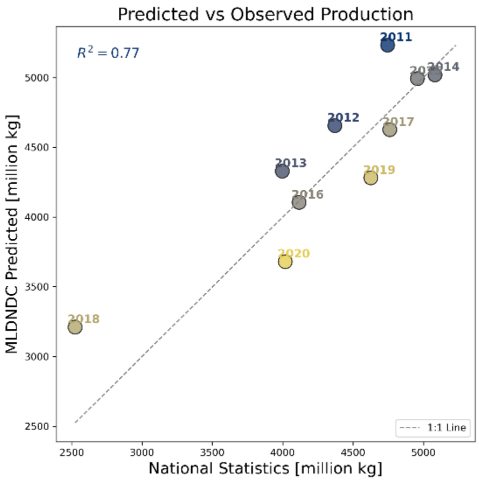
Supplementary information

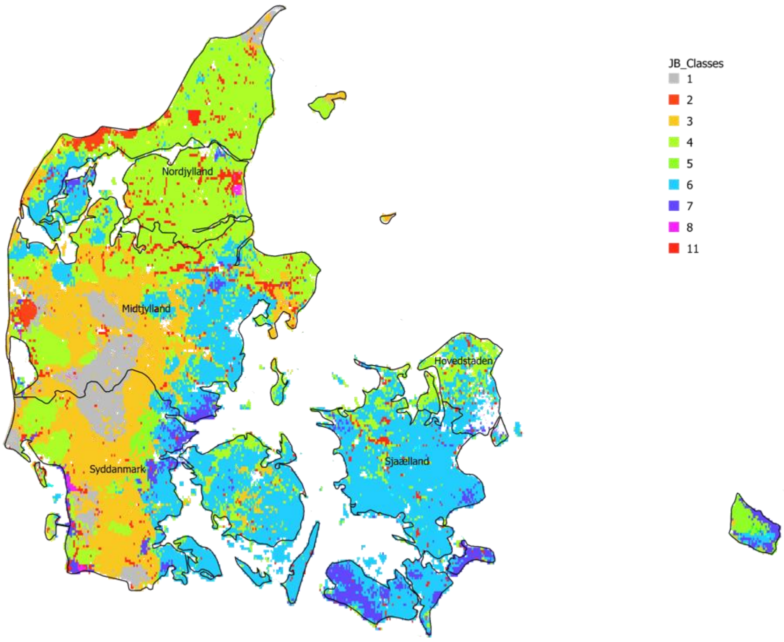
mLDNDC: A Machine Learning-based Surrogate for Optimising Cropping Systems in Denmark

Meshach Ojo Aderele[a], Edwin Haas[b], Licheng Liu[d], João Serra[a,c,e], Klaus Butterbach-Bahl [a,b], Jaber Rahimi[a,b*]

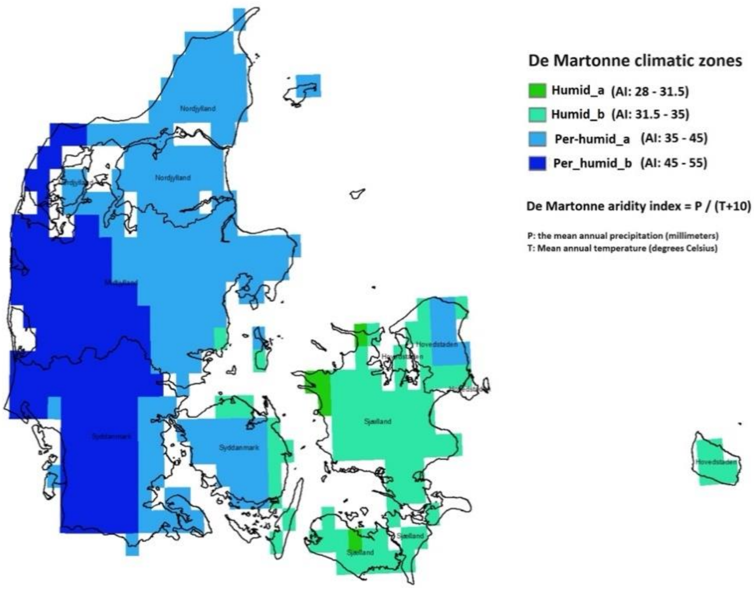
[a] Pioneer Center Land-CRAFT, Department of Agroecology, Aarhus University, Aarhus, Denmark
[b] Institute of Meteorology and Climate Research, Atmospheric Environmental Research (IMK-IFU), Karlsruhe Institute of Technology (KIT), Garmisch-Partenkirchen, Germany
[c] Department of Agroecology, Aarhus University, Blichers Allé 20, 8830 Tjele, Denmark
[d] Department of Bioproducts and Biosystems Engineering, University of Minnesota, St. Paul, MN 55108, USA.
[e] Forest Research Centre CEF, Associate Laboratory TERRA, Instituto Superior de Agronomía, Universidade de Lisboa, 1349-017, Lisbon, Portugal



