International Institute for Applied Systems Analysis
IIASA www.iiasa.ac.at

Address: IIASA, Schlossplatz 1, A-2361 Laxenburg, Austria
Email: permissions@iiasa.ac.at

Young Scientists Summer Program

Investigating the Biophysical Determinants of Soil Organic Carbon Response to Management in the EPIC-IIASA Model

Tara Ippolito, Tara.Ippolito@colorado.edu
Juraj Balkovic, balkovic@iiasa.ac.at
Rastislav Skalsky, skalsky@iiasa.ac.at
Christian Folberth, folberth@iiasa.ac.at
Jason Neff, Jason.C.Neff@colorado.edu

Approved by:

Mentor(s): Juraj Balkovic
Program: BNR / AFE
Date: 29 September, 2022

This report represents the work completed by the author during the IIASA Young Scientists Summer Program (YSSP) with approval from the YSSP mentor.

It was finished by ___Tara Ippolito___ and has not been altered or revised since.

Mentor signature:

Table of contents

ZVR 524808900

UPDATE OR DELETE
For IIASA Working Papers/Reports/Policy Briefs funded by the IIASA core budget we ask the authors to include the following:
The authors gratefully acknowledge funding from IIASA and the National Member Organizations that support the institute (The Austrian Academy of Sciences; The Brazilian Federal Agency for Support and Evaluation of Graduate Education (CAPES); The National Natural Science Foundation of China (NSFC); The Academy of Scientific Research and Technology (ASRT), Egypt; The Finnish Committee for IIASA; The Association for the Advancement of IIASA, Germany; The Technology Information, Forecasting and Assessment Council (TIFAC), India; The Indonesian National Committee for IIASA; The Iran National Science Foundation (INSF); The Israel Committee for IIASA; The Japan Committee for IIASA; The National Research Foundation of Korea (NRF); The Mexican National Committee for IIASA; The Research Council of Norway (RCN); The Russian Academy of Sciences (RAS); Ministry of Education, Science, Research and Sport, Slovakia; The National Research Foundation (NRF), South Africa; The Swedish Research Council for Environment, Agricultural Sciences and Spatial Planning (FORMAS); The Ukrainian Academy of Sciences; The Research Councils of the UK; The National Academy of Sciences (NAS), USA; The Vietnam Academy of Science and Technology (VAST).

For research funded by an external third party we ask the authors to include the full name of the funder, the title of the project and the grant number, e.g.
The authors gratefully acknowledge funding from the European Research Council for the research project 'Forecasting Societies Adaptive Capacities to Climate Change' (FUTURESOC, FP7 230195).

Abstract

The management of Soil Organic Carbon (SOC) is a critical component of both nature-based solutions for climate change mitigation and global food security. Agriculture has contributed substantially to a reduction in global SOC through cultivation, thus there has been renewed focus on management practices which minimize SOC losses as a pathway towards maintaining healthy soils and reducing greenhouse gas emissions. Mechanistic models are frequently used to aid in identifying these pathways due to their scalability and cost-effectiveness. In this analysis, we aim to understand the spatial patterns and the background biophysical determinants of SOC responses to agricultural management such as nitrogen fertilizer, addition of crop residues, and crop choice in Europe through the use of a multifactorial crop model and a statistical meta-model approach. Using 35 years of factorial, spatially-explicit simulation data from the gridded Environmental Policy Integrated Climate-based modeling system (EPIC-IIASA), we build multiple polynomial regression ensemble meta-models to quantify SOC responses to varying management intensities for unique combinations of climate and soils across Europe. We find that the biophysically determined meta-model is a highly accurate ($R^2 = .97$) and a sufficiently robust simplification of EPIC-IIASA for the estimation of SOC responses to cropland management. Model stratification by means of climate and soil clustering improved the meta-model's performance compared to the full EU-scale model. Furthermore, we find notable differences between biophysically specific models throughout Europe, which point to spatially distinct SOC responses to management choices such as nitrogen fertilizer application rates and residue retention. This underlines the importance of appropriate regionalization in the large-scale SOC meta model. The proposed solution provides a framework for downscaling results from IIASA large-scale agricultural model for regional assessments and its extension to farmers and land owners.

About the author

Tara Ippolito is a PhD Candidate at the University of Colorado, Boulder, Boulder, CO U.S.A in the Mortenson Center for Global Engineering. She is a National Science Foundation Graduate Research Fellow.
(Contact: Tara.Ippolito@colorado.edu)

Acknowledgments

Thank you to the U.S. Committee for IIASA at the National Academy of Sciences and the Friends of IIASA for the continued support of this research. This material is based upon work supported by the National Science Foundation Graduate Research Fellowship Program. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation. Finally, thank you to the IIASA AFE and BNR groups for their lasting collaboration and support.

Introduction

Climate change presents a tremendous threat to global agriculture, and thus the livelihoods and economies which depend on agricultural productivity. In the time period 1983-2009, roughly 75% of the global harvested area experienced drought-induced yield losses corresponding to 166 billion USD in losses (Kim et al., 2019). Maize and wheat, which together account for 50% of global cereal production, are expected to experience a global decrease in production over time as climate variability increases (Daryanto et al., 2016; Lobell et al., 2011). In Europe, climate trends have been attributed to ~10% of the yield stagnation observed since the early 1990s (Moore & Lobell, 2015). It is expected that European grain maize yields will decline by up to 22% and wheat yields in Southern Europe are expected to decrease by up to 49% (Hristov et al., 2020). Threats from adverse weather events (Trnka et al., 2016), pests (Skendžić et al., 2021), and crop diseases (Gautam et al., 2013) are all expected to increase with climate change. These challenges are compounded by land degradation which has affected over 20% of the global vegetated land area (UNCCD, 2017). While global agriculture is highly vulnerable to climate change, it is also a substantial contributor to global greenhouse gas emissions through input resources and changes in soil organic carbon (SOC). Greenhouse gas emissions from agriculture make up 21% of the global aggregate and agriculture is the second largest source of global emissions (FAO, 2016). Input resources, such as synthetic fertilizer, emit carbon throughout their production, storage, and transportation processes (Walling and Vaneeckhaute, 2020). One of the major mechanisms of agricultural greenhouse gas emissions is the loss of SOC (Baumann et al., 2017; Lal, 2004).

Soil Organic Carbon (SOC), which represents a carbon stock of roughly 1,500–2,400 Gt C (~5500–8800 Gt CO2) globally, constitutes a substantial portion of global carbon stocks (Sanderman et al., 2017). Agriculture has historically contributed to a loss in SOC, primarily through conversion of native soil to agricultural uses (Sanderman et al. 2017) - soil carbon stocks decline substantially when cropland replaces native forest (−42%) and pasture (−59%) (Guo & Gifford, 2002). Once soils have been converted for cultivation, further soil carbon losses may occur as a result of management practices such as tillage, choice of crop, and nutrient inputs (Lal, 2002). Jastrow et al. (2007) highlight that management practices "mainly control the spatiotemporal distribution of Organic Matter (OM) inputs to soil, along with the sensitivity of OM to mineralisation, both of which affect soil OM stocks". Although agricultural soils are often a source of carbon emissions, they can serve as sinks for carbon emissions depending on factors such as soil properties and climate (Eglin et al., 2010). There has been a renewed focus on global soil carbon sequestration and its potential co-benefits as global policy makers attempt to mitigate climate change through initiatives such as the '4p1000' initiative (Chabbi et al., 2017). While estimation is tenuous, recent analysis suggests a soil C sequestration annual technical potential of 2–5 Gt CO2/year (Fuss et al., 2018). Smith et al. (2019) note that there is still an "incomplete understanding on how SOC changes are influenced by climate, land use, management and edaphic factors". It has been observed that these factors may cause an asymmetry in the mechanisms of SOC changes (Attard et al., 2016). As decision- and policy-makers renew focus on adopting agricultural practices which reduce or reverse soil C losses, proper attention must be paid to the complexity and spatial variability of SOC changes resulting from environmental boundary conditions and the effects of management practices. One obstacle to understanding this complexity is reliable, scalable, and cost-efficient monitoring of SOC.

A number of methods exist to estimate, and thus monitor, SOC. These methods fall into two main categories - direct measurement through sampling, and indirect measurement through statistical inference or process-based models (Smith et al., 2019). Direct measurement, while often the most trusted source of SOC estimates, relies on appropriate study design, sufficient sampling depth, and proper processing and storage of samples (Minasny et al., 2017; Smith et al., 2019; Nelson & Sommers, 1996). Furthermore, to capture the spatial variability and long-term nature of SOC changes, direct measurements must be repeated sufficiently which may prove to be cost-prohibitive (Minasny et al., 2017; Vanguelova et al., 2017; Smith, 2004). Indirect measurements such as inference from flux measurements and spectral methods vary in their accuracy and cost of implementation (Smith et al., 2019). One highly scalable and cost-efficient method in estimating SOC changes is the use of process-based models. There are a wide variety of models which vary in the processes they represent and how they are structured mathematically (Manzoni & Porporato, 2009). The IPCC, in their 2019 Refinement to the 2006 IPCC Guidelines for National Greenhouse Gas Inventories, provide SOC quantification methodologies at varying levels of detail from simple empirical calculations to more involved process-based models such as DAYCENT or DNDC (Whitaker et al., 2013). The robustness and accuracy of these models are limited by the underlying assumptions and initialization processes, and the availability of reliable calibration and validation data (Keel et al., 2017; Toudert et al., 2018). Additionally, large-scale model applications are limited by accuracy of spatially-explicit model inputs, including initial soil properties and crop management assumptions (Balkovič et al., 2020). In spite of these limitations, process-based models are frequently used for a wide range of applications such as assessment of policy goals (Minasny et al., 2017) and estimation of landscape-scale SOC dynamics (Pennock & Frick, 2001). Gridded biophysical models, including the well-established gridded model EPIC-IIASA (Balkovič et al. 2014), have been evaluated as robust solutions for the agriculture sector assessments at large scales, even globally (Müller et al., 2016; Jägermeyr et al., 2021). These models are increasingly used in the EU-scale assessments to support land use policies, such as carbon emissions and removals from land use and land use change (Petrescu et al., 2021; Frank et al., 2015). One major benefit of using models like EPIC-IIASA for estimation of SOC changes is the ability to simulate both existing and potential agricultural practices across large areas. As the effects of climate change continue to affect farmers, management practices may change to reflect new growing conditions (Iglesias et al., 2012). For example, a northward shift in cultivable zones may allow farmers in some areas of Europe to grow a wider variety of crops while other areas of Europe may see a reduction in crop growing potential of certain varieties (Ewert et al., 2005). Models allow us to investigate such scenarios and the SOC dynamics of these novel management choices.

While there are many benefits to using process-based models, large-scale agricultural models like EPIC-IIASA may demonstrate limited accuracy for regional SOC assessments or the extension of projected impacts to farmers and land owners due to their spatial granularity, uncertain inputs, and a robust regional calibration of biophysical processes (Balkovič et al. 2020). In order to navigate these limitations, users of these models often build meta-models which train statistical models on simulation data from process-based models in order to emulate process-based models in a simplified format. These meta-models, which so far focus foremost on crop yield responses, have been utilized to evaluate differences between process-based models (Ringeval et al., 2021), explore yield responses to climate change (Franke et al., 2020; Blanc, 2017; Oyebamiji et al., 2015), and downscale process-based model estimates (Folberth et al., 2019). These meta-models alleviate the substantial data requirements and computational cost of running, calibrating, and validating a process-based model, while preserving their robust scientific capabilities. A meta-model framework that allows inputting fine-scale data together with parameterization for different regions, while still accounting for the main biophysical and management determinants of SOC dynamics, would be a significant improvement for regional SOC modeling.

In this paper, we use multifactorial modeling and a statistical approach to investigate the biophysical determinants and spatial patterns of SOC responses to management in Europe. Through analyses of inputs and outputs from the well-established gridded crop model EPIC-IIASA, we design a robust and scalable statistical meta-model to simulate SOC response to changing management intervention such as N fertilization, organic C inputs, and crop choice at the EU scale. This approach allows us to explore a wide variety of management scenarios across the European continent and identify nuances in the SOC responses across different climates and soil properties. The proposed framework improves the capacity of the IIASA agricultural modeling suite to contribute to regional carbon assessments.

Methodology and Data

2.1 Methodology Overview

As outlined in Figure 1, the framework we utilize to explore the biophysical determinants of SOC responses to management consists of 3 major components: (i) multifactorial gridded EPIC-IIASA modeling (EPIC Hypercube), (ii) biophysical clustering of spatially-explicit, gridded simulation units (SimU), and (iii) cluster-specific regression meta-models. We begin by generating modeled data through the use of a gridded version of the Environmental Policy Integrated Climate (EPIC) model - the EPIC-IIASA Gridded Agronomic Model (GAM, Balkovič et al., 2020, 2013) - in order to simulate a wide variety of growing conditions and subsequent crop growth throughout Europe. From the EPIC-IIASA GAM multifactorial modeling approach, we generate millions of unique crop growth simulations spanning Europe at 1km x 1km resolution over 40 years (1985-2019) under a variety of management practices. In the factorial design, crop growth, yields and SOC were simulated for increasing levels of N fertilization, organic C inputs and a variety of crops. In order to explore the unique biophysical determinants of SOC responses to management, we split SimU into distinct clusters based on the spatio-temporal climate and soil properties. For each of these biophysically-distinct clusters, we build an interpretable meta-model of the EPIC-IIASA GAM using a machine-learning approach. While the EPIC model directly simulates SOC dynamics, we utilize EPIC meta-models in order to identify the nature of the dynamics which drive changes in SOC and their biophysical determinants.

