VITA: Variational Pretraining of Transformers for Climate-Robust Crop Yield Forecasting

Adib Hasan[1], **Mardavij Roozbehani**[2], **Munther A. Dahleh**[2]
[1]Independent Researcher
[2]Massachusetts Institute of Technology
notadib@alum.mit.edu, mardavij@mit.edu, dahleh@mit.edu

Abstract

Accurate crop yield forecasting is essential for global food security. However, current AI models systematically underperform when yields deviate from historical trends. We attribute this to the lack of rich, physically grounded datasets directly linking atmospheric states to yields. To address this, we introduce *VITA (**V**ariational **I**nference **T**ransformer for **A**symmetric data)*, a variational pretraining framework that learns representations from large satellite-based weather datasets and transfers to the ground-based limited measurements available for yield prediction. VITA is trained using detailed meteorological variables as proxy targets during pretraining and learns to predict latent atmospheric states under a seasonality-aware sinusoidal prior. This allows the model to be fine-tuned using limited weather statistics during deployment. Applied to 763 counties in the U.S. Corn Belt, VITA achieves state-of-the-art performance in predicting corn and soybean yields across all evaluation scenarios, particularly during extreme years, with statistically significant improvements (paired t-test, $p < 0.0001$). Importantly, VITA outperforms prior frameworks like GNN-RNN without soil data, and larger foundational models (e.g., Chronos-Bolt) with less compute, making it practical for real-world use—especially in data-scarce regions. This work highlights how domain-aware AI design can overcome data limitations and support resilient agricultural forecasting in a changing climate.

Code & Data — https://github.com/neehan/VITA
Extended Paper — https://arxiv.org/pdf/2508.03589

Introduction

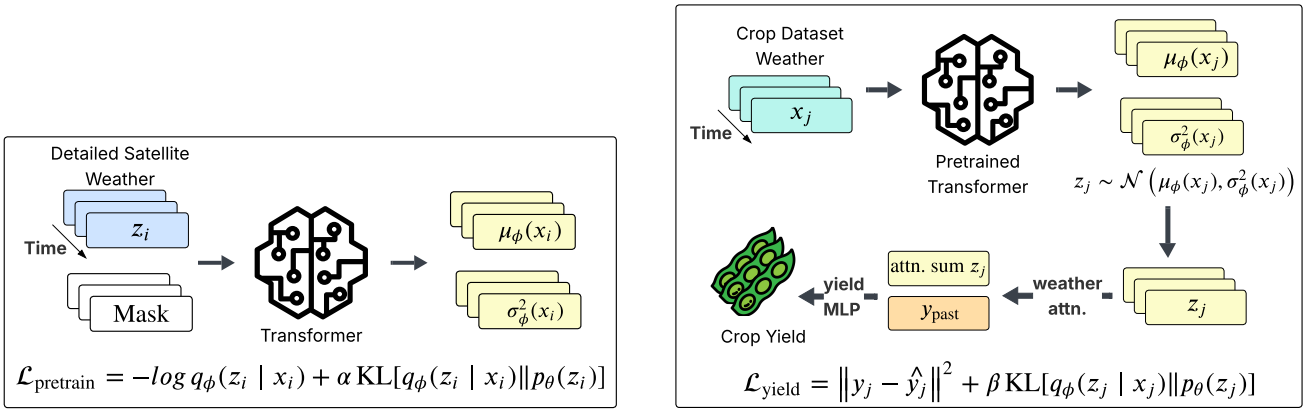
Climate change is transforming agriculture, with extreme weather causing billions in annual crop losses (Lobell, Schlenker, and Costa-Roberts 2011). In 2012, U.S. drought reduced corn yields by 13%, while 2019 flooding prevented planting on 19.4 million acres (USDA National Agricultural Statistics Service 2013; USDA Farm Service Agency 2019). Accurate yield prediction under such volatility is critical for agricultural risk management and long-term food security (Beddington 2010). Yet current operational models—including regression-based approaches from USDA ERS (Westcott and Jewison 2013)—often fail when yields diverge from historical trends.

The challenge stems from fundamental data limitations of existing methods in yield forecasting. First, many models train on less than 10 years of historical data (Gandhi, Petkar, and Armstrong 2016; Lin et al. 2024), insufficient for capturing rare but increasingly critical extreme weather patterns. Second, multi-modal approaches (You et al. 2017; Lin et al. 2024; Khaki, Wang, and Archontoulis 2019; Fan et al. 2021) rely on extensive auxiliary data—satellite imagery, soil surveys, planting records, and weather records—which limits their applicability in regions that lack detailed agricultural monitoring infrastructure.

Third, general-purpose time-series pretraining methods like SimMTM (Dong et al. 2023), PatchTST (Wu et al. 2023), and foundational models, such as Chronos (Ansari et al. 2024) assume consistent input features between pretraining and fine-tuning. This is suboptimal in weather domains, where pretraining can leverage rich satellite datasets with dozens of variables (e.g., 31 meteorological variables from NASA POWER (NASA 2024)), but fine-tuning must rely on smaller, accessible subsets (e.g., 6 basic weather variables from Khaki, Wang, and Archontoulis (2019)). We term this the *data asymmetry problem*, in which pretraining and operational feature sets fundamentally differ—an issue largely unaddressed by existing AI approaches.

To overcome this limitation, we introduce VITA, a variational pretraining framework for weather–yield prediction that transfers knowledge from rich satellite data to limited ground-based inputs. Trained with a variational loss and feature mask, VITA learns latent weather representations that generalize to settings with fewer variables. Since many crop yields are largely governed by weather, this objective enables rich representation learning essential for yield forecasting.

