

A Large-Scale Simulated Dataset for Potato Crop with AI-Relevant Challenges

Satya Krishna Pothapragada

spothapragada@ufl.edu

University of Florida

Rishabh Gupta

University of Florida

Prateek Kumar Goel

University of Florida

Joel B Harley

University of Florida

Lincoln Zoteralli

University of Florida

Alina Zare

University of Florida

data-descriptor

Keywords:

Posted Date: October 27th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11237950/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.
[Read Full License](#)

Additional Declarations: No competing interests reported.

A Large Scale Simulated Dataset for Potato Crop with AI-Relevant Challenges

Satya Krishna Pothapragada ^{1*}
spothapragada@ufl.edu

Dr. Rishabh Gupta ^{2*}
r.gupta@ufl.edu

Prateek Kumar Goel¹
pkgprateek@gmail.com

Dr. Joel B Harley¹
joel.harley@ufl.edu

Dr. Lincoln Zoteralli ³
lzota@ufl.edu

Dr. Alina Zare¹
azare@ufl.edu

¹ Electrical and Computer Engineering ² Agricultural and Biological Engineering
³ Horticultural Sciences University of Florida, Gainesville, FL 32611

October 7, 2026

Abstract

Artificial Intelligence (AI) is increasingly influential in modeling and optimizing complex agricultural systems. However, publicly available large-scale agricultural datasets that present unique and significant AI challenges remain limited. In this paper, we introduce an extensive simulated dataset for potato crop production targeting Florida-like off-season growing conditions: a setting characterized by sandy soils, limited nutrient retention, and heightened sensitivity to weather and management variability. This dataset is generated using the DSSAT (Decision Support System for Agrotechnology Transfer) modeling framework and its Potato module. A key contribution of this work is the careful calibration of DSSAT parameters against real ground truth observations, a process that is well known to be non-trivial and time consuming, with no single standardized protocol typically requiring iterative trial-and-error adjustment of crop-specific genetic coefficients. This dataset provides a reproducible, agronomically calibrated starting point for researchers to build upon directly. The dataset comprises approximately 4.9 million crop-management scenarios varying planting dates, fertilizer rates, irrigation strategies, and historical weather yielding over 675 million daily data points across 24 years. Beyond its agricultural scope, this dataset presents a range of AI-specific challenges, including multivariate time-series forecasting, rare-event prediction, and early forecasting, making it a valuable benchmark for the broader machine learning and time-series research communities.

Background and Summary

Agriculture represents one of the largest real-world, yet underexplored, opportunities for modern AI, posing fundamental challenges in long-horizon time-series modeling, distribution shift, and heterogeneous data integration. High-quality, consistent datasets are essential for applying AI in agriculture, enabling models to learn complex relationships driven by weather, management, and environmental variation and to support sustainability-focused decision-making [24]. However, current agricultural datasets are often heterogeneous (numerical, categorical, images, text) [3, 5],

*Corresponding authors : spothapragada@ufl.edu (S.K.Pothapragada), r.gupta@ufl.edu (R.Gupta)

manually collected, and frequently rely on destructive, labor-intensive sampling, which limits temporal coverage and consistency. As a result, many datasets are small, short-term, and poorly standardized, making them difficult to integrate and insufficient for training robust AI models [5]. Therefore, large-scale, publicly available datasets spanning multiple seasons and capturing crop, soil, and weather management interactions are needed to accelerate progress in agricultural AI (see details in Table 2 for simulation scenarios and Table 3 for all variables used in the dataset). In this paper, we introduce a comprehensive simulated dataset-focused specifically on potato crop, generated from a process based model calibrated on real world crop management scenarios. Subsequently, these real world crop management scenarios were extrapolated to approximately 4.9 million crop management hypothetical scenarios such as varied planting dates, fertilizer application rates and timings, and irrigation strategies subjected to weather conditions from last 24 years, resulting in over 675 million rows of simulated data at daily time-step, providing a robust foundation for agricultural and AI research (see Fig 1 for a sample time series of a crop management scenario along with some covariates). The dataset contains variables related to plant growth, plant stress, soil nitrogen (N) balance, and soil water balance.

