UTS UNIVERSITY OF TECHNOLOGY SYDNEY

Machine learning and uncertainty quantification in soil science and agricultural land economic

by SHUAI LI

Thesis submitted in fulfilment of the requirements for the degree of

Master of Engineering

under the supervision of Dr Xuzhen He and Dr Danial Jahed Armaghani

University of Technology Sydney
Faculty of School of Civil and Environmental Engineering

November 2025

Certificate of Original Authorship

I, Shuai Li, declare that this thesis is submitted in fulfilment of the requirements for the award of Master of Engineering, in the School of Civil and Environmental Engineering at the University of Technology Sydney.

This thesis is wholly my own work unless otherwise referenced or acknowledged. In addition, I certify that all information sources and literature used are indicated in the thesis.

This document has not been submitted for qualifications at any other academic institution.

This research was supported by an Australian Government Research Training Program (RTP) Scholarship doi.org/10.82133/C42F-K220.

Signature: Shuai Li

Production Note:
Signature removed prior to publication.

Date:07/November/2025

Publications

Journal papers related to the theme of this thesis:

- Li, S.; He, X. (2025). Feature-driven estimation of soil organic carbon with uncertainty quantification: Mitigating Overconfidence in Machine-Learning with Variational Bayesian Neural Networks. *Environmental Geotechnics*. Under review.

Table of Contents

Abbreviations

AWC	Available Water Capacity
BEA	Bureau of Economic Analysis
COD	Coefficient of Determination
CEC	Cation Exchange Capacity
CI	Credible Interval
CrI	Credible Interval
CME	Chicago Mercantile Exchange
DXY	USA Dollar Index
EC	Electricity Conductive
EIA	Energy Information Administration
ELBO	Evidence Lower Bound
GDP	Gross Domestic Product
GWR	Geographically Weighted Regression
IQR	Interquartile Range
NASA	National Aeronautics and Space Administration
NDVI	Normalized Difference Vegetation Index
NLL	Negative Log-Likelihood
NN	Neural Network
PCE	Personal Consumption Expenditures

RH2M	Mean Relative Humidity at 2 Meters Above the Surface
SE	Standard Error
SHAP	SHapley Additive exPlanations
SOC	Soil Organic Carbon
SSURGO	Soil Survey Geographic Database
SWCC	Soil Water Characteristic Curve
T2M	Mean Air Temperature at 2 Meters Above the Surface
UNSODA	Unsaturated Soil hydraulic Database
USDA	United States Department of Agriculture
VB	Variational Bayesian
VBNN	Bayesian Neural Network
VIF	Variance Inflation Factor
WoSIS	World Soil Information Service
WTI	West Texas Intermediate
XGBoost	Extreme Gradient Boosting

List of Symbols

$\boldsymbol{\sigma}$	Standard deviation
$\boldsymbol{\mu}$	Mean value
$\boldsymbol{\mathcal{N}}$	Normal distribution
$\boldsymbol{\epsilon}$	A small constant value
$\boldsymbol{\theta_s}$	Saturated water content
$\boldsymbol{\theta_r}$	Residual water content
$\boldsymbol{\epsilon}$	A small constant value (ϵ=1)
$\boldsymbol{\beta}$	Spatially Varying Coefficients
$\boldsymbol{u}$	Latitude
$\boldsymbol{v}$	Longitude
$\boldsymbol{\rho_k}$	Pearson Correlation Coefficient at Lag k
$\boldsymbol{W_{ij}}$	Spatial Weighting Matrix

List of Figures

List of Tables

Abstract

This thesis integrates three research components to address key challenges in agricultural prediction and land valuation through advanced machine learning and spatial modelling techniques. Despite the growing use of data-driven methods, few studies have systematically quantified uncertainty and spatial heterogeneity across soil, yield, and farmland valuation models. This research bridges that gap by developing probabilistic and spatially explicit frameworks.

The first component focuses on the prediction of Soil Organic Carbon (SOC), a critical indicator of soil health and productivity. Given the limitations of traditional measurement methods, five models were developed: linear regression, neural networks, XGBoost, variational Bayesian linear regression, and variational Bayesian neural networks, using accessible soil properties such as bulk density, pH, and texture. The models were evaluated under two scenarios, with and without nitrogen data, revealing that nitrogen is a critical predictor and that most models tend to be overconfident when key inputs are missing. This demonstrates the necessity of uncertainty-aware modelling in soil property prediction. Among the models, variational Bayesian neural networks performed best in balancing accuracy and uncertainty quantification.

The second component predicts maize yields across 842 counties in the U.S. Corn Belt from 2014 to 2023, incorporating soil, climate, remote sensing, and phenological variables. Four models were assessed: linear regression, geographically weighted regression (GWR), a variational Bayesian GWR (VB-GWR), and a neural-augmented VB-GWR (VBNN-GWR). VB-GWR outperformed others by accurately capturing spatially varying relationships and predicting yield ranges within 150–220 bu/acre. pH was identified as the most influential factor, while precipitation had limited impact. This part of the study enhances yield prediction and supports land valuation, especially for emerging agricultural areas. The VB-GWR model offers a novel integration of spatial regression with Bayesian inference, improving interpretability and uncertainty quantification.

The third component explores the spatial determinants of farmland rental prices using GWR and macroeconomic variables such as GDP, oil prices, and Treasury yields. The model effectively captures regional differences in rental price sensitivity, with oil prices

having a strong positive correlation in ethanol-producing states and maize yield showing weaker effects in core production areas. Uncertainty assessment based on standard error-derived confidence intervals reveals that spatial variation in rent uncertainty is minimal, suggesting that economic factors play a more dominant role. This provides new evidence linking macroeconomic volatility to regional rent disparities. The model is further used to estimate rental prices for 2026 and 2027, projecting an average annual growth rate of 10.40%, with a 68% confidence interval ranging from 10.41% to 12.39%.

Overall, this thesis demonstrates the value of integrating machine learning, spatial analysis, and probabilistic modelling to support decision-making in agriculture. By jointly modelling uncertainty and spatial heterogeneity across soil, yield, and land markets, the research provides a unified framework for agricultural prediction and valuation under complex environmental and economic conditions. The findings offer practical tools for predicting soil carbon, crop yield, and farmland rent under uncertain and spatially heterogeneous conditions, contributing to both land management and agricultural investment strategy.

1. Chapter 1: Introduction

This chapter introduces the background, motivation, and objectives of the study. It outlines the research problem and provides a summary of the study structure.

1.1 Introduction

Agricultural production systems are inherently complex, characterised by spatial heterogeneity, uncertain environmental conditions, and interconnected biophysical and economic drivers (Baldos et al., 2019). Predicting key variables such as soil organic carbon (SOC), crop yields, and farmland rental prices is essential for effective land management, food security, environmental protection, and investment planning (Lobell et al., 2009; Wang et al., 2019). Traditionally, such predictions have relied on deterministic models that offer single-point estimates, assuming that relationships between variables are consistent across time and space. However, these assumptions are increasingly untenable in the face of rapid environmental change, data limitations, and regionally varying agricultural practices, often resulting in overconfident model outputs. In this context, there is a growing need for models that not only predict outcomes accurately but also quantify the uncertainty associated with those predictions and recognise spatial variability.

Soil organic carbon is a critical component of soil health, influencing water retention, permeability, and structural stability—all of which are relevant to both agriculture and civil engineering (Rawls et al., 2003; Rumpel & Kögel-Knabner, 2010; Wu et al., 2003). While laboratory-based methods for measuring SOC are accurate, they are also costly and labour-intensive, making them unsuitable for large-scale implementation. As a result, indirect models using easily obtained soil properties such as bulk density, texture, and pH have gained popularity. Yet, most existing models fail to express the uncertainty inherent in both the input data and the modelling process, leading to overconfident predictions, which is particularly problematic when critical inputs such as nitrogen are missing or unreliable.

Similarly, maize yield prediction remains a key challenge in agricultural modelling due to its dependence on a diverse set of variables, including soil characteristics, climate conditions, and vegetation indices. These relationships are not only non-linear but also highly location-specific, as the same factor (e.g., temperature or pH) may affect yield differently in different parts of the U.S. Corn Belt. While statistical models and neural networks have improved yield forecasting, they often overlook spatial heterogeneity and fail to provide robust measures of predictive uncertainty. This shortcoming limits their

utility in land valuation and agricultural planning, where confidence in prediction intervals is as important as the prediction itself.

Beyond production, the economic dimension of agriculture, particularly farmland rental pricing, is shaped by both agronomic and macroeconomic factors. Rental prices are affected by yield potential, input costs, commodity prices (e.g., maize and oil), and broader indicators such as GDP or Treasury yields. These relationships are not constant across space; they vary significantly between core production zones and peripheral regions, as well as between states with different energy policies and industrial bases. While geographically weighted regression (GWR) has proven effective in modelling spatial variation, its deterministic structure still lacks the ability to represent uncertainty, which is essential for investors and policymakers engaged in risk-sensitive decision-making.

Taken together, these three challenges, including modelling SOC, maize yield, and farmland rental prices, highlight a common limitation in current agricultural modelling practices: difficult measurement and uncertainty quantification. This thesis addresses this gap by exploring the use of probabilistic machine learning, specifically variational Bayesian models, in conjunction with spatial modelling techniques such as GWR. By applying these methods across three case studies, this research seeks to build a unified framework that can adapt to regional variation while also quantifying predictive uncertainty, thereby enhancing the robustness and applicability of agricultural decision-making models.

1.2 Objectives

The primary aim of this thesis is to investigate how the integration of spatial modelling and uncertainty quantification can enhance predictive performance in agricultural systems and mitigate the problem of overconfidence in machine learning, through a unified analysis of: (1) to develop and evaluate probabilistic models for SOC estimation using indirect indicators under both complete and incomplete data scenarios; (2) to construct a hybrid GWR-Bayesian modelling framework for maize yield prediction that accounts for both non-linearity and spatial structure; and (3) to analyse the determinants of farmland rental prices using macroeconomic and agronomic data, and to assess how

uncertainty manifests across different regions. In doing so, this work makes methodological, applied, and theoretical contributions to the fields of environmental modelling, spatial data science, and agricultural economics.

1.3 Thesis Structure

The thesis is structured as follows. Chapter 2 presents a comprehensive review of relevant literature, covering uncertainty quantification in machine learning, spatial regression methods, and the applications in SOC, yield, and farmland valuation. Chapters 3 to 5 each present a case study: Chapter 3 focuses on SOC estimation using probabilistic models; Chapter 4 examines maize yield prediction using a spatial-Bayesian hybrid approach; and Chapter 5 explores the economic and spatial determinants of farmland rental prices. Chapter 6 synthesises the findings across these three applications, and concludes the thesis by summarising key contributions, identifying limitations, and outlining directions for future research.

2. Chapter 2: Literature Review

This chapter critically reviews the existing literature related to machining learning methods, soil organic carbon estimation, spatial and probabilistic models in maize yield prediction, spatial dynamics of farmland rental prices and economic indicators, and research gap.

2.1 Literature review

2.1.1 Machine Learning Methods overview

Environmental and agricultural systems are characterised by strong nonlinearity, spatial heterogeneity, and significant uncertainty (Snow et al., 2014). The inherent randomness in physical systems and lack of complete knowledge about a physical system also contribute to uncertainty in modelling these systems (Sadeghi Tabas & Samadi, 2022). Traditional linear models often struggle to capture the complex interactions among variables in such systems (Sahoo et al., 2017). The complexity of systems makes accurate physical process simulations challenging, leading to the appeal of data-driven and machine learning methods based on nonlinear interdependencies (Sahoo et al., 2017).

In recent years, machine learning methods have gained widespread adoption in tasks such as soil property estimation, crop yield prediction, and farmland rental analysis due to their flexibility, adaptability, and powerful function approximation capabilities (Garanayak et al., 2021; Xiao et al., 2022). From the reviewed literature, (Sadeghi Tabas & Samadi, 2022)machine learning approaches can be categorised into three main groups: deep learning models based on artificial neural networks (ANNs), ensemble learning methods such as random forests (RF) and gradient boosting, and probabilistic models rooted in Bayesian inference (Jin et al., 2022; Melendez-Pastor et al., 2023; Sahoo et al., 2017; Xiao et al., 2022).

To address the spatial non-stationarity of agricultural systems, Geographically Weighted Regression (GWR) has emerged as a valuable tool (McCord et al., 2012). GWR estimates local regression coefficients for each spatial unit by weighting nearby observations, thus allowing model parameters to vary over space (Sagan et al., 2021). This makes it particularly suited to applications where the relationship between predictors and response variables changes geographically. However, It is typically limited to linear formulations, making it inadequate for capturing complex nonlinear relationships.

Artificial Neural Networks (ANNs) are multi-layer function approximators inspired by the structure of biological neurons (Jin et al., 2022). They are particularly suitable for regression and classification tasks with complex, high-dimensional input features (Sahoo et al., 2017). ANNs have been widely applied in agricultural modelling, including for

estimating soil organic carbon (SOC), predicting crop yields, and modelling aspects of agricultural systems such as groundwater level changes which are influenced by environmental factors (Garanayak et al., 2021). The strength of ANNs lies in the capacity to model nonlinear relationships without requiring prior knowledge of the functional form between input and output variables (Kanwisher et al., 2023). However, ANNs also present several limitations. They typically involve many parameters, require significant computational resources for training, and lack interpretability (Végh, 2021). Most critically, they produce point estimates without any indication of uncertainty or confidence in the predictions, making them less suitable for risk-aware decision-making (Jin et al., 2022).

Ensemble learning methods, such as Random Forests (RF) and Extreme Gradient Boosting (XGBoost), combine the outputs of multiple weaker models to achieve stronger predictive performance (AlThuwaynee et al., 2021). RF constructs numerous decision trees trained on bootstrapped samples and random feature subsets, which helps reduce variance and improve robustness (Xu et al., 2024). In agricultural studies, RF has been effectively used for feature selection and prediction in tasks such as soil classification and SOC estimation (Lin & Liu, 2022). XGBoost, an efficient implementation of gradient boosting, enhances performance by iteratively minimising prediction errors through additive model updates and regularisation (Sarkhani Benemaran, 2023). These methods offer strong performance in medium-sized datasets and structured input settings. However, both RF and XGBoost are not naturally equipped to model spatial dependencies, which can be a limitation in applications involving spatially distributed data.

In contrast, Bayesian machine learning approaches emphasise probabilistic inference and the quantification of uncertainty. A study highlight that Bayesian methods provide a natural framework for uncertainty quantification by specifying prior distributions and updating them with observed data to obtain posterior distributions (Sadeghi Tabas & Samadi, 2022). Bayesian Linear Regression introduces prior distributions over parameters and updates these beliefs considering observed data, yielding posterior distributions that can inform prediction intervals (Fallah Mortezanejad & Mohammad-Djafari, 2024; Sadeghi Tabas & Samadi, 2022; Yang et al., 2017). Gaussian Processes, which model distributions over functions rather than parameters, offer flexible, non-

parametric representations of spatial and temporal relationships, though they are computationally intensive and less scalable for large datasets (Yang et al., 2017).

Among recent developments, Variational Bayesian Neural Networks (VBNNs) have emerged as a promising hybrid approach that integrates the expressive power of deep learning with the rigour of Bayesian inference (Le et al., 2020). VBNNs treat network weights as random variables with learnable distributions and apply variational inference techniques to approximate the posterior (Chen et al., 2022). This is a core characteristic of VBNNs, where instead of fixed weights, probability distributions over the weights are learned using variational inference. The goal of variational inference is to find a tractable distribution that closely approximates the intractable posterior distribution over the parameters of network. As a result, VBNNs produce probabilistic predictions, capturing both model and input uncertainty while retaining the ability to learn complex, nonlinear relationships (Chen et al., 2022). the ability of VBNNs to handle uncertainty and noise, which would be beneficial in situations with noisy or incomplete data. For instance, one source discusses enhancing the anti-noise ability of a network using variational inference (Jin et al., 2022). However, VBNNs come with increased computational cost, require careful tuning of hyperparameters, and demand a strong understanding of Bayesian inference, which may limit their widespread adoption in applied agricultural settings (Chen et al., 2022; Sadeghi Tabas & Samadi, 2022).

In summary, ANN, RF, and XGBoost offer strong predictive capabilities and have been widely used in agricultural modelling. However, their inability to quantify uncertainty or incorporate spatial structure limits their application in contexts requiring reliability and interpretability. Bayesian models, particularly VBNNs, offer a compelling alternative by enabling the generation of confidence intervals and risk-aware outputs. The choice of method should be guided by the specific characteristics of the data and the modelling objectives. This thesis evaluates and compares these methods across multiple use cases, aiming to construct a unified modelling framework that is not only accurate but also capable of representing uncertainty and spatial variation in agricultural systems.

2.1.2 Soil Organic Carbon Estimation Using Probabilistic Machine Learning

Soil organic carbon (SOC) plays a central role in agricultural productivity, soil health, and global carbon cycling (Emadi et al., 2020). Accurate estimation of SOC is therefore critical for both scientific and policy objectives, particularly in the context of carbon sequestration, sustainable land management, and climate mitigation (Minasny et al., 2017). However, SOC is notoriously difficult and expensive to measure directly, often requiring intensive laboratory analysis of soil samples. As a result, indirect estimation methods using machine learning have gained traction, wherein more readily available soil properties (e.g., texture, bulk density, pH) are used to infer SOC content (Chen et al., 2024; Makovníková et al., 2017; Xu et al., 2015).

Due to the cost and complexity of direct SOC measurements, numerous studies have explored the use of surrogate variables to predict SOC content (Makovníková et al., 2017; Xu et al., 2015). Commonly used predictors include physical and chemical soil properties such as bulk density, sand, silt, clay fractions, soil pH, cation exchange capacity (CEC), and electrical conductivity (Bell & van Keulen, 1995; Chen et al., 2024; Hollis et al., 2011). These properties are more easily and cheaply acquired, often through remote sensing or field surveys, and have been shown to correlate with SOC under various climatic and land-use conditions (Richer-de-Forges et al., 2023).

Indirect modelling approaches, therefore, aim to learn a mapping between these observable properties and SOC values (Emadi et al., 2020). Models such as random forests (RF), support vector machines (SVM), and artificial neural networks (ANN) have been widely employed for this task, achieving reasonably high Coefficient of Determination (COD) in many studies (Emadi et al., 2020; Liu et al., 2023). For example, the study evaluated four machine-learning techniques (multiple linear regression, artificial neural networks, support vector machine, and random forest) for SOC quantification at a high spatial resolution (1 m). The random forest model performed best, with the high-resolution land cover classification being the most relevant predictive variable (Siewert, 2018). The study estimated the landscape mean SOC storage and highlighted the importance of wetland areas, particularly peat plateaus, for SOC storage. It also investigated the effect of reduced spatial resolutions on SOC estimation, noting a

significant loss of detail and underestimation at coarser resolutions due to the omission of small-scale wetland hotspots. Another study used data from the second national soil survey in China to evaluate the impact of eight selected PTFs for bulk density on the estimation of SOC storage at regional scale (Xu et al., 2015). The results showed that different PTFs can lead to higher uncertainty in SOC storage estimation, with the coefficient of variation varying significantly across the PTFs. The study found that PTFs had significant effects on SOC density in different soil types and that these effects could vary with soil depth.

While traditional machine learning models have shown promising results in SOC prediction, their deterministic nature limits their ability to account for the inherent uncertainty in soil data and model generalisability, their outputs remain limited to point predictions, failing to reflect uncertainty in both the input data and the learned model parameters (Emadi et al., 2020). This is a significant limitation in the context of spatial soil mapping, where data sparsity, heterogeneity, and measurement noise are commonplace.

For that limitation, Probabilistic approaches, especially those based on variational Bayesian methods, offer a powerful alternative by incorporating uncertainty directly into the learning process through the derivation of probability distributions for parameters and predictions (Contreras et al., 2018).

2.1.3 Integrating Spatial and Probabilistic Models in Maize Yield Prediction

Maize is one of the most important cereal crops globally, and accurate yield prediction plays a critical role in ensuring food security, guiding agricultural management, and informing policymaking (Gaffney et al., 2015; Kusmec & Schnable, 2024; Sarzaeim & Muñoz-Arriola, 2024). While traditional models, both statistical and machine learning based, have achieved notable success in yield estimation, their performance is often limited in the face of spatial heterogeneity and uncertainty in agricultural data (Rizzo et al., 2018). Recent advances have introduced hybrid approaches that integrate spatial modelling techniques, such as Geographically Weighted Regression (GWR), with probabilistic frameworks like Variational Bayesian Neural Networks (VBNNs), offering

a more robust and interpretable modelling pipeline (Deines et al., 2021; Leng, 2021; Peng et al., 2018).

2.1.3.1 Drivers of Maize Yield: From Deterministic to Probabilistic Modelling

Maize yield is influenced by a complex interplay of climatic, soil, management, and socio-economic factors (Deines et al., 2021; Leng, 2019; Massigoge et al., 2023; Mourtzinis et al., 2016). These include temperature, precipitation, solar radiation, soil fertility, planting and harvesting dates, and input levels. Importantly, these drivers often exhibit strong spatial variability, and their impact on yield is seldom uniform across different geographic locations (Deines et al., 2021; Morell et al., 2016).

Conventional yield prediction models which include linear regression, random forests, or artificial neural networks, typically assume global parameter constancy, treating all regions under a single model structure (Morell et al., 2016). This assumption of stationarity limits the capacity of models to adapt to spatial variation, leading to significant biases in regions where the local relationship between inputs and yield deviates from the global trend (Qin et al., 2021). Moreover, traditional models often do not provide any measure of prediction uncertainty, which is especially problematic in real-world applications where data may be noisy, incomplete, or derived from indirect sources (Morell et al., 2016).

One study explored the benefits of using seasonal climate predictions and satellite data for forecasting US maize yield. While not strictly a purely probabilistic model, the use of climate predictions inherently involves dealing with uncertainty in future weather conditions, which would influence the uncertainty in yield forecasts (Joshi et al., 2021). Another study developed a framework to aid risk assessment of droughts on crop yields in the US, defining risk by the probability of a drought event and the probability of yield loss under a given drought intensity (Leng, 2021). The research found that the risk of maize yield loss increased with drought intensity, with specific probabilities of yield reduction under moderate, severe, extreme, and exceptional drought events.

2.1.3.2 Geographically Weighted Regression for Spatially Varying Relationships

In the context of maize yield modelling, GWR can identify region-specific influences of key variables—such as the importance of temperature in one zone versus soil organic

carbon in another (Zhao et al., 2024). By capturing such spatial dynamics, GWR provides deeper insights into the agronomic and environmental drivers of yield variation (Massigoge et al., 2023). However, despite its strengths in representing spatial variation, GWR remains a deterministic approach and does not quantify uncertainty in its predictions. Some study has been developed use this method, for example, GWR is used as a local variation modelling technique to analyse the determinants of house prices, contrasting it with the traditional global perspective offered by Multiple Regression Analysis (MRA) (McCord et al., 2012). However, the case does not explicitly detail the consideration of uncertainty.

2.1.4 Spatial Dynamics of Farmland Rental Prices and Economic Indicators

Understanding the spatial dynamics of farmland rental prices is critical for land valuation, investment decision-making, and agricultural policy development. In regions such as the U.S. Corn Belt, farmland rental values are shaped by a combination of biophysical and socio-economic factors, including soil productivity, crop yields, commodity prices, climate variability, and regional economic conditions (Uludere Aragon, 2019). However, the relationships between these factors and land value are rarely homogeneous across space (Hendricks et al., 2014). As such, spatial modelling approaches that account for geographic heterogeneity are increasingly employed to uncover the nuanced drivers of farmland rental price variation (McNunn et al., 2020).

2.1.4.1 Agricultural and Economic Drivers of Farmland Rental Prices

Farmland rental prices are influenced by both supply-side and demand-side factors (Uludere Aragon, 2019). On the supply side, land quality attributes such as soil organic carbon, nutrient content, bulk density, and water retention capacity play a fundamental role in determining the productive potential of a parcel (Wang et al., 2021). On the demand side, crop profitability, proximity to markets, population density, and macroeconomic variables such as interest rates and inflation also contribute to land value determination (Moss, 1997).

Importantly, the interactions between these factors can vary significantly between regions (Hendricks et al., 2014). in high-yielding zones, rental prices may closely track

commodity price fluctuations, whereas in marginal areas, climatic risk or access to irrigation infrastructure may be more dominant For example, a study on farmland valuation highlighted that while inflation generally explains much of the variation in farmland values, the Northeast and Southern Plains regions showed significantly lower sensitivity to inflation compared to the national average (Moss, 1997). In the Northeast, the cost of debt capital was a more significant explanatory factor, while in the Southern Plains, the return on agricultural assets played a more crucial role (Moss, 1997). This spatial variability challenges the assumptions of global models and necessitates the use of location-sensitive analytical techniques (Hendricks et al., 2014). Similarly, the optimal maize planting date in Iowa was found to be highly dependent on geographic location and fluctuate with weather patterns, underscoring the need for location-specific agronomic recommendations rather than static ones (Baum et al., 2020).

2.1.4.2 Application of GWR in Land Economics

Geographically Weighted Regression (GWR) has been increasingly adopted in land economics to capture spatially varying relationships between land value and its determinants (Neelawala et al., 2012). Unlike traditional regression models, which assume constant parameters across the entire study area, GWR allows coefficients to vary geographically by calibrating local models at each spatial location (Ribeiro & Pereira, 2018). For instance, one study estimated Ordinary Least Squares (OLS) and Geographically Weighted Regression (GWR) models to analyse the determinants of property values. GWR is described as an auxiliary regression that checks for the robustness of spatial models and weights spatially distributed observations across Cartesian coordinate points (Ribeiro & Pereira, 2018). This approach acknowledges that the influence of factors on property prices can differ across locations (Neelawala et al., 2012).

Studies applying GWR to farmland rental pricing have revealed significant spatial heterogeneity in the importance of biophysical and economic variables (Neelawala et al., 2012). For example, variables such as corn yield, precipitation, and soil texture may explain a large proportion of land value variation in one part of the Corn Belt, while in other areas, socio-economic factors such as rural income or population density may be more influential (Hendricks et al., 2014). For example, the study on land quality and corn

cultivation in the Western Corn Belt demonstrates that the responsiveness of corn acreage to price changes varies significantly based on land quality, indicating spatial differences in the influence of economic variables (Uludere Aragon, 2019).

By revealing these spatial patterns, GWR offers valuable insights for land managers, investors, and policymakers seeking to identify undervalued regions, assess rental price risk, or understand regional competitiveness. However, like other deterministic models, GWR does not incorporate uncertainty in its outputs, which limits its utility in risk-sensitive applications such as land investment under climate change or carbon offset initiatives (Basnet et al., 2021).

2.1.4.3 Integrating Agricultural and Economic Indicators with Uncertainty Modelling

Recent developments in spatial modelling have begun to explore the integration of economic variables with uncertainty-aware machine learning techniques (Zhang et al., 2019). For example, models that combine climate data, soil health indicators, and commodity prices with probabilistic frameworks such as VBNNs or Bayesian geo-statistics have demonstrated the ability to provide both accurate predictions and credible uncertainty bounds for land valuation (Basnet et al., 2021; Kingwell & Xayavong, 2016).

This integration is particularly relevant in the context of carbon markets and ecosystem service payments, where the valuation of land is not solely based on its agricultural output, but also on its environmental attributes (e.g., SOC sequestration potential) and associated financial incentives (Ssegane et al., 2016). The analysis of carbon farming in China demonstrates that providing financial incentives (in the form of a carbon tax) can induce changes in land use and farm practices to reduce on-farm greenhouse gas emissions (Tang et al., 2019). This indicates a direct link between financial incentives tied to environmental outcomes and land management decisions, which would necessitate valuation models that account for these factors. The marginal abatement costs of GHG emissions in agriculture are also being studied, highlighting the economic aspect of environmental performance(Tang et al., 2016). As the agricultural sector increasingly interfaces with financial instruments, the need for transparent, data-driven, and uncertainty-aware land valuation models becomes more urgent.

While the literature to date has largely examined land value from either a biophysical or economic lens, future work could benefit from a unified framework that jointly models soil, yield, and rental dynamics under uncertainty. Such an approach could provide a more holistic understanding of land productivity and value, particularly in regions experiencing rapid climatic or market shifts.

2.2. Research Gaps and Positioning

The previous sections have reviewed a wide range of studies covering uncertainty quantification, soil organic carbon (SOC) estimation, maize yield prediction, and spatial modelling of farmland rental prices. These studies demonstrate the growing use of machine learning in agricultural and environmental systems. However, despite notable progress in improving predictive accuracy, two critical limitations persist in current research: a lack of explicit uncertainty quantification in most models, and insufficient treatment of spatial heterogeneity in modelling frameworks. These issues directly limit the generalisability, interpretability, and real-world applicability of existing models, especially in regions characterised by data scarcity, environmental variability, or missing inputs.

In terms of uncertainty-aware modelling, Bayesian machine learning methods, such as variational Bayesian neural networks (VBNNs), have recently emerged as promising tools. They offer principled ways to quantify predictive confidence and capture model uncertainty. Yet, applications of these methods in core agricultural problems like SOC prediction, yield forecasting, and land valuation remain limited. Even when used, probabilistic models are often applied in isolated scenarios and are rarely evaluated across different levels of data completeness or compared systematically against deterministic methods. The broader potential of Bayesian models to improve decision-making under uncertainty has not been fully explored.

Meanwhile, spatial modelling tools such as Geographically Weighted Regression (GWR) have shown strong capabilities in addressing spatial non-stationarity. These models adapt parameter estimates to local contexts, offering enhanced interpretability across geographic regions. However, GWR remains fundamentally a deterministic and linear approach. It does not model nonlinear relationships, nor does it produce prediction

intervals, making it less suitable for contexts where both spatial variation and uncertainty must be addressed simultaneously.

Another key gap in the literature lies in the separation of spatial and probabilistic modelling. While both dimensions are critical in agricultural contexts, they are often studied independently. Very few studies have developed unified modelling frameworks that jointly capture spatial heterogeneity and predictive uncertainty. Additionally, most research tends to focus on a single target, such as yield or SOC, without testing whether the proposed approach generalises across different types of agricultural outputs. This limits the development of versatile tools that can support integrated land management, policy analysis, or investment planning.

To address these gaps, this thesis proposes a unified modelling framework that integrates spatial sensitivity and uncertainty quantification. The framework is applied to three key agricultural use cases: (1) indirect estimation of SOC under complete and missing input scenarios; (2) maize yield prediction across a geographically diverse region using a spatial-Bayesian hybrid approach; and (3) spatial modelling of farmland rental prices with integrated economic and agronomic indicators. This study contributes not only methodologically—by systematically combining VBNNs and GWR—but also empirically, by applying this framework across distinct yet related agricultural domains. The models developed in this thesis produce probabilistic outputs, identify spatial sensitivity patterns, and provide actionable insights for land valuation, risk analysis, and agricultural decision-making.

By addressing both spatial heterogeneity and uncertainty within a single framework, this research bridges a significant gap between data-driven modelling and the real-world complexity of agricultural systems. It advances the state of the art in environmental and agricultural machine learning, while offering practical tools for more informed, transparent, and robust decision-making under uncertainty.

3. Chapter 3: Indirect models for soil organic carbon with uncertainty quantification: Mitigating overconfidence in machine-learning with variational Bayesian neural networks

Abstract

Building upon the insights identified in the previous chapter, which highlighted the limitations of traditional Soil Organic Carbon (SOC) estimation approaches and the emerging potential of machine-learning techniques, this chapter presents the development of indirect models for Soil Organic Carbon (SOC) prediction with explicit uncertainty quantification. While the literature review underscored the growing need to move beyond deterministic models that ignore uncertainty, the current chapter translates these conceptual gaps into a practical modelling framework. Soil Organic Carbon (SOC) plays a crucial role in soil health, influencing properties such as water retention, permeability, and structural stability, which are critical in civil engineering and agriculture applications. Traditional methods for measuring SOC, while accurate, are resource-intensive and unsuitable for large-scale implementation, and thus indirect models provide an efficient alternative by utilising readily available soil properties such as bulk density, soil texture, and pH, etc. This study builds and examines three machine-learning models (linear regression, neural networks, XGBoost) and two models directly providing uncertainty quantifications (variational Bayesian linear regression and variational Bayesian neural networks). We examine two application scenarios: Scenario #1 with nitrogen as an input variable and Scenario #2 with nitrogen missing. The results indicate a significant decrease in predictive accuracy across all models when nitrogen is excluded, underscoring its strong correlation with SOC and its critical importance in enhancing performance. In scenarios where limited information is available—such as when nitrogen data is missing—all machine-learning models tend to exhibit overconfidence, with predicted variability being smaller than the actual variability. In such cases, only variational Bayesian neural networks, which account for both non-linearity and uncertainty, demonstrate satisfactory performance.

Keywords: Machine-learning; Nitrogen; Soil data base; Soil organic carbon; Variational Bayesian; WoSIS 2019.

3.1 Introduction

Soil Organic Carbon (SOC) is a critical component of the soil ecosystem, influencing numerous physical and chemical properties such as water retention capacity, permeability, and structural stability (Rawls et al., 2003; Rumpel & Kögel-Knabner, 2010; Wu et al., 2003). These properties are fundamental to civil engineering applications like infrastructure design, drainage system planning, and overall structural integrity. In particular, SOC plays a vital role in shaping the Soil Water Characteristic Curve (SWCC), affecting parameters like θ_s (saturated water content) and θ_r (residual water content), which are essential for predicting soil moisture dynamics (Bauer & Black, 1981; Bell & van Keulen, 1995; Lal, 2020; Liu et al., 2023; Miller & Naeth, 2019; Rawls et al., 2003; Wang et al., 2020; Xu et al., 2016). Given its widespread impact, SOC has significant implications not only for soil science but also for climate change mitigation and sustainable land management.

SOC is primarily formed through the decomposition of organic matter from plant and animal residues that are incorporated into the soil. When plants and organisms die, the remains are broken down by soil microorganisms in a complex process that converts organic materials into humus, a stable form of organic matter rich in carbon (Schmidt et al., 2011). This process is influenced by various factors such as climate, vegetation type, soil texture, and land management practices (Garg et al., 2021). For instance, cooler temperatures and higher moisture levels generally slow down decomposition, leading to greater SOC accumulation. Conversely, warmer temperatures and certain agricultural practices can accelerate decomposition, reducing SOC levels (Davidson & Janssens, 2006).

As climate change and carbon emissions become more pressing global issues, SOC, as a major carbon storage medium, is closely linked to the global carbon cycle and climate change (Longbottom et al., 2014; Poeplau & Dechow, 2023). The accumulation and decomposition of SOC directly affect atmospheric CO_2 concentrations, creating feedback loops that influence climate patterns. Accurate SOC predictions not only optimise soil management practices but also contribute to global carbon neutrality efforts (Ju et al., 2014). In geotechnical applications, SOC dynamics interact with soil-atmosphere processes, where soil water retention and evaporation influence both soil stability and

carbon sequestration potential. These interactions are crucial for engineered structures, as they impact soil strength, permeability, and overall environmental sustainability in construction projects (Cui, 2022). Furthermore, SOC prediction can provide critical data for assessing the carbon storage capacity of soils, thereby helping policymakers develop effective carbon emission and compensation policies (Cushing et al., 2018). Moreover, SOC research is not isolated; it intersects with multiple disciplines, including hydrology, ecology, climatology, and civil engineering (Jiang et al., 2022; Minasny et al., 2017; Smith et al., 2013). Machine-learning (ML) methods can integrate data from these fields and provide a more comprehensive SOC prediction framework through cross-disciplinary analysis (Longbottom et al., 2014; Shogren et al., 2024).

While SOC measurement in controlled laboratory settings is relatively straightforward, traditional methods such as wet chemical analysis, elemental analysers, and the loss on ignition (LOI) method are time-consuming, resource-intensive, and costly (Pribyl, 2010). These methods become particularly impractical for large-scale applications. Additionally, SOC data is often missing in major soil databases such as Unsaturated Soil Hydraulic Database (UNSODA) and World Soil Information Service (WOSIS), limiting the predictive power and interpretability of SWCC models (Es-haghi et al., 2023; Pham et al., 2023). Thus, while SOC may not be difficult to measure on a small scale, large-scale and efficient estimation remains a challenge, especially in resource-constrained environments.

Rapid environmental changes, such as extreme weather events and exacerbated climate change, can cause soil properties and SOC levels to fluctuate significantly within short periods (Smith et al., 2008). The increasing demand for more efficient, scalable, and cost-effective SOC prediction methods has driven researchers to explore alternative approaches, particularly machine-learning. ML has demonstrated great potential in predicting SOC using other readily available soil properties such as bulk density, soil texture, and pH (Emadi et al., 2020).

In recent years, many Pedotransfer Functions (PTFs) for SOC estimation have been widely used (Emadi et al., 2020; Siewert, 2018; Song et al., 2005; Wu et al., 2003). ML offers a more efficient alternative, using accessible soil properties such as bulk density, soil texture, and pH to predict SOC, thereby greatly reducing data acquisition time and costs. Various ML techniques, including Random Forest, Support Vector Machines,

XGBoost, and deep neural networks, have demonstrated high predictive accuracy for SOC (Song et al., 2005; Wu et al., 2003). Ensemble methods like Random Forest, for instance, have shown excellent performance, particularly well-suited to capturing the non-linear relationships in SOC data (Hengl et al., 2017). Gradient boosting models, such as XGBoost, further improve accuracy, especially in datasets with missing values(Padarian et al., 2020). Additionally, some studies have integrated ML models with geographical information systems data, meteorological data, and other environmental factors, which enhances the spatial accuracy of SOC predictions (Padarian et al., 2020).

Measurement errors, soil heterogeneity, experimental design, sampling errors, and environmental variations dynamically influence soil properties, leading to the inherent uncertainty in predictions, further complicating SOC prediction (Conant et al., 2010). Many existing studies have overlooked the uncertainty in SOC predictions, limiting the reliability of these models for practical applications (Taalab et al., 2015; Xu et al., 2015). In the past few years, studies have highlighted the importance of uncertainty in SOC predictions (Li et al., 2024). However, significant limitations remain in quantifying this uncertainty. Moreover, current research has yet to adequately account for these uncertainties, restricting the reliability and accuracy of SOC models in practical applications.

This study addresses SOC data scarcity by estimating SOC values using probabilistic models based on variational Bayesian theory and compare it against traditional methods like linear regression, NN and XGBoost. The structure of this paper is organised as follows: Section 1 introduces the research background and significance of SOC. Section 2 outlines the mathematical framework and models employed in the study, detailing their theoretical basis and methodological design. Section 3 describes the datasets utilised, including the sources, preprocessing steps, and the modelling design, highlighting the scenarios considered in the experiments. Section 4 presents the results, comparing the outcomes of two scenarios and quantifying uncertainties to provide a comprehensive evaluation. Section 5 discusses the implications of the findings and potential limitations. Finally, Section 6 concludes the paper by summarising key insights.