Supplementary Figure 1. Relationship between mLDNDC (surrogate model) predicted and observed national winter wheat production in Denmark.



Supplementary Figure 2. Jodbær soil classification map



Supplementary Figure 3. Climate classification based on De Martonne aridity index

Supplementary Section 1. Gradient Boosting Models Used in mLDNDC

To develop the machine learning surrogate model, three state of the art gradient boosting algorithms were evaluated: XGBoost, LightGBM, and CatBoost. These ensemble methods belong to the family of gradient boosted decision trees, a modelling approach that has consistently demonstrated high performance on structured tabular datasets such as those derived from agroecosystem simulations. Gradient boosting builds an additive collection of decision trees by fitting successive learners to the residuals of previous trees, thereby capturing complex nonlinear relationships, interactions among features, and heterogeneous response surfaces (Friedman, 2001).

The selection of these three models reflects their documented advantages in predictive accuracy, computational efficiency, and interpretability for large and complex datasets. Below, we describe the fundamental principles and algorithmic characteristics of each model, with emphasis on their relevance for environmental and agricultural applications.

Supplementary Section 1.1. XGBoost

XGBoost (Extreme Gradient Boosting) is an optimised gradient boosting framework introduced by (Chen & Guestrin, 2016). It has become a widely adopted algorithm in scientific and industrial applications due to its accuracy, scalability, and flexibility. XGBoost incorporates several innovations that improve training speed and generalisation, including regularised tree construction, sparsity-aware learning, weighted quantile sketching for distributed computation, and parallelised tree growing.

A key strength of XGBoost is its ability to handle high-dimensional datasets with mixed feature types while maintaining robust performance across large simulation domains. Regularisation terms applied to both tree depth and leaf weights reduce overfitting and improve stability, which is particularly valuable when modelling biogeochemical outputs that exhibit nonlinear responses and threshold behaviour. The algorithm also supports histogram based tree building and GPU acceleration, enabling efficient training on datasets with millions of samples, such as the synthetic factorial dataset developed for LandscapeDNDC. Because agroecosystem processes often contain complex interactions among soil, climate, and management variables, XGBoost's ability to approximate nonlinear relationships through successive residual fitting makes it highly suitable for surrogate modelling.

Supplementary Section 1.2. LightGBM

LightGBM (Light Gradient Boosting Machine) was introduced by (Ke et al., 2017) as a highly efficient gradient boosting framework designed to improve training speed and reduce memory consumption. It uses two core algorithmic innovations: Gradient Based One Side Sampling and Exclusive Feature Bundling. These techniques reduce the number of samples and features evaluated at each splitting step without sacrificing accuracy. A defining characteristic of LightGBM is its leaf-wise tree growth strategy, in which trees are expanded by splitting the leaf that produces the largest reduction in loss rather than by level wise growth. This allows the model to capture deep and specific interactions that are common in soil hydrochemistry and nitrogen cycling processes. The algorithm also supports GPU accelerated training, which significantly reduces computational burden during optimisation and cross validation.

