Downscaling Crop Yield Estimates Using Geospatial Foundation Embeddings

Joel Ferguson[a], Kangogo Kibarar[b], and James E. Sayre[c*]

[a]*Nelson Institute for Environmental Studies, University of Wisconsin–Madison, Madison, WI 53706, USA*
[b]*Department of Environmental Science, Policy, and Management, University of California, Berkeley, Berkeley, CA 94720, USA*
[c]*Department of Agricultural and Resource Economics, University of California, Davis, One Shields Avenue, Davis, CA 95616, USA*

Abstract

Crop production statistics are critical to food security planning, yet they are costly to collect and are typically reported at large administrative units such as states, counties, or municipalities. Existing methods that use satellite imagery to produce high-resolution yield estimates rely on complex pipelines and/or disaggregated ground truth that may be inaccessible to practitioners and researchers. Our study develops an alternative approach that uses histograms and quantiles of Google AlphaEarth Foundations (AEF) embedding distributions to downscale yields using only publicly available aggregate data. We apply our model to downscale municipality (admin-2) maize yields to near-farm scales and validate its performance using disaggregated ground truth from the Mexican Agricultural Census, attaining an overall R^2 of 0.61 and within-R^2 of 0.23. We find that our best model, trained with AEF features, explains 78% of the within-municipality yield variation that can be predicted by a model trained on the ground truth labels. Approaches of the kind used in the literature have limited within-municipality prediction performance, with within-R^2 below 0.10, less than half that of our best AEF specification. We highlight within-region R^2 as a complementary metric for evaluating whether models recover fine-scale yield variation beyond what is already observable in regional-level official statistics. To aid in this evaluation, we propose a set of diagnostics to help others anticipate downscaling performance in the absence of disaggregated ground truth data.
Keywords: remote sensing, yield prediction, geospatial foundation models, downscaling

1 Introduction

High-resolution crop area and yield maps can provide critical information for agricultural planning and policy evaluation under a changing climate (Lobell et al., 2008; Hultgren et al., 2025). However, producing such maps often depends on field- or sub-field-level ground-truth data that are

*Corresponding author. E-mail address: jsayre@ucdavis.edu

costly to collect and rarely available at scale, leaving decision-makers in many low- and middle-income countries reliant on spatially aggregated, outdated, or uncertain statistics (Carletto et al., 2013). Combining freely available satellite imagery with ground observations offers a path toward more spatially resolved agricultural statistics (Burke et al., 2021; Muruganantham et al., 2022; Joshi et al., 2023), but conventional approaches can require substantial preprocessing and harmonization across sensors, spatial resolutions, and time periods (Mai et al., 2022). For the promises of machine learning and remote sensing to become accessible to resource-constrained countries, researchers, and policy makers, an approach is required that minimizes training costs and utilizes relatively abundant publicly available ground truth. Geospatial foundation models (GFMs) have recently emerged as a promising way to reduce these data-processing burdens (Mai et al., 2023).

In this paper, we develop a method to generate fine-scale yield estimates from Google DeepMind's AlphaEarth Foundations (AEF) embeddings, using only aggregate municipality-level statistics publicly reported by the *Dirección General del Servicio de Información Agroalimentaria y Pesquera* (DGSIAP) for model training. AEF embeddings provide compact pixel-level representations derived from multiple Earth-observation and environmental data sources (Brown et al., 2025). The scale and self-supervised training of GFMs allow representations such as AEF embeddings to integrate diverse multisensor Earth-observation and climate data into a compact embedding space, creating a rich and reusable feature space for downstream agricultural prediction (Yang et al., 2025). By using AEF embeddings, our approach avoids repeated processing of large multisensor remote-sensing datasets and the need to separately incorporate climate covariates, while preserving rich spatial and environmental information for fine-scale yield downscaling.

To evaluate the output of these models, we use restricted-access data from the 2022 Mexican census of agriculture, georeferenced at the *área de control* (ADC) level, approximately two orders of magnitude finer than the municipality-level data. These highly disaggregated yield measurements allow us to directly compare our predictions to ground truth yields for regions smaller than those present in the publicly accessible aggregate data. In assessing downscaling performance, we report a number of evaluation metrics but focus on the within-R^2, a measure of the variation explained by predictions after accounting for differences across elements of some partition of observations. There are two main advantages of the within-R^2 over traditional metrics like the coefficient of determination (R^2) or squared correlation coefficient (r^2) for the particular task of downscaling when the partition is the set of units observed in the training data. The first is that there is a very good natural benchmark model that is costless to implement: the training data. We demonstrate that it is possible to attain high R^2 and r^2 in the disaggregated data by guessing the value in the municipality-level data that each ADC belongs to. In contrast, the within-R^2 mechanically evaluates as 0 for this prediction strategy. Second, the within-R^2 reports the expected share of the variation within a training data unit accounted for by predictions, which naturally corresponds to a notion of downscaling performance.

We show that compared to a number of alternative models and input data choices, a model that uses AEF embeddings-histogram and quantile features together with regularization achieves the highest within-$R^2 = 0.23$ at essentially the highest total R^2 (0.61). This represents about 78% of the downscaling performance of an "oracle" model trained directly on disaggregated ADC-level ground truth. Many models that do not utilize AEF embeddings inputs produce negative within-R^2, indicating that the entirety of their total-R^2 performance comes from predicting differences in municipality-level means, rather than truly downscaling. Importantly for practitioners, we show that downscaling performance, as measured by within-R^2, can be reliably predicted using a number of summary statistics of the variation in features across ADCs. We also show that our disaggregated predictions can be aggregated to the municipality level to produce yield estimates that correlate more strongly with the Agricultural Census measurements than the DGSIAP estimates used to train the model. This finding suggests the potential for AEF or other geospatial foundation model embeddings to improve annual sample-based yield estimates.

1.1 Related literature

Geospatial foundation models have progressively expanded the types of Earth-observation information encoded into reusable representations. Early models such as SatMAE and Scale-MAE demonstrated that self-supervised pretraining could learn transferable representations from multi-spectral imagery and across spatial scales (Cong et al., 2022; Reed et al., 2023). Multi-temporal models such as Prithvi-EO-2.0 further incorporated seasonal and interannual information from Harmonized Landsat–Sentinel-2 time series (Szwarcman et al., 2024), while TESSERA extended this idea by combining annual Sentinel-2 optical and Sentinel-1 radar time series into reusable 10-m pixel-level embeddings that transfer efficiently to downstream tasks such as crop classification (Feng et al., 2026). More recent models such as DOFA and SkySense++ have broadened representation learning across increasingly heterogeneous sensors and modalities (Xiong et al., 2024; Wu et al., 2025). AlphaEarth Foundations extends this further by learning representations informed not only by optical and radar observations but also by elevation, climate, vegetation structure, and other environmental information, and distributing these representations directly as analysis-ready annual embedding layers (Brown et al., 2025).

Geospatial foundation models have now been applied across several agricultural tasks, including crop classification, semantic segmentation, and continuous regression. For crop classification, models such as Prithvi and CropSTS have been used to distinguish crop types using multi-temporal satellite observations across growing seasons and heterogeneous agricultural regions (Szwarcman et al., 2024; Yan et al., 2025). For semantic segmentation, foundation models have been applied to crop and field-boundary mapping at very different spatial scales, from large commercial agricultural regions using 10-m Sentinel-2 imagery to smallholder systems with fields smaller than 0.1 ha. SAM, for example, has been tested for zero-shot field delineation without local fine-tuning (Ferreira et al., 2025; Tripathy et al., 2026). More recently, these models have also been applied to continuous regression tasks, with Prithvi-EO-2.0 used for field-scale canola yield prediction and AEF embeddings evaluated for crop yield, tillage, and cover-crop prediction across multiple locations and years (Nejadshamsi et al., 2026; Ma et al., 2026). These applications show that GFM representations can support agricultural prediction across different tasks, spatial scales, and time periods, but most studies still train or adapt the model using labels collected close to the scale of prediction. This leaves open how well these representations can transfer across scales when only coarse aggregate labels are available for training.

Scale transfer also poses a domain-adaptation problem: spatial aggregation alters feature variance, so models trained on aggregated features face a distribution shift at fine-scale inference (Ma et al., 2024; Paudel et al., 2023). Ma et al. (2024) address it with the Quantile loss Domain Adversarial Neural Network (QDANN) framework, which uses adversarial learning to extract scale-invariant features and a variational autoencoder (VAE) to filter out outlier training samples. Our approach instead sidesteps the mismatch at the featurization stage, with distributional summaries whose dimension does not depend on the size of the spatial unit (Section 3.2).

2 Context and data

Our study is based in Mexico, where maize is a dietary staple and the country's most widely produced crop. Maize production is essential to the livelihoods of millions of people in Mexico, particularly smallholder farmers who depend on the crop for their sustenance and income. Maize production varies across the country, with Jalisco and Sinaloa being the most productive states and Oaxaca having far lower yields. Fortunately, Mexico has agricultural yield data available at several scales: aggregated data at the municipality level (similar to US counties), as well as more disaggregated data close to, or at, the farm level. Our objective here is to train our model solely on aggregated yield data publicly available for Mexico's 32 states and 2,469 municipalities and validate at finer scales which may be unavailable in other contexts. We describe the datasets used below as well as in Table 1.

Table 1: Yield datasets used for model training and fine-scale evaluation.

Provider	Description	Spatial unit	Years	Role
DGSIAP	Statistical Yearbook of Agricultural Production, reported annually by crop cycle	Municipality (admin-2)	2003–2025	Training
INEGI	National Agricultural Census	*Área de control* (admin-4)	1991, 2007, 2022[a]	Evaluation
CIMMYT	Recorded maize-grain yields on georeferenced farmer trial plots	Plot	2017–2022	Evaluation

Notes: DGSIAP yields are public aggregate statistics (2017–2024 used for training, the years with AlphaEarth Foundations embeddings). [a]Only the 2022 Census is used here for our evaluation; such information was obtained via the INEGI microdata laboratory. CIMMYT plot records span 2012–2022, of which the 2017–2022 plot-years overlap the embeddings (23,670 matched maize-grain plot-years) and serve as external validation.

2.1 Aggregate yield information

Our primary source of agricultural productivity information is from the Directorate General of the Agrifood and Fisheries Information Service (DGSIAP, by its initials in Spanish: *Dirección General del Servicio de Información Agroalimentaria y Pesquera*). DGSIAP conducts monthly surveys of farmers to produce estimates of production (in terms of tons) and hectares planted for each crop in a municipality. As of writing, these data are publicly available at the municipality level from 2003 to 2025. We consider this information to be comparable to the aggregated agricultural survey data available in many countries.

2.2 Near-farm yield information

Our primary source of disaggregated yield data comes from restricted-access microdata from the 2022 Agricultural Census conducted by Mexico's National Institute of Statistics and Geography

(INEGI), accessed by request through a microdata laboratory project. This provides us more disaggregated information than publicly available data at the municipality level from the same census. In the microdata, crop yields can be disaggregated down to the level of the area of control (*área de control* in Spanish), which we refer to here as ADCs. ADCs represent clusters of farms with similar agricultural and ownership characteristics, such as private or communal land tenure.

The median ADC comprises 10 farm units[1], with a minimum of one farm unit, and a maximum of 700 farm units, reached in the state of Oaxaca, where smallholder agriculture and land fragmentation are prevalent. In terms of area, the median ADC covers about 129 hectares, with a minimum of 16 square meters, and a maximum of 635,229 hectares (the largest is located in Baja California). For reference, the median municipality in terms of area is 23,178 hectares, the minimum is 221 hectares, and the maximum is 5,325,617 hectares. Although there is some overlap in this distribution, 62% of ADCs are smaller than the smallest municipality in Mexico.[2] Figure 1 contrasts yields at the two scales: its panel (e) plots the DGSIAP municipal yield assigned to each ADC against that ADC's census yield, and panel (f) maps the difference between the ADC and municipality yield surfaces, the within-municipality variation our downscaling exercise targets.

2.3 Plot-level information

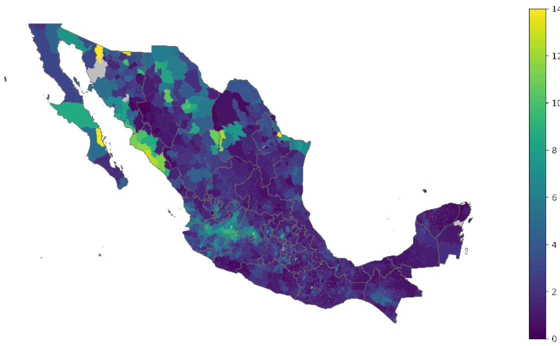
We also use an independent dataset from the International Maize and Wheat Improvement Center (CIMMYT, by its initials in Spanish) in Mexico, available at the individual-plot scale, to provide an additional source of fine-resolution validation (Gardeazábal-Monsalve et al., 2025). These observations are used exclusively for external validation and never enter model training, allowing us to test whether a model trained only on municipality-level aggregates can recover productivity variation at the scale of individual fields. The CIMMYT data are available for 2012–2022 and include planting and harvest dates, plot area, latitude, longitude, and maize grain yield. From 2020 on, they also report plot boundaries, allowing for extraction of satellite features over each field's actual footprint. In total the network comprises 18,582 georeferenced plots distributed across 29 states and 926 municipalities. As such, the CIMMYT sample spans much of the country's maize-growing geography.

The fields in the CIMMYT dataset are small: the median plot is 1.6 hectares (mean 3.1), with the middle 80% between 0.6 and 5.4 hectares, roughly two orders of magnitude smaller than the median census ADC. Intersecting the plots with the years for which AEF embeddings are available (2017 onward) within cropland areas (see Section 3.1 for a definition of cropland area) gives us 23,670 maize-grain plot-years with recorded yields used for validation. The CIMMYT plots are often not representative of average local production (see Table 7, Panel B), and we do not expect a model trained on area-average yields to reproduce their absolute levels. Our external-validation exercise instead asks whether the model correctly ranks fields by productivity and resolves within-area yield differences at true farm scale (Section 4.9).

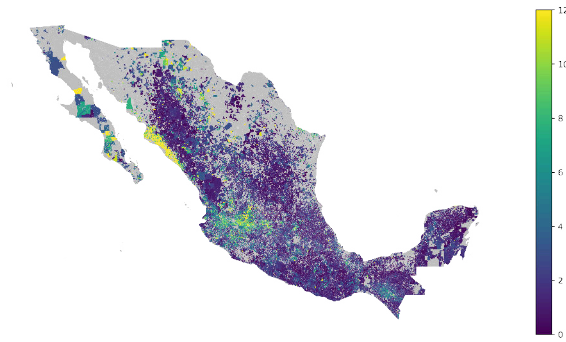
[1] A farm unit is INEGI's technical definition for a single set of one or more plots held by the same owner within a given municipality.

[2] Only 2 ADCs are smaller than 30 square meters – the pixel size of the coarser Landsat imagery.