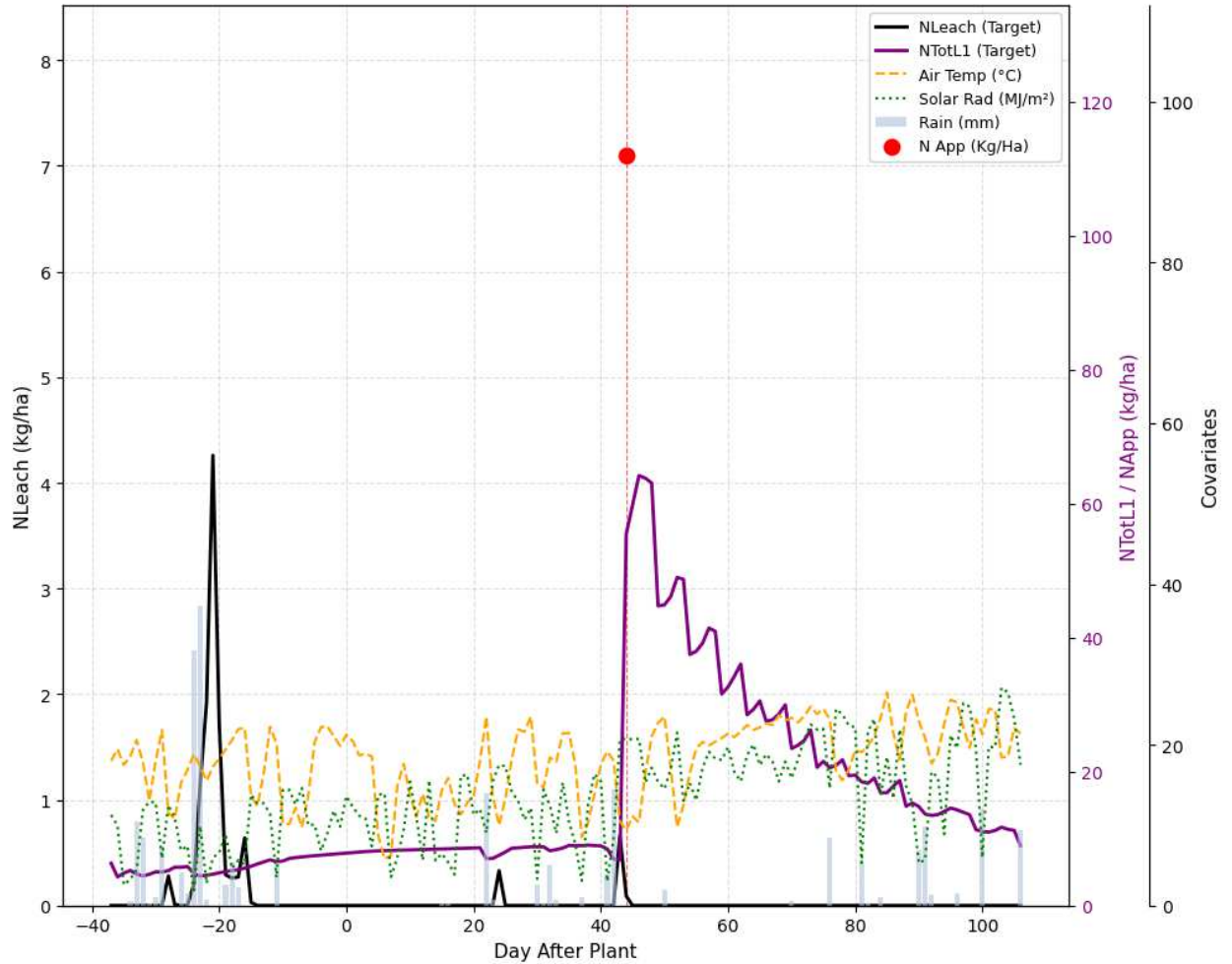


Figure 1: Interaction between Nitrogen leaching($NLeach$ shown on left axes, soil Nitrogen $NTotL1$ shown on right axes(purple), and some covariates (Air Temp, Solar Radiation, Rain and Nitrogen Application) over time for a sample crop management scenario (scenario details in Table 1)

Parameter	Crop Management Scenario
Year	2020
Planting Day	15
Nitrogen First Application	Npre
Irrigation Depth (mm)	60
Irrigation Threshold (%)	80
Treatment(Kg/Ha)	0–0–112

Table 1: Crop management scenarios used for visualization and time-series forecasting. The context length is 30 days, meaning the forecasting starts from day 31 through end of the time series.

We have chosen to develop a dataset focused on potatoes, as they offer a compelling case study. Potatoes are the fourth most staple food crop in the world, and in the United States, they are the fifth most widely grown crop, generating around \$100.8 billion in revenue annually, and play an essential role in food security [18]. Because potato growth is heavily influenced by complicated interactions among applied fertilizer, weather, and soil conditions, exhibit unique underground growth characteristics and demonstrate varying sensitivity to fertilizers and environmental factors throughout their growth stages [7]. Florida, which leads the production of off-season potatoes, adds further complexity through its sandy soils, which hinder nutrient retention and heighten sensitivity to fertilizer and weather variability. Compounding this, ground truth observations are scarce typically spanning only 5–6% of crop production days making this an important and challenging setting [12].

To address this scarcity of real-world observations, we turn to DSSAT, a well-established and powerful crop simulation platform capable of generating detailed synthetic data across a wide range of management and environmental scenarios [14]. However, calibrating DSSAT’s crop-specific genetic coefficients is itself a non-trivial undertaking: there is no universally accepted protocol, and researchers typically rely on iterative trial-and-error approaches, adjusting parameter combinations until simulated outputs align with observed field data [16, 17]. This challenge is amplified for off-season potato production on sandy soils, where distinct water and nutrient dynamics demand careful, location-specific parameterization [6]. The dataset presented here reflects this effort; it has been carefully calibrated to Florida-like sandy soil conditions and curated across a broad range of management scenarios, providing a reproducible starting point that other researchers can directly build upon without repeating this costly calibration process. It comprises a rich collection of potato crop simulations that cover key growth variables and the environmental and management factors that influence them, offering exciting opportunities to develop advanced ML and DL models that predict potato tuber growth, soil nitrogen content, nitrogen leaching, and other dynamics of the agroecosystem. In particular, variables such as soil nitrogen and nitrogen leaching exhibit strong temporal structure, making them well-suited for time-series modeling. Beyond standard modeling, the dataset supports challenging tasks, including rare-event prediction and early prediction under limited historical observations. As a result, it serves as a valuable benchmark for advancing methods in partial-observation and rare-event time-series forecasting, with relevance to both the time-series and machine learning communities, as well as agricultural decision-making.

AI in Time series modeling

A primary motivation for sharing this dataset is to aid the development of forecasting models under realistic covariate-driven conditions for rare event prediction (Nitrogen leaching, *NLeach* variable in the dataset) and early time series forecasting any other crop variables (like Soil Nitrogen -*NTotL1*, *NTotL2*, *NTotL3*, Soil Water-*SWatL1*, *SwatL2*, *SWatL3* variables). Overcoming these

challenges vital in agriculture, and there are few open datasets for studying them. In addition, these challenges are of growing concern within the broader AI community. Recent AI advances have reshaped time-series forecasting, exposing core AI challenges (generalization, representation learning) due to the continuous, noisy, domain-specific nature of series and frequent temporal distribution shifts [19, 15, 10, 13].

The successful prediction of rare events in this time series data is to accurately predict rare occurrences of N leaching. Existing research on rare-event prediction predominantly frames the problem as a classification task, typically identifying only when a rare event occurs. Rare events are defined as properties of the future trajectory. However, the problem considered here in this dataset is more demanding, involving not just timing predictions but also precise estimation of event magnitudes or real-valued outcomes, significantly increasing complexity. This is challenging because: (i) their rarity leads to data dominated by zeros, limiting useful pattern learning, and (ii) their extreme magnitudes follow long-tailed distributions, which machine-learning models struggle to predict accurately. Also, there is also a notable lack of publicly accessible datasets designed specifically for rare-event prediction within time series contexts, further constraining progress in this area [25, 1].

Early time series forecasting involves identifying the earliest possible point within a time series, potentially including exogenous (or covariate) inputs (e.g., weather, prices, interventions, calendar effects) where a reliable forecast can be made. This forecast may pertain either to a future value within the same series, i.e., early time series forecasting [4] or to a separate real-valued target variable not associated with that series, i.e., early time series extrinsic regression (TSER) [23]. In these challenges, exogenous inputs can play a critical role in improving prediction of target variables, particularly when these covariates are not reliably available in advance. There are a limited number of ML models, foundational models, and benchmarks that systematically evaluate covariate-aware forecasting. In the present dataset, these covariates are provided explicitly, allowing modelers to incorporate exogenous information into the forecasting process.