3.2 Mathematical Framework

This study utilises linear regression, XGBoost, and NN to predict SOC. Additionally, VB linear and VBNN were employed to quantify the uncertainty of predictions, enabling a comprehensive comparison between deterministic and probabilistic approaches in SOC modelling.

3.2.1 Outlier Detection, Removal and Data transformation

To ensure the quality and reliability of the dataset, the interquartile range (IQR) method was employed for outlier detection and removal. The IQR method is a widely used statistical approach that identifies outliers based on the spread of the central 50% of the data (Xie et al., 2022). Specifically, the IQR is calculated as the difference between the third quartile (*Q3*) and the first quartile (*Q1*):

$$IQR = Q3 - Q1 \tag{3.1}$$

Data points falling below the lower bound ($Q1 - 1.5IQR$) or exceeding the upper bound ($Q3 + 1.5IQR$) are considered outliers. These thresholds demarcate observations that deviate substantially from the interquartile range, which represents the central and most representative portion of the dataset.

Standardisation: Standardisation is a widely used data preprocessing technique in machine-learning and statistical modelling. It involves rescaling the features of a dataset. The primary objective of standardisation is to eliminate the influence of varying feature scales, which can otherwise bias the training process. For instance, features with larger magnitudes can dominate the optimisation of model, reducing its ability to appropriately learn from smaller-scale features. Mathematically, it is expressed as:

$$x' = \frac{x - \mu}{\sigma} \tag{3.2}$$

where x represents the original feature value, μ is the mean of the feature, σ is the standard deviation, and x' is the standardised value.

Logarithmic Transformation: Variables with skewed distributions, such as soil depth, SOC, and other soil properties, were transformed using a natural logarithm with a small

constant (ϵ=1) added to avoid issues with values less than 1 and to prevent excessive scale dispersion.

$$x' = log(x + \epsilon) \tag{3.3}$$

Both transformations aimed to improve model performance by normalising the feature distributions and reducing skewness, albeit in different ways to cater to the specific needs of each algorithm. Extensive testing demonstrated the transformation methods for each parameter of the models.

3.2.2 Machine-learning models

Linear regression, as one of the most fundamental models in machine-learning, assumes a linear relationship between input features and the target variable. Compared to more complex machine-learning models, its primary advantage lies in its simplicity and resistance to overfitting, making it a robust choice for well-behaved datasets. The predicted output is given by:

$$\mathrm{y} = \mathrm{Wx} + \mathrm{b} \tag{3.4}$$

where W represents the weights and b the bias. The objective of linear regression is to find the line that minimises the error between predictions and actual target values.

Building upon the simplicity of linear models, which assume a linear relationship between input features and the target variable, neural networks (NN) excel in capturing complex, nonlinear relationships, making results well-suited for estimating intricate patterns in data. A NN consists of interconnected layers of nodes (neurons), each performing weighted sums of inputs followed by a nonlinear activation function (Chen et al., 2024). This structure enables NN to model the complexities that linear models cannot. The forward pass through a single layer is expressed as:

$$y = F(Wx + b) \tag{3.5}$$

where W is the weight matrix, x the input vector, b the bias term, and F an activation function (e.g., ReLU or sigmoid). Training involves optimising these weights and biases to minimise the loss function.

XGBoost is another powerful model capable of capturing nonlinear relationships in data. As an ensemble learning method, it builds a series of decision trees to improve prediction accuracy (Emadi et al., 2020). Each tree attempts to correct the errors made by the previous trees by learning from the residuals. The model optimises the objective function:

$$Obj(\theta) = L(\theta) + \Omega(\theta) \tag{3.6}$$

where $L(\theta)$ is the loss function (e.g., MSE) and $\Omega(\theta)$ is a regularisation term to prevent overfitting. XGBoost effectively models non-linear relationships and can be tuned for high accuracy with low computational cost.

In this study, a variational Bayesian linear model (VB linear) is employed to estimate posterior distributions of model parameters. The VB linear model incorporates uncertainty quantification by learning the posterior distributions of weights and biases using the VB approach. This method has been widely validated and applied in numerous studies, demonstrating its effectiveness in capturing parameter uncertainty and improving predictive robustness across various domains, including geotechnical engineering, environmental modelling (Contreras & Brown, 2019; Contreras et al., 2018). The implementation and methodological steps are detailed as follows:

The spike-and-slab prior is a widely used probabilistic approach for inducing sparsity in parameter spaces (He et al., 2025). It excels at identifying and excluding irrelevant parameters by driving their values toward zero (via the spike component). By penalising unnecessary parameters, this prior reduces the risk of overfitting, especially in cases where the number of features exceeds the number of observations. This prior is modelled as a mixture of two independent Gaussian distributions, which balance between enforcing sparsity and allowing flexibility:

$$\theta_{\text{prior}} \sim \pi \cdot \mathcal{N}(\theta \mid 0, \sigma_{\text{spike}}{}^2) + (1 - \pi) \cdot \mathcal{N}(\theta \mid 0, \sigma_{\text{slab}}{}^2) \tag{3.7}$$

Where θ_{prior} is the spike and slab prior, representing the probability density function for the parameter θ. π is the weight for 'spike' component and $(1 - \pi)$ is the weight of 'slab'. σ represents the variance of each component. The distribution of spike and slab prior is shown as Figure 1.

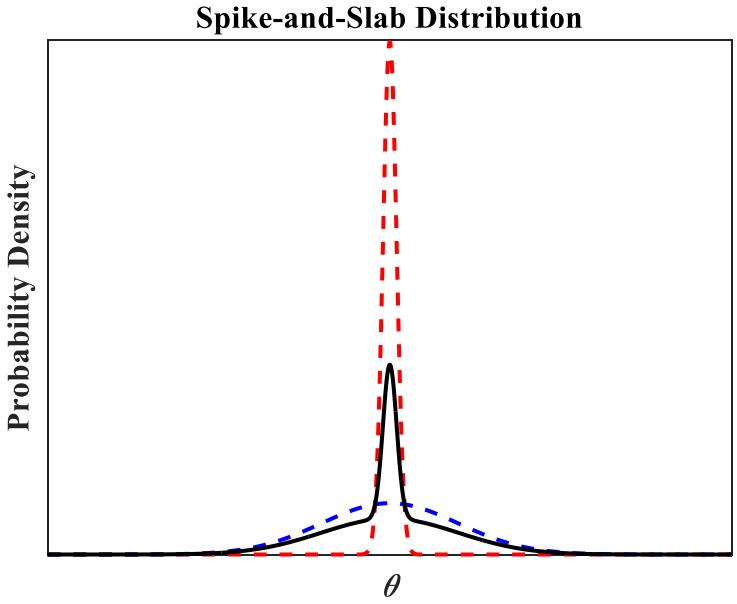


Figure 1 Spike and slab distribution. (The red dashed line represents the spike component, the blue dashed line represents the slab component, and the black solid line represents their weighted mixture. The horizontal axis denotes the variable θ , and the vertical axis denotes the probability density).

The posterior distributions of weights and biases were modelled using the mean-field variational family. Specifically, the posterior mean field function was defined to parameterise the posterior distributions with independent Gaussian distributions, where the mean and standard deviation were learned during training. A Softplus transformation ensured the standard deviation remained positive, while a small additive term stabilised the learning process. The posterior distribution was implemented as:

$$\boldsymbol{\theta}_{\mathbf{posterior}} \sim \boldsymbol{\mathcal{N}}(\boldsymbol{\mu}, \boldsymbol{\sigma}) \tag{3.8}$$

$$\sigma = softpuls(c + \sigma_{raw}) \tag{3.9}$$

Where c ensures numerical stability in the scale transformation.

The VB linear model was constructed as follows:

1. Input Layer: The input shape matched the feature dimension of the training data.
2. Dense Variational Layer: A variational dense layer was used to learn the posterior distributions of weights and biases. The Kullback-Leibler (KL) divergence was

scaled by a weight w_{KL} equal to the inverse of the training sample size to ensure proper regularisation.

3. Output Distribution: The final layer used a distribution to parameterise the output as a Gaussian distribution. The mean and scale of the Gaussian were parameterised using separate learned components, with the scale constrained to be positive using a Softplus transformation.

The variational Bayesian neural network (VBNN) described in this section implements Bayesian principles to capture uncertainty in both model parameters and predictions. The architecture incorporates probabilistic reasoning into a neural network framework. The key components of the model are outlined below:

For prior of VBNN, there is a challenge encountered when using a spike-and-slab prior in VBNN implementation by extensive testing. Specifically, the spike-and-slab prior leads to results oscillating between two extremes: overconfidence in predictions and low accuracy. These issues arise due to the complexity of tuning the spike-and-slab prior in high-dimensional NN architectures, where the mixture nature of the distribution complicates optimisation. Therefore, a standard normal distribution was chosen as the prior for the VBNN:

$$\theta_{\text{prior}} \sim \mathcal{N}(\mu, \sigma^2) \tag{3.10}$$

The posterior distribution of the network parameters is defined using the same method with VB linear, assuming that each parameter is independent and follows a Gaussian distribution.

The VBNN architecture incorporates multiple dense variational layers to parameterise the posterior distributions of weights and biases, as well as nonlinear activation functions to model complex relationships in the data. The key components of the model are:

1. Input Layer: Processes input features with dimensionality matching the dataset.
2. Hidden Layer: A Dense Variational layer with units, using the defined posterior and prior distributions. This layer applies Bayesian reasoning to the first level of feature transformation.
3. Nonlinearity: An activation function introduces nonlinearity, enabling the network to learn hierarchical feature representations.

4. Output Layer: A second Dense Variational layer with two outputs parameterises the predictive distribution (mean and variance).
5. Distribution Output: The final output is modelled as a Gaussian distribution, where the mean and variance are learned from the output layer parameters.

3.2.3 Evaluation Method

The Mean Squared Error (MSE) is a common loss function used to measure the difference between predicted and actual target values. It is used to optimise models such as NN, linear regression, and XGBoost, aiming to reduce the average squared error between the predictions and true values. The MSE is defined as:

$$\text{MSE} = \frac{1}{N}\sum(y_i - \hat{y}_i)^2 \tag{3.11}$$

where N is the number of data points, y_i the actual target value, and $\hat{y}_i$ the predicted value. Minimising MSE corresponds to reducing the average squared error between the predictions and observed values of model.

The coefficient of determination (COD) is a critical metric for evaluating the performance of machine-learning models, as it quantifies the overlap between predicted and actual values. COD is calculated based on the proportion of variance explained by the model, providing insights into its ability to capture the variability of the target variable. The formula is defined as:

$$COD = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \overline{y})^2} \tag{3.12}$$

Where $\overline{y}$ is the mean of the actual values. A *COD* value close to 1 indicates that the model performs well in fitting the data, whereas values near 0 or negative suggest weak predictive ability, potentially worse than a baseline mean prediction.

The negative Log-Likelihood (NLL) is employed as the primary loss function to optimise the probabilistic predictions of the model. This loss function is particularly suitable for models that output probability distributions, allowing both point predictions and uncertainty quantification. The NLL measures how well the predicted distribution explains the observed data by penalising unlikely outcomes.

The NLL is defined as the negative logarithm of the probability of the observed data under the predicted distribution. For a predicted output $\hat{y}$ with corresponding standard deviation σ, the NLL is defined as:

$$NLL = -\log p\,(y \mid \hat{y}, \sigma) = 0.5 \times \left[\frac{(y-\hat{y})^2}{\sigma^2} + log(2\pi\,\sigma^2)\right] \tag{3.13}$$

This function penalises large errors while ensuring that the predicted uncertainty $\boldsymbol{\sigma}$ remains balanced.

Although NLL is often used as a criterion to evaluate uncertainty models, relying solely on NLL is insufficient. A very small standard deviation can significantly reduce the NLL value; however, if the standard deviation is excessively small and fails to cover a sufficient proportion of the observed data, it becomes meaningless.

In this study, the reliability of predictive uncertainty is evaluated using the 95% Credible Interval (CrI) coverage. This metric assesses whether the predicted uncertainty intervals align with the expected proportion of the data. Specifically, a well-calibrated model should produce a 95% CrI that contains approximately 95% of the actual observed values. This method provides a straightforward way to evaluate the consistency and reliability of uncertainty estimates across different models.

$$CrI_{95} = \mu - 1.96 \cdot \sigma, \mu + 1.96 \cdot \sigma \tag{3.14}$$

A coverage rate close to 95% indicates that the uncertainty estimates are well-calibrated, while deviations suggest potential overconfidence or over-conservation of the model – if it is smaller than 95%, than prediction variability is smaller than the true variability and the model is too confident. This metric is particularly useful in comparing the performance of different models in terms of their ability to quantify uncertainty accurately.

Due to the nonlinear transformation applied to the target variable in this study, the transformed data conforms to a normal distribution. However, when the results are reverted to the original scale, the distribution on the original scale no longer retains normality due to the nature of the nonlinear transformation. This change is primarily attributed to the introduction of skewness, causing the CrI bounds to become asymmetric.

Nonetheless, the probabilistic interpretation of the 95% CrI remains valid. Specifically, the posterior cumulative probability ensures that 95% of the parameter values fall within the CrI, and the amount of data contained within the CrI remains unchanged.

To quantify the importance of input variables and understand their contributions to the predictive performance of the models, SHAP (SHapley Additive exPlanations) was employed. SHAP is a game-theoretic approach that calculates the marginal contribution of each feature to the output of model, ensuring a consistent and interpretable measure of feature importance (Mitchell et al., 2022).

$$\phi_i = \sum_{S \subseteq N \setminus i} \frac{|S|!\,(|N| - |S| - 1)!}{|N|!} [f(S \cup i) - f(S)] \tag{3.15}$$

Where ϕ_i represents the importance score of features i, quantifying its average marginal contribution to the output across all possible subsets of features. N denotes the set of all features, and S is any subset of N that excludes feature i. The contribution of feature i is calculated as the change in the output of model when feature i is added to subset S, expressed as $f(S \cup i) - f(S)$. These contributions are then averaged across all subsets S, weighted by $\frac{|S|!(|N|-|S|-1)!}{|N|!}$, which ensures that all possible orders and combinations of feature inclusion are considered.

3.3 Methodology

3.3.1 Data Source

This study primarily utilises data from the WoSIS 2019 database, a globally recognised repository offering rigorously standardised and quality-controlled soil profile data. WoSIS provides data for large-scale digital soil mapping and environmental research, compiling 196,498 georeferenced soil profiles across 173 countries, with over 832,000 soil horizons and 5.8 million records (Batjes et al., 2020). This extensive dataset encompasses a wide range of essential soil physical and chemical properties, including SOC, total carbon, total carbonate, total nitrogen, phosphorus content, soil pH, cation exchange capacity, electrical conductivity, soil texture (e.g., sand, silt, and clay fractions), bulk density, coarse fragment content, and water retention capacity. All data points are georeferenced with high precision and adhere to international classification standards, such as United States Department of Agriculture (Batjes et al., 2020). The comprehensive nature of this dataset makes it highly suitable for digital soil mapping and large-scale environmental modelling applications.

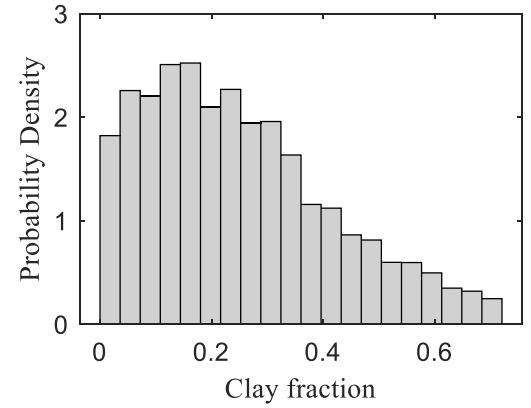
3.3.2 Study design

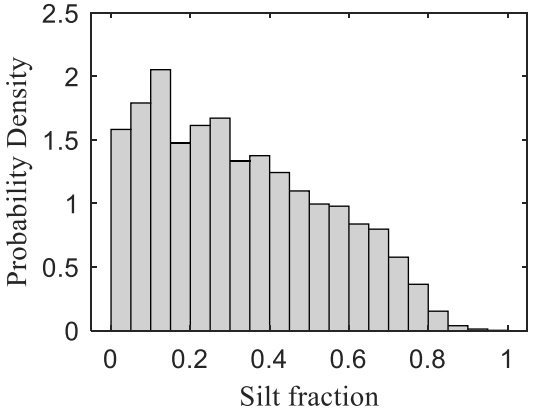
Guided by the existing literature (Chen et al., 2024; Patton et al., 2019; Rawls et al., 2003; Zhao et al., 2021), key soil properties were selected based on their relevance to SOC and availability within the WoSIS database. The features have been shown as Table 1. Nitrogen serves as a critical predictor for SOC estimation. In this study, SOC estimation was addressed under two distinct Scenarios: Scenarios #1 is nitrogen included in input features; Scenarios #2 is without nitrogen, a setup that aligns with datasets lacking nitrogen information, such as UNSODA. This approach allows for model adaptability based on dataset availability and reflects real-world variations in soil dataset completeness.

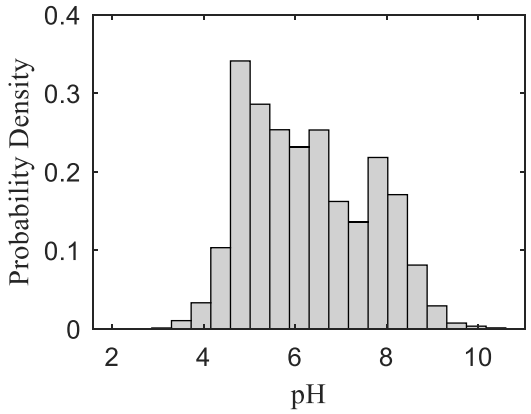
Table 1 Feature selection

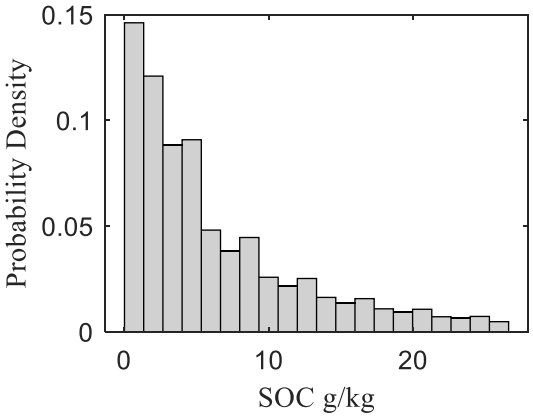
Feature	Influence on SOC	Unit
Bulk Density	Indicates soil structure and porosity	kg/dm^3

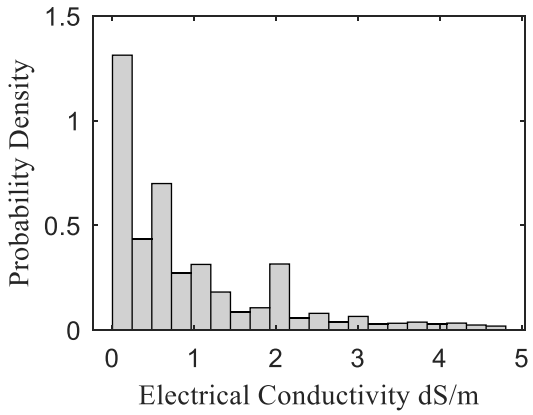
Electrical Conductivity (EC)	Reflects the concentration of soluble salts in the soil, which can indirectly influence SOC by affecting microbial activity and nutrient availability.	cmol(c)/kg
Nitrogen	A nutrient essential for organic matter decomposition, closely related to SOC cycling.	g/kg
pH	Influences SOC decomposition rates, with acidic or alkaline conditions affecting organic matter breakdown.	N/A
Clay fraction	Slay fraction, relevant for characterising soil composition affecting SOC.	N/A
Soil Depth	Captures variations in SOC at different soil horizons, with depth impacting organic matter presence.	cm
Silt fraction	Silt fraction, relevant for characterising soil composition affecting SOC.	N/A

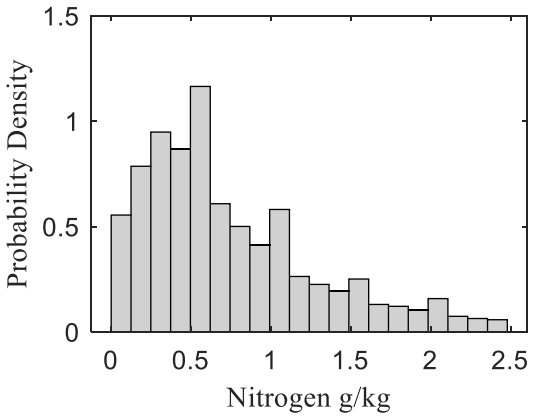





Probability Density
0.4
0.3
0.2
0.1
0
2
4
6
8
10
pH


Probability Density
0.15
0.1
0.05
0
0
10
20
SOC g/kg


Probability Density
1.5
1
0.5
0
0
1
2
3
4
5
Electrical Conductivity dS/m


Probability Density
1.5
1
0.5
0
0
0.5
1
1.5
2
2.5
Nitrogen g/kg

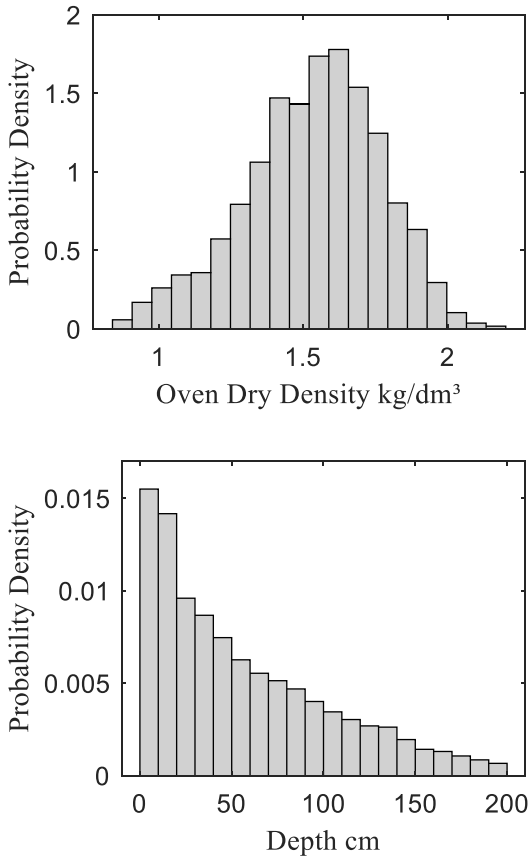


Figure 2 Probability density distributions of key soil properties.

The WoSIS dataset includes SOC measurements obtained through three different methods: Loss on Ignition (LOI), wet oxidation, and unknown methods. Given that 77% of SOC data lacks explicit measurement method details, this study treats all SOC values uniformly for modelling purposes, rather than attempting to differentiate between measurement methods. This approach minimises potential biases stemming from measurement variability, ensuring a consistent SOC dataset for model training.

3.3.3 Data Processing and Preprocessing

In data-driven research, data quality directly determines the performance of the model and the credibility of its predictive results. However, raw data often contains missing values, and outliers, which can lead to biases in the model training process or even result in distorted outcomes. To ensure the robustness and reliability of the model, it is crucial to conduct scientific and systematic data preprocessing. By removing outliers, and imputing missing values, the overall data quality can be significantly improved, providing

a stable foundation for model training. To ensure model robustness and optimise predictive performance, the IQR method was implemented to remove outliers in this study. This method is widely applied in the field of data analysis (Ditzhaus et al., 2021). First, outlier removal was conducted using the IQR method, which isolates the central portion of the data distribution while eliminating outliners that could otherwise skew model outcomes. This approach effectively minimises the impact of outliers and ensures a more stable data foundation for modelling.

To address multicollinearity, representative features were excluded when high correlations were detected among variables. For example, total organic carbon was excluded due to its close correlation with SOC, while bulk density was retained as a proxy for soil structure and porosity based on its established linear relationship with these properties. Furthermore, given the absence of certain features relevant to SOC prediction, such as soil structure, porosity, soil temperature, and permeability, bulk density was also used as a substitute due to its known correlation with these attributes.

For the bulk density measurement, oven-dry bulk density was selected as the primary density variable. Despite air-dry bulk density potentially offering advantages for SOC studies, oven-dry density aligns more closely with civil engineering applications and easily measured. Its strong correlation with SOC and widespread use in engineering contexts justified its use in this analysis. The distribution of selected inputs is shown as Figure 2Error! Reference source not found.

For the machine-learning models, the input features were standardised using standardisation method, while for the VB models, a logarithmic transformation was applied to the input features. For the target variable, both machine-learning and VB models used a logarithmic transformation. These transformations were determined based on experimental results to optimise model performance and stability.

3.3.4 Model Development and Comparative Analysis

To rigorously evaluate SOC prediction performance, five distinct models were developed and compared: linear regression, XGBoost, NN, VB linear and VBNN. To ensure a fair comparison among models, hyperparameters for each model were optimised through grid

search. A consistent learning rate optimisation process was applied to all models, testing values of 0.1, 0.01, and 0.001.

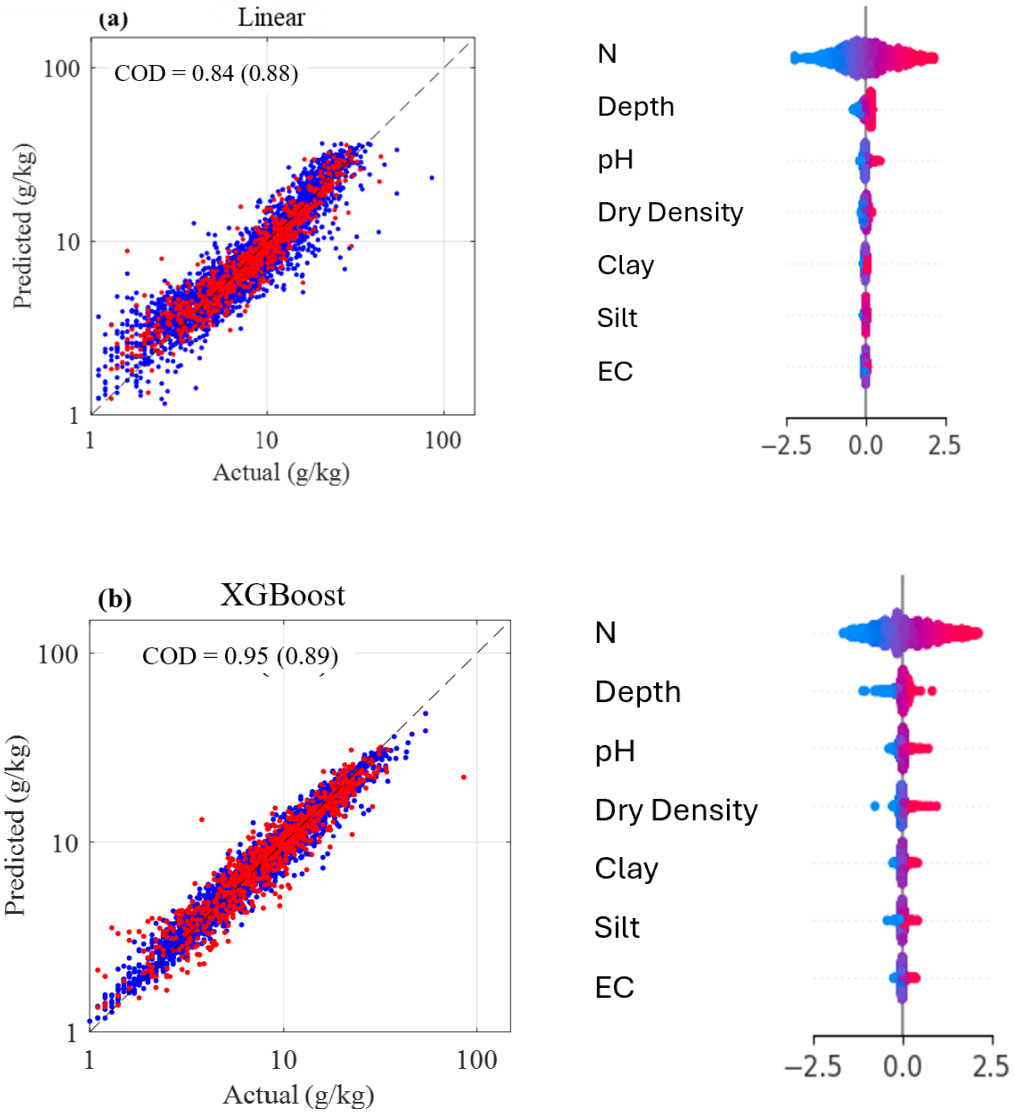
For NN, the architectures included 1 to 2 hidden layers, with each layer containing 16, 32, 64, or 128 neurons. Early stopping was applied based on validation loss to mitigate overfitting.

For XGBoost, grid search was conducted on the number of estimators (10, 50, 100), maximum depth (3, 5, 7, 9), and column subsampling ratios (0.6, 0.8, 1.0). These parameters were selected to balance model complexity and robustness, ensuring an efficient and thorough evaluation of hyperparameters.

For the VB models, the KL divergence weight w_{KL} was systematically adjusted to evaluate its effect on predictive accuracy and uncertainty quantification. Tested values included $\frac{0.1}{N_{samples}}$, $\frac{1}{N_{samples}}$, $\frac{10}{N_{samples}}$ where $N_{samples}$ represents the number of input samples.

All models were trained on identical input datasets, and hyperparameter performance was evaluated on a validation set to ensure consistency and robustness. However, due to the limited amount of data, grid search alone might not guarantee the discovery of optimal hyperparameters, as the validation set may not fully represent the underlying data distribution. Some hyperparameters were manually fine-tuned such as the standard deviation of prior distributions in the Bayesian models. This manual adjustment allowed for targeted exploration of parameter settings, leveraging empirical observations to enhance the predictive performance of model.

3.4 Results


(a)
Linear
COD = 0.84 (0.88)
Predicted (g/kg)
Actual (g/kg)
100
10
1
1
10
100
N
Depth
pH
Dry Density
Clay
Silt
EC
−2.5
0.0
2.5
(b)
XGBoost
COD = 0.95 (0.89)
Predicted (g/kg)
Actual (g/kg)
100
10
1
1
10
100
N
Depth
pH
Dry Density
Clay
Silt
EC
−2.5
0.0
2.5

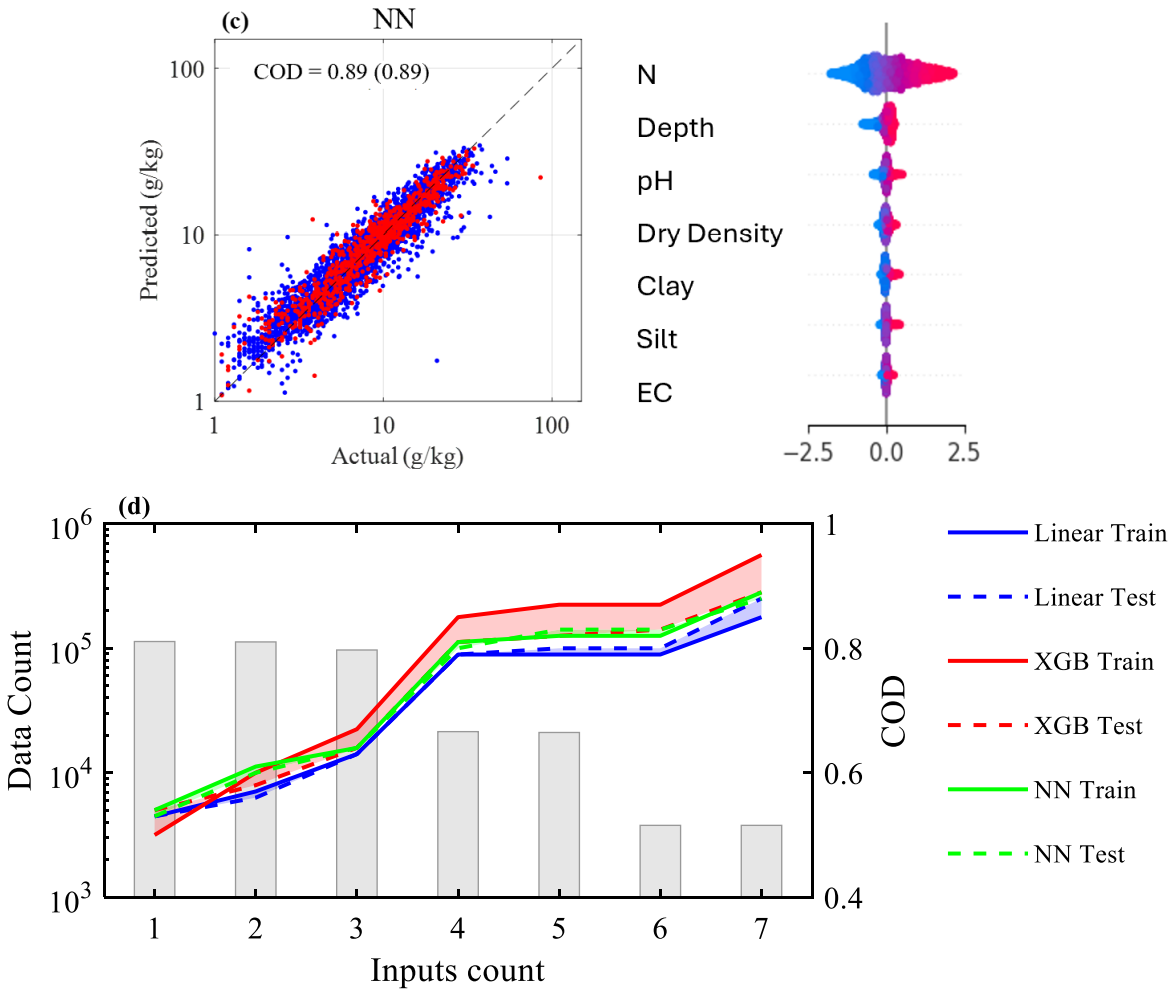
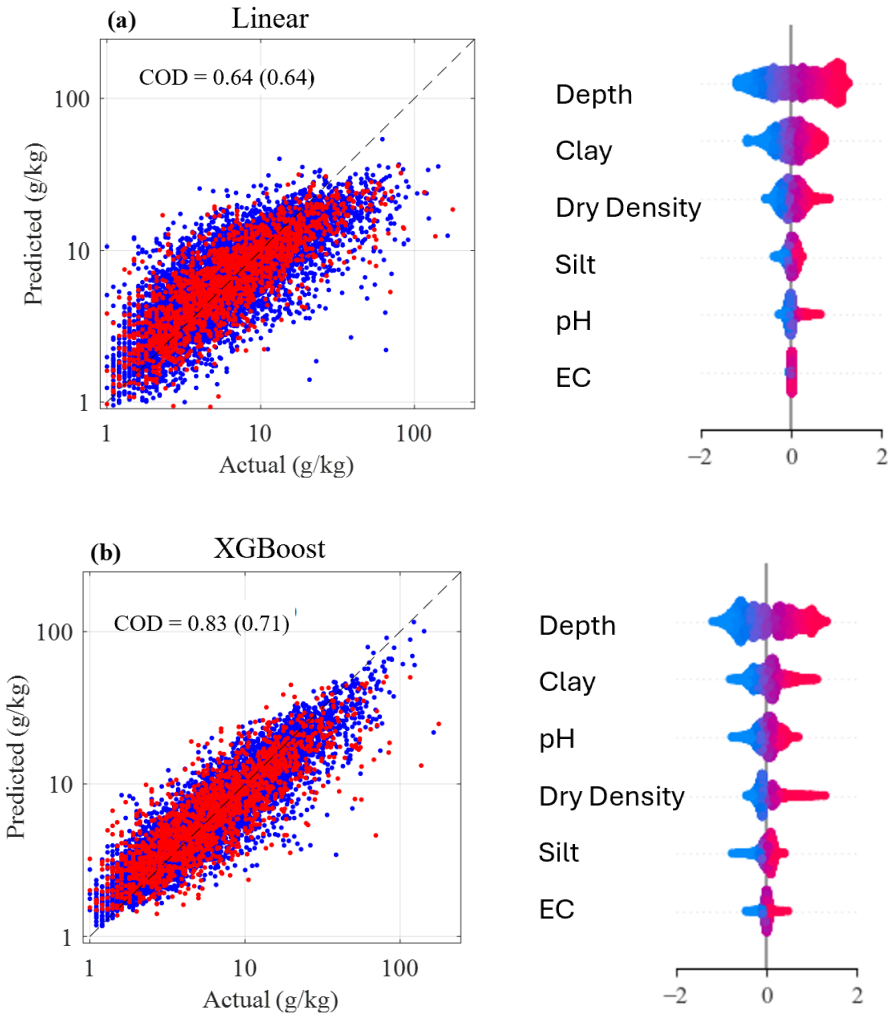


Figure 3 Performance of machine-learning models for Scenario #1 with nitrogen in input features. ((a), (b), (c) represent the results of linear regression, XGBoost, and NN with the values in parentheses represent the COD for the testing set, (d) is the variation of COD and available data size for using different numbers of input features e.g., 1 for horizontal axis shows results of using only 1 feature, which is the most important for modelling.).

The predictive performance of three models—linear regression, XGBoost, and NN, in estimating SOC of Scenario #1 is evaluated in Figure 3. Linear regression achieves COD of 0.84 and 0.88 for the training and testing sets, respectively. While these results indicate a reasonable fit, the simplicity of model limits its capacity to fully capture the underlying complexity of SOC prediction, leading to underfitting. In contrast, the XGBoost model demonstrates exceptional accuracy on the training set, with a COD of 0.95, but a slightly lower value of 0.89 on the testing set. This discrepancy suggests that XGBoost is prone

to overfitting, particularly. Meanwhile, the NN model exhibits the most consistent performance, achieving COD values of 0.89 on both the training and testing sets. This balance highlights its superior ability to generalise across datasets, making it the most reliable model among the three.

A more nuanced understanding of model behaviour is provided through the analysis of SHAP values, which reveal the relative importance of input features. Nitrogen emerges as the most influential predictor, followed by variables such as depth and pH. Figure 3 (d) further explores the interplay between model performance, the number of input variables, and data availability. As the number of input features increases, the amount of available data diminishes due to sparsity. This trade-off is evident in the responses: XGBoost increasingly overfits as input dimensionality grows, while linear regression struggles with underfitting, unable to capture complex interactions. In contrast, NN maintains robust performance, effectively balancing complexity and generalisability.