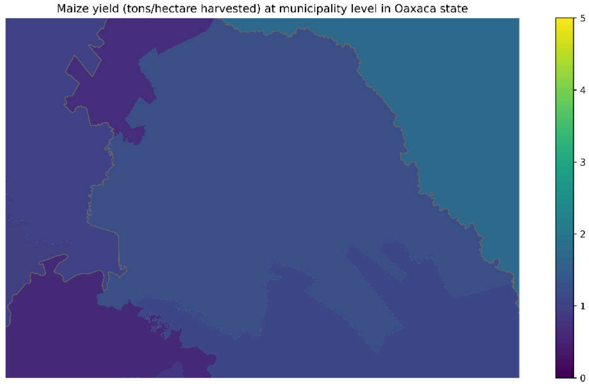
Figure 1: Comparison between maize yield in municipalities and *áreas de control* (ADCs)



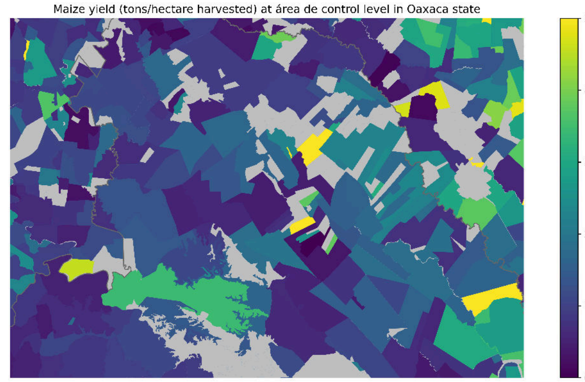
(a) Municipalities in Mexico



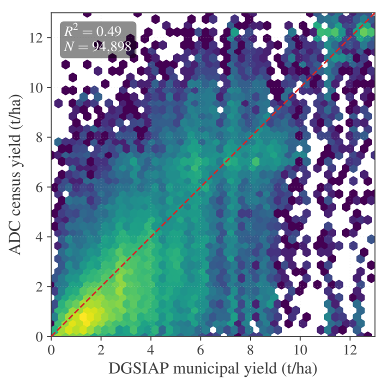
(b) ADCs in Mexico



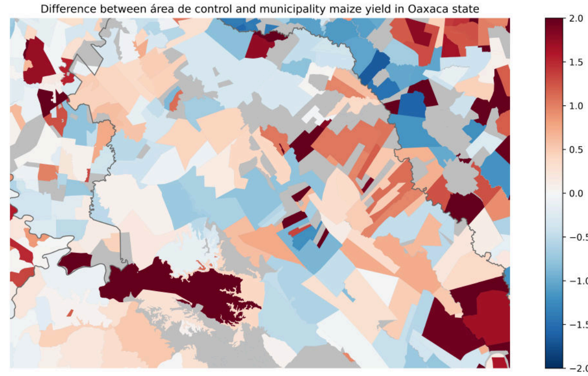
(c) Municipalities in Oaxaca state



(d) ADCs in Oaxaca state



(e) DGSIAP municipal yield vs. ADC census yield



(f) ADC yield minus municipality average, Oaxaca state

Source: INEGI Agricultural Census (panels (a)–(d) and (f)) and DGSIAP (panel (e)). Average maize yield plotted at the municipality and área de control (ADC) level. Panels (c) and (d) show a selection taken from Oaxaca state. Grey areas correspond to zero or missing maize production. Panel (e) plots the DGSIAP municipal maize yield (2022) assigned to each ADC against the ADC-level census maize yield (hexbin density on a log color scale; dashed line is the 1:1 line). Panel (f) maps the difference between the yield surfaces of panels (d) and (c) – each ADC's census yield minus its municipality's census average (tons/hectare) – the within-municipality variation that municipal statistics cannot resolve.

2.4 AlphaEarth Foundations embeddings

Our primary methodology leverages embeddings from the AlphaEarth Foundations (AEF) model, a geospatial foundation model developed by Google DeepMind (Brown et al., 2025). AEF has approximately 480 million parameters and generates near-global geospatial representations by assimilating spatial, temporal, climate, topographic, and measurement contexts across multiple sources, including optical imagery (Landsat 8/9, Sentinel-2), synthetic aperture radar (Sentinel-1), spaceborne LiDAR (GEDI), climate reanalysis (ERA5-Land), and elevation data (GLO-30). The model is then trained across a range of downstream geospatial tasks, such as land-use and land-cover estimation, allowing the learned representations to generalize across applications and environmental outcomes. The resulting annual embeddings are publicly available on Google Earth Engine (GEE) with tiered access, stored as a 64-band raster image for each year (bands labeled A00 through A63). AEF embeddings offer several practical advantages for our setting: (i) they inherently harmonize data from multiple sensor modalities, eliminating the need for manual data fusion, (ii) they provide consistent, gap-free spatial coverage not subject to cloud contamination, and (iii) they require no pre-processing before use in downstream modeling.

3 Methodology

Our methodology addresses the common problem of making fine-scale predictions when the outcome of interest is only observed at a coarser, aggregated spatial scale. The prediction model must therefore be capable of performing inference at a geographically finer scale than the data used for training, i.e. scale transfer. A key challenge in scale transfer is that spatial aggregation alters feature variance, producing different distributions at coarse and fine scales; consequently, models trained on aggregated features face a severe distribution shift when applied during fine-scale inference (Ma et al., 2024; Paudel et al., 2023).

Compounding this challenge, differences in land area and feature variance between the aggregated and disaggregated scales can obscure localized environmental signals. Spatial aggregation over large administrative units smooths within-unit heterogeneity, reducing the fine-scale variation available to inform disaggregated predictions (Ma et al., 2026). Therefore, we use aggregation techniques that produce consistently sized feature representations across municipalities, ADCs, and field plots, regardless of differences in their spatial extent or geometry. These representations summarize the distribution of features within each spatial unit using statistics such as the arithmetic mean or the 25th percentile. Figure 2 shows our core approach of feature engineering distributional aggregation methods invariant to the scale of our areas of interest that can be easily used for the purposes of estimation and inference.

3.1 Cropland masking

One challenge that arises when using satellite imagery is that the user does not know where agricultural production, namely maize production, occurs within the area of interest. However, studies such as You et al. (2017) suggest that precise discrimination of these areas is not always necessary to receive some signal about crop production practices. The basic principle is the need to develop an aggregation that preserves enough information about the features of pixels employed in cropland production to predict yields.

In order to loosely constrain the set of inputs, we use the ESA WorldCover v200 (2021) cropland mask to attempt to restrict to pixels within cropland production areas. It is worth noting that this cropland measure does not identify maize fields specifically, although of course maize is the most commonly grown crop in Mexico. That our approach works with so coarse a mask indicates that it does not hinge on high-quality crop maps; refined crop-type products would, if anything, improve it.

DGSIAP similarly publishes a cropland mask for Mexico, the *Frontera Agrícola Serie II*, a 1:10,000 map of agricultural land that classifies each agricultural polygon as irrigated or rainfed. However, we largely refrain from its use in all but a few ways, restricted to the aggregation of our downscaled predictions back to more aggregated scales (Sections 3.5.1 and 4.7) and the development of the model weights and skill indices described later (Sections 3.5.2 and 4.4). This is to demonstrate that most users can use globally available gridded products rather than high-quality, context-specific information.

3.2 Input aggregation

One concern is that our areas of interest vary widely. Municipalities in our sample differ in area by more than two orders of magnitude (the two examples in Figure 2 differ by a factor of 160), and the ADCs at which we ultimately perform inference are smaller still, down to a few dozen pixels at 10 meter resolution. Therefore, we must standardize input sizes for each spatial unit.

To do so, for each feature of the input (i.e. dimensions or bands of the image) we count the number of cropland pixels each unit contains across a predetermined number of bins separated by value. Then, histograms record the proportion of pixels in each bin, allowing a unit with a few dozen pixels and a unit with millions of pixels to be compressed into the same fixed-length vector. A single model can therefore be trained at the scale where labels exist (municipalities) and applied without modification at any other scale (ADCs, or the individual farmer plots in our external validation).

Concatenating the per-feature histograms across all features produces a fixed-size input that we feed into a gradient boosting regression model (Friedman, 2001; Ke et al., 2017). This approach can also be expanded to include other summary statistics that are scale invariant, such as per-dimension distributional summaries like quantiles that can be combined with the histograms as inputs.

This approach is generalizable to different sources of inputs. To compare how our aggregation procedure performs across a number of inputs, our primary model bins the 64 dimensions of the AlphaEarth Foundations embeddings, and as a comparison we also bin harmonic regression gap-filled NDVI time series extracted from Landsat imagery evaluated at planting-anchored growth stages (see Section 3.4.2 for more information). Save for the inputs, the aggregation procedure is otherwise identical.

We next describe in more detail the development of each specific model we evaluate.

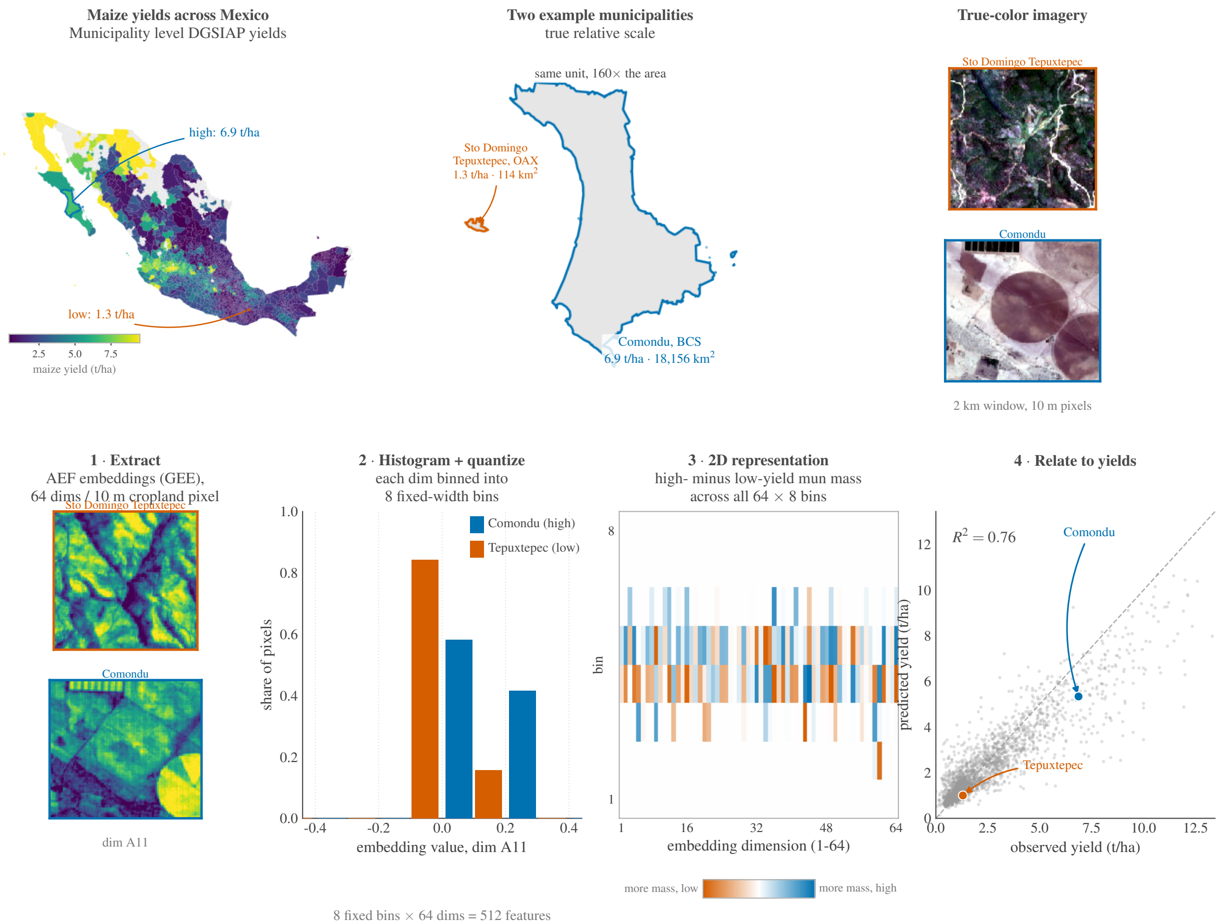


Figure 2: Overview of the AEF distributional methodology. *Top row:* municipal maize yields across Mexico (DGSIAP, 2022), with two example municipalities highlighted — low-yield Santo Domingo Tepuxtepec, Oaxaca (1.3 t/ha, rainfed) and high-yield Comondú, Baja California Sur (6.9 t/ha, irrigated). The right panel shows true-color imagery over a 2 km cropland window in each municipality. *Bottom row:* (1) We extract 64-dimensional AlphaEarth Foundations embeddings at 10 m resolution over cropland pixels; the panel shows embedding dimension A11 over the same 2 km windows as the imagery above. (2) Each embedding dimension is summarized as a histogram over eight fixed bins on $[-0.8, 0.8]$, quantizing the distribution of pixel values within a spatial unit; the two example municipalities occupy visibly different regions of the embedding space. (3) Stacking all 64 dimensions yields a 64×8 histogram representation per unit-year (panel shows the high- minus low-yield difference in bin mass across all dimensions, in their natural order). (4) A gradient-boosting model maps these histogram features to yields, trained on DGSIAP municipal yields and validated at the near-farm (ADC) level against the Agricultural Census; the panel shows leave-one-year-out cross-validated municipal predictions for 2022 (a municipality-level fit, not comparable with the held-out-municipality validation of Table 2); the two example municipalities are marked.

3.3 Model development

3.3.1 AEF distributional scale-transfer model

For each municipality m and year y, we retrieve the 64-dimensional AEF embedding at every 10 m pixel p, $\mathbf{x}_{p,my}^{\mathrm{AEF}} \in \mathbb{R}^{64}$, within the unit that is classified as cropland by the ESA WorldCover v200 (2021) land-cover map. Within each spatial unit, we count the cropland pixel values into eight histogram bins for each of the 64 AEF embedding dimensions, as shown in Figure 2, yielding a $64 \times 8 = 512$-dimensional feature vector for each unit-year. We consider two sets of bin edges: fixed-width with bounds $[-0.8, 0.8]$, and equal-mass (quantile) bin bounds[3]. We refer to this model as *AEF Hist*. Because the model is trained on municipality-level histograms but applied to far smaller units, it is trained on municipal histograms subsampled to resemble ADC-level histograms, as described in Section 3.3.2.

3.3.2 AEF histogram ensemble

The ensemble complements the histogram representation with a second model whose features are additional moments of the embedding distributions, capturing the within-region heterogeneity of the input values. The distributional features for each of the dimensions include five percentiles (10%, 25%, 50%, 75%, 90%) and the standard deviation across all agricultural pixels within a spatial unit, yielding $64 \times 6 = 384$ distributional features per observation. These are combined with the 64 mean features to produce a total of 448 input features for this percentile model. Together these distributional features capture information lost by averaging: for example, a municipality with uniformly productive cropland will have tightly concentrated embedding distributions (low standard deviation, similar percentiles), while a municipality with heterogeneous land use will exhibit dispersed embedding distributions.

In addition, we aim to address a dilution bottleneck that arises in the scale-transfer setting, where downscaled predictions exhibit less variance than true disaggregated values (Tripathy et al., 2026). When models are trained on municipality-level distributional features (percentiles and standard deviations computed from many pixels), the resulting feature distributions are smooth because there are many pixels relative to the number of bins. At prediction time, however, ADC-level features are computed from far fewer pixels, producing much sparser, spikier distributions. To bridge this gap, the AEF Hist model is a HistGradientBoosting model trained on binned histogram features (64 dimensions $\times$ 8 bins = 512 features per unit) that have been subsampled from the municipality distributions to mimic the higher variance in distributions of ADC-level data: for each municipality-year, we draw $K = 5$ synthetic observations by sampling $N = 2$ pixels per AEF embedding dimension from the municipality's binned distribution, creating training data whose distributional characteristics match those of real ADCs. We then ensemble the ADC-level predictions from this subsampled-bins model (AEF Hist) with a second HistGradientBoosting model trained on the 448 percentile, dispersion, and mean features described above (the percentile model) using a weighted average ($w = 0.5$ on AEF Hist, $1 - w = 0.5$ on the percentile model). We refer to this model as the *AEF Hist Ens.* model.

The weight is chosen from public data alone, by the irrigation-projection argument we use for the shrinkage factor in Section 3.5.2: under that argument the within-municipality validity of the

[3] The quantile bin bounds are based on an empirical distribution of AEF values based on a representative sample.

blended prediction is proportional to its within-municipality correlation with the DGSIAP irrigated share, which is maximized at $w = 0.5$.[4] The choice of w is not terribly sensitive: the projected within-municipality R^2 is flat for w between 0.35 and 0.70. The two models contribute complementary signals: the subsampled-bins model (AEF Hist), trained on ADC-like distributions, provides strong between-municipality predictions, while the percentile model—whose distributional features vary naturally across ADCs within a municipality—provides within-municipality differentiation. Training a single model on all 960 combined features performs worse than this ensemble, as the two separately-trained models capture complementary patterns.