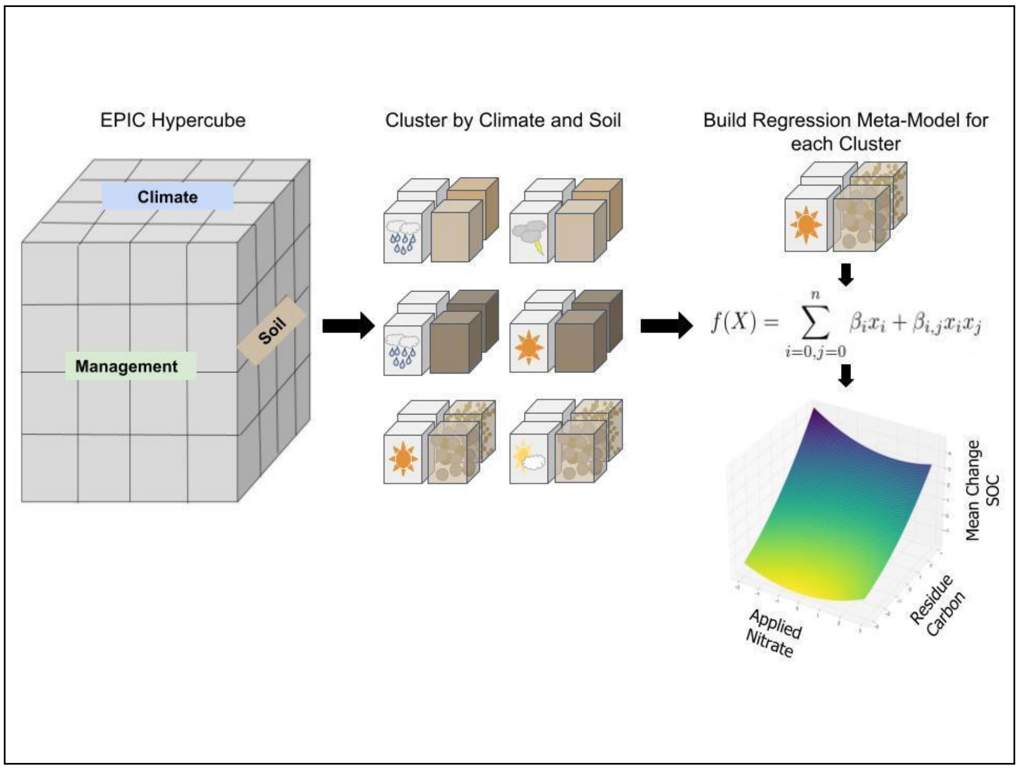


Figure 1: Framework utilized in this paper

2.2 Multifactorial Gridded EPIC-IIASA Modeling

2.2.1 Model Description

The Environmental Policy Integrated Climate (EPIC) Model is a process-based model which simulates crop growth under a wide range of crop management options such as tillage, fertilization, irrigation, pesticides, and liming (Izaurralde et al., 2006; Williams, 1995). The EPIC model has been used in a variety of studies investigating soil organic matter cycling (Izaurralde et al., 2006), irrigation timing (Bryant et al., 1992), and the impact of climate change on the agricultural sector (White et al., 2011) among others (Gassman et al., 2004). The EPIC-IIASA GAM (the EU version) was built by coupling EPIC (v. 0810) with EU-wide datasets on land cover, soils, topography and crop management practices (Balkovič et al., 2013, 2018)., The EPIC-IIASA GAM is one of 14 models in the Inter-Sectoral Impact Model Intercomparison Project (ISIMIP) which provides ensemble projections of climate change impacts on agriculture (Warszawski et al., 2014; Jägermeyr et al., 2021). In this work, we use the EPIC-IIASA GAM which runs the EPIC model for a given set of gridded SimU utilizing soil, topography, and climate data (Skalský et al., 2008).

2.2.2 Multifactorial Modeling Framework

Using the EPIC-IIASA GAM, we build a multifactorial simulation across ~86,000 gridded SimU covering Europe simulating a wide variety of management factors. Our modeling design closely follows the framework implemented by Balkovič et al. (2020). For each SimU, we model annual crop yield (in tDM/ha), crop residue carbon (in kgC/ha) and annual change in SOC (tC/ha/year) for the time period 1980-2019. For each SimU x year, we simulate a multifactorial combination of the following management factors: crop type (maize, rape, rice, barley, soya, sunflower, rye, wheat), maximum annual nitrogen fertilizer application (0, 50, 100, 250 kgN/ha, and a crop-specific business-as-usual N application rate from Balkovič et al., 2013), and crop residue retention (0, 30, 60, and 90% of crop residues retained). We discard the first 5 years (1980-1985) of simulation data to allow for model spin-up. Across the ~86,000 SimU, 35 years, 8 crop types, 5 nitrogen application rates, and 4 residue management scenarios, the multifactorial simulation results in over 512 million unique crop growth simulations across Europe.

2.2.3 Crop Model Input Data

Climate Data

Gridded climate data from the CRU ERA v.2.0 source at a spatial resolution of 0.1° from the time period 1980-2019 (provided by the VERIFY project, https://verify.lsce.ipsl.fr) were utilized for EPIC-IIASA GAM simulations. Variables used include daily precipitation, minimum and maximum temperature, vapor pressure deficit, solar radiation, and annual atmospheric CO2 concentration.

Soil Data

Topsoil (0-30cm) and subsoil (30-100cm) properties were calculated from underlying datasets for each SimU. The soil datasets utilized include the European Soil Bureau Database v. 2.0 (ESBD, https://esdac.jrc.ec.europa.eu), the Database of Hydraulic Properties of European Soils (Wösten et al., 1999), and the map of organic carbon content in the topsoil (Lugato et al., 2014). A total of 13 soil properties were used, following the framework of Balkovič et al. (2013, 2018).

Management Data

Management parameters which were varied in the multifactorial modeling approach were crop type, maximum annual nitrogen fertilizer application rates, and residue retention. The crops simulated in this analysis included maize, rape, rice, barley, soya, sunflower, rye, and winter wheat. The maximum annual nitrogen fertilizer application rates considered were 0 kg N/ha, 50 kg N/ha, 100 kg N/ha, 250 kg N/ha, and business-as-usual (BAU) where N is automatically applied as needed throughout the season. In this simulation, nitrate is the Nitrogen fertilizer applied. Crop residue retentions simulated included 0% residues retained, 30% residues retained, 60% residues retained, and 90% residues retained. Under no N application rate, 0 kg/ha, we only simulate 0% residue retained. For all other variables, default crop growing parameters from the EPIC-IIASA model were used (Balkovič et al., 2018, 2013). All simulations were carried out under rainfed conditions only. Nitrogen application is completed automatically by the EPIC-IIASA model based on plant stress, up to the maximum annual application rate specified. We assume a conventional tillage consisting of two soil cultivation operations and a 30-cm deep mouldboard plowing, and an offset disking shortly after harvesting of cereals. In addition, two row cultivations were simulated for maize. Soil erosion was not accounted for in our simulations.

2.3 Biophysical Clustering of Simulation Units

Using a combination of a-priori and unsupervised clustering techniques, we generate biophysical clusters of all ~86,000 SimU across Europe. We consider soil properties and climate properties as the two main components of biophysical characterization. Soil clusters and climate clusters are produced separately, and then combined, so that each overall biophysical cluster is a unique combination of distinct soil and climate conditions. From the 5 SimU soil clusters and 10 SimU climate clusters, there are a total of 50 potential biophysical clusters of which 43 exist among the SimU. Biophysical clusters with less than 1000 observations were omitted from analysis due to concerns of model overfitting which are raised when there is a lack of sufficient data.

2.3.1 Soil Clustering

SimU soil clusters (Figure 2) were selected a-priori using the EPIC-IIASA defined texture classification. There are 6 discrete soil texture classes within EPIC-IIASA - coarse, medium, medium-fine, fine, very fine, and peat. Peat was excluded from this analysis. Each simulation unit has one soil texture classification.

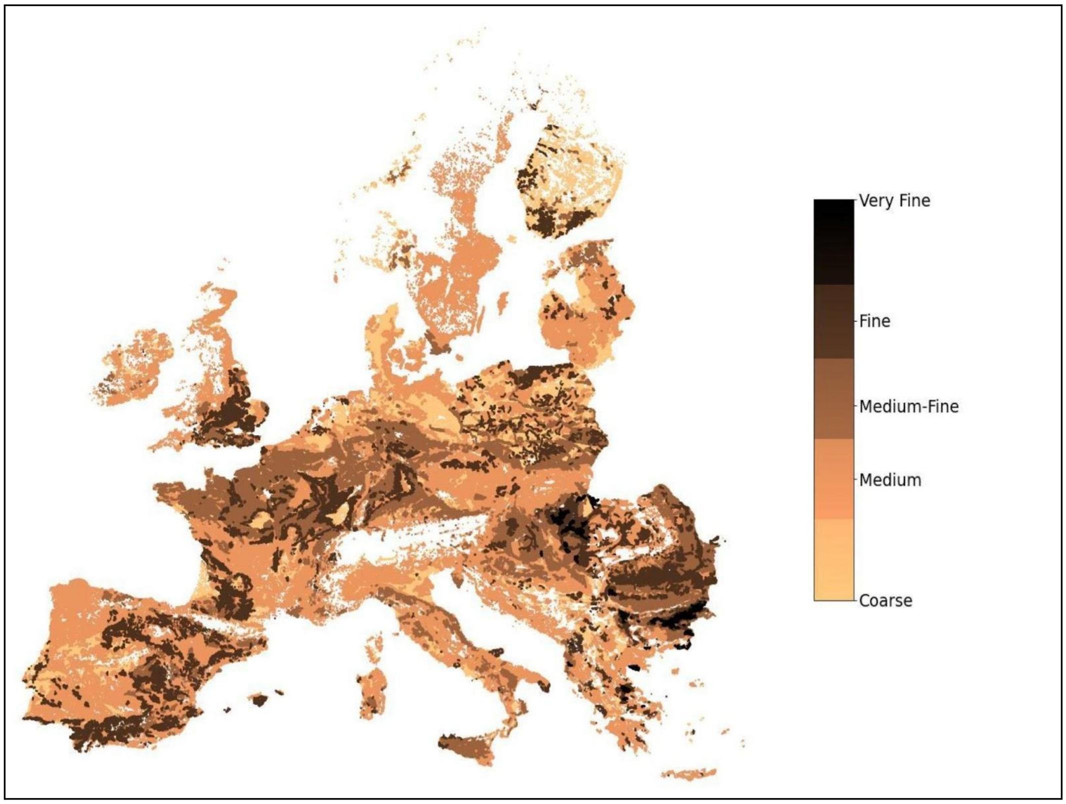


Figure 2: Soil clusters of SimU across Europe - clusters consist of 5 discrete classes based on soil texture defined by EPIC-IIASA.

2.3.2 Climate Clustering

SimU climate clusters (Figure 3) were discovered using an unsupervised machine learning approach - K-means clustering via Principal Component Analysis - which has been used for a variety of applications such as clustering of DNA gene expressions and internet news articles (Ding & He, 2004). The climate data used for the SimU climate clusters is the same data which is used as input to the EPIC-IIASA model, but aggregated from daily to monthly averages and sums depending on the variable. Climate parameters used in the clustering algorithm include monthly precipitation, temperature minimum and maximum, vapor pressure deficit, solar radiation, and potential evapotranspiration over all years (1980-2019). Due to the large number of climate parameters (6 variables x 12 months/year x 40 years), we use Principal Component Analysis as a dimensionality reduction technique to minimize the size of the data while preserving a large amount of information. We calculate principal components and retain those which explain a combined minimum of 90% of the variance in the climate data. We then use the PCA-reduced climate data as input to a K-means clustering algorithm (Pedregosa et al., 2011). The K-means clustering algorithm aims to find cluster centroids by optimizing a squared error function, typically the sum of the squared Euclidean distances (Likas et al., 2003). The algorithm operates by arbitrarily placing cluster centroids and then adjusting centroid position at each step to minimize distance between cluster centroids and members of the cluster (Likas et al., 2003). We tested k = 2 to k = 12 clusters and selected the optimal number of clusters based on a variety of statistical metrics (Inertia, Caliński-Harabasz Score, Davies-Bouldin Score, etc.) in combination with manual inspection and comparison to well known climate classifications such as Köppen-Geiger Climate Classification (Caliński & Harabasz, 1974; Davies & Bouldin, 1979, Peel et al., 2007). We find the optimal number of climate clusters to be 10.

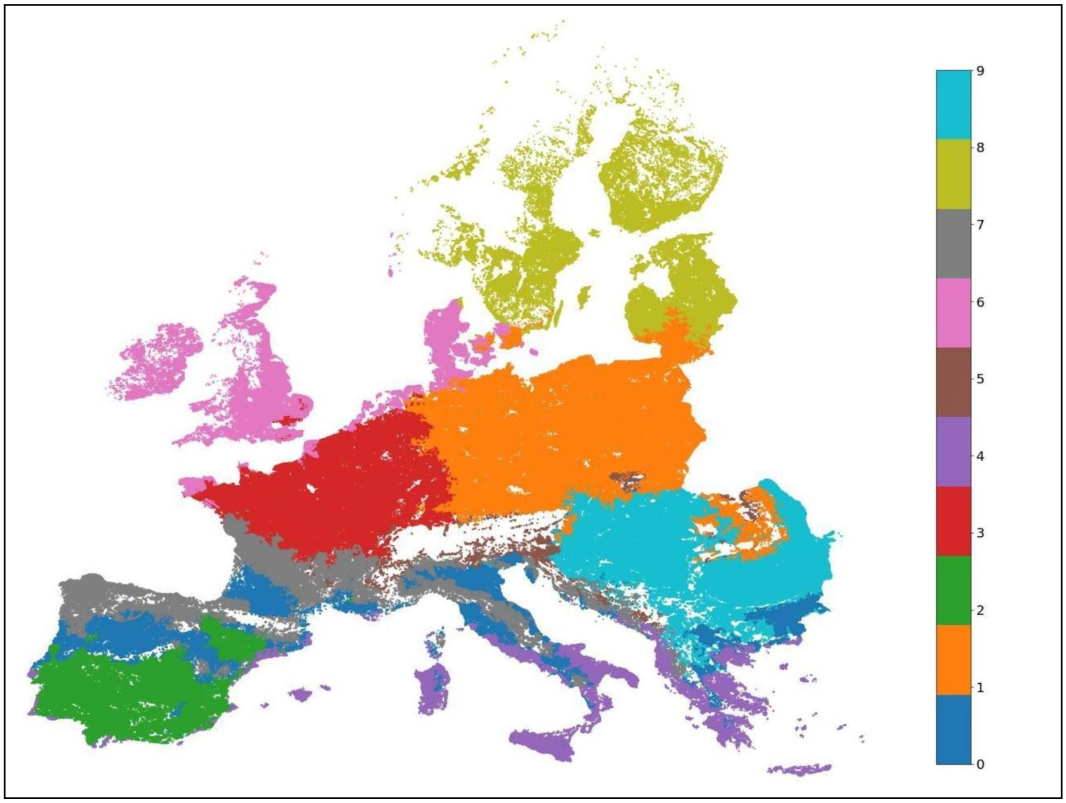


Figure 3: Climate clusters of SimU across Europe - clusters consist of 10 discrete classes based on 40 years of monthly climate data.

2.4 Cluster-Specific Regression Meta-Models

2.4.1 Machine-Learning Framework

We utilize a bagging meta-estimator of multiple polynomial regression models together with inputs and outputs of the multifactorial EPIC-IIASA simulations (Section 2.2) in order to build a robust, yet highly interpretable machine learning model for prediction of mean annual change in SOC. Ensemble methods such as this combine a large number of base-estimators built with a specific algorithm (here, a multiple polynomial regression model) in order to improve on the performance of a model built with a single estimator. In order to avoid overfitting and the challenges of multicollinearity we utilize n-fold cross validation, L2 regularization, out-of-bag error monitoring, and testing on the 25% of observations held out from the training set (Kohavi, 1995; Chong & Jun, 2005).