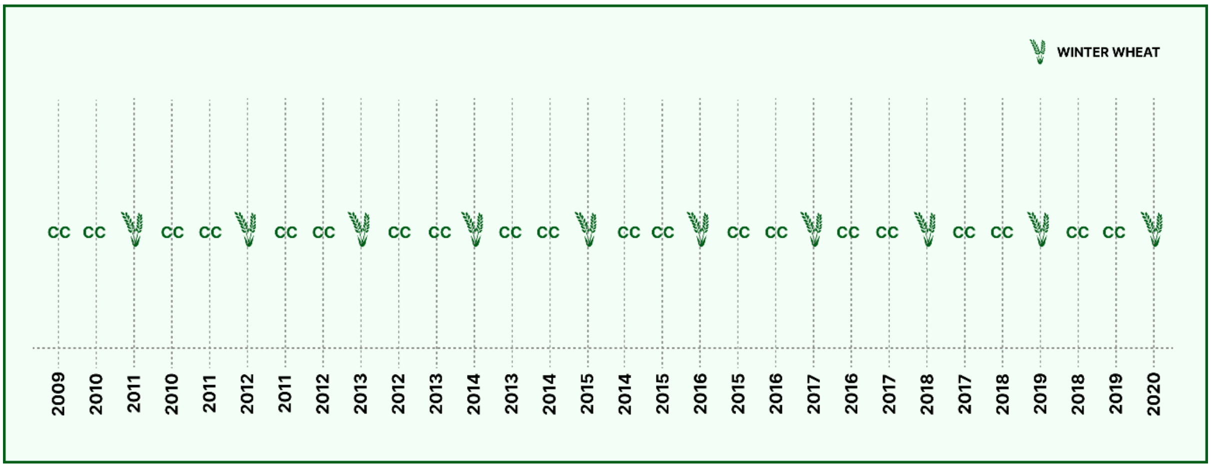
LightGBM has been shown to outperform or match XGBoost in many tabular data applications, particularly when the dataset contains large numbers of continuous variables with wide numeric ranges, as is typical in climate and soil datasets. These properties make it an effective candidate for surrogate modelling of agricultural systems.

Supplementary Section 1.3. CatBoost

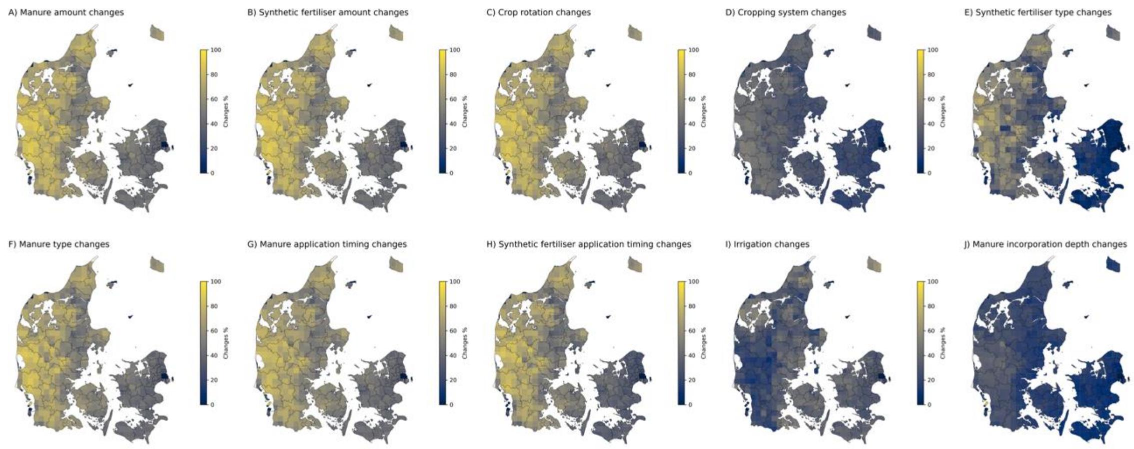
CatBoost (Categorical Boosting) is a gradient boosting algorithm introduced by (Prokhorenkova et al., 2017), designed to handle categorical features natively through ordered target statistics and permutation driven training. Unlike most boosting methods that require one hot encoding or arbitrary numeric transformations of categorical variables, CatBoost applies a principled method that reduces target leakage and improves generalisation. This characteristic is particularly relevant for agroecosystem modelling, where many variables such as crop rotation type, synthetic fertiliser type, irrigation decision, or tillage practice are categorical. CatBoost relies on oblivious decision trees, which split on the same feature across a given tree depth. This symmetric tree structure reduces model variance and improves stability, making CatBoost well suited for applications with mixed feature types and moderate dataset sizes. Although CatBoost is often competitive with XGBoost and LightGBM, it can be more computationally intensive and less scalable in very large datasets. In the present study, CatBoost performed consistently well but did not exceed the predictive accuracy of XGBoost or LightGBM for any of the four target variables. Nonetheless, its structural advantages for categorical inputs justified its inclusion in the model comparison.