Methods

Simulation Framework and Data Sources

Background on DSSAT: DSSAT is a framework of multiple models that can simulate daily crop phenological development, biomass accumulation, and crop yield for over 42 crops by integrating weather, soil, and crop management data. For this dataset, we use SUBSTOR-Potato model of DSSAT to simulate the growth and development of potato crop as a response of weather variability and a wide-range of potato management scenarios, such as varied planting days, fertilizer application rates and timings, and irrigation strategies [8, 11]. The SUBSTOR-Potato model requires three primary input files to function. The weather file (*.wth*) provides essential environmental data, including temperature, rainfall, and solar radiation, which are necessary for estimating photosynthesis, evapotranspiration, and crop phenology. The soil file (*.sol*) contains information on soil properties, such as soil texture, wilting point, field capacity, and saturation limits, root growth factor, bulk density, and organic carbon content at various soil depths, that drive water-balance calculations, root development, and nutrient dynamics. The crop management specifies key details, including planting dates, fertilizer types, rates, and timings, cultivar choice, and field location, which define the simulation setup. In addition, the model uses genotype-specific parameters to capture the effects of crop genetics on growth and initial soil conditions to establish the starting state of the simulation.

Scenarios	Descriptor	Data Type	Description	Count
Planting Dates	PlantingDay	Int32	Jan 1, Jan 8, Jan 15, Jan 22, Jan 29, Feb 5, Feb 12, Feb 19, Feb 26, Mar 5	10
N Spplication	NFirstApp	Categorical	N at Preplant and N at Planting	2
N Rate-Timing	Treatment	Categorical	Npre/Npl = 0-196 kg/ha Nem = 0-196 kg/ha Nti = 0-196 kg/ha Max N Rate Capped = 400 kg/ha	428
Irrigation Depth	IrrgDep	Int32	30, 40, 50, 60 cm	4
Irrigation Threshold	IrrgThresh	Int32	50, 60, 70, 80, 90, 100%	6
Years	Year	Int32	2001-2024	24
		Total Simulations:		4,930,560
		Total Data Rows:		675,486,720

Table 2: Summary of simulation scenarios

Dataset Description

Dataset Variables Overview: The simulated dataset captures approximately 4.9 million diverse crop management practices. Each scenario refers to a unique combination of crop management inputs defined by six key parameters: year, planting day, applied N fertilizer rates, timing of first N application, irrigation depth, and irrigation threshold (see details in Table 2). Specifically, our scenarios incorporate ten planting days scenarios (weekly apart from Jan 1 to Mar 5) to reflect seasonal variability for 24 years (2001-2024) with harvesting fixed at 106 days after planting to represent typical potato growth season length in northeastern Florida. A total of 428 unique combinations of applied fertilizer rate and timing treatments were developed, with each fertilizer application ranged from 0-196 kg/ha in increments of 28 kg/ha, while ensuring that total N fertilizer application rate does not exceed 400 kg/ha. Anhydrous ammonia is used as N fertilizer and is applied at three critical phases: pre-planting (36 days before planting, Npre) or at planting (Npl), plant emergence (22 days after planting, Nemg), and tuber initiation (44 days after planting, Ntu). Farmers in northeastern Florida typically apply fertilizer either before (pre-planting) or during planting stage. Therefore, in our simulation dataset, two N fertilizer application strategies were included, one with the first application being at 36 before planting and other with the first application at planting, later followed by emergence and tuber-initiation N fertilizer applications. The dataset includes various irrigation strategies with different depths (30-60 cm with 10 cm increments) and thresholds (50-100% with 10% increments) triggering irrigation to maintain set percentage of plant available water at set soil depth to simulate diverse soil moisture conditions. In this way, the dataset has exactly 4,930,560 crop management scenarios with 24 years, 10 planting days, 428 applied N fertilizer rate-timings combinations, 2 different timings of first N application, 4 irrigation depths, and 6 irrigation thresholds.

Dataset Variables Description: The DSSAT model produced daily crop growth and soil water-N balance related outputs across an entire crop growth cycle for each unique management

scenario, resulting in approximately 675 million data rows. The comprehensive set of output variables includes daily measurements of soil N (nitrate-N, ammonium-N, and sum of these two N types) across the top three soil layers, soil water content (for top three soil layers), plant N uptake, accumulated biomass for both tuber and plant (dry and fresh weights), N leaching, and crop growth stage index. Additional data on irrigation amounts, weather descriptors from the FAWN station, and plant stress indicators related to N and water availability are also included (full details provided in Table 3). The dataset’s extensive variety and scale make it highly suitable for researchers and machine learning practitioners interested in modeling complex interactions among plant growth, soil nutrients, water availability, and weather factors. The structured organization of this dataset allows easy filtering, grouping, and selection of data subsets for targeted analyses.

Data Records

The hosted dataset referenced in Section Data availability has a well-organized and clean structure. It consists of a single top-level directory containing (i) a weather file with information for all years and (ii) a folder named *potsim_yearly*. The *potsim_yearly* folder contains 24 files, each representing one year of simulation data covering all scenario combinations listed in Table 2. All files are provided in parquet format (`.parquet`).

Technical Validation

For technical validation, we employ DSSAT, a well established crop simulation platform, specifically its SUBSTOR-Potato module, which is designed to simulate the growth and development dynamics of potato crops. The mechanistic structure of SUBSTOR-Potato itself serves as an implicit form of validation, as its underlying equations are built to capture the physiological and bio-geochemical processes governing crop variables such as tuber growth, biomass accumulation, soil nitrogen dynamics, and nitrogen leaching [14, 11]. As discussed earlier, calibrating DSSAT parameters is a non-trivial process with no single standardized approach; to address this, we carefully calibrated the model parameters against available ground truth data points to ensure that the simulated outputs are representative of real-world conditions in Florida-like sandy soil environments which can be seen in Fig 2. In this figure, the calibrated simulations are compared against observed ground truth for critical crop variables, namely *Tuber Dry Weight*, *Tuber Nitrogen*, and *Plant Nitrogen*. The results show a reasonably good agreement between the simulated and observed values with R^2 around 0.5 on these variables. This calibration step is critical for producing simulations that are representative of Florida growing conditions and consistent with ground truth observations, thereby strengthening the dataset’s reliability for downstream machine learning and deep learning research.