The results are robust to the two tuning choices in the bins model. The first is the coarseness of the subsampling: the $N = 2$ multinomial draw reflects the extreme concentration of real ADC histograms—in the census evaluation year, 77% of ADC bin cells are empty and the largest bin holds 89% of an embedding dimension's mass on average—and an alternative that draws pixel counts from the empirical ADC distribution performs essentially identically (overall R^2 of 0.59 under either, within-R^2 of 0.17 vs. 0.16, combined season). The second is the bin edges: because AEF embeddings are near unit-norm, a typical component is roughly $1/\sqrt{64} \approx 0.125$ in magnitude, so most cropland pixel mass falls in the central bins of the fixed $[-0.8, 0.8]$ grid, while the quantile-bin variant spreads it evenly across all eight levels. With the deployed shrinkage applied, the two versions are essentially equivalent on the within-municipality margin (within-R^2 of 0.22 fixed vs. 0.23 quantile, combined season) and the fixed-width version retains the better overall fit (0.61 vs. 0.58).

3.4 Reference models

3.4.1 AEF mean model

As a benchmark comparison methodology utilizing the AEF embeddings as features, we use per-band means as tested by Ma et al. (2026), $\mathbf{x}_{my}^{\mathrm{AEF}} \in \mathbb{R}^{64}$. As in their paper, we input these features into a random forest model, one of the two learners they evaluate, with the 64 per-band means as the only features.[5] We refer to this model as *AEF mean*. This serves as a baseline comparison for the models we develop.

3.4.2 Remote sensing baseline features

As a useful point of comparison between our AEF-based methods and more established remote sensing (RS) methods, we deploy the same architecture on conventional sources of satellite data

[4] Write the blended within-municipality deviation as $\tilde{p}(w) = w\,\tilde{p}_{\mathrm{bins}} + (1-w)\,\tilde{p}_{\mathrm{pct}}$. Under the projection of Section 3.5.2, $\tilde{y} = \beta\,\widetilde{\mathrm{irr}} + \tilde{v}$, the blend's within-municipality correlation with yields decomposes as $\rho(w) = \rho_{\mathrm{irr}}(w)\,\beta\sigma_{\mathrm{irr}}/\sigma_y + \rho_v(w)\,\sigma_v/\sigma_y$, where only $\rho_{\mathrm{irr}}(w) = \mathrm{corr}(\tilde{p}(w), \widetilde{\mathrm{irr}})$ is observable from public data. Both the lower bound ($\rho_v = 0$) and the midpoint estimate used for λ (which sets ρ_v proportional to ρ_{irr}) are therefore $\rho_{\mathrm{irr}}(w)$ multiplied by a factor that does not depend on w, so the public-data argmax of $\rho(w)$—and of the shrunk within-municipality R^2, $\rho(w)^2$—is simply the w that maximizes the within-municipality correlation between the blended prediction and the DGSIAP irrigated share. Computed on the 2022 ADC predictions of the two component models (56,509 ADCs with an irrigated share), this correlation peaks at $w = 0.5$. Neither the census nor the CIMMYT data enter this choice.

[5] The learner matters little for this baseline: plain OLS on the same 64 means delivers nearly identical downscaling accuracy (spring-summer ADC-level R^2 of 0.47 for either learner, and 0.51 vs. 0.50 with the deployed shrinkage), so the AEF mean results reflect the information content of the mean-pooled embeddings rather than the choice of learner.

(You et al., 2017; Deines et al., 2021; Ma et al., 2024). We use Landsat Collection 2 Tier 1 Level-2 surface reflectance data from Google Earth Engine, computing the Normalized Difference Vegetation Index (NDVI), defined as $\text{NDVI} = \frac{\text{NIR}-\text{Red}}{\text{NIR}+\text{Red}}$, the canonical measure of vegetation greenness and crop vigor. Individual Landsat scenes are frequently covered by clouds, and during the summer rainy season parts of Mexico may have no usable observation at all in the weeks surrounding peak maize growth. Rather than discard contaminated scenes or composite over fixed calendar windows, we fit a smoothed curve to each pixel's growing cycle. For each pixel p and year, we pool all clear-sky NDVI observations from Landsat 7 and Landsat 8, masking cloud and cloud-shadow pixels with the Collection 2 QA_PIXEL (CFMask) flags, and estimate a third-order harmonic regression,

$$\text{NDVI}_p(t) = \beta_{0,p} + \sum_{k=1}^{3} \left[\beta^c_{k,p} \cos(2\pi kt) + \beta^s_{k,p} \sin(2\pi kt) \right] + \varepsilon_p(t), \tag{1}$$

where t is the date of the observation expressed as a fraction of the year. The seven coefficients are estimated by ordinary least squares separately for each pixel-year, so that year-specific growing conditions are retained rather than averaged away. The fitted curve is a smooth approximation of the pixel's annual phenology, tracing green-up, peak greenness, and senescence over the growing cycle while filtering out noise from residual cloud and atmospheric contamination. Pixels are retained wherever the year's clear observations suffice to identify the seven coefficients; pixels with fewer observations remain masked and drop out of the unit's distributional summary[6].

An important consideration is the timing of maize cultivation in Mexico. We seek to assemble imagery that captures maize at peak growth right before harvest, as well as imagery showing those same fields while fallowed. Understanding the timing of the growing season is crucial for using vegetation indices to discriminate between crops, as crops have different spectral profiles according to their phenology, maturity, and planting date. Within the appropriate time span, vegetation indices can be used to differentiate between crops, which is crucial as we do not know which specific parts of our images maize is planted in (Yang et al., 2023; Gao and Zhang, 2021). Therefore, we draw upon information on the municipality-specific monthly harvesting and planting of maize from crop progress (in Spanish, *avance de siembras y cosechas*) data from DGSIAP, which we display in Figure A.1. Summing up the hectares planted and harvested across all municipalities during the period of 2018 to 2024, on average we find that planting of maize in Mexico starts in May, peaks in July, and trails off by September. In contrast, maize harvesting starts in October, peaks in December, and trails off by February[7].

We use this local planting date information from DGSIAP together with our harmonically smoothed NDVI series to obtain estimates of crop vigor at key dates during the growing season (Dado et al., 2020). In particular, we reconstruct NDVI at three growth stages: one, three, and five months after the start of the local planting month, roughly corresponding to early vegetative growth, peak greenness, and maturity. Harmonic smoothing of the NDVI series ensures that we can calculate these values even when cloud cover prohibits observation in the relevant time period (Gao et al., 2017).

Our baseline remote sensing features are then built by exactly the same distributional methodology as the AEF models, so we compare the prediction value of the embeddings rather than

[6] Landsat 9, operational from early 2022, is excluded from the pool to ensure a consistent sample of sensors when training the model across years.

[7] This planting and harvesting cycle is referred to as the "spring-summer" crop cycle, although a smaller maize crop is also planted during the "fall-winter" crop cycle.

the summarization method. Where the AEF features summarize the distribution of each of the 64 embedding dimensions across a unit's cropland pixels, here we summarize ten NDVI features consisting of the predicted values at dates corresponding to 3 growth stages and 7 harmonic coefficients generated from the daily time series. We calculate the same set of statistics as for the *AEF Hist* model for each NDVI feature, yielding 150 inputs per unit-year.

This approach is closest to You et al. (2017) who use histograms of NDVI measured by MODIS imagery to predict corn and soy yield in the United States corn belt. Our methodology here differs by substituting harmonically gap-filled Landsat imagery for MODIS and using fewer time steps for which histograms are constructed. As with the AEF features, the histogram bins record pixel shares rather than counts, so the representation will be similar for a small ADC as for the largest municipality, allowing the approach to transfer across scales without modification. The pixels entering these summaries are restricted to the ESA WorldCover v200 cropland class, allowing us to examine model performance attributable to the input representation itself, that is, masked NDVI versus AEF.

3.4.3 Aggregation-constrained neural network

As a further comparison, we train an *aggregation-constrained* neural network (Agg-NN) following population-mapping approaches that supervise fine-scale predictions only through their spatial aggregates (Metzger et al., 2022, 2024). The network predicts a bounded ADC-level adjustment to a random-forest baseline; ADC predictions are aggregated to the municipality level with area weights and the loss is computed against DGSIAP municipal yields. In practice, this attempts to ensure the ADC-level predictions aggregate to the municipality-level average yield. Appendix A.3 details the architecture and training protocol.

3.5 Ex-post correction procedure

3.5.1 Aggregate-consistency correction

There are many important features relevant to maize yields that may not be observable via satellites, such as production and harvesting techniques, the maize varieties planted, local climatic conditions or management practices in the region. These features may be inherent to given regions and generate differences in observed maize yields, despite showing little differences on satellite imagery. Furthermore, if these techniques or knowledge are autocorrelated over time, there may be important spatiotemporal differences in maize yield that we may not be able to detect, and thus predict.

Therefore, we apply a simple correction procedure to account for these differences in average yield that are not observable to satellites. In keeping with our general approach, such a correction procedure must only rely on the aggregated data available to other researchers. Our procedure aggregates the downscaled predictions back to the regional level and then adjusts them so that the aggregate matches the outcome measured at the regional level in the training data. This procedure relies on the accuracy of the aggregated data, which may itself contain measurement error.

To be precise, consider that in municipality m there are a number of sub-regions $i \in \{1, 2, \ldots, N_m\}$. For each sub-region i, we produce the yield predictions $\widehat{z_1}, \widehat{z_2}, \ldots, \widehat{z_{N_m}}$. Since yield is measured as the tons of maize harvested per hectare, in order to aggregate these predictions up to the municipality-level image we need to take a stand on how many hectares of maize are planted in each sub-region. We opt to use readily-observable proxies for the area of maize planted in the aggregation process. In particular, in place of the true amount of maize planted in each sub-region, $a_1, a_2, \ldots, a_{N_m}$, we use either the total area of the sub-region or the amount of agricultural land found in the sub-region to produce proxies for the area planted in each region $\widetilde{a_1}, \widetilde{a_2}, \ldots, \widetilde{a_{N_m}}$. We compute the predicted average maize yield in the municipality based on our sub-region predictions with these proxies using the formula

$$\widehat{z_m}^{\text{Aggregated ADC prediction}} = \frac{\sum_{i=1}^{N_m} \widehat{z_i}\widetilde{a_i}}{\sum_{i=1}^{N_m} \widetilde{a_i}} = \frac{\sum_{i=1}^{N_m} \widehat{q_i}}{\sum_{i=1}^{N_m} \widetilde{a_i}}, \tag{2}$$

where $\widehat{q_1}, \widehat{q_2}, \ldots, \widehat{q_{N_m}}$ are the predictions for the quantity harvested of maize in each sub-region.

We then compare these aggregated ADC predictions to the DGSIAP municipality-level yield observations, and compute the difference $e_m = \widehat{z_m}^{\text{Aggregated ADC prediction}} - z_m^{\text{DGSIAP}}$. The ex-post corrected predictions are then simply the set $\{\widehat{z_1} - e_m, \widehat{z_2} - e_m, \ldots, \widehat{z_{N_m}} - e_m\}$ for the sub-regions in municipality m, enforcing that aggregated predictions are equal to the observed municipality-level yield.

3.5.2 Within-municipality shrinkage

We examine the dispersion of predictions within municipalities to evaluate whether the model captures not only which places have relatively higher or lower yields, but also the magnitude of those differences. The ex-post correction adjusts the level of each municipality's predictions but leaves their within-municipality spread untouched. We develop a final post-processing step that uses a simple form of regularization to adjust the dispersion of our predictions:

$$\widehat{z_r}^{\text{shrink}} = \bar{\hat{z}}_m + \lambda\left(\widehat{z_r} - \bar{\hat{z}}_m\right), \qquad \lambda \in [0, 1], \tag{3}$$

where $\bar{\hat{z}}_m$ denotes the mean yield prediction among the sub-regions of municipality m. The motivation is a variance-compression argument. The within-municipality R^2 can be written as $2\rho r - r^2$, where ρ is the within-municipality correlation between predicted and observed yield deviations and $r = \text{SD}(\hat{y}_{\text{dev}})/\text{SD}(y_{\text{dev}})$ is the ratio of predicted to observed within-municipality dispersion. When a model's predicted deviations are over-dispersed relative to their accuracy ($r > \rho$), shrinking them raises the within-municipality R^2, and the optimal factor $\lambda^\star = \rho/r$ attains the maximum value ρ^2. Shrinkage requires no retraining and no additional satellite inputs, and it leaves the between-municipality component of accuracy unchanged, since the municipality-mean predictions are untouched.

We aim to select the optimal factor $\lambda^\star$ based on public data alone. To estimate the dispersion ratio r, we compare the model's predicted within-municipality dispersion across ADCs (1.11 t/ha) with the within-state dispersion of DGSIAP municipal yields (1.66 t/ha),[8] which yields $\hat{r} \approx 0.67$.

Estimating ρ, the within-municipality correlation of the outcome variable, without knowledge of the downscaled data is more difficult. To do so, we look for a variable that can be measured at the downscaled level that likely correlates with yields. One such measure is information on the irrigated share of agricultural land. As a separate output from yield statistics, DGSIAP provides geolocated information on irrigation adoption. Correlating this measure against DGSIAP

[8] In this setting, we find this dispersion closely tracks the census within-municipality dispersion of 1.53 t/ha.

municipality-level yields suggests that going from no irrigation to full irrigation is associated with a 4.82 t/ha yield increase for maize, suggesting that this public measure correlates strongly with productivity. Projecting yields on irrigation, and assuming only that the predictions do not correlate negatively with the non-irrigation component of yields, gives $\rho \geq 0.36$, or $\lambda^\star \geq 0.54$.

Turning this lower bound into a point estimate requires a stand on how good the model is on the within-municipality yield variation that public data do not measure, relative to the irrigation channel they do. We can assume the true ρ lies between two polar cases: no validity on the unmeasured channel ($\rho = 0.36$), and validity equal to the measured one ($\rho = 0.61$).[9] Without taking a definitive stand, the midpoint of the two cases yields $\hat{\rho} = 0.48$ and $\hat{\lambda} = 0.72$. Fortunately, our results are not terribly sensitive to the precise choice of ρ: ρ between the lower and upper bounds maps to $\lambda^\star = \rho/\hat{r}$ between 0.54 and 0.91. The lower half of this interval is exactly the flat region of the within-municipality R^2 (any $\lambda \in [0.55, 0.80]$ comes within 0.01 of the maximum). With knowledge of the downscaled data, a leave-municipalities-out cross-validation on the census evaluation data selects $\lambda = 0.68$, close to our estimate, and like it, inside the flat region. For our deployed specification we set the shrinkage factor to the point estimate derived from our procedure, which yields $\lambda = 0.72$.