Supplementary Table 1. Hyperparameters Used for XGBoost, LightGBM, and CatBoost Models

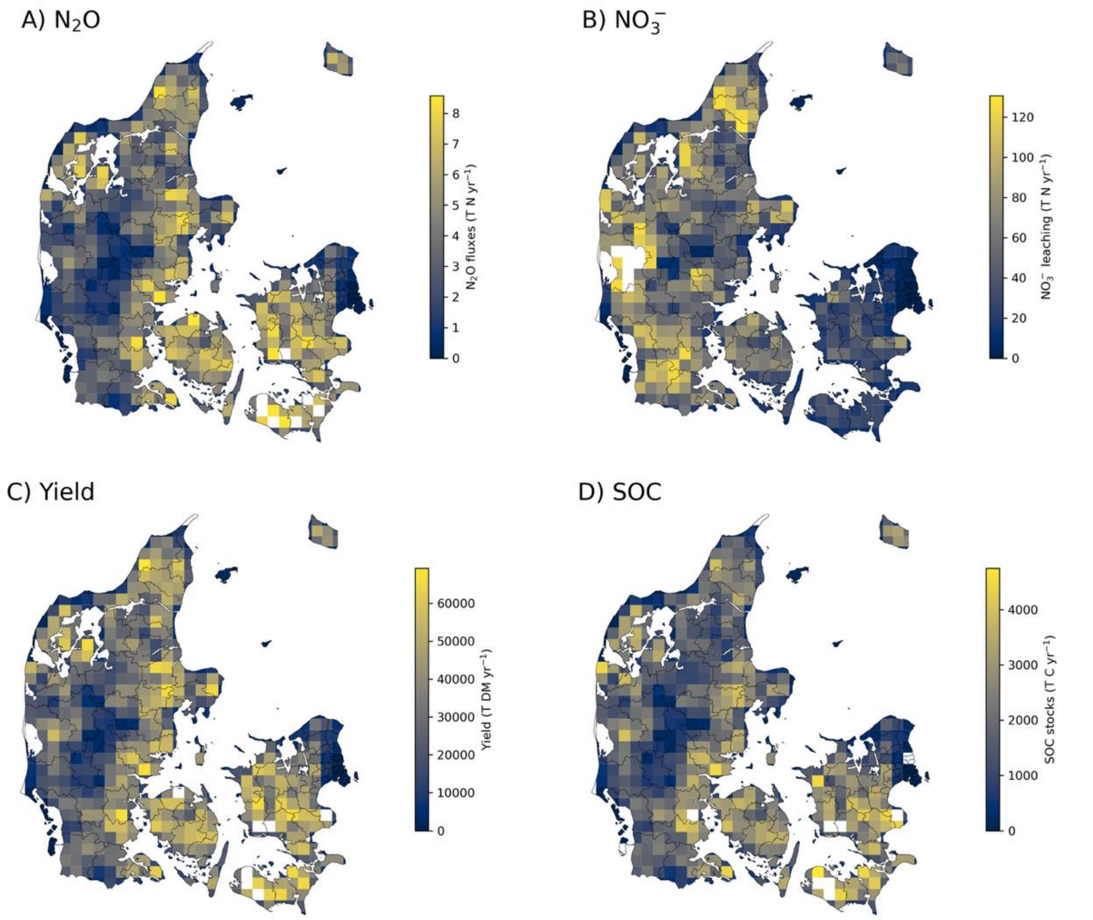
Parameter	XGBoost	LightGBM	CatBoost
objective	reg:squarederror	regression	RMSE
eval_metric / metric	rmse	rmse	RMSE
boosting / tree method	hist	gbdt	Lossguide
device	cuda	gpu	GPU
n_estimators / num_boost_round / iterations	10000	5000	2000
learning_rate	0.013556	0.01	0.013556
max_depth	9	-1	6
num_leaves	—	256	—
min_child_weight	48	—	—
subsample	0.8	0.8	—
colsample_bytree	0.6	0.8	—
gamma	1.074052	—	—
reg_alpha	0.151021	0.0	—
reg_lambda	11.097351	1.0	—
border_count	—	—	64
l2_leaf_reg	—	—	3
grow_policy	—	—	Lossguide
random_state / seed	—	42	42
sampling_method	gradient_based	—	—
max_bin	256	—	—
task_type	—	—	GPU
devices	—	—	0
verbose	—	—	False



