of complete human mitochondrial genomes. BMC Genom. *15*, 68.
40. Choi, J., Yang, F., Stepanauskas, R., Cardenas, E., Garoutte, A., Williams, R., Flater, J., Tiedje, J.M., Hofmockel and, K.S., and Gelder, B. (2016). RefSoil: A reference database of soil microbial genomes. ISME J. *761*, 1–7.
41. Xiao, K.Q., Ge, T.D., Wu, X.H., Peacock, C.L., Zhu, Z.K., Peng, J., Bao, P., Wu, J.S., and Zhu, Y.G. (2021). Metagenomic and 14C tracing evidence for autotrophic microbial $CO_2$ fixation in paddy soils. Environ. Microbiol. *23*, 924–933.
42. Topcuoglu, B.D., Lesniak, N.A., Ruffin, M.I., Wiens, J., and Schlossa, P.D. (2020). A Framework for Effective Application of Machine Learning to Microbiome-Based Classification Problems. Preprint at bioRxiv. https://doi.org/10.1101/816090.
43. Ilyas, I.F., Beskales, G., and Soliman, M.A. (2008). A Survey of Top-k Query Processing Techniques in Relational Database Systems. ACM Comput. Surv. *40*, 1–58.
44. Still, C.J., Berry, J.A., Collatz, G.J., and DeFries, R.S. (2003). Global Distribution of $C_3$ and $C_4$ Vegetation: Carbon Cycle Implications. Global Biogeochem. Cycles *17*, 6.1–6.14.

## STAR★METHODS

### KEY RESOURCES TABLE

| REAGENT or RESOURCE | SOURCE | IDENTIFIER |
|---|---|---|
| Deposited data | | |
| The Python codes used in this study are available at GitHub | GitHub | https://github.com/duck9427/Predicition-of-Global-Soil-Microbial-Function.git |
| Software and algorithms | | |
| XGBoost regression | Scikit-learn (version1.0.1): Machine Learning in Python 3.8 | https://scikit-learn.org/stable/index.html |
| Random forests regression | Scikit-learn (version1.0.1): Machine Learning in Python 3.8 | https://scikit-learn.org/stable/index.html |
| K-nearest neighbors regression | Scikit-learn (version1.0.1): Machine Learning in Python 3.8 | https://scikit-learn.org/stable/index.html |
| Ordinary least squares regression | Scikit-learn (version1.0.1): Machine Learning in Python 3.8 | https://scikit-learn.org/stable/index.html |
| Linear regression | Scikit-learn (version1.0.1): Machine Learning in Python 3.8 | https://scikit-learn.org/stable/index.html |
| SHapley Additive exPlanations (SHAP) model | Scikit-learn (version1.0.1): Machine Learning in Python 3.8 | https://scikit-learn.org/stable/index.html |
| R software version 4.1.2 | R software | https://www.r-project.org/ |
| Python version 3.7 | Python Software | https://www.python.org/ |
| ArcGIS 10.7 | ArcGIS Desktop | https://desktop.arcgis.com/ |

### RESOURCE AVAILABILITY

#### Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact, Li Mu (Email: muli@caas.cn).

#### Materials availability

This study did not generate new unique materials.

#### Data and code availability

- The experimental data have been presented in Data S1 and S2.
- The Python codes used in this study are available at (GitHub:https://github.com/duck9427/Predicition-of-Global-Soil-Microbial-Function.git), are publicly available as of the date of publication. .
- Any additional information required to reanalyze the data reported in this work paper is available from the lead contact upon request.

### METHOD DETAILS

#### Data collection

Given that the second-generation sequencing technology established in 2005 matured and began to be applied to a variety of studies by 2010,[39] an extensive literature survey was conducted through the Web of Science platform. The soil microbial data were from relevant studies published from January 2010 to December 2021. The keywords we used to search the Web of Science platform included "soil," "bacteri*" and "fung*", and the initial search returned 1,931,418 studies. The following criteria were used to screen for appropriate studies: (1) only field studies were selected, and laboratory experimental studies were excluded; (2) microbial composition at the phylum level was reported; and (3) soil microbial community information was quantified by high-throughput sequencing techniques.