(a)
Linear
COD = 0.64 (0.64)
Predicted (g/kg)
Actual (g/kg)
100
10
1
1
10
100
Depth
Clay
Dry Density
Silt
pH
EC
−2
0
2
(b)
XGBoost
COD = 0.83 (0.71)
Predicted (g/kg)
Actual (g/kg)
100
10
1
1
10
100
Depth
Clay
pH
Dry Density
Silt
EC
−2
0
2

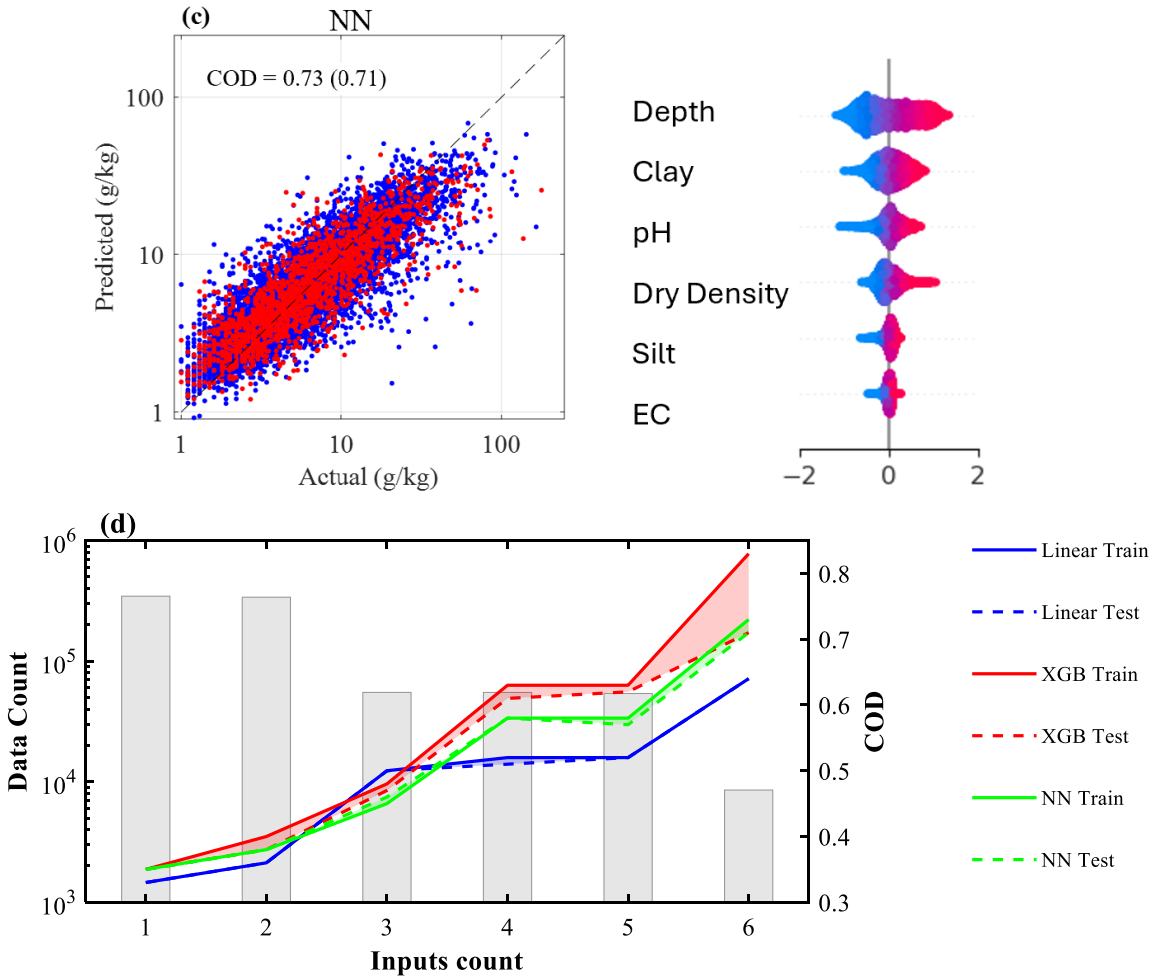


Figure 4 Performance of machine-learning models for Scenario #2.

In Scenario #2, a noticeable decrease in predictive accuracy is observed across all models compared to Scenarios #1. For instance, in Figure 4, the linear model, the COD decreases significantly from 0.88 to 0.64, reflecting a substantial loss of explanatory power. Similarly, the XGBoost model experiences a reduction in COD from 0.89 to 0.71, though it still exhibits overfitting tendencies relative to other models. The NN also sees a decline in accuracy, with COD dropping from 0.88 to 0.71 when nitrogen is not included.

When the analysis is further restricted to panel (d), the results show a consistent, albeit slight, reduction in COD across all models. Interestingly, the NN maintains comparable COD values to linear regression while demonstrating stability without significant overfitting. Meanwhile, XGBoost continues to achieve the highest COD but retains noticeable overfitting issues, as indicated by the red shaded area in the figure. Linear

regression, however, struggles to effectively capture the relationship between inputs and the target variable, resulting in substantially lower COD performance compared to NN and XGBoost.

These results underscore the critical role of nitrogen as an input feature in improving model performance, particularly for the linear and NN models. At the same time, it highlights the persistent overfitting challenge associated with XGBoost.

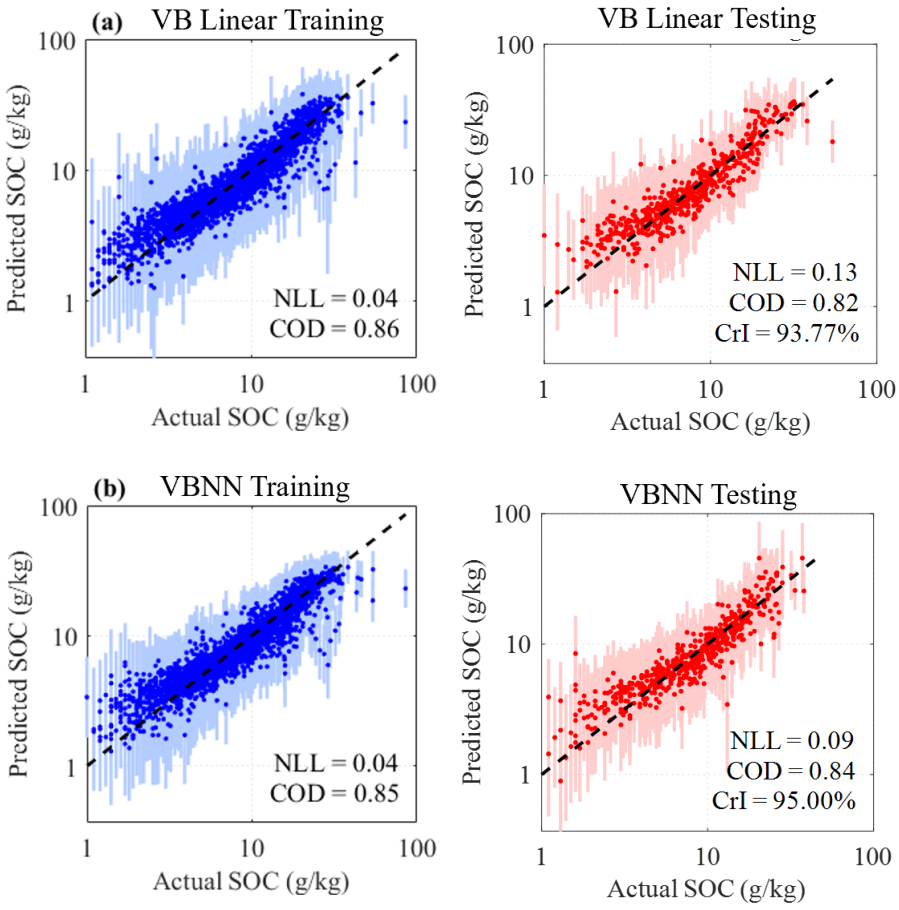


Figure 5 Performance of variational Bayesian models for Scenario #1. ((a), (b) represent the results of VB linear and VBNN)

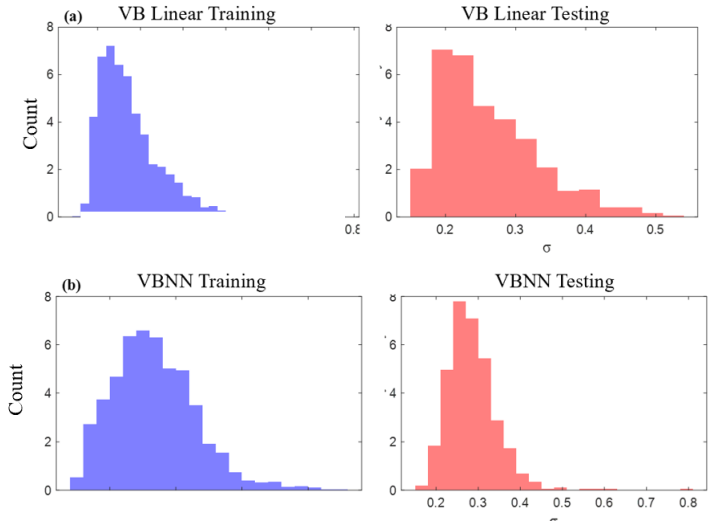
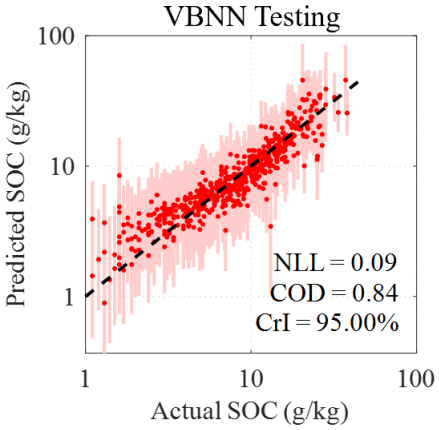


Figure 6 Probability density distribution of standard deviations in Scenario #1. (a) is VB Linear model. (b) is VBNN model.

In Scenario #1, the VB linear and VBNN models exhibit notable contrasts in predictive performance and uncertainty quantification. In Figure 5, the VB linear model achieves a COD of 0.82 on the test set, indicating moderate predictive accuracy. However, its NLL value of 0.13 reveals higher uncertainty, suggesting limitations in generalising to unseen data. By comparison, the VBNN model outperforms with a COD of 0.84, accompanied by a consistently lower NLL of 0.09. This improvement is further emphasised by the 95% CrI coverage, where VBNN achieves the expected 95% coverage, whereas VB linear falls short, indicating that VBNN provides more reliable uncertainty estimates.



The standard deviation distributions further illustrate these differences in uncertainty quantification. In Figure 6, VB linear shows a broader spread, reflecting greater variability and less precise uncertainty estimation. In contrast, VBNN demonstrates a narrower and more uniform distribution, signalling its capability to model complex data relationships while maintaining calibrated and robust predictions. These results highlight clear advantage of VBNN model in balancing predictive accuracy and uncertainty reliability, making it better suited for applications demanding both precision and confidence in model outputs.

Due to the use of logarithm transformation in this study, the data values became significantly smaller, resulting in very small standard deviations. Consequently, the NLL values are close to zero. Based on the standard deviation distribution, this is a normal phenomenon observed in datasets with small numerical values.

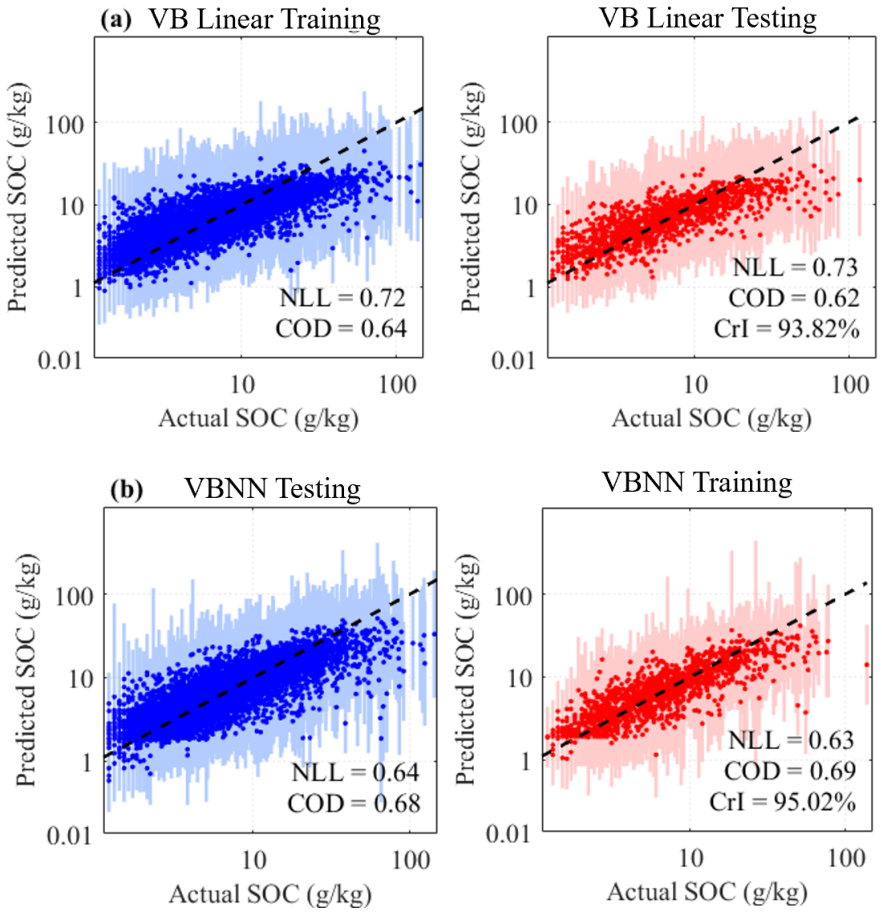


Figure 7 Performance of variational Bayesian models for Scenario #2. (a) is VB Linear model. (b) is VBNN model.

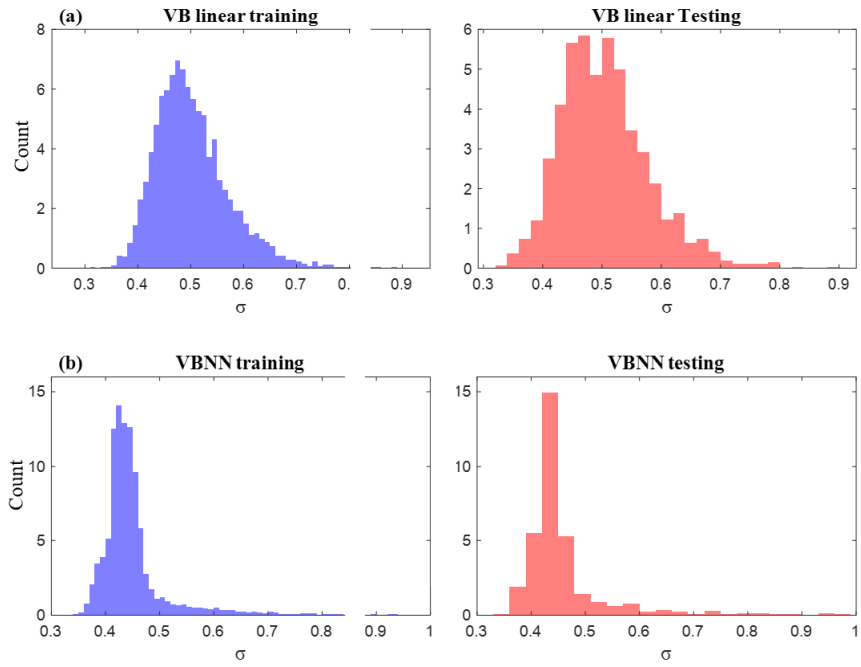


Figure 8 The probability density distribution of standard deviations in #2. (a) is VB Linear model. (b) is VBNN model.

Scenario #2 of VB results in notable differences in predictive performance and uncertainty between the VB linear and VBNN models. For the test set in Figure 7, the VB linear model yields an NLL of 0.73, indicating higher uncertainty compared to Scenarios #1. The VBNN model, on the other hand, demonstrates relatively better uncertainty quantification, with an NLL of 0.63 and CrI of 95.02%. However, both models show decreased predictive accuracy, as evidenced by the wider spread of predictions, highlighting the critical role of nitrogen in improving model performance. The predicted standard deviation distributions of Figure 8 provide further insights into the uncertainty handling. For VB linear, the standard deviations are narrowly distributed around 0.5, reflecting a more uniform but potentially underestimated uncertainty across predictions. In contrast, VBNN displays a more skewed distribution, particularly in the test set, where standard deviations occasionally approach 1 but are more frequently distributed between 0.4 and 0.5. This suggests that VBNN is better equipped to capture complex uncertainty patterns but may occasionally overestimate uncertainty for certain predictions.

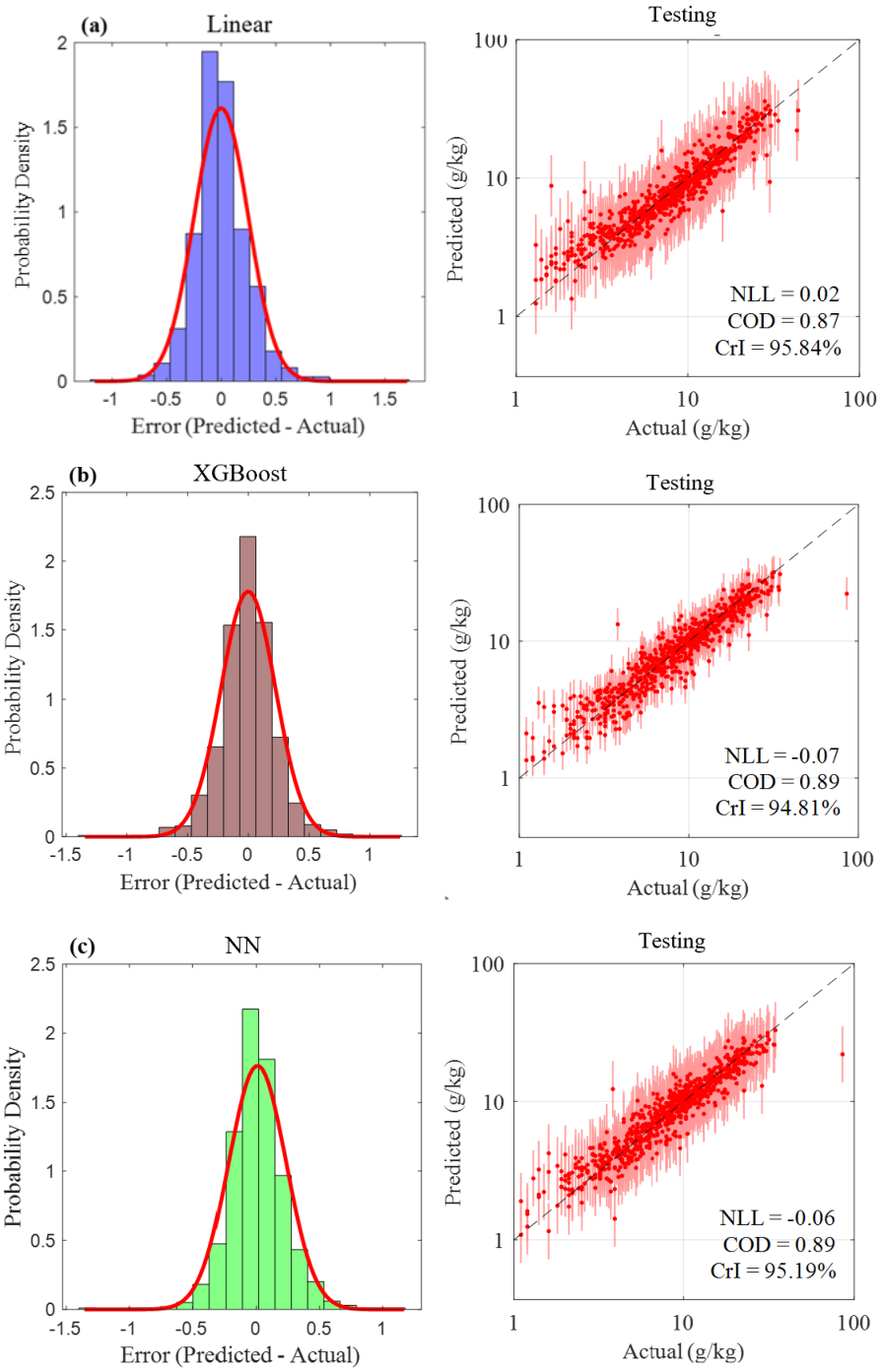


Figure 9 Performance of uncertainty quantification of machine-learning models for Scenario #1. (a) is Linear model. (b) is XGBoost model. (c) is NN model

The machine-learning models can also be used for probabilistic predictions by estimating a scale of prediction error between predictions and true values on training data (***σ*** in the left column of Figure 9). With the machine-learning prediction as mean values, the

estimated scale as standard deviation, we obtain a probabilistic model that can be used on new data – the testing data (right column of Figure 9).

The evaluation of these probabilistic models from the linear regression, XGBoost, and NN models in Scenario #1 reveals key insights into their performance. The probability density distributions of prediction errors on the training set indicate that the linear regression, XGBoost, and NN models achieve standard deviations of 0.25, 0.22, and 0.23, respectively, suggesting similar levels of error dispersion among the models.

The predicted SOC values versus actual values, visualised with error bars representing predictive uncertainty, highlight differences in the ability to quantify uncertainty of models. The test set NLL values are 0.02 for linear regression, -0.07 for XGBoost, and -0.06 for NN. The negative NLL for NN and XGB are due to the combination of a relatively small standard deviation and low prediction errors, resulting in a well-calibrated uncertainty quantification as per the NLL formula $log(2\pi\sigma^2)$. The NN and linear regression models also achieve 95% CrI data coverage rates of 95.06% and 95.32% separately, aligning closely with the expected value and confirming the reliability of its uncertainty estimates. By comparison, XGBoost achieves a lower data coverage rate of 94.81%, indicating an underestimation of uncertainty and challenges in producing reliable uncertainty estimates despite its strong predictive power.

These results, as summarised in Figure 9, demonstrate that while all three models exhibit comparable prediction error distributions. The NN model achieves a robust balance between predictive accuracy and reliable uncertainty quantification, surpassing XGBoost and matching linear regression in 95% CrI coverage reliability.

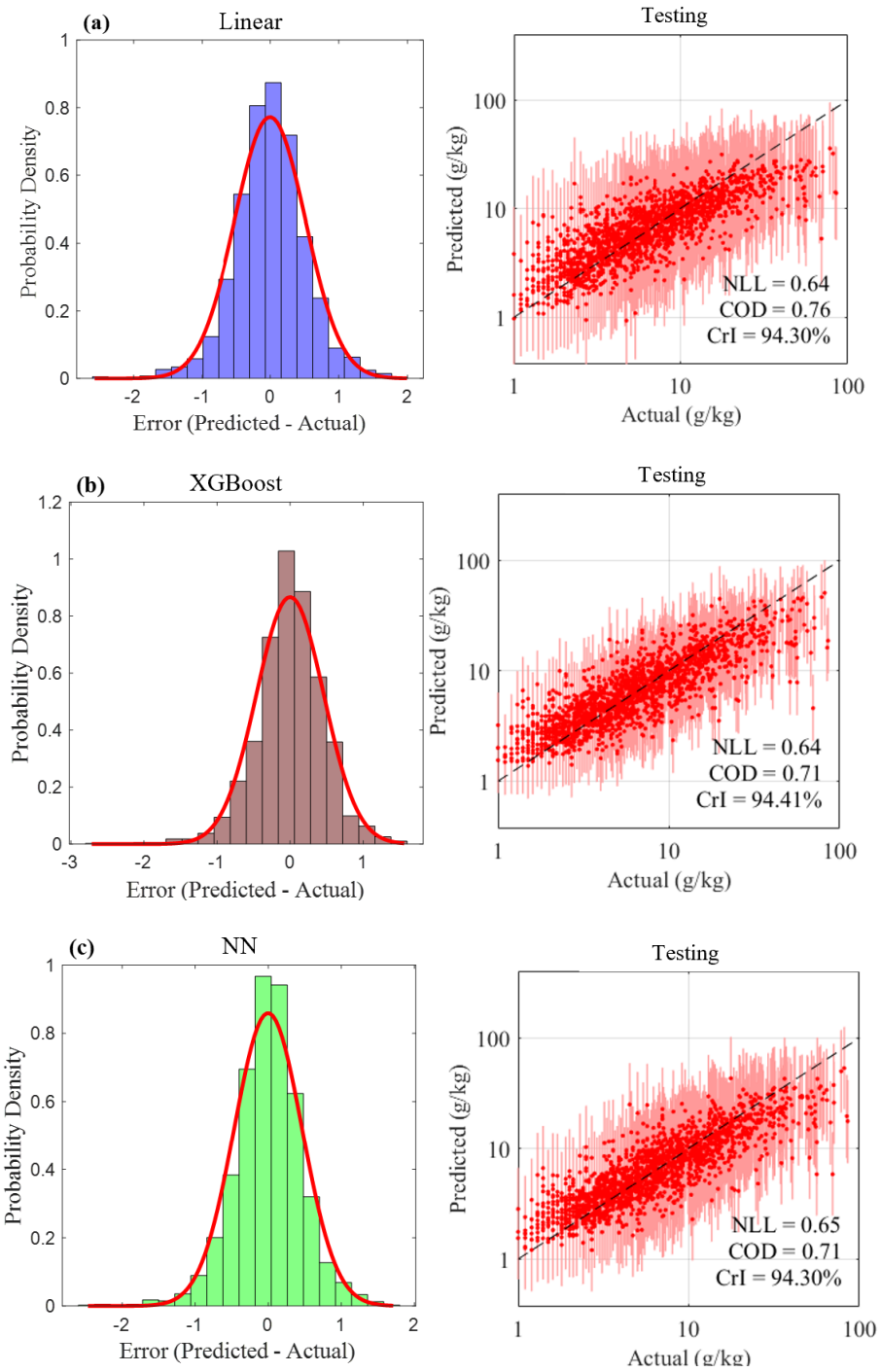


Figure 10 Performance of uncertainty quantification of machine-learning models for Scenario #2. (a) is Linear model. (b) is XGBoost model. (c) is NN model

In Scenario #2, the predictive uncertainty increases, and accuracy declines slightly across all models compared to Scenario #1. In Figure 10, all three models show notable challenges in maintaining predictive accuracy and reliable uncertainty quantification. For

the linear regression model, the standard deviation rises significantly from 0.25 to 0.52, accompanied by an increase in NLL from 0.02 to 0.64, reflecting a marked deterioration in both accuracy and uncertainty estimation. This indicates that the model struggles to provide reliable uncertainty estimates under the new conditions. Similarly, XGBoost experiences a substantial rise in NLL, from -0.07 to 0.64, with a 95% CrI coverage rate of 94.41%, suggesting that its uncertainty quantification is also less effective in this scenario.

The NN model, while showing a slight advantage in Scenario #1, demonstrates comparable performance to the other models in Scenario #2. Its NLL increases from -0.06 to 0.65, and the 95% CrI coverage rate decreases marginally from 95.06% to 94.30%. While this indicates relative stability compared to the linear regression and XGBoost models, the NN fails to exhibit a significant edge in uncertainty quantification under these less favourable conditions. Notably, all three models achieve coverage rates below the expected 95%, highlighting their overconfidence in this scenario of limited information and their shared limitations in providing fully reliable uncertainty estimation.

3.5 Discussion

Table 2 Results of models

	Scenario #1			Scenario #2		
	COD	NLL	Cover rate (%)	COD	NLL	Cover rate (%)
Linear	0.87	0.02	95.32	0.64	0.76	94.30
XGBoost	0.89	-0.07	94.81	0.71	0.64	94.41
NN	0.89	-0.06	95.06	0.71	0.65	94.30
VB linear	0.82	0.13	93.77	0.63	0.73	93.82
VBNN	0.84	0.09	95.00	0.69	0.62	95.02

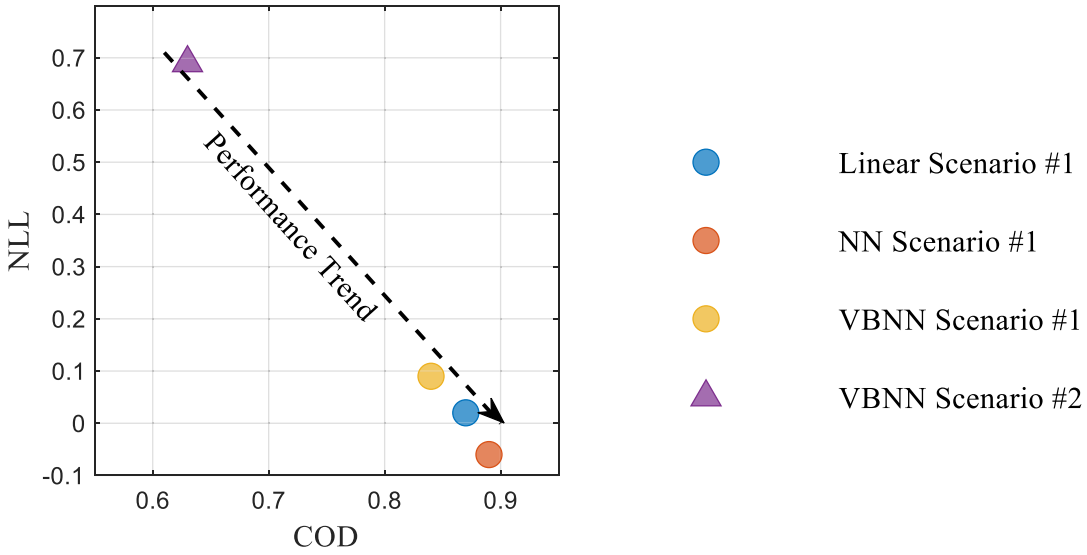


Figure 11 Comparison of the models that only satisfy the criteria of 95% CrI based on performance in terms of COD and NLL.

The analysis reveals significant variations in the performance of machine-learning models and VB machine-learning models between Scenario #1 and Scenario #2, emphasising the role of nitrogen as a key feature for enhancing both predictive accuracy and uncertainty

quantification. These findings also highlight the comparative strengths and weaknesses of different models in predicting SOC mean values and their associated uncertainties. A closer examination of the results, as shown in Table 2 and Figure 11 provides further insight into these performance differences and the critical influence of nitrogen on model outcomes.

3.5.1 Implications of Missing Key Inputs

The comparison between the two scenarios highlights the critical role of nitrogen in model performance. Without nitrogen, there is a noticeable increase in NLL values across all models, indicating higher uncertainty in predictions. For instance, the NLL of the linear regression model increases from 0.02 to 0.76, while its COD drops from 0.87 to 0.64, suggesting that the model struggles to establish meaningful relationships between the remaining inputs and the target variable.

3.5.2 Cover Rate and Uncertainty Estimation

For Scenario #2 with limited information, all the machine-learning models and VB linear are too confident, achieving a credible interval cover below 95%. However, only the VBNN achieves a credible interval that covers more than 95% of the total data points, meeting the threshold for reliable uncertainty estimation. This result indicates that VBNN is better equipped to capture the uncertainty inherent in the predictions when key input variables are missing. This suggests that these models are less robust in handling scenarios with reduced input information, leading to overconfidence and underestimation of predictive uncertainty.

For Scenario #1, there are three models that achieve cover rates where more than 95% of the total data points are within 95% CrI, including linear regression, NN, and VBNN., indicating that the inclusion of nitrogen significantly enhances the reliability of uncertainty estimation across all model types. This improvement can be attributed to the additional predictive power provided by nitrogen, which reduces the uncertainty in the predictions and enables more accurate modelling of the target variable.

3.5.3 Predictive Accuracy and Model Performance

In terms of predictive accuracy, as reflected by the COD, NN and XGBoost perform the best in both Scenarios. However, when analysing uncertainty quantification through NLL,

NN stands out as the best model. NN consistently achieves the lowest NLL values with 95% CrI criteria in Scenario #1, indicating its superior ability to balance predictive accuracy and uncertainty estimation. In Scenario #2, only VBNN satisfy the 95% CrI criteria.

3.5.4 Trend Line Analysis and Model Selection

The performance trend line further illustrates the trade-offs between predictive accuracy and uncertainty quantification across models. In Figure 11, the ideal model that satisfy the criteria of 95% CrI lies in the bottom-right corner of the plot, where COD is maximised, and NLL is minimised. NN consistently positions itself closest to this region in Scenario #1, making it the optimal choice for scenarios where both accuracy and uncertainty estimation are critical.

3.5.5 Practical Implications

From a practical perspective, these findings emphasise the importance of including relevant input variables, such as nitrogen, in predictive models for soil-related studies. The ability of NN and VBNN to balance accuracy and uncertainty makes them valuable tools for tasks such as SOC prediction, especially when uncertainty estimation is a priority. Additionally, the results suggest that model selection should be guided not only by accuracy metrics like COD but also by the reliability of uncertainty estimates, as quantified by NLL and cover rates.

3.5.6 Limitations and Future Work

While the study provides valuable insights, some limitations should be noted. First, the performance of models was evaluated using a specific dataset, WoSIS 2019, and the results may vary with different datasets or application contexts. Additionally, the use of 6 and 7 input features significantly reduced the amount of available data, which negatively impacted the performance of machine-learning models by limiting the training capacity and potentially increasing overfitting risks. Future research could explore the impact of these factors and investigate the generalisability of the findings to other predictive tasks. Moreover, addressing the trade-offs between the number of input variables and data availability could lead to more robust modelling strategies in Scenarios with limited data.

3.6 Conclusion

This study provides a comprehensive comparison of machine-learning models in predicting SOC and estimating predictive uncertainty. It focuses on the quantification of SOC uncertainty and the evaluation of model performance under varying conditions. By analysing both the COD and NLL, the study highlights the trade-offs between predictive accuracy and uncertainty quantification across different models. Furthermore, this research adopts a flexible approach by treating nitrogen as a removable input variable, allowing SOC predictions to be made without relying on datasets that include nitrogen. This approach addresses the common challenge of incomplete soil datasets, enabling the applicability of the models to a broader range of datasets and scenarios. By reducing the dependence on nitrogen, the study enhances the versatility of SOC prediction methods and provides a framework for more inclusive soil data analysis.

The results demonstrate that NN consistently achieve the best balance between accuracy and uncertainty estimation, making them the optimal choice for SOC prediction when both aspects are critical. VBNN, on the other hand, excel in uncertainty quantification, particularly in scenarios with limited input variables, as evidenced by their superior performance in datasets without nitrogen. These findings emphasise the robustness of VBNN in handling predictive uncertainty and its suitability for applications requiring reliable uncertainty estimates.

Moreover, the study underscores the challenges posed by increasing the number of input variables, which can reduce data availability and negatively impact machine-learning performance. This highlights the importance of addressing the trade-offs between input complexity and data volume in predictive modelling.

In conclusion, this research provides new insights into SOC uncertainty estimation and model comparison, offering practical guidance for selecting and optimising models in soil-related studies. Future work could further explore the integration of uncertainty quantification into decision-making processes and assess the generalisability of these findings to other predictive tasks in soil science and beyond.

4. Chapter 4: Incorporating Uncertainty into Maize Yield Prediction Using Variational Bayesian Theorem and Geographically Weighted Regression

Abstract

Following the uncertainty quantification framework developed in the previous chapter, this chapter applies similar probabilistic modelling principles to maize yield prediction. Extending from point-based soil modelling to regional-scale crop yield estimation, it integrates spatial dependence and uncertainty into a unified framework.

Accurately predicting maize yields is crucial for optimising agricultural production as well as investment returns and risk analysis. This study applies advanced machine learning and statistical models to predict maize yields across 842 counties in the U.S. Corn Belt from 2014 to 2023. Key input variables include soil characteristics such as pH, available water capacity, bulk density, electrical conductivity, and cation exchange capacity, as well as climate factors like air temperature, humidity, precipitation, and shortwave radiation, alongside additional data such as normalized difference vegetation index from remote sensing and planting and harvesting dates. Four models were evaluated, including linear regression as a baseline, geographically weighted regression (GWR), a model combining variational Bayesian with GWR to consider uncertainty quantification (denoted as VB-GWR), and a model further utilises neural network to account for nonlinearity (denoted as VBNN-GWR). VB-GWR demonstrates the best overall performance. The analysis shows that VB-GWR accurately predicts yields within 150–220 bu/acre, with pH as the most influential factor and precipitation having minimal impact. Sensitivity analysis identifies optimal input ranges, aiding land valuation and decisions for new lands. This study underscores the importance of integrating soil, climate, and geographical data for yield prediction, enhancing land valuation and agricultural management. This study quantifies uncertainty across the core and peripheral maize-producing regions by analysing geospatially relevant parameters. The findings provide valuable insights for assessing land costs, investment returns, and risk analysis, particularly for newly developed agricultural lands. By integrating soil, climate, and geographical data, this study enhances yield prediction accuracy, land valuation, and agricultural management strategies.

Keywords: Corn Belt; Machine-learning; Maize yield; GWR; Variational Bayesian; USDA; Uncertainty quantification.

4.1 Introduction

The rapid growth of global population and demand for biofuels has significantly increased the need for agriculture production, placing unprecedented pressure on agricultural systems worldwide (Tilman et al., 2011; Timilsina et al., 2012). As one of the foremost agricultural producers, the United States has undergone substantial intensification and expansion in agricultural practices and land use, particularly in maize cultivation. The US is the largest producer and exporter of maize of the world, significantly influencing global supply and pricing (Tollenaar et al., 2017). It accounts for a substantial 30% to 40% of global maize production, with the Corn belt region contributing approximately 90% of the US national output (Leng, 2019; Morell et al., 2016; Zhao et al., 2024; Zilberman et al., 2012). Since 2000, maize has experienced significant growth in both yield and acreage (Annan et al., 2024; Hertel et al., 2010). Beyond its traditional uses in food and feed, the increasing demand for biofuel production, driven by the Renewable Fuel Standard, has further amplified the need for maize as a key input for ethanol (Dhoubhadel et al., 2015; Roberts & Schlenker, 2013). This shift has also strengthened the interconnections between energy and agricultural commodity prices, particularly in the biofuels era, where structural changes in natural gas, corn, and fertilizer price relationships have been observed (Beckman & Riche, 2015; Boyer et al., 2015).

The significant increase in maize production in the US over recent decades underscores the importance of yield prediction (Rotundo et al., 2024). Accurate yield predictions in the US have direct implications for global food security and agricultural land value. Predicting crop yield remains a complex task due to many interrelated factors, including soil characteristics, climatic conditions, and vegetation indices derived from remote sensing data (Jithitikulchai et al., 2018; Lobell & Burke, 2010). In recent years, many study highlight the potential of integrating machine learning and remote sensing to enhance yield prediction accuracy (Joshi et al., 2021; Peng et al., 2018). These technologies can leverage vast datasets, capturing complex interactions between environmental variables and yield.