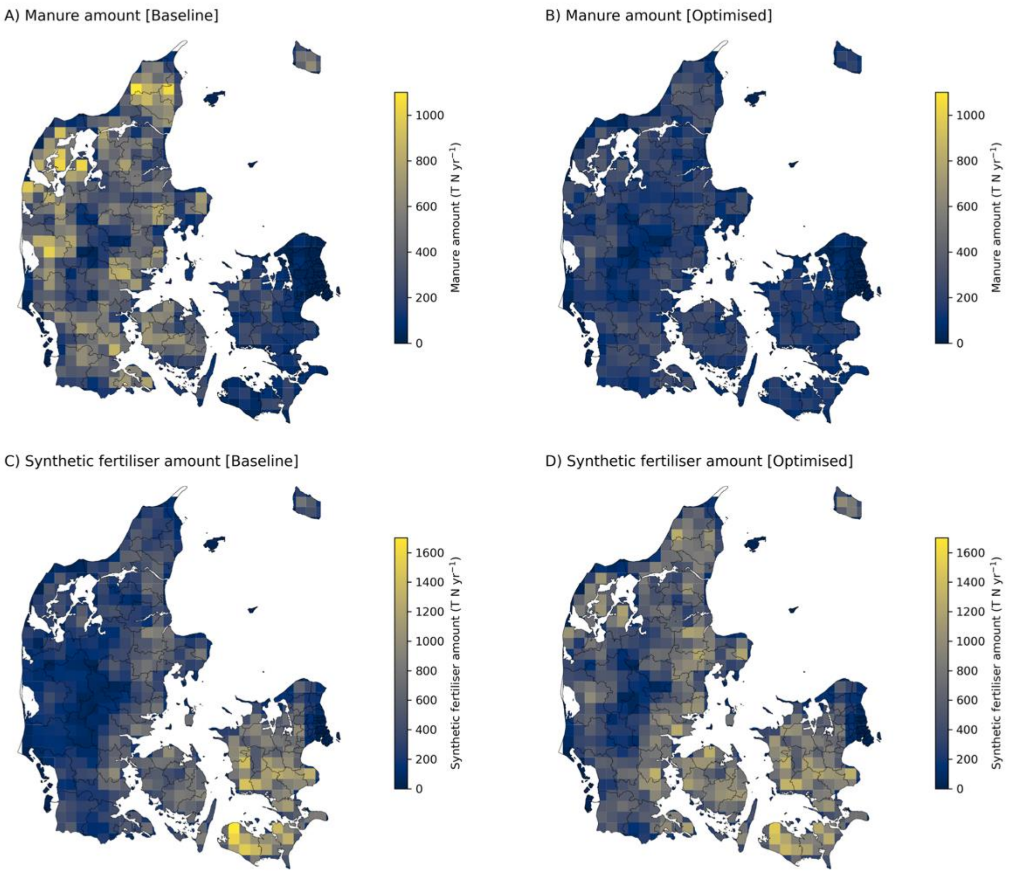
Supplementary Figure 4. Schematic representation of the crop rotation structure used in this study, illustrating that winter wheat is planted after two preceding crops. Rotations are grouped into five categories: cereal crops (CC), legume crops (BS), leafy crops (LC), root crops (RS), and grass crops (GR). An example shown in the figure is a CC–CC rotation, where two consecutive cereal crops precede winter wheat.