3.6 Model evaluation

For our yield prediction regression tasks, we report the coefficient of determination (R^2) and root mean squared error (RMSE) as evaluation metrics; Appendix A.1 gives the standard definitions. In addition to these two metrics, we emphasize the importance of a less-standard measure of predictive performance: the within-R^2. Given aggregate units (e.g. municipalities) indexed by m, the within-R^2 is calculated as

$$\text{within-}R^2 = 1 - \frac{\sum\limits_{m}\sum\limits_{i \in m} \left[(\hat{Y}_i - \bar{\hat{Y}}_m) - (Y_i - \bar{Y}_m)\right]^2}{\sum\limits_{m}\sum\limits_{i \in m} (Y_i - \bar{Y}_m)^2} \tag{4}$$

where i indexes disaggregated units (e.g. ADCs), $\hat{\cdot}$ denotes predicted values, and $\bar{\cdot}$ denotes arithmetic averages. The within-R^2 is similar to the ordinary coefficient of determination, except that all terms are differenced from their within-aggregated unit average. In this way, the within-R^2 is a measure of the variation within aggregated units that is explained by the predictions. To see this,

[9] Within a municipality, we project yield deviations on the irrigated share: $\tilde{y} = \beta\,\widetilde{\text{irr}} + \tilde{v}$ with $\text{cov}(\widetilde{\text{irr}}, \tilde{v}) = 0$, where $\beta = 4.82$ t/ha is the DGSIAP municipal yield gradient on irrigated share (within state-year). The predictions' within-municipality correlation with yields then decomposes as $\rho = \rho_{\text{irr}} \frac{\beta\sigma_{\text{irr}}}{\sigma_y} + \rho_v \frac{\sigma_v}{\sigma_y}$, where $\rho_{\text{irr}} = \text{corr}(\tilde{p}, \widetilde{\text{irr}}) = 0.44$ is observable from public data, $\beta\sigma_{\text{irr}}/\sigma_y = 0.82$ is the share of within-municipality yield dispersion carried by irrigation, $\sigma_v/\sigma_y = \sqrt{1 - 0.82^2} = 0.57$, and $\rho_v = \text{corr}(\tilde{p}, \tilde{v})$ is the predictions' validity on the non-irrigation component. Two cases can help bound ρ_v. Setting $\rho_v = 0$ gives the lower end, $\rho = 0.44 \times 0.82 = 0.36$: falling below it would require $\rho_v < 0$, i.e. a model that systematically anti-predicts the non-irrigation component of yields. Setting $\rho_v = \rho_{\text{irr}}$, or $\rho = 0.36 + 0.44 \times 0.57 = 0.61$, provides an upper bound. This is equivalent to assuming that the predictions' validity on the unmeasured channel of yields equals their validity on the measured (irrigation) channel. In all likelihood, ρ_v is closer to this upper bound, as irrigation is the most satellite-visible yield determinant, and the predictions correlate with the irrigated share more strongly than a uniform yield signal would imply, so these features load on visible channels over more subtle management-driven ones. Attenuation across scales would independently cap ρ at its municipal-scale value of roughly 0.65.

consider a simple baseline in which every disaggregated unit within a municipality is assigned the municipality's observed mean yield. The first term in parentheses in the numerator of equation (4) evaluates to 0, leading to a within-R^2 of 0, because every disaggregated unit prediction is equal to the aggregated unit mean prediction, $\hat{Y}_i = \bar{\hat{Y}}_m = \bar{Y}_m$. Intuitively, the within-R^2 evaluates downscaling performance by asking whether the model can explain variation within municipalities that aggregated data alone cannot predict.

We evaluate predictions at two spatial scales. For the municipality-level validation (Table 2), we randomly hold out 20% of municipalities (all of their 2017–2024 municipality-years) as the test set. All models are trained with those municipalities excluded and scored on the same held-out municipality-years over 2017–2024. The ADC-level results are obtained by applying the municipality-trained models to ADC features and validating against the 2022 INEGI Agricultural Census. For this exercise the deployed models are trained on the full municipality sample while the census provides an independent test set. Where a model involves stochastic neural-network training, we average results over five random seeds. Figure 3 summarizes the full pipeline, from the input data through model training, scale transfer, the two ex-post corrections, and the accuracy metrics.

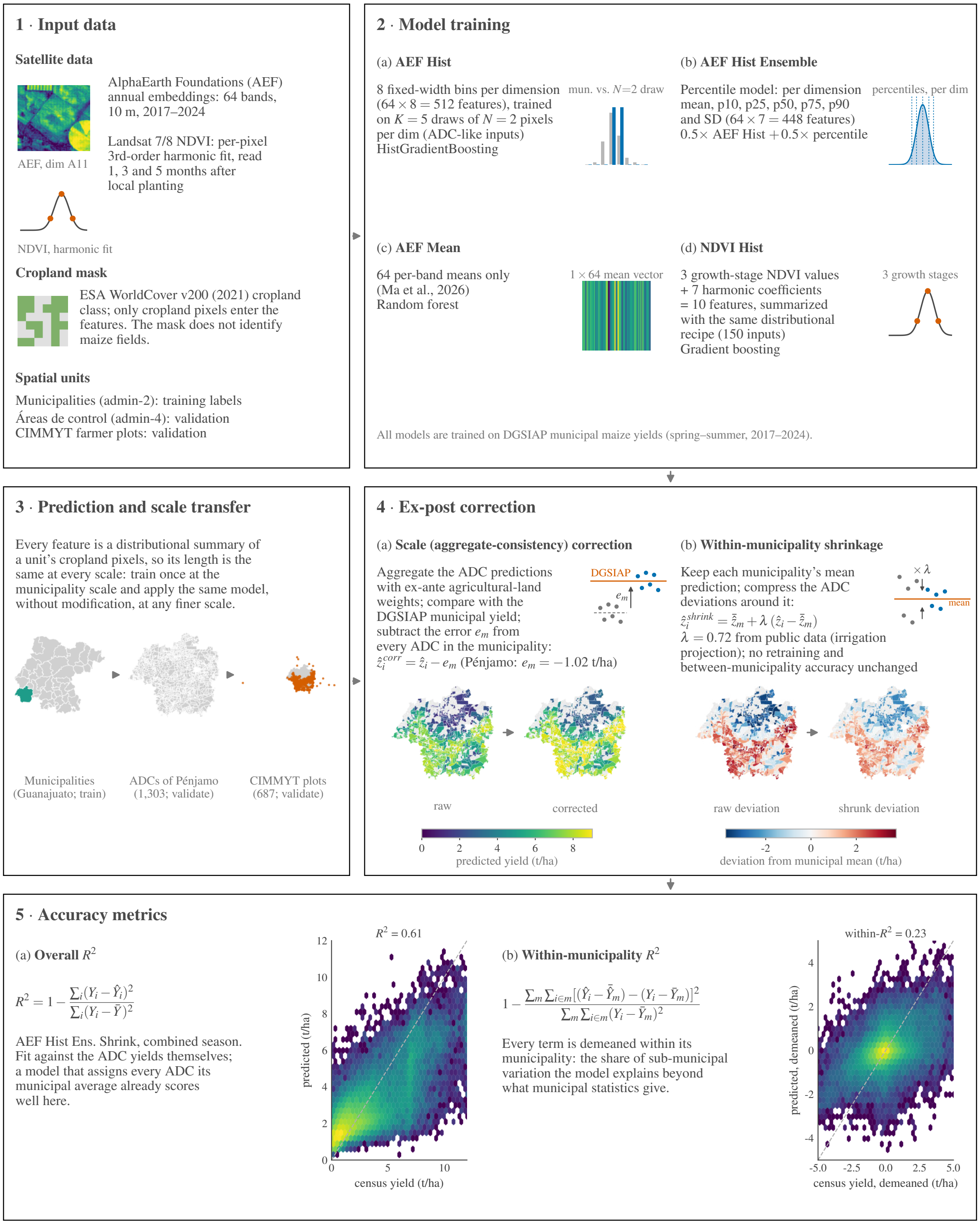


Figure 3: Overview of the prediction pipeline.

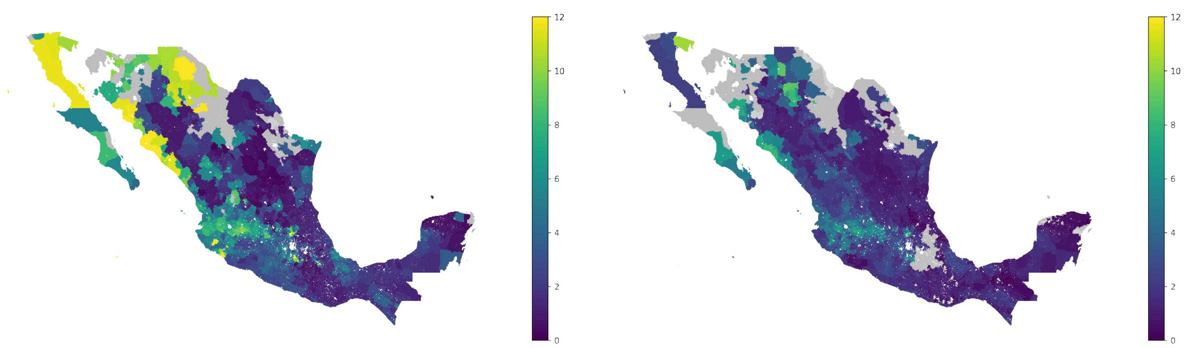
4 Results

We begin by reporting our models' performance at the municipality level, which is the aggregated scale at which they are trained. At this level, all models achieve strong validation statistics. We then take the models' predictions to the disaggregated farm-level ground truth from the Agricultural Census and find that the most performant municipality-level model is the worst in terms of within-municipality performance. Practitioners selecting among models based on aggregate performance could therefore choose the model with the worst downscaling performance. This motivates us to provide an ex-ante downscaling skill index that recovers farm-level reliability for each model from satellite and survey inputs alone, with no farm-level ground truth.

4.1 Municipality-level yield prediction

We first evaluate the performance of AEF and RS models at the municipality level, the same level at which the algorithms were trained and the level at which a standard supervised learning task would be evaluated.

Figure 4: Comparison between DGSIAP municipal maize yields and municipality-level predictions, 2022 (t/ha). Predictions are the AEF Hist Ensemble Shrink ADC predictions aggregated with ex-ante agricultural-land weights.



(a) DGSIAP municipality-level yields

(b) Prediction of municipality yields

Table 2 reports municipality-level prediction accuracy for each model under a single standardized protocol: every model is trained with the same randomly held-out 20% of municipalities excluded and all models are scored on the same 3,571 held-out municipality-years (2017–2024) against DGSIAP spring-summer maize yields. The AEF models' ADC-level predictions are aggregated to the municipality level using agricultural-land-area weights; the NDVI model is trained and scored directly at the municipality level. The Agg-NN, which is trained with an aggregation constraint that computes loss at the municipality level, achieves an R^2 of 0.79, the best of any specification, ahead of the AEF Hist Ensemble (0.63), the AEF Hist model (0.61), and the AEF mean baseline (0.52). The AEF distributional models cluster above the Landsat-derived NDVI baseline (0.53), while the means-only AEF baseline performs on par with it. Figure 4 displays the aggregated predictions against DGSIAP municipal yields.

These results suggest that GFMs like AEF have some advantage in predicting yields over traditional remote sensing sources, but only once the embeddings' distributional structure is used. The

NDVI baseline shares its lineage with the model developed by You et al. (2017); the simplest model using AEF features, AEF mean, corresponding to the model tested by Ma et al. (2026), performs on par with it. Relative to that most basic AEF model, our best-performing AEF model (Agg-NN) achieves 34% lower RMSE and explains 27 percentage points more of the variation. While these findings may be interesting in their own right, we will show next that they do not necessarily imply anything about downscaling performance.

Table 2: Municipality-level validation performance (DGSIAP spring-summer maize, 2017–2024). Every model is trained with the same randomly held-out 20% of municipalities excluded from training and evaluated on the same held-out municipality-years (the intersection of municipality-years for which all models produce a prediction), so N is identical across rows. AEF models' ADC-level predictions are aggregated to the municipality level using agricultural-land-area weights; the NDVI model is trained and scored directly at the municipality level. RMSE in t/ha.

Model	N	R^2	RMSE
Landsat-derived features			
NDVI	3,571	0.534	1.490
AEF-derived features			
AEF mean	3,571	0.520	1.512
Agg-NN	3,571	0.790	1.001
AEF Hist	3,571	0.608	1.366
AEF Hist Ens.	3,571	0.627	1.334

4.2 ADC-level yield downscaling

We now evaluate the same models on the downscaling task: predicting at the *área de control* (ADC) level. Models trained on municipality-level data are applied to ADC-level features (a scale transfer), and validated against the 2022 INEGI Agricultural Census, which provides ground-truth yield observations at the ADC level. We compare five families of predictions. One is the Landsat-derived NDVI baseline: for each spatial unit we fit a per-year three-harmonic model to the Landsat NDVI time series to gap-fill clouds, reconstruct NDVI at three phenology-anchored growth stages keyed to municipality-specific planting dates, and summarize the resulting cropland-pixel distributions with the same per-dimension recipe used for the AEF features (Section 3.4.2). The other four are AEF-derived: the AEF mean-pooled embedding model (AEF mean), an AEF-based aggregation-constrained neural network (Agg-NN), an AEF distributional model (AEF Hist) that replaces the mean-pooled embeddings with the 512 subsampled-histogram features described in the methodology, and an AEF distributional ensemble model (AEF Hist Ens., described in Section 3.3.2). For each, we report the raw (uncorrected) predictions, the ex-post additively corrected predictions, as well as the shrink corrected results. In the results tables, models are grouped by feature source—Landsat-derived (the NDVI baseline) and AEF-derived. As a benchmark, we also report the accuracy of simply assigning each ADC the DGSIAP municipal average yield, which cannot explain any within-municipality variation but may perform well under traditional metrics like RMSE and R^2.

To decompose prediction accuracy, we report the standard R^2, as well as a between-municipality R^2 (computed on municipality-level group means of predicted and actual yields) and a within-municipality R^2 (computed on deviations from group means). The between-R^2 captures how well the model ranks municipalities, while the within-R^2 captures how well the model differentiates yields across ADCs within the same municipality (see equation (4)). By construction, the DGSIAP municipal average has a within-R^2 of zero, as it assigns identical predicted yields to all ADCs in a municipality.

Table 3 presents ADC-level accuracy metrics for the spring-summer (P-V) season, the cycle that accounts for the majority of Mexican maize production, with models grouped by feature source (the combined-season and fall-winter results appear in the appendix). Several results stand out. First, the AEF-based models substantially outperform the Landsat NDVI baseline: the AEF Hist and AEF Hist Ens. models both reach a raw R^2 of 0.60 (0.605 and 0.596), compared to 0.44 for the corrected NDVI baseline (NDVI Corr.). Second, the additive ex-post correction sharply improves the baseline whose raw predictions carry little municipal structure (NDVI from 0.31 to 0.44) but penalizes the models that already capture it, since the correction overwrites their fitted municipal means with the noisier season-matched DGSIAP anchor (AEF mean from 0.47 to 0.44; AEF Hist Ens. from 0.60 raw to 0.50 corrected). Third, the AEF-based models outperform the DGSIAP municipal average benchmark ($R^2 = 0.41$), demonstrating that the downscaled predictions add information beyond what is contained in the aggregate survey data. Figure 5 presents hexbin scatter plots of predicted versus reported yields for the combined season, showing the deployed Shrink specification of each model alongside the DGSIAP municipal-average benchmark (the spring-summer and fall-winter scatter plots appear separately in Figure A.3). The scatter plots visually confirm the patterns observed in the accuracy metrics: the AEF-based models produce predictions that cluster more tightly around the 1:1 line, while the NDVI baseline exhibits greater dispersion and the municipal-average benchmark collapses onto a small set of municipality-wide values. Figure 6 maps census and predicted yields on the matched 2022 sample; Appendix Figure A.2 maps the prediction errors, which are less spatially clustered for the corrected predictions than for the DGSIAP municipal-average benchmark.

Most importantly, the AEF Hist Ens. model achieves the highest within-R^2 of any specification (0.18 raw, 0.17 corrected), whereas the within-R^2 of the AEF mean model is essentially zero (AEF mean Corr., 0.04). Applying the within-municipality shrinkage of Section 3.5.2 (AEF Hist Ens. Shrink) raises this within-R^2 further to 0.23 while also lifting the overall R^2 to 0.61, at no additional data or training cost. As we show below, this is close to the intrinsic ceiling: an oracle model trained directly on ADC-level census labels reaches a within-municipality R^2 of only 0.30. This suggests that there is an inherent limit in the amount of within-municipality variation that can be explained by the AEF features. On the overall and between-municipality margins the ensemble matches the AEF Hist model (0.60 and 0.75 raw for both; after shrinkage AEF Hist reaches 0.63 against the ensemble's 0.61): the gain from ensembling is concentrated in the within-municipality margin ($0.10 \rightarrow 0.18$ raw; $0.20 \rightarrow 0.23$ shrunk), the margin that municipal statistics cannot supply.