Recent advancements in machine learning have demonstrated significant potential in enhancing yield estimation accuracy by leveraging remote sensing data (Johnson, 2016; Schwalbert et al., 2020; Shahhosseini et al., 2020; Wu et al., 2022). Remote sensing

technologies offer novel tools for large-scale crop monitoring and yield estimation, with indices such as the Normalized Difference Vegetation Index (NDVI) being widely adopted for assessing vegetation health and predicting crop yield (Burke & Lobell, 2017; Hao et al., 2016; Lobell et al., 2015; Qiao et al., 2021). High-resolution sensors like Landsat 8 provide detailed spatial information suited for analysing specific agricultural regions, while sensors like MODIS offer more frequent temporal coverage at lower spatial resolutions (Khan et al., 2019). NDVI data, which indirectly reflects regional vegetation growth conditions, serves as a valuable input for yield prediction models. Additionally, climatic variables are key determinants of maize yield (Lobell & Burke, 2010).

A study used satellite data and machine learning to evaluate the impact of tillage practices on maize yields in the western U.S. Corn Belt (Cambron et al., 2024). The results indicated that low-tillage practices were correlated with higher yields in semi-arid regions, especially during periods of water stress. This research underscores the utility of remote sensing and advanced machine learning methodologies in informing sustainable agricultural practices, particularly in regions vulnerable to the impacts of climate change.

While remote sensing methods coupled with machine learning are gaining traction for yield prediction, limitations exist, particularly regarding the integration of soil features. The lack of detailed soil information in some yield prediction models, particularly those relying solely on remote sensing data, can limit their accuracy and applicability, especially at local scales. This limitation is highlighted in a study where satellite-based crop yield mapping showed reduced sensitivity in detecting yield variations caused by soil properties (Deines et al., 2021).

Soil properties, such as texture, organic matter content, and available water capacity (AWC), significantly influence maize yield. A study in the western US Corn Belt found that the yield benefits of conservation tillage practices were more pronounced in fields with higher soil organic carbon levels (Cambron et al., 2024). Climate conditions is another factor influence maize yields by year-to-year variations and long-term trends (Jithitikulchai et al., 2018; Le, 2016). The key factors include temperature, precipitation, and solar radiation (Maltais‐Landry & Lobell, 2012).

Machine learning models have gained significant attention for their capacity to address the complexities of maize yield prediction by leveraging diverse datasets and uncovering non-linear relationships. Among these, random forests are frequently cited for the robustness and ability to model interannual climate variability, proving effective at both county and national scales (Ruiz et al., 2023). Support vector machines, known for handling high-dimensional data and complex interactions, have demonstrated success in predicting maize yields across various regions, including the US Corn Belt (Joshi et al., 2021). Some deep learning models, including long short-term memory (LSTM) networks and convolutional neural networks, hold promise for capturing intricate patterns (Sarzaeim & Muñoz-Arriola, 2024).

However, these models demonstrate notable limitations, particularly in the lack of integration of geospatial information and the inability to quantify prediction uncertainty. The absence of geospatial features may hinder the capacity of the models to account for spatial variability in yield predictions, while the lack of uncertainty estimation limits their utility in risk-sensitive decision-making processes. Bayesian modelling is one approach to quantifying uncertainty and has been widely applied in agricultural research, such as assessing the economic effects of varietal improvement and technological changes on flue-cured tobacco yield and quality (Ramsey & Rejesus, 2021). In addition, geographically weighted regression (GWR) is a spatial statistical technique that provides a more localised insight into how coefficients are influenced by specific variables (McCord et al., 2012).

In this study, data from the U.S. Corn Belt will be utilised to develop and refine the yield prediction models. GWR will be employed to integrate geospatial information with soil characteristic and climate condition, while the Variational Bayesian (VB) approach will be applied to analyse and quantify prediction uncertainty. This framework aims to provide a more comprehensive understanding of yield variability and enhance the reliability of the predictions.

This paper begins by presenting the academic background of maize yield prediction, outlining the importance and challenges associated with this area of research. It then introduces the mathematical methods employed in this study, providing a detailed

explanation of the techniques used. Following this, the paper describes the dataset utilised, including its key characteristics and relevance to the analysis. In the results section, the performance of four different models is evaluated, and the best-performing model is selected for an in-depth analysis, offering valuable insights into its strengths and implications.

4.2 Materials and Methods

This study focuses on the traditional Corn Belt region of the United States, encompassing 842 counties across ten key states where per-acre maize yields exceeded 90 bushels between 2014 and 2023. By selecting a ten-year timeframe characterised by relatively stable yield growth, the analysis minimises the confounding effects of technological advancements and efficiency improvements. This temporal scope ensures that observed yield variations are more directly attributable to the variables under investigation rather than external technological factors. The final dataset includes 6,945 samples, providing a robust foundation for statistical analysis and model development.

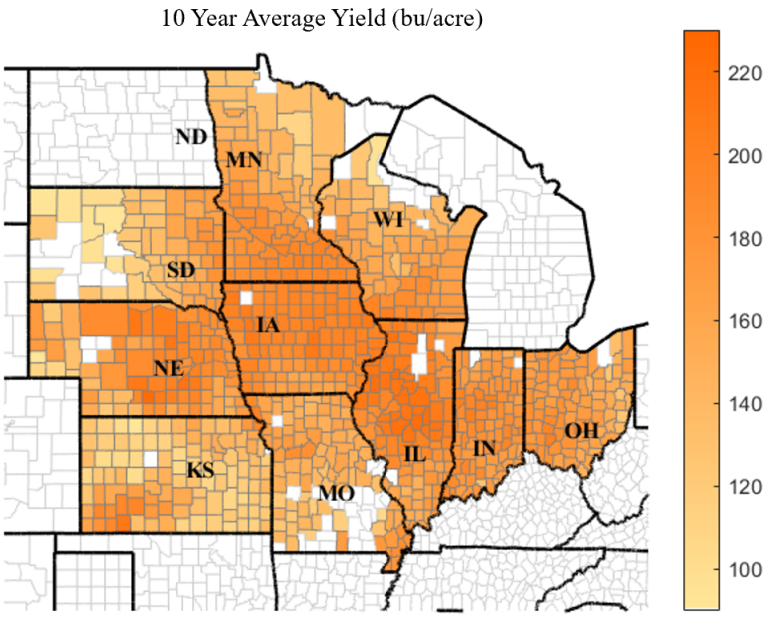


Figure 12 The average annual yield for selected counties from 2014 to 2023.

4.2.1 Data Collection and Preprocessing

The data utilised in this research are categorised into two primary groups: soil feature data and climate condition data. Additionally, remote sensing data, planting and harvesting timings are included as part of the input variables. This structured approach ensures that both environmental and management factors are accounted for in the modelling process, enhancing the robustness and reliability of the results.

All datasets were carefully aligned both temporally and spatially to ensure consistency and accuracy in the analysis. The climate data and NDVI values were averaged over the specific planting to harvest period for each county and year, capturing the environmental conditions during the maize growing season. Soil parameters, being relatively stable over time due to the low update frequency of the United States Department of Agriculture (USDA) soil surveys, were treated as constant values for each county throughout the ten-year period.

4.2.1.1 Soil Parameters

Soil data were obtained from the Soil Survey Geographic Database (SSURGO) through the USDA Web Soil Survey platform. SSURGO provides detailed soil geographic and attribute data for the United States at a high spatial resolution. Key soil property parameters for each county were collected, including available water capacity (AWC), organic matter content, pH, bulk density, electrical conductivity (EC), and cation exchange capacity (CEC).

AWC, organic matter content, and bulk density are critical for assessing soil properties related to air and water infiltration, nutrient retention, permeability, and erodibility (Acharjee et al., 2023; Carter, 2002). Soil pH, on the other hand, plays a vital role in nutrient availability and plant productivity, with the optimal pH range varying depending on crop requirements (Whetton et al., 2017). Both EC and CEC are closely tied to the ability of soil to facilitate nutrient uptake by plants, thereby influencing overall plant health and agricultural yields (Acharjee et al., 2023; El Bilali & Taleb, 2020).

For each county, mean values of these soil parameters were calculated to represent the overall soil characteristics.

Table 2 Selected input features

Category	Variable	Source	Unit	Level
Target	Yield	USDA	bu/acre	County
Climate	T2M	NASA	K	County

Climate	RH2M	NASA	%	County
Climate	Shortwave	NASA	J/m^2	County
Climate	Precipitation	NASA	m	County
Management	Planting Week	USDA	-	State
Management	Harvest Week	USDA	-	State
Remote sensing	NDVI	Landsat 8 Level 2	-	County
Soil	CEC	SSURGO	cmol (+)/kg	County
Soil	Organic Matter	SSURGO	%	County
Soil	pH	SSURGO	-	County
Soil	Bulk density	SSURGO	g/cm^3	County
Soil	AWC	SSURGO	cm/m	County
Soil	$CaCO_3$	SSURGO	%	County
Soil	Electrical conductivity	SSURGO	dS/m	County

4.2.1.2 Climate Data

Climatic variables, including temperature, precipitation, and solar radiation, further influence crop development (Jhajharia & Mathur, 2022; Shekoofa et al., 2014). Climate data were sourced from the National Aeronautics and Space Administration (NASA) via the prediction of worldwide energy resource project. This dataset provides meteorological parameters derived from satellite observations and assimilation models, suitable for agricultural applications. The variables included in this study are:

Temperature (T2M): Average air temperature at 2 meters above the surface, influencing photosynthesis and respiration rates.

Precipitation: Total corrected precipitation, impacting soil moisture and water availability for crops.

Relative Humidity (RH2M): Relative humidity at 2 meters above the surface, affecting evapotranspiration and plant water stress.

Shortwave: Downward shortwave radiation under all-sky conditions, critical for photosynthetic activity.

4.2.1.3 Remote Sensing Data

Although weekly remote sensing data is not applicable to new farmland study, the average remote sensing data for the entire region can be included as part of the input parameters. Remote sensing data were acquired using the Google Earth Engine platform to process imagery from the Landsat 8 satellite, operated by the United States Geological Survey. Landsat 8 provides high-resolution multispectral imagery suitable for detailed agricultural analysis (Roy et al., 2014). NDVI was calculated for the agricultural regions within each county. NDVI is a widely used indicator of vegetation health, providing indirect information about crop conditions and biomass accumulation (Eisfelder et al., 2023).

4.2.1.4 Exclusion of Fertilizer Data

Although some studies have shown that fertiliser use significantly impacts crop yield (Villacis et al., 2020), this study did not consider the effect of fertilisers on maize yield and assumed that all samples contained adequate levels of nitrogen (N), phosphorus (P), and potassium (K). This decision was made because fertiliser data were only available at the state level, lacking the spatial resolution required for county-level analysis. Incorporating state-level fertiliser data could introduce inaccuracies due to spatial mismatches and heterogeneity within states. Therefore, to maintain the accuracy and integrity of the county-level analysis, fertiliser usage was excluded from consideration.

4.2.1.5 Timing of planting and harvesting

The timing of planting and harvesting, as a management input, provides indirect insights into the climatic conditions of a given year, making it an important variable for predicting maize yield. Due to the structural limitations of USDA reports, certain variables were only available at the state level. This study utilised the specific weeks during which 50% of planting and harvesting were completed in each county to define the start and end of the growing season. These data facilitated a standardised approach to defining the growing season across states, ensuring temporal consistency throughout the data integration process.

As these timing variables primarily serve as supplementary inputs with relatively low expected importance, state-level data were included to enhance the richness of the informational foundation. By incorporating these specific weeks as input variables, the model was able to capture temporal variations in crop development, thereby supporting yield prediction efforts.

4.2.1.6 Geolocation Information

The geographic location data utilised in this study were sourced from the TIGER/Line dataset of U.S. Census Bureau. This dataset provides precise geospatial information represented as latitude and longitude coordinates.

These coordinates are used to calculate spatial distances between observations, which are essential for determining the bandwidth in the Geographically Weighted Regression (GWR) model. The bandwidth governs the spatial extent of localised regression analyses by defining which observations contribute to the weighting function.

It is important to note that these geographic coordinates are not directly included as input features in the regression model. Instead, they serve a critical role in establishing the spatial relationships and weight matrix necessary for the localised modelling process.

4.2.1.7 Yield Data

Yield data were obtained from the USDA quick stats database, providing county-level maize yield measurements expressed in bushels per acre. These data represent the target variable for the machine learning model.

4.2.2 Data Integration and Preprocessing

In this study, data were first aggregated into distinct county-level geographic units based on the coordinates. The temporal coverage of each unit was then assessed by counting the number of represented years. Units that contained data for only a single year were fully included in the training set. In contrast, for units spanning multiple years, one year was randomly selected for the test set, while the remaining years from these units were allocated to the training set. This approach ensures temporal independence between the testing and training sets while prioritising the inclusion of all regions with more than two years of data in the test set, thereby maximising the coverage across different counties. Subsequently, the combined training data were randomly split into the final training and validation sets. The training, validation, and test sets remain consistent across different models, ensuring comparability and fairness in performance evaluation.

4.2.3 Machine Learning Method Selection

While LSTM method is widely used for yield prediction due to the ability to capture temporal dependencies in time-series data, they are unsuitable for the objectives of this study—namely, the valuation of newly cultivated land in transactions.

This study uses traditional linear regression as a baseline model and employs Geographically Weighted Regression (GWR) to demonstrate performance improvements. Furthermore, it explores the capabilities of combining Variational Bayesian with Geographically Weighted Regression (VB-GWR) and Variational Bayesian Neural Network Geographically Weighted Regression (VBNN-GWR) in capturing and quantifying uncertainty. By evaluating intrinsic land attributes—such as soil properties, geographical location, and climatic conditions—these models provide robust yield prediction and land valuation while addressing predictive uncertainty, ensuring broader applicability across diverse regions.

4.2.3.1 Linear regression

Linear regression, as one of the most fundamental models in machine-learning, assumes a linear relationship between input features and the target variable. Compared to more

complex machine-learning models, its primary advantage lies in its simplicity and resistance to overfitting, making it a robust choice for well-behaved datasets.

4.2.3.2 Geographically Weighted Regression

GWR is a widely used spatial statistical method designed to capture spatial heterogeneity in relationships between dependent and independent variables. Unlike traditional regression models, which assume that parameter estimates are globally constant, GWR allows coefficients to vary across geographic locations, making it particularly effective for analysing phenomena with significant spatial variability. By incorporating a spatial weighting matrix, GWR estimates local regression coefficients for each observation, emphasizing the influence of nearby data points while down weighting more distant observations.

In agriculture, yield data often exhibit substantial variability across regions due to differences in soil properties, climate, and management practices. Traditional models may fail to account for this spatial heterogeneity, leading to biased or oversimplified predictions. GWR addresses this challenge by tailoring the regression model to each geographic location, offering more insights into local yield-driving factors. The GWR model can be expressed as:

$$y_i = \beta_0(u_i, v_i) + \sum_{k=1}^{K} \beta_k(u_i, v_i) x_{ik} + \epsilon_i \qquad (4.1)$$

where y_i is the dependent variable, x_{ik} represents the K independent variable, and $\beta_k(u_i, v_i)$ denotes the spatially varying coefficients at location (u_i, v_i). $\beta_0(u_i, v_i)$ interpreted as the fundamental contribution in the absence of any influence from the independent variables. $\boldsymbol{\epsilon_i}$ is the error term assumed to follow a normal distribution with constant variance.

A key component of GWR is the spatial weighting matrix W_{ij} which plays a crucial role in capturing the local relationships between variables across geographic space. The spatial weighting matrix W_i determines how much influence each observation j exerts on the estimation of parameters β_k at a specific location i.

The elements of W are derived based on the spatial proximity between observations, typically calculated using a kernel function. A common choice is the Gaussian kernel, where the weight W_{ij} between locations *i* and *j* is defined as:

$$W_{ij} = exp\left(-\frac{d_{ij}^2}{h^2}\right) \tag{4.2}$$

where d_{ij} is the distance between locations *i* and *j*, and h is the bandwidth parameter controlling the range of influence. Observations closer to the location of interest ***i*** are assigned larger weights, while those farther away contribute less.

In a GWR model, bandwidth h optimization is critical as it balances local and global model behaviour. Smaller bandwidths focus on localized features, capturing fine-scale spatial variations, while larger bandwidths lean towards global patterns, smoothing over regional differences. To ensure optimal performance, bandwidth can be optimised during model training using gradient descent. In this case, bandwidths can be treated as a learnable parameter, adjusted iteratively to minimise the loss function.

The spatial weighting matrix is incorporated into the local regression process through weighted least squares. At each location ***i***, the local regression coefficients $\hat{\beta}(u_i, v_i)$ are estimated as:

$$\hat{\beta}(u_i, v_i) = (X^T W_i X)^{-1} X^T W_i y \tag{4.3}$$

where X represents the matrix of independent variables, y is the dependent variable vector, and W_i is the spatial weighting matrix cantered at location *i*.

By enabling parameter estimates to vary geographically, the spatial weighting matrix allows GWR to address spatial heterogeneity effectively. This characteristic makes GWR particularly valuable for studying phenomena where relationships are expected to change across space, such as environmental processes, urban development, and regional economics.

4.2.3.3 The Variational Bayesian Geographically Weighted Regression (VB-GWR)

However, despite its advantages, GWR and many traditional models rarely incorporate uncertainty into their framework. In contexts such as agricultural yield prediction, where data variability is inherent, the ability to account for uncertainty becomes critical. Incorporating uncertainty not only enhances the robustness of the model and reduces the risk of overfitting but also provides a foundation for more informed decision-making. For example, in the context of maize yield, prediction intervals derived from uncertainty-aware models allow investors to better evaluate risks and devise precise trading strategies.

Recent advancements have sought to integrate Bayesian frameworks with GWR to further address these challenges. Their approach effectively quantified uncertainty in parameter estimates while maintaining the ability to model spatial heterogeneity, demonstrating the potential for improved robustness and interpretability in spatial analysis.

To address this limitation, a VB-based estimation method is introduced, imposing probabilistic distributions over regression parameters and approximate their posterior through variational inference. Specifically, the model employs a trainable prior distribution θ_{prior} and a posterior approximation $\theta_{\text{posterior}}$, defined as:

$$\theta_{prior} \sim \mathcal{N}(\mu_{prior}, \sigma_{prior}) \tag{4.4}$$

$$\theta_{posterior} \sim \mathcal{N}(\mu_{posterior}, \sigma_{posterior}) \tag{4.5}$$

Where μ is mean value and σ is standard deviation, these values can be optimised with ELBO during training.

In VB-GWR, the bandwidth h optimization approach remains consistent with traditional GWR, where the bandwidth is treated as a trainable, deterministic parameter rather than being integrated into the uncertainty model. This design choice stems from the role of bandwidth as a parameter that determines the spatial weighting range in the geographic weighting process. Unlike probabilistic parameters in the uncertainty model, the bandwidth produces deterministic geographic weights that reflect the spatial influence of training points on query points.

By keeping the bandwidth *h* independent of the uncertainty framework, the model avoids unnecessary complexity while preserving the clear and intuitive physical interpretation of the bandwidth. The optimized bandwidth value directly corresponds to the effective range of geographical weighting, ensuring that spatial relationships remain interpretable and consistent with the underlying geographic structure.

In variational Bayesian models, the optimization objective is to maximize the approximation of the posterior distribution $\theta_{\text{posterior}}$. This process is achieved by minimizing the negative Evidence Lower Bound (ELBO).

The VB-GWR model was constructed as follows:

- Input Layer**:** The input shape matched the feature dimension of the training data, along with an additional input for geographic coordinates to account for spatial information.

- Geographic Weighting Layer: A custom geographic weighting layer was implemented to compute spatial weights based on the distances between the geographic coordinates of query points and training points. The weights were parameterised using a trainable bandwidth, which controlled the spatial range of influence. The weighted inputs were computed by aggregating the training features using these spatial weights.

- Dense Variational Layer: A variational dense layer was employed to learn the posterior distributions of the regression weights and biases. The KL divergence between the posterior and prior distributions was scaled by a weight w_{KL} , set as the inverse of the training sample size, to achieve proper regularisation.

- Output Distribution**:** The final layer parameterised the output as a Gaussian distribution. The mean and scale of the Gaussian were learned separately, with the scale constrained to be positive using a Softplus transformation. This output distribution allowed the model to capture both the predicted value and the associated uncertainty.

4.2.3.4 The Variational Bayesian Neural Network Geographically Weighted Regression (VBNN-GWR)

VBNN-GWR extends the VB-GWR framework by incorporating hidden layers to increase the model capacity for capturing non-linear relationships in data. While retaining the fundamental principles of VB-GWR, including uncertainty quantification and spatial heterogeneity modelling, the addition of hidden layers tests potential underfitting that may arise from the simpler linear structure of VB-GWR. This enhancement evaluates whether the added complexity improves predictive performance while preserving the benefits of the variational Bayesian approach.

4.2.4 Loss Function:

4.2.4.1 Mean Squared Error (MSE)

MSE is used for the linear regression training, a loss function MSE is minimized:

$$\mathcal{L}_{Linear} = \frac{1}{N}\sum(y_i - \hat{y}_i)^2 \tag{4.6}$$

$$\mathcal{L}_{GWR} = \frac{1}{N}\sum_i W_{ij}(h) \cdot (y_i - \hat{y}_i)^2 \tag{4.7}$$

where y_i is the true value, $\hat{y}_i$ is the predicted value, and *N* is the number of samples. This approach effectively learns the average pattern in data.

4.2.4.2 Negative Log-Likelihood (NLL)

The NLL is defined as the negative logarithm of the probability of the observed data under the predicted distribution. It is not only a component of the VB loss function but also serves as one of the evaluation metrics for comparing different models. For a predicted output $\hat{y}$ with corresponding standard deviation σ, the NLL is defined as:

$$NLL = -\log p\,(\,y \mid \hat{y}, \sigma\,) = 0.5 \times \left[\frac{(y-\hat{y})^2}{\sigma^2} + log(2\pi\,\sigma^2)\right] \tag{4.8}$$

This function penalises large errors while ensuring that the predicted uncertainty σ remains balanced.

4.2.4.3 Evidence Lower Bound (ELBO)

The negative $ELBO$ is employed as the primary loss function of VB to optimise the probabilistic predictions of the model. This loss function is particularly suitable for models that output probability distributions, allowing both point predictions and uncertainty quantification with the optimization target typically expressed as:

$$\mathcal{L}_{VB} = -ELBO = NLL + \lambda \cdot KL\left[\theta_{posterior} \mid \theta_{prior}\right] \tag{4.9}$$

Where NLL is the loss function to optimise the probabilistic predictions of the model, λ is a hyperparameter that controls the weight of the KL. $KL\left[\theta_{posterior} \mid \theta_{prior}\right]$ measures the difference between the posterior distribution $\theta_{posterior}$ and the prior distribution θ_{prior}. Simultaneously, the trainable prior is updated during this step, with its parameters optimised through the same process.

4.2.5 Sensitivity Analysis

For sensitivity analysis on the mean prediction, the parameter weights from the GWR model are leveraged. As GWR belongs to the class of linear models, its parameter weights can directly indicate the importance of each input feature. By analysing these weights across different geographical locations, the spatially varying significance of input features on the predicted mean can be revealed.

To evaluate the impact of input features on prediction uncertainty, the method analyses how the prediction standard deviation changes with varying feature values. The approach involves dividing the standardized input values of each feature into 10 bins and calculating the mean prediction uncertainty within each bin using multiple sampling iterations. This process identifies the relationship between each value range and the resulting uncertainty, enabling the visualization of uncertainty trends and highlighting the features that contribute the most to variability in the predictions.

4.2.6 Model Hyperparameter Tuning

Given that GWR is based on a linear modelling framework, the number of tuneable hyperparameters is limited. In contrast, VB-GWR and VBNN-GWR provide more flexibility with additional hyperparameters, including the KL weight $\boldsymbol{\lambda}$ and the learning

rate. To optimise these hyperparameters, a grid search approach was employed, comparing three different KL weight configurations: $\frac{0.1}{N_{samples}}$, $\frac{1}{N_{samples}}$, $\frac{10}{N_{samples}}$ where $N_{samples}$ represents the number of input samples. Additionally, learning rates of 0.1, 0.01, and 0.001 were tested to evaluate their impact on model performance. For VBNN-GWR, the model has tested one and two hidden layers with 2, 10, 20 neurons. This comparative analysis aimed to identify the optimal combination of hyperparameters for both VB-GWR and VBNN-GWR to ensure robust predictive performance.

4.3 Model Results

This study built four models with data from 842 counties in the U.S. Corn Belt over the period from 2014 to 2023.The predictive performance of linear regression and GWR is shown as Figure 13. Linear regression demonstrates significant underfitting, as evidenced by its inability to adequately capture the relationship between the input features and the target variable. This limitation is reflected in its COD of only 0.31 on both the training and test sets, suggesting poor predictive performance and a failure to generalise. In comparison, GWR shows improved predictive accuracy with a test COD of 0.62, indicating a better ability to account for spatial heterogeneity in the data. However, the distribution of predictions in Figure 13 (b) suggests clear signs of overfitting, as the model appears to overly conform to the training data patterns, limiting its robustness when applied to unseen data.

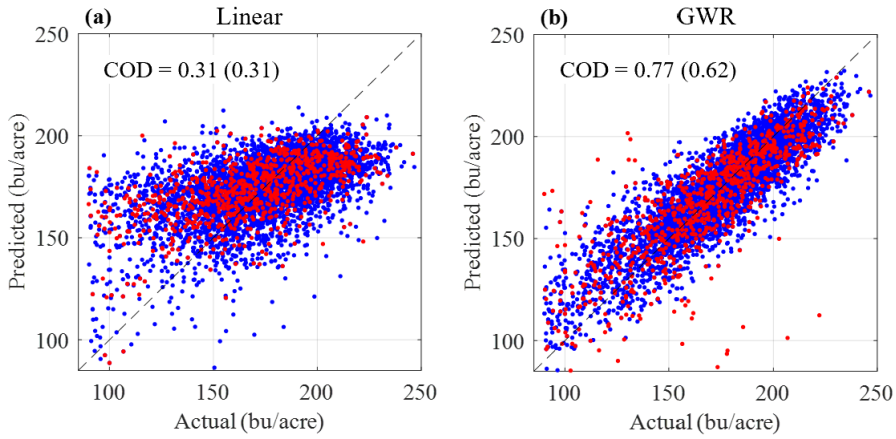


Figure 13 Performance of linear regression and GWR model

The performance of VB models is assessed using multiple metrics to ensure its reliability and effectiveness. VB-GWR demonstrates robust performance, with COD values remaining consistent between the training set (0.59) and the testing set (0.57). The NLL also shows minimal variation, increasing slightly from 0.88 for training to 0.98 for testing in Figure 14 (a). Furthermore, the 95% credible interval (CrI) on the testing set covers 94.4% of the data, which is very close to the ideal coverage of 95%. These results indicate that VB-GWR achieves a balance between predictive accuracy and uncertainty quantification, making it a reliable model.

In contrast, VBNN-GWR reveals clear issues in its testing performance in Figure 14 (b). The testing COD drops significantly to 0.29, and the NLL rises from 0.95 for training to 1.33 for testing. Additionally, the 95% CrI only covers 91.35% of the data, indicating suboptimal uncertainty calibration. Even with experiments involving both complex and extremely simple architectures, such as a single hidden layer with a small number of neurons, VBNN-GWR fails to achieve reliable performance. These results suggest that the introduction of hidden layers has led to instability and a lack of generalisability, likely due to overfitting or insufficient regularisation. This highlights the challenges of extending VB-GWR with neural network components without further tuning and optimisation.

In terms of handling outliers or dispersed points, VB-GWR demonstrates greater robustness, with only a single clearly dispersed point observed in its predictions (circled in Figure 3a). In contrast, VBNN-GWR shows a larger number of such dispersed points.

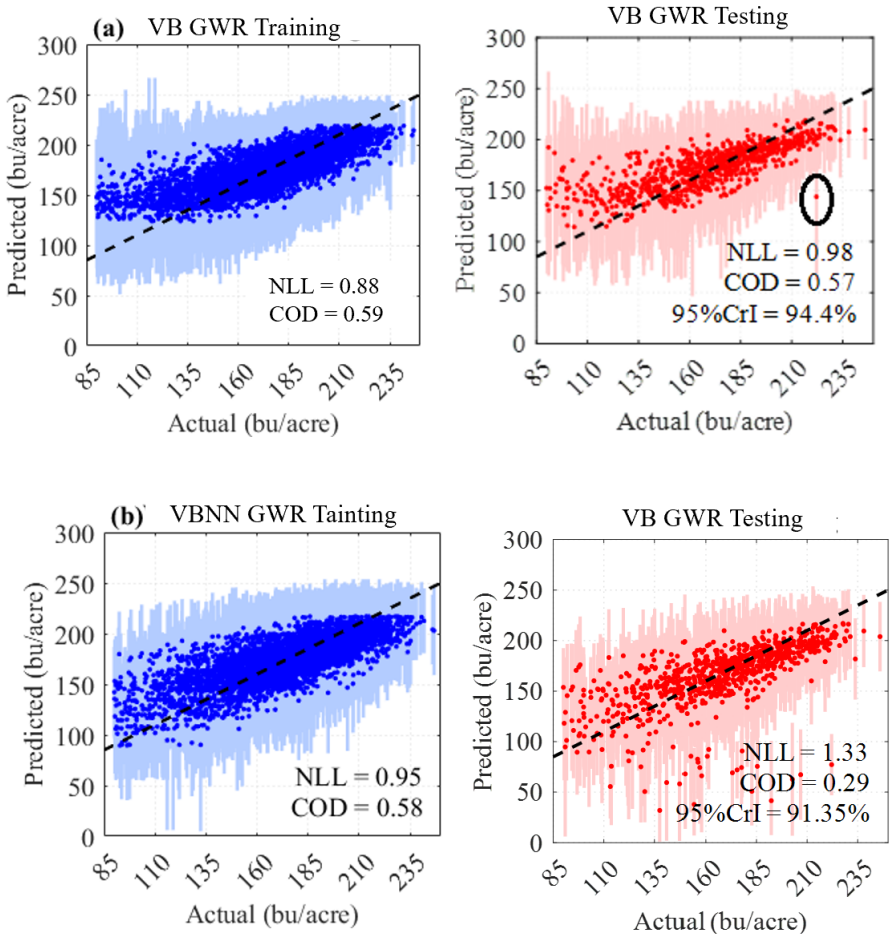


Figure 14 Performance of variational Bayesian models. ((a), (b) represent the results of VB-GWR and VBNN-GWR; vertical lines represent 95% credible interval)

The density distributions of the predicted $\boldsymbol{\sigma}$ of VB-GWR and VBNN-GWR are shown in the figure. For VB-GWR, the predicted $\boldsymbol{\sigma}$ exhibits a broader spread in both training and testing datasets, reflecting a more conservative uncertainty estimation that captures a wider range of variability.

In contrast, VBNN-GWR demonstrates a narrower distribution of $\boldsymbol{\sigma}$, indicating a more concentrated and consistent uncertainty estimation. However, despite this narrower spread, 95%CrI coverage of VBNN-GWR is notably lower than VB-GWR, highlighting an issue of overconfidence. This overconfidence suggests that VBNN-GWR underestimates the true uncertainty in its predictions, which may contribute to its poorer generalisation performance compared to VB-GWR.

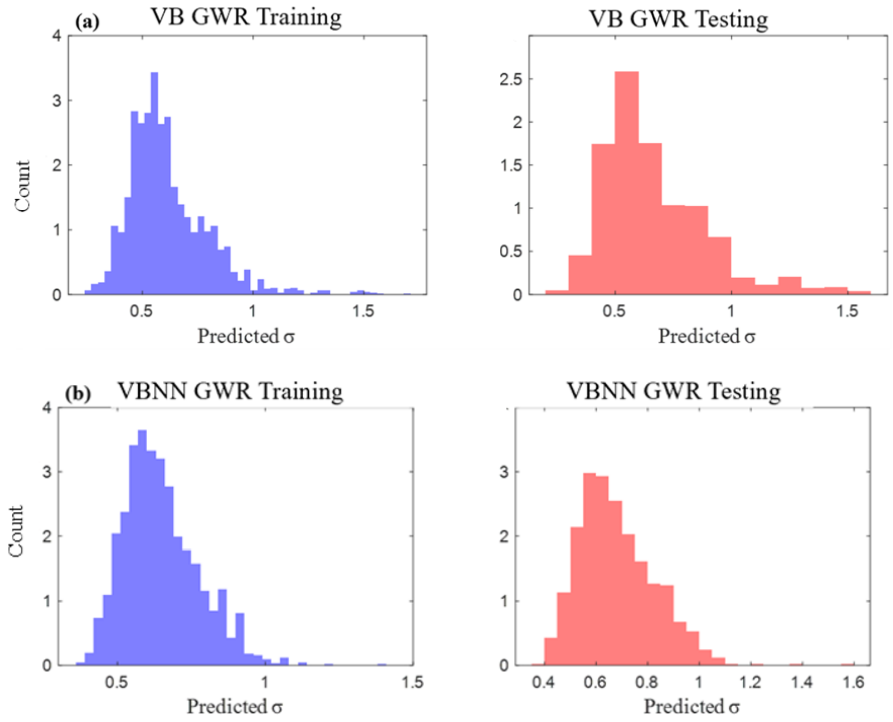


Figure 15 Probability density distribution of standard deviations

The linear regression and GWR, can be used for probabilistic predictions by estimating the scale of prediction error between predictions and true values on training data ($\boldsymbol{\sigma}$ as shown in the left column of Figure 16). With the predicted mean values and the estimated $\boldsymbol{\sigma}$ as standard deviations, these models generate probabilistic predictions that can be applied to unseen data—the testing set (right column of Figure 16).

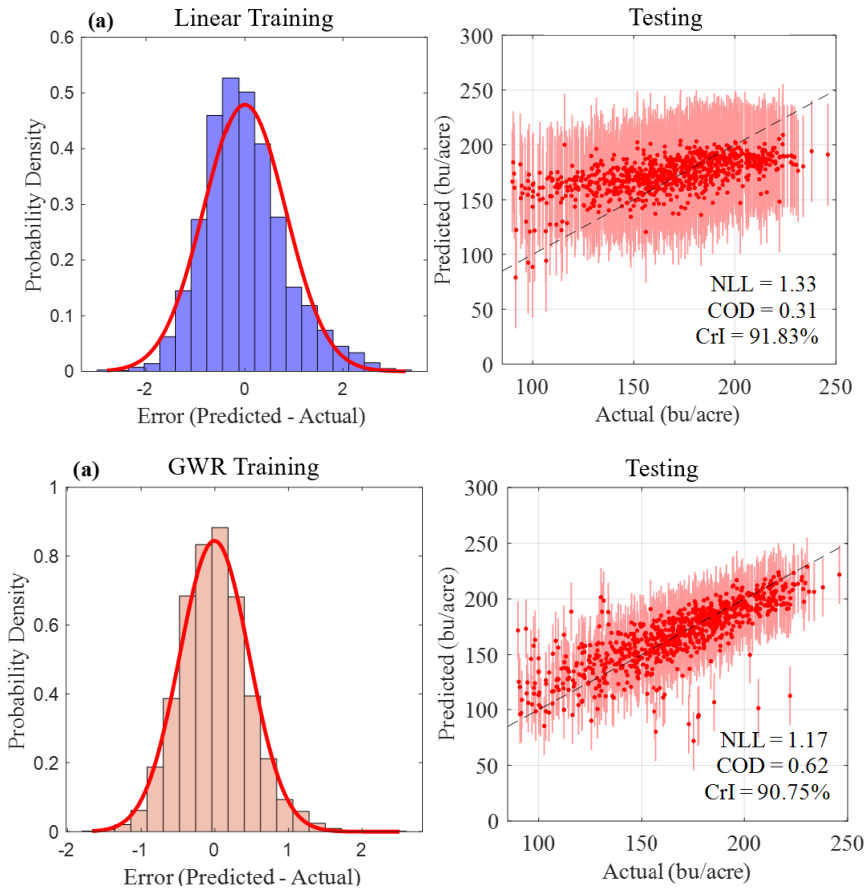


Figure 16 Performance of uncertainty quantification of linear regression and GWR model

The evaluation of these probabilistic models from linear regression and GWR reveals key insights into their performance. The probability density distributions of prediction errors on the training set show that linear regression and GWR achieve standard deviations of 0.83 and 0.47, respectively. The smaller standard deviation of GWR highlights its ability to better reduce error dispersion compared to linear regression, reflecting its improved capacity to capture spatial heterogeneity.

The predicted yield values versus actual values, visualised with error bars representing predictive uncertainty, highlight differences in the ability of these models to quantify uncertainty. The test set NLL values are 1.33 for linear regression and 1.17 for GWR. The improved NLL in GWR suggests a more accurate and better-calibrated uncertainty quantification compared to linear regression. However, while linear regression achieves a 95% CrI data coverage rate of 91.83%, GWR achieves a slightly lower CrI coverage

rate of 90.75%, indicating a modest underestimation of uncertainty despite its stronger predictive power.

By comparison, among the four models, GWR demonstrates the best performance in terms of precise prediction values, while mean prediction of VB-GWR is slightly inferior to GWR. However, VB-GWR significantly outperforms the other three models in uncertainty quantification. VB-GWR not only exhibits greater stability in the distribution of prediction errors but also achieves superior credible interval coverage and balanced NLL values. This indicates that, although VB-GWR is slightly less accurate in mean predictions, it achieves an excellent balance between predictive accuracy and uncertainty quantification, making it highly promising for practical applications.

4.4 Result Analysis

Through the comparison of results obtained from four models, it was observed that the VB-GWR model provided the best overall performance. Therefore, the subsequent analysis is based on the results from the VB-GWR model, which showed the most accurate predictions and better adaptability to the dataset.

4.4.1 Prediction Accuracy

The VB-GWR model prediction accuracy was assessed by analysing the differences between actual maize yields and predicted values across various yield ranges, as illustrated in Figure 17. The model performed best within the 140–220 bu/acre range, where prediction differences were minimal, indicating high accuracy and consistency. Outside this range, however, the model exhibited larger biases: in the lower yield range (<140 bu/acre), predictions were significantly overestimated, while in the higher yield range (>220 bu/acre), predictions tended to be underestimated. These observations indicate that the 140–220 bu/acre range corresponds to typical yields under normal agricultural conditions, where the model is most reliable.