Ultimately, the dataset (Data 1) includes 1726 observations from 640 locations (sample locations are shown in Figure S1). The relative abundance of soil microorganisms at the phylum level was the most common data type in different studies, was comparable between studies, and was collected as the microbial feature. The dataset covered variations in 14 environmental factors related to soil CFM abundance, which are mean annual temperature (MAT), mean annual precipitation (MAP), pH, soil total carbon (TC), soil organic matter (OM), soil total N (TN), soil total P (TP), soil water (SW), organic carbon (OC), texture and artificial management status (*i.e.*, use of N/P/K fertilizers). Longitude and latitude

data were also collected. In cases where the studies did not report MAT or MAP, the values were derived from the historical monthly weather data from the Database: WorldClim (https://www.worldclim.org) using site geographic location (*i.e.*, latitude and longitude). Any absent data pertaining to properties of global soils (*e.g.*, pH, TC, OM, OC, TN and TP) were filled in by retrieving data from the Database: International Soil Reference and Information Centre (ISRIC) World Soil Information Data Hub (http://www.tpdc.ac.cn). SW and texture data were collected from the Database: Harmonized World Soil Database (HWSD v1.2, https://previous.iiasa.ac.at/web/home/research/researchPrograms/water/HWSD.html). In cases where the studies did not report the latitude or longitude, the approximate latitude and longitude were derived by geocoding the name of the location in Google Earth 7.0. We took into consideration that various climate and soil properties exist in any given habitat and that further subdivision will lead to insufficient data. Therefore, we did not include the habitats in the machine learning model.

## Calculation of the CFMs and microbial tolerance

The relative abundance of CFMs was defined as the proportion of microbes with corresponding functions (shown in Data 2):

$$K = \frac{\sum_{i=1}^{n} B_i * \frac{N_k}{N}}{\sum_{i=1}^{n} B_i} \quad \text{(Equation 1)}$$

where $N_k$ is the number of functional microbes in phylum $i$; $N$ is the total species number in phylum $i$; and $B_i$ is the relative abundance of the bacterial phylum.

Information about common soil microbial species ($n$=851) was obtained from the RefSoil database.[40] Five common marker genes (*cbbL*, *aclA*, *acsA*, *accA* and *hcd*)[41] involved in microbial $CO_2$ fixation pathways were used to select the functional microbes based on the genomes of the above species from the National Center for Biotechnology Information (NCBI). Those species that do not have a complete genome in NCBI but have been proven to be autotrophic microorganisms, such as Cyanobacteria, were also considered functional microorganisms (Data 2).

Microbial tolerance is the proportion of tolerant species in the microbial community under extreme disturbance (*e.g.*, excessive temperature or extreme pH) (Figure 1D)[26] (Allison and Martiny, 2008). We selected 6 extreme environmental conditions, including high temperature, low temperature, barrenness, drought, and peracidic and peralkaline conditions. Microbes that can survive in these extreme environments are considered to be stress-tolerant microbes (shown in Data 2). According to information from the NCBI (National Center for Biotechnology Information (nih.gov)), thermophilic microbes are defined as species with an optimum temperature greater than 30°C; psychrophilic microbes are defined as species with an optimum temperature lower than 15°C; acidophilic microbes are defined as having an optimal survival pH less than 5; alkalophilic microbes are defined as having an optimal survival pH higher than 9; and drought- and barrenness-tolerant microbes are the dominant species in desert and infertile areas. We defined microbial tolerance as follows:

$$T = \frac{\sum_{i=1}^{n} B_i * \frac{N_t}{N}}{\sum_{i=1}^{n} B_i} \quad \text{(Equation 2)}$$

where $N_t$ represents the number of tolerant microbes in phylum $i$; $N$ represents the total species number in phylum $i$; and $B_i$ represents the relative abundance of bacterial phylum $i$.

## Data preparation and model building

We used the interquartile range (IQR) criterion to exclude outlier data.[42] As shown in Figure S2D, IQR is defined as Q3-Q1, and (Q1, Q3) covers the middle 50% of the data in the data distribution. The data outside the range (Q1-1.5·IQR, Q3+1.5·IQR) were considered outlier data. We upsampled the positive samples to obtain a balanced dataset. There were 1726 samples used for the model construction. We randomized our dataset into a training subsample (72%, $n$=1243) and a test subsample (18%, $n$=311). Both the training set and test set were divided by 5-fold cross-validation. The validation set accounted for 10% of the dataset ($n$=172). Then, the data dependence caused by upsampling and data division was eliminated, making the results credible. XGBoost, random forests, decision trees, K-nearest neighbors, ordinary least squares and linear regression were tested to find a suitable model. The validation set accounted for 10% of the training set. The sampling ratio of random sampling was 0.9. The maximum depth and *colsample_bytree* were 12 and 0.8, respectively. The number of samples in the leaf node was 13. Seed was set at 2 to make the results the same every time.