We build a separate EPIC-IIASA meta-model for each biophysical cluster x crop type in order to investigate SOC responses to management which are unique to given soil and climate conditions. By building meta-models of the EPIC-IIASA simulation data, we can explore the nature of SOC dynamics under a wide-variety of management choices. We chose to separate out crop types to avoid tangling the signals of different crops within our models. In these models, we predict 35-year long term averages of mean annual change in SOC and crop yield. Since SOC changes occur slowly, long-term averages are more appropriate for our investigation. For each model, EPIC-IIASA simulation data was randomly split into training (75% of data) and testing (25% of data) sets.

2.4.2 Regression Meta-Model Structure

The bagging meta-estimator of multiple polynomial regression models is an ensemble method which estimates a number of base estimators and combines these base estimators into one final model. The multiple polynomial regression base we utilize in this work includes linear and quadratic terms for each parameter, as well as interaction terms between parameters. The structure of the multiple polynomial regression base estimator is as follows

Eq. 1 $$f(X) = \sum_{i=0,\, j=0}^{n} \beta_i x_i + \beta_{i,j} x_{i,j}$$

where $f(X)$ is the predicted variable,

$\beta_i x_i$ are the linear terms,

$\beta_{i,j} x_{i,j}$ are the interaction terms when $i \neq j$ and $\beta_{i,j} x_{i,j}$ are the quadratic terms when $i = j$.

This model was selected for its interpretability and the ability to explore response relationships of interest from the learned model. We chose to use quadratic terms for our multiple polynomial regression as quadratic effects of management parameters such as applied nitrogen are well studied and accepted in the literature (Bullock & Bullock, 1994; Puntel et al., 2016). We use L2 regularization in our base models which prevents overfitting and alleviates the challenges of multicollinearity. Our bagging meta-estimator, built with scikit learn, fits 20 base estimators on randomly selected subsets of the training data and then combines the predictions of all estimators into a final model (Breiman, 1996). We tested up to 60 base estimators and found marginal improvements in model performance past 20 base estimators.

2.4.3 Regression Meta-Model Features

Features which are used to train the regression meta-models are based on EPIC-IIASA GAM input data, as described in section 2.2.3. We use a limited set of climate, soil, and management features which are of interest in exploring the biophysically-determined SOC response to management (Table 1). For climate data, we engineer a number of growing season (GS) specific climate variables following the framework of Folberth et al. (2019). For soil data, we engineer full-profile parameters as a weighted sum of topsoil- and subsoil-specific variables. Applied nitrogen fertilizer is calculated as the sum of all applied nitrogen throughout the season as dictated by the automatic application mechanism in EPIC. Applied nitrogen fertilizer does not exceed the specified maximum annual N rate which is specified in the multifactorial simulation. Residue C treatment is calculated as a function of the C in standing aboveground residues, the C in the dead roots, and C in residues from harvest losses. For our target variable we use EPIC-IIASA predicted mean annual change in SOC (meanOCPD), calculated as the average month to month difference in organic carbon in the plowing depth (OCPD) over a given year. If meanOCPD is positive, soils absorbed C throughout the year, if negative, soils lost C throughout the year. As a final step, we collapse data into 35-year long term averages. For each unique simulation design - SimU x crop type x N application rate x residue management - all features, including soil, climate, management, and target variables, are averaged. Since averaging across 35 years dampens the variability of parameters, we also calculate long-term standard deviations of temporally dynamic features (all features except soil variables) to retain information on variability which may be important to SOC dynamics. Thus each data point used to train the model is a long-term record of a given SimU and cropping scenario.

Variable Type	Variable (unit)	Variable Abbreviation
Soil	Profile Sand Content (%)	SAND
	Profile Silt Content (%)	SILT
	Organic C in topsoil (%)	OCTOP
	Profile Field Water Capacity at 33 kPa (cm3/cm3)	FWC
Climate	GS Precipitation Sum (mm)	GSsumPRCP
	GS Precipitation Skew (mm)	GSskewPRCP
	GS Temperature Mean (℃)	GSmeanTEMP
	GS Temperature Skew (℃)	GSskewTEMP
	GS Radiation Mean (MJ/m2)	GSmeanRAD
	GS Radiation Skew (MJ/m2)	GSskewRAD
	GS Potential Evapotranspiration Mean (mm)	GSmeanPET
Management	Applied nitrogen Fertilizer (kgN/ha)	FNO3
	Applied Residue C content (kgC/ha)	RSDCa
Target	Mean Annual Change in SOC (tC/ha)	meanOCPDΔ

Table 1: Regression Meta-model input variables

2.4.4 Feature Importance

In order to identify the feature importance within each biophysically-specific model, we implement leave-one-out and feature subset selection strategies (James et al., 2013). In the leave-one-out feature selection strategy, we test the importance of each feature individually by training the model using all features except the feature of interest. The change in model performance, measured by change in R^2 value, resulting from the loss of the feature is the measured importance of that feature. Larger performance losses (negative change in R^2) indicate a feature is important to the accuracy of the model. The main challenge of using this methodology is the effect of multicollinearity on measured feature importance. If there are highly collinear variables in the model, removing one of the collinear variables may not result in a substantial decrease in model performance. Since many features in this model such as climate and soil parameters covary, we also implement a feature subset selection where we remove all soil parameters, all climate parameters, all nitrogen management parameters and all residue management parameters from the model to test the importance of the groupings of variables.

2.4.5 Performance Metrics and Meta-Model Evaluation

To assess the accuracy of the regression meta-models, we evaluate the meta-models with the held out testing data (25% of observations). We also benchmark the biophysically-specific models against a regression meta-model built on the full set of simulation data to investigate the achieved benefit of soil and climate clustering. The coefficient of determination, R^2, was calculated as

Eq. 2 $$R^2 = (1 - \frac{u}{v})$$

where $u = \sum_{i=1}^{n} (y_{true,i} - y_{prediction,i})^2 = residual\ sum\ of\ squares$

and $v = \sum_{i=1}^{n} (y_{true,i} - \mu_{y\ true})^2 = total\ sum\ of\ squares$

Mean Absolute Error (MAE) was used as a measurement of model bias and is calculated as

Eq. 3 $$MAE = (\sum_{i=1}^{n} (|y_{prediction,i} - y_{true,i}|)/n .$$

Furthermore, the structure of the regression meta-models is an important indicator of biophysically-specific differences in the SOC response to management parameters. In order to investigate the structure of the model, we look at feature importance, regression coefficient magnitudes and directions, and modeled response curves between target variables and management parameters.

2.5 Computational Framework

All data engineering, computations, modeling, and plotting were completed with Python 3 software. Data processing and computations were completed using the Numpy and pandas data analysis libraries (Harris et al., 2020; Reback et al., 2020; McKinney, 2010). Machine learning models were built and evaluated using the scikit learn package (Pedregosa et al., 2011). Figures were produced using matplotlib (Hunter, 2007). All code available upon request.

Results

3.1 Accuracy and Model Bias

The accuracy of the biophysically-specific meta-models in predicting EPIC-IIASA simulated meanOCPD is close to perfect across all biophysical clusters and all crop types with low mean bias. Models without stratification of biophysical properties achieved an average R^2 of .97 across all crops. Meta-models trained on biophysical clusters of SimU achieve an average $R^2 = .99$ with all meta-models achieving an $R^2 > .97$ across all crops and clusters. The size of training data has a slight negative impact on the accuracy (R^2) and a positive impact on the bias (MAE) of the meta-models (Figure 4).

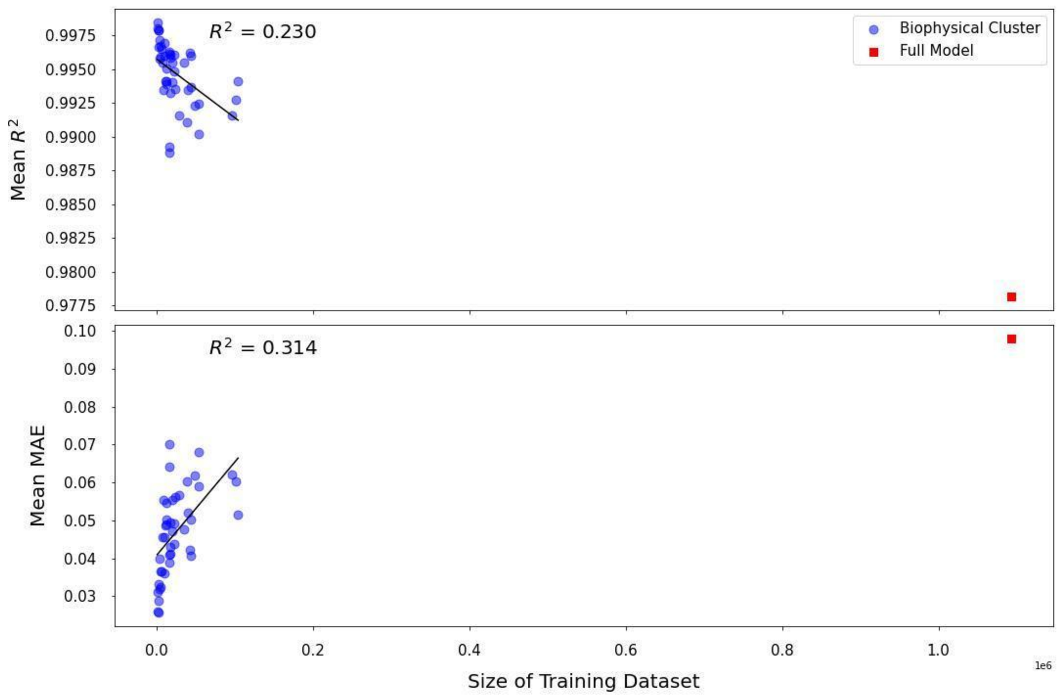


Figure 4a: Mean R^2 across all crops versus size of training dataset for each biophysically-specific model and the full model without biophysical stratification.
Figure 4b: Mean MAE across all crops versus size of training dataset for each biophysically-specific model and the full model without biophysical stratification.

We also investigated the relationship between biophysical cluster properties and meta-model accuracy. While there is little relationship between meta-model accuracy and cluster climate properties (e.g., precipitation), there is a positive relationship between accuracy and average soil field water-holding capacity within the cluster (Figure 5).

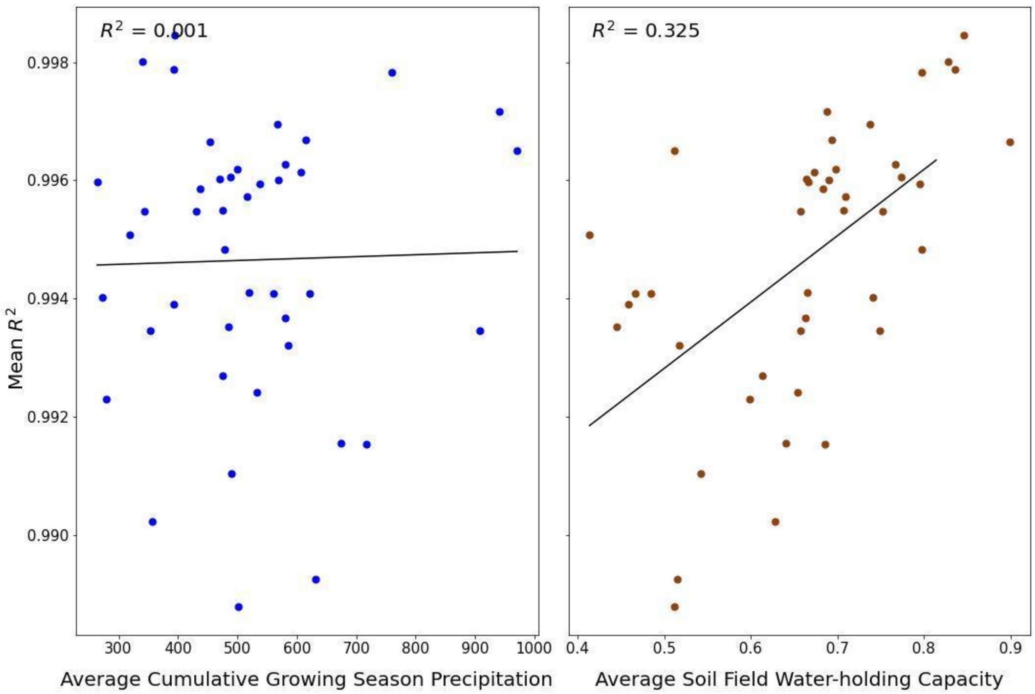


Figure 5a: Mean R² across all crops versus average cumulative growing season precipitation for each biophysically-specific meta-model

Figure 5b: Mean R² across all crops versus average soil field water-holding capacity for each biophysically-specific meta-model

3.2 Feature Importance

In our leave-one-out feature importance analysis, we find that on average, mean C content of residues applied (RSDCa) is by a wide-margin the most important variable in predicting meanOCPD across clusters and crop types with an average change in R² of -.11 resulting from removal (Figure 6). In this analysis, feature subset importance was more informative likely due to the covarying nature of features used to train our meta-models.