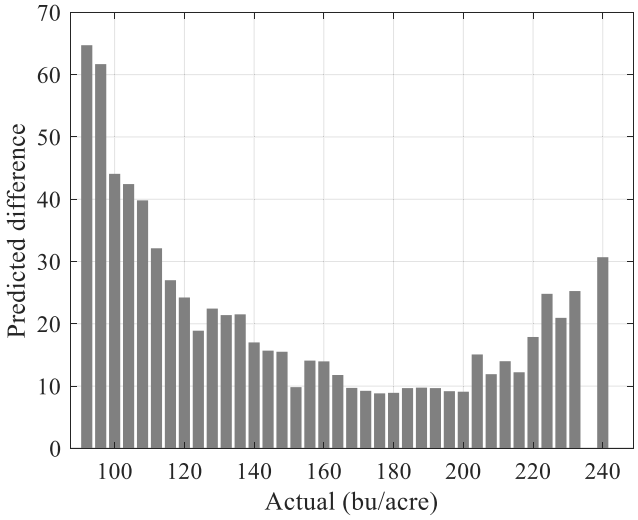


Figure 17 Prediction error based on yield

4.4.2 Feature Importance Analysis

This study utilised the local regression coefficients from the VB-GWR model as indicators of feature importance. By calculating the mean absolute importance of each input feature, the contributions to the model predictions were quantified in Figure 18. The results reveal distinct levels of importance among the features, providing insights into relative influence on maize yield predictions.

Among all features, pH emerged as the most influential, demonstrating the highest mean absolute importance. This highlights the critical role of soil pH in determining maize yield, likely due to its impact on nutrient availability and root development. Following pH, a group of features including T2M, bulk density, RH2M, shortwave radiation, $CaCO_3$, and CEC showed consistently high importance. These features collectively represent key climatic and soil properties affecting crop performance.

Features such as AWC and electrical conductivity exhibited moderate importance. While these factors are relevant for water retention and salinity, the influence may vary geographically depending on local soil conditions and irrigation practices. Conversely, features like organic matter and precipitation showed relatively low importance. The limited contribution of organic matter may stem from the absence of soil fertiliser data in this study, which might have restricted the ability to capture its full impact on soil

productivity. Similarly, the low importance of precipitation could reflect consistent rainfall patterns across the study region or the prevalence of irrigation systems especially in NE, which mitigate reliance on natural rainfall.

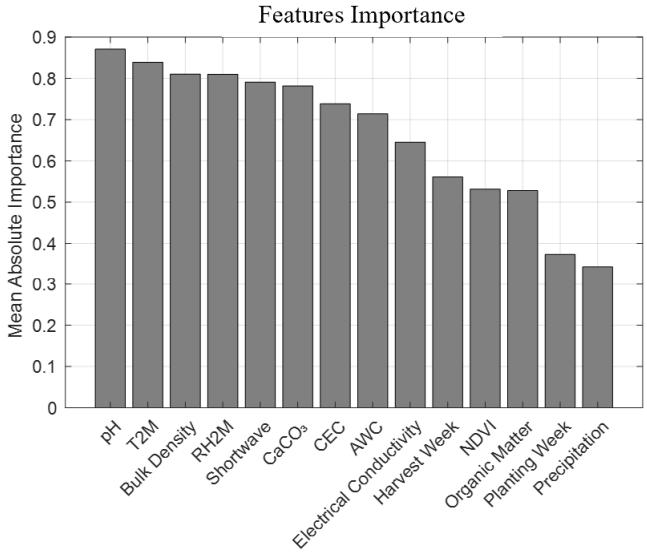


Figure 18 Feature importance rank

4.4.3 Uncertainty Analysis

The uncertainty analysis evaluates the variability in model predictions at the county level for the VB-GWR testing set. Figure 19 illustrates the standard deviation of the standardized prediction results, which is used as a quantitative measure of uncertainty.

From the map, it can be observed that the uncertainty varies across different counties in the U.S. Corn Belt. Certain regions, particularly those with darker blue shades, exhibit higher prediction uncertainties. Conversely, regions with lighter shades demonstrate lower uncertainties, indicating that the model performs more consistently in these areas. From the map, regions such as MN and ND exhibit relatively higher uncertainty compared to IA and IL, which show more consistent predictions. States like KS and MO display moderate levels of uncertainty. This regional variation highlights differences in the prediction consistency across the Corn Belt. The higher uncertainty in arid regions is closely related to maize acreage and yield data in these areas. For instance, SD and KS exhibit greater uncertainty, whereas NE, despite also being an arid region, benefits from

its well-developed irrigation system, resulting in significantly lower uncertainty compared to the other two states.

By comparing the map with the annual average maize yield distribution in Figure 12, it can be observed that regions with higher average yields tend to have lower uncertainty, while areas with lower yields often correspond to higher uncertainty.

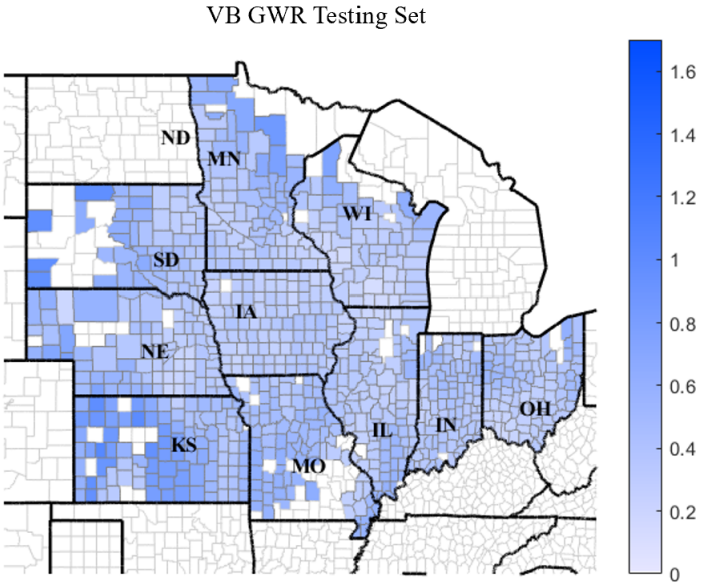
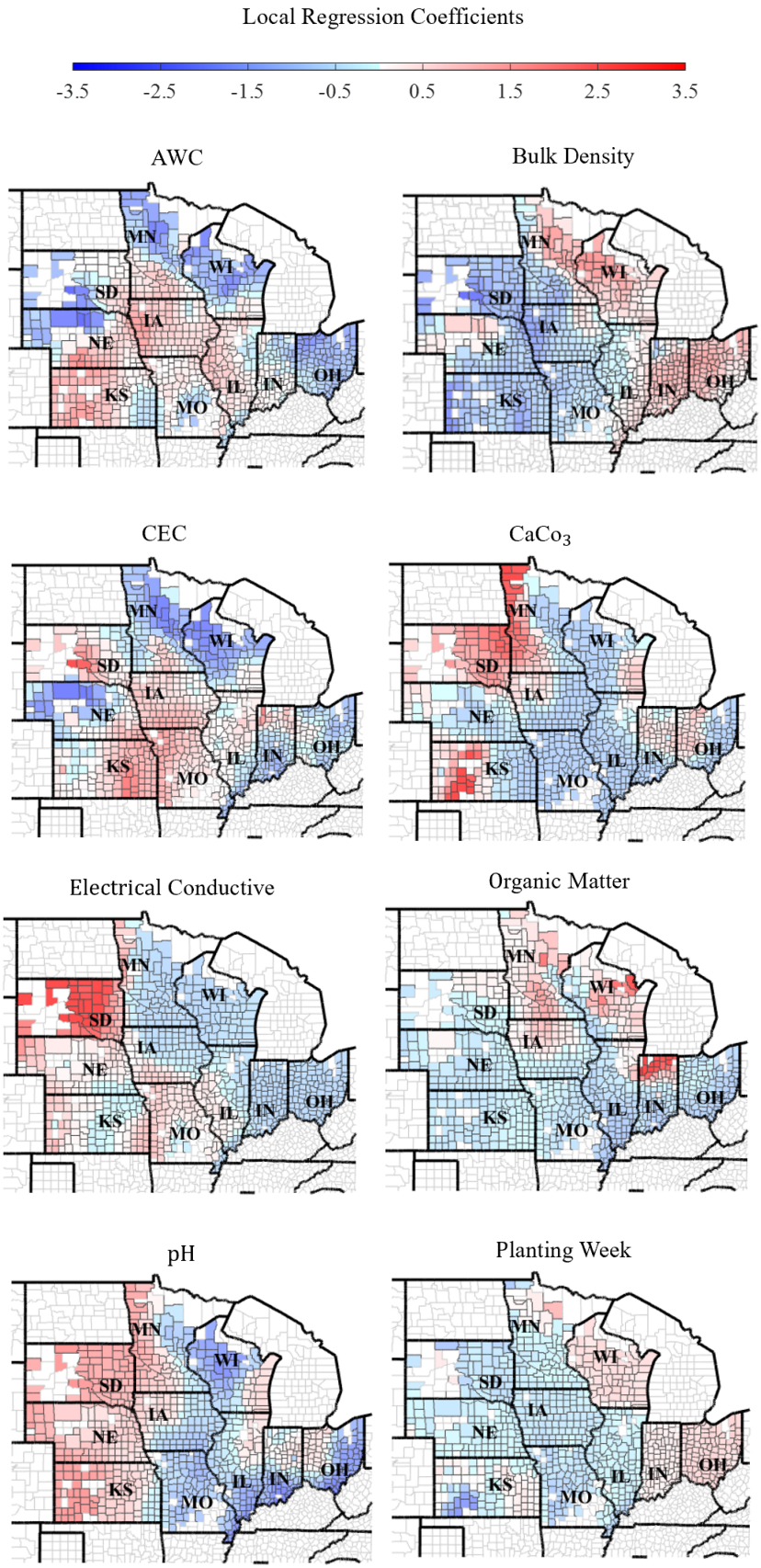


Figure 19 Uncertainty distribution of counties in Corn Belt

4.4.4 Uncertainty sensitive Analysis

The analysis of uncertainty utilised local regression coefficients from different regions as a reference in Figure 20. These coefficients provide a detailed understanding of spatial variability in model predictions, offering insights into how input features influence predictive uncertainty across diverse geographical locations. By focusing on the regional heterogeneity captured through local regression, the analysis highlights the varying significance of input features and their associated uncertainties in different areas.


Local Regression Coefficients
-3.5
-2.5
-1.5
-0.5
0.5
1.5
2.5
3.5
AWC
Bulk Density
CEC
CaCo3
Electrical Conductive
Organic Matter
pH
Planting Week
MN
WI
SD
IA
NE
KS
MO
IL
IN
OH

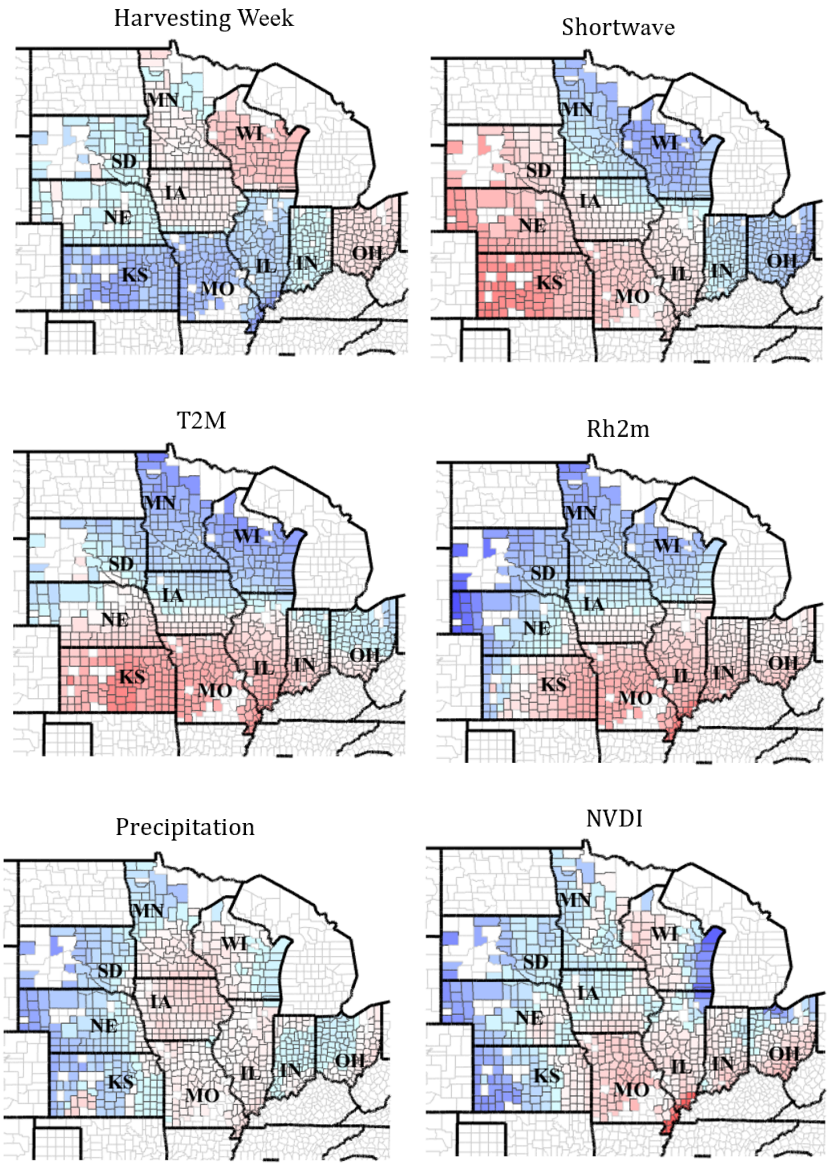


Figure 20 Local regression coefficient of every feature in Corn Belt

4.4.4.1 Regions with Low Uncertainty

Regions such as IA, IL, and parts of IN exhibit consistently low uncertainty in the testing set, as indicated by the lighter shades on the uncertainty map. These regions also display balanced and stable local regression coefficients for key features such as AWC, pH, and CEC. The even contribution of these inputs to the model likely leads to more accurate predictions and reduced uncertainty. Additionally, the balanced importance of features reflects optimal growing conditions, where no single factor dominates or limits yield, resulting in reliable model performance.

4.4.4.2 Regions with High Uncertainty

Northern regions like MN and ND, as well as western regions like SD and NE, show higher uncertainty, represented by darker shades on the map. In these regions, the local regression coefficients for critical features such as AWC, pH, and CEC exhibit higher variability, with some areas even showing negative coefficients. This inconsistency suggests that the relationships between input features and yield are less stable, contributing to higher prediction uncertainty. For instance, in MN, strong positive coefficients for T2M and shortwave radiation contrast with weaker or negative coefficients for precipitation and RH2M, indicating complex interactions that challenge the model.

4.4.4.3 Role of Specific Features – AWC

Regions such as IA and IL, where AWC has consistently positive coefficients, correspond to areas with low uncertainty. This highlights the importance of water availability in driving stable yield predictions. Conversely, in regions like KS and SD, where AWC coefficients vary or become negative, prediction uncertainty increases. This suggests that inconsistent water availability or irrigation practices in these regions make yield predictions more challenging for the model.

4.4.4.4 Role of Specific Features – pH

In areas such as IL and IN, where pH coefficients are balanced and positive, low uncertainty is observed. However, in KS and SD, pH coefficients vary widely, with some areas showing negative values. This variability aligns with higher uncertainty, indicating that extreme or inconsistent soil pH conditions are more difficult for the model to predict accurately. Stable pH values in central regions contribute to higher prediction reliability.

4.4.4.5 Role of Specific Features – T2M and Shortwave Radiation

Regions like MN, where T2M and shortwave radiation have high positive coefficients, show moderate to high uncertainty. This reflects the importance of these features in northern regions with shorter growing seasons, but also suggests that variability introduces challenges for accurate predictions. The reliance on climatic factors like temperature and radiation in these areas highlights the need for adaptive models that can handle such variability.

4.4.4.6 Impact of Feature Interactions

Regions with high uncertainty, such as KS and SD, tend to exhibit stronger or more variable interactions between features like precipitation, RH2M, and soil properties. These interactions complicate yield predictions as environmental conditions fluctuate, or one factor, such as salinity or compaction, disproportionately affects yields. In contrast, regions like IA and IL show smoother and more uniform interactions between features, resulting in stable environmental conditions and reduced prediction uncertainty.

4.4.4.7 General Observations

Areas with consistent local regression coefficients across key features, such as IA and IL, align with regions of low uncertainty, reflecting more reliable model predictions. In contrast, regions with more variable or extreme coefficients for features such as AWC, pH, and T2M correspond to areas of higher uncertainty. This highlights the importance of consistent environmental and soil conditions in improving prediction accuracy.

4.4.5 Feature Uncertainty Quantification

The relationship between input feature values and corresponding prediction uncertainty is shown as Figure 21, measured as the mean σ within bins of feature values. These figures provide insights into how different input ranges contribute to the variability in predictions. However, edge values in the bins may show slight deviations due to fewer data points within those bins, which could impact the robustness of the results. Planting and harvest time are not included in this analysis as it is state-level data.

The relationship between pH and prediction uncertainty indicates that the model performs best when pH values are between 5.5 and 7.0, where uncertainty is lowest. Beyond a pH of 7.2, uncertainty increases sharply, reflecting the challenges of modelling alkaline soils or the limited representation of high-pH environments in the dataset.

For bulk density, uncertainty decreases steadily as density increases from very low values, reaching the lowest point around 1.5 g/cm³. However, uncertainty is higher for values below 1.4 g/cm³, reflecting the difficulty in predicting yields for very loose soils, likely due to limited representation in the dataset or increased variability in these conditions.

The uncertainty associated with organic matter is remarkably low, with all mean σ values remaining below 0.8 across the range, suggesting that the model performs consistently well within typical organic matter ranges. The low overall uncertainty implies that organic matter contributes minimally to prediction variability, and its impact on model uncertainty is negligible, even at extreme levels. The low overall uncertainty implies that organic matter contributes minimally to prediction variability, and its impact on model uncertainty is lower than the average uncertainty of model, even at extreme levels.

Uncertainty for electrical conductivity increases with electrical conductivity growth, highlighting the difficulty in predicting yields in high-salinity soils. This could be attributed to fewer samples in saline regions or the complex interaction of salinity with other soil properties, leading to greater variability in predictions.

Compared to other input features, AWC demonstrates the lowest overall uncertainty across its range, as indicated by the relatively lower values on the y-axis. While there is a slight increase in uncertainty for very low values below 0.14 cm^3/cm^3 and a minor fluctuation above 0.17 cm^3/cm^3, these variations are modest. The current range of AWC values does not significantly contribute to model uncertainty, suggesting that this feature has a stable and consistent influence on predictions. The performance for AWC is reliable, and its variations are less likely to affect prediction robustness compared to other inputs.

For CEC, uncertainty is lowest when values are between 10 and 25 cmol (+)/kg, reflecting consistent model predictions in soils with moderate nutrient retention. However, uncertainty rises sharply for values exceeding 28 cmol (+)/kg, indicating challenges in predicting yield outcomes in regions with extremely high CEC, which are less common in the dataset.

The model exhibits stable uncertainty for $CaCO_3$ values up to 11%, beyond which uncertainty rises sharply. This suggests difficulties in predicting yields in soils with high calcium carbonate content, likely due to limited representation of such conditions in the dataset.

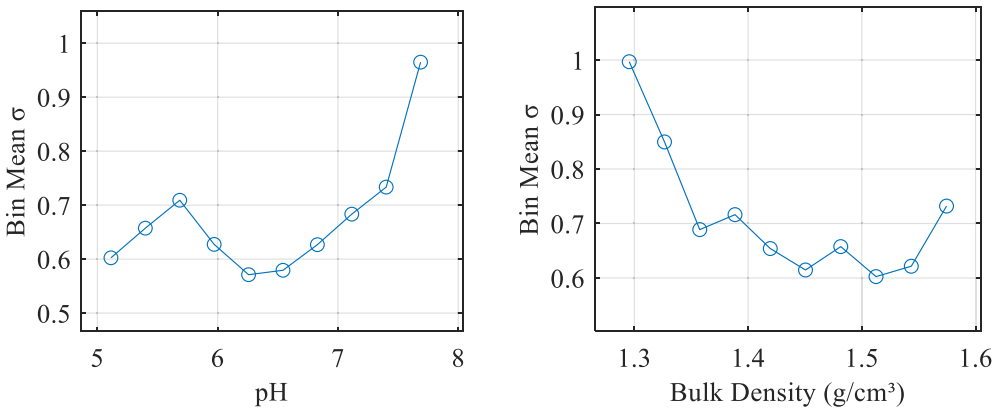
Uncertainty is lowest for T2M values between 288 K and 294 K, reflecting consistent model predictions under moderate temperature conditions. Higher uncertainty is observed for temperatures above 296 K, likely because of heat stress on crop growth.

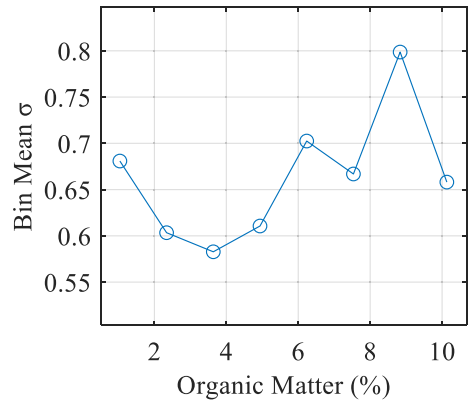
Prediction uncertainty decreases with increasing precipitation, reaching its lowest point between 0.6 and 1.0 m. This suggests the model performs well in regions with adequate rainfall, while higher uncertainty in low-precipitation areas indicates challenges in predicting yields under water-limited conditions.

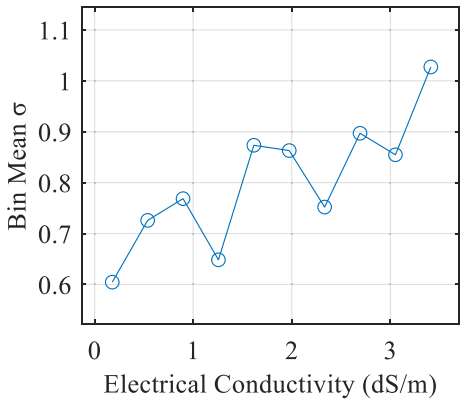
For shortwave radiation, the best range is observed between 1.6×10^7 J/m² and 1.9×10^7 J/m², corresponding to moderate sunlight levels. Higher radiation levels are associated with increased uncertainty, due to heat stress effects or the rarity of extreme sunlight

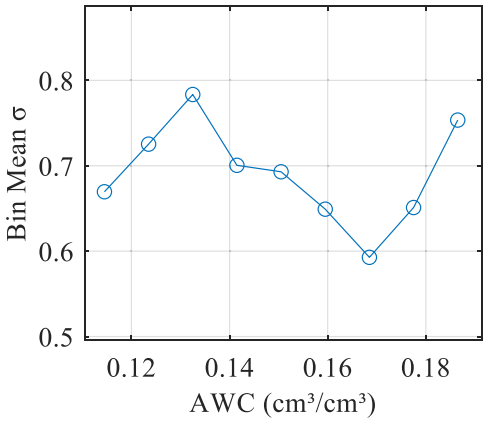
Uncertainty is lowest at moderate RH2M levels around 28.5%, where the model performs consistently. The dataset shows a relatively narrow range of RH2M values, which contributes to stable predictions in the central range. However, when RH2M falls below

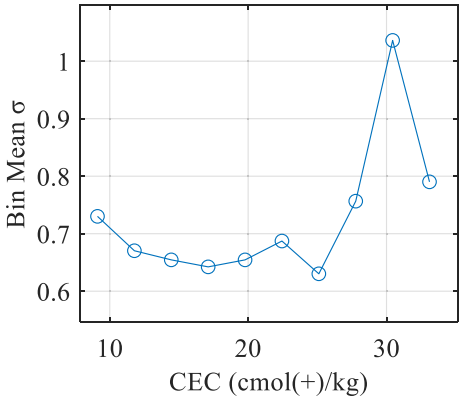
For NDVI, uncertainty is lowest between 0.4 and 0.6, typically corresponding to healthy vegetation conditions. Higher uncertainty at lower NDVI values suggests difficulties in predicting yields in areas with sparse vegetation or poor crop health.

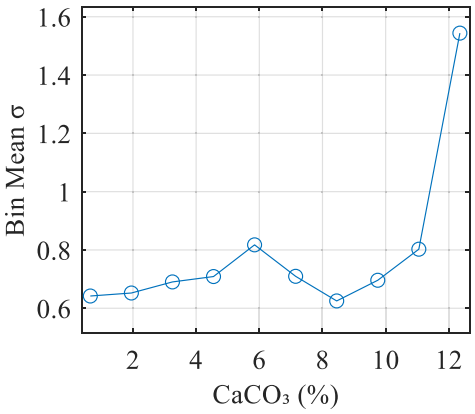


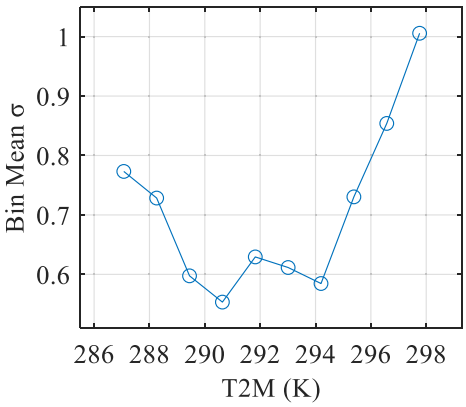

Bin Mean σ
0.8
0.75
0.7
0.65
0.6
0.55
2
4
6
8
10
Organic Matter (%)


Bin Mean σ
1.1
1
0.9
0.8
0.7
0.6
0
1
2
3
Electrical Conductivity (dS/m)


Bin Mean σ
0.8
0.7
0.6
0.5
0.12
0.14
0.16
0.18
AWC (cm³/cm³)


Bin Mean σ
1
0.9
0.8
0.7
0.6
10
20
30
CEC (cmol(+)/kg)


Bin Mean σ
1.6
1.4
1.2
1
0.8
0.6
2
4
6
8
10
12
CaCO₃ (%)


Bin Mean σ
1
0.9
0.8
0.7
0.6
286
288
290
292
294
296
298
T2M (K)

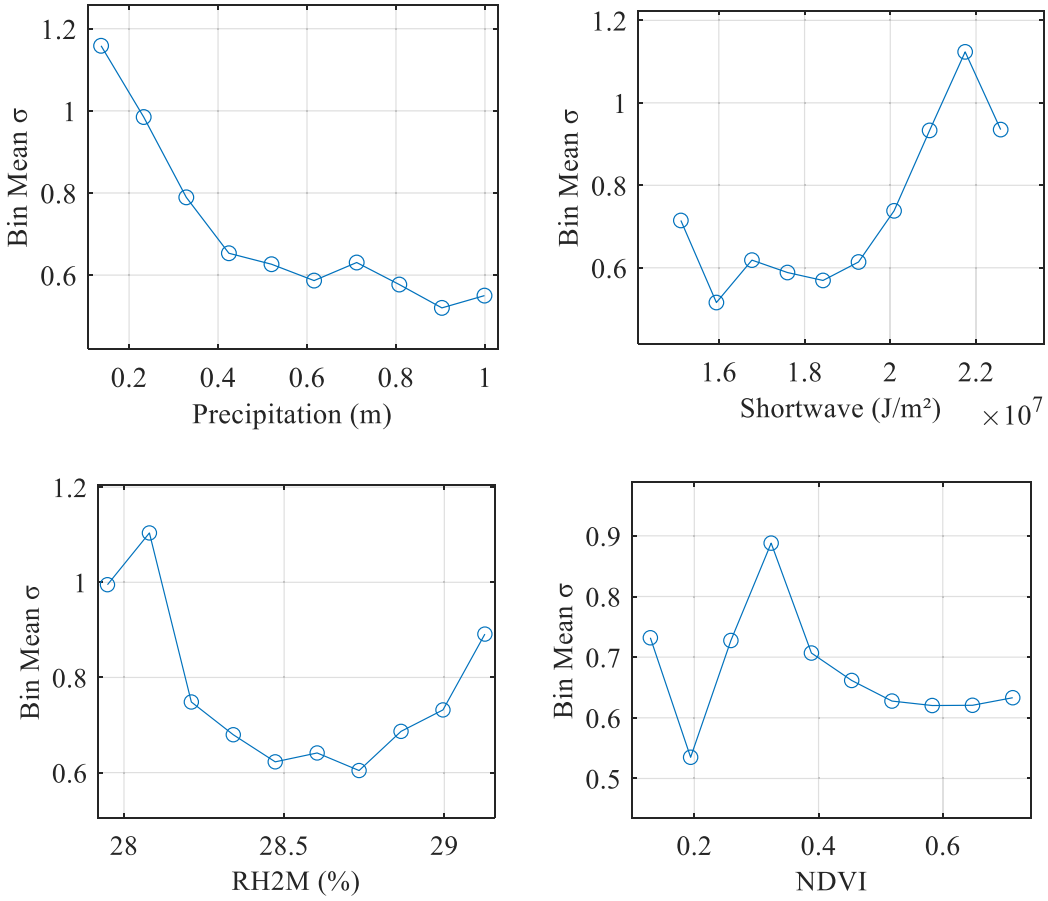


Figure 21 The relationship between input feature values and corresponding prediction uncertainty

4.4.6 Limitation

This study uses data from 842 counties from 2014 to 2023; however, not all regions have a complete 10 years of data. As shown in Figure 22, the dataset is centred around IA, with the data density decreasing as move outward. This distribution aligns with higher average annual yields, as these regions tend to exhibit more consistent agricultural production. This geographic bias reflects the benefits of focusing on regions with reliable data. However, the sparsity of data in certain areas contributes to model prediction biases, as the model struggles to generalise effectively in underrepresented regions.

Due to the constraints of data scale and temporal coverage, the study excludes characteristics related to irrigation and fertiliser use. The 10-year period was selected to minimise the potential impact of production efficiency changes over longer time horizons.

As a result, certain key factors that could influence predictions, such as changes in soil management practices or technological advances, are not accounted for.

Additionally, the model demonstrates limited accuracy in predicting extreme conditions or low-probability events, such as droughts, floods, and other natural disasters. These factors, though rare, can significantly impact agricultural yields and introduce variability that is not captured by the current dataset and modelling approach.

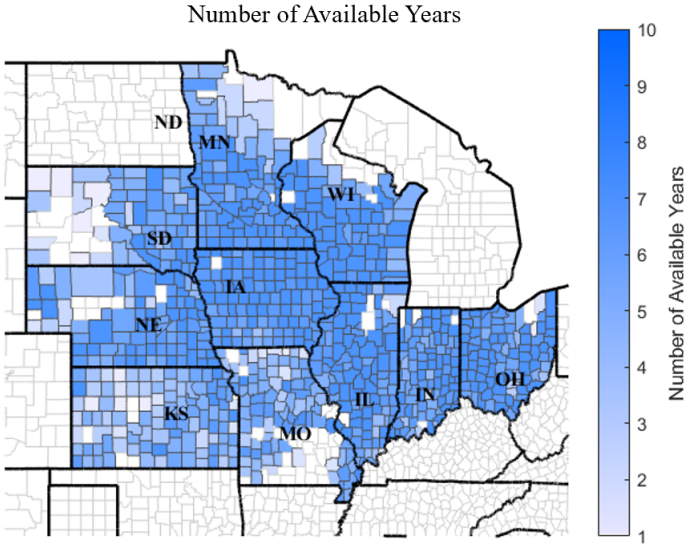


Figure 22 Available data for each county in 10 years

4.5 Conclusion

This study analysed maize yields across 842 counties in the U.S. Corn Belt from 2014 to 2023, alongside corresponding soil parameters such as pH, AWC, bulk density, electrical conductivity, and CEC, as well as climate parameters including T2M, RH2M, precipitation, and shortwave radiation. Additional inputs such as NDVI and planting and harvesting dates were also incorporated. Using four models—Linear Regression, GWR, VB-GWR, and VBNN-GWR—the analysis included uncertainty assessments, with VB-GWR demonstrating the best overall performance.

The results highlight that the VB-GWR model achieves relatively accurate predictions for yields between 150–220 bu/acre. Feature importance analysis revealed that pH had the highest weight across the Corn Belt, indicating its critical role in yield prediction, while precipitation had the lowest weight, likely due to the relatively uniform rainfall distribution and advancements in irrigation technology. The contributions of soil value and environmental value to yield predictions were found to be comparable. Additionally, the model quantifies the yield uncertainty for each county, providing insights into regional variations and the confidence associated with yield estimates.

Through uncertainty sensitivity analysis, the optimal ranges for each feature were quantified. The results showed that regions with higher uncertainty were often associated with lower yields. This finding suggests that identifying suitable ranges for key parameters can support land valuation and improve decision-making for agricultural expansion and land management in new areas.

Geographic analysis revealed that IA and IL are the regions with the highest land value, characterised by relatively balanced local regression coefficients across all inputs and the lowest uncertainty. As distance increases from IA and IL, maize yields decline, and prediction uncertainty rises. This pattern reflects the increasing dominance of specific inputs, which leads to greater model variability and reduced accuracy in peripheral regions.

In summary, this study demonstrates the importance of soil and climate factors in predicting maize yields, highlights the robustness of VB-GWR in handling uncertainty,

and provides actionable insights for land valuation and agricultural decision-making. These findings emphasise the need for region-specific modelling to address spatial variability and improve prediction accuracy across diverse agricultural landscapes.

5. Chapter 5: Spatial Analysis of Farmland Rental Prices in the U.S. Corn Belt Using a Geographically Weighted Regression Model

Abstract

The preceding chapter demonstrated the integration of uncertainty quantification and spatial modelling for maize yield prediction. Building upon that methodological foundation, this chapter applies the Geographically Weighted Regression (GWR) framework to analyse farmland rental prices.

Understanding the determinants of farmland rental prices is essential for agricultural stakeholders, policymakers, and investors. While previous studies have examined farmland rental prices using conventional regression models, they often overlook the spatial heterogeneity of economic and agricultural influences, limiting their ability to capture localised variations. Moreover, the factors that directly influence farmland rental prices are often difficult to obtain and may be affected by noise or measurement errors. This paper addresses this gap by employing Geographically Weighted Regression (GWR) to analyse how regional variations in macroeconomic indicators, commodity prices, and agricultural productivity impact farmland rental prices using the Corn Belt as a case study. The trained model is found to effectively captures the spatial sensitivity of key variables such as maize yield, oil prices, and GDP, revealing significant heterogeneity in rental price determinants. Findings indicate that maize yield has a weaker influence on rent in central of core production areas than other area of core Corn Belt, while oil prices exhibit a strong positive correlation, particularly in ethanol-producing states. Additionally, GDP and 10-year Treasury yields contribute to rental price fluctuations, highlighting the role of broader economic conditions. By evaluating uncertainty through confidence intervals (CI) derived from standard errors, the results reveal no clear spatial variation in the uncertainty of farmland rent. This suggests that the uncertainty is not strongly associated with spatial heterogeneity but is instead primarily driven by underlying economic and financial factors. By leveraging high-quality and readily available economic data, this research provides a practical framework for assessing farmland rental trends and risk. The findings offer valuable insights for financial planning, risk management, and capital-return modelling in agricultural investment.

Keywords: Corn Belt; Machine-learning; GWR; Rental Cost; Commodity; macroeconomic; USDA.

5.1 Introduction

The U.S. Corn Belt plays a crucial role in global agricultural production, with farmland rental rates and operational costs significantly impacting farm profitability and sustainability (Key, 2019; Mishra et al., 2009; Panagopoulos et al., 2015). Understanding the economic and agronomic factors influencing farmland rent is essential for policymakers, landowners, and farmers striving to optimise land use and financial returns (Annan et al., 2024; Paulson et al., 2010; Uludere Aragon, 2019).

Agricultural modelling and analysis provide valuable tools for examining the factors influencing farmland rental rates and yield. By incorporating data on climate, soil quality, market trends, and policy changes, these models enable researchers to predict shifts in land value and rental prices (Hendricks et al., 2014; Panagopoulos et al., 2015; Teste et al., 2024; Uludere Aragon, 2019). Recent advancements in econometrics and machine learning further enhance the ability to assess long-term profitability and sustainability of leased farmland. Such analytical approaches are critical for understanding how macroeconomic and environmental variables interact with localised farming conditions to shape economic outcomes (Ait Sidhoum, 2023; Zhang et al., 2019).

Farm performance and efficiency are closely tied to land tenure, input costs, and yield optimisation (Paulson et al., 2010; West et al., 2024). Differences in rental agreements，such as cash rent versus sharecropping, can affect productivity and investment decisions (Paulson et al., 2010). Land tenure security often influences willingness of framers to adopt innovative practices that enhance soil fertility and conservation efforts (Wang et al., 2021). A well-functioning rental market should therefore facilitate efficiency gains, while policies supporting equitable land access and resource allocation play an essential role in improving farm performance (Paulson et al., 2010).

Risk management in agriculture is a fundamental consideration for farmers operating under lease agreements (Ait Sidhoum, 2023; Arata et al., 2017). Market volatility, extreme weather events, and policy shifts all contribute to uncertainty in agricultural profitability (Arata et al., 2017; Uludere Aragon, 2019). Effective risk management strategies, including crop insurance, forward contracting, and flexible lease terms, help mitigate these risks (Archer & Reicosky, 2009). High rental rates may exacerbate

financial vulnerability, making it crucial for farmers to adopt adaptive strategies that safeguard their economic viability while ensuring resilience against external shocks (West et al., 2024).

Environmental considerations are increasingly shaping farmland rental agreements and operational strategies. Sustainable land management practices, such as cover cropping and reduced tillage, contribute to long-term soil health and water conservation (Brock et al., 2021; McNunn et al., 2020; Oh & Gramig, 2023; Tessema et al., 2018). However, tenant farmers may have limited incentives to invest in such practices if rental terms discourage long-term planning. Policy mechanisms that align environmental sustainability with economic incentives, such as conservation-focused rental agreements, can enhance the ecological benefits of agricultural land use while maintaining farm profitability (Arata et al., 2017; McNunn et al., 2020).

Land use and crop management decisions are directly influenced by rental costs, with implications for soil health, crop diversity, and long-term sustainability (Uludere Aragon, 2019). Higher rental prices may drive farmers toward high-return crops, potentially affecting crop rotation patterns and environmental outcomes. Additionally, regional variations in soil quality and climatic conditions necessitate localised approaches to land management and rental agreements (Hendricks et al., 2014). A deeper understanding of these dynamics can inform policies that balance economic efficiency with environmental stewardship.

Off-farm work is a crucial factor affecting farmland rental and operational costs. As rental prices and input costs rise, many farmers supplement their income with off-farm employment (Bouchakour & Saad, 2019). This shift has implications for farm management, as time and labour constraints may limit the adoption of resource-intensive farming practices (Ma et al., 2018). The increasing prevalence of off-farm income highlights broader structural shifts in rural economies and the evolving nature of farm household livelihoods (Bouchakour & Saad, 2019).

Agricultural profitability and finance remain at the core of farmland rental considerations. Access to credit, interest rates, and government subsidies influence farm financial stability and long-term investment decisions (West et al., 2024). Rental costs represent a

significant component of farm budgets, affecting liquidity and profitability. Research on farm finance underscores the importance of policies that enhance credit accessibility and provide financial safeguards for tenant farmers (Mishra et al., 2009). Understanding these financial dynamics is essential for developing strategies that support profitable and resilient farming operations.