Supplementary Figure 5. Map showing the percentage of specific management changes from baseline to optimised in per 10 by 10km grid



Supplementary Figure 6. Spatial distribution of baseline values (mLDNDCv1.0 Predicted) for nitrous oxide emissions, nitrate leaching, crop yield, and soil organic carbon across Denmark in a 10 by 10km grid.



Supplementary Figure 7. Spatial distribution comparing baseline vs optimized manure and synthetic fertilizer amount across Denmark in a 10 by 10km grid.

Supplementary Table 2. Management Summary of Winter Wheat

Timing	Planting DOY	Harvest DOY	Manure Days to Planting (1)	Manure Days to Planting (2)	Manure Days to Planting (3)	Synthetic Fertiliser Days to Planting (1)	Synthetic Fertiliser Days to Planting (2)	Synthetic Fertiliser Days to Planting (3)
0	270	226	NA	NA	NA	NA	NA	NA
1	270	226	201	NA	NA	190	NA	NA
2	270	226	0	203	NA	185	217	NA
3	270	226	-10	80	199	181	203	226

Supplementary Table 3. Data Cleaning Conditions Summary to remove unrealistic scenarios.

Logical condition	Explanation
n_org_amount < 70 ∧ n_org_replication > 1	Removes cases where very low organic nitrogen is applied more than once, which is agronomically unrealistic.
70 ≤ n_org_amount ≤ 140 ∧ n_org_replication > 2	Removes scenarios where moderate organic nitrogen is applied too many times within a season.
n_synthamount < 100 ∧ n_synth_replication > 1	Prevents multiple synthetic fertilizer applications when the total synthetic nitrogen amount is low.
100 ≤ n_synthamount ≤ 200 ∧ n_synth_replication > 2	Removes excessive synthetic fertilizer replications at moderate application levels.
n_org_amount = 0	Sets organic fertilizer type to "none" and replication count to zero when no organic nitrogen is applied.
n_synthamount = 0	Sets synthetic fertilizer type to "none" and replication count to zero when no synthetic nitrogen is applied.
n_synthamount + n_org_amount > 400	Removes records where total nitrogen input exceeds realistic agronomic limits.
Drop duplicate rows	Removes identical rows to ensure each management scenario is unique.