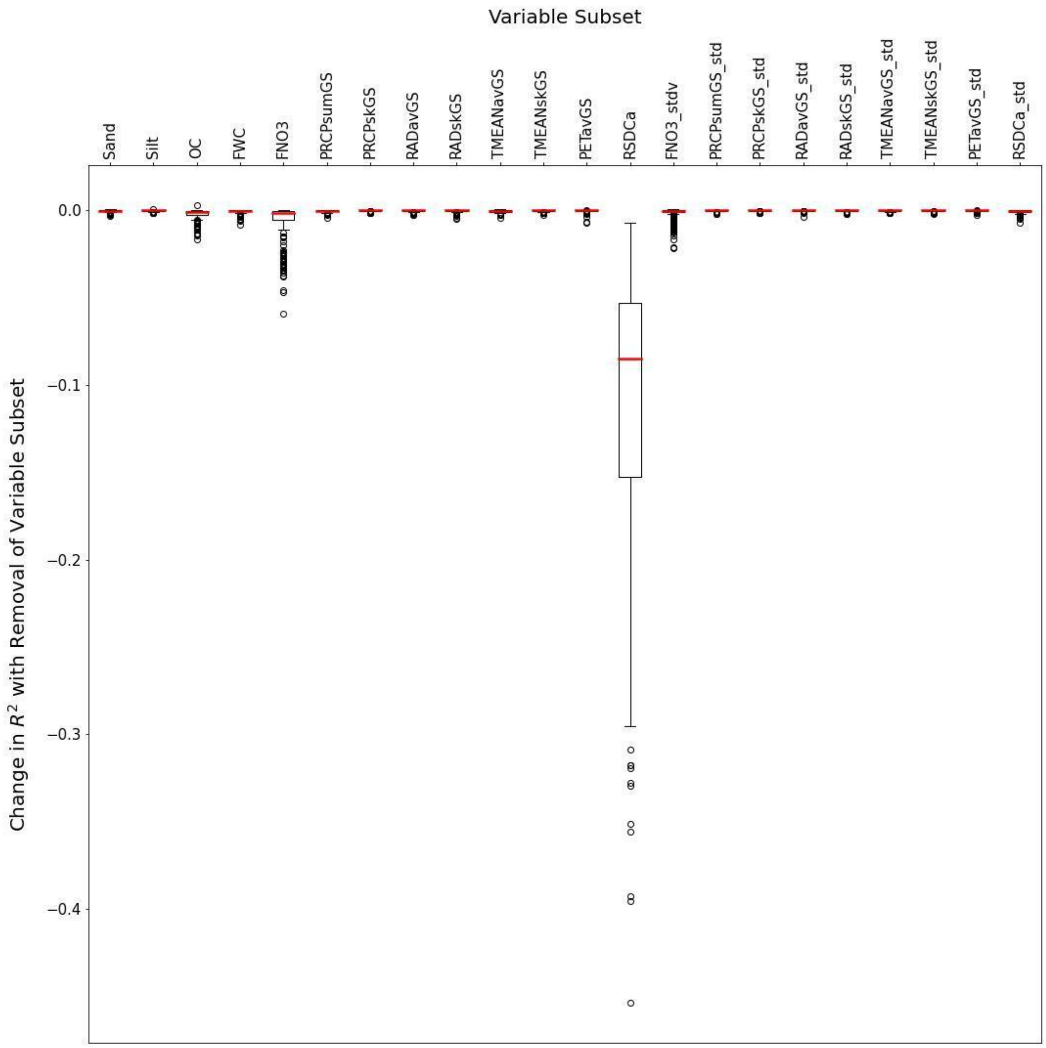


Figure 6: Leave-One-Out feature importance for all crops and biophysical-specific meta-models. Whiskers correspond to 1.5 times the upper and lower quartiles. Box corresponds to the interquartile range, red line corresponds to median value.

On average across all biophysically-specific meta-models, we find that the subset of residue features is the most important in predicting meanOCPD over the 35 year time period followed by climate, soil, and N fertilizer feature subsets (Figure 7). Removal of residue features results in an average change in R^2 of -.45. We find that the subset of residue features also has the highest variation in feature performance across all crops and biophysical clusters.

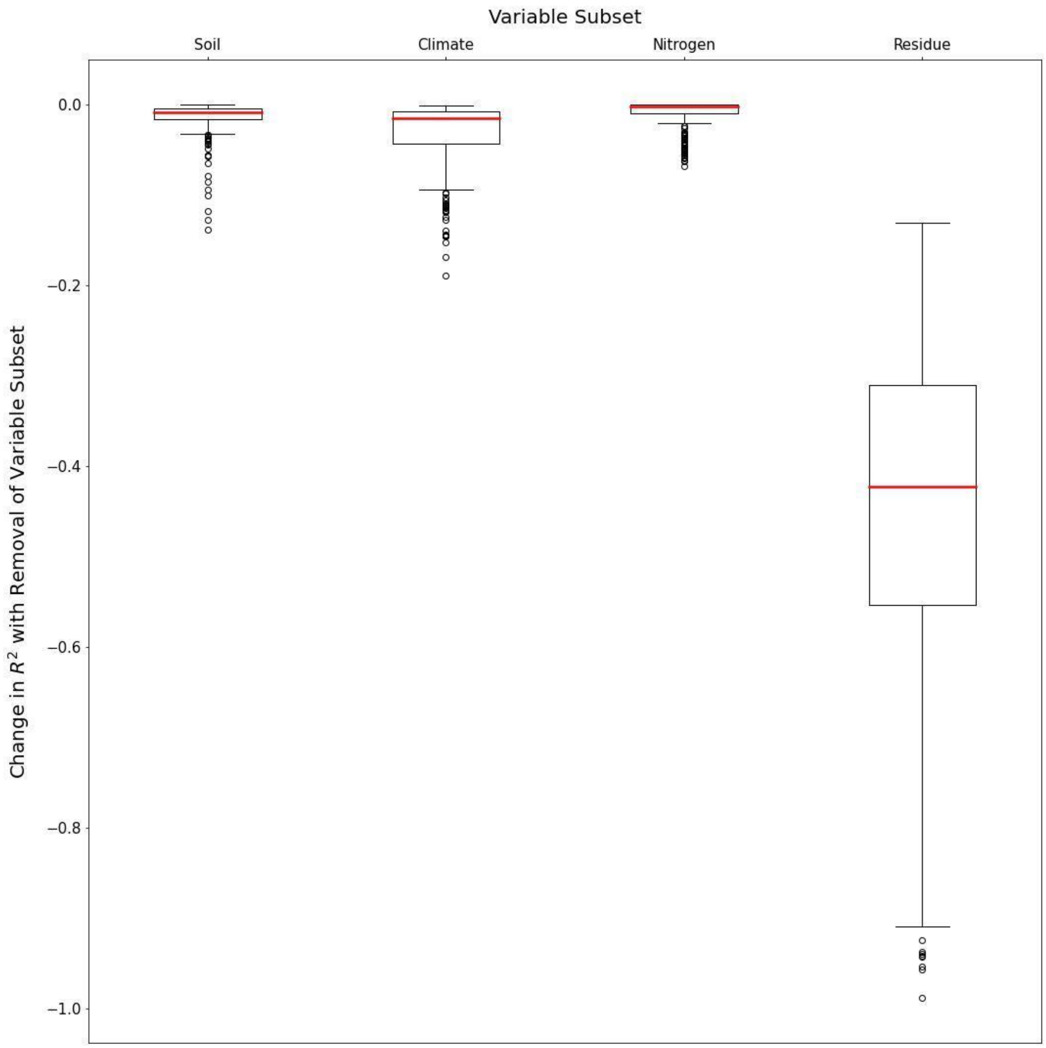


Figure 7: Feature subset importance for all crops and biophysical-specific meta-models. Whiskers correspond to 1.5 times the upper and lower quartiles. Box corresponds to the interquartile range, red line corresponds to median value.

There are distinct differences in feature subset importance which are dependent on biophysical cluster properties. We find that the importance of residue features is larger among clusters with lower cumulative precipitation during the growing season (Figure 8). The importance of other feature subsets have little relationship with cluster climate parameters. We also find that the importance of climate and soil feature subsets is larger among clusters with low soil field water-holding capacity (Figure 9).

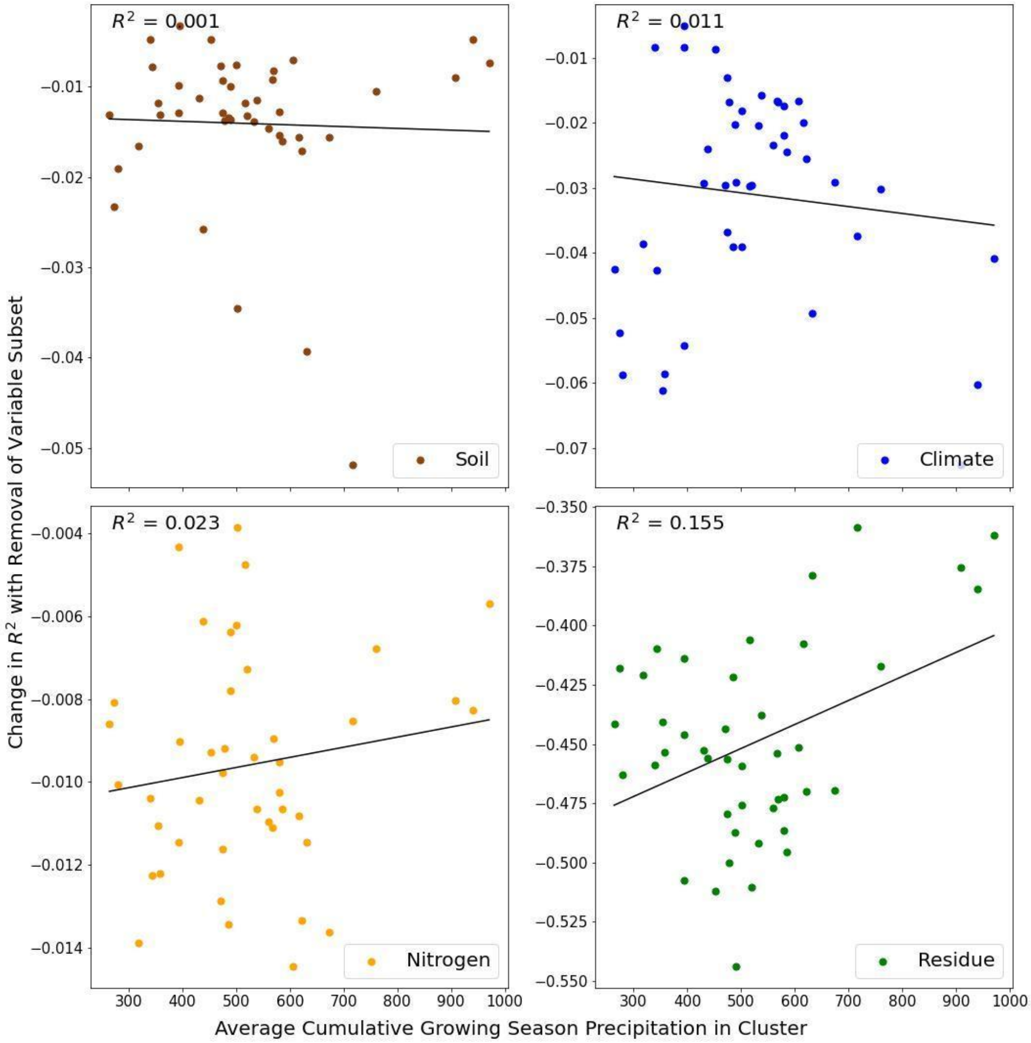


Figure 8: Average feature subset importance across all crops of soil, climate, nitrogen, and residue feature subsets versus average cumulative growing season precipitation for each biophysically-specific meta-model.

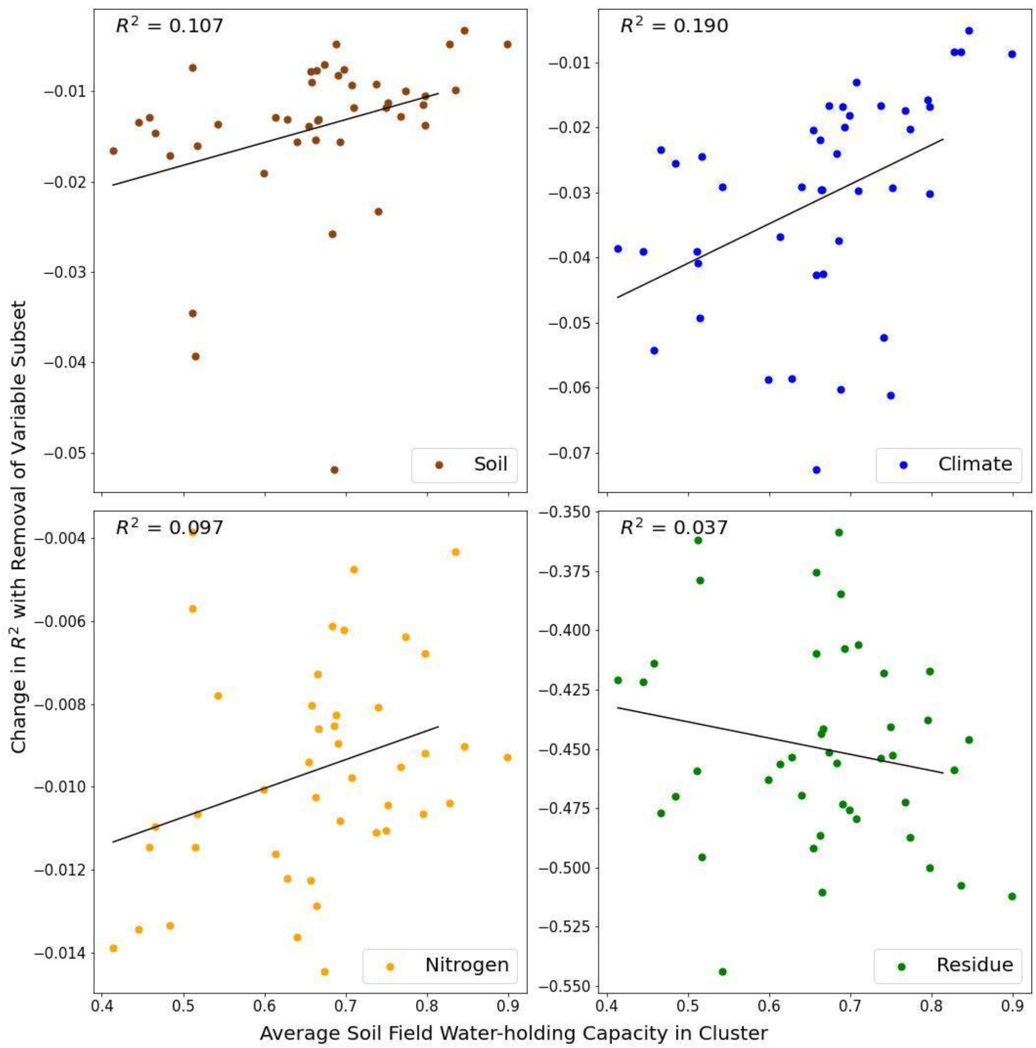


Figure 9: Average feature subset importance across all crops of soil, climate, nitrogen, and residue feature subsets versus average soil field water-holding capacity for each biophysically-specific meta-model.

We find that feature subset importance also varies by crop - the average change in R^2 across biophysically-specific meta-models with removal of the residue feature subset is the largest when modeling soybean growth (R^2 = .75) and smallest when modeling winter wheat (R^2 = .29) (Figure 10). We can also visualize the crop-specific differences in feature subset importance geospatially (for more crop-specific feature importance plots, see Appendix). For example, when predicting meanOCPD for maize versus winter wheat, there are substantial spatial distinctions in the importance of residue features (Figure 11).