In the process of data cleaning, it is possible to remove data that are useful but not robust after removing outlier data.[43] We introduced adversarial samples to test the robustness of the models. Adversarial samples that were generated from the critical environmental factors remained unchanged, and other features were randomly generated within the disturbance range. The new data were mixed with the original and adversarial data at a ratio of 7:3 and then put into the models as a new dataset. The robustness of the models was evaluated by the change in $R^2$ between the original dataset and the new dataset.

### SHAP analysis and identification of tipping points

SHapley Additive exPlanations (SHAP) values estimate the contribution of each feature by averaging over all the possible marginal contributions to a prediction task and are a unified framework for interpreting machine learning models.[27] The SHAP model can be used for not only global interpretation but also local interpretation. The possible relationship between a predicted value given by a model and some features can be explained by the SHAP model.[27]

The SHAP model is a post hoc model.[27] The SHAP model calculates the marginal contribution of features to the model output and then explains the black box model at the global and local levels. It constructs an additive interpretation model, and all features are regarded as contributors. For each prediction sample, the model generates a prediction value. Various combinations of features (players) are taken to form coalitions, while each SHAP value measures the average contribution of each player across all possible combinations. The calculation of the SHAP value of a single feature, eliminating cross effects, is as follows:

$$\Phi_{i,i} = \varphi_i - \sum_{j \neq i} \Phi_{i,j} \quad \text{(Equation 3)}$$

Here, $\Phi_{i,i}$ represents the contribution of the feature $i$; $\varphi_i$ represents the calculated SHAP value of feature $i$ based on the tree-based model; and $\Phi_{i,j}$ represents the cross influence of features $i$ and $j$. The dataset X_train was used to calculate the SHAP value.

Two types of tipping points (discontinuous and continuous) were identified and analyzed. A continuous tipping point indicates that the relationship between independent and dependent variables changes significantly; a discontinuous tipping point indicates that the value of the dependent variable changes abruptly due to the independent variable.[12] A generalized additive model (GAM) and segmented linear regression (SLR) model were used for regression analysis. The tipping point of the GAM was defined as the point with a second derivative of 0 in the continuous curve; the tipping point of the SLR model was defined by the overall change in the intercept and slope of the linear regression before and after the tipping point.[12] Tipping points were derived from the *segmented* and *mgcv* packages in R. The model with the highest $R^2$ was used for further analysis. We also used the GAM and the SLR model to calculate the tipping points of the independent variables and the corresponding SHAP values. Only when the tipping points from the raw data and the SHAP values were similar were they considered to be the actual tipping points.

### Global prediction

The monthly temperature and precipitation data for 2021 were obtained from the Database: National Oceanic and Atmospheric Administration (https://psl.noaa.gov/data/gridded/index.html), with a resolution of 0.5°×0. 5°. The future MAP data were obtained from the Database: Climate Model Intercomparison Project Phase 6 (CMIP6,https://esgf-node.llnl.gov/search/cmip6/) at a 1°×1° resolution. The fertilizer data at present and in the future were derived from the Database: Land-Use Harmonization (LUH2) project (https://luh.umd.edu). The resolution was 1°×1°. In the latitude range of -30 degrees to 10 degrees, $C_4$ plants are dominant,[44] so $C_4$ plant fertilizer data were used. $C_3$ plant data were employed for the remaining regions. Global soil properties with 10 km resolution were obtained from the Database: Global Soil Dataset for Earth System Modeling (2014) (http://data.tpdc.ac.cn/zh-hans/data). Because there is no suitable soil property database that is based on future climate models, current soil properties were used for future predictions in our analysis. The data were resampled to a spatial resolution of 1°×1°, and all terrestrial grid cells were input into the established models to obtain the global distribution pattern.The global cropland maps for 2019 are publicly available at Database: Global cropland expansion in the 21st century (https://glad.umd.edu/dataset/croplands).

We repeatedly predicted values for different climatic zones and continents to quantify the uncertainty in the machine learning models. The global data were divided into the following regions to test predictive repeatability: Africa, Asia, Europe, North America, South America, Oceania, North Cold Zone, North Temperate Zone, Tropical Zone and South Temperate Zone. The stability of the models was assessed by calculating the differences between repeated predictions for the same location.

## QUANTIFICATION AND STATISTICAL ANALYSIS

The confidence interval of prediction and significance analysis of different habitats were conducted in Python 3.8.6. All confidence levels not otherwise specified were 0.95.

## ADDITIONAL RESOURCES

This study did not generate additional resources.