The Landsat NDVI baseline exhibits the same distinction between aggregate and sub-municipal accuracy. While it attains a municipality-level R^2 of 0.53 (Table 2), its raw ADC-level R^2 drops to 0.31. After the additive ex-post correction its overall R^2 rises to 0.44, driven by a strong municipality-level signal (between-R^2 of 0.61). Its within-municipality R^2, however, remains negative (-0.03, raw and corrected): the correction shifts all ADCs in a municipality by the same amount and cannot manufacture sub-municipal signal that the features do not contain. Applying the shrinkage regularization lifts this baseline's within-R^2 to a modestly positive 0.09, roughly a third of the 0.23 attained by the best AEF specification. This result contrasts with Ma et al. (2026), who report that AEF-based models significantly underperformed RS-based models in scale-transfer experiments in the U.S. context. One potential explanation is that they compare to a model specifically developed for scale transfer, whereas we use a relatively simple model in line with the design goals of our approach.

Another notable finding is that municipality-level prediction accuracy does not necessarily predict downscaling performance. While most models tested do exhibit a stable ranking, the Agg-NN model is a clear exception (Figure 8). While Agg-NN outperformed the other models in the traditional supervised learning setting using municipality-level data, it has the worst within-R^2, at -0.39 (Table 3). This reversal suggests that practitioners should interpret aggregate performance cautiously when the ultimate goal is downscaling. It also motivates the ex-ante downscaling skill index of Section 4.4, which estimates downscaled reliability without any farm-level ground truth.

All rows of the accuracy tables are evaluated on the sample where the AEF Hist Ensemble is defined: ADCs with at least one cropland pixel under the ESA WorldCover mask (58,022 census-matched ADCs in the spring-summer season). Census-matched ADCs with no cropland pixels—12,173 in the spring-summer season—are excluded for every model rather than being scored on features that carry no information there: on these ADCs the masked histogram features are entirely missing, and a model scored on them anyway attains an R^2 of -0.07, while the land involved is overwhelmingly marginal (mean census yield of 1.28 t/ha, against 3.00 t/ha on the retained sample). The NDVI baseline's slightly smaller N within this sample reflects its own extraction coverage. The common-sample comparison in the appendix further intersects all models onto a single shared set of ADCs and leaves the ranking unchanged; the appendix also gives a full accounting of each row's sample size.

4.3 The role of aggregated data quality

A natural question is whether the quality of municipality-level training data limits prediction accuracy, particularly for within-municipality variation. Our model is trained on DGSIAP municipal yield estimates, which are survey-based and may contain measurement error. To investigate this, we conduct a thought experiment: we retrain the same architecture using INEGI 2022 census municipality-level yields as the training target instead of DGSIAP. Since the census is the same data source used for ADC-level evaluation, this exercise isolates the role of training data quality from other sources of prediction error.

Table 4 presents the results for all models. For each model we report the raw predictions, the additive correction toward DGSIAP municipality yields, and the additive correction toward the INEGI census municipality yields.

The additive correction toward census data substantially improves between-R^2 for every model (e.g., AEF mean: $0.63 \to 0.95$; Agg-NN: $0.68 \to 0.94$; AEF Hist Ens.: $0.75 \to 0.95$). However, within-R^2 is essentially unchanged by the correction (AEF mean: $0.02 \to 0.04$; Agg-NN: $-0.41 \to -0.37$). This is by construction: the additive correction shifts all ADCs within a municipality by the same amount, so it cannot alter within-municipality variation.

Retraining the Agg-NN directly on census municipal yields (Table 4, final panel) achieves comparable between-R^2 (0.93) but improves within-R^2 only modestly, from -0.41 to -0.21: it remains

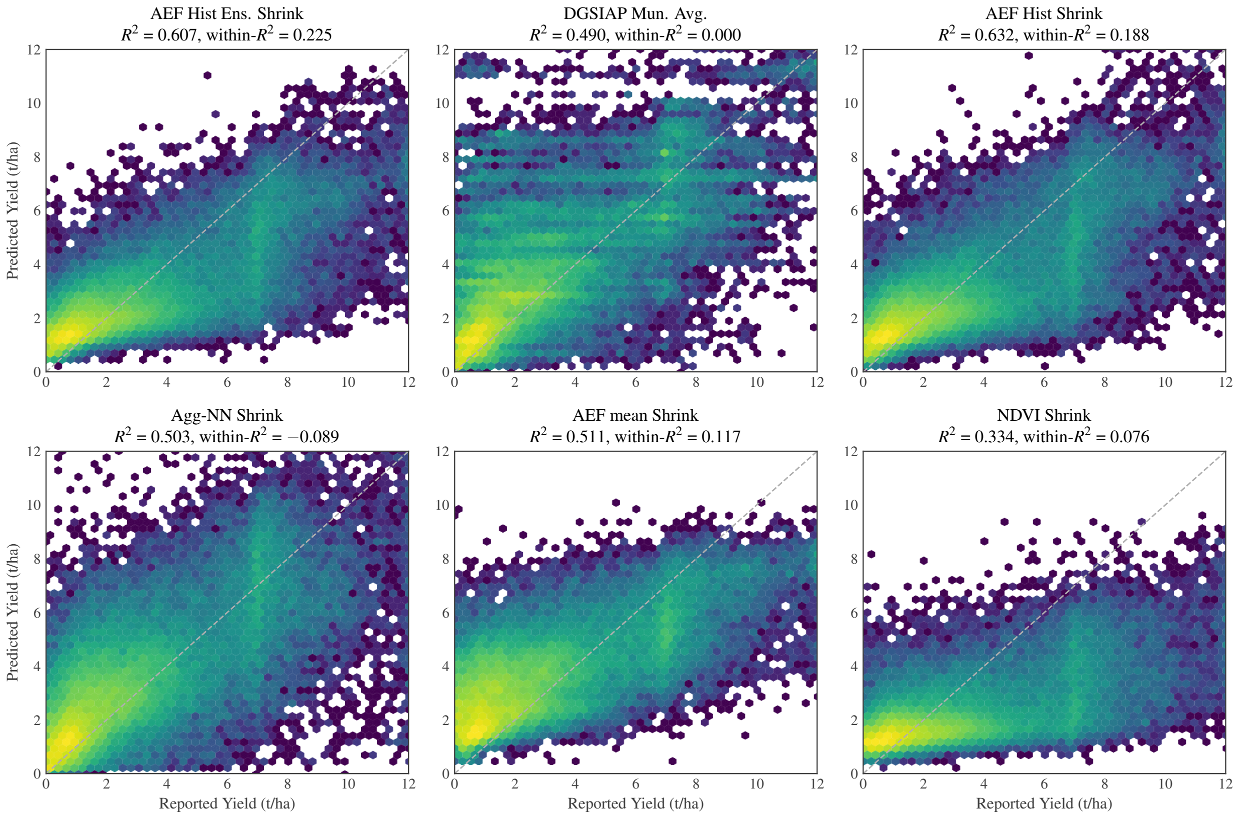
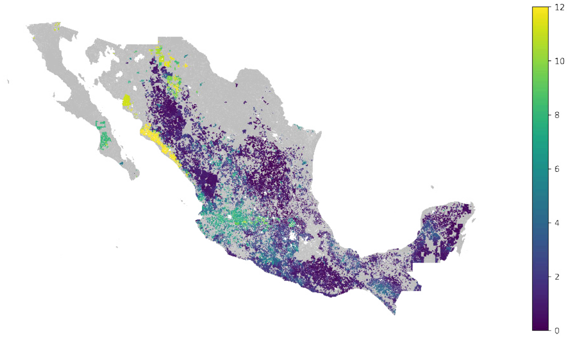
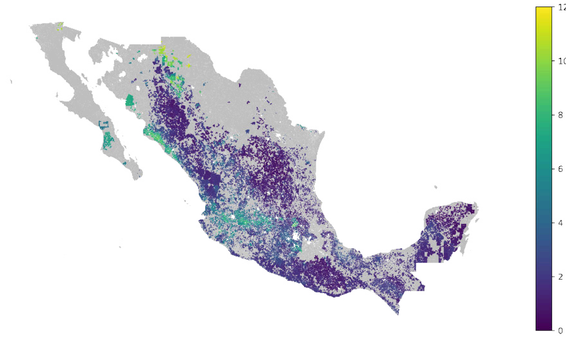


Figure 5: Predicted vs. reported maize yield at the ADC level (combined season, 2022; t/ha). Each panel is a hexbin density plot for the deployed shrinkage-corrected (Shrink) specification of one model, together with the DGSIAP municipal-average benchmark, which assigns every ADC its municipality's DGSIAP yield and therefore has a within-municipality R^2 of zero by construction. Panel titles report the overall and within-municipality R^2 of Table A.1; the dashed line indicates perfect prediction and color intensity reflects observation density.

Figure 6: Census ground truth and predictions at the *área de control* (ADC) level, 2022 (t/ha, shared 0–12 scale). Both panels show the same matched sample of ADCs carrying both a 2022 census maize yield and an AEF Hist Ens. Shrink prediction; ADCs outside the matched sample are grey in both panels.



(a) Census ADC yields (2022)



(b) Predicted ADC yields (2022)

Table 3: Accuracy metrics for maize yield predictions vs. INEGI 2022 census, ADC level — spring-summer (P-V) season. Corrected rows use ex-ante agricultural-land weights for the municipal anchor, and the anchor is the DGSIAP municipal maize yield for the spring-summer season only, matching the census target scored here. All rows are evaluated on the ADCs for which the AEF Hist Ensemble is defined (at least one cropland pixel; see the sample accounting in the appendix). RMSE in t/ha. The Oracle (ADC-trained) row is a non-deployable upper bound that trains HistGradientBoosting directly on ADC-level census labels (5-fold GroupKFold over municipalities); it bounds how much yield signal the embeddings carry.

Model	N	R^2	Between R^2	Within R^2	RMSE
Landsat-derived features					
NDVI Raw	56,101	0.313	0.526	−0.025	2.448
NDVI Corr.	55,817	0.437	0.605	−0.028	2.215
NDVI Shrink	56,101	0.341	0.525	0.085	2.398
AEF-derived features					
AEF mean Raw	58,022	0.474	0.632	0.032	2.123
AEF mean Corr.	57,731	0.436	0.573	0.039	2.196
AEF mean Shrink	58,022	0.499	0.632	0.123	2.073
Agg-NN Raw	58,022	0.413	0.679	−0.387	2.244
Agg-NN Corr.	57,731	0.387	0.650	−0.382	2.290
Agg-NN Shrink	58,022	0.498	0.680	−0.075	2.075
AEF Hist Raw	58,022	0.605	0.754	0.098	1.840
AEF Hist Corr.	57,731	0.475	0.591	0.090	2.120
AEF Hist Shrink	58,022	0.633	0.755	0.200	1.774
AEF Hist Ens. Raw	58,022	0.596	0.748	0.181	1.861
AEF Hist Ens. Corr.	57,731	0.498	0.601	0.172	2.073
AEF Hist Ens. Shrink	58,022	0.610	0.748	0.234	1.828
Benchmark					
DGSIAP Mun. Avg.	57,731	0.409	0.555	0.000	2.250
Oracle (ADC-trained)	57,774	0.754	0.871	0.299	1.454

negative even with perfect municipality-level training data, and even the deployed shrinkage only just turns it positive ($-0.21 \to 0.08$).

Correcting toward census yields lifts the overall R^2 to 0.73–0.76 for the strongest models (AEF Hist Ens. 0.76, AEF Hist 0.74, AEF mean 0.73), but only the AEF distributional specifications retain a positive within-R^2 (AEF Hist Ens. 0.17, AEF Hist 0.08), whereas the AEF mean model (0.04), the Agg-NN, and the NDVI baseline keep zero or negative within-R^2 even after correcting to perfect municipality-level targets. This reflects the fact that the AEF distributional predictions already capture meaningful within-municipality variation, while the additive correction (and the Agg-NN's municipality-level training objective) cannot. Together, these results indicate that improving within-municipality prediction will require either finer-grained training data, the within-municipality shrinkage (the Shrink rows of Tables 3 and A.1), or architectural innovations that preserve sub-municipal feature heterogeneity during training.

The shrinkage of Section 3.5.2 improves within-municipality resolution at no cost to overall fit. For the AEF Hist Ensemble (within-municipality R^2 of 0.17, combined season), the census evaluation data imply $\rho = 0.48$ and $r = 0.72$ (the evaluation-data counterparts of the public estimates $\hat{\rho} = 0.48$ and $\hat{r} = 0.67$ of Section 3.5.2): because r exceeds ρ, the predicted within-municipality

Table 4: Thought experiment: ADC-level yield prediction accuracy when correcting toward INEGI census municipality yields vs. DGSIAP municipality yields, for all models. All evaluated against the INEGI 2022 census at the ADC level (combined season), on the ADCs for which the AEF Hist Ensemble is defined (see the sample accounting in the appendix). "+ Shrink" rows additionally scale each ADC's predicted deviation from its municipality-mean prediction by the deployed $\lambda = 0.72$ (Section 3.5.2). RMSE in t/ha.

Model	N	R^2	Between R^2	Within R^2	RMSE
Panel: NDVI					
Raw	57,638	0.304	0.518	−0.039	2.531
Raw + Shrink	57,638	0.334	0.518	0.076	2.477
Corr. (→ DGSIAP)	57,421	0.531	0.632	−0.034	2.079
Corr. (→ DGSIAP) + Shrink	57,421	0.560	0.632	0.078	2.014
Corr. (→ Census)	57,638	0.698	0.930	−0.032	1.666
Corr. (→ Census) + Shrink	57,638	0.727	0.930	0.079	1.585
Panel: AEF mean					
Raw	59,641	0.487	0.634	0.024	2.154
Raw + Shrink	59,641	0.511	0.634	0.117	2.103
Corr. (→ DGSIAP)	59,417	0.516	0.591	0.037	2.093
Corr. (→ DGSIAP) + Shrink	59,417	0.539	0.591	0.123	2.044
Corr. (→ Census)	59,641	0.728	0.954	0.041	1.569
Corr. (→ Census) + Shrink	59,641	0.750	0.954	0.125	1.505
Panel: Agg-NN					
Raw	59,641	0.420	0.683	−0.413	2.290
Raw + Shrink	59,641	0.503	0.683	−0.089	2.120
Corr. (→ DGSIAP)	59,417	0.450	0.657	−0.406	2.232
Corr. (→ DGSIAP) + Shrink	59,417	0.532	0.657	−0.086	2.059
Corr. (→ Census)	59,641	0.613	0.940	−0.367	1.871
Corr. (→ Census) + Shrink	59,641	0.689	0.940	−0.069	1.676
Panel: AEF Hist					
Raw	59,641	0.604	0.753	0.079	1.893
Raw + Shrink	59,641	0.632	0.753	0.188	1.824
Corr. (→ DGSIAP)	59,417	0.545	0.605	0.081	2.031
Corr. (→ DGSIAP) + Shrink	59,417	0.572	0.605	0.189	1.968
Corr. (→ Census)	59,641	0.740	0.953	0.083	1.535
Corr. (→ Census) + Shrink	59,641	0.767	0.953	0.190	1.452
Panel: AEF Hist Ens.					
Raw	59,641	0.592	0.746	0.169	1.921
Raw + Shrink	59,641	0.607	0.746	0.225	1.886
Corr. (→ DGSIAP)	59,417	0.575	0.620	0.170	1.963
Corr. (→ DGSIAP) + Shrink	59,417	0.589	0.620	0.226	1.930
Corr. (→ Census)	59,641	0.762	0.951	0.171	1.469
Corr. (→ Census) + Shrink	59,641	0.776	0.951	0.226	1.425
Panel: Agg-NN (census-trained)					
Raw	59,641	0.661	0.933	−0.213	1.752
Raw + Shrink	59,641	0.736	0.933	0.081	1.545
Corr. (→ DGSIAP)	59,417	0.436	0.552	−0.205	2.259
Corr. (→ DGSIAP) + Shrink	59,417	0.510	0.552	0.083	2.106
Corr. (→ Census)	59,641	0.669	0.958	−0.206	1.729
Corr. (→ Census) + Shrink	59,641	0.744	0.958	0.084	1.522

deviations are over-dispersed relative to their accuracy, and scaling each ADC's predicted deviation from its municipality mean by $\lambda^\star = \rho/r \approx 0.66$ would maximize the within-municipality R^2 at $\rho^2 = 0.23$—confirming, on the evaluation data, the public estimate $\lambda = 0.72$ deployed throughout (Section 3.5.2; the flat within-R^2 curve makes the two indistinguishable). The "shrink" rows of Tables 3 and A.1 report that the deployed shrinkage raises the within-municipality R^2 from 0.18 to 0.23 and the overall R^2 from 0.60 to 0.61 (spring-summer season), with similar gains in the combined season (0.17 to 0.22). Because the shrinkage is a generic post-processing step, we apply the same deployed $\lambda = 0.72$ to every model. It improves every model's within-municipality R^2 (NDVI $-0.04 \to 0.08$; AEF mean $0.02 \to 0.12$; Agg-NN $-0.41 \to -0.09$) and raises the overall R^2 of the Agg-NN from 0.42 to 0.50 (combined season), without requiring additional data or training; but a common factor cannot rescue a model whose within-municipality spread is mostly noise—the Agg-NN's within-R^2 remains negative, and would turn positive only under much harder, model-specific shrinkage.