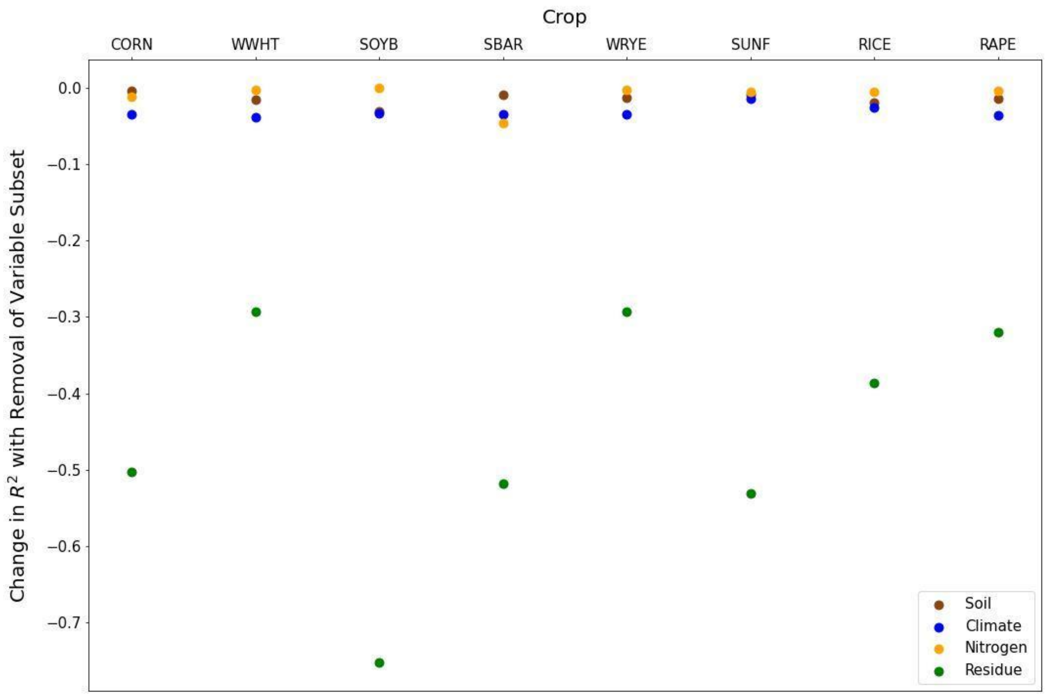


Figure 10: Average feature subset importance across biophysical clusters for each simulated crop.

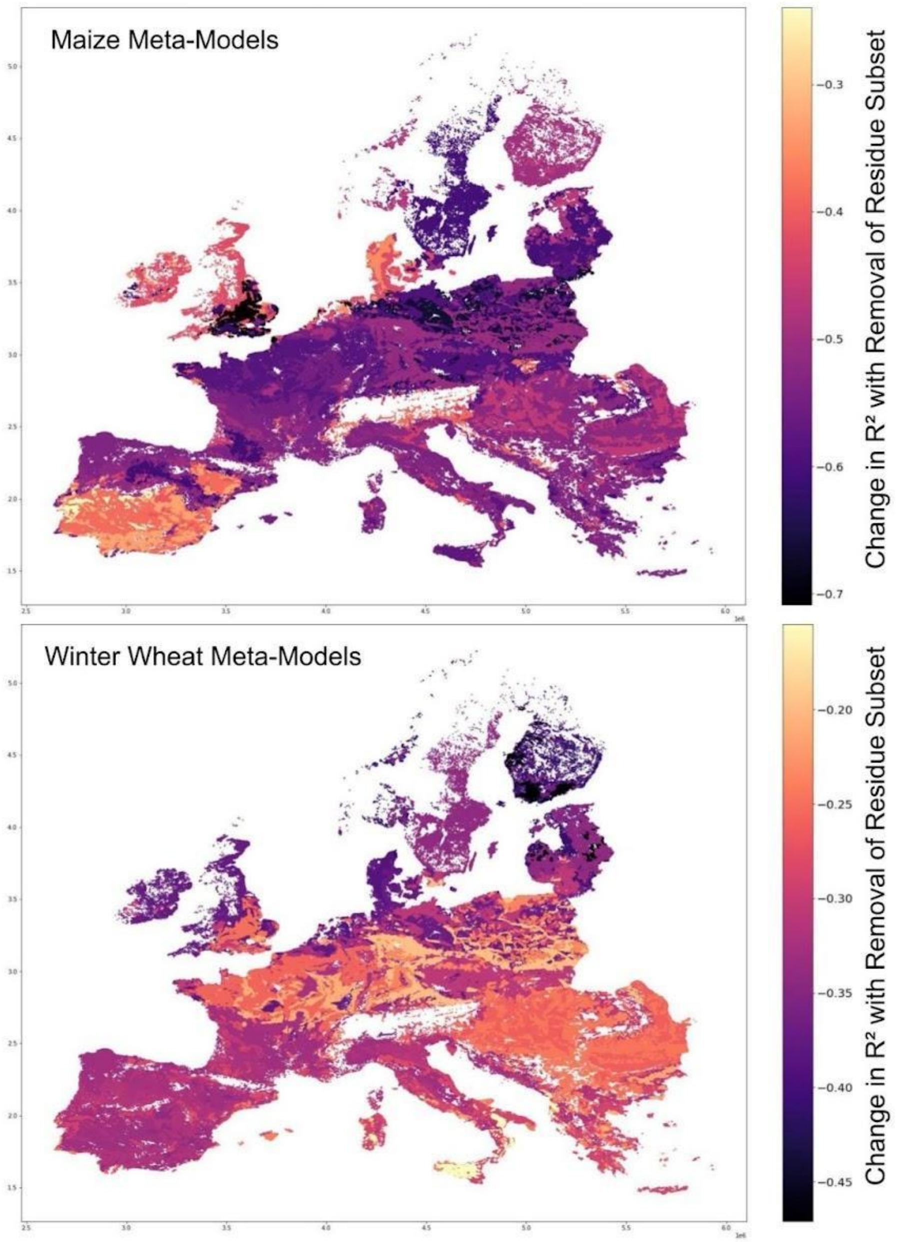


Figure 11: Geospatial differences of importance of residue features between meta-models for different crop types across all biophysical clusters.

3.3 Model Structure

Of great interest are the modeled response relationships between features and meanOCPD. Due to the interpretable nature of the multiple polynomial regression model, we can unpack the meta-model structure through model coefficients. While we cannot directly interpret meta-model coefficients as effects on the target variable, we can utilize coefficients to understand the structure of the meta-model and the nature of the feature effects. We find that among all biophysically-specific meta-models and crop types, there is agreement on the nature of the response relationship between mean RSDCa and meanOCPD (Figure 12). Although the response curve includes linear and quadratic terms, the response of meanOCPD to mean RSDCa is strongly linear across all biophysically-specific meta models and crop types, as are the meta-models without biophysical stratification.

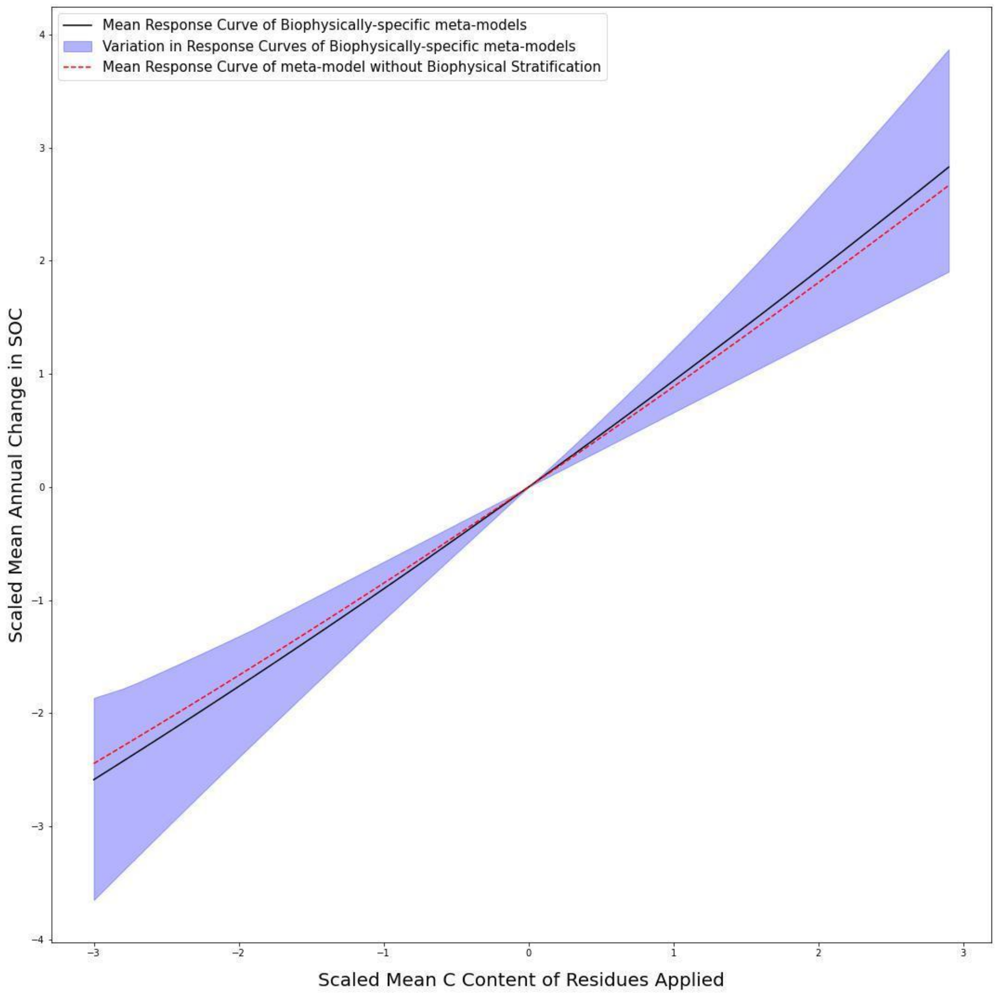


Figure 12: Response relationships between RSDCa and meanOCPDΔ, variables are scaled to have unit variance and mean centered at zero.

We find less overall agreement on the response relationships between meanOCPD and climatic and soil properties. Response relationships within given climate clusters are often similar, with minor distinctions resulting from the different underlying soil texture clusters (black lines in Figure 13 panels) of the biophysically-specific meta-models. For example, the response relationship between mean cumulative growing season precipitation and meanOCPD is very similar across soils in the continental climate region (denoted in orange, Figure 13) while in other climates such as the coastal mediterranean climate (denoted in purple, Figure 13) there are wider discrepancies between response relationships depending on the soil texture cluster.

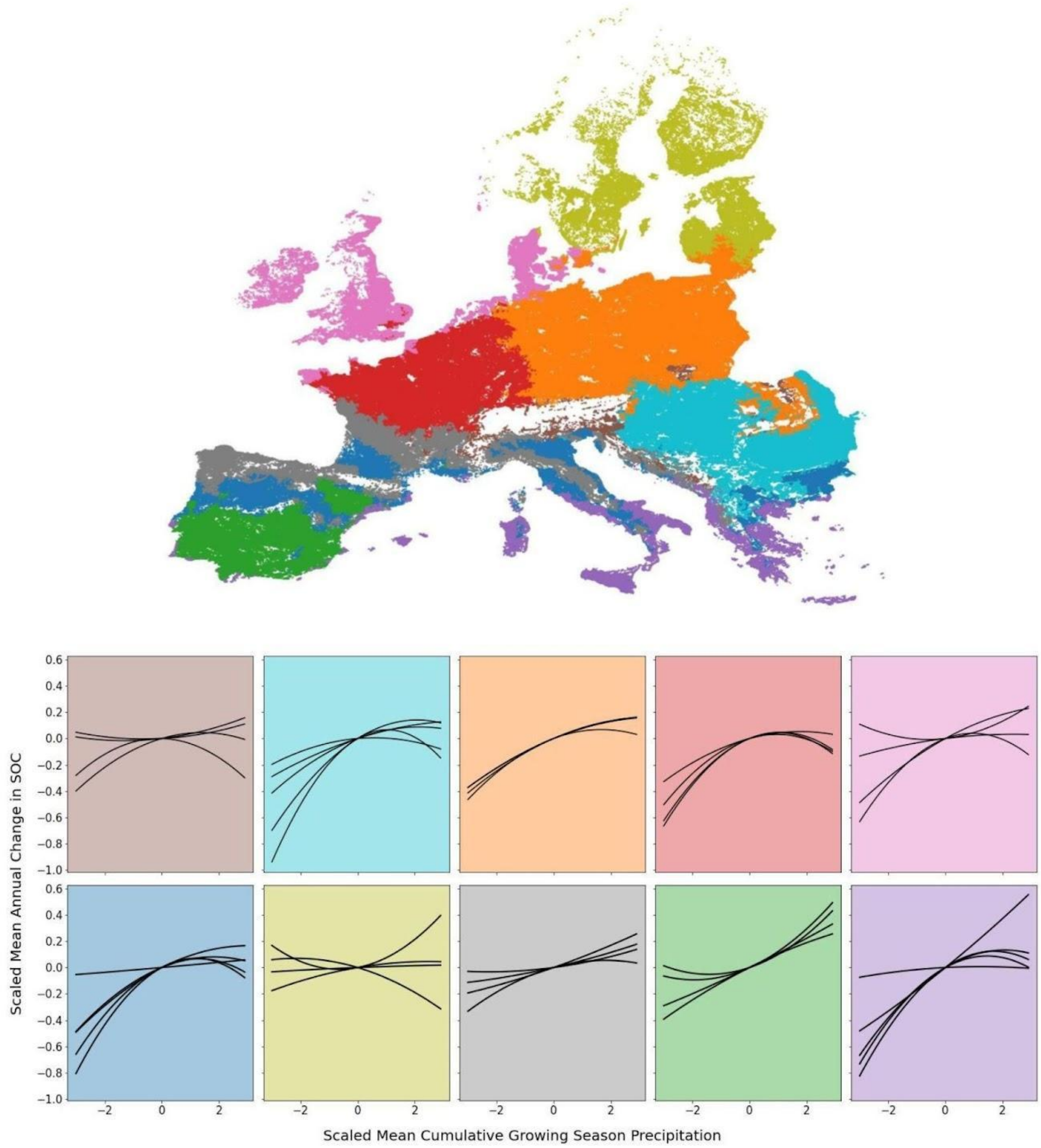


Figure 13: Response relationships between GSsumPRCP and meanOCPDΔ. Each subplot shows multiple climate x soil specific response curves for the given climate cluster, denoted by color. Variables are scaled to have unit variance and mean centered at zero.