Calibration of DSSAT to real-world conditions: We generated simulated data for various management scenarios mentioned in Table 2 throughout a 24 year period using the calibrated DSSAT model for Hastings, Florida. For our simulations, we used daily weather data including rainfall, air temperature (maximum and minimum) (at 60 cm), relative humidity, dew point, solar radiation, and wind speed collected over 24 years (2001–2024) from a Florida Automated Weather Network (FAWN) station in Hastings, FL [9]. The DSSAT model was calibrated using observations from two field trials conducted from 2011–2014 on four commercial farms in Hastings, FL under eight different N fertilizer rates and timing treatments applied during multiple stages of potato growth—pre-planting, emergence, and tuber-initiation [21, 20, 27, 28]. Key soil properties (wilting point, field capacity, saturated water content, root growth factor, bulk density, organic carbon

Variable	Description
Date	Date of measurement or simulation event
NApp	Rate of N fertilizer applied (kg/ha)
NLeach	Amount of N leached (kg/ha)
NO3L1	Nitrate concentration at soil layer 1 (mg/kg)
NO3L2	Nitrate concentration at soil layer 2 (mg/kg)
NO3L3	Nitrate concentration at soil layer 3 (mg/kg)
NH4L1	Ammonium concentration at soil layer 1 (mg/kg)
NH4L2	Ammonium concentration at soil layer 2 (mg/kg)
NH4L3	Ammonium concentration at soil layer 3 (mg/kg)
NTotL1	Total N concentration at soil layer 1 (mg/kg)
NTotL2	Total N concentration at soil layer 2 (mg/kg)
NTotL3	Total N concentration at soil layer 3 (mg/kg)
Irrg	Amount of irrigation water applied (mm)
SWatL1	Soil water content at soil layer 1 (cm ³ /cm ³)
SWatL2	Soil water content at layer 2 (cm ³ /cm ³)
SWatL3	Soil water content at layer 3 (cm ³ /cm ³)
PlantN	N content of plant canopy (kg/ha)
TuberN	N content of tubers (kg/ha)
NPlantUp	Total N uptake by plants (kg/ha)
LeafAreaIdx	Leaf area index of the plant canopy (dimensionless)
TuberFW	Fresh weight of harvested tubers (Mg/ha)
TuberDW	Dry weight of harvested tubers (kg/ha)
TopsDW	Dry weight of aboveground plant biomass (kg/ha)
GroStage	Growth stage of the crop : 3 stages, takes values (0-2) 0 for before plant emergence, 1 for between emergence and tuber-initiation, and 2 for tuber-initiation and harvest
WatStressPhs	Water stress index affecting photosynthesis (fraction)
WatStressGro	Water stress index affecting growth (fraction)
NStress	N stress index of the crop (fraction)
IrrgDep	Depth of irrigation water applied per event (cm)
IrrgThresh	Threshold soil water level triggering irrigation (%)
NFirstApp	First N fertilizer application timings: 'Npre' for at pre-planting or 'Npl' for at planting
Treatment	Experimental treatment identifier
DayAfterPlant	Days elapsed since planting (0 is day of planting)
Year	Year of the crop growing season
PlantingDay	Calendar day when planting occurred

Table 3: Description of dataset output variables

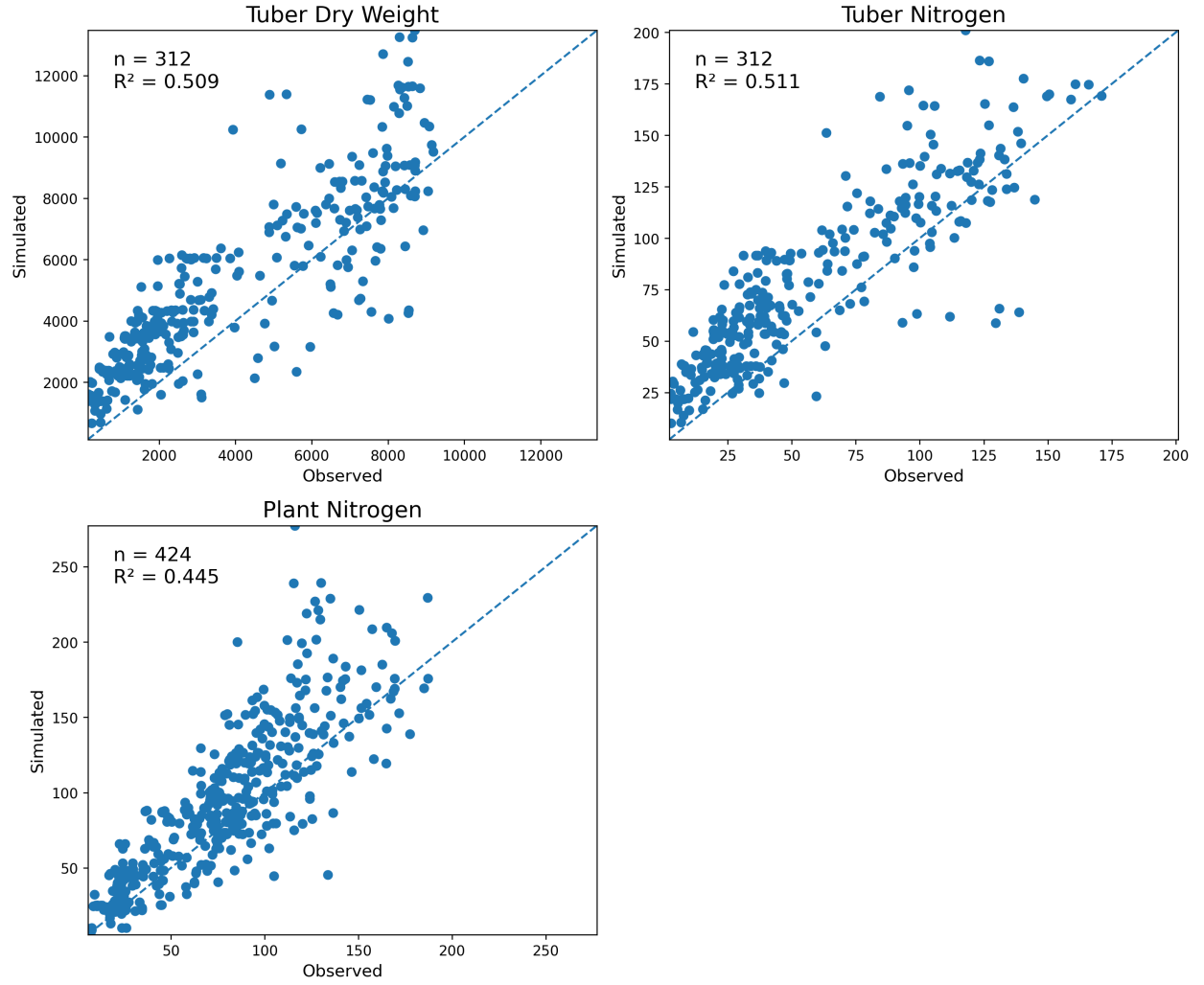


Figure 2: This Plot compared Simulations generated in this dataset (y-axis) vs ground truth observed values (x-axis) for several crop variables: Tuber Dry Weight(kg/Ha), Tuber Nitrogen(kg/Ha), Plant Nitrogen(kg/Ha) from 2011– 2014. Here n is the number of ground truth obtained points for a given target variable and R^2 is the correlation between observed ground truth and the simulated values for a given target variable.