Despite the critical role of farmland operational costs in agricultural decision-making, research on this topic remains limited. A primary challenge in this area is the difficulty in obtaining high-quality, detailed regional data regarding the factors mentioned above. The obtained data may often be rare, subject to outliers, or influenced by confidentiality constraints for example, private leasing agreements and contractual terms. Additionally, micro-scale climatic variations, local labour market conditions, and difficult-to-quantify factors, including geopolitical tensions and trade tariff conflicts, further complicate accurate measurement and analysis. In contrast, macroeconomic activity can often be effectively captured through broader asset classes, such as the 10-year US Treasury yield and WTI crude oil prices, which serve as reliable proxies for underlying economic dynamics (Coeurdacier & Rey, 2013).

Given that farmland rents exhibit distinct regional patterns, this study hypothesises that the sensitivity of rental costs to macroeconomic and commodity indicators is similarly characterised by regional variation. Therefore, to address the aforementioned gap, this research aims to develop a spatially weighted analytical model using readily accessible and high-quality macroeconomic and commodity data, integrated with agricultural indicators, to model regional farmland operational costs, particularly rents. By validating the effectiveness of this model, the study seeks to provide practical insights into agricultural operating costs derived from broader economic indicators, facilitating improved financial planning, risk management, and capital-return modelling for agricultural stakeholders.

This paper first presents the academic background of farmland rental costs, highlighting the significance and challenges associated with this research area. It then provides a detailed description of the dataset utilised in the analysis, including its key characteristics and relevance to the study. Subsequently, the mathematical methods employed in this

research are introduced, offering a comprehensive explanation of the analytical techniques applied. The validity of the proposed model is then assessed through empirical evaluation. In the results analysis section, the paper quantitatively examines the impact of each factor on farmland rental costs. Additionally, projections for 2026 and 2027 farmland rental costs are provided, offering insights into potential future market trends and their broader economic implications.

5.2 Dataset

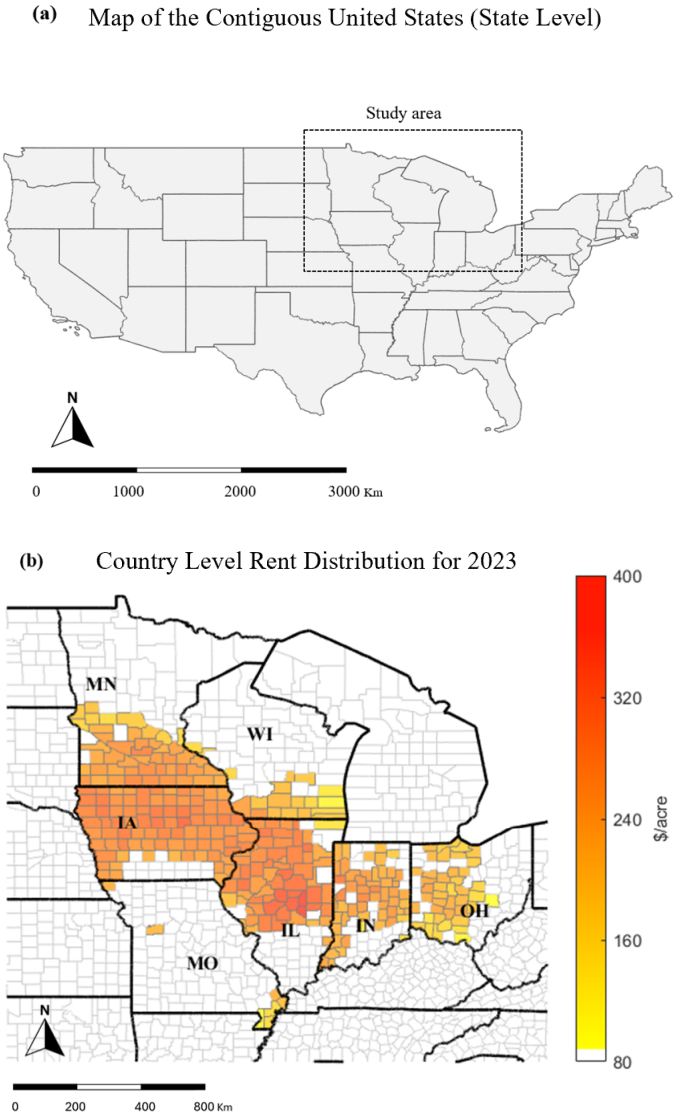


Figure 23 County-level rental cost distribution in 2023. ((a) the study area of contiguous U.S. (b) rental cost distribution of core Corn Belt)

This study selects regions within the core Corn Belt where the ten-year average annual yield exceeds 160 bushels per acre as the sample, encompassing a total of 332 counties. Due to the unavailability of comprehensive operational cost data, this study uses farmland rent as the target variable. As shown in Figure 23, the 2023 rental rates exhibit distinct regional characteristics. While the proportion of rent in overall operational costs is not fixed, historical data from 2015 to 2023 indicate that the rent-to-operational cost ratio has exhibited only narrow fluctuations across states, as illustrated in Figure 24. Therefore,

using regional rent levels as a proxy for estimating operational costs is a feasible approach in this context.

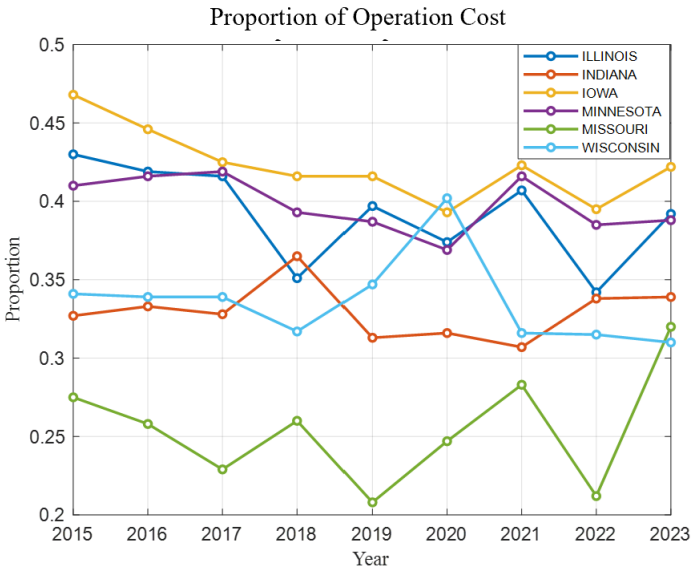


Figure 24 Ratio of rental cost and operation cost from 2015 to 2023

5.2.1 Data Selection

The data used in this research are categorised into three primary groups: agricultural data, commodity data, and macroeconomic data. This structured approach ensures that inputs are readily accessible and accurately incorporated into the modelling process, enhancing the robustness and reliability of the results. To ensure consistency and comparability within the modelling process, annual averages of these variables were calculated and used as input variables (Table 3).

5.2.1.1 Macroeconomic Indicators

Macroeconomic indicators used in this study were obtained from the Bureau of Economic Analysis (BEA) of the United States Department of Commerce. These indicators include Gross Domestic Product (GDP), Personal Consumption Expenditures (PCE), and the U.S. Dollar Index (DXY), all of which serve as key measures of overall economic activity, consumer spending patterns, and currency strength.

GDP reflects the overall economic prosperity and productivity of the country, influencing investment decisions, land values, and agricultural market stability. PCE, on the other hand, captures household consumption trends and purchasing power, which can indirectly

impact agricultural rental markets by affecting demand for agricultural products and farm incomes (Jiang et al., 2017). DXY measures the value of the U.S. dollar relative to a basket of major foreign currencies, influencing trade competitiveness, import and export prices, and global commodity markets (Arfaoui & Ben Rejeb, 2017). As agriculture is highly interconnected with international trade, fluctuations in DXY can affect farm revenues, input costs, and farmland rental prices.

Table 3 Selected input features

Category	Variable	Source	Unit	Level	Final Selection	Reason
Target	Rent	USDA	$/acre/year	County	Include	None
Agriculture	Yield	USDA	bu/acre	County	Include	None
Economic	Maize Price	USDA	$/bu	State	Include	Interdependence
Economic	10Y Treasury	FRED	%	Nation	Include	None
Economic	PCE	BEA	%	Nation	Exclude	Multicollinearity
Economic	GDP	BEA	%	Nation	Include	None
Economic	DXY	BEA	$	Nation	Exclude	Multicollinearity
Commodity	Oil Price	EIA	$/barrel	Nation	Include	None

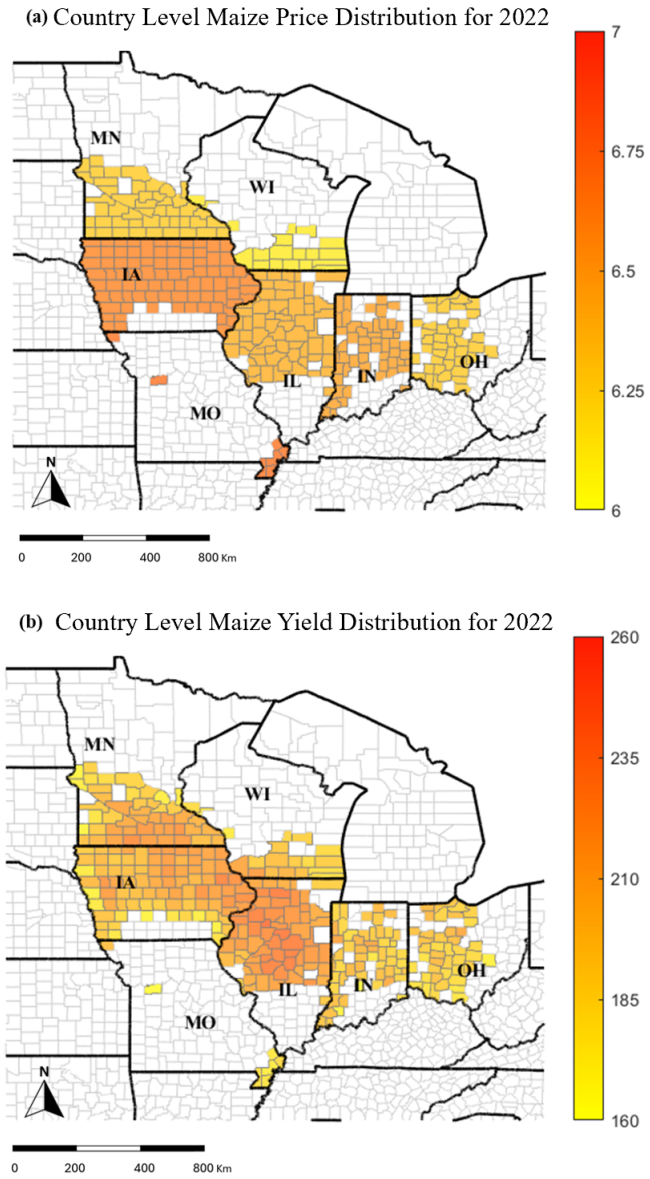


Figure 25 the distribution of (a) Maize price (state-level) and (b) yield (county-level) for 2022

5.2.1.2 Agriculture Data

The USDA Yield Data were sourced from the USDA Database, which provides county-level maize yield measurements expressed in bushels per acre Figure 25(a). These data play a crucial role in the analysis, offering valuable insights into the spatial distribution and regional variability of maize production.

To compare the distribution of rent in 2023, the study utilises the distribution of maize price and yield from the previous year (2022), as shown in Figure 25(b). By calculating

the coefficient of determination (COD), the correlations between maize price, yield, and rent were found to be 0.18 and 0.47, respectively. While yield exhibits a stronger correlation with rent, the results indicate that a single input alone is insufficient to accurately predict farmland rent for the following year.

5.2.1.3 Commodity Price Indicators

The commodity price indicators used in this study include West Texas Intermediate (WTI) crude oil prices and maize revenue prices, obtained from the U.S. Energy Information Administration (EIA) and the United States Department of Agriculture (USDA) (EIA, 2024; USDA, 2024).

WTI crude oil prices serve as a key indicator of energy costs, which have significant implications for agricultural production expenses, including fuel for machinery, transportation, and fertiliser costs. As energy prices fluctuate, they can directly impact the cost structure of farming operations, influencing land rental prices accordingly (Narani et al., 2019).

Maize revenue prices, sourced from USDA, reflect the market value of maize production, capturing the economic returns that farmers receive from their crops. These prices are influenced by various factors, including supply and demand conditions, global trade policies, and macroeconomic trends. Higher maize prices generally indicate stronger agricultural profitability, which can, in turn, affect land values and rental rates (Baum et al., 2020).

5.2.1.4 Geolocation Information

The geographic location data used in this study were obtained from the Topologically Integrated Geographic Encoding and Referencing dataset provided by the U.S. Census Bureau. This dataset offers precise geospatial information, represented in latitude and longitude coordinates, ensuring accuracy in spatial analysis.

These coordinates play a crucial role in calculating spatial distances between observations, which are fundamental for determining the bandwidth in the Geographically Weighted Regression (GWR) model. The bandwidth defines the spatial scope of localised regression, determining the extent to which neighbouring observations influence the

weighting function. By incorporating precise geographic data, the model effectively captures spatial dependencies and regional variations in rental price dynamics.

5.2.2 Data Integration and Preprocessing

In this study, data were first aggregated into county-level geographic units based on the coordinates. To evaluate temporal coverage, the number of years available for each unit was counted. Counties with data for only a single year were fully assigned to the training set to maximise sample utilisation.

For counties with data spanning multiple years, one year was randomly selected for inclusion in the test set, while the remaining years were allocated to the training set. This approach ensures temporal independence between the training and test sets while also prioritising the inclusion of counties with more than two years of data in the test set, thereby enhancing regional representativeness.

Following this stratification, the training data were further randomly split into final training and validation sets. The division of training, validation, and test sets remained consistent across all models, ensuring comparability and fairness in performance evaluation.

5.3 Methodology

This section outlines the methodological framework adopted in this study, highlighting each key stage from data collection to model application. The process involves temporal alignment of variables, feature selection through statistical diagnostics, and spatial modelling using Geographically Weighted Regression (GWR). The methodological flow chart is shown in Figure 26.

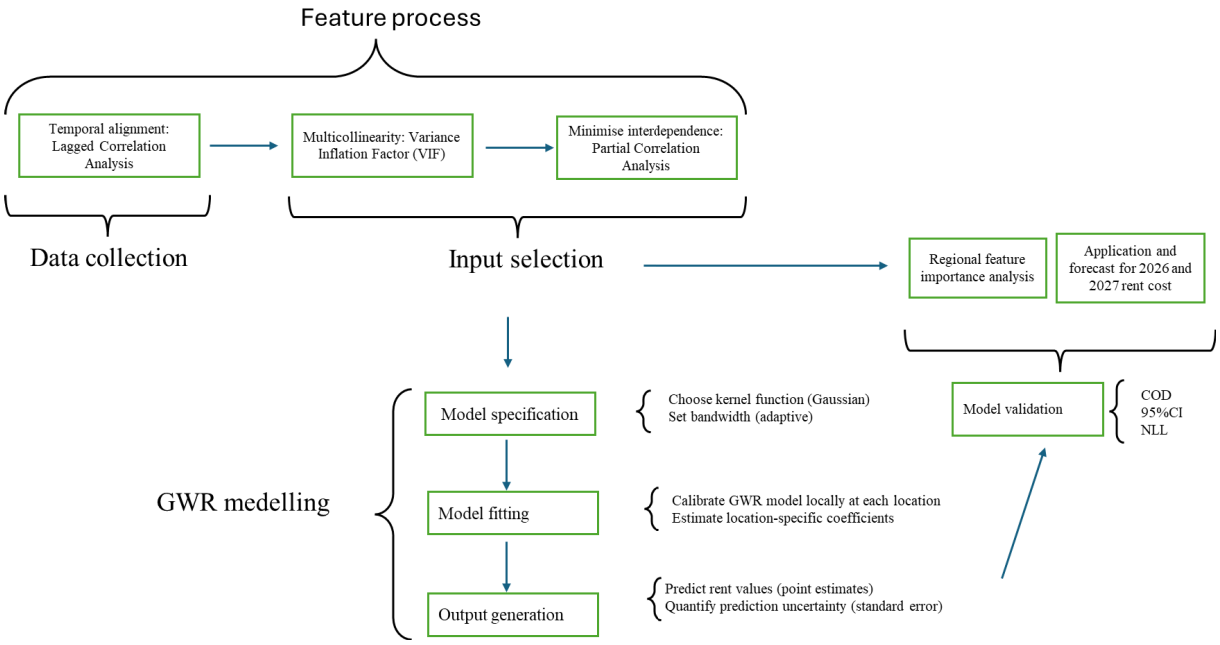


Figure 26 Methodological flow chart

5.3.1 Lagged Correlation Analysis

Lagged correlation analysis is used to examine the delayed effects of inputs on a target variable over time (Wang et al., 2018). In this study, the influence of past values of agricultural, financial, and economic indicators on farmland rent is assessed by introducing time lags in the data. For each input feature X_t, its correlation with farmland rent Y_t at different lag periods k is calculated:

$$\rho_k = corr(X_{t-k}, Y_t) \tag{5.1}$$

Where ρ_k represents the Pearson correlation coefficient at lag k, By shifting feature values backward by k years within each state and county, this method captures the temporal dependencies between historical changes in yield, commodity prices, interest

rates, and macroeconomic factors with current farmland rent. Identifying significant lagged relationships provides insights into long-term trend. The larger the absolute value of ρ_k the more significant the effect at the given lag period.

5.3.2 Input Cleaning

5.3.2.1 Variance Inflation Factor (VIF) for Detecting Multicollinearity

Variance Inflation Factor (VIF) is a statistical measure used to detect multicollinearity among input variables in a regression model (Salmerón-Gómez et al., 2024). Multicollinearity occurs when outputs are highly correlated, leading to inflated standard errors and unstable coefficient estimates, which can undermine the interpretability and reliability of the model (Chalkias et al., 2020).

$$VIF(X_j) = \frac{1}{1 - COD_j} \tag{5.2}$$

Where COD_j is the coefficient of determination obtained by regressing X_j on all other outputs. A VIF below 5 suggests low multicollinearity and is generally acceptable. A VIF between 5 and 10 indicates moderate multicollinearity, which may require further investigation. When VIF exceeds 10, it suggests a high degree of multicollinearity, meaning the variable is highly redundant and should potentially be removed or adjusted to improve model stability.

5.3.2.2 Partial Correlation Analysis for Feature Relationships

In this study, each individual variable is analysed for its impact on the model. However, when a variable changes and subsequently influences other variables, effective analysis becomes challenging. One approach to address this issue is to construct a Bayesian network to model the joint distribution (Basnet et al., 2021; Tiffin & Balcombe, 2011). However, the limited sample size within individual regions, this method is not suitable for the current study. Alternatively, partial correlation analysis is employed to refine the input variables by identifying and mitigating redundant or highly outputs, ensuring a more robust and interpretable model (Wang et al., 2024).

Partial correlation analysis measures the direct relationship between two variables while controlling for the influence of other variables. Unlike Pearson correlation, which

captures both direct and indirect associations, partial correlation isolates the unique contribution of each variable by removing the confounding effects of others. In this study, the individual input will be analysed for the impact on the model.

$$r_{X_i,X_j|Z} = corr\left(\widehat{\boldsymbol{X_i}}, \widehat{\boldsymbol{X_j}}\right) \tag{5.3}$$

Where $\widehat{\boldsymbol{X_i}}$ and $\widehat{\boldsymbol{X_j}}$ are the residuals obtained by regressing X_i and X_j on the control variables Z. A higher partial correlation coefficient $r_{X_i,X_j|Z}$ indicates a stronger unique association between $\widehat{\boldsymbol{X_i}}$ *and* $\widehat{\boldsymbol{X_j}}$ after controlling for other factors.

5.3.3 Machine Learning Method Selection

5.3.3.1 Geographically Weighted Regression

Geographically Weighted Regression (GWR) is a statistical method designed to analyse spatial relationships between outputs and inputs. Unlike traditional regression models, which assume globally constant parameters, GWR allows coefficients to vary across different geographic locations, providing a more localised understanding of spatial heterogeneity (Neelawala et al., 2012). By incorporating a spatial weighting matrix, GWR assigns higher importance to nearby observations while down-weighting more distant data points, making it particularly useful for capturing regional variations.

In farmland rental price research, factors such as location, proximity to infrastructure, transportation accessibility, and regional economic conditions can lead to significant spatial differences in rental prices (Paulson et al., 2010). Traditional regression models may fail to account for these localised variations, resulting in oversimplified or biased predictions. GWR addresses this issue by constructing separate regression models for different geographic areas, allowing for a more precise analysis of local market dynamics and rental price determinants. This approach provides valuable insights for land leasing decisions and regional planning. The general formulation of the GWR model can be expressed as follows (Fotheringham et al., 2001) :

$$y_i = \beta_0(u_i, v_i) + \sum_{k=1}^{K} \beta_k(u_i, v_i) x_{ik} + \epsilon_i \tag{5.4}$$

In a Geographically Weighted Regression (GWR) model, the spatially varying coefficients $\beta_k(u_i, v_i)$ are estimated at each geographic location (u_i, v_i), allowing for localised relationships between variables. The intercept term, $\beta_0(u_i, v_i)$ represents the baseline contribution in the absence of inputs influence. The error term $\boldsymbol{\epsilon_i}$ follows a normal distribution with constant variance.

The spatial weighting matrix W_{ij} which captures the influence of neighbouring observations on local parameter estimates. This matrix determines how much weight each observation *j* contributes to the estimation of parameters β_k at a specific location *i*.

The elements of W are based on the spatial proximity between observations and are typically computed using a kernel function. A commonly used approach is the Gaussian kernel, which defines the weight W_{ij} between locations *i* and *j* as (Fotheringham et al., 2001):

$$W_{ij} = exp\left(-\frac{d_{ij}^2}{h^2}\right) \tag{5.5}$$

The distance between locations *i* and *j* denoted as d_{ij} plays a crucial role in determining spatial influence within the GWR model. The bandwidth parameter, *h,* controls the extent of this influence, with observations closer to the focal location *i* receiving greater weights, while those further away contribute less.

Optimising the bandwidth h is essential for balancing local and global model behaviour. A smaller bandwidth emphasises fine-scale spatial variations, capturing highly localised patterns, whereas a larger bandwidth smooths over regional differences, favouring global trends. To achieve optimal performance, the bandwidth can be fine-tuned during model training using gradient descent. In this approach, bandwidth is treated as a learnable parameter, iteratively adjusted to minimise the loss function.

The spatial weighting matrix is applied in the local regression process using weighted least squares (Fotheringham et al., 2001).

$$\hat{\beta}(u_i, v_i) = (X^T W_i X)^{-1} X^T W_i y \tag{5.6}$$

where X represents the matrix of inputs, y is the output vector, and W_i is the spatial weighting matrix cantered at location i.

5.3.3.2 Uncertainty Modelling in GWR Using Standard Errors

In Geographically Weighted Regression (GWR), uncertainty modelling is essential for assessing the reliability of local parameter estimates. One common approach to quantify uncertainty is through standard errors (SEs), which provide insight into the variability of estimated coefficients across geographic locations (Ribeiro & Pereira, 2018). Standard errors help determine whether the spatial variations in coefficients are statistically significant or merely the result of random fluctuations.

The standard errors of the local regression coefficients are calculated based on the hat matrix **S**, which captures the influence of each observation on the estimated parameters (Yu et al., 2019). The standard error of a coefficient $\beta_k(u_i, v_i)$ at location i is given by:

$$SE\big(\beta_k(u_i, v_i)\big) = \sqrt{\sigma^2 \cdot S_{ii}} \tag{5.7}$$

Where σ^2 is the estimated variance of the residuals, S_{ii} is the diagonal element of the hat matrix, representing the leverage of observation *i.*

5.3.4 Evaluation Criteria

5.3.4.1 Feature importance based on coefficient variability

In this study, feature importance is assessed by computing the standard deviation of regression coefficients, measuring the variability of each input influence across different geographic locations (Acal et al., 2023). The coefficient variability can be calculated:

$$\sigma_{X_i} = \sqrt{\frac{1}{n}\sum_{j=1}^{n}\left(\beta_{X_i}(u_j, v_j) - \frac{1}{n}\sum_{j=1}^{n}\beta_{X_i}(u_j, v_j)\right)^2} \tag{5.8}$$

Unlike traditional global regression models (e.g., linear regression), which assume that feature effects remain constant across all regions, this approach utilises Geographically Weighted Regression (GWR), allowing regression coefficients to vary spatially.

5.3.4.2 Confidence Interval (CI)

Since this study does not employ a Bayesian framework, the confidence interval (CI) is used to define the uncertainty range. To preserve the integrity of predictive uncertainty assessment, standard errors for the test set are derived from the training set to prevent data leakage. Estimating standard errors directly from the test set would introduce bias and undermine the generalisability of the model. By employing the average variance of the training set coefficients, we ensure a robust and independent evaluation of prediction uncertainty, adhering to best practices in statistical modelling and enhancing the reliability of uncertainty quantification in out-of-sample predictions. The general formulation of confidence interval in this study can be expressed as follows (Abdalla et al., 2024):

$$CI_{95\%} = \beta_k(u_i, v_i) \pm 1.96 \cdot SE\big(\beta_k(u_i, v_i)\big) \tag{5.9}$$

By evaluating the coverage rate of the prediction results, the 95% confidence interval (CI) should ideally encompass 95% of the predicted samples. Given the inherent variability in data and model estimation, a reasonable coverage range is typically between 92% and 97%. If the observed coverage falls significantly below this range, it may indicate that the model underestimates uncertainty, leading to overly narrow confidence intervals. Conversely, if the coverage exceeds this range, the model may be overestimating uncertainty, resulting in excessively wide intervals.

5.3.5 Sensitivity Analysis

For the sensitivity analysis of mean farmland rent predictions, the parameter weights from the Geographically Weighted Regression (GWR) model are employed. Since GWR is a linear modelling approach, its parameter estimates directly represent the contribution of each input variable to rental price variations. By analysing these weights across different geographic regions, the spatial heterogeneity in the significance of influencing factors can be observed. This approach provides valuable insights into how local economic and environmental conditions shape farmland rent dynamics, highlighting regional disparities in key determinants.

5.4 Results and discussion

5.4.1 Lagged Correlation Analysis

The lagged correlation analysis results between the target rent and inputs such as yield, maize price, oil price, Dollar Index, 10-year Treasury Yield, PCE, GDP, are represented as Table 4. The results indicate that yield exhibits the strongest positive correlation with rent in the 4th year, suggesting that farmland rental prices may adjust to past agricultural production levels over a longer period. However, PCE and GDP show the highest positive correlations with rent in the 1st year, implying that macroeconomic growth and consumer spending changes are quickly reflected in the agricultural rental market. Additionally, the 10-year Treasury Yield has the most significant negative correlation with rent in the 1st year, indicating that rising long-term interest rates may increase agricultural financing costs, thereby suppressing rental price growth.

Table 4 Lagged Correlation for 5 years

	1st yr	2nd yr	3rd yr	4th yr	5th yr
Yield	0.40	0.39	0.41	0.45	0.42
Maize Price	0.25	0.11	-0.07	-0.18	-0.18
Oil Price	-0.08	-0.18	-0.18	-0.19	-0.20
DXY	0.33	0.26	0.23	0.23	0.25
10yr Treasury	-0.41	-0.40	-0.26	-0.12	-0.10
PCE	0.42	0.36	0.30	0.25	0.25
GDP	0.41	0.36	0.30	0.25	0.25

Although different variables influence rent over varying timeframes, this study adopts a one-year lag for modelling, considering that five input variables exhibit a one-year lag effect. This choice balances short-term and long-term impacts while capturing the primary driving factors behind rent dynamics. Furthermore, selecting a shorter lag period

reduces historical data uncertainty and enhances model robustness, making it more suitable for policy analysis and market decision-making.

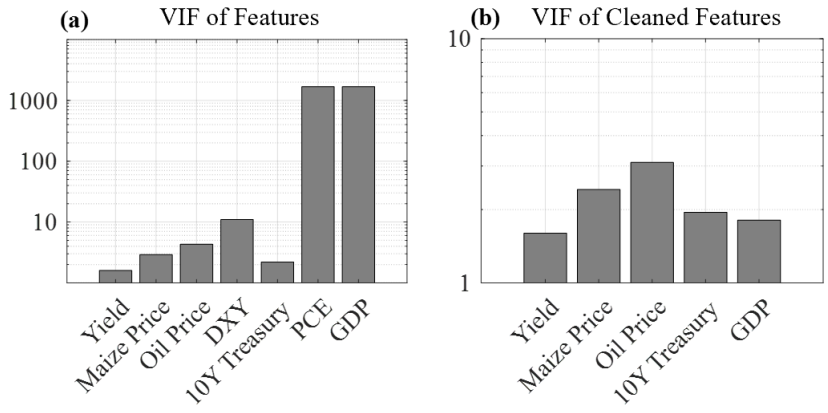


Figure 27 VIF for original feature and cleaned feature ((a) and (b) show the variance inflation factors (VIFs) prior to and following the removal of the selected feature.)

The Variance Inflation Factor (VIF) in Figure 27 represents the results before and after removing highly collinear features. In chart (a), three input variables, DXY, PCE, and GDP, exhibit VIF values exceeding 10, indicating severe multicollinearity. To address this, only one among these highly correlated variables was retained while eliminating the others to reduce redundancy. Chart (b) shows the revised feature set, where all remaining variables now have VIF values below 10, ensuring reduced multicollinearity and a more stable regression model.

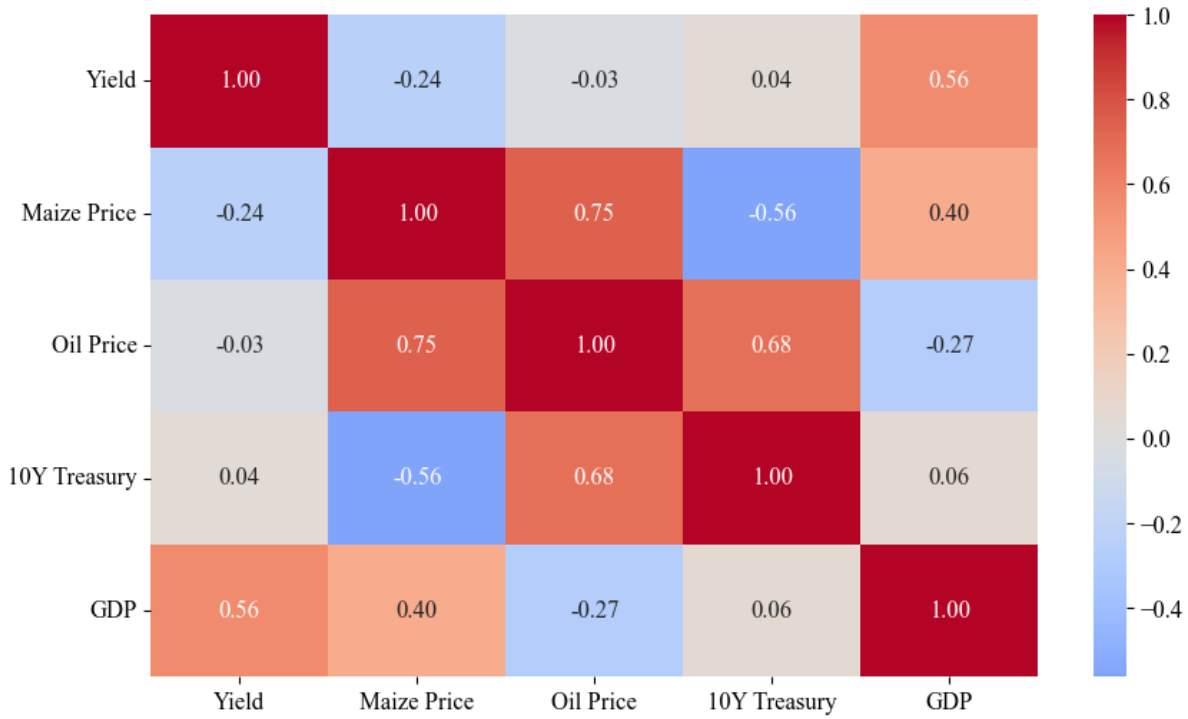


Figure 28 Partial correlation analysis for cleaned features

After eliminating multicollinearity, a partial correlation analysis was conducted to examine the direct relationships between variables, and results are presented in Figure 28. The results indicate that oil price, 10-year U.S. Treasury yield, and maize price exhibit interdependencies, making it challenging to isolate individual effects in the analysis. Among these, maize price and oil price have a partial correlation coefficient exceeding 0.7, indicating a strong correlation. Due to this redundancy, maize price was removed to reduce collinearity issues.

Additionally, GDP and yield show some level of correlation, as yield is a component of GDP (Xie et al., 2017). However, in this study, yield is not classified as an economic variable, and its correlation with GDP remains below 0.6 in Figure 28, suggesting that it does not introduce significant collinearity. Therefore, both GDP and yield were retained in the final model to preserve relevant economic and agricultural insights.

5.4.2 Model Validation

The results of rent prediction using linear regression (Figure 13(a)) indicate that the model fails to capture underlying patterns effectively. Additionally, the test set COD (0.45) is higher than the training set COD (0.38), suggesting that the model suffers from underfitting, as it does not generalise well across different samples. Hence, linear regression model is not suitable for this study.

In contrast, the GWR model (Figure 13(b)) demonstrates significantly improved performance, with a training and test COD of 0.85, indicating that the model generalises well across different datasets. This suggests that the influence of input variables varies across geographic locations, reinforcing the need for a spatially adaptive model. The strong agreement between training and test performance highlights that GWR effectively captures the spatial heterogeneity in farmland rent determinants.

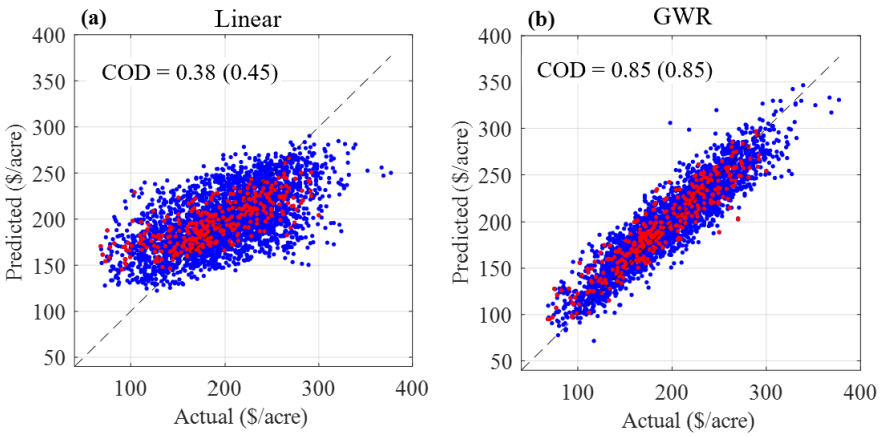


Figure 29 Performance of linear regression and GWR model ((a) the result of linear regression. (b) the result of GWR model)

For the uncertainty GWR model, the results are shown as Figure 14. The model achieves a COD of 0.85 for both training and testing sets, indicating a strong predictive capability. The NLL remains at 4.38, suggesting a consistent likelihood estimation across both training and testing datasets. Additionally, the 95% Confidence Interval (CI) covers 96% samples in the test set, demonstrating that the model provides well-calibrated uncertainty estimates.

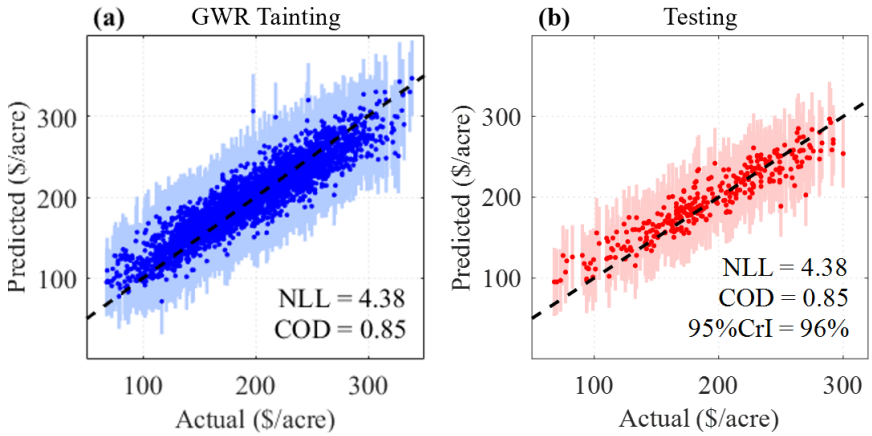


Figure 30 Uncertainty performance of GWR models (Vertical lines represent 95% credible interval. (a) the result of GWR training set (b) the result of GWR testing set)

The probability density distribution of standard errors (SE) for both training and testing datasets is shown as Figure 31. The SE values exhibit a similar distribution across both training and testing datasets, cantering around 21 with a spread between 18 and 25, further confirming the consistency in uncertainty estimation. The similarity between training and testing standard error distributions reinforces that the GWR model maintains a stable uncertainty structure, avoiding overfitting or excessive confidence in predictions.

The validation results from both the test and validation sets indicate that the model is both reliable and robust. The consistency in predictive performance across different datasets suggests that the model effectively captures the underlying relationships between the input variables and farmland rental prices.