Discussion and Conclusions

4.1 Model Accuracy

The high degree of accuracy achieved by all meta-models, including those with and without biophysical-stratification, shows the ability to utilize highly interpretable meta-models to simplify and explore the EPIC-IIASA GAM. In a prior study, Folberth et al. (2019) used an ensemble method to build a meta-model of EPIC-IIASA GAM in order to spatio-temporally downscale yield predictions with a high degree of accuracy. The multiple polynomial regression model used in this study, while less sophisticated than other models such as that used in Folbert et al. (2019), does not suffer from the oft-maligned black-box nature of machine learning models (Rudin, 2019). We also show that the accuracy of our meta-model does not suffer as a result of the simplified statistical model used. Furthermore, the results of our analysis show that biophysical stratification does modestly improve the accuracy and reduce the bias compared to the meta-models built without biophysical stratification. By biophysically stratifying our training data and building separate meta-models on these subsets of observations, our meta-models may learn more nuanced response relationships which are unique to the biophysical conditions, thus resulting in higher accuracy.

4.2 Feature Importance

Feature importance analysis allows us to better understand the meta-model prediction mechanisms and signals which variables within EPIC are critical to estimation of SOC dynamics. From our feature analysis, we find that the residue feature subset is the most important to meta-model accuracy and has the highest mean feature subset importance and highest variance in feature subset importance across biophysical clusters and across crop types. Since residue management acts as a direct input of organic material to the cropping system, residue features should have a substantial impact on prediction of EPIC-IIASA GAM meanOCPD values. The lower feature subset importance of Nitrogen fertilizer features across all crop types and biophysical clusters in our study could be because these are partly collinear with residue C inputs since more N means more root residues at a site, and more straw at each retention intensity. The importance of the N fertilization feature would become clearer if only a certain retention scenario (e.g., no residue retention) was analyzed alone. While nitrogen inputs to the cropping system may have an effect on SOC dynamics, studies have shown that the effect of nitrogen inputs is largely dependent on the quantity, quality, and source of residues (e.g. residues from roots only) (Lugato et al., 2006).

Trends in feature subset importance point to the impact of biophysical parameters on differences between meta-models. As shown in Figures 8, as average cumulative growing season precipitation of clusters increases, the subset feature importance of residue features decreases. Thus residue features are more important in predicting meanOCPD of clusters with low rainfall during the growing season. Additionally, as soil field water-holding capacity increases, the subset feature importance of soil and climate subsets decreases. In sum, as the biophysical cluster edges towards more drought-prone conditions (low water storage and low rainfall), the importance of residue, soil, and climate parameters increases. This may be a result of a decrease in plant litter input as a result of lower yields. These trends in feature subset importance signal important distinctions between biophysical clusters and the dynamics of SOC represented by the EPIC-IIASA model and highlight the utility of biophysical stratification to better understand meta-model dynamics within specific clusters of interest.

4.3 Model Structure

The biophysically-specific meta-models we build pay specific attention to the unique soil and climate dynamics which may affect the responses of SOC to management choices. These biophysically-specific meta-models are also built in a way that allows for further downscaling through increased stratification of climate or soil properties resulting in a larger number of clusters. While this analysis clusters the entire region of Europe, clustering could be performed within a given country to investigate finer-scale SOC dynamics. The interpretable model structure chosen for this analysis, the multiple polynomial regression model, allows us to investigate response dynamics between parameters and SOC dynamics through the functional model form and learned coefficients. As is shown in Figures 12 and 13, we can pull out response curves of interest, such as the response between residue C content and meanOCPD, to better understand the mechanisms of the EPIC-IIASA model, and how the model changes given biophysical stratification.

The linear nature of meanOCPD response to residue C content within our meta-models is likely a facet of the EPIC model which does not specify a plateau point of SOC saturation, thus the meanOCPD can continue to grow without limit. The close agreement of biophysically-specific meta-model residue C content response curves shows that biophysical parameters have little effect on the nature of the response relationship. For other response dynamics, such as the response of meanOCPD to precipitation during the growing season, biophysical stratification produces curves which differ between climate and soil clusters. Climate clusters drive most of the major differences in response curves, and soil clusters sometimes add additional variation within a climate cluster. Under certain climate conditions, such as the continental zone highlighted in orange in Figure 13, soil clusters do not contribute to a substantial amount of variation in response dynamics within the climate cluster. In other regions, such as the mediterranean regions highlighted in blue and purple, we see more variation as a result of soil clustering. Furthermore, across climates, response curves vary in their steepness and shape with some curves showing more linear responses and others showing a plateauing effect.

4.4 Outlook

The approach employed in this analysis achieves high accuracy and provides high interpretability, thus allowing for prediction of EPIC-IIASA model output and the exploration of the model mechanisms. This technique is a scalable way to parse the complexities of the EPIC model and to explore the impact of biophysical parameters on important relationships within the model. First, it allows the downscaling of large-scale EPIC-IIASA projections to finer regional scales by supplying spatially-detailed and accurate input information for the most sensitive features. Second, this study identifies a framework to increase the capacity of GAMs for regional SOC modeling by combining factorial simulations with statistical response modeling. Lastly, for end-users of EPIC-IIASA output, these meta-models point to nuances in the model which may help guide the appropriate use of crop model simulations by highlighting areas where the model may misrepresent known ground dynamics. This could, in turn, help reduce the challenges of calibration and validation of EPIC-IIASA by providing a simplified version of the model to survey. For example, if an end-user knows that a specific area should see a large impact of climate on SOC dynamics, they can utilize the meta-model to look at nature and magnitude of response relationships to see where sources of error in the model may originate from. Our novel approach of biophysical-stratification highlights the importance of the soil and climate systems in driving EPIC-IIASA model dynamics. The improved accuracy and divergences in meta-model structure as a result of this stratification signal a benefit of explicitly investigating the ways that biophysical characteristics specify SOC responses.

Appendix

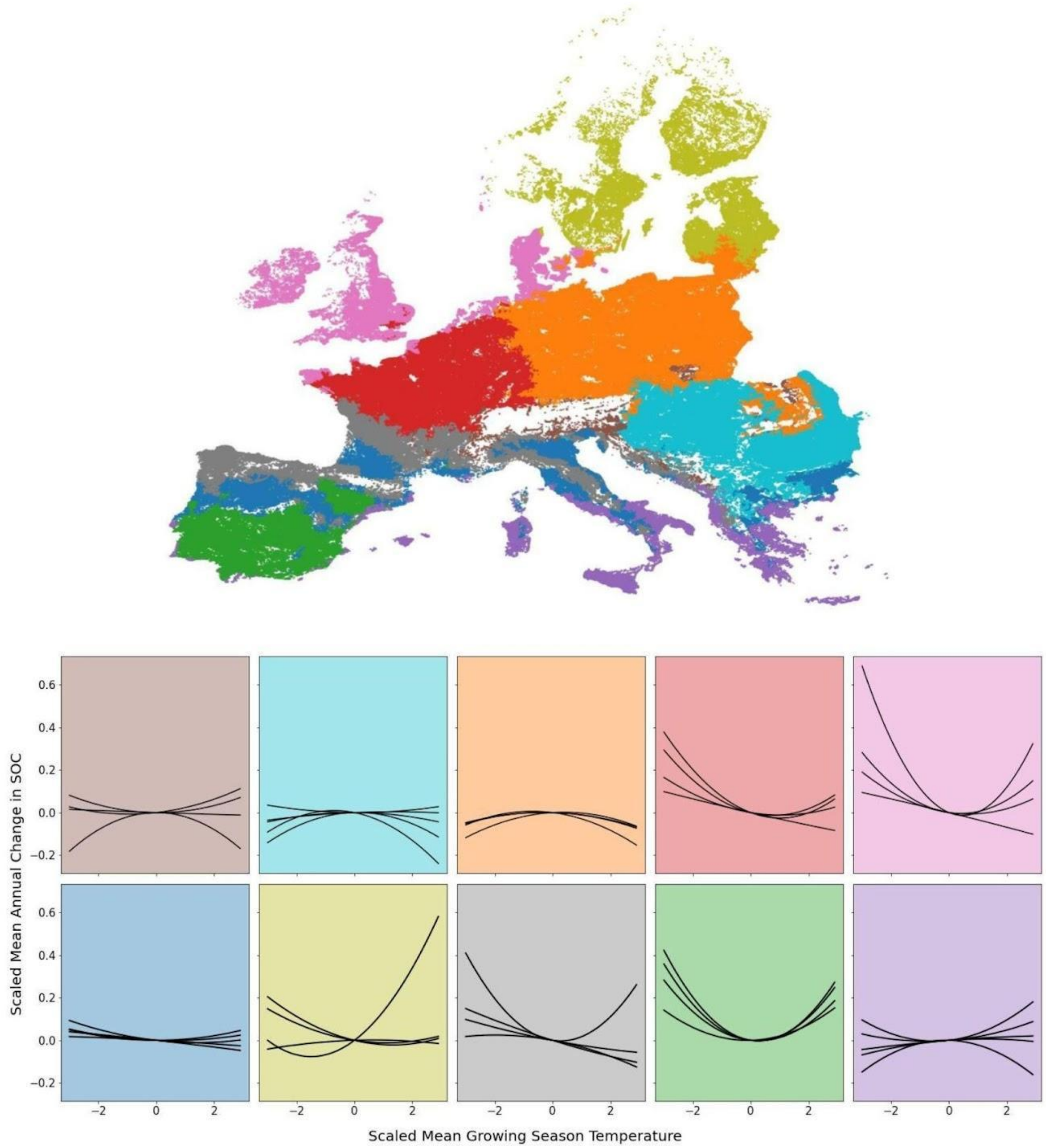
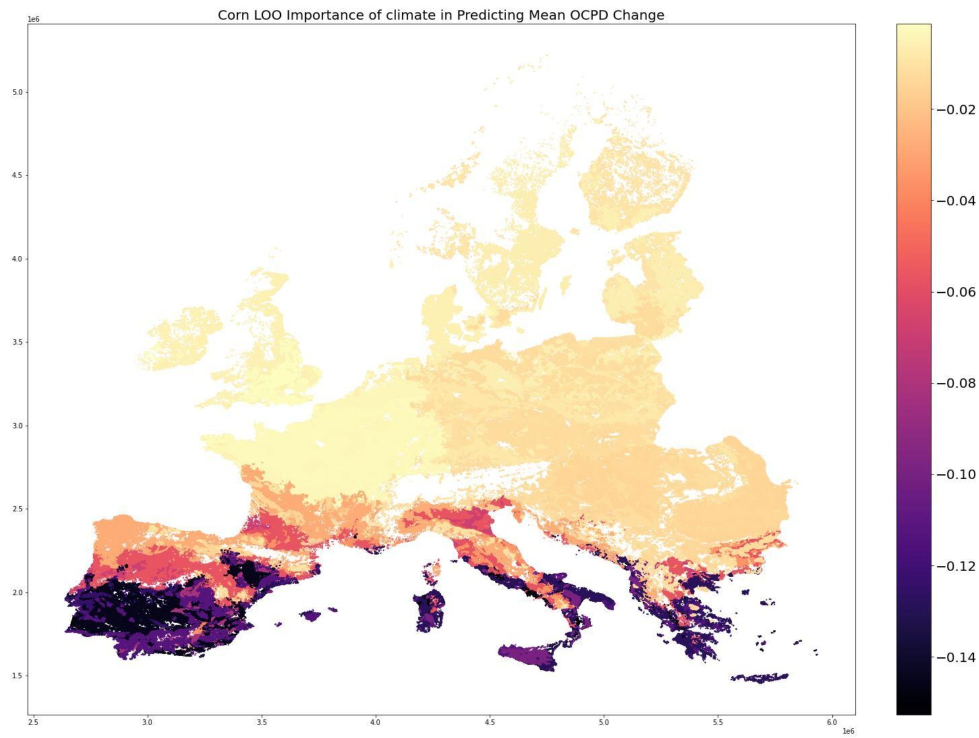
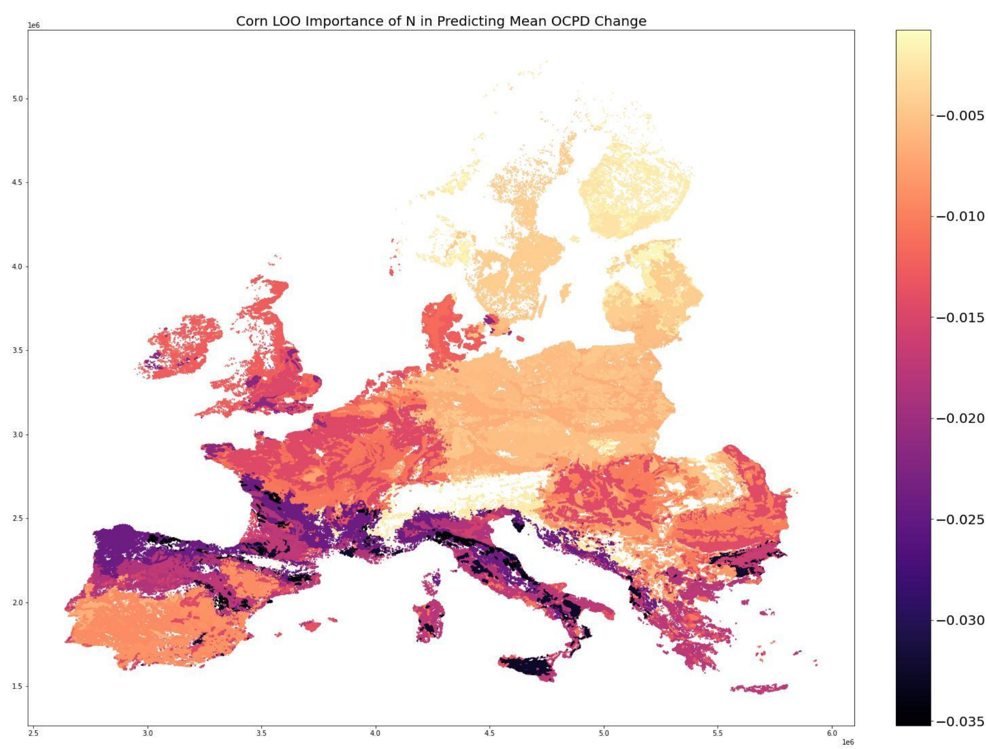
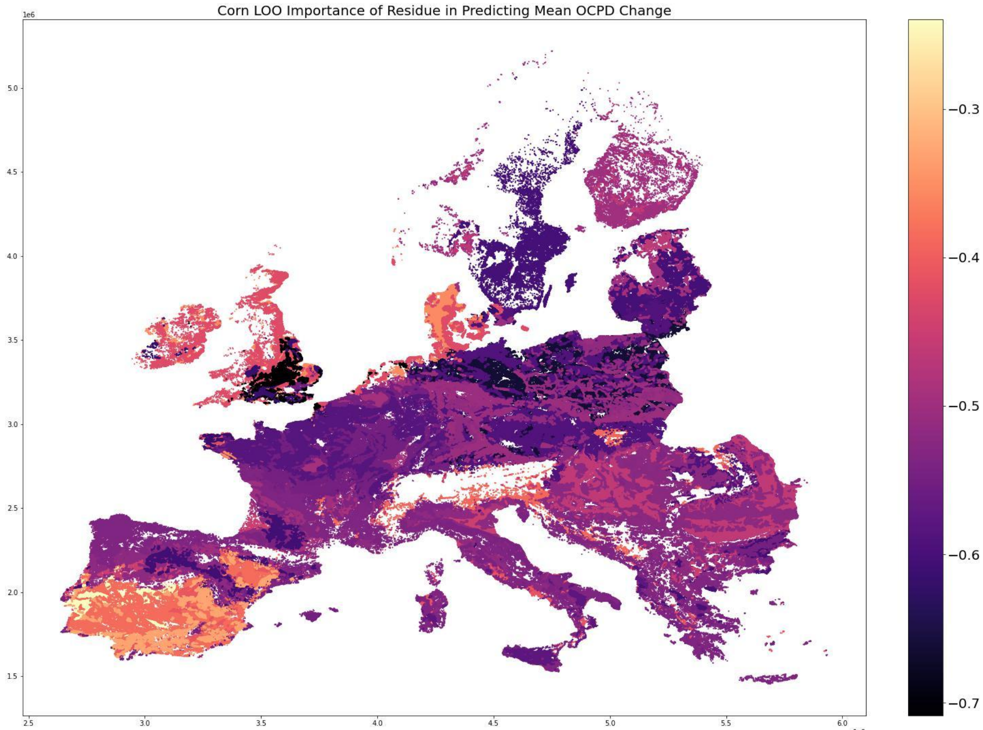