To bound how much signal the embeddings carry in principle, we turn back to our oracle model, reported as the "oracle (ADC-trained)" row under the benchmark sections of Tables 3 and A.1; it achieves an overall R^2 of 0.75 but a within-municipality R^2 of only 0.30 (spring-summer season). Two conclusions follow. First, the within-municipality ceiling is low: even with farm-level labels, the embeddings resolve only $R^2 \approx 0.30$ of within-municipality variation, somewhat above the 0.23 that the shrinkage attains from aggregate-only training. The binding constraint on within-municipality resolution is therefore the information content of the annual embeddings, not the training scheme or the resolution of the labels. Second, the gains from finer labels accrue almost entirely to the level (between-municipality) component: ADC labels would raise the overall R^2 from our deployed 0.61 to 0.75, while the within-municipality component is nearly saturated by the aggregate-only model plus shrinkage. This is consistent with the census thought experiment above: better training data improves prediction of levels more than of sub-municipal variation.

4.4 Ex-ante targeting: predicting where downscaling works

A practical concern for any downscaling procedure is knowing, before any farm-level ground truth is observed, where its predictions are reliable. Our results suggest that the method's ability to resolve within-area yield variation is governed by the share of within-area yield variance that is agro-ecological (e.g. terrain, water availability, soil, and crop fraction), and therefore potentially encoded by the AEF embeddings, rather than management-driven (e.g. fertilizer, seed variety, agrochemical use). Our external CIMMYT validation below (Section 4.9) illustrates this pattern: prediction skill is strong only where plot yields track municipal averages and is worse where management-driven yield premiums dominate.

We thus develop an ex-ante downscaling skill index where a fixed score for each municipality is derived from features requiring no ADC-level ground truth. The index is based on the four representativeness drivers suggested by the logic presented above: the within-municipality spread of the AEF embeddings (the square root of the trace of the within-municipality embedding covariance matrix), their effective dimension (the participation ratio of the covariance eigenvalues), the within-municipality heterogeneity of irrigation, and the number of constituent ADCs. We standardize each across municipalities, fix its sign a priori (more agro-ecological heterogeneity that the satellite can observe, and more sub-areas over which to resolve it, raise the index), and average them. No farm-level data enters the index; it is a deterministic function of satellite embeddings and public survey aggregates. Validated against the held-out INEGI census, this purely ex-ante score orders the deployed model's realized within-municipality skill: it correlates with realized within-municipality R^2 at $r = 0.21$ (Spearman $\rho = 0.24$) and with per-municipality rank skill (the Spearman correlation between predicted and observed ADC yields) at $\rho = 0.25$, and it separates municipalities with positive from negative within-municipality skill with an AUC of 0.62. The drivers enter with the predicted signs: within-municipality skill rises with irrigation heterogeneity (bivariate $\rho = +0.21$), ADC count ($+0.18$), AEF embedding spread ($+0.18$), and, most weakly, embedding effective dimension ($+0.08$).

Appendix A.2 gives a step-by-step guide to constructing the index in a new setting from satellite and survey inputs alone.

Figure 7 confirms that realized within-municipality skill rises monotonically across quintiles of the ex-ante downscaling skill index (panel a) and across quintiles of AEF within-municipality embedding spread (panel b): pooled within-municipality R^2 climbs from about -0.1 in the bottom quintile to roughly $+0.3$ in the top, and the median within-municipality rank correlation from about 0.1 to about 0.37. This suggests a simple rule that requires no farm-level ground truth: report farm-level predictions in municipalities the index flags as positive-skill, and treat predictions in low-index municipalities as more speculative. Because the index depends only on satellite embeddings and aggregate survey data, it can be computed in any new setting where the method might be deployed.

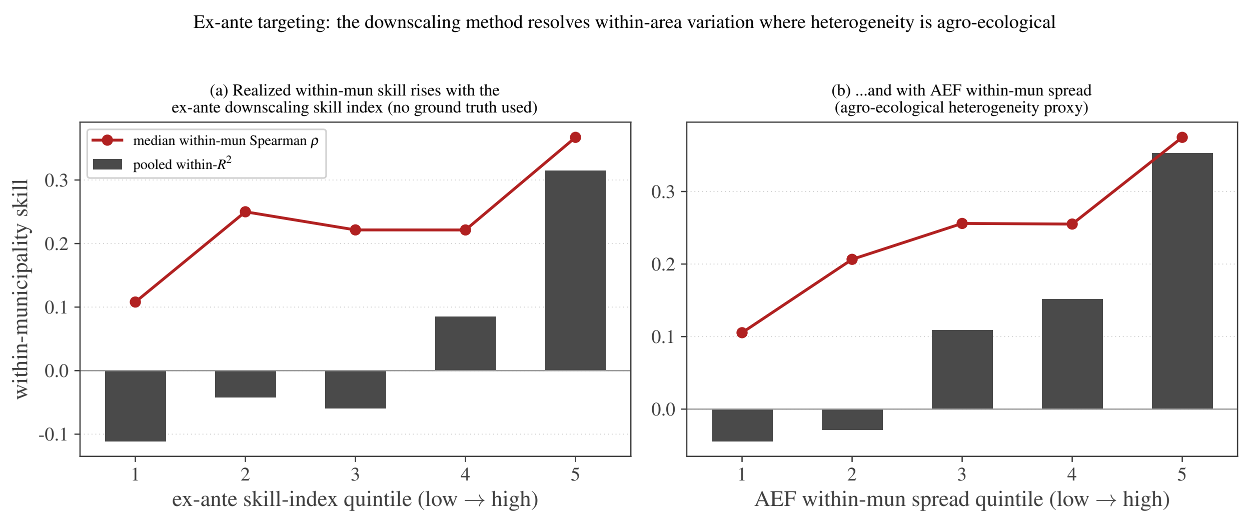


Figure 7: Ex-ante targeting of the downscaling method. Realized within-municipality skill—pooled within-municipality R^2 (bars) and median within-municipality Spearman ρ (line)—rises monotonically across quintiles of (a) the ex-ante downscaling skill index (a composite of four representativeness features of AEF embedding spread and effective dimension, irrigation heterogeneity, and ADC count, computed without any ADC-level ground truth) and (b) AEF within-municipality embedding spread alone, an ex-ante proxy for agro-ecological heterogeneity. Realized skill is that of the deployed AEF Hist Ens. Shrink model ($\lambda = 0.72$, Section 3.5.2). Municipalities with at least five ADCs.

Finally, the downscaling skill index also discriminates across models, making it a tool for ex-ante model selection as well as targeting, and helping to explain the apparent paradox of Table 2.

Ranked by realized farm-level (within-municipality) skill, the five models rank very differently than at the municipality level: the AEF distributional models lead (AEF Hist Ens. pooled within-R^2 of 0.17, AEF Hist 0.08), the AEF mean model (0.02) and the Landsat NDVI baseline (-0.04) are near zero, and the Agg-NN, first at the municipality level, is last at the farm level (-0.41), because its aggregation-constrained objective sacrifices sub-municipal resolution for municipality-level fit.[10] The same ex-ante index correlates with each model's realized per-municipality within-skill, at $+0.45$ to $+0.48$ across the four satellite models (0.48 for the AEF Hist Ens.), with the weakest correlation ($+0.38$) for the Agg-NN. A practitioner can therefore use the index, computed from satellite and survey inputs alone, to decide both where farm-level predictions are reliable and which model to use there. Figure 8 visualizes this inversion: the Agg-NN falls from first at the municipality level to last at the farm level, while the AEF distributional models rise.

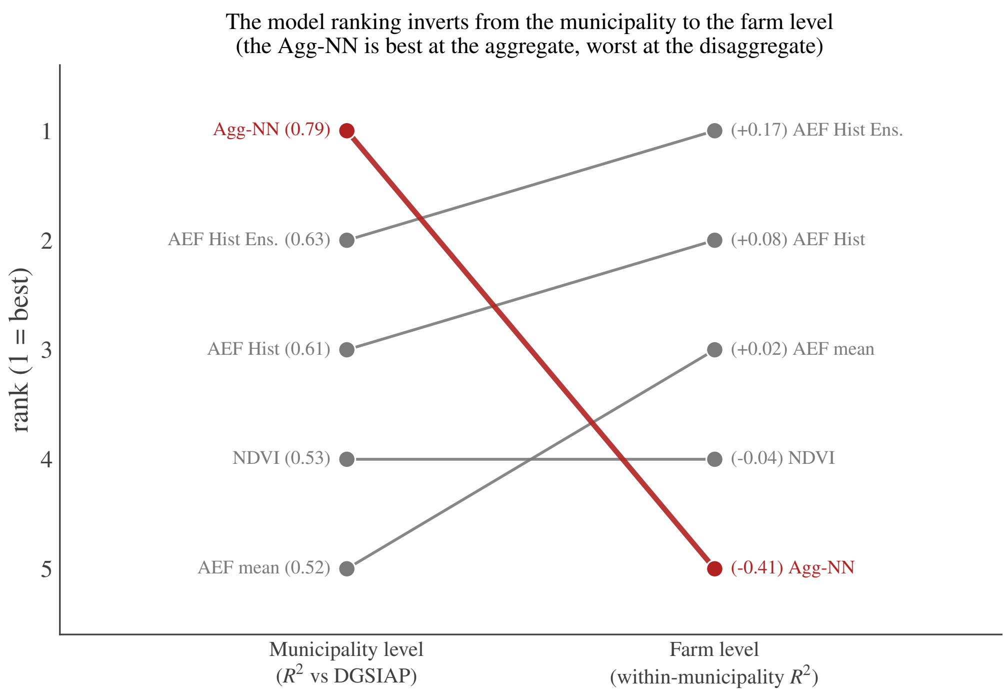


Figure 8: The model ranking inverts between the municipality and farm levels. Left: rank by municipality-level R^2 against DGSIAP (Table 2). Right: rank by farm-level (within-municipality) R^2 against the INEGI census. The Agg-NN (red), trained with a municipality-level aggregation constraint, is best at the aggregate but worst at the disaggregate; the AEF distributional models that lead at the farm level trail only the Agg-NN at the municipality level. Parenthetical values are the underlying R^2.

4.5 Seasonal heterogeneity

We additionally examine the robustness of our results across growing seasons, ADC characteristics, additional crops, and external plot-level data, as well as their application to improving aggregated

[10]These pooled values are computed on the common combined-season sample used by the ex-ante analysis and therefore differ slightly from the corresponding entries of Table A.1.

survey data. We begin with growing seasons. Mexico has two main maize crop cycles: the spring-summer (primavera-verano, P-V) cycle, which accounts for the large majority of maize production and is the season our headline results (Table 3) evaluate, and the fall-winter (otoño-invierno, O-I) cycle, which is concentrated in irrigated areas of northern Mexico. Two features disadvantage the fall-winter evaluation. First, AEF embeddings are calendar-year composites: the 2022 embedding covers the maturation and harvest months of the fall-winter cycle evaluated against the census, but the cycle's planting and establishment months fall in the 2021 embedding year, whereas the spring-summer cycle sits almost entirely within a single embedding year. Second, the deployed models are trained on spring-summer DGSIAP yields. The training season, however, is not the binding constraint: retraining the ensemble on fall-winter DGSIAP yields (10,305 municipality-years) raises the overall fall-winter R^2 from 0.62 to 0.73, but does so by inheriting DGSIAP's fall-winter municipal structure, which itself diverges from the census in this cycle—the retrained model's between-municipality R^2 is -0.58, close to the DGSIAP benchmark's own -0.61 (Table A.2), and its within-municipality R^2 remains negative. Fall-winter accuracy is therefore limited less by the training domain than by DGSIAP–census divergence for this smaller, irrigated cycle. Table A.2 in the appendix presents the corresponding fall-winter accuracy metrics, and Table A.1, also in the appendix, pools both cycles.

The seasonal decomposition reveals important differences. For the fall-winter season, the additive ex-post correction has a particularly large impact. Uncorrected AEF Hist Ens. predictions achieve an R^2 of 0.62, which rises to 0.74 after additive correction, approaching the DGSIAP benchmark of 0.76. This is likely because fall-winter maize is concentrated in a small number of highly productive, irrigated municipalities (primarily in Sinaloa and Sonora), where the additive correction's direct anchoring to DGSIAP regional averages is particularly effective. For the spring-summer season, which captures the bulk of Mexican maize production and features greater heterogeneity in growing conditions, the correction offers no level gains (AEF mean R^2 moves from 0.47 to 0.44). Across both seasons, within-R^2 values remain low or negative for all models, with only the AEF distributional specifications in the spring-summer season reaching clearly positive values, indicating that fine-scale yield variation within municipalities remains the primary challenge, and the one the AEF Hist Ens. Shrink specification addresses most directly (spring-summer within-R^2 of 0.23).

Figure A.3 presents the seasonal scatter plots, showing the predicted versus reported yield distributions for each model separately for the fall-winter and spring-summer seasons.

4.6 Heterogeneity in prediction accuracy by ADC characteristics

We examine how prediction accuracy varies with observable characteristics of the ADC. Table 5 stratifies the AEF Hist Ensemble predictions by terciles of maize planted area, total ADC planted area, maize share of total ADC area, and number of production units. The model performs substantially better on larger, more commercially oriented ADCs. Within-municipality R^2 increases from 0.08 in the bottom tercile of maize area to 0.22 in the top tercile, and from 0.10 to 0.27 moving from ADCs with few production units to those with many. ADCs with intermediate maize shares (40–79% of planted area) achieve the highest within-R^2 (0.22), suggesting the model works best where maize is a dominant but not exclusive crop. These patterns indicate that prediction accuracy is highest where the satellite signal is most representative of the crop of interest, with larger areas with more maize providing a stronger and more stable spectral signature for the model to exploit.

Table 5: AEF Hist Ensemble prediction accuracy by ADC characteristics, maize vs. INEGI 2022 census. RMSE in t/ha.