Supplementary Table 4. List of features used for model training

Category	Short name	Descriptive name
Soil	soil	Soil class
	bd	Soil bulk density
	corg	Soil organic carbon content
	norg	Soil organic nitrogen content
	sand	Sand content (%)
	silt	Silt content (%)
	clay	Clay content (%)
	ph	Soil pH
	sks	Saturated hydraulic conductivity
	wcmax	Field capacity
	wcmin	Wilting point
Climate	climate	Climate class
	prec_days	Number of precipitation days per year
	total_precipitation_year	Annual total precipitation
	total_average_temperature_year	Annual mean air temperature
	total_precipitation_growing_season	Total precipitation during the growing season

	total_average_temperature_growing_season	Mean temperature during the growing season
	total_precipitation_autumn	Total autumn precipitation
	total_average_temperature_autumn	Mean autumn temperature
	total_precipitation_winter	Total winter precipitation
	total_average_temperature_winter	Mean winter temperature
	total_precipitation_spring	Total spring precipitation
	total_average_temperature_spring	Mean spring temperature
	total_precipitation_3_after_fert_1	Precipitation during 3 days after fertilizer application 1
	total_precipitation_3_after_fert_2	Precipitation during 3 days after fertilizer application 2
	total_precipitation_3_after_fert_3	Precipitation during 3 days after fertilizer application 3
	total_precipitation_3_after_manu_1	Precipitation during 3 days after manure application 1
	total_precipitation_3_after_manu_2	Precipitation during 3 days after manure application 2
	total_precipitation_3_after_manu_3	Precipitation during 3 days after manure application 3
	total_precipitation_7_before_fert_1	Precipitation during 7 days before fertilizer application 1
	total_precipitation_7_before_fert_2	Precipitation during 7 days before fertilizer application 2
	total_precipitation_7_before_fert_3	Precipitation during 7 days before fertilizer application 3
	total_precipitation_7_before_manu_1	Precipitation during 7 days before manure application 1
	total_precipitation_7_before_manu_2	Precipitation during 7 days before manure application 2
	total_precipitation_7_before_manu_3	Precipitation during 7 days before manure application 3
	precipitation_clay_interaction	Interaction between precipitation and clay content
	precip_n_interaction	Interaction between precipitation and total nitrogen applied
Management	cropping_systems	Cropping system (residue and catch crop incorporation or none)
	crop_rotation	Crop rotation scheme (crop categories in the two years preceding winter wheat)

	n_synth_type	Synthetic nitrogen fertilizer type
	n_org_type	Organic nitrogen source type (e.g., compost, farmyard manure, slurry)
	n_org_replication	Number of organic nitrogen applications
	n_synth_replication	Number of synthetic nitrogen applications
	irrigation	Irrigation regime (irrigated or rainfed)
	manu_depth	Manure incorporation depth
	n_org_amount	Total organic nitrogen applied (kg N ha^{-1})
	n_synthamount	Total synthetic nitrogen applied (kg N ha^{-1})
	fert_amount_1	Synthetic nitrogen amount, application 1 (kg N ha^{-1})
	fert_amount_2	Synthetic nitrogen amount, application 2 (kg N ha^{-1})
	fert_amount_3	Synthetic nitrogen amount, application 3 (kg N ha^{-1})
	manu_amount_1	Manure nitrogen amount, application 1 (kg N ha^{-1})
	manu_amount_2	Manure nitrogen amount, application 2 (kg N ha^{-1})
	manu_amount_3	Manure nitrogen amount, application 3 (kg N ha^{-1})
	total_nitrogen	Total nitrogen applied (kg N ha^{-1})
	synth_org_ratio	Ratio of synthetic to organic nitrogen
	fert_amount_1_sq	Squared synthetic nitrogen amount, application 1
	fert_amount_2_sq	Squared synthetic nitrogen amount, application 2
	fert_amount_3_sq	Squared synthetic nitrogen amount, application 3
	manu_amount_1_sq	Squared manure nitrogen amount, application 1
	manu_amount_2_sq	Squared manure nitrogen amount, application 2
	manu_amount_3_sq	Squared manure nitrogen amount, application 3
	total_nitrogen_sq	Squared total nitrogen applied