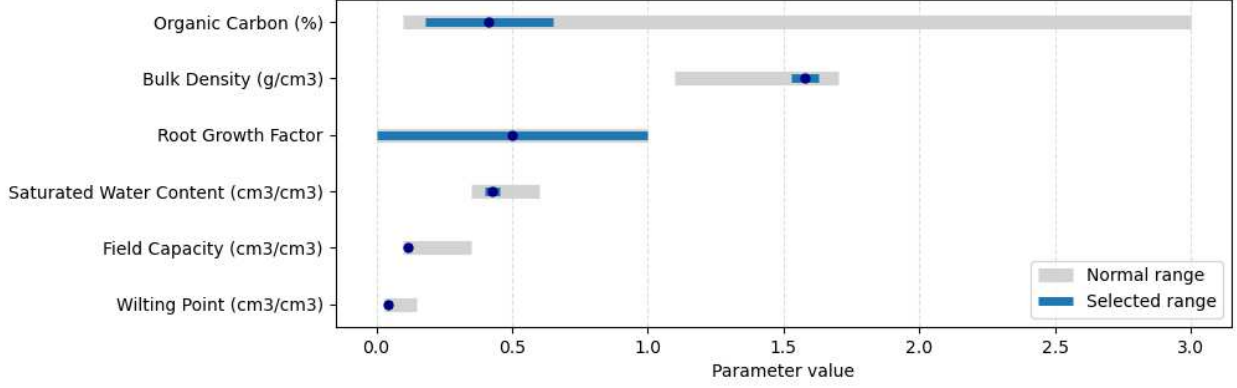


Figure 3: This plot shows the soil-layered parameters commonly used for potato crop , along with the selected ranges used for our simulations that broadly represent sandy soil conditions in Florida for depths ranging from 0 to 140 cm.

content whose ranges mentioned in Fig. 3) and crop genotype coefficients were tuned to closely mimic field conditions and accurately simulate plant aboveground biomass, tuber yield, and soil N balance (please refer to [12] and [22] for more details on the calibration and evaluation of the DSSAT model).

Crop Variable Prediction Performance: One of the primary motivations for developing this dataset is to stimulate AI research on agricultural problems, which pose distinct AI challenges. Therefore, these experiments illustrate the AI challenges described in the AI in Time series modeling section. First, we investigate whether models trained from scratch can effectively learn key crop and soil variables from this dataset by predicting a target variable at time t using available covariates. This is formulated as a regression problem, where all models share the same feature set and predict the value at time t . Sequence-based models (TCN, LSTM, Transformer) consume a history of t days, while simpler models (MLP, linear regression) use only features at time t . For all of our evaluation results we report two summary statistics mean absolute error (MAE) and the coefficient of determination (R^2) that are computed by aggregating the predictions over all scenarios in the test set. Regression based results are reported in Tables 4 and 5, with dataset splits and training details provided in Supplementary section. When trained from scratch, deep learning models are able to capture meaningful patterns for most crop variables, indicating that the dataset contains learnable structure. Nitrogen leaching ($NLeach$) is a notable exception. While DSSAT provides a consistent and reliable data source, and while most target variables in this setting prove learnable under a regression framework, nitrogen leaching systematically fails to be modeled accurately. Importantly, this failure is not a consequence of data quality or model selection, but rather an expected and theoretically grounded outcome. Nitrogen leaching is an instantaneous, event-driven process governed by transient weather conditions such as rainfall and soil moisture, combined with residual soil nitrogen at that moment. This infers that leaching relies on future dependent exogenous inputs which is unavailable to regression models at inference, making such approaches structurally infeasible resulting in a poor performance. Consequently, we develop forecasting models that incorporate future exogenous inputs, using this dataset to address the challenge. Where in contrast, soil nitrogen ($NTotL1$), which exhibits smoother temporal dynamics, can be learned more effectively in the regression setting, achieving higher predictive accuracy across models.

Early Time series forecasting and Rare event Prediction Results: Second, we evaluate the ability of forecasting models to capture temporal dynamics under realistic conditions. Trans-

Target	CNN1D	Transformer	LSTM	LinearRegression	MLP	TCN
NLeach	0.894	0.923	0.774	1.077	1.000	1.415
NPlantUp	0.685	0.865	0.717	1.443	0.719	0.720
NTotL1	9.023	11.491	9.215	17.403	10.652	9.428
NTotL2	7.848	9.870	8.054	16.409	9.594	8.943
SWatL1	0.006	0.009	0.006	0.024	0.016	0.009
SWatL2	0.005	0.009	0.005	0.009	0.008	0.006

Table 4: Mean squared error (MAE) metrics for different models

Target	CNN1D	Transformer	LSTM	LinearRegression	MLP	TCN
NLeach	0.432	-0.02	0.343	0.002	0.014	0.265
NPlantUp	0.803	0.733	0.794	0.322	0.753	0.791
NTotL1	0.843	0.764	0.831	0.481	0.779	0.823
NTotL2	0.861	0.799	0.849	0.489	0.792	0.843
SWatL1	0.973	0.949	0.972	0.620	0.841	0.950
SWatL2	0.944	0.783	0.928	0.700	0.816	0.914

Table 5: Coefficient of determination (R^2) metrics for different models

formers have become a common choice for time-series modeling in the ML community. Accordingly, we use a transformer-based forecaster. A key consideration in practical agricultural settings is future availability of covariate information. Nitrogen leaching, for instance, is strongly driven by rainfall events as discussed before. If future rainfall were known in advance, forecasting leaching would be relatively easier. To explicitly capture this distinction, we develop two variants of our forecasting model: one that operates exclusively on past and current covariates in an autoregressive setting (the realistic setting: *Trained model (without future exog.)*), and one that is additionally provided with future covariates, such as weather and fertilizer inputs, to generate the forecast for the next horizon days (the oracle-like setting: *Trained model (with future exog.)*). The latter serves as an reference for the predictive performance attainable with future covariate knowledge (further details on the forecasting models are provided in Supplementary Section. To contextualize these oracle-level results, we benchmark against foundational time-series models such as **MOIRAI** (*M*asked enc*O*der based *UnI*ve*R*s*AI* t*I*me series forecaster) and **CHRONOS**, which likewise have access to the true future exogenous covariates during inference [26, 2].