Subsample	N	R^2	Between R^2	Within R^2	RMSE
All ADCs	59,641	0.592	0.746	0.169	1.921
Spring-summer only	58,022	0.596	0.748	0.181	1.861
Maize area: Bottom tercile [0.00-14.27]	19,881	0.421	0.597	0.082	1.939
Maize area: Middle tercile [14.27-55.98]	19,879	0.552	0.705	0.188	1.882
Maize area: Top tercile [56.00-15434.45]	19,881	0.690	0.784	0.219	1.941
Total ADC area: Bottom tercile [0.00-32.66]	19,880	0.472	0.576	0.133	1.986
Total ADC area: Middle tercile [32.66-146.71]	19,880	0.584	0.698	0.168	1.877
Total ADC area: Top tercile [146.72-27138.35]	19,881	0.673	0.771	0.223	1.898
Maize share: Bottom tercile [0.00-0.40]	19,880	0.478	0.690	0.042	1.881
Maize share: Middle tercile [0.40-0.79]	19,881	0.662	0.738	0.223	1.773
Maize share: Top tercile [0.79-1.00]	19,880	0.593	0.643	0.138	2.095
Num. prod. units: Bottom tercile [3.00-8.00]	20,770	0.467	0.633	0.102	2.291
Num. prod. units: Middle tercile [9.00-26.00]	19,070	0.629	0.748	0.213	1.854
Num. prod. units: Top tercile [27.00-8287.00]	19,801	0.708	0.792	0.268	1.515

4.7 Improving aggregated survey data

Next, we examine the performance of our predictions when aggregated from the ADC level back to the municipality level. This exercise tests whether the downscaling and re-aggregation procedure can improve upon existing survey data. We aggregate ADC-level predictions to the municipality level using each ADC's agricultural-land area as an ex-ante weight, and compare the resulting municipality-level predictions against both the INEGI 2022 census and DGSIAP municipal estimates.

A key methodological point is that the aggregation weights must be *ex-ante*. To combine ADC-level predictions into a municipality estimate we weight each ADC by its agricultural-land area (a proxy for maize area available without the census), not by census-reported planted area, which would use the ground truth we are testing against. Table 6 presents the municipality-level accuracy metrics under this ex-ante weighting. When compared against the INEGI census (Panel A), the AEF Hist Ensemble achieves an R^2 of 0.75 across 1,983 municipalities, above the DGSIAP municipal estimates ($R^2 = 0.62$, $N = 2{,}039$) and with the lowest RMSE (1.13) of any aggregated model. The AEF mean model, by contrast, reaches only 0.65, close to DGSIAP, while the AEF Hist model attains 0.75, matching the ensemble: the improvement over the survey comes from the histogram features rather than the mean-pooled embeddings. The Shrink (agg.) rows confirm that the deployable shrink specification gives up almost nothing at the municipal level (AEF Hist Ens. from 0.75 to 0.75) and even improves the Agg-NN (0.60 to 0.63); its within-municipality gains therefore come at no cost to aggregate accuracy. The exception is the NDVI baseline, whose aggregated accuracy dips under shrinkage (0.61 to 0.58), consistent with its within-municipality deviations being mostly noise. The NDVI baseline aggregated from ADC predictions (0.61) and the Agg-NN (0.60) fall short of DGSIAP, while the municipality-trained NDVI model (0.68) edges past it, all well below the AEF Hist Ensemble.

When compared against DGSIAP (Panel B), the aggregated INEGI census yields reach an R^2 of 0.68 and the AEF Hist Ensemble 0.71, while the AEF mean model reaches 0.72 and the Agg-NN, which is trained to match DGSIAP at the municipality level, correlates most strongly with DGSIAP

(0.89), as expected. The discrepancy between DGSIAP and the census ($R^2 = 0.68$) confirms that meaningful measurement differences exist between the two data sources. The AEF Hist Ensemble correlates more strongly with the census (0.75) than DGSIAP does (0.62); on the common set of municipalities, this difference carries a municipality-bootstrap 95% confidence interval of [0.09, 0.22], excluding zero. This suggests that the downscale-then-re-aggregate procedure, using only public aggregate data and an ex-ante area proxy, produces municipality-level yield estimates that are more accurate than the existing survey.

Table 6: Municipality-level maize yield prediction results, 2022. ADC-level predictions are aggregated to the municipality level weighting each ADC by its *ex-ante* agricultural-land area (a proxy for maize area that does not use the census), except NDVI GB (mun.), which is trained directly at the municipality level (random municipality-year cross-validation). Shrink (agg.) rows first apply the deployed within-municipality shrinkage ($\lambda = 0.72$, deployable from predictions alone), then aggregate the shrunk ADC predictions with the same ex-ante weights. RMSE in t/ha.

Panel A: vs. INEGI Census (aggregated)			
Model	N	R^2	RMSE
NDVI (agg.)	1,968	0.613	1.416
NDVI Shrink (agg.)	1,968	0.583	1.470
NDVI GB (mun.)	1,949	0.677	1.261
AEF mean (agg.)	2,054	0.645	1.333
AEF mean Shrink (agg.)	2,054	0.645	1.334
AEF Hist (agg.)	1,983	0.749	1.135
AEF Hist Shrink (agg.)	1,983	0.749	1.135
AEF Hist Ens. (agg.)	1,983	0.752	1.129
AEF Hist Ens. Shrink (agg.)	1,983	0.749	1.136
Agg-NN (agg.)	2,054	0.604	1.409
Agg-NN Shrink (agg.)	2,054	0.632	1.358
DGSIAP Mun. Avg.	2,039	0.623	1.370
Panel B: vs. DGSIAP Municipal Estimates			
Model	N	R^2	RMSE
NDVI (agg.)	1,938	0.492	1.743
NDVI Shrink (agg.)	1,938	0.445	1.822
NDVI GB (mun.)	1,949	0.672	1.380
AEF mean (agg.)	2,023	0.720	1.274
AEF mean Shrink (agg.)	2,023	0.715	1.285
AEF Hist (agg.)	1,953	0.691	1.353
AEF Hist Shrink (agg.)	1,953	0.685	1.366
AEF Hist Ens. (agg.)	1,953	0.706	1.319
AEF Hist Ens. Shrink (agg.)	1,953	0.693	1.348
Agg-NN (agg.)	2,023	0.892	0.793
Agg-NN Shrink (agg.)	2,023	0.892	0.790
INEGI Census (agg.)	2,039	0.677	1.370

4.8 Extension to other crops

To evaluate the generalizability of our approach beyond maize, we predict yields for four additional crops: sorghum, sugarcane, wheat, and avocados. These crops span different growing seasons (annual, fall-winter, perennial) and production systems, and all are extensively grown in Mexico, although to a lesser degree than maize.

Table A.3 presents ADC-level accuracy metrics for each crop, comparing our predictions against the DGSIAP municipal average benchmark. Performance varies substantially across crops. Sorghum, the most similar crop to maize in terms of growing season and production geography, achieves the best results among non-maize crops (AEF Hist Ens. $R^2 = 0.46$) and is the only crop besides maize whose uncorrected predictions carry positive within-municipality signal (within-R^2 of 0.14, rising to 0.16 after correction), indicating that the AEF embeddings capture some within-municipality yield variation for this crop. The ensemble's between-R^2 (0.56) clears the DGSIAP benchmark (0.37) even without the ex-post correction. Sugarcane shows an interesting pattern: while the raw AEF Hist Ensemble outperforms the DGSIAP benchmark at the ADC level ($R^2 = 0.28$ vs. 0.16), the additive correction pushes the overall R^2 to -0.02. This occurs because DGSIAP and census measurements diverge for sugarcane, so correcting toward DGSIAP can increase error relative to the census. For wheat, the AEF Hist Ensemble reaches an R^2 of 0.42 (0.53 after correction), though accuracy remains below the DGSIAP benchmark (0.62). Avocados remain challenging for all methods, with negative R^2 values even after correction, likely reflecting the difficulty of predicting perennial tree crop yields where year-to-year variation depends on tree age, alternate bearing cycles, and orchard management practices not captured in annual satellite embeddings. As with maize, the deployed within-municipality shrinkage (the "AEF Hist Ens. Shrink" rows, $\lambda = 0.72$) moves each crop's within-R^2 toward ρ^2 at no additional cost, roughly halving the negative within-R^2 of sugarcane ($-0.36 \rightarrow -0.15$) and wheat ($-0.56 \rightarrow -0.27$); because their within-municipality spread carries less signal than maize's, a common factor narrows but does not close the gap.

When these ADC-level predictions are aggregated back to the municipality level, a clear pattern emerges for the non-maize crops. The AEF-aggregated predictions correlate far more strongly with DGSIAP municipal estimates than with the INEGI census (Table A.4 in the appendix; e.g., avocado $R^2 = 0.63$ against DGSIAP vs. -0.69 against the census; sugarcane 0.75 vs. 0.40, aggregating with the same ex-ante agricultural-land weights as Table 6). Since the model is trained on DGSIAP, it inherits DGSIAP's measurement characteristics, and the systematic DGSIAP–census divergence is more pronounced for these crops than for maize, reinforcing the importance of ground-truth source selection when training and evaluating remote sensing models.

4.9 External validation: CIMMYT farmer trial plots

As an external validation exercise, we apply the AEF Hist Ensemble model to predict maize yields at CIMMYT farmer trial plots in Mexico. CIMMYT maintains a network of managed farmer plots across Mexico with georeferenced yield measurements, providing an independent source of ground-truth data at a finer spatial scale than the INEGI census ADCs. For each plot we extract plot-level AEF features—mean embeddings, distributional statistics, and binned histograms—directly from the pixels within each plot boundary, and generate predictions using the same ensemble model trained on DGSIAP municipality-level data. Because the fields are small relative to the land-cover product, 28% of matched plot-years contain no WorldCover cropland pixels inside the plot boundary and thus carry empty AEF features; excluding these plots leaves the results essentially unchanged (see the notes to Table 7). We evaluate on the 23,670 matched maize-grain plot-years over 2017–2022, after excluding the 3.2% of observations with implausible recorded yields (outside 0.3–20 t/ha, reflecting evident data-entry errors of up to 16,600 t/ha in the raw records).

Table 7 presents the results. The ensemble achieves an overall R^2 of 0.29 (between-R^2 0.44, within-R^2 0.18), with a Pearson correlation of 0.79 and a Spearman rank correlation of 0.69. The informative comparison is the naive baseline that assigns every plot its DGSIAP municipal mean yield: it posts a similar overall R^2 (0.25) but, by construction, zero within-municipality skill (within-R^2 of -0.03 in the matched sample). The ensemble's within-R^2 of 0.18 is therefore its marginal plot-level skill beyond the municipal anchor, close to the 0.17–0.18 within-R^2 range it attains against the census at the ADC level, now confirmed on fully independent plot-level ground truth. The comparison across model families mirrors the census results: the AEF Hist model also retains positive within-R^2 (0.14), whereas the AEF mean model, despite the highest overall R^2 in the panel (0.37), has none (-0.02). Applying the within-municipality shrinkage exactly as deployed in the ADC pipeline ($\lambda = 0.72$, estimated from public data and ex ante with respect to the CIMMYT data) leaves the ensemble's R^2 essentially unchanged but improves its rank agreement with plot yields (Pearson from 0.79 to 0.80; Spearman from 0.69 to 0.72).

CIMMYT plots are intensively managed trials, and their yields sit systematically above the yields of the areas that surround them. Aggregated to the municipality level, the model's predictions track DGSIAP closely ($R^2 = 0.61$, $r = 0.78$), as they should given municipal-level training. CIMMYT municipal means also rank municipalities similarly to DGSIAP ($r = 0.72$) but with a substantial level premium, so their municipality-level R^2 against DGSIAP is strongly negative (-0.62): trial plots inherit the local productivity gradient while out-yielding the local average.

Performance accordingly varies with how representative local trial yields are of their surroundings, which Panel B of Table 7 reports by grouping plots according to their municipality's mean CIMMYT-to-DGSIAP yield ratio. Within-municipality skill is positive in every band (within-R^2 of 0.10–0.19), but overall accuracy falls monotonically as the local management premium grows, from $R^2 = 0.52$ where trial yields sit below the municipal average to 0.06 where they exceed twice it and satellite features cannot see the management driving yields. Even in that band the correlation remains high (Pearson $r = 0.82$, Spearman 0.80), indicating that the model still orders plots by underlying land productivity.

The model also exhibits meaningful within-municipality differentiation where plot yields are representative of their surroundings. In municipalities whose mean CIMMYT yield lies between 0.9 and 1.3 times the DGSIAP average (the representative band of Table 7; 6,425 plot-years), the within-municipality R^2 is 0.15, with a Pearson correlation of 0.81 and a Spearman rank correlation of 0.71. The strongest within-municipality performance appears in states where trials sit close to area averages, such as Durango (within-R^2 of 0.41), Guanajuato (0.35), and Querétaro (0.48). These patterns reinforce the interpretation that the model captures area-level agro-ecological conditions rather than plot-level management effects, and that its predictions are most informative where plot yields are representative of broader agricultural conditions.

Table 7: Predictions evaluated against CIMMYT plot-level maize yields, 2017–2022

	N	R^2	Btw-R^2	Wtn-R^2	Pearson	Spearman
Panel A: Plot-level accuracy						
AEF Hist Ens.	23,670	0.288	0.439	0.182	0.786	0.687
AEF Hist Ens. Shrink	23,670	0.290	0.439	0.186	0.802	0.723
AEF Hist	23,670	0.360	0.518	0.137	0.784	0.699
AEF mean	23,670	0.371	0.231	−0.023	0.615	0.498
DGSIAP Mun. Avg. (naive)	23,670	0.251	0.459	−0.033	0.629	0.681
Panel B: By municipality mean CIMMYT/DGSIAP ratio (AEF Hist Ens. Shrink)						
Mun. ratio < 0.9 (below avg.)	3,806	0.522	0.510	0.101	0.741	0.748
Mun. ratio 0.9–1.3 (representative)	6,425	0.371	0.620	0.146	0.809	0.713
Mun. ratio 1.3–2.0	8,371	0.153	0.396	0.193	0.812	0.732
Mun. ratio > 2.0 (management premium)	5,068	0.057	−0.004	0.114	0.822	0.798

Notes: models trained on DGSIAP municipal spring-summer maize yields (2017–2022) and applied to plot-level AEF features. CIMMYT yields are self-reported; observations outside 0.3–20.0 t/ha ($\approx$3.2%, evident data-entry errors) are excluded, and 28% of matched plot-years contain no WorldCover cropland pixels inside the plot polygon (empty features); results are insensitive to excluding the latter. Between- and within-R^2 use municipality groupings. The naive baseline assigns every plot its DGSIAP municipal mean, so its within-R^2 is zero by construction; the ensemble's within-$R^2 = 0.182$ is its marginal plot-level skill beyond the municipal level data alone. The Shrink row applies the deployed $\lambda = 0.72$ (Section 3.5.2), estimated from public data and ex ante with respect to the CIMMYT data. Panel B groups plots by their municipality-year's mean CIMMYT-to-DGSIAP yield ratio, an indicator of how representative local trial yields are of area averages.

5 Conclusion

We present and validate an agricultural productivity downscaling approach that relies only on publicly available input and training data while minimizing storage and computation burden. By leveraging AlphaEarth Foundations embeddings (Brown et al., 2025), we demonstrate a methodology that eliminates the need for extensive image pre-processing and sensor-specific feature engineering while achieving competitive prediction accuracy. Importantly for practitioners who lack access to disaggregated ground truth, often a key motivation for pursuing downscaling, we show how aggregated summary statistics of disaggregated features can reasonably predict downscaling accuracy.

Our results deliver several key findings. First, AEF-based models substantially outperform the traditional Landsat NDVI baseline in scale-transfer experiments, achieving an R^2 of 0.60 at the ADC level (spring-summer season) compared to 0.44 for the corrected NDVI baseline (NDVI Corr.). This contrasts with the findings of Ma et al. (2026), who report that AEF underperforms RS in scale-transfer settings in the U.S. This difference may be attributable to our improved featurization, as an AEF model analogous to that of Ma et al. (2026) demonstrates approximately equal performance to the best-performing Landsat input model tested. Second, the subsampled-histogram AEF Hist model, and its ensemble with a percentile model, demonstrate that addressing the distribution mismatch between training (municipality-level) and prediction (ADC-level) features can substantially improve within-municipality differentiation (within-R^2 of 0.10 and 0.18). Third, when ADC-level predictions are aggregated back to the municipality level using an ex-ante area proxy, the AEF Hist Ensemble ($R^2 = 0.75$ against the census) outperforms DGSIAP survey estimates ($R^2 = 0.62$), suggesting that the downscaling procedure, or geospatial foundation models more generally, can be used to improve and augment existing survey data collection efforts.