Unlike a standard variational autoencoder (VAE) (Kingma and Welling 2013), VITA is trained without a decoder – it maximizes the variational likelihood using detailed weather variables as proxy targets, with a sinusoidal prior to capture seasonality. During fine-tuning, six years of historical yield serve as a proxy for soil and management factors, reducing data requirements. In total, VITA adds under 2% more parameters than a standard Transformer encoder (Vaswani et al. 2017) and trains end-to-end in under 2.5 hours on a single L40S GPU. Its efficiency and exclusive reliance on public data (NASA POWER, USDA



(a) Variational pretraining

(b) Yield prediction fine-tuning

Figure 1: Two-stage variational training framework for asymmetric weather features. (a) A transformer encoder is pretrained on 31-variable weather time series by randomly masking $10 \leq k \leq 25$ features and predicting them from remaining context. The model learns a variational posterior $q_\phi(z_i \mid x_i)$ over weather representations by directly maximizing variational likelihood. (b) During fine-tuning, only 6 weather features are available. The pretrained transformer encodes these into a latent distribution $q_\phi(z_j \mid x_j)$, from which $z_j \sim \mathcal{N}(\mu_\phi(x_j), \sigma^2_\phi(x_j))$ is sampled. It is aggregated with learnable attention across time dimension and concatenated with historical yield y_{past} for final yield prediction.

yield) make it practical for operational use, including crop insurance and subsidy programs.

Empirically, VITA delivers state-of-the-art accuracy, particularly in years with extreme yield deviations, with statistically significant gains ($p < 0.0001$) over other pretraining strategies. It also generalizes strongly across space and time—models pretrained on non-U.S. weather improve U.S. fine-tuning, and those trained on 1994–2009 remain accurate for 2014–2018—demonstrating robust, transferable weather representations.

In summary, the key contributions of this work are:

- A decoder-free variational pre-training framework for modeling asymmetric features (see Equation 8).
- A seasonality-aware sinusoidal prior that captures structured temporal patterns.
- State-of-the-art performance in years with extreme yield deviations, validated through rigorous, statistically grounded evaluation.

The remainder of this paper is organized as follows: Related Work reviews prior methods; Methodology details our variational framework; Experiments presents our experimental setup; Results analyzes performance across various conditions; and Discussion examines implications and limitations.

Related Work

Crop Yield Prediction. Researchers have proposed various different approaches for crop yield forecasting, including mechanistic modeling, CNN-RNN architectures, graph neural networks, Deep Gaussian Processes, and Vision Transformers (Keating et al. 2003; Khaki, Wang, and Archontoulis 2019; Fan et al. 2021; You et al. 2017; Lin et al. 2024). Multi-modal approaches integrate satellite imagery, weather data, soil surveys, management records, and vegetation indices (Wu et al. 2021; Sun et al. 2019; Gandhi, Petkar, and Armstrong 2016; Oliveira et al. 2018; Basir et al. 2021; Hasan, Roozbehani, and Dahleh 2024; Ferraz et al. 2024; Cao, Ma, and Zhang 2022). However, extensive data requirements limit deployment in regions with limited agricultural monitoring infrastructure, and many approaches rely on temporally short regional datasets that may not capture sufficient weather variability for extreme events (Gandhi, Petkar, and Armstrong 2016; Lin et al. 2024; McFarland et al. 2020; Chu and Yu 2020).

Time Series Pretraining. Recent work develops unsupervised representation learning through contrastive methods, transformer frameworks, and masked reconstruction (Franceschi, Dieuleveut, and Jaggi 2019; Yue et al. 2022; Zerveas et al. 2021; Dong et al. 2023; Wu et al. 2023; Woo et al. 2022). These methods achieve strong forecasting performance but do not explicitly model data asymmetry, which is critical for weather-based yield forecasting.

Variational Methods. Variational autoencoders have been applied to weather prediction, climate forecasting, and agricultural data generation (Kingma and Welling 2013; Higgins et al. 2017; Kwok and Qi 2021; Wang et al. 2024; Palma et al. 2025; Razavi et al. 2024). However, existing variational methods are not used for asymmetric data learning. Our approach is tailored to this problem, learning latent states without input reconstruction and using a seasonality-aware sinusoidal prior.

Methodology

VITA incorporates a two-stage approach: (1) self-supervised pretraining on extensive weather data to learn robust weather representations, and (2) variational fine-tuning with basic weather statistics and past yields.

Problem Formulation

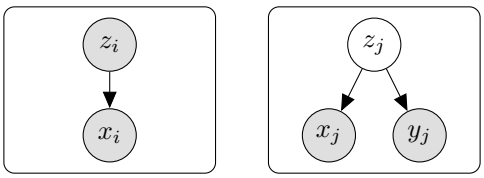


Figure 2: Graphical model showing the data structure of pretraining and prediction phases in VITA.

We formulate crop yield prediction as semi-supervised learning with latent weather representations. Suppose $z \in \mathbb{R}^{364\times 31}$ denotes the detailed meteorological variables from pretraining, and $x \in \mathbb{R}^{364\times 6}$ denotes basic weather statistics (temperature, precipitation, etc.) available in both pretraining and downstream tasks. Each weather input is a 364-week sequence representing seven years of weekly means.

We consider two datasets: (1) a pretraining dataset $\mathcal{D}_w = \{(x_i, z_i)\}_{i=1}^{N_w}$, where both basic and detailed weather states are observed for each 364-week *non-overlapping* window i across the NASA POWER grid, but no yield information is available; and (2) a finetuning dataset $\mathcal{D}_y = \{(x_j, y_j, y_{\text{past},j})\}_{j=1}^{N_y}$, where detailed states z_j remain latent for each 364-week *overlapping* sequence j across U.S. Corn Belt counties. Each training example uses weather from years $[t-6, t]$ and historical yields from $[t-6, t-1]$ to predict yield $y_{c,t}$ for county c. Test-year yields are strictly held out during training and additional details are provided in the Appendix. We model each county independently as a 1D temporal sequence and spatial context enters only via (latitude,longitude) features concatenated to each timestep (Equation 1).