Our forecasting approach is designed to address two core AI challenges: early time series forecasting and rare event prediction. We unify these challenges by formulating our task as a one-dimensional time series problem, where target variables are modeled using a set of covariates: nitrogen fertilizer application *N App* (kg/ha), rainfall (cm), solar radiation (*W/m2*) and air temperature ($^{\circ}C$). In this framework, the model uses past covariate information and current variable states over a specific observation window to predict the target variable across a defined forecast horizon. To evaluate the impact of the available historical context on model performance, we employ two different window sizes (30 and 60 days) for both target variables, with a rolling forecast horizon of 30 days extending through the end of the crop cycle. This setup allows us to examine two distinct dynamics. For *N TotL1*, a smoothly continuous variable, we explore the transition from simple regression which we previously demonstrated can achieve decent performance, to true forecasting, which requires predicting future values with limited historical windows. For *N Leach*

Model	SoilN (<i>NTotL1</i>) MAE/ R^2		N Leaching (<i>N Leach</i>) MAE/ R^2	
	window=30	window=60	window=30	window=60
Trained model (without future exog.)	11.603/0.634	9.525/0.752	1.298/0.000	1.523/-0.001
Trained model (with future exog.)	4.069/0.958	4.319/0.956	0.622/0.466	0.839/0.469
Moirai <i>1.0-R large</i>	27.712/-0.646	30.408/-0.509	0.858/-0.101	0.957/-0.090
Chronos <i>V2</i>	25.025/-0.457	27.089/-0.240	0.923/-0.295	0.923/-0.043

Table 6: MAE and R^2 Performance comparison for Soil Nitrogen and Nitrogen Leaching Forecasting across trained models (without access to future exogenous inputs and with future exogenous inputs) and latest foundation models (with future exogenous input setting) for two input window sizes. All models have the same forecasting horizon of 30 days at once.

however, the task naturally evolves into a rare event prediction problem because leaching is a sparse, intermittent event heavily dependent on rainfall; forecasting it without future covariate knowledge is considerably more challenging, providing a rigorous test of the model’s predictive limits

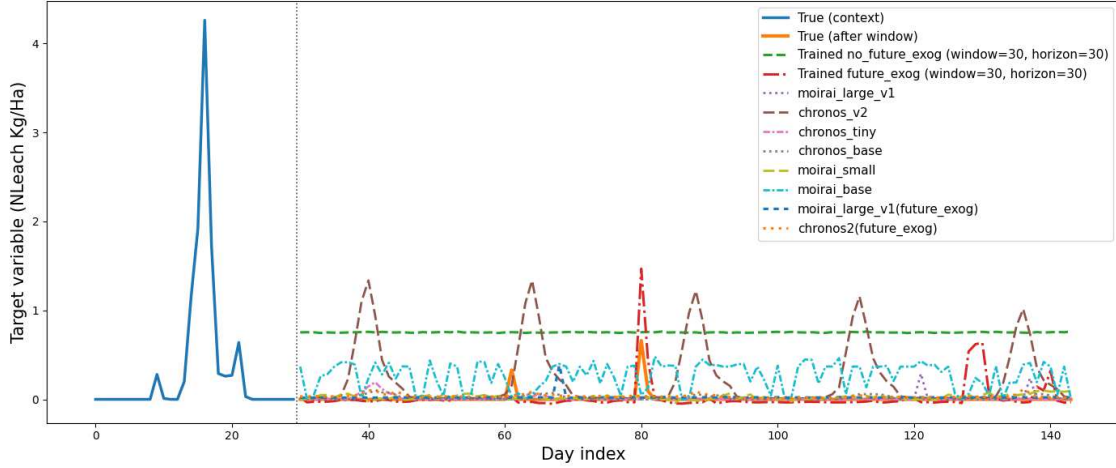
The performance of our models across different configurations is detailed in Table 6 , 7 and Figures 4 ,5 (30 and 60-day input window respectively). As shown in these results, increasing the input window from 30 to 60 days generally improves performance for the continuous smooth variable *N TotL1* by providing more historical context. However, in a practical agricultural setting, waiting to collect more data delays critical decision-making and can jeopardize crop yields, emphasizing the urgent need for accurate early-season forecasting with minimal observations. In contrast, for the rare event variable *N Leach*, increasing the window size yields negligible improvement. Performance remains poor except in the oracle-like setting, where access to future covariates allows for a some boost, highlighting the extreme difficulty of forecasting rare events without future knowledge.

In addition to trained models, we also evaluated large pre-trained foundation models, MOIRAI and CHRONOS. We evaluated *MOIRAI-1.0-R-large* and *CHRONOS v2* variants on our test set, which are designed to handle future exogenous features and we used the same experimental setting we applied to our trained models. Each of these foundation models receives the first 30 or 60 days of data and is then asked to predict the remaining days of the scenario for the target variable in 30-day horizons. Despite their immense capacity and access to richer input information, these models do not show a substantial improvement over our trained models, as illustrated in Table 6. We also compared several other variants of Moirai and Chronos; the results for those alternatives are plotted in the figures 4 and 5 where they consistently fall behind in capturing the dynamics of the target variable and yield poor forecast accuracy.

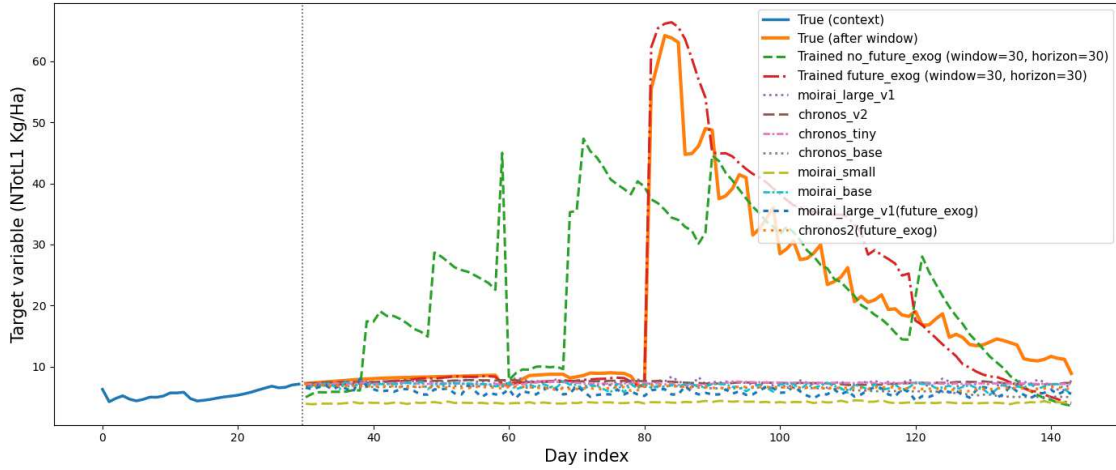
Input Window	Forecast Horizon	SoilN (<i>NTotL1</i>) MAE/ R^2	N Leaching (<i>N Leach</i>) MAE/ R^2
30	10	6.604 / 0.839	1.255 / -0.000
30	20	9.642 / 0.709	1.272 / -0.001
30	30	11.603 / 0.634	1.298 / 0.000
60	10	7.249 / 0.842	1.476 / -0.000
60	20	8.464 / 0.794	1.496 / -0.000
60	30	9.525 / 0.752	1.523 / -0.001

Table 7: Test MAE and R^2 performance for soil nitrogen and nitrogen leaching forecasting across different window and horizons (days) across test set.