Fourth, the limits of within-municipality resolution are predictable in advance. We show that an ex-ante downscaling skill index can predict where farm-level predictions are reliable (positive-skill classification AUC of 0.62).

Our findings also contribute to the growing literature on geospatial foundation models by evaluating AEF embeddings in a developing country agricultural context, where data quality challenges and institutional constraints differ from the U.S. settings studied in prior work (Ma et al., 2026).

There are at least two notable directions along which we believe this work can be expanded in the future. The first is to apply our downscaling approach to alternative outcomes. As Ma et al. (2026) showed, there is substantial heterogeneity in the AEF features' capacity to predict different agricultural quantities. The methodology we develop here may prove useful for downscaling other variables, including those outside the agricultural domain. Second, applying our downscaling approach in other developing countries may shed light on agricultural processes at unprecedented levels of spatial and temporal granularity, given the often limited survey data collection in low-income contexts (Ahmad et al., 2024). Indeed, a key characteristic of the methodology we develop here is its accessibility and limited resource requirements.

Data and code availability

The code used in this paper is available at `https://github.com/jaysayre/predicting_yields_rs`. The DGSIAP municipal production statistics and crop-progress data used for training are openly available from DGSIAP (`https://nube.agricultura.gob.mx/datosAbiertos/Agricola.php`), and the AlphaEarth Foundations embeddings are publicly available on Google Earth Engine. The ADC-level Agricultural Census microdata are restricted access, but other researchers wishing to replicate the census-based evaluation can request access to the Agricultural Census through an INEGI microdata laboratory project, as we did under project LM2304. The CIMMYT plot data are described in Gardeazábal-Monsalve et al. (2025).

A Additional Figures and Tables

A.1 Evaluation metric definitions

The evaluation metrics reported throughout the paper are defined as

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2}, \tag{A.1}$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}, \tag{A.2}$$

where N denotes the total number of samples in the evaluation set; y_i is the ground truth yield value; $\hat{y}_i$ is the predicted value; and $\bar{y}$ is the mean of the ground truth values.

A.2 Constructing the ex-ante downscaling skill index in a new setting

Constructing the index in a new setting. Scoring a unit never uses that unit's farm-level yields; the index is a function of satellite and DGSIAP survey inputs alone. *(i) Features.* Take the municipality (or whatever aggregation unit reports survey yields) as the observation and assemble label-free predictors in three groups: survey moments (the cross-year standard deviation and mean of the unit's reported yield, its irrigation share, and its crop share of planted area); unit geometry (the number of constituent sub-areas and their mean size and size dispersion, from an administrative or agricultural-land partition); and, most importantly, the within-unit heterogeneity of the satellite embeddings and of irrigation, summarized as the square root of the trace of the within-unit embedding covariance (total agro-ecological spread), its effective dimension (the participation ratio $(\sum_k \lambda_k)^2/\sum_k \lambda_k^2$ of the covariance eigenvalues), and the standard deviation of sub-area irrigation shares. *(ii) Score.* Standardize each feature across units and average them with signs fixed a priori by theory; because no weights are fit to realized skill, the index requires no reference sample of ground truth. Our main index averages the four representativeness drivers (embedding spread, effective dimension, irrigation heterogeneity, and sub-area count), each entered with a positive sign; a one-feature version based only on the AEF embedding spread (Figure 7b) performs nearly as well. *(iii) Validation.* Farm-level ground truth enters only to verify that the index is meaningful: against the held-out INEGI census, the score tracks realized within-area skill, the per-unit rank correlation, and the sign of within-area skill (the correlations and AUC reported above). It never enters as an input to the score itself. *(iv) Use.* Apply the index to new units from satellite and survey features alone: those it flags as positive-skill are where farm-level predictions can be reported, while the rest are trusted only for between-area levels (ranking aggregation units). Because every feature is computable from satellite embeddings and public aggregate survey data, the index can be computed for regions and crops where no farm-level data exists.

A.3 Aggregation-constrained neural network details

The Agg-NN predicts a bounded residual adjustment on top of a random-forest baseline trained on mean AEF embeddings. Each ADC is described by 67 features: the 64 mean AEF embedding dimensions, the ADC's irrigation share, its log land area, and its agricultural-land share. A residual multilayer perceptron (input projection to 128 hidden units, two residual blocks with dropout 0.2, and a linear head) maps these features to an adjustment bounded via a tanh activation scaled to ± 5 t/ha; the final ADC prediction is the random-forest baseline plus this adjustment, clipped to $[0, 25]$ t/ha. During training, ADC predictions are aggregated to the municipality level by an agricultural-land-area weighted mean, and the loss is the mean squared error between the aggregated prediction and the DGSIAP municipal yield, plus a penalty $0.01 \cdot \overline{\text{residual}^2}$ that discourages gratuitous sub-municipal adjustments; ADC-level yields never enter the objective. The network is trained with Adam (learning rate 10^{-4}, weight decay 10^{-4}, batch size of 128 municipality-years, gradient-norm clipping at 5), a plateau-based learning-rate schedule, and early stopping on validation loss over at most 300 epochs; municipalities with more than 200 ADCs are subsampled to 200 per batch. Results are averaged over five random seeds. This design parallels aggregation-supervised disaggregation approaches in population mapping (Metzger et al., 2022, 2024).

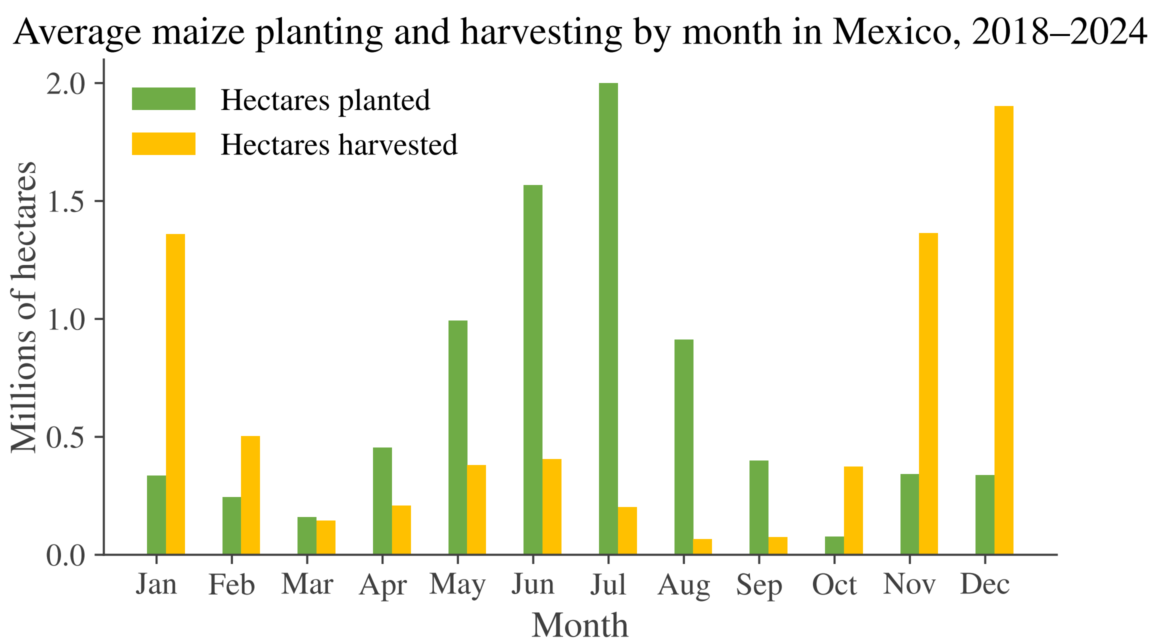
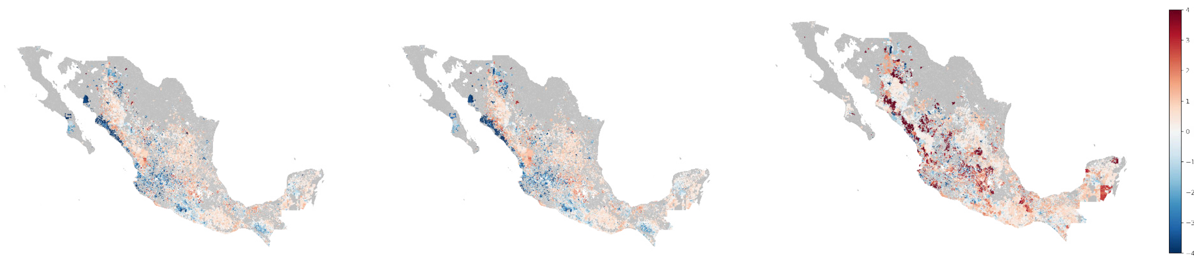


Figure A.1: Average monthly maize harvesting and planting across all municipalities in Mexico.

A.4 Common-sample comparison

The headline accuracy metrics in Section 4.2 are computed on each model's full set of available ADCs, which differs across models because of coverage: in the spring-summer season the NDVI baseline covers 68,604 ADCs, the AEF mean and Agg-NN models 70,195, and the AEF Hist and AEF Hist Ensemble models the fewest (58,022). To verify that the cross-model comparison is not an artifact of differing sample composition, Tables A.5 and A.6 re-evaluate every model on a single common set of ADCs—the intersection of ADCs for which all Landsat- and AEF-derived models produce a prediction (and a DGSIAP municipal yield exists). On this common sample the ranking is unchanged: the AEF distributional models remain the strongest, with the ensemble retaining

Figure A.2: ADC-level maps of maize yield prediction error (predicted − census, 2022, t/ha, shared ±4 scale; the colorbar appears in the right panel). Left: AEF Hist Ensemble, raw predictions (uncorrected and unshrunk). Center: the deployed AEF Hist Ensemble Shrink predictions. Right: DGSIAP municipal-average benchmark.



(a) AEF Hist Ens., raw (b) AEF Hist Ens. Shrink (c) DGSIAP mun. yield

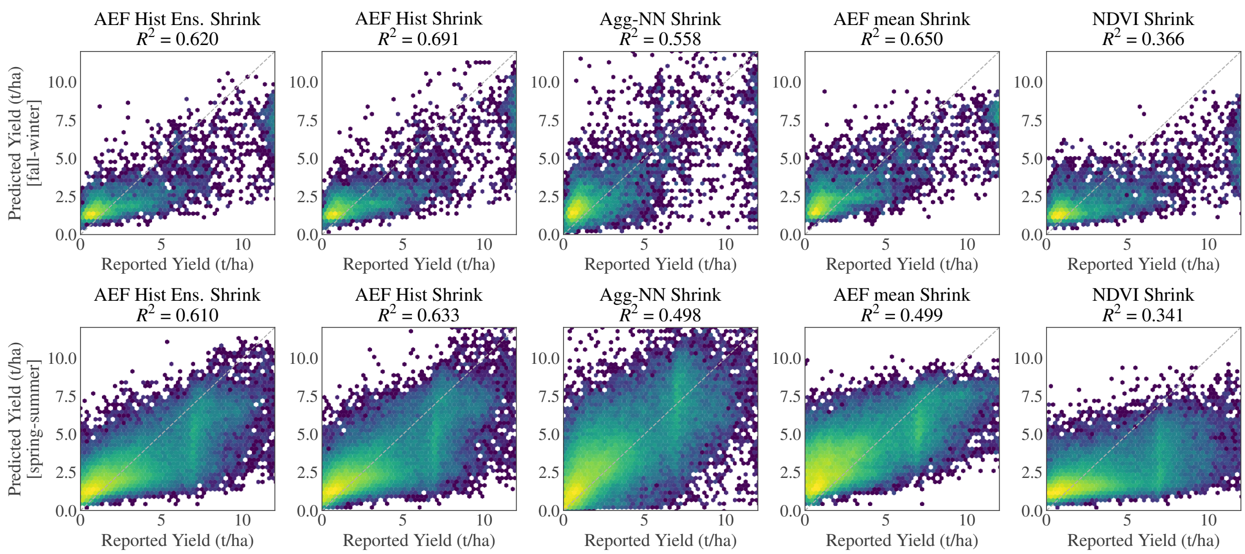


Figure A.3: Predicted vs. reported maize yield at the ADC level by season (2022), deployed Shrink specification of each model. Top row: fall-winter (O-I) season. Bottom row: spring-summer (P-V) season.

the highest within-municipality R^2, and both continue to outperform the NDVI baseline and the DGSIAP municipal-average benchmark.

The sample sizes themselves have a transparent accounting. The 2022 census reports a spring-summer maize yield for 91,441 ADCs (95,264 for the combined season). The DGSIAP municipal-average benchmark requires only that an ADC's municipality report a season-matched DGSIAP yield (which 90,962 spring-summer ADCs do; 94,898 combined); the benchmark row in the tables is nonetheless scored on the ensemble-eligible sample, like every model row (57,731 spring-summer; 59,417 combined). Every model row is then evaluated on the ensemble-eligible sample: ADCs with at least one cropland pixel under the ESA WorldCover mask, for which both of the ensemble's constituent feature sets exist—58,022 of the spring-summer census ADCs (59,641 combined). The excluded ADCs (12,173 spring-summer) have no cropland pixels, so the masked

Table A.1: Accuracy metrics for maize yield predictions vs. INEGI 2022 census, ADC level — Combined season. Corrected rows use ex-ante agricultural-land weights for the municipal anchor, and the anchor is the DGSIAP municipal maize yield for all growing seasons, matching the census target scored here. All rows are evaluated on the ADCs for which the AEF Hist Ensemble is defined (at least one cropland pixel; see the sample accounting in the appendix). RMSE in t/ha. The Oracle (ADC-trained) row is a non-deployable upper bound that trains HistGradientBoosting directly on ADC-level census labels (5-fold GroupKFold over municipalities); it bounds how much yield signal the embeddings carry.

Model	N	R^2	Between R^2	Within R^2	RMSE
Landsat-derived features					
NDVI Raw	57,638	0.304	0.518	−0.039	2.531
NDVI Corr.	57,421	0.531	0.632	−0.034	2.079
NDVI Shrink	57,638	0.334	0.518	0.076	2.477
AEF-derived features					
AEF mean Raw	59,641	0.487	0.634	0.024	2.154
AEF mean Corr.	59,417	0.516	0.591	0.037	2.093
AEF mean Shrink	59,641	0.511	0.634	0.117	2.103
Agg-NN Raw	59,641	0.420	0.683	−0.413	2.290
Agg-NN Corr.	59,417	0.450	0.657	−0.406	2.232
Agg-NN Shrink	59,641	0.503	0.683	−0.089	2.120
AEF Hist Raw	59,641	0.604	0.753	0.079	1.893
AEF Hist Corr.	59,417	0.545	0.605	0.081	2.031
AEF Hist Shrink	59,641	0.632	0.753	0.188	1.824
AEF Hist Ens. Raw	59,641	0.592	0.746	0.169	1.921
AEF Hist Ens. Corr.	59,417	0.575	0.620	0.170	1.963
AEF Hist Ens. Shrink	59,641	0.607	0.746	0.225	1.886
Benchmark					
DGSIAP Mun. Avg.	59,417	0.490	0.572	0.000	2.150
Oracle (ADC-trained)	59,395	0.765	0.875	0.290	1.459

features carry no information about them; the unmasked mean-embedding models can nominally score them, but with no skill (AEF mean R^2 of -0.07 on this group), and the land is marginal (mean census yield of 1.28 t/ha, against 3.00 t/ha on the retained sample). The NDVI rows are slightly smaller (56,101 spring-summer; 57,638 combined) because the Landsat extraction covers marginally fewer of these ADCs. Corrected