

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

Adib Hasan¹ Mardavij Roozbehani² Munther A. Dahleh²

¹Independent Researcher ²Massachusetts Institute of Technology

AAAI 2026

<https://github.com/Neehan/VITA>

Motivation: Climate Change Threatens Agricultural Forecasting

The Challenge

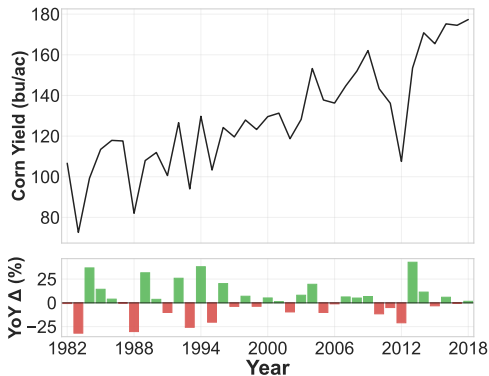
Climate change is increasing extreme weather frequency.
When predictions matter most, current models fail.

Real-world consequences:

- ▶ **2012 U.S. drought:** Corn yields dropped **27%** from 5-year mean
- ▶ **2019 Mid-West flooding:** 19.4M acres unplanted
- ▶ Federal Crop Insurance manages **billions** in premiums

Current State

OLS model similar to USDA ERS operational models achieve only **0.227 R^2** on extreme years



Mean corn yield across 763 U.S. Corn Belt counties. Note the sharp 2012 deviation.

The Data Asymmetry Problem: A New Challenge

We Identify a Fundamental Limitation

Available pretraining and deployment datasets have **different feature sets**—a problem we term **data asymmetry**.

Pretraining data (NASA POWER satellite):

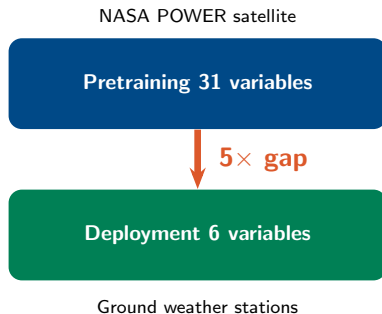
- ▶ **31 meteorological variables** at 0.5° resolution
- ▶ Radiation fluxes, humidity, wind speed, surface pressure...

Deployment data (Daymet ground stations):

- ▶ Only **6 basic variables** at $\sim 0.01^\circ$ resolution
- ▶ Min/max temperature, precipitation, solar radiation

Why This Matters

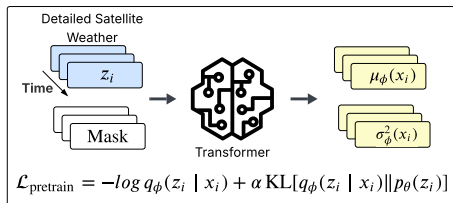
SimMTM, Chronos, PatchTST all assume **consistent features**. This assumption **fundamentally limits** their applicability to real-world agricultural forecasting.



Our Contribution: We formalize this problem and propose a principled solution.

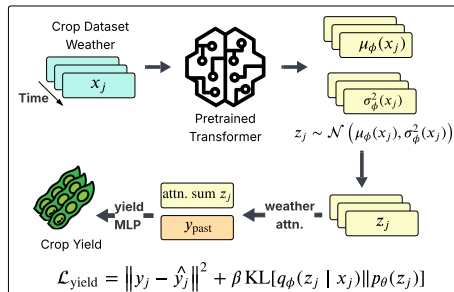
VITA: Two-Stage Variational Framework

Stage 1: Variational Pretraining



- ▶ NASA POWER (31 variables)
- ▶ Progressive masking: 10 \rightarrow 25 features
- ▶ Learn to infer atmospheric state from limited context

Stage 2: Yield Prediction Fine-tuning



- ▶ Only 6 weather features available
- ▶ Historical yields proxy soil/management
- ▶ Latent distribution captures uncertainty

Core Idea: VITA learns to **reconstruct masked features with uncertainty** during pretraining, then transfers this representation to deployment where only sparse features are available.