The results show that modeling these agricultural variables remains challenging, even when future exogenous variables are supplied in an oracle-like setting. For instance, the model attains a modest R^2 of 0.4694 for the rare event variable $N\ Leach$. This modest performance highlights the limitations of standard foundational models and the need for more effective approaches. While the dataset does not solve the problem outright, its large scale and strong alignment with ground-truth measurements provide high-quality simulations that closely mirror real observations. Consequently, agricultural researchers can generate reliable daily estimates of crop variables under specific management scenarios, and AI/ML engineers can use the data to develop and benchmark methods for prediction, forecasting, and rare event detection, all within a single, coherent framework that encourages the creation of better algorithms for this demanding domain.

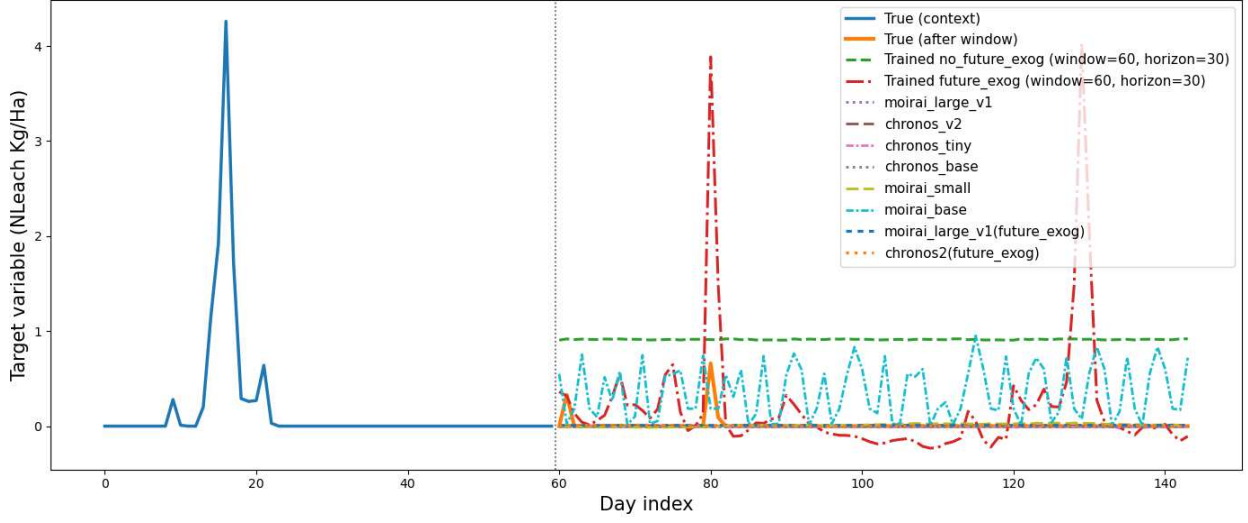


(a) Nitrogen leaching forecasting.

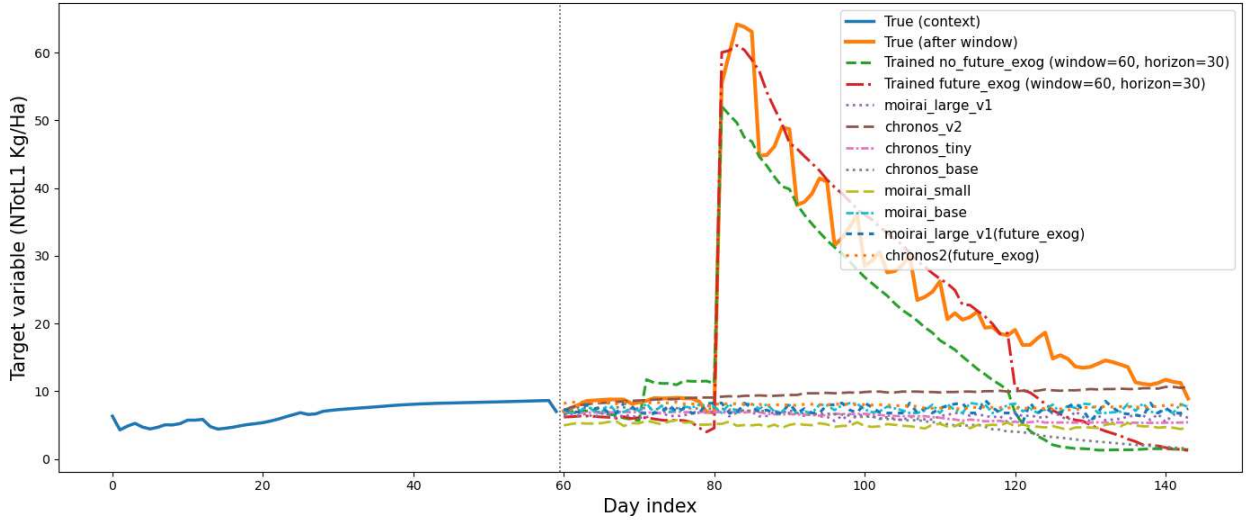


(b) Soil nitrogen forecasting.

Figure 4: Time-series forecasting of nitrogen leaching using various versions of MOIRAI, CHRONOS models, and two trained models using rolling-horizon forecasting model and a forecasting model that uses future exogenous variables for **input window = 30 days** on a crop scenario mentioned in Table 1. Corresponding quantitative results are reported in Table 6.



(a) Nitrogen leaching forecasting.



(b) Soil nitrogen forecasting.

Figure 5: Time-series forecasting of soil nitrogen using various versions of MOIRAI, CHRONOS models, and two trained models using rolling-horizon forecasting model and a forecasting model that uses future exogenous variables for **input window = 60 days** on a crop scenario mentioned in Table 1. Corresponding quantitative results are reported in Table 6.

Data availability

The dataset is publicly accessible at <https://doi.org/10.7910/DVN/GQMD0V>. This dataset contains simulation files yearly and also contains weather information for all these years, separately in parquet file format.

Code availability

To facilitate the use of the dataset, we have developed a GitHub repository (<https://github.com/GatorSense/PotSim>.) that includes built-in functions for efficient data loading and data processing, like loading a specific crop management scenario from available options. In addition to easing data processing, the library includes all codes for the models used in this paper and results are attached. Furthermore, we offer startup code that integrates several ML and DL algorithms with the dataset. This setup enables researchers to quickly begin experimentation and model development with minimal overhead.