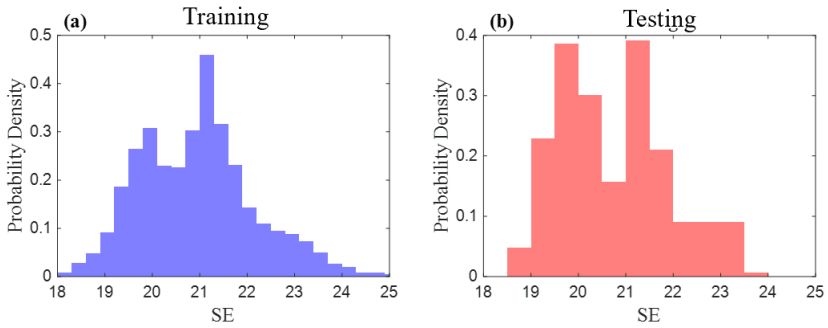


Figure 31 Probability density distribution of standard Errors (SE) ((a) the probability density distribution of SE in training set. (b) the probability density distribution of SE in testing set)

5.5. Model Analysis

5.5.1 Feature impact importance

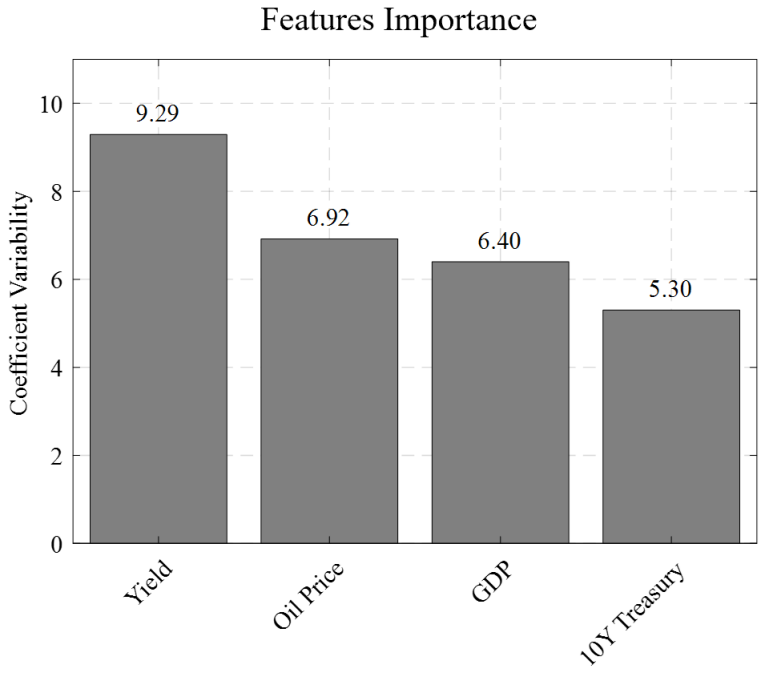


Figure 32 The rank of feature importance

The result of feature importance in Figure 32 indicate that yield emerges as the most significant determinant (9.29), aligning with the understanding that agricultural productivity plays a crucial role in shaping rental market dynamics. In contrast, Oil Price (6.92) exhibits strong influence, as fuel and energy costs directly impact agricultural production expenses, making it a key cost-side driver. Among macroeconomic variables, GDP (6.40) and 10Y Treasury (5.30) show notable contributions, reflecting the role of consumer expenditure and currency fluctuations in shaping market conditions and investment sentiment.

5.5.2 Regional uncertainty map

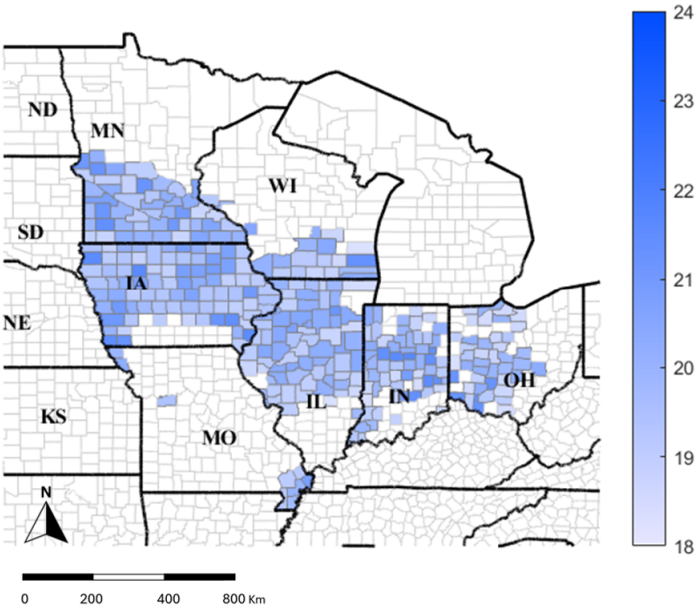


Figure 33 County-level uncertainty quantification map

County-level uncertainty in rental price predictions, with varying shades of blue representing different levels of uncertainty in Figure 33 and Figure 33. Based on the selected input variables, rental price uncertainty does not display a clear regional pattern, indicating that geographic factors may have a limited impact within the current model.

In the core Corn Belt region, which includes Iowa, Illinois, Indiana, and Ohio, uncertainty levels remain relatively consistent across counties, with no strong spatial clustering of high or low uncertainty areas. This suggests that farmland rental prices in this region are relatively stable, potentially due to well-established market conditions, consistent agricultural productivity, and uniform land valuation practices. The absence of strong spatial differentiation in uncertainty highlights those non-geographic factors, such as economic conditions or policy influences, may play a more significant role in driving rental price variability.

5.5.3 Sensitive analysis

The distribution of weight of each feature is presented as Figure 34 illustrating the relative importance of different factors in determining farmland rental prices. The results

highlight variations in feature significance across regions, indicating that some variables exert a stronger influence in certain locations.

The local regression coefficients illustrate that yield generally has a positive relationship with rental prices, particularly along the periphery of the core Corn Belt (Figure 34(a)). This pattern suggests that in regions where agricultural expansion is still occurring or where land productivity is more variable, higher yields are associated with increased land rental values.

However, in mature corn-producing areas, the influence of yield on rental price variation is more limited. These regions benefit from a well-established agricultural infrastructure, stable production systems, and lower yield uncertainty, which contribute to more efficient market pricing. As a result, fluctuations in yield have already been fully incorporated into rental agreements, reducing their direct impact on rental price dynamics.

Oil price generally exhibits a positive correlation with rental prices and shows a distinct west-to-east gradient in its influence (Figure 34(b)). This pattern is likely tied to transportation costs and varying energy demands across states. In Minnesota, for example, one of major ethanol-producing states—rising oil prices can drive up ethanol prices, thereby encouraging greater production and pushing up overall land costs, including rents. Additionally, colder climates demand more fuel for machinery, further reinforcing the effect of oil on rental markets.

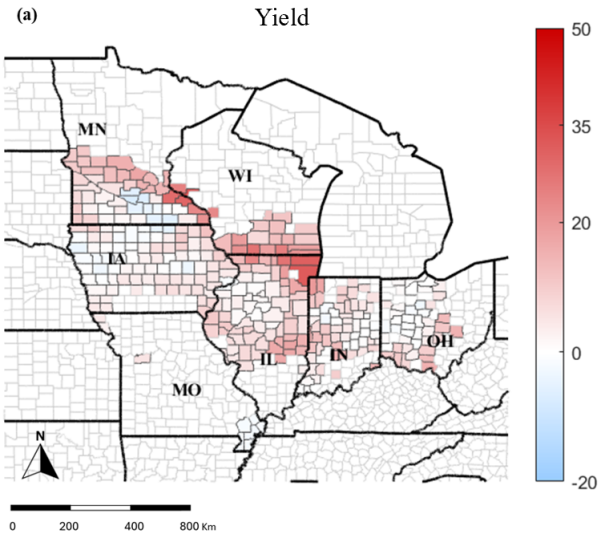
In contrast, states such as Ohio have fewer ethanol plants and thus lower ethanol demand, leading to diminished sensitivity to fuel price fluctuations. Consequently, the influence of oil price on rental values wanes toward the eastern part of the region.

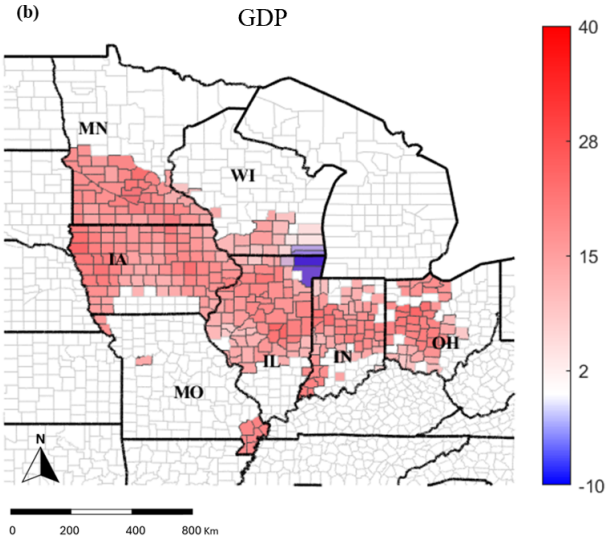
The impact of the 10-year Treasury yield on rental prices exhibits a spatial pattern like that of oil prices, largely due to the high correlation between Treasury yields and oil prices (Figure 34(c)). This results in a comparable regional effect. However, the sensitivity to the 10-year Treasury yield also reflects differences in local economic structures. Compared to agriculture, industrial activities tend to be more sensitive to long-term interest rates, making regions with significant ethanol processing facilities particularly affected.

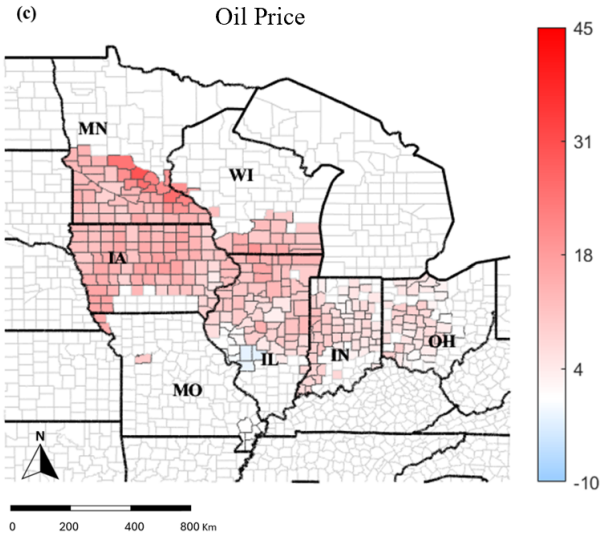
In these areas, higher interest rates lead to increased financing costs for businesses, potentially limiting expansion plans or raising production costs, which can influence rental prices. Conversely, in agricultural regions, the effect of the 10-year Treasury yield is more likely to be transmitted through macroeconomic conditions and investment expectations rather than directly impacting farm-level production decisions. These variations in industrial composition and financing needs result in differing levels of sensitivity to Treasury yield fluctuations, contributing to the observed spatial heterogeneity in rental price responses.

The spatial relationship between GDP and farmland rent, where red-shaded areas indicate a positive association, and blue-shaded areas represent a negative or weaker relationship which only occur in metropolitan area such as Chicago (Figure 34(d)). The results suggest that GDP is positively correlated with farmland rent, reflecting the role of economic prosperity in shaping rental market dynamics.

In the core agricultural regions, such as Iowa, Illinois, and Indiana, GDP exhibits a stronger influence, suggesting that farmland rent in these areas is more responsive to economic conditions. This heightened sensitivity may be attributed to higher agricultural productivity, stronger market integration, and greater investment in agribusiness infrastructure, which link land values more closely to broader economic performance. Conversely, the presence of isolated blue regions suggests that in certain counties, local economic activity may not be the primary driver of farmland rent, possibly due to alternative land-use priorities or differences in agricultural dependency.


(a)
Yield
MN
WI
IA
IL
IN
OH
MO
N
0
200
400
800 Km
50
35
20
0
-20


(b)
GDP
MN
WI
IA
IL
IN
OH
MO
N
0
200
400
800 Km
40
28
15
2
-10


(c)
Oil Price
MN
WI
IA
IL
IN
OH
MO
N
0
200
400
800 Km
45
31
18
4
-10

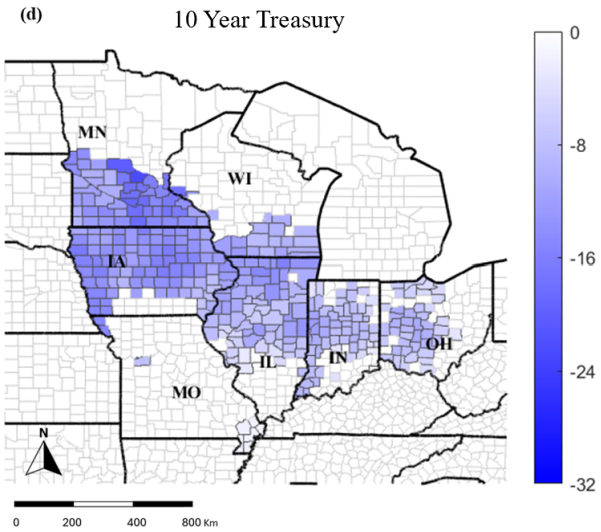


Figure 34 Country-level weight distribution of different features ((a) maize yield, (b) GDP, (c) oil, (d) 10-year treasury yield)

5.6 Application and forecast

This study utilises economic forecasts for the upcoming two years to conduct a systematic rent projection. Data from 2025 and 2026 are used, while 2026 and 2027 are assumed as the base years for extrapolation.

Table 5 forecast input collection

	2025	2026	Source
Yield	Regionals	Regionals	USDA/ previously study
GDP	1.95%	1.75%	The Conference Board
Oil price	$74	$66	EIA
10Y Treasury	4.15%	4.12%	Bloomberg

Note: Economic forecasts are subject to continuous updates.

The input forecast data are collected from several trustable source in Table 5. The GDP projections are based on The Conference Board estimates, while oil price projections are derived from EIA data, and the 10-year Treasury yield forecast is sourced from Bloomberg. By integrating the model with the previously quantified yield uncertainty and yield data from USDA, this study derives the rent estimates based on mean value and 68% CI.

Table 6 Average forecast results of 2026 and 2027

	Mean	Mean of 68% Lower	Mean of 68% Upper
2026	$250	$226	$269
2027	$276	$254	$297
Growth Rate	10.40%	12.39%	10.41%

The projected rental results for 2026 and 2027, as presented in Table 4 and Figure 35, indicate a clear upward trajectory, with the rental mean value rising from $250 in 2026 to $276 in 2027, representing a growth rate of approximately 10.4%. Given the scenario assumptions—including a decline in oil prices, reduced volatility in the 10-year US

Treasury yields, and stable agricultural yield—this increase in rental prices primarily reflects the cumulative effects of rising GDP and decreasing Treasury yields, indicative of inflation-driven rent growth.

Notably, the lower bound of the 68% confidence interval demonstrates a greater growth rate (12.39%) than the mean. This anomaly arises due to the inherent skewness in yield uncertainty distributions: although yield uncertainty appears normally distributed on the standardised scale, it exhibits distinct skewness in absolute terms, the result is collected from prior research (chapter 4). Consequently, the prediction intervals become asymmetric, with a narrower lower bound and an upper bound closely aligned with the mean. Such asymmetric intervals are typical in agricultural and rental markets, where constraints such as environmental conditions, government subsidies, and market mechanisms limit downside volatility, while favourable weather or market dynamics provide comparatively greater flexibility to the upside.

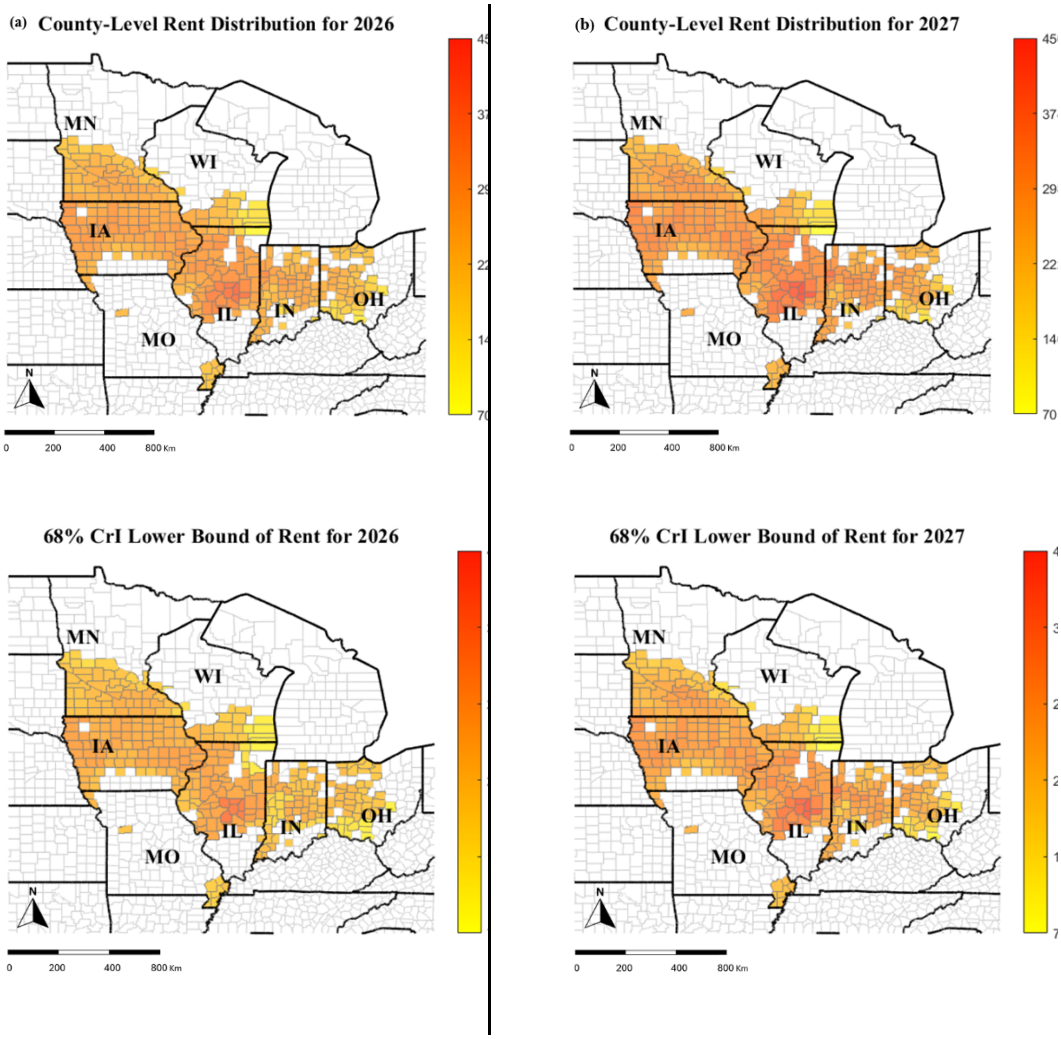
In this context, the significant increase observed in the lower bound for 2027 indicates that the worst-case rental scenario has notably improved, reducing downside risk and elevating the baseline expectations of market. The smaller change in the upper bound, meanwhile, suggests that potential upside movements have already been adequately priced by the market. Figure 35 illustrates the spatial distribution of the mean farmland rental prices along with the 68% probability interval. Overall, these results point towards enhanced market stability and maturity, reflecting more effective risk management within the agricultural rental sector.

5.6.1 Limitation

While this study provides valuable insights, several limitations should be noted. Firstly, the model was trained using a relatively small dataset consisting of multi-year data from only 332 counties. This limited spatial and temporal coverage may affect the generalisability of the results, particularly when applied to broader regions or under different economic contexts.

Secondly, the model incorporates forward-looking economic indicators, such as GDP, oil prices, and interest rates, which are inherently uncertain and subject to market volatility. The predicted rent cost is therefore not fixed, but heavily dependent on future economic

trajectories—many of which are shaped by investment bank forecasts and macroeconomic assumptions. This introduces a level of uncertainty into the outputs, which should be considered when interpreting the results or applying them for policy and planning purposes.



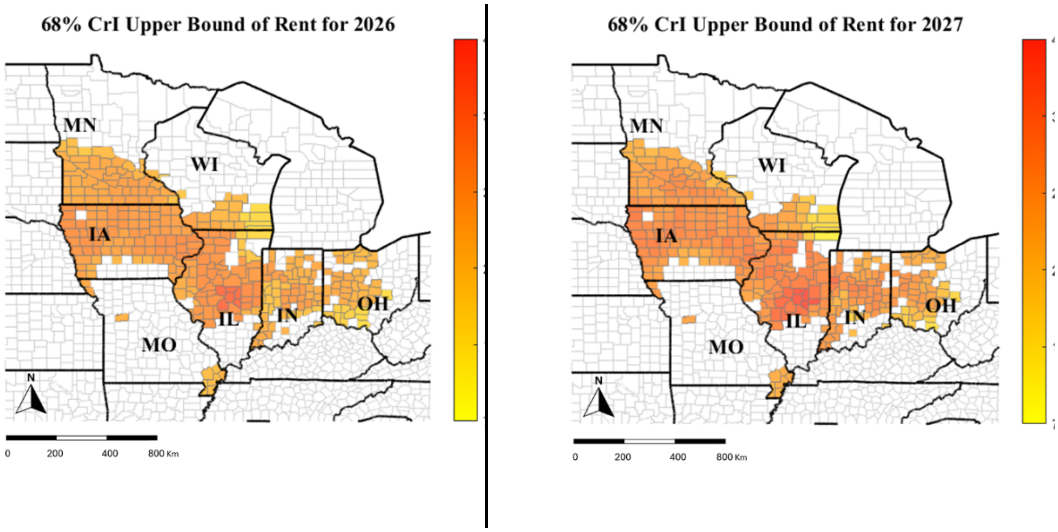


Figure 35 County-level rental cost for (a) 2026 and (b) 2027

5.7 Conclusion

This study explores the regional dynamics of farmland rental prices by integrating macroeconomic indicators, commodity prices, and agricultural data within a geographically weighted regression (GWR) framework. The findings highlight the spatial heterogeneity in rental price determinants, with key variables such as maize yield, oil prices, and GDP exhibiting distinct regional sensitivities. The results indicate that rental price uncertainty does not follow a clear geographical pattern, suggesting that economic and financial factors play a more dominant role than localised spatial characteristics. Additionally, the model validation process confirms the reliability and robustness of the proposed approach, with strong and consistent predictive performance across different datasets (training, validation and test sets), reinforcing its applicability for real world agricultural analysis.

By leveraging high-quality, readily available economic indicators, this research provides a practical framework for assessing farmland rental trends and uncertainty. The insights gained from this study contribute to improved risk assessment and financial planning in agricultural investments, offering valuable guidance for policymakers, investors, and stakeholders navigating farmland markets amid shifting economic landscapes and evolving policy frameworks. Future work could further refine the model by incorporating higher-resolution microeconomic data, policy-related variables, and machine learning approaches to enhance predictive accuracy and its applicability for strategic decision-making in agricultural land markets.

6. Chapter 6: Conclusion

6.1 Conclusion

This thesis presents an integrated exploration of soil and agricultural prediction challenges through three distinct yet interrelated applications: predicting Soil Organic Carbon (SOC), modelling maize yield, and estimating farmland rental prices. Last 2 studies applies advanced machine learning and spatial statistical methods to address uncertainties inherent in soil and agricultural data, and together they demonstrate the value of combining predictive accuracy with uncertainty quantification to support data-driven land management and investment decisions.

The SOC modelling study highlights the potential of data-driven methods for estimating SOC using commonly available soil attributes, particularly under scenarios where nitrogen data is absent. It emphasises the trade-off between prediction accuracy and uncertainty estimation, with variational Bayesian neural networks (VBNNs) proving effective in balancing both, especially in data-limited conditions. This offers a practical approach to SOC prediction in cases where complete soil datasets are unavailable, thereby enhancing the generalisability and applicability of SOC estimation models.

The maize yield prediction study applies spatially aware and probabilistic models across the U.S. Corn Belt, demonstrating the superiority of VB-GWR in capturing spatial heterogeneity and providing robust uncertainty estimates. The findings underscore the relative influence of key soil and climate variables, with pH identified as the most significant predictor and precipitation playing a limited role. Sensitivity and geographic analyses reveal how local yield variation and model confidence are spatially structured, offering actionable insights for land valuation and regional crop planning.

The third study examines farmland rental prices through a spatial economic lens, integrating macroeconomic variables within a GWR framework. The analysis confirms regional variation in the influence of yield, GDP, oil prices, and interest rates on rental values, while also showing that the uncertainty in predicted rents is more closely linked to economic variability than to geographic patterns. This component not only contributes to understanding land market dynamics but also illustrates how forward-looking

economic indicators can be incorporated into spatial modelling for policy and investment purposes.

Together, these studies contribute to the growing body of research at the intersection of soil science, spatial analysis, and machine learning by demonstrating scalable methods for prediction and uncertainty assessment. The combined findings support more informed decision-making in land use, agricultural planning, and investment risk assessment.

6.2 **Limitations**

While this thesis offers meaningful insights into SOC prediction, maize yield modelling, and farmland rent estimation, several limitations should be acknowledged.

First, the overall performance of the models is constrained by limited data availability, particularly in spatial and temporal dimensions. For SOC and farmland rent prediction, the models were trained on relatively small datasets, with only 332 counties included in the rent study and varying temporal coverage across regions. In the yield modelling task, although 842 counties were covered, not all had complete 10-year records. As a result, model generalisability to broader geographies or long-term projections remains uncertain.

Second, increasing the number of input features significantly reduced the usable data due to missing values. This trade-off between model complexity and data availability limited training capacity and increased the risk of overfitting, particularly for data-intensive machine learning models like neural networks. Future work should further explore this trade-off to achieve optimal model robustness in low-data environments.

Third, the study encountered spatial data resolution mismatches. Some inputs, such as GDP or oil prices, were collected at the state or national level, whereas outcomes like maize yield and rent were available at the county level. These inconsistencies may introduce structural errors in spatial analyses, limiting the accuracy of geographically localised predictions.

Fourth, although the study incorporated a wide range of input variables, it did not assess multicollinearity or potential redundancy among features. Variables such as pH, CEC, and NDVI may share overlapping information, which can inflate model complexity without improving accuracy.

Fifth, the models are purely data-driven, with no underlying causal framework. While predictive performance is prioritised, the absence of causal reasoning reduces interpretability, especially in economic and policy-related contexts where understanding variable interactions is critical.

Sixth, forward-looking economic indicators used in the rent prediction model, such as GDP, oil prices, and Treasury yields, are inherently volatile and subject to global market and policy shifts. These variables are often influenced by investment bank forecasts and sentiment-driven factors, making model outputs highly sensitive to short-term economic fluctuations and challenging to use for long-term planning without adjustment.

Seventh, due to data limitations, important agronomic management practices—such as irrigation and fertiliser use—were not included. Their exclusion may limit model completeness, particularly in regions where such practices significantly influence yields or soil carbon levels.

Finally, the models exhibited reduced accuracy in predicting extreme or rare events, such as droughts, floods, and other climate anomalies. These events can strongly affect agricultural systems but are difficult to capture with average-based or historically focused approaches.

Collectively, these limitations highlight areas for future refinement in data acquisition, spatial integration, model interpretability, and the treatment of uncertainty in real-world agricultural modelling applications.

6.3 **Future Work**

Building upon the findings and limitations outlined in this thesis, several avenues for future research are recommended to improve model robustness, enhance practical relevance, and expand the applicability of predictive frameworks in soil and agricultural modelling.

First, future work should focus on expanding the volume and coverage of data, both spatially and temporally. The performance and generalisability of machine learning models, particularly those relying on high-dimensional inputs, are currently constrained by missing values and limited regional representation. Collaborations with national soil

databases, remote sensing platforms, and agricultural monitoring systems may help in establishing more comprehensive datasets suitable for predictive analysis across broader contexts.

Second, addressing spatial resolution inconsistencies remains a key priority. As some input variables (e.g., GDP, oil prices) are provided at state or national levels, while others (e.g., yield, rent) are available at the county level, the integration of multi-scale data introduces structural uncertainty. Future research could explore hierarchical or multi-level modelling techniques to bridge these gaps, thereby improving spatial consistency and predictive accuracy.

Third, while this thesis focused on data-driven approaches, there is a need to explore the integration of causal inference frameworks, particularly in economic and policy-related modelling. Understanding the directionality and interaction among variables can improve model interpretability and allow the outputs to be more effectively incorporated into decision-making processes.

Fourth, future models should incorporate climate variability and extreme events, such as droughts, floods, and heatwaves, which are becoming increasingly frequent due to climate change. Scenario-based modelling and the inclusion of indices like ENSO, SPI, or soil moisture anomalies may enhance model responsiveness and risk assessment capability.

Fifth, the treatment of uncertainty can be further refined. Although this study applied standard error-based intervals and variational Bayesian methods, future work may consider ensemble modelling, Monte Carlo dropout, or hybrid Bayesian-deep learning architectures to provide more robust uncertainty quantification, particularly in under-sampled regions or during rare events.

Finally, the proposed frameworks hold promise for application in broader agricultural and environmental settings. Future studies could adapt the methodologies to different crops, regions, or socio-economic conditions, and integrate additional factors such as fertiliser use, irrigation efficiency, or land-use change. In doing so, predictive tools can better support land valuation, resource allocation, and sustainable agricultural planning across diverse and evolving landscapes.

References

Abdalla, O., Walker, C., & Ishimori, K. (2024). R-code for calculating fluctuation assay results and 95% confidence intervals based on Ma–Sandri–Sarkar Maximum Likelihood. *Software Impacts*, *21*. https://doi.org/10.1016/j.simpa.2024.100661

Acal, C., Maldonado, D., Aguilera, A. M., Zhu, K., Lanza, M., & Roldan, J. B. (2023). Holistic Variability Analysis in Resistive Switching Memories Using a Two-Dimensional Variability Coefficient. *ACS Appl Mater Interfaces*, *15*(15), 19102-19110. https://doi.org/10.1021/acsami.2c22617

Acharjee, D., Mallik, N., Das, D., Aktar, M., & Majumdar, P. (2023). Crop Yield and Soil Moisture Prediction Using Machine Learning Algorithms. In *Machine Learning Applications* (pp. 183-194). https://doi.org/10.1002/9781394173358.ch11

Ait Sidhoum, A. (2023). Measuring farm productivity under production uncertainty. *Australian Journal of Agricultural and Resource Economics*, *67*(4), 672-687. https://doi.org/10.1111/1467-8489.12520

AlThuwaynee, O. F., Kim, S. W., Najemaden, M. A., Aydda, A., Balogun, A. L., Fayyadh, M. M., & Park, H. J. (2021). Demystifying uncertainty in PM10 susceptibility mapping using variable drop-off in extreme-gradient boosting (XGB) and random forest (RF) algorithms. *Environ Sci Pollut Res Int*, *28*(32), 43544-43566. https://doi.org/10.1007/s11356-021-13255-4

Annan, K., Fausti, S. W., Van der Sluis, E., & Kolady, D. E. (2024). Corn Acreage Intensification Levels in U.S. Corn Belt States. *Journal of Agricultural and Applied Economics*, *56*(3), 353-372. https://doi.org/10.1017/aae.2024.14

Arata, L., Donati, M., Sckokai, P., & Arfini, F. (2017). Incorporating risk in a positive mathematical programming framework: a dual approach. *Australian Journal of Agricultural and Resource Economics*, *61*(2), 265-284. https://doi.org/10.1111/1467-8489.12199

Archer, D. W., & Reicosky, D. C. (2009). Economic Performance of Alternative Tillage Systems in the Northern Corn Belt. *Agronomy Journal*, *101*(2), 296-304. https://doi.org/10.2134/agronj2008.0090x

Arfaoui, M., & Ben Rejeb, A. (2017). Oil, gold, US dollar and stock market interdependencies: a global analytical insight. *European Journal of Management and Business Economics*, *26*(3), 278-293. https://doi.org/10.1108/ejmbe-10-2017-016

Baldos, U. L. C., Hertel, T. W., & Moore, F. C. (2019). Understanding the Spatial Distribution of Welfare Impacts of Global Warming on Agriculture and its Drivers. *American Journal of Agricultural Economics*, *101*(5), 1455-1472. https://doi.org/10.1093/ajae/aaz027

Basnet, S. K., Jansson, T., & Heckelei, T. (2021). A Bayesian econometrics and risk programming approach for analysing the impact of decoupled payments in the European Union*. *Australian Journal of Agricultural and Resource Economics*, *65*(3), 729-759. https://doi.org/10.1111/1467-8489.12430

Batjes, N. H., Ribeiro, E., & van Oostrum, A. (2020). Standardised soil profile data to support global mapping and modelling (WoSIS snapshot 2019). *Earth System Science Data*, *12*(1), 299-320. https://doi.org/10.5194/essd-12-299-2020

Bauer, A., & Black, A. L. (1981). Soil Carbon, Nitrogen, and Bulk Density Comparisons in Two Cropland Tillage Systems after 25 Years and in Virgin Grassland. *Soil*

Science Society of America Journal, *45*(6), 1166-1170. https://doi.org/10.2136/sssaj1981.03615995004500060032x

Baum, M. E., Licht, M. A., Huber, I., & Archontoulis, S. V. (2020). Impacts of climate change on the optimum planting date of different maize cultivars in the central US Corn Belt. *European Journal of Agronomy*, *119*. https://doi.org/10.1016/j.eja.2020.126101

Beckman, J., & Riche, S. (2015). Changes to the Natural Gas, Corn, and Fertilizer Price Relationships from the Biofuels Era. *Journal of Agricultural and Applied Economics*, *47*(4), 494-509. https://doi.org/10.1017/aae.2015.22

Bell, M. A., & van Keulen, H. (1995). Soil Pedotransfer Functions for Four Mexican Soils. *Soil Science Society of America Journal*, *59*(3), 865-871. https://doi.org/10.2136/sssaj1995.03615995005900030034x

Bouchakour, R., & Saad, M. (2019). Farm and farmer characteristics and off-farm work: evidence from Algeria. *Australian Journal of Agricultural and Resource Economics*, *64*(2), 455-476. https://doi.org/10.1111/1467-8489.12349

Boyer, C. N., Larson, J. A., Roberts, R. K., McClure, M. A., Tyler, D. D., & Smith, S. A. (2015). Effects of Recent Corn and Energy Prices on Irrigation Investment in the Humid Climate of Tennessee. *Journal of Agricultural and Applied Economics*, *47*(1), 105-122. https://doi.org/10.1017/aae.2014.4

Brock, C., Jackson-Smith, D., Kumarappan, S., Culman, S., Herms, C., & Doohan, D. (2021). Organic Corn Production Practices and Profitability in the Eastern U.S. Corn Belt. *Sustainability*, *13*(16). https://doi.org/10.3390/su13168682

Burke, M., & Lobell, D. B. (2017). Satellite-based assessment of yield variation and its determinants in smallholder African systems. *Proc Natl Acad Sci U S A*, *114*(9), 2189-2194. https://doi.org/10.1073/pnas.1616919114

Cambron, T. W., Deines, J. M., Lopez, B., Patel, R., Liang, S.-Z., & Lobell, D. B. (2024). Further adoption of conservation tillage can increase maize yields in the western US Corn Belt. *Environmental Research Letters*, *19*(5). https://doi.org/10.1088/1748-9326/ad3f32

Carter, M. R. (2002). Soil Quality for Sustainable Land Management. *Agronomy Journal*, *94*(1), 38-47. https://doi.org/10.2134/agronj2002.3800

Chalkias, C., Polykretis, C., Karymbalis, E., Soldati, M., Ghinoi, A., & Ferentinou, M. (2020). Exploring spatial non-stationarity in the relationships between landslide susceptibility and conditioning factors: a local modeling approach using geographically weighted regression. *Bulletin of Engineering Geology and the Environment*, *79*(6), 2799-2814. https://doi.org/10.1007/s10064-020-01733-x

Chen, S., Chen, Z., Zhang, X., Luo, Z., Schillaci, C., Arrouays, D., Richer-de-Forges, A. C., & Shi, Z. (2024). European topsoil bulk density and organic carbon stock database (0–20 cm) using machine-learning-based pedotransfer functions. *Earth System Science Data*, *16*(5), 2367-2383. https://doi.org/10.5194/essd-16-2367-2024

Chen, X., Zhao, Y., & Liu, C. (2022). Medical image segmentation using scalable functional variational Bayesian neural networks with Gaussian processes. *Neurocomputing*, *500*, 58-72. https://doi.org/10.1016/j.neucom.2022.05.055

Coeurdacier, N., & Rey, H. (2013). Home Bias in Open Economy Financial Macroeconomics. *Journal of Economic Literature*, *51*(1), 63-115. https://doi.org/10.1257/jel.51.1.63

Conant, R. T., Ogle, S. M., Paul, E. A., & Paustian, K. (2010). Measuring and monitoring soil organic carbon stocks in agricultural lands for climate mitigation. *Frontiers in Ecology and the Environment*, *9*(3), 169-173. https://doi.org/10.1890/090153

Contreras, L.-F., & Brown, E. T. (2019). Slope reliability and back analysis of failure with geotechnical parameters estimated using Bayesian inference. *Journal of Rock Mechanics and Geotechnical Engineering*, *11*(3), 628-643. https://doi.org/10.1016/j.jrmge.2018.11.008

Contreras, L. F., Brown, E. T., & Ruest, M. (2018). Bayesian data analysis to quantify the uncertainty of intact rock strength. *Journal of Rock Mechanics and Geotechnical Engineering*, *10*(1), 11-31. https://doi.org/10.1016/j.jrmge.2017.07.008

Cui, Y. (2022). Soil–atmosphere interaction in earth structures. *Journal of Rock Mechanics and Geotechnical Engineering*, *14*(1), 35-49. https://doi.org/10.1016/j.jrmge.2021.11.004

Cushing, L., Blaustein-Rejto, D., Wander, M., Pastor, M., Sadd, J., Zhu, A., & Morello-Frosch, R. (2018). Carbon trading, co-pollutants, and environmental equity: Evidence from California's cap-and-trade program (2011-2015). *PLoS Med*, *15*(7), e1002604. https://doi.org/10.1371/journal.pmed.1002604

Davidson, E. A., & Janssens, I. A. (2006). Temperature sensitivity of soil carbon decomposition and feedbacks to climate change. *Nature*, *440*(7081), 165-173. https://doi.org/10.1038/nature04514

Deines, J. M., Patel, R., Liang, S.-Z., Dado, W., & Lobell, D. B. (2021). A million kernels of truth: Insights into scalable satellite maize yield mapping and yield gap analysis from an extensive ground dataset in the US Corn Belt. *Remote Sensing of Environment*, *253*. https://doi.org/10.1016/j.rse.2020.112174

Dhoubhadel, S. P., Azzam, A. M., & Stockton, M. C. (2015). The Impact of Biofuels Policy and Drought on the U.S. Grain and Livestock Markets. *Journal of Agricultural and Applied Economics*, *47*(1), 77-103. https://doi.org/10.1017/aae.2014.6

Ditzhaus, M., Fried, R., & Pauly, M. (2021). QANOVA: quantile-based permutation methods for general factorial designs. *Test*, *30*(4), 960-979. https://doi.org/10.1007/s11749-021-00758-y

Eisfelder, C., Asam, S., Hirner, A., Reiners, P., Holzwarth, S., Bachmann, M., Gessner, U., Dietz, A., Huth, J., Bachofer, F., & Kuenzer, C. (2023). Seasonal Vegetation Trends for Europe over 30 Years from a Novel Normalised Difference Vegetation Index (NDVI) Time-Series—The TIMELINE NDVI Product. *Remote Sensing*, *15*(14). https://doi.org/10.3390/rs15143616

El Bilali, A., & Taleb, A. (2020). Prediction of irrigation water quality parameters using machine learning models in a semi-arid environment. *Journal of the Saudi Society of Agricultural Sciences*, *19*(7), 439-451. https://doi.org/10.1016/j.jssas.2020.08.001