Innovation 1: Decoder-Free Variational Learning

Standard VAE Problem

Requires a decoder term $p(x|z)$ to reconstruct inputs during training

Our insight: Physical laws deterministically link basic weather to detailed atmospheric states:

- ▶ Clausius–Clapeyron / Tetens (vapor pressure)
- ▶ Humidity Identities (relative humidity)
- ▶ Stefan–Boltzmann (thermal longwave emission)

Empirical validation: MLP reconstructs basic from detailed weather with $R^2 > 0.9999$

Result

This allows $p(x|z) \approx 1$, eliminating decoder term entirely.

Simplified Objective:

$$\mathcal{L}_{\text{yield}} = \|y - \hat{y}\|^2 + \beta \cdot \text{KL}[q_{\phi}(z|x) \| p_{\theta}(z)]$$

Generalizable Recipe:

Applies to any domain where $z \rightarrow x$ is deterministic (no decoder needed)

Innovation 2: Sinusoidal Prior for Seasonality

Problem with Standard Priors

$p(z) \sim \mathcal{N}(0, I)$ ignores temporal structure

Sinusoidal prior captures seasonality:

$$p_{\theta}(z) \sim \mathcal{N}(A \sin(\omega t + \phi), \sigma^2 I)$$

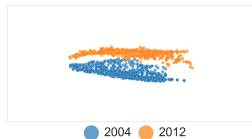
- ▶ Parameters ($A, \omega, \phi, \sigma^2$) learned during pretraining
- ▶ Explicitly models weather periodicity
- ▶ Enables richer latent representations

Key Metric

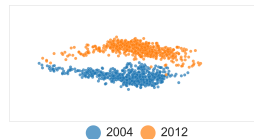
PCA variance in top-2 components:

T-BERT: 84% → VITA-Sinusoid: **15.7%**

Latent Space Comparison:



T-BERT (84% collapsed)



VITA-Sinusoid (15.7%)

Sinusoidal prior **prevents latent collapse**, enabling better separation between record-breaking yield years (2004, blue) and extreme drought years (2012, orange).

Pretraining:

- ▶ NASA POWER (1984–2022)
- ▶ 116 grids across Americas, 100K sequences
- ▶ 31 meteorological variables, weekly resolution

Fine-tuning:

- ▶ **763 Corn Belt counties**
- ▶ Corn & soybean yields (1982–2018)
- ▶ Only 6 weather variables (no soil data)

Extreme Years (by z-score deviation):

- ▶ Corn: 2002, 2004, 2009, 2012, 2014
- ▶ Soybean: 2003, 2004, 2009, 2012, 2016

Strict Evaluation

Train on 15 preceding years only. Test years **held out**.

Baselines:

Type	Method
Traditional	OLS, XGBoost [†]
Deep Learning	CNN-RNN [†] GNN-RNN [†]
Foundation	Chronos-Bolt
Pretraining	SimMTM T-BERT (ours)

[†] Uses soil data; VITA does not

Target: **hardest prediction scenario**—years deviating most from trends

Results: State-of-the-Art on Extreme Years

Method	Corn R^2	Soy R^2
OLS	0.227	0.460
XGBoost [†]	0.135	0.377
CNN-RNN [†]	0.256	0.498
GNN-RNN [†]	0.564	0.640
Chronos-Bolt	0.525	0.621
SimMTM	0.642	0.687
T-BERT	0.660	0.693
VITA	0.729	0.722

[†] Uses soil data; **VITA does not**

+29% corn, +13% soybean
vs GNN-RNN ($p < 0.0001$)

VITA *without soil data* outperforms GNN-RNN
with soil data—enabling use in data-scarce regions