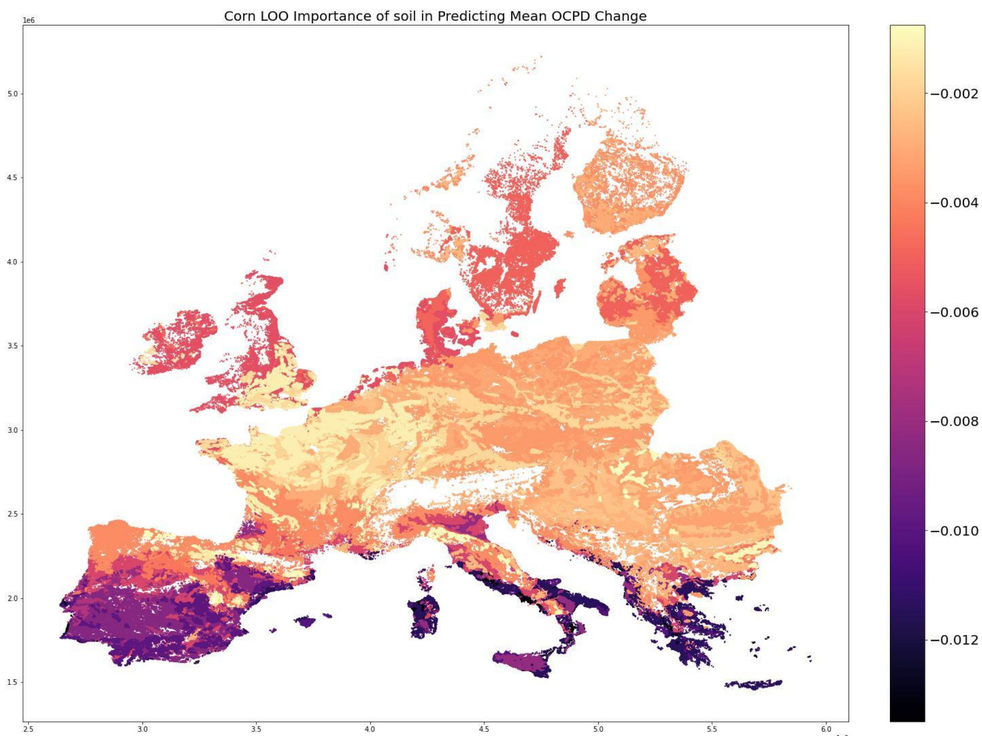
Figure A.1: Response relationships between GSavTEMP and meanOCPDΔ. Each subplot shows multiple climate x soil specific response curves for the given climate cluster, denoted by color. Variables are scaled to have unit variance and mean centered at zero.

Figures A.2, A.3, A.4, A.5: Feature subset importance of climate, N, residue, and soil feature subsets for predicting meanOCPDΔ in maize systems. Color shown indicates Change in R^2 resulting from removal of feature subset

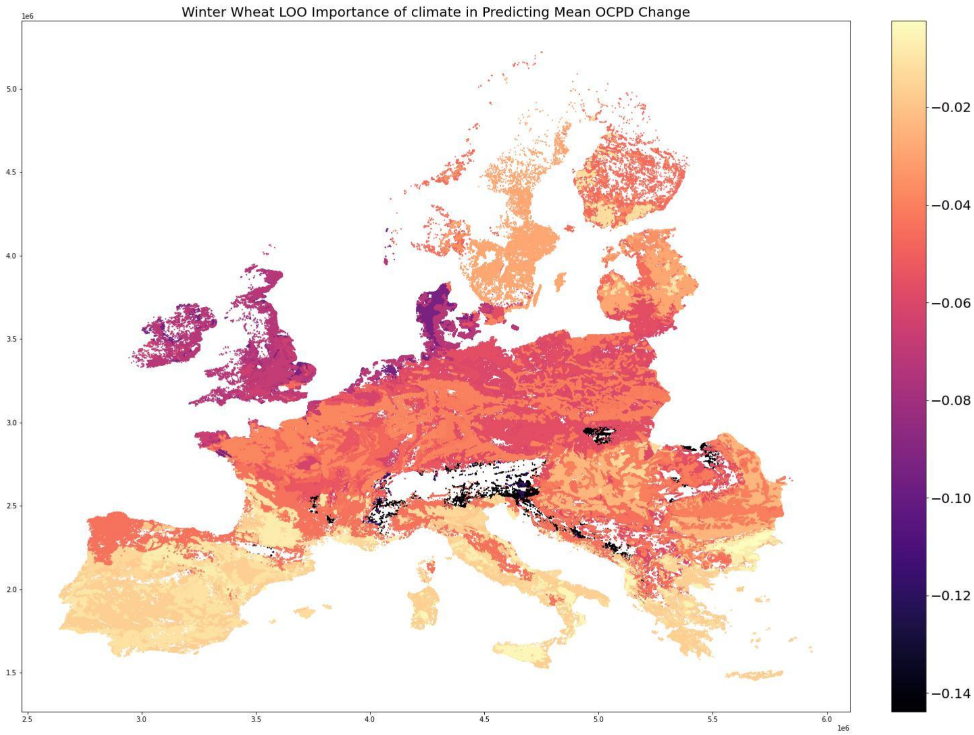

Corn LOO Importance of climate in Predicting Mean OCPD Change
1e6
5.0
4.5
4.0
3.5
3.0
2.5
2.0
1.5
2.5
3.0
3.5
4.0
4.5
5.0
5.5
6.0
1e6
−0.02
−0.04
−0.06
−0.08
−0.10
−0.12
−0.14


Corn LOO Importance of N in Predicting Mean OCPD Change
1e6
5.0
4.5
4.0
3.5
3.0
2.5
2.0
1.5
2.5
3.0
3.5
4.0
4.5
5.0
5.5
6.0
1e6
−0.005
−0.010
−0.015
−0.020
−0.025
−0.030
−0.035

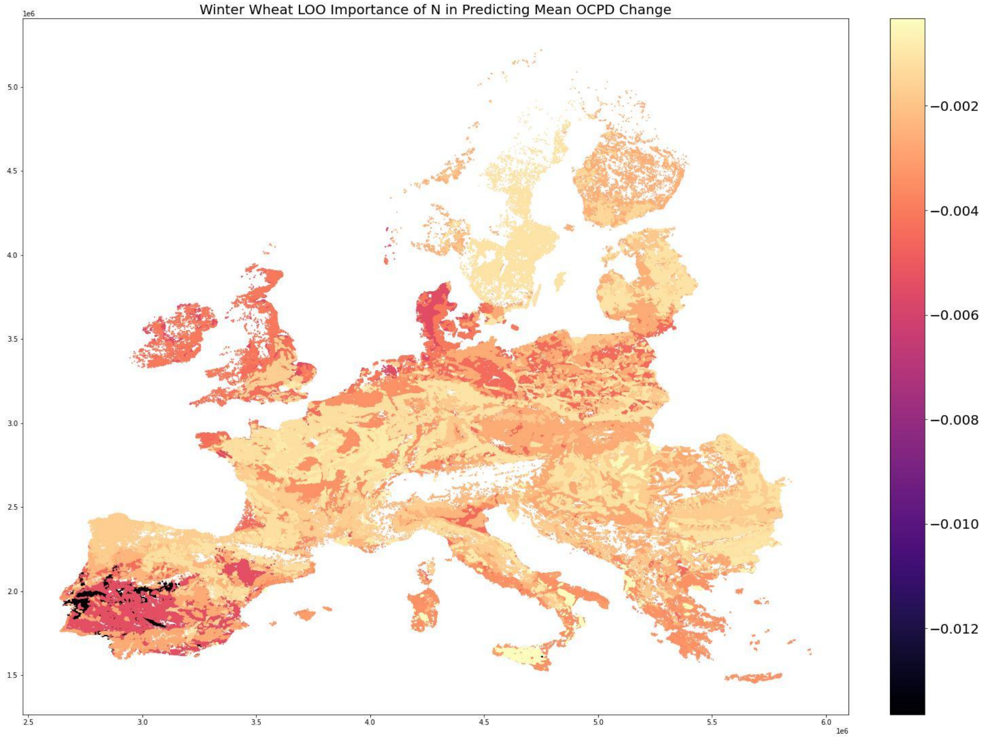

Corn LOO Importance of Residue in Predicting Mean OCPD Change
1e6
5.0
4.5
4.0
3.5
3.0
2.5
2.0
1.5
2.5
3.0
3.5
4.0
4.5
5.0
5.5
6.0
1e6
−0.3
−0.4
−0.5
−0.6
−0.7


Corn LOO Importance of soil in Predicting Mean OCPD Change
1e6
5.0
4.5
4.0
3.5
3.0
2.5
2.0
1.5
2.5
3.0
3.5
4.0
4.5
5.0
5.5
6.0
1e6
−0.002
−0.004
−0.006
−0.008
−0.010
−0.012

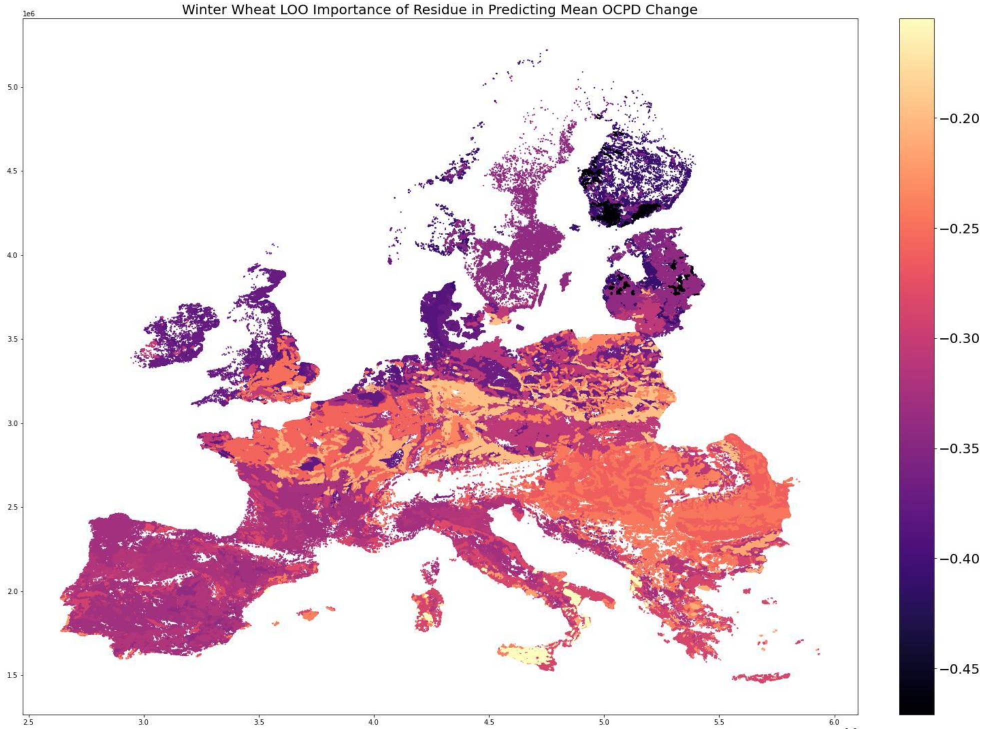
Figures A.6, A.7, A.8, A.9: Feature subset importance of climate, N, residue, and soil feature subsets for predicting meanOCPDΔ in winter wheat systems. Color shown indicates Change in R^2 resulting from removal of feature subset

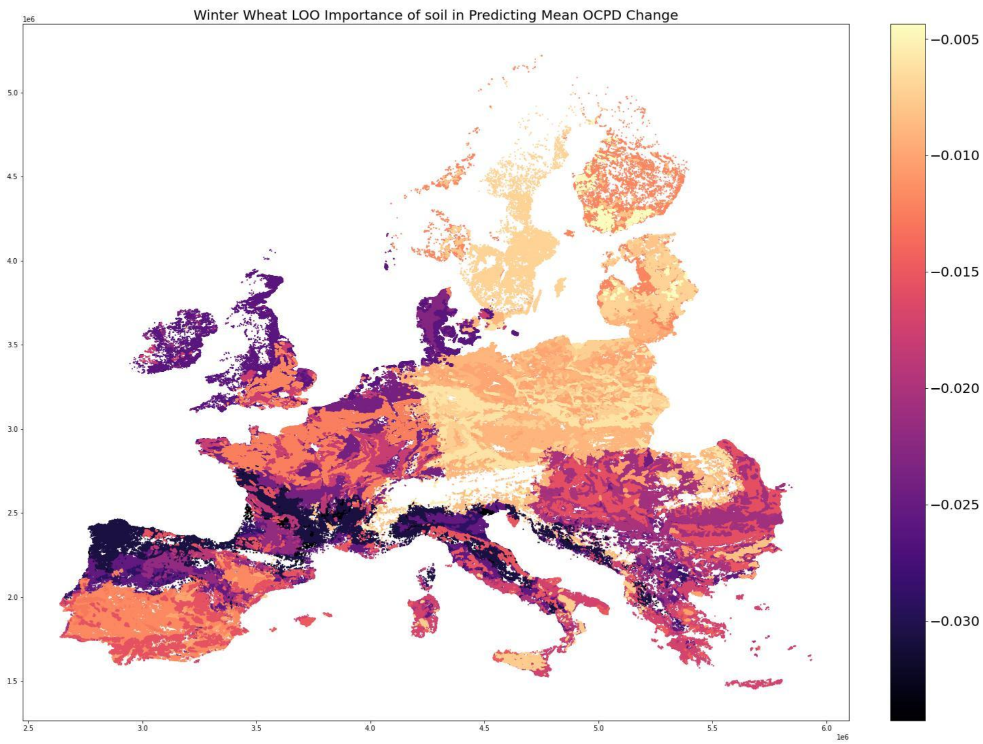


Winter Wheat LOO Importance of N in Predicting Mean OCPD Change



Winter Wheat LOO Importance of Residue in Predicting Mean OCPD Change




Winter Wheat LOO Importance of soil in Predicting Mean OCPD Change
1e6
5.0
4.5
4.0
3.5
3.0
2.5
2.0
1.5
2.5
3.0
3.5
4.0
4.5
5.0
5.5
6.0
1e6
−0.005
−0.010
−0.015
−0.020
−0.025
−0.030

References

Attard, Eléonore, et al. "Delayed and asymmetric responses of soil C pools and N fluxes to grassland/cropland conversions." Soil Biology and Biochemistry 97 (2016): 31-39.