Emadi, M., Taghizadeh-Mehrjardi, R., Cherati, A., Danesh, M., Mosavi, A., & Scholten, T. (2020). Predicting and Mapping of Soil Organic Carbon Using Machine Learning Algorithms in Northern Iran. *Remote Sensing*, *12*(14). https://doi.org/10.3390/rs12142234

Es-haghi, M. S., Rezania, M., & Bagheri, M. (2023). Machine learning-based estimation of soil's true air-entry value from GSD curves. *Gondwana Research*, *123*, 280-292. https://doi.org/10.1016/j.gr.2022.06.012

Fallah Mortezanejad, S. A., & Mohammad-Djafari, A. (2024). Variational Bayesian Approximation (VBA): Implementation and Comparison of Different Optimization Algorithms. *Entropy (Basel)*, *26*(8). https://doi.org/10.3390/e26080707

Fotheringham, A. S., Charlton, M. E., & Brunsdon, C. (2001). Spatial Variations in School Performance: A Local Analysis Using Geographically Weighted Regression. *Geographical and Environmental Modelling*, *5*(1), 43-66. https://doi.org/10.1080/13615930120032617

Gaffney, J., Schussler, J., Löffler, C., Cai, W., Paszkiewicz, S., Messina, C., Groeteke, J., Keaschall, J., & Cooper, M. (2015). Industry-Scale Evaluation of Maize Hybrids Selected for Increased Yield in Drought-Stress Conditions of the US Corn Belt. *Crop Science*, *55*(4), 1608-1618. https://doi.org/10.2135/cropsci2014.09.0654

Garanayak, M., Sahu, G., Mohanty, S. N., & Jagadev, A. K. (2021). Agricultural Recommendation System for Crops Using Different Machine Learning Regression Methods. *International Journal of Agricultural and Environmental Information Systems*, *12*(1), 1-20. https://doi.org/10.4018/IJAEIS.20210101.oa1

Garg, A., Huang, H., Cai, W., Reddy, N. G., Chen, P., Han, Y., Kamchoom, V., Gaurav, S., & Zhu, H.-H. (2021). Influence of soil density on gas permeability and water retention in soils amended with in-house produced biochar. *Journal of Rock Mechanics and Geotechnical Engineering*, *13*(3), 593-602. https://doi.org/10.1016/j.jrmge.2020.10.007

Hao, P., Wang, L., Zhan, Y., & Niu, Z. (2016). Using Moderate-Resolution Temporal NDVI Profiles for High-Resolution Crop Mapping in Years of Absent Ground Reference Data: A Case Study of Bole and Manas Counties in Xinjiang, China. *ISPRS International Journal of Geo-Information*, *5*(5). https://doi.org/10.3390/ijgi5050067

He, X., Cai, G., & Sheng, D. (2025). Indirect models for SWCC parameters: reducing prediction uncertainty with machine learning. *Computers and Geotechnics*, *177*. https://doi.org/10.1016/j.compgeo.2024.106823

Hendricks, N. P., Sinnathamby, S., Douglas-Mankin, K., Smith, A., Sumner, D. A., & Earnhart, D. H. (2014). The environmental effects of crop price increases: Nitrogen losses in the U.S. Corn Belt. *Journal of Environmental Economics and Management*, *68*(3), 507-526. https://doi.org/10.1016/j.jeem.2014.09.002

Hengl, T., Mendes de Jesus, J., Heuvelink, G. B., Ruiperez Gonzalez, M., Kilibarda, M., Blagotic, A., Shangguan, W., Wright, M. N., Geng, X., Bauer-Marschallinger, B., Guevara, M. A., Vargas, R., MacMillan, R. A., Batjes, N. H., Leenaars, J. G., Ribeiro, E., Wheeler, I., Mantel, S., & Kempen, B. (2017). SoilGrids250m: Global gridded soil information based on machine learning. *PLoS One*, *12*(2), e0169748. https://doi.org/10.1371/journal.pone.0169748

Hertel, T. W., Golub, A. A., Jones, A. D., O'Hare, M., Plevin, R. J., & Kammen, D. M. (2010). Effects of US Maize Ethanol on Global Land Use and Greenhouse Gas Emissions: Estimating Market-mediated Responses. *BioScience*, *60*(3), 223-231. https://doi.org/10.1525/bio.2010.60.3.8

Hollis, J. M., Hannam, J., & Bellamy, P. H. (2011). Empirically-derived pedotransfer functions for predicting bulk density in European soils. *European Journal of Soil Science*, *63*(1), 96-109. https://doi.org/10.1111/j.1365-2389.2011.01412.x

Jhajharia, K., & Mathur, P. (2022). Machine Learning Approaches to Predict Crop Yield Using Integrated Satellite and Climate Data. *International Journal of Ambient Computing and Intelligence*, *13*(1), 1-17. https://doi.org/10.4018/ijaci.300799

Jiang, X., Deng, S., Li, H., & Zuo, H. (2022). Characterization of 3D pore nanostructure and stress-dependent permeability of organic-rich shales in northern Guizhou Depression, China. *Journal of Rock Mechanics and Geotechnical Engineering*, *14*(2), 407-422. https://doi.org/10.1016/j.jrmge.2021.08.019

Jiang, Z., Dai, Y., Luo, X., Liu, G., Wang, H., Zheng, H., & Wang, Z. (2017). Assessment of bioenergy development potential and its environmental impact for rural household energy consumption: A case study in Shandong, China. *Renewable and Sustainable Energy Reviews*, *67*, 1153-1161. https://doi.org/10.1016/j.rser.2016.09.085

Jin, X. B., Gong, W. T., Kong, J. L., Bai, Y. T., & Su, T. L. (2022). A Variational Bayesian Deep Network with Data Self-Screening Layer for Massive Time-Series Data Forecasting. *Entropy (Basel)*, *24*(3). https://doi.org/10.3390/e24030335

Jithitikulchai, T., McCarl, B. A., & Wu, X. (2018). Decadal Climate Variability Impacts on Climate and Crop Yields. *Journal of Agricultural and Applied Economics*, *51*(1), 104-125. https://doi.org/10.1017/aae.2018.25

Johnson, D. M. (2016). A comprehensive assessment of the correlations between field crop yields and commonly used MODIS products. *International Journal of Applied Earth Observation and Geoinformation*, *52*, 65-81. https://doi.org/10.1016/j.jag.2016.05.010

Joshi, V. R., Kazula, M. J., Coulter, J. A., Naeve, S. L., & Garcia, Y. G. A. (2021). In-season weather data provide reliable yield estimates of maize and soybean in the US central Corn Belt. *Int J Biometeorol*, *65*(4), 489-502. https://doi.org/10.1007/s00484-020-02039-z

Ju, Y., Wang, G., Bu, H., Li, Q., & Yan, Z. (2014). China organic-rich shale geologic features and special shale gas production issues. *Journal of Rock Mechanics and Geotechnical Engineering*, *6*(3), 196-207. https://doi.org/10.1016/j.jrmge.2014.03.002

Kanwisher, N., Khosla, M., & Dobs, K. (2023). Using artificial neural networks to ask 'why' questions of minds and brains. *Trends Neurosci*, *46*(3), 240-254. https://doi.org/10.1016/j.tins.2022.12.008

Key, N. (2019). Farm size and productivity growth in the United States Corn Belt. *Food Policy*, *84*, 186-195. https://doi.org/10.1016/j.foodpol.2018.03.017

Khan, M. S., Semwal, M., Sharma, A., & Verma, R. K. (2019). An artificial neural network model for estimating Mentha crop biomass yield using Landsat 8 OLI. *Precision Agriculture*, *21*(1), 18-33. https://doi.org/10.1007/s11119-019-09655-9

Kingwell, R. S., & Xayavong, V. (2016). How drought affects the financial characteristics of Australian farm businesses. *Australian Journal of Agricultural and Resource Economics*, *61*(3), 344-366. https://doi.org/10.1111/1467-8489.12195

Kusmec, A., & Schnable, P. S. (2024). Phenological Adaptation Is Insufficient to Offset Climate Change-Induced Yield Losses in US Hybrid Maize. *Glob Chang Biol*, *30*(10), e17539. https://doi.org/10.1111/gcb.17539

Lal, R. (2020). Soil organic matter and water retention. *Agronomy Journal*, *112*(5), 3265-3277. https://doi.org/10.1002/agj2.20282

Le, T. D., Noumeir, R., Quach, H. L., Kim, J. H., Kim, J. H., & Kim, H. M. (2020). Critical Temperature Prediction for a Superconductor: A Variational Bayesian Neural

Network Approach. *IEEE Transactions on Applied Superconductivity*, *30*(4), 1-5. https://doi.org/10.1109/tasc.2020.2971456
Le, T. T. H. (2016). Effects of Climate Change on Rice Yield and Rice Market in Vietnam. *Journal of Agricultural and Applied Economics*, *48*(4), 366-382. https://doi.org/10.1017/aae.2016.21
Leng, G. (2019). Uncertainty in Assessing Temperature Impact on U.S. Maize Yield Under Global Warming: The Role of Compounding Precipitation Effect. *Journal of Geophysical Research: Atmospheres*, *124*(12), 6238-6246. https://doi.org/10.1029/2018jd029996
Leng, G. (2021). Maize yield loss risk under droughts in observations and crop models in the United States. *Environmental Research Letters*, *16*(2). https://doi.org/10.1088/1748-9326/abd500
Li, T., Cui, L., Kuhnert, M., McLaren, T. I., Pandey, R., Liu, H., Wang, W., Xu, Z., Xia, A., Dalal, R. C., & Dang, Y. P. (2024). A comprehensive review of soil organic carbon estimates: Integrating remote sensing and machine learning technologies. *Journal of Soils and Sediments*, *24*(11), 3556-3571. https://doi.org/10.1007/s11368-024-03913-8
Lin, L., & Liu, X. (2022). Mixture-based weight learning improves the random forest method for hyperspectral estimation of soil total nitrogen. *Computers and Electronics in Agriculture*, *192*. https://doi.org/10.1016/j.compag.2021.106634
Liu, G., Tian, S., Xu, G., Zhang, C., & Cai, M. (2023). Combination of effective color information and machine learning for rapid prediction of soil water content. *Journal of Rock Mechanics and Geotechnical Engineering*, *15*(9), 2441-2457. https://doi.org/10.1016/j.jrmge.2022.12.029
Lobell, D. B., & Burke, M. B. (2010). On the use of statistical models to predict crop yield responses to climate change. *Agricultural and Forest Meteorology*, *150*(11), 1443-1452. https://doi.org/10.1016/j.agrformet.2010.07.008
Lobell, D. B., Cassman, K. G., & Field, C. B. (2009). Crop Yield Gaps: Their Importance, Magnitudes, and Causes. *Annual Review of Environment and Resources*, *34*(1), 179-204. https://doi.org/10.1146/annurev.environ.041008.093740
Lobell, D. B., Thau, D., Seifert, C., Engle, E., & Little, B. (2015). A scalable satellite-based crop yield mapper. *Remote Sensing of Environment*, *164*, 324-333. https://doi.org/10.1016/j.rse.2015.04.021
Longbottom, T. L., Townsend-Small, A., Owen, L. A., & Murari, M. K. (2014). Climatic and topographic controls on soil organic matter storage and dynamics in the Indian Himalaya: Potential carbon cycle–climate change feedbacks. *Catena*, *119*, 125-135. https://doi.org/10.1016/j.catena.2014.03.002
Ma, W., Renwick, A., & Grafton, Q. (2018). Farm machinery use, off-farm employment and farm performance in China. *Australian Journal of Agricultural and Resource Economics*, *62*(2), 279-298. https://doi.org/10.1111/1467-8489.12249
Makovníková, J., Širáň, M., Houšková, B., Pálka, B., & Jones, A. (2017). Comparison of different models for predicting soil bulk density. Case study – Slovakian agricultural soils. *International Agrophysics*, *31*(4), 491-498. https://doi.org/10.1515/intag-2016-0079
Maltais-Landry, G., & Lobell, D. B. (2012). Evaluating the Contribution of Weather to Maize and Wheat Yield Trends in 12 U.S. Counties. *Agronomy Journal*, *104*(2), 301-311. https://doi.org/10.2134/agronj2011.0220

Massigoge, I., Carcedo, A., Lingenfelser, J., Hefley, T., Prasad, P. V. V., Berning, D., Lira, S., Messina, C. D., Rice, C. W., & Ciampitti, I. (2023). Maize planting date and maturity in the US central Great Plains: Exploring windows for maximizing yields. *European Journal of Agronomy*, *149*. https://doi.org/10.1016/j.eja.2023.126905

McCord, M., Davis, P. T., Haran, M., McGreal, S., & McIlhatton, D. (2012). Spatial variation as a determinant of house price. *Journal of Financial Management of Property and Construction*, *17*(1), 49-72. https://doi.org/10.1108/13664381211211046

McNunn, G., Karlen, D. L., Salas, W., Rice, C. W., Mueller, S., Muth, D., & Seale, J. W. (2020). Climate smart agriculture opportunities for mitigating soil greenhouse gas emissions across the U.S. Corn-Belt. *Journal of Cleaner Production*, *268*. https://doi.org/10.1016/j.jclepro.2020.122240

Melendez-Pastor, I., Lopez-Granado, O. M., Navarro-Pedreno, J., Hernandez, E. I., Jordan Vidal, M. M., & Gomez Lucas, I. (2023). Environmental factors influencing DDT-DDE spatial distribution in an agricultural drainage system determined by using machine learning techniques. *Environ Geochem Health*, *45*(12), 9067-9085. https://doi.org/10.1007/s10653-023-01486-y

Miller, V. S., & Naeth, M. A. (2019). Hydrogel and Organic Amendments to Increase Water Retention in Anthroposols for Land Reclamation. *Applied and Environmental Soil Science*, *2019*, 1-11. https://doi.org/10.1155/2019/4768091

Minasny, B., Malone, B. P., McBratney, A. B., Angers, D. A., Arrouays, D., Chambers, A., Chaplot, V., Chen, Z.-S., Cheng, K., Das, B. S., Field, D. J., Gimona, A., Hedley, C. B., Hong, S. Y., Mandal, B., Marchant, B. P., Martin, M., McConkey, B. G., Mulder, V. L., . . . Winowiecki, L. (2017). Soil carbon 4 per mille. *Geoderma*, *292*, 59-86. https://doi.org/10.1016/j.geoderma.2017.01.002

Mishra, A. K., Moss, C. B., & Erickson, K. W. (2009). Regional differences in agricultural profitability, government payments, and farmland values. *Agricultural Finance Review*, *69*(1), 49-66. https://doi.org/10.1108/00021460910960462

Mitchell, R., Frank, E., & Holmes, G. (2022). GPUTreeShap: massively parallel exact calculation of SHAP scores for tree ensembles. *PeerJ Comput Sci*, *8*, e880. https://doi.org/10.7717/peerj-cs.880

Morell, F. J., Yang, H. S., Cassman, K. G., Wart, J. V., Elmore, R. W., Licht, M., Coulter, J. A., Ciampitti, I. A., Pittelkow, C. M., Brouder, S. M., Thomison, P., Lauer, J., Graham, C., Massey, R., & Grassini, P. (2016). Can crop simulation models be used to predict local to regional maize yields and total production in the U.S. Corn Belt? *Field Crops Research*, *192*, 1-12. https://doi.org/10.1016/j.fcr.2016.04.004

Moss, C. B. (1997). Returns, Interest Rates, and Inflation: How They Explain Changes in Farmland Values. *American Journal of Agricultural Economics*, *79*(4), 1311-1318. https://doi.org/10.2307/1244287

Mourtzinis, S., Ortiz, B. V., & Damianidis, D. (2016). Climate Change and ENSO Effects on Southeastern US Climate Patterns and Maize Yield. *Sci Rep*, *6*, 29777. https://doi.org/10.1038/srep29777

Narani, A., Konda, N., Chen, C. S., Tachea, F., Coffman, P., Gardner, J., Li, C., Ray, A. E., Hartley, D. S., Simmons, B., Pray, T. R., & Tanjore, D. (2019). Simultaneous application of predictive model and least cost formulation can substantially benefit biorefineries outside Corn Belt in United States: A case study in Florida. *Bioresour Technol*, *271*, 218-227. https://doi.org/10.1016/j.biortech.2018.09.103

Neelawala, P., Wilson, C., & Athukorala, W. (2012). The impact of mining and smelting activities on property values: a study of Mount Isa city, Queensland, Australia. *Australian Journal of Agricultural and Resource Economics*, *57*(1), 60-78. https://doi.org/10.1111/j.1467-8489.2012.00604.x

Oh, S., & Gramig, B. M. (2023). Valuing Ecosystem Services and Downstream Water Quality Improvement in the U.S. Corn Belt. *Environmental and Resource Economics*, *85*(3-4), 823-872. https://doi.org/10.1007/s10640-023-00784-4

Padarian, J., Minasny, B., & McBratney, A. B. (2020). Machine learning and soil sciences: a review aided by machine learning tools. *Soil*, *6*(1), 35-52. https://doi.org/10.5194/soil-6-35-2020

Panagopoulos, Y., Gassman, P. W., Jha, M. K., Kling, C. L., Campbell, T., Srinivasan, R., White, M., & Arnold, J. G. (2015). A refined regional modeling approach for the Corn Belt – Experiences and recommendations for large-scale integrated modeling. *Journal of Hydrology*, *524*, 348-366. https://doi.org/10.1016/j.jhydrol.2015.02.039

Patton, N. R., Lohse, K. A., Seyfried, M., Will, R., & Benner, S. G. (2019). Lithology and coarse fraction adjusted bulk density estimates for determining total organic carbon stocks in dryland soils. *Geoderma*, *337*, 844-852. https://doi.org/10.1016/j.geoderma.2018.10.036

Paulson, N. D., Schnitkey, G. D., & Sherrick, B. J. (2010). Rental arrangements and risk mitigation of crop insurance and marketing. *Agricultural Finance Review*, *70*(3), 399-413. https://doi.org/10.1108/00021461011088512

Peng, B., Guan, K., Pan, M., & Li, Y. (2018). Benefits of Seasonal Climate Prediction and Satellite Data for Forecasting U.S. Maize Yield. *Geophysical Research Letters*, *45*(18), 9662-9671. https://doi.org/10.1029/2018gl079291

Pham, K., Kim, D., Le, C. V., & Won, J. (2023). Machine learning-based pedotransfer functions to predict soil water characteristics curves. *Transportation Geotechnics*, *42*. https://doi.org/10.1016/j.trgeo.2023.101052

Piepho, H. P. (2023). An adjusted coefficient of determination (R(2)) for generalized linear mixed models in one go. *Biom J*, *65*(7), e2200290. https://doi.org/10.1002/bimj.202200290

Poeplau, C., & Dechow, R. (2023). The legacy of one hundred years of climate change for organic carbon stocks in global agricultural topsoils. *Sci Rep*, *13*(1), 7483. https://doi.org/10.1038/s41598-023-34753-0

Pribyl, D. W. (2010). A critical review of the conventional SOC to SOM conversion factor. *Geoderma*, *156*(3-4), 75-83. https://doi.org/10.1016/j.geoderma.2010.02.003

Qiao, M., He, X., Cheng, X., Li, P., Luo, H., Tian, Z., & Guo, H. (2021). Exploiting Hierarchical Features for Crop Yield Prediction Based on 3-D Convolutional Neural Networks and Multikernel Gaussian Process. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *14*, 4476-4489. https://doi.org/10.1109/jstars.2021.3073149

Qin, Z., Guan, K., Zhou, W., Peng, B., Villamil, M. B., Jin, Z., Tang, J., Grant, R., Gentry, L., Margenot, A. J., Bollero, G., & Li, Z. (2021). Assessing the impacts of cover crops on maize and soybean yield in the U.S. Midwestern agroecosystems. *Field Crops Research*, *273*. https://doi.org/10.1016/j.fcr.2021.108264

Ramsey, A. F., & Rejesus, R. M. (2021). Bayesian Hierarchical Models for Measuring Varietal Improvement in Tobacco Yield and Quality. *Journal of Agricultural and Applied Economics*, *53*(4), 563-586. https://doi.org/10.1017/aae.2021.25

Rawls, W. J., Pachepsky, Y. A., Ritchie, J. C., Sobecki, T. M., & Bloodworth, H. (2003). Effect of soil organic carbon on soil water retention. *Geoderma*, *116*(1-2), 61-76. https://doi.org/10.1016/s0016-7061(03)00094-6

Ribeiro, M. C., & Pereira, M. J. (2018). Modelling local uncertainty in relations between birth weight and air quality within an urban area: combining geographically weighted regression with geostatistical simulation. *Environ Sci Pollut Res Int*, *25*(26), 25942-25954. https://doi.org/10.1007/s11356-018-2614-x

Richer-de-Forges, A. C., Arrouays, D., Poggio, L., Chen, S., Lacoste, M., Minasny, B., Libohova, Z., Roudier, P., Mulder, V. L., NÉDÉLec, H., Martelet, G., Lemercier, B., Lagacherie, P., & Bourennane, H. (2023). Hand-feel soil texture observations to evaluate the accuracy of digital soil maps for local prediction of soil particle size distribution: A case study in Central France. *Pedosphere*, *33*(5), 731-743. https://doi.org/10.1016/j.pedsph.2022.07.009

Rizzo, G., Edreira, J. I. R., Archontoulis, S. V., Yang, H. S., & Grassini, P. (2018). Do shallow water tables contribute to high and stable maize yields in the US Corn Belt? *Global Food Security*, *18*, 27-34. https://doi.org/10.1016/j.gfs.2018.07.002

Roberts, M. J., & Schlenker, W. (2013). Identifying Supply and Demand Elasticities of Agricultural Commodities: Implications for the US Ethanol Mandate. *American Economic Review*, *103*(6), 2265-2295. https://doi.org/10.1257/aer.103.6.2265

Rotundo, J. L., Salinas, A., Gomara, N., Borras, L., & Messina, C. (2024). Maize outyielding sorghum under drought conditions helps explain land use changes in the US. *Field Crops Research*, *308*. https://doi.org/10.1016/j.fcr.2024.109298

Roy, D. P., Wulder, M. A., Loveland, T. R., C.E, W., Allen, R. G., Anderson, M. C., Helder, D., Irons, J. R., Johnson, D. M., Kennedy, R., Scambos, T. A., Schaaf, C. B., Schott, J. R., Sheng, Y., Vermote, E. F., Belward, A. S., Bindschadler, R., Cohen, W. B., Gao, F., . . . Zhu, Z. (2014). Landsat-8: Science and product vision for terrestrial global change research. *Remote Sensing of Environment*, *145*, 154-172. https://doi.org/10.1016/j.rse.2014.02.001

Ruiz, A., Trifunovic, S., Eudy, D. M., Sciarresi, C. S., Baum, M., Danalatos, G. J. N., Elli, E. F., Kalogeropoulos, G., King, K., dos Santos, C., Thies, A., Pico, L. O., Castellano, M. J., Schnable, P. S., Topp, C., Graham, M., Lamkey, K. R., Vyn, T. J., & Archontoulis, S. V. (2023). Harvest index has increased over the last 50 years of maize breeding. *Field Crops Research*, *300*. https://doi.org/10.1016/j.fcr.2023.108991

Rumpel, C., & Kögel-Knabner, I. (2010). Deep soil organic matter—a key but poorly understood component of terrestrial C cycle. *Plant and Soil*, *338*(1-2), 143-158. https://doi.org/10.1007/s11104-010-0391-5

Sadeghi Tabas, S., & Samadi, S. (2022). Variational Bayesian dropout with a Gaussian prior for recurrent neural networks application in rainfall–runoff modeling. *Environmental Research Letters*, *17*(6). https://doi.org/10.1088/1748-9326/ac7247

Sagan, V., Maimaitijiang, M., Bhadra, S., Maimaitiyiming, M., Brown, D. R., Sidike, P., & Fritschi, F. B. (2021). Field-scale crop yield prediction using multi-temporal WorldView-3 and PlanetScope satellite data and deep learning. *ISPRS Journal of Photogrammetry and Remote Sensing*, *174*, 265-281. https://doi.org/10.1016/j.isprsjprs.2021.02.008

Sahoo, S., Russo, T. A., Elliott, J., & Foster, I. (2017). Machine learning algorithms for modeling groundwater level changes in agricultural regions of the U.S. *Water Resources Research*, *53*(5), 3878-3895. https://doi.org/10.1002/2016wr019933

Salmerón-Gómez, R., García-García, C. B., & García-Pérez, J. (2024). A Redefined Variance Inflation Factor: Overcoming the Limitations of the Variance Inflation Factor. *Computational Economics*, *65*(1), 337-363. https://doi.org/10.1007/s10614-024-10575-8

Sarkhani Benemaran, R. (2023). Application of extreme gradient boosting method for evaluating the properties of episodic failure of borehole breakout. *Geoenergy Science and Engineering*, *226*. https://doi.org/10.1016/j.geoen.2023.211837

Sarzaeim, P., & Muñoz-Arriola, F. (2024). A Method to Estimate Climate Drivers of Maize Yield Predictability Leveraging Genetic-by-Environment Interactions in the US and Canada. *Agronomy*, *14*(4). https://doi.org/10.3390/agronomy14040733

Schmidt, M. W., Torn, M. S., Abiven, S., Dittmar, T., Guggenberger, G., Janssens, I. A., Kleber, M., Kogel-Knabner, I., Lehmann, J., Manning, D. A., Nannipieri, P., Rasse, D. P., Weiner, S., & Trumbore, S. E. (2011). Persistence of soil organic matter as an ecosystem property. *Nature*, *478*(7367), 49-56. https://doi.org/10.1038/nature10386

Schwalbert, R., Amado, T., Nieto, L., Corassa, G., Rice, C., Peralta, N., Schauberger, B., Gornott, C., & Ciampitti, I. (2020). Mid-season county-level corn yield forecast for US Corn Belt integrating satellite imagery and weather variables. *Crop Science*, *60*(2), 739-750. https://doi.org/10.1002/csc2.20053

Shahhosseini, M., Hu, G., & Archontoulis, S. V. (2020). Forecasting Corn Yield With Machine Learning Ensembles. *Front Plant Sci*, *11*, 1120. https://doi.org/10.3389/fpls.2020.01120

Shekoofa, A., Emam, Y., Shekoufa, N., Ebrahimi, M., & Ebrahimie, E. (2014). Determining the most important physiological and agronomic traits contributing to maize grain yield through machine learning algorithms: a new avenue in intelligent agriculture. *PLoS One*, *9*(5), e97288. https://doi.org/10.1371/journal.pone.0097288

Shogren, A. J., Zarnetske, J. P., Abbott, B. W., Grose, A. L., Rec, A. F., Nipko, J., Song, C., O'Donnell, J. A., & Bowden, W. B. (2024). Hydrology Controls Dissolved Organic Carbon and Nitrogen Export and Post-Storm Recovery in Two Arctic Headwaters. *Journal of Geophysical Research: Biogeosciences*, *129*(2). https://doi.org/10.1029/2023jg007583

Siewert, M. B. (2018). High-resolution digital mapping of soil organic carbon in permafrost terrain using machine learning: a case study in a sub-Arctic peatland environment. *Biogeosciences*, *15*(6), 1663-1682. https://doi.org/10.5194/bg-15-1663-2018

Smith, P., Haberl, H., Popp, A., Erb, K. H., Lauk, C., Harper, R., Tubiello, F. N., de Siqueira Pinto, A., Jafari, M., Sohi, S., Masera, O., Bottcher, H., Berndes, G., Bustamante, M., Ahammad, H., Clark, H., Dong, H., Elsiddig, E. A., Mbow, C., . . . Rose, S. (2013). How much land-based greenhouse gas mitigation can be achieved without compromising food security and environmental goals? *Glob Chang Biol*, *19*(8), 2285-2302. https://doi.org/10.1111/gcb.12160

Smith, P., Martino, D., Cai, Z., Gwary, D., Janzen, H., Kumar, P., McCarl, B., Ogle, S., O'Mara, F., Rice, C., Scholes, B., Sirotenko, O., Howden, M., McAllister, T., Pan,

G., Romanenkov, V., Schneider, U., Towprayoon, S., Wattenbach, M., & Smith, J. (2008). Greenhouse gas mitigation in agriculture. *Philos Trans R Soc Lond B Biol Sci*, *363*(1492), 789-813. https://doi.org/10.1098/rstb.2007.2184

Snow, V. O., Rotz, C. A., Moore, A. D., Martin-Clouaire, R., Johnson, I. R., Hutchings, N. J., & Eckard, R. J. (2014). The challenges – and some solutions – to process-based modelling of grazed agricultural systems. *Environmental Modelling & Software*, *62*, 420-436. https://doi.org/10.1016/j.envsoft.2014.03.009

Song, G., Li, L., Pan, G., & Zhang, Q. (2005). Topsoil organic carbon storage of China and its loss by cultivation. *Biogeochemistry*, *74*(1), 47-62. https://doi.org/10.1007/s10533-004-2222-3

Ssegane, H., Zumpf, C., Cristina Negri, M., Campbell, P., Heavey, J. P., & Volk, T. A. (2016). The economics of growing shrub willow as a bioenergy buffer on agricultural fields: A case study in the Midwest Corn Belt. *Biofuels, Bioproducts and Biorefining*, *10*(6), 776-789. https://doi.org/10.1002/bbb.1679

Taalab, K., Corstanje, R., Mayr, T. M., Whelan, M. J., & Creamer, R. E. (2015). The application of expert knowledge in Bayesian networks to predict soil bulk density at the landscape scale. *European Journal of Soil Science*, *66*(5), 930-941. https://doi.org/10.1111/ejss.12282

Tang, K., Hailu, A., Kragt, M. E., & Ma, C. (2016). Marginal abatement costs of greenhouse gas emissions: broadacre farming in the Great Southern Region of Western Australia. *Australian Journal of Agricultural and Resource Economics*, *60*(3), 459-475. https://doi.org/10.1111/1467-8489.12135

Tang, K., He, C., Ma, C., & Wang, D. (2019). Does carbon farming provide a cost-effective option to mitigate GHG emissions? Evidence from China. *Australian Journal of Agricultural and Resource Economics*, *63*(3), 575-592. https://doi.org/10.1111/1467-8489.12306

Tessema, Y. M., Asafu-Adjaye, J., & Shiferaw, B. (2018). The impact of conservation tillage on maize yield and input demand: the case of smallholder farmers in north-west Ethiopia. *Australian Journal of Agricultural and Resource Economics*, *62*(4), 636-653. https://doi.org/10.1111/1467-8489.12270

Teste, F., Makowski, D., Bazzi, H., & Ciais, P. (2024). Early forecasting of corn yield and price variations using satellite vegetation products. *Computers and Electronics in Agriculture*, *221*. https://doi.org/10.1016/j.compag.2024.108962

Tiffin, R., & Balcombe, K. (2011). The determinants of technology adoption by UK farmers using Bayesian model averaging: the cases of organic production and computer usage. *Australian Journal of Agricultural and Resource Economics*, *55*(4), 579-598. https://doi.org/10.1111/j.1467-8489.2011.00549.x

Tilman, D., Balzer, C., Hill, J., & Befort, B. L. (2011). Global food demand and the sustainable intensification of agriculture. *Proc Natl Acad Sci U S A*, *108*(50), 20260-20264. https://doi.org/10.1073/pnas.1116437108

Timilsina, G. R., Beghin, J. C., van der Mensbrugghe, D., & Mevel, S. (2012). The impacts of biofuels targets on land-use change and food supply: A global CGE assessment. *Agricultural Economics*, *43*(3), 315-332. https://doi.org/10.1111/j.1574-0862.2012.00585.x

Tollenaar, M., Fridgen, J., Tyagi, P., Stackhouse, P. W., Jr., & Kumudini, S. (2017). The Contribution of Solar Brightening to the US Maize Yield Trend. *Nat Clim Chang*, *7*, 275-278. https://doi.org/10.1038/nclimate3234

Uludere Aragon, N. Z. (2019). Role of land quality in corn acreage response to price and policy changes: evidence from the Western Corn Belt. *Environmental Research Communications*, *1*(6). https://doi.org/10.1088/2515-7620/ab2c3f

Végh, J. (2021). Which scaling rule applies to large artificial neural networks. *Neural Computing and Applications*, *33*(24), 16847-16864. https://doi.org/10.1007/s00521-021-06456-y

Villacis, A. H., Ramsey, A. F., Delgado, J. A., & Alwang, J. R. (2020). Estimating Economically Optimal Levels of Nitrogen Fertilizer in No-Tillage Continuous Corn. *Journal of Agricultural and Applied Economics*, *52*(4), 613-623. https://doi.org/10.1017/aae.2020.23

Wang, F., Wang, L., & Chen, Y. (2018). Lagged multi-affine height correlation analysis for exploring lagged correlations in complex systems. *Chaos*, *28*(6), 061102. https://doi.org/10.1063/1.5030563

Wang, F., Zhang, Z., Wang, M., & Ling, G. (2024). Detrended partial cross-correlation analysis-random matrix theory for denoising network construction. *Applied Intelligence*, *55*(1). https://doi.org/10.1007/s10489-024-05975-0

Wang, H., Shen, M., Hui, D., Chen, J., Sun, G., Wang, X., Lu, C., Sheng, J., Chen, L., Luo, Y., Zheng, J., & Zhang, Y. (2019). Straw incorporation influences soil organic carbon sequestration, greenhouse gas emission, and crop yields in a Chinese rice (Oryza sativa L.) –wheat (Triticum aestivum L.) cropping system. *Soil and Tillage Research*, *195*. https://doi.org/10.1016/j.still.2019.104377

Wang, M., Pande, G. N., Pietruszczak, S., & Zeng, Z. X. (2020). Determination of strain-dependent soil water retention characteristics from gradation curve. *Journal of Rock Mechanics and Geotechnical Engineering*, *12*(6), 1356-1360. https://doi.org/10.1016/j.jrmge.2020.03.005

Wang, T., Jin, H., Fan, Y., Obembe, O., & Li, D. (2021). Farmers' adoption and perceived benefits of diversified crop rotations in the margins of U.S. Corn Belt. *J Environ Manage*, *293*, 112903. https://doi.org/10.1016/j.jenvman.2021.112903

West, S. C., Mugera, A. W., & Kingwell, R. S. (2024). The impact of repayment obligations arising as a by-product of input use on partial inefficiency: Evidence from Western Australian farm businesses. *Australian Journal of Agricultural and Resource Economics*, *68*(3), 678-700. https://doi.org/10.1111/1467-8489.12568

Whetton, R., Zhao, Y., Shaddad, S., & Mouazen, A. M. (2017). Nonlinear parametric modelling to study how soil properties affect crop yields and NDVI. *Computers and Electronics in Agriculture*, *138*, 127-136. https://doi.org/10.1016/j.compag.2017.04.016

Wu, H., Guo, Z., & Peng, C. (2003). Distribution and storage of soil organic carbon in China. *Global Biogeochemical Cycles*, *17*(2). https://doi.org/10.1029/2001gb001844

Wu, X., Shi, J., Zhang, T., Zuo, Q., Wang, L., Xue, X., & Ben-Gal, A. (2022). Crop yield estimation and irrigation scheduling optimization using a root-weighted soil water availability based water production function. *Field Crops Research*, *284*. https://doi.org/10.1016/j.fcr.2022.108579

Xiao, L., Wang, G., Zhou, H., Jin, X., & Luo, Z. (2022). Coupling agricultural system models with machine learning to facilitate regional predictions of management practices and crop production. *Environmental Research Letters*, *17*(11). https://doi.org/10.1088/1748-9326/ac9c71

Xie, W., Ali, T., Cui, Q., & Huang, J. (2017). Economic impacts of commercializing insect-resistant GM maize in China. *China Agricultural Economic Review*, *9*(3), 340-354. https://doi.org/10.1108/caer-06-2017-0126

Xie, W., Huang, G., Fu, W., Shu, B., Cui, B., Li, M., & Yue, F. (2022). A quality control method based on improved IQR for estimating multi-GNSS real-time satellite clock offset. *Measurement*, *201*. https://doi.org/10.1016/j.measurement.2022.111695

Xu, L., He, N. P., Yu, G. R., Wen, D., Gao, Y., & He, H. L. (2015). Differences in pedotransfer functions of bulk density lead to high uncertainty in soil organic carbon estimation at regional scales: Evidence from Chinese terrestrial ecosystems. *Journal of Geophysical Research: Biogeosciences*, *120*(8), 1567-1575. https://doi.org/10.1002/2015jg002929

Xu, Q., Yang, F., Hu, S., He, X., & Hong, Y. (2024). Tree Height–Diameter Model of Natural Coniferous and Broad-Leaved Mixed Forests Based on Random Forest Method and Nonlinear Mixed-Effects Method in Jilin Province, China. *Forests*, *15*(11). https://doi.org/10.3390/f15111922

Xu, X., Shi, Z., Li, D., Rey, A., Ruan, H., Craine, J. M., Liang, J., Zhou, J., & Luo, Y. (2016). Soil properties control decomposition of soil organic carbon: Results from data-assimilation analysis. *Geoderma*, *262*, 235-242. https://doi.org/10.1016/j.geoderma.2015.08.038

Yang, K., Guha, N., Efendiev, Y., & Mallick, B. K. (2017). Bayesian and variational Bayesian approaches for flows in heterogeneous random media. *Journal of Computational Physics*, *345*, 275-293. https://doi.org/10.1016/j.jcp.2017.04.034

Yu, H., Fotheringham, A. S., Li, Z., Oshan, T., Kang, W., & Wolf, L. J. (2019). Inference in Multiscale Geographically Weighted Regression. *Geographical Analysis*, *52*(1), 87-106. https://doi.org/10.1111/gean.12189

Zhang, C., Di, L., Lin, L., & Guo, L. (2019). Machine-learned prediction of annual crop planting in the U.S. Corn Belt based on historical crop planting maps. *Computers and Electronics in Agriculture*, *166*. https://doi.org/10.1016/j.compag.2019.104989

Zhao, H., Jiang, Y., Xiao, Q., Zhang, C., & Behzad, H. M. (2021). Coupled carbon-nitrogen cycling controls the transformation of dissolved inorganic carbon into dissolved organic carbon in karst aquatic systems. *Journal of Hydrology*, *592*. https://doi.org/10.1016/j.jhydrol.2020.125764

Zhao, H., Zhang, L., Wan, N., Avenson, T. J., Welch, S. M., & Lin, X. (2024). Sensitivity changes of US maize yields to extreme heat through timely precipitation patterns. *Environmental Research Communications*, *6*(7). https://doi.org/10.1088/2515-7620/ad6404

Zilberman, D., Hochman, G., Rajagopal, D., Sexton, S., & Timilsina, G. (2012). The Impact of Biofuels on Commodity Food Prices: Assessment of Findings. *American Journal of Agricultural Economics*, *95*(2), 275-281. https://doi.org/10.1093/ajae/aas037