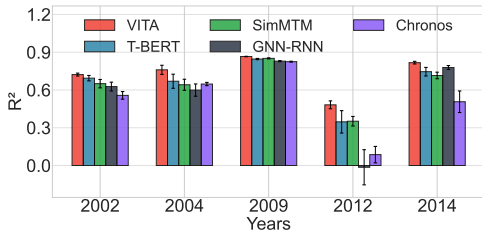
Key Findings

- ▶ **3.2×** improvement over OLS baseline
- ▶ Low variance: $\pm 0.008 R^2$ across seeds
- ▶ Chronos-Bolt ($4.5\times$ larger) fails on **data asymmetry**

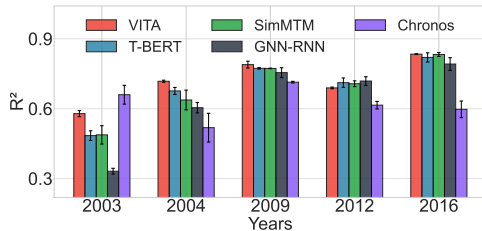
Ablation insight:

- ▶ Variational objective: **+7%** over MSE
- ▶ Sinusoidal prior: **+3%** over $\mathcal{N}(0, I)$

Corn (5 extreme years):



Soybean (5 extreme years):



Statistical Validation

8/10 extreme years

($p < 0.0001$)

2012 drought: **+39%**

Forward Gap Test

Train 1994–2009, Test 2014–2018

0.797 corn, **0.819** soy

(not memorizing)

Early Season Forecast

Weather cutoff: July

0.689 corn R^2

(operationally useful)

Experiment

Pretrain on **Central & South American weather only**
(completely exclude U.S. data)

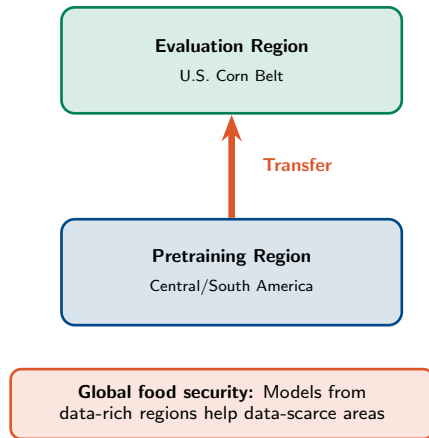
Result: Still improves U.S. predictions significantly:

- ▶ Corn: **+34%** improvement ($p < 0.01$)
- ▶ Soybean: **+17%** improvement ($p < 0.01$)

Interpretation

VITA learns **universal weather-agriculture relationships**:

- ▶ Temperature stress patterns
- ▶ Precipitation deficit impacts
- ▶ Radiation anomaly effects



\$4.75 Billion

Potential value of improved predictions

Calculation:

- ▶ **11.4 bu/ac RMSE reduction** over OLS
- ▶ At \$4.70/bushel
- ▶ Across 88.7M Corn Belt acres

Applications:

- ▶ Federal Crop Insurance pricing
- ▶ USDA operational forecasts
- ▶ Global food security planning

Practical Deployment

- ▶ Single GPU training
- ▶ <2.5 hours total
- ▶ **Only public data:**
 - ▶ NASA POWER (satellite weather)
 - ▶ USDA NASS (historical yields)

Model Efficiency

- ▶ 2M parameters (vs 9M Chronos)
- ▶ 4-layer transformer
- ▶ <2% overhead vs standard encoder

Conclusion: Domain-Aware AI Overcomes Data Limitations

Summary

1. **Data asymmetry** is a fundamental challenge in weather-based prediction
2. **VITA**: decoder-free variational pretraining + sinusoidal prior
3. **State-of-the-art** on extreme years:
 $R^2 = 0.729$ corn, 0.722 soybean
4. **Cross-continental transfer** enables global deployment

Broader Applicability

Framework generalizes to any setting with:
rich sensors at training → **sparse at inference**

- ▶ ICU monitoring vs bedside vitals
- ▶ Industrial IoT telemetry
- ▶ Environmental sensor networks

Key Takeaway

Foundation models assume feature consistency. When real deployments don't offer it, variational pretraining with domain knowledge can bridge the gap.

Code & Data:



Thank You!

Questions?