References

- [1] Yahaya Idris Abubakar et al. “A systematic review of rare events detection across modalities using machine learning and deep learning”. In: *IEEE Access* (2024).
- [2] Abdul Fatir Ansari et al. “Chronos-2: From Univariate to Universal Forecasting”. In: *ArXiv abs/2510.15821* (2025).
- [3] Showkat Ahmad Bhat and Nen-Fu Huang. “Big data and ai revolution in precision agriculture: Survey and challenges”. In: *Ieee Access* 9 (2021), pp. 110209–110222.
- [4] Chris Chatfield. *Time-series forecasting*. Chapman and Hall/CRC, 2000.
- [5] Ania Cravero et al. “Challenges to use machine learning in agricultural big data: a systematic literature review”. In: *Agronomy* 12.3 (2022), p. 748.
- [6] Andre Luiz Biscaia Ribeiro da Silva et al. “Assessing the impact of irrigation and nitrogen management on potato performance under varying climate in the state of Florida, USA”. In: *Agricultural Water Management* 295 (2024), p. 108769. ISSN: 0378-3774. DOI: <https://doi.org/10.1016/j.agwat.2024.108769>. URL: <https://www.sciencedirect.com/science/article/pii/S0378377424001045>.
- [7] André Devaux et al. “Global food security, contributions from sustainable potato agri-food systems”. In: *The potato crop: Its agricultural, nutritional and social contribution to humankind* (2020), pp. 3–35.
- [8] KL Divya et al. “Crop simulation models as decision-supporting tools for sustainable potato production: a review”. In: *Potato Research* 64.3 (2021), pp. 387–419.
- [9] FAWN. *Florida Automated Weather Network*. <https://fawn.ifas.ufl.edu/>. 2022.
- [10] Rakshitha Godahewa et al. *Monash Time Series Forecasting Archive*. 2021. arXiv: 2105.06643 [cs.LG]. URL: <https://arxiv.org/abs/2105.06643>.
- [11] Timothy S Griffin, Bradley S Johnson, and Joe T Ritchie. *A simulation model for potato growth and development: Substor-potato Version 2.0*. Vol. 2. Michigan State University, Department of Crop and Soil Sciences East Lansing ..., 1993.
- [12] Rishabh Gupta et al. “Estimating soil mineral nitrogen from data-sparse field experiments using crop model-guided deep learning approach”. In: *Computers and Electronics in Agriculture* 225 (2024), p. 109355.
- [13] Xiaochuang Han et al. “A Survey on Distribution Shift in Machine Learning”. In: *Proceedings of the International Conference on Machine Learning (ICML)*. 2023.

- [14] Gerrit Hoogenboom et al. “The DSSAT crop modeling ecosystem”. In: *Advances in crop modelling for a sustainable agriculture*. Burleigh Dodds Science Publishing, 2019, pp. 173–216.
- [15] hubli kottapalli hubli et al. “Foundation Models for Time Series: A Survey”. In: *arXiv preprint* (2023). <https://arxiv.org/html/2504.04011v1>.
- [16] Prakash Kumar Jha et al. “A comparison of multiple calibration and ensembling methods for estimating genetic coefficients of CERES-Rice to simulate phenology and yields”. In: *Field Crops Research* 284 (2022), p. 108560. ISSN: 0378-4290. DOI: <https://doi.org/10.1016/j.fcr.2022.108560>. URL: <https://www.sciencedirect.com/science/article/pii/S0378429022001319>.
- [17] James W Jones et al. “Estimating DSSAT cropping system cultivar-specific parameters using Bayesian techniques”. In: *Methods of introducing system models into agricultural research* 2 (2011), pp. 365–393.
- [18] William Knudson and Steven R Miller. “Measuring the economic significance of the US potato industry”. In: *Report for the National Potato Council*. Available online: <https://www.nationalpotatocouncil.org/wp-content/uploads/2023/02/NPCSpudNationReport.pdf> (accessed on 12 March 2024) (2023).
- [19] Bryan Lim and Stefan Zohren. “Time-Series Forecasting with Deep Learning: A Survey”. In: *Philosophical Transactions of the Royal Society A* 379.2194 (2021). DOI: 10.1098/rsta.2020.0209.
- [20] Libby R Rens et al. “Optimizing nitrogen fertilizer rates and time of application for potatoes under seepage irrigation”. In: *Field Crops Research* 215 (2018), pp. 49–58.
- [21] Libby R Rens et al. “Rate and timing of nitrogen fertilizer application on potato ‘FL1867’ part II: Marketable yield and tuber quality”. In: *Field Crops Research* 183 (2015), pp. 267–275.
- [22] Andre Luiz Biscaia Ribeiro da Silva et al. “Assessing the impact of irrigation and nitrogen management on potato performance under varying climate in the state of Florida, USA”. In: *Agricultural Water Management* 295 (2024), p. 108769.
- [23] Chang Wei Tan et al. “Time series extrinsic regression: Predicting numeric values from time series data”. In: *Data Mining and Knowledge Discovery* 35.3 (2021), pp. 1032–1060.
- [24] Rui-Feng Wang and Wen-Hao Su. “The application of deep learning in the whole potato production Chain: A Comprehensive review”. In: *Agriculture* 14.8 (2024), p. 1225.
- [25] Gary M Weiss and Haym Hirsh. “Learning to Predict Rare Events in Event Sequences.” In: *KDD*. Vol. 98. 1998, pp. 359–363.
- [26] Haixu Wu et al. “MOIRAI: A Foundation Model for Time Series Forecasting”. In: *arXiv preprint arXiv:2402.02592* (2024).
- [27] Lincoln Zotarelli et al. “Nitrogen fertilizer rate and application timing for chipping potato cultivar Atlantic”. In: *Agronomy Journal* 106.6 (2014), pp. 2215–2226.
- [28] Lincoln Zotarelli et al. “Rate and timing of nitrogen fertilizer application on potato ‘FL1867’. Part I: Plant nitrogen uptake and soil nitrogen availability”. In: *Field Crops Research* 183 (2015), pp. 246–256.

Authors Contributions

S.K. Pothapragada: Writing – original draft, Software, Visualization, Methodology, Investigation, Formal analysis; **R. Gupta:** Writing – original draft, Software, Methodology, Investigation, Formal analysis, Conceptualization; **P.K. Goel:** Data curation; **J.B. Harley:** Writing – review & editing, Project administration, Methodology, Funding acquisition, Conceptualization; **L. Zotarelli:** Writing – review & editing, Project administration, Methodology, Funding acquisition, Conceptualization; **A. Zare:** Writing – review & editing, Project administration, Methodology, Funding acquisition, Conceptualization.

Competing Interests

The authors declare no competing interests.

Funding

This work was supported by the Florida Department of Agriculture and Consumer Services (FDACS) under funding Award ID: AWD17077. We gratefully acknowledge the University of Florida Hastings Agricultural Extension Center for their field support.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [supplementary.tex](#)