Balkovič, Juraj, et al. "Pan-European crop modelling with EPIC: Implementation, up-scaling and regional crop yield validation." Agricultural Systems 120 (2013): 61-75.

Balkovič, Juraj, et al. "Global wheat production potentials and management flexibility under the representative concentration pathways." Global and Planetary Change 122 (2014): 107-121.

Balkovič, Juraj, et al. "Impacts and uncertainties of+ 2 C of climate change and soil degradation on European crop calorie supply." Earth's future 6.3 (2018): 373-395.

Balkovič, Juraj, et al. "Verifiable soil organic carbon modelling to facilitate regional reporting of cropland carbon change: A test case in the Czech Republic." Journal of Environmental Management 274 (2020): 111206.

Balsalobre-Lorente, Daniel, et al. "Do agricultural activities induce carbon emissions? The BRICS experience." Environmental Science and Pollution Research 26.24 (2019): 25218-25234.

Baumann, Matthias, et al. "Carbon emissions from agricultural expansion and intensification in the Chaco." Global change biology 23.5 (2017): 1902-1916.

Bates, Douglas M., and Donald G. Watts. Nonlinear regression analysis and its applications. Wiley, 1988.

Blanc, Élodie. "Statistical emulators of maize, rice, soybean and wheat yields from global gridded crop models." Agricultural and Forest Meteorology 236 (2017): 145-161.

Breiman, Leo. "Bagging predictors." Machine learning 24.2 (1996): 123-140.

Bryant, K. J., et al. "Simulating corn yield response to irrigation timings: Validation of the EPIC model." Journal of Production Agriculture 5.2 (1992): 237-242.

Bullock, Donald G., and David S. Bullock. "Quadratic and quadratic-plus-plateau models for predicting optimal nitrogen rate of corn: A comparison." Agronomy Journal 86.1 (1994): 191-195.

Caliński, Tadeusz, and Jerzy Harabasz. "A dendrite method for cluster analysis." Communications in Statistics-theory and Methods 3.1 (1974): 1-27.

Chabbi, Aligning, et al. "Aligning agriculture and climate policy." Nature climate change 7.5 (2017): 307-309.

Chong, Il-Gyo, and Chi-Hyuck Jun. "Performance of some variable selection methods when multicollinearity is present." Chemometrics and intelligent laboratory systems 78.1-2 (2005): 103-112.

Daryanto, Stefani, Lixin Wang, and Pierre-André Jacinthe. "Global synthesis of drought effects on maize and wheat production." PloS one 11.5 (2016): e0156362.

Davies, David L., and Donald W. Bouldin. "A cluster separation measure." IEEE transactions on pattern analysis and machine intelligence 2 (1979): 224-227.

Ding, Chris, and Xiaofeng He. "K-means clustering via principal component analysis." Proceedings of the twenty-first international conference on Machine learning. 2004.

Eglin, Thierry, et al. "Historical and future perspectives of global soil carbon response to climate and land-use changes." Tellus B: Chemical and Physical Meteorology 62.5 (2010): 700-718.

Evenson, Robert E., and Douglas Gollin. "Assessing the impact of the Green Revolution, 1960 to 2000." science 300.5620 (2003): 758-762.

Ewert, F., et al. "Future scenarios of European agricultural land use: I. Estimating changes in crop productivity." Agriculture, Ecosystems & Environment 107.2-3 (2005): 101-116.

FAO (2016) Food and Agriculture Organization: The State of Food and Agriculture. http://www.fao.org/3/a-i6030e.pdf

Francis, Hannah R., Ting Fung Ma, and Matthew D. Ruark. "Toward a standardized statistical methodology comparing optimum nitrogen rates among management practices: A bootstrapping approach." Agricultural & Environmental Letters 6.2 (2021): e20045.

Frank, Stefan, et al. "The dynamic soil organic carbon mitigation potential of European cropland." Global Environmental Change 35 (2015): 269-278.

Franke, James A., et al. "The GGCMI Phase 2 emulators: global gridded crop model responses to changes in CO 2, temperature, water, and nitrogen (version 1.0)." Geoscientific Model Development 13.9 (2020): 3995-4018.

Fuss, Sabine, et al. "Negative emissions—Part 2: Costs, potentials and side effects." Environmental Research Letters 13.6 (2018): 063002.

Gautam, H. R., M. L. Bhardwaj, and Rohitashw Kumar. "Climate change and its impact on plant diseases." Current science (2013): 1685-1691.

Gassman, Philip W., et al. "Historical development and applications of the EPIC and APEX models." 2004 ASAE Annual Meeting. American Society of Agricultural and Biological Engineers, 2004.

Guo, Lanbin B., and Roger M. Gifford. "Soil carbon stocks and land use change: a meta analysis." Global change biology 8.4 (2002): 345-360.

Harris, Charles R., et al. "Array programming with NumPy." Nature 585.7825 (2020): 357-362.

Houghton, Richard A., et al. "Carbon emissions from land use and land-cover change." Biogeosciences 9.12 (2012): 5125-5142.

Hristov, J. et al. "Analysis of climate change impacts on EU agriculture by 2050", EUR 30078 EN, Publications Office of the European Union, Luxembourg, 2020, ISBN 978-92-76-10617-3, doi:10.2760/121115, JRC119632.

Hunter, John D. "Matplotlib: A 2D graphics environment." Computing in science & engineering 9.03 (2007): 90-95.

Iglesias, Ana, et al. "From climate change impacts to the development of adaptation strategies: challenges for agriculture in Europe." Climatic Change 112.1 (2012): 143-168.

Izaurralde, R. C., et al. "Simulating soil C dynamics with EPIC: Model description and testing against long-term data." Ecological Modelling 192.3-4 (2006): 362-384.

Jägermeyr, Jonas, et al. "Climate impacts on global agriculture emerge earlier in new generation of climate and crop models." Nature Food 2.11 (2021): 873-885.

James, Gareth, et al. An introduction to statistical learning. Vol. 112. New York: springer, 2013.

Jastrow, Julie D., James E. Amonette, and Vanessa L. Bailey. "Mechanisms controlling soil carbon turnover and their potential application for enhancing carbon sequestration." Climatic Change 80.1 (2007): 5-23.

Keel, Sonja G., et al. "Large uncertainty in soil carbon modelling related to method of calculation of plant carbon input in agricultural systems." European Journal of Soil Science 68.6 (2017): 953-963.

Kim, Wonsik, Toshichika Iizumi, and Motoki Nishimori. "Global patterns of crop production losses associated with droughts from 1983 to 2009." Journal of Applied Meteorology and Climatology 58.6 (2019): 1233-1244.

Kohavi, Ron. "A study of cross-validation and bootstrap for accuracy estimation and model selection." Ijcai. Vol. 14. No. 2. 1995.

Lal, Rattan. "Carbon emission from farm operations." Environment international 30.7 (2004): 981-990.

Likas, Aristidis, Nikos Vlassis, and Jakob J. Verbeek. "The global k-means clustering algorithm." Pattern recognition 36.2 (2003): 451-461.

Lobell, David B., Wolfram Schlenker, and Justin Costa-Roberts. "Climate trends and global crop production since 1980." Science 333.6042 (2011): 616-620.

Lugato, E., A. Berti, and L. Giardini. "Soil organic carbon (SOC) dynamics with and without residue incorporation in relation to different nitrogen fertilisation rates." Geoderma 135 (2006): 315-321.

Lugato, Emanuele, et al. "A new baseline of organic carbon stock in European agricultural soils using a modelling approach." Global change biology 20.1 (2014): 313-326.

Manzoni, Stefano, and Amilcare Porporato. "Soil carbon and nitrogen mineralization: Theory and models across scales." Soil Biology and Biochemistry 41.7 (2009): 1355-1379.

McKinney, Wes. "Data structures for statistical computing in python." Proceedings of the 9th Python in Science Conference. Vol. 445. No. 1. 2010.

Minasny, Budiman, et al. "Soil carbon 4 per mille." Geoderma 292 (2017): 59-86.

Moore, Frances C., and David B. Lobell. "The fingerprint of climate trends on European crop yields." Proceedings of the National Academy of sciences 112.9 (2015): 2670-2675.

Müller, Christoph, et al. "Global gridded crop model evaluation: benchmarking, skills, deficiencies and implications." Geoscientific Model Development 10.4 (2017): 1403-1422.

Nelson, Darrell W., and Lee E. Sommers. "Total carbon, organic carbon, and organic matter." Methods of soil analysis: Part 3 Chemical methods 5 (1996): 961-1010.

Oyebamiji, Oluwole K., et al. "Emulating global climate change impacts on crop yields." Statistical Modelling 15.6 (2015): 499-525.

Paustian, Keith, et al. "Quantifying carbon for agricultural soil management: from the current status toward a global soil information system." Carbon Management 10.6 (2019): 567-587.

Peel, Murray C., Brian L. Finlayson, and Thomas A. McMahon. "Updated world map of the Köppen-Geiger climate classification." Hydrology and earth system sciences 11.5 (2007): 1633-1644.

Pennock, D. J., and A. H. Frick. "The role of field studies in landscape-scale applications of process models: an example of soil redistribution and soil organic carbon modeling using CENTURY." Soil and Tillage Research 58.3-4 (2001): 183-191.

Pedregosa, Fabian, et al. "Scikit-learn: Machine learning in Python." the Journal of machine Learning research 12 (2011): 2825-2830.

Puntel, Laila A., et al. "Modeling long-term corn yield response to nitrogen rate and crop rotation." Frontiers in plant science 7 (2016): 1630.

Ray, Deepak K., et al. "Yield trends are insufficient to double global crop production by 2050." PloS one 8.6 (2013): e66428.

Reback, Jeff, et al. "pandas-dev/pandas: Pandas 1.0. 5." Zenodo (2020).

Ringeval, Bruno, et al. "Potential yield simulated by global gridded crop models: using a process-based emulator to explain their differences." Geoscientific Model Development 14.3 (2021): 1639-1656.

Rosenzweig, Cynthia, et al. "Assessing agricultural risks of climate change in the 21st century in a global gridded crop model intercomparison." Proceedings of the national academy of sciences 111.9 (2014): 3268-3273.

Rudel, Thomas K., et al. "Agricultural intensification and changes in cultivated areas, 1970–2005." Proceedings of the National Academy of Sciences 106.49 (2009): 20675-20680.

Rudin, Cynthia. "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead." Nature Machine Intelligence 1.5 (2019): 206-215.

Sanderman, Jonathan, Tomislav Hengl, and Gregory J. Fiske. "Soil carbon debt of 12,000 years of human land use." Proceedings of the National Academy of Sciences 114.36 (2017): 9575-9580.

Skalský, R., et al. "Geo-bene global database for bio-physical modeling v. 1.0. Concepts, methodologies and data." The GEO-BENE database report (2008): 58.

Smith, Pete. "How long before a change in soil organic carbon can be detected?." Global Change Biology 10.11 (2004): 1878-1883.

Smith, P., et al. "A comparison of the performance of nine soil organic matter models using datasets from seven long-term experiments." Geoderma 81.1-2 (1997): 153-225.

Smith, Pete, et al. "How to measure, report and verify soil carbon change to realize the potential of soil carbon sequestration for atmospheric greenhouse gas removal." Global Change Biology 26.1 (2020): 219-241.

Skendžić, Sandra, et al. "The impact of climate change on agricultural insect pests." Insects 12.5 (2021): 440.

Toudert, A., Braimoh, A., Bernoux, M. M. Y., St-Louis, M., Abdelmagied, M., Bockel, L., ... Zhao, Y. (2018). Carbon accounting tools for sustainable land management. Washington, DC: The World Bank, 135 pp. Retrieved from https://openknowledge.worldbank.org/bitstream/handle/10986/31062/132767-6-12-2018-14-1-54-SLMFullReportFINAL.pdf?sequence=1

Trnka, Miroslav, et al. "Adverse weather conditions for European wheat production will become more frequent with climate change." Nature Climate Change 4.7 (2014): 637-643.

UNCCD, 2017. Global Land Outlook, 1st ed. Bonn.

Vanguelova, E. I., et al. "Sources of errors and uncertainties in the assessment of forest soil carbon stocks at different scales—review and recommendations." Environmental monitoring and assessment 188.11 (2016): 1-24.

Walling, Eric, and Céline Vaneeckhaute. "Greenhouse gas emissions from inorganic and organic fertilizer production and use: A review of emission factors and their variability." Journal of Environmental Management 276 (2020): 111211.

Warszawski, Lila, et al. "The inter-sectoral impact model intercomparison project (ISI–MIP): project framework." Proceedings of the National Academy of Sciences 111.9 (2014): 3228-3232.

Whittaker, Carly, Marcelle C. McManus, and Pete Smith. "A comparison of carbon accounting tools for arable crops in the United Kingdom." Environmental Modelling & Software 46 (2013): 228-239.

White, Jeffrey W., et al. "Methodologies for simulating impacts of climate change on crop production." Field Crops Research 124.3 (2011): 357-368.

Williams, Jimmy R. "The EPIC model." Computer models of watershed hydrology. (1995): 909-1000.

Wösten, J. H. M., et al. "Development and use of a database of hydraulic properties of European soils." Geoderma 90.3-4 (1999): 169-185.