Architecture

VITA employs a transformer encoder that maps 364-week weather sequences into latent representations for yield prediction. The forward process is:

$$x_{\text{input}} = \text{concat}(x_{\text{weather}}, \text{year}, \text{coordinates}) \tag{1}$$

$$h_{\text{weather}} = E_\phi\big(\text{LinearProj}(x_{\text{input}}) + \text{PosEmbed}(\cdot)\big) \tag{2}$$

$$[\mu, \log\sigma^2] = \text{LinearProj}_{\mu,\sigma^2}(h_{\text{weather}}) \tag{3}$$

$$z = \mu + \sigma \odot \epsilon, \quad \epsilon \sim \mathcal{N}(0, I) \tag{4}$$

$$z_{\text{agg}} = \sum_{k=1}^{364} \alpha_k z_k;\ \alpha_k = \text{softmax}\big(\text{MLP}_a(z_k)\big) \tag{5}$$

$$\hat{y}_t = \text{MLP}_y\big([\, z_{\text{agg}},\, y_{\text{past}}\,]\big) \tag{6}$$

Here, E_ϕ is a transformer encoder applied to weekly weather inputs x_{input} with positional embeddings, and coordinates refer to the spatial latitude and longitude. The $\text{LinearProj}_{\mu,\sigma^2}$ outputs the mean and log-variance of each latent state; MLP_a computes attention weights for temporal aggregation into z_{agg}; and MLP_y maps the aggregated latent representation and historical yields to the predicted yield $\hat{y}_t$.

Self-Supervised Pretraining

We pretrain on the NASA POWER dataset (1984-2022) using progressive feature-wise masking, starting with $k = 10$ masked features and increasing by 1 every 2 epochs until 25 out of 31 features are masked. The pretraining objective balances reconstruction with regularization:

$$\begin{aligned}\mathcal{L}_{\text{pretrain}} &= -\,\mathbb{E}_{(x_i,z_i)\sim\mathcal{D}_w}\,[\log q_\phi(z_i \mid x_i) \\ &\quad + \alpha \cdot \text{KL}[q_\phi(z_i \mid x_i) \| p_\theta(z_i)]] \\ &= -\,\mathbb{E}_{(x_i,z_i)\sim\mathcal{D}_w}\,\big[\log \mathcal{N}(z_i; \mu_\phi(x_i), \sigma^2_\phi(x_i))\big] \\ &\quad + \alpha \cdot \text{KL}[q_\phi(z_i \mid x_i) \| p_\theta(z_i)]\end{aligned} \tag{7}$$

The first term maximizes the Gaussian likelihood, which encourages the posterior distribution $q_\phi(z_i \mid x_i)$ to accurately predict the observed detailed weather state z_i. The second term is a regularizer preventing overfitting by imposing a prior structure.

We investigate two prior distributions: standard normal $p_\theta(z) \sim \mathcal{N}(0, I)$ and sinusoidal prior $p_\theta(z) \sim \mathcal{N}(\mu_{\sin} = A\sin(\theta \cdot \text{pos} + \theta_0), \sigma^2 I)$ to capture seasonal patterns. The Sinusoidal prior has additional prior parameters that are learned during both pretraining and finetuning. This prior explicitly models the periodicity in weather variables, allowing a more structured latent space.

Decoder-Free Fine-Tuning Objective

Standard semi-supervised VAEs require a decoder term $\log p(x_j|z_j)$ to model input reconstruction (Kingma et al. 2014). However, in our meteorological context, established principles like the Tetens equation, Penman-Monteith formulation, Clausius-Clapeyron equation and Stefan-Boltzmann radiation balance link basic weather statistics to the detailed atmospheric state (O. Tetens 1930; Ndulue and Ranjan 2021; Brown 1951; Murray-Tortarolo 2023), making the decoder term unnecessary.

In the Appendix, we empirically validated this deterministic relationship, training a small MLP to reconstruct basic weather variables from detailed ones with near-perfect accuracy ($R^2 > 0.9999$). This enables us to model $p(x_j|z_j) \approx 1$ and derive the simplified variational objective:

$$\mathcal{L}_{\text{yield}} = \|y_j - \hat{y}_j\|^2 + \beta \cdot \text{KL}[q_\phi(z_j \mid x_j) \| p_\theta(z_j)] \tag{8}$$

where $\beta > 0$ is a hyperparameter. Note that the β in Equation (8) does not weaken the variational objective and the full evidence lower bound (ELBO) is still optimized. The full derivation of this objective is shown in the Appendix.

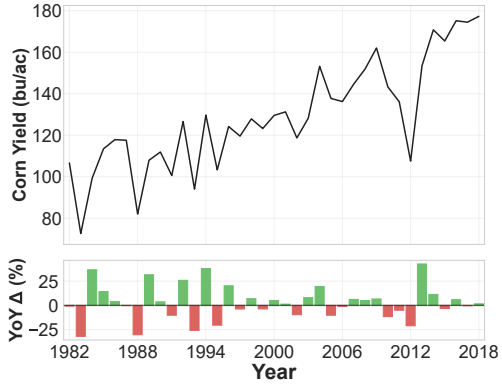
Baselines

We compare against several types of baselines. (1) *Non-deep learning methods:* OLS linear regression with

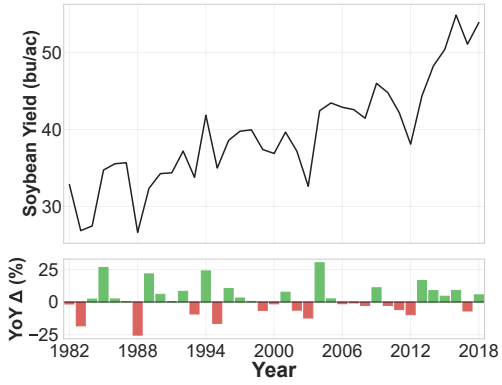
agronomically-motivated features following USDA Economic Research Service (ERS) methodology (Westcott and Jewison 2013), and XGBoost (Chen and Guestrin 2016). (2) *Deep learning methods:* CNN-RNN (Khaki, Wang, and Archontoulis 2019) and GNN-RNN (Fan et al. 2021). (3) *Masked time series pretraining:* SimMTM (Dong et al. 2023). (4) *Pre-trained foundational time series models:* Chronos-Bolt-tiny-9M (Ansari et al. 2024) with full fine-tuning. To isolate the effect of variational pretraining from architecture, we also include T-BERT, which is identical to VITA but trained with a standard MSE reconstruction loss. Notably, XGBoost, GNN-RNN, and CNN-RNN have access to soil data, while the pretrained transformer models and OLS do not. All models except the OLS use identical temporal windows and train-test splits.

Experiments

We test three hypotheses: (1) weather pretraining improves yield prediction on extreme weather years, (2) variational objectives with sinusoidal priors outperform standard approaches, and (3) these benefits generalize to standard years and forward temporal gaps. We evaluate on county-level corn and soybean yield prediction across 763 US Corn Belt counties.



(a) Corn yield trends by year



(b) Soybean yield trends by year

Figure 3: Mean crop yield in bushels per acre (bu/ac) 763 US Corn Belt counties showing extreme weather years as sharp deviations from historical patterns.

Data. We pretrain on the NASA POWER dataset (NASA 2024), comprising 39 years (1984–2022) of climate data at 0.5° resolution across 116 grids over the Americas. It includes 31 meteorological variables aggregated weekly (100K sequences). However, its resolution is too coarse for county-level yield forecasting, as the median U.S. county spans 622 square miles (U.S. Census Bureau 2025). We therefore evaluate on the dataset of Khaki, Wang, and Archontoulis (2019), containing weather, soil, and county-level corn and soybean yields for 763 Corn Belt counties (1982–2018) (U.S. Department of Agriculture, National Agricultural Statistics Service 2023). Corn (C4) and soybean (C3) represent distinct physiological and weather-sensitivity regimes and jointly account for over 60% of U.S. row-crop acreage (Williams and Pounds-Barnett 2024).

The Khaki, Wang, and Archontoulis (2019) weather data include six ground-based weekly weather measurements: (1) minimum temperature, (2) maximum temperature, (3) solar radiation, (4) precipitation, (5) snow water equivalent, and (6) vapor pressure, and 11 soil properties averaged over county areas. We exclude soil data to test deployment in data-sparse regions. This presents a key challenge, as pretraining contains satellite measurements of 31 meteorological variables, while fine-tuning uses only 6 ground-based variables averaged over county areas.

Training Configuration. Full hyperparameters, schedules, and data splits are detailed in Appendix; all models share identical splits and computational budgets. All experiments, including pretraining and grid search, share random seed 1234, unless otherwise noted.

Hyperparameter Optimization. We performed a 27-configuration grid search to optimize hyperparameters, with full details and robustness analysis in Appendix. Best hyperparameters are used for all subsequent experiments.

Extreme Year Evaluation (Primary Contribution)

We identify the five most weather-extreme years for each crop between 2000 and 2018 by computing absolute z-scores from 5-year rolling means of yields. These include known drought years (2002, 2003, 2012) and years with favorable conditions and record-breaking yields (2004, 2009). (National Drought Mitigation Center, NOAA, and USDA 2025)

We also conduct an early-season forecasting experiment using our top four models (two VITA variants, T-BERT, and SimMTM) and OLS, truncating weather data at the end of July in the final year (week 30). This design follows the USDA Economic Research Service (ERS) framework (Westcott and Jewison 2013), which is conceptually similar to NASS models that also rely on regression-driven estimates grounded in observed weather and crop conditions. The OLS baseline normally incorporates July–August temperature and precipitation for soybean, but for consistency across methods we limit it to July in this experiment. More details is provided in the Appendix.

Method	Corn R^2 (RMSE)	Soybean R^2 (RMSE)	Mean R^2
OLS	0.227 (27.7)	0.460 (7.0)	0.344
XGBoost	0.135 ± 0.033 (29.0 ± 0.6)	0.377 ± 0.039 (7.6 ± 0.2)	0.256
CNN-RNN	0.256 ± 0.030 (26.5 ± 0.5)	0.498 ± 0.023 (6.8 ± 0.2)	0.377
GNN-RNN	0.564 ± 0.051 (20.2 ± 1.0)	0.640 ± 0.007 (5.7 ± 0.1)	0.602
Chronos-Bolt-tiny	0.525 ± 0.015 (21.6 ± 0.3)	0.621 ± 0.017 (6.0 ± 0.1)	0.573
SimMTM	0.642 ± 0.028 (18.8 ± 0.7)	0.687 ± 0.018 (5.3 ± 0.1)	0.665
T-BERT (ours)	0.660 ± 0.041 (18.3 ± 1.0)	0.693 ± 0.011 (5.3 ± 0.1)	0.677
VITA-Std. Normal (ours)	0.706 ± 0.025 (17.1 ± 0.7)	0.698 ± 0.020 (5.2 ± 0.2)	0.702
VITA-Sinusoidal (ours)	**0.729 ± 0.008 (16.3 ± 0.2)**	**0.722 ± 0.005 (5.0 ± 0.1)**	**0.726**

Table 1: Performance on the 5 most extreme years. Results averaged across 3 random seeds (1234, 5678, 2025) and best results **bolded**. The OLS baseline shows no change since it is deterministic.

Standard Years Generalization

To validate that extreme weather optimization doesn't compromise standard performance, we evaluate on 2014–2018 using hyperparameters optimized for extreme years. We test both 15-year and 30-year training periods to assess data efficiency requirements.

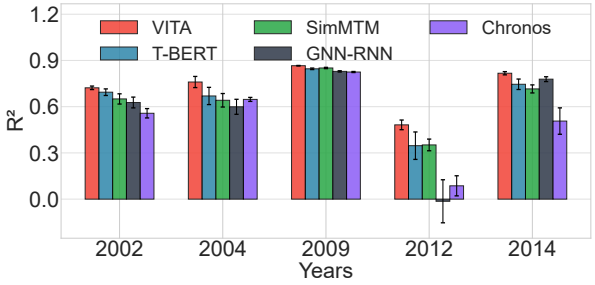
Forward Gap Robustness

We also evaluate forward gap robustness across five experiments with 5-year gaps: train on 1994–2009/test on 2014, train on 1995–2010/test on 2015, and so forth through 2018.

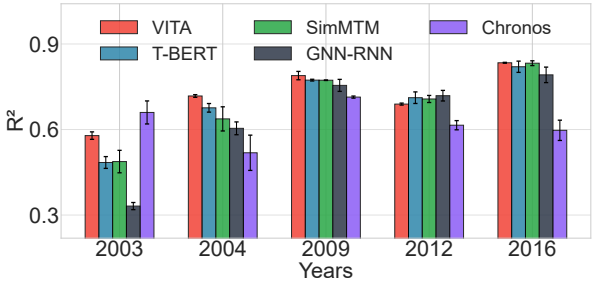
Ablation Studies

We ablate: (1) pretraining vs. random initialization, (2) variational vs. MSE objective, (3) sinusoidal vs. normal priors, and (4) spatial generalization by pretraining on weather grids excluding continental USA, focusing on extreme weather performance.

All pretraining experiments run on four L40S GPUs and all finetuning experiments run on one L40S GPU with identical computational budgets across methods. The code, pretrained models, and datasets will be publicly released upon publication to facilitate reproducibility and broader agricultural AI research.



(a) Corn



(b) Soybean

Figure 4: VITA-Sinusoidal shows consistent improvement over other baselines.

Results

Extreme Year Performance

VITA-Sinusoidal achieves 0.729 ± 0.008 R^2 for corn and 0.722 ± 0.005 R^2 for soybean on the five most extreme weather years (Table 1), representing +10.5% and +4.2% improvements over T-BERT (0.660 and 0.693 respectively). These gains translate to 2.0 bu/ac corn and 0.3 bu/ac soybean RMSE reductions over T-BERT—critically important during droughts like 2012 when accurate forecasts inform billions in crop insurance decisions. The remarkably low variance across random seeds demonstrates the approach's stability.

Figure 4 reveals VITA outperforms T-BERT on 8/10 individual extreme years and surpasses both SimMTM and Chronos-Bolt on 9/10 evaluations (paired t-test across 30 instances from 3 seeds, $p < 0.0001$). The rare underperformances occur when baselines already achieve high accuracy (soybean 2012: 0.689 R^2, 2016: 0.834 R^2), leaving minimal room for improvement. Notably, Chronos-Bolt—despite being 4.5× larger (9M vs 2M parameters) and pretrained 890k sequences (Ansari et al. 2024)—struggles due to the data asymmetry problem.

We also note that the traditional methods fail catastrophically (OLS: 0.227 corn R^2, XGBoost: 0.135 R^2 with soil data) during extreme years, while soil-enriched deep learning shows moderate success (GNN-RNN: 0.564 corn, CNN-RNN: 0.256). Among transformer models, pretraining strategy determines success—SimMTM's temporal masking (0.642 corn) and Chronos's general-purpose approach (0.525 corn) both underperform domain-specific variational pretraining. Comparing T-BERT (0.660) to VITA-Std Normal (0.706) and VITA-Sinusoidal (0.729), we observe that, variational objectives provide +7% improvement, with si-

Method	Corn 15yr R^2 (RMSE)	Corn 30yr R^2 (RMSE)	Soybean 15yr R^2 (RMSE)	Soybean 30yr R^2 (RMSE)
OLS	0.515 (24.5)	0.508 (24.1)	0.673 (6.0)	0.660 (6.1)
XGBoost	0.439 (27.4)	0.310 (29.2)	0.602 (6.5)	0.564 (7.1)
CNN-RNN	0.659 (20.7)	0.635 (20.8)	0.721 (5.6)	0.671 (6.0)
GNN-RNN	0.788 (16.6)	0.785 (16.5)	0.800 (4.7)	0.810 (4.6)
Chronos-Bolt-tiny	0.704 (19.5)	0.693 (19.7)	0.704 (5.7)	0.724 (5.6)
SimMTM	0.753 (17.9)	0.768 (17.2)	0.814 (4.6)	0.822 (4.5)
T-BERT (ours)	0.791 (16.5)	0.780 (16.8)	0.831 (4.4)	**0.837 (4.3)**
VITA-Sinusoidal (ours)	**0.827 (16.0)**	**0.837 (15.5)**	**0.833 (4.3)**	**0.837 (4.2)**

Table 2: Standard years (2014-2018) performance. Best results **bolded**.

Crop	Model	Full R^2 (RMSE)	Week 30 R^2 (RMSE)
Corn	OLS	0.227 (27.7)	0.227 (27.7)
	SimMTM	0.642 (18.8)	0.568 (20.5)
	T-BERT	0.660 (18.3)	0.589 (20.2)
	VITA-Std. Norm.	0.706 (17.1)	0.642 (18.9)
	VITA-Sinusoidal	0.729 (16.3)	**0.689 (17.6)**
Soybean	OLS	0.460 (7.0)	0.382 (7.5)
	SimMTM	0.687 (5.3)	0.481 (6.8)
	T-BERT	0.693 (5.3)	0.508 (6.7)
	VITA-Std. Norm.	0.698 (5.2)	0.551 (6.3)
	VITA-Sinusoidal	0.722 (5.0)	**0.560 (6.2)**

Table 3: Early-season forecasting on extreme years for the top 4 models and OLS. Full: 7 years weather through end of season (52 weeks). Week 30: cutoff at end of July, ie. week 30 of the final year.

nusoidal priors adding another +3% by capturing seasonal structure.

Operational viability is confirmed through two stress tests. With weather cut off at the end of July (week 30 of the final year), VITA-Sinusoidal maintains strong performance, achieving 0.689 R^2 on corn and 0.560 R^2 – nearly 3.0× and 1.5× better than OLS baselines, respectively (Table 3). The performance drop for soybean is particularly sharp across all models, as its late pod-filling stage makes yields highly sensitive to August rainfall and temperature stress. (Westcott and Jewison 2013)

Lastly, we note that the XGBoost model underperforms in all instances due to correlated weather features, and will require feature engineering to get competitive performance.

Reduced temporal context (5 years vs 7 years) shows VITA-Sinusoidal achieves 0.697 R^2 corn and 0.684 R^2 soybean, outperforming most baselines despite using 28% less data (Appendix).

Standard Years and Forward Gap Robustness

Standard years (2014-2018) validate that extreme-weather optimization doesn't compromise performance under normal conditions. VITA achieves 0.827 R^2 for corn and 0.833 R^2 for soybean with 15-year training—improvements of +4.6% and +0.2% over T-BERT respectively. The contrast with extreme years (+10.5% corn improvement) supports our core hypothesis: variational uncertainty modeling provides greatest value precisely when predictions are hardest. With 30-year training, corn performance rises to 0.837 R^2 (+7.3% over T-BERT), suggesting additional historical data helps but yields diminishing returns compared to better representations.

Five-year temporal shift (training on 1994-2009, testing on 2014-2018) stresses whether learned patterns generalize beyond training conditions or merely memorize era-specific correlations. VITA achieves 0.797 R^2 corn and 0.819 R^2 soybean—maintaining the +1.9% and +2.0% margins over T-BERT seen in standard evaluation. Meanwhile, the CNN-RNN and GNN-RNN models suffer noticeable degradation (0.718 corn, down from 0.788), revealing their soil-enhanced features do not help with temporal robustness.

Ablation Studies

Pretraining and Spatial Transfer

Pretraining is critical for VITA's performance. Without pretraining, both VITA variants show high variance across hyperparameters (±0.23 corn R^2, ±0.14 soybean R^2) and achieve only 0.47 and 0.57 mean R^2 respectively (Table 5). Pretraining on Central and South American weather—excluding all US continental data—provides substantial improvements of +34-37% corn and +17-19% soybean ($t = 3.61 - 3.82, p < 0.01$), demonstrating that VITA learns universal weather-agriculture relationships rather than region-specific patterns.

Full Americas pretraining (including US weather outside Corn Belt counties and target years) further improves performance to approximately 0.70 R^2 for both crops with continued low variance (±0.015-0.019). The +50% corn and +22% soybean improvements over no pretraining ($t = 5.35 - 5.49, p < 0.001$) highlight that variational objectives are particularly initialization-sensitive—they either converge to strong solutions with pretrained weights or fail to escape poor local minima without them. For comparison, T-BERT shows smaller but still significant pretraining gains (+10.8% corn, +5.0% soybean, $t = 9.13, p < 0.001$) with consistently lower variance across all conditions, as detailed in Appendix.

Method	Corn R^2 (RMSE)	Soybean R^2 (RMSE)	Mean R^2
OLS	0.471 (25.6)	0.634 (6.4)	0.552
XGBoost	0.159 (34.4)	0.433 (8.2)	0.296
CNN-RNN	0.556 (23.9)	0.659 (6.2)	0.608
GNN-RNN	0.718 (18.9)	0.785 (4.9)	0.752
Chronos-Bolt-tiny	0.631 (21.7)	0.685 (5.9)	0.658
SimMTM	0.705 (19.7)	0.776 (5.0)	0.741
T-BERT (ours)	0.782 (16.9)	0.803 (4.7)	0.793
VITA-Sinusoidal (ours)	**0.797 (16.3)**	**0.819 (4.5)**	**0.808**

Table 4: Forward gap robustness: 5-year temporal shift (train: 1994-2009, test: 2014-2018). Best results **bolded**.

Prior	Pretraining	Corn R^2	Soybean R^2	t-stat / p-value
Std. Normal	None	0.463 ± 0.230	0.575 ± 0.134	-
	Non-US	0.632 ± 0.018	0.674 ± 0.020	$3.82 / 7.5 \times 10^{-4}$
	Full	0.706 ± 0.015	0.698 ± 0.017	$5.49 / 9.3 \times 10^{-6}$
Sinusoidal	None	0.469 ± 0.227	0.569 ± 0.139	-
	Non-US	0.627 ± 0.020	0.669 ± 0.021	$3.61 / 1.3 \times 10^{-3}$
	Full	0.703 ± 0.015	0.698 ± 0.019	$5.35 / 1.3 \times 10^{-5}$

Table 5: VITA pretraining ablation across 27 hyperparameter configurations (random seed 1234).

Discussion

VITA-Sinusoidal achieves statistically significant improvements over architecturally identical T-BERT baselines ($p < 0.0001$), with gains most pronounced during extreme weather when accurate predictions are most critical. This superior performance stems from rich latent representations that avoid the collapse seen in T-BERT (15.7% vs. 84.0% variance in top two PCA components; Appendix), enabling better differentiation between normal and extreme conditions. VITA exceeds GNN-RNN despite lacking soil data, demonstrating that historical yields can proxy soil characteristics when paired with rich weather representations.

Operational Context and Social Impact. Current USDA operational forecasts—including ERS regression models (Westcott and Jewison 2013)—rely on simpler models with hand-crafted features, achieving 0.227 R^2 for extreme corn years with our OLS baseline. VITA's 3.2× improvement (0.729 R^2) translates to substantial impact: 11.4 bu/ac RMSE reduction over OLS and 2.0 bu/ac over T-BERT across 88.7M Corn Belt acres (National Corn Growers Association 2024). At \$4.70/bushel (DTN Progressive Farmer 2025), this translates to \$4.75B and \$800M in value, respectively. This accuracy is critical for the Federal Crop Insurance Program managing billions in premiums (USDA Economic Research Service 2024) and for policy responses during droughts. Deployment requires minimal infrastructure (single GPU, 2.5 hours training) with only public data, enabling integration into existing agricultural statistics systems.

Spatial Transfer and Global Food Security. Pretraining on Central/South American weather—climatically distinct from the U.S. Corn Belt—significantly improves U.S. predictions ($t = 3.61$, $p < 0.01$; Table 5). This cross-continental transfer demonstrates that VITA learns universal weather-agriculture relationships (temperature stress, precipitation deficits, radiation anomalies) rather than region-specific patterns. For global food security, models pretrained on data-rich regions can enhance predictions in data-scarce areas despite climatic differences.

Broader Applicability and Limitations. The decoder-free variational framework can ba applied beyond agriculture to any setting with rich sensors at training and sparse sensors at inference (e.g., ICU vitals vs. labs, IoT vs. industrial telemetry). Current evaluation focuses on U.S. Corn Belt corn and soybean (Khaki, Wang, and Archontoulis 2019; Fan et al. 2021); extending to other crops and regions remains future work. We open-sourced implementation, preprocessing scripts, model weights, and documentation for new region adaptation. Data privacy concerns are negligible as we use aggregated public meteorological data, ensuring equitable access through transparency.

Conclusion

We introduced VITA, a variational pretraining framework for forecasting crop yield under extreme weather. It is pretrained on satellite weather datasets through decoder-free variational learning and achieves state-of-the-art performance on extreme weather years (R^2 = 0.729 for corn, 0.722 for soybeans). VITA requires only basic weather variables available globally, allowing deployment in data-scarce regions where multi-modal approaches are infeasible. Our comprehensive evaluation across 763 U.S. Corn Belt counties with statistical validation and spatial transfer experiments demonstrates robust, real performance.

References

Ansari, A. F.; Stella, L.; Turkmen, C.; Zhang, X.; Mercado, P.; Shen, H.; Shchur, O.; Rangapuram, S. S.; Arango, S. P.; Kapoor, S.; Zschiegner, J.; Maddix, D. C.; Wang, H.; Mahoney, M. W.; Torkkola, K.; Wilson, A. G.; Bohlke-Schneider, M.; and Wang, Y. 2024. Chronos: Learning the Language of Time Series. arXiv:2403.07815.

Basir, M. S.; Chowdhury, M.; Islam, M. N.; and Ashik-E-Rabbani, M. 2021. Artificial neural network model in predicting yield of mechanically transplanted rice from transplanting parameters in Bangladesh. *Journal of Agriculture and Food Research*, 5: 100186.

Beddington, J. 2010. Food security: contributions from science to a new and greener revolution. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 365(1537): 61–71.

Brown, O. L. I. 1951. The Clausius-Clapeyron equation. *Journal of Chemical Education*, 28(8): 428.

Cao, Z.; Ma, Y.; and Zhang, Z. 2022. Corn Yield Prediction based on Remotely Sensed Variables Using Variational Autoencoder and Multiple Instance Regression. arXiv:2211.13286.

Chen, T.; and Guestrin, C. 2016. XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, 785–794. ACM.

Chu, Z.; and Yu, J. 2020. An end-to-end model for rice yield prediction using deep learning fusion. *Computers and Electronics in Agriculture*, 174: 105471.

Dong, J.; Wu, H.; Zhang, H.; Zhang, L.; Wang, J.; and Long, M. 2023. SimMTM: A Simple Pre-Training Framework for Masked Time-Series Modeling. arXiv:2302.00861.

DTN Progressive Farmer. 2025. 2025 Spring Crop Insurance Price Discovery. Accessed: 2025-10-02.

Fan, J.; Bai, J.; Li, Z.; Ortiz-Bobea, A.; and Gomes, C. P. 2021. A GNN-RNN Approach for Harnessing Geospatial and Temporal Information: Application to Crop Yield Prediction. *CoRR*, abs/2111.08900.

Ferraz, M. A. J.; Barboza, T. O. C.; Piza, M. R.; Von Pinho, R. G.; and dos Santos, A. F. 2024. Sorghum grain yield estimation based on multispectral images and neural network in tropical environments. *Smart Agricultural Technology*, 9: 100661.

Franceschi, J.-Y.; Dieuleveut, A.; and Jaggi, M. 2019. Unsupervised scalable representation learning for multivariate time series. In *Advances in Neural Information Processing Systems*, volume 32.

Gandhi, N.; Petkar, O.; and Armstrong, L. J. 2016. Rice crop yield prediction using artificial neural networks. In *2016 IEEE Technological Innovations in ICT for Agriculture and Rural Development (TIAR)*, 105–110.

Hasan, A.; Roozbehani, M.; and Dahleh, M. 2024. WeatherFormer: A Pretrained Encoder Model for Learning Robust Weather Representations from Small Datasets. arXiv:2405.17455.

Higgins, I.; Matthey, L.; Pal, A.; Burgess, C.; Glorot, X.; Botvinick, M.; Mohamed, S.; and Lerchner, A. 2017. beta-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework. In *International Conference on Learning Representations*.

Keating, B.; Carberry, P.; Hammer, G.; Probert, M.; Robertson, M.; Holzworth, D.; Huth, N.; Hargreaves, J.; Meinke, H.; Hochman, Z.; McLean, G.; Verburg, K.; Snow, V.; Dimes, J.; Silburn, M.; Wang, E.; Brown, S.; Bristow, K.; Asseng, S.; Chapman, S.; McCown, R.; Freebairn, D.; and Smith, C. 2003. An overview of APSIM, a model designed for farming systems simulation. *European Journal of Agronomy*, 18(3): 267–288. Modelling Cropping Systems: Science, Software and Applications.

Khaki, S.; Wang, L.; and Archontoulis, S. V. 2019. A CNN-RNN Framework for Crop Yield Prediction. *Frontiers in Plant Science*, 10.

Kingma, D. P.; Mohamed, S.; Jimenez Rezende, D.; and Welling, M. 2014. Semi-supervised learning with deep generative models. *Advances in neural information processing systems*, 27.

Kingma, D. P.; and Welling, M. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Kwok, P. H.; and Qi, Q. 2021. A Variational U-Net for Weather Forecasting. *CoRR*, abs/2111.03476.

Lin, F.; Guillot, K.; Crawford, S.; Zhang, Y.; Yuan, X.; and Tzeng, N.-F. 2024. An Open and Large-Scale Dataset for Multi-Modal Climate Change-aware Crop Yield Predictions. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 5375–5386. ACM.

Lobell, D. B.; Schlenker, W.; and Costa-Roberts, J. 2011. Climate trends and global crop production since 1980. *Science*, 333(6042): 616–620.

McFarland, B. A.; AlKhalifah, N.; Bohn, M.; Bubert, J.; Buckler, E. S.; Ciampitti, I.; Edwards, J.; Ertl, D.; Gage, J. L.; Falcon, C. M.; Flint-Garcia, S.; Gore, M. A.; Graham, C.; Hirsch, C. N.; Holland, J. B.; Hood, E.; Hooker, D.; Jarquin, D.; Kaeppler, S. M.; Knoll, J.; Kruger, G.; Lauter, N.; Lee, E. C.; Lima, D. C.; Lorenz, A.; Lynch, J. P.; McKay, J.; Miller, N. D.; Moose, S. P.; Murray, S. C.; Nelson, R.; Poudyal, C.; Rocheford, T.; Rodriguez, O.; Romay, M. C.; Schnable, J. C.; Schnable, P. S.; Scully, B.; Sekhon, R.; Silverstein, K.; Singh, M.; Smith, M.; Spalding, E. P.; Springer, N.; Thelen, K.; Thomison, P.; Tuinstra, M.; Wallace, J.; Walls, R.; Wills, D.; Wisser, R. J.; Xu, W.; Yeh, C. T.; and de Leon, N. 2020. Maize genomes to fields (G2F): 2014-2017 field seasons: genotype, phenotype, climatic, soil, and inbred ear image datasets. *BMC Research Notes*, 13(1): 71.

Murray-Tortarolo, G. 2023. A breviary of Earth's climate changes using Stephan-Boltzmann law. *Atmósfera*, 37.

NASA. 2024. NASA Power API.

National Corn Growers Association. 2024. Corn Production: USDA expects U.S. farmers to harvest about 88.7 million acres of corn for grain. Accessed: 2025-10-02.

National Drought Mitigation Center; NOAA; and USDA. 2025. U.S. Drought Monitor. Accessed: 2025-07-31.

Ndulue, E.; and Ranjan, R. S. 2021. Performance of the FAO Penman-Monteith equation under limiting conditions and fourteen reference evapotranspiration models in southern Manitoba. *Theoretical and Applied Climatology*, 143(3): 1285–1298.

O. Tetens. 1930. Über einige meteorologische Begriffe (On some meteorological terms). *Z. Geophys.*, 6: 297–309.

Oliveira, I.; Cunha, R. L. F.; Silva, B.; and Netto, M. A. S. 2018. A Scalable Machine Learning System for Pre-Season Agriculture Yield Forecast. arXiv:1806.09244.

Palma, L.; Peraza, A.; Civantos, D.; Duarte, A.; Materia, S.; Ángel G. Muñoz; Peña-Izquierdo, J.; Romero, L.; Soret, A.; and Donat, M. G. 2025. Data-driven Seasonal Climate Predictions via Variational Inference and Transformers. arXiv:2503.20466.

Razavi, M. A.; Nejadhashemi, A. P.; Majidi, B.; Razavi, H. S.; Kpodo, J.; Eeswaran, R.; Ciampitti, I.; and Prasad, P. V. 2024. Enhancing crop yield prediction in Senegal using advanced machine learning techniques and synthetic data. *Artificial Intelligence in Agriculture*, 14: 99–114.

Sun, J.; Di, L.; Sun, Z.; Shen, Y.; and Lai, Z. 2019. County-Level Soybean Yield Prediction Using Deep CNN-LSTM Model. *Sensors*, 19(20): 4363.

U.S. Census Bureau. 2025. Gazetteer Files: National Counties Gazetteer File. https://www.census.gov/geographies/reference-files/time-series/geo/gazetteer-files.html. County land and water area (ALAND_SQMI, AWATER_SQMI); median county land area can be computed from this file (author's calculation). Page last revised 2025-09-10.

U.S. Department of Agriculture, National Agricultural Statistics Service. 2023. Quick Stats Database. https://quickstats.nass.usda.gov/. Accessed: 2025-07-08.

USDA Economic Research Service. 2024. Crop Insurance at a Glance. Accessed: 2025-10-02.

USDA Farm Service Agency. 2019. Report: Farmers Prevented from Planting Crops on 19 Million Acres. https://www.fsa.usda.gov/news-events/news/08-12-2019/report-farmers-prevented-planting-crops-19-million-acres. Accessed: 2025-07-22.

USDA National Agricultural Statistics Service. 2013. Crop Production 2012 Summary. https://www.nass.usda.gov/Newsroom/archive/2013/01_11_2013.php. Accessed: 2025-07-22.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; and Polosukhin, I. 2017. Attention Is All You Need. arXiv:1706.03762.

Wang, W.; Zhang, J.; Su, Q.; et al. 2024. Accurate initial field estimation for weather forecasting with a variational constrained neural network. *npj Climate and Atmospheric Science*, 7(1): 223.

Westcott, P. C.; and Jewison, M. 2013. Weather Effects on Expected Corn and Soybean Yields. Technical Report FDS-13g-01, U.S. Department of Agriculture, Economic Research Service.

Williams, B.; and Pounds-Barnett, G. 2024. As U.S. Farmers Respond to Crop Price Changes, Trends in Planted Acreage Emerge. *Amber Waves*.

Woo, G.; Liu, C.; Sahoo, D.; Kumar, A.; and Hoi, S. 2022. CoST: Contrastive Learning of Disentangled Seasonal-Trend Representations for Time Series Forecasting. arXiv:2202.01575.

Wu, H.; Xu, J.; Wang, J.; and Long, M. 2023. Patchtst: A general-purpose patch-based transformer for time series forecasting. In *The Eleventh International Conference on Learning Representations*.

Wu, X.; Xiao, X.; Steiner, J.; Yang, Z.; Qin, Y.; and Wang, J. 2021. Spatiotemporal Changes of Winter Wheat Planted and Harvested Areas, Photosynthesis and Grain Production in the Contiguous United States from 2008–2018. *Remote Sensing*, 13(9).

You, J.; Li, X.; Low, M.; Lobell, D.; and Ermon, S. 2017. Deep Gaussian process for crop yield prediction based on remote sensing data. *arXiv preprint arXiv:1704.02720*.

Yue, Z.; Wang, Y.; Pang, J.; Zhang, F.; Yang, W.; Sun, L.; Li, J.; Wang, J.; and Zhang, Y. 2022. Ts2vec: Towards universal representation of time series. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, 8980–8988.

Zerveas, G.; Jayaraman, S.; Patel, D.; Bhamidipaty, A.; and Eickhoff, C. 2021. A transformer-based framework for multivariate time series representation learning. In *Proceedings of the 27th ACM SIGKDD conference on knowledge discovery & data mining*, 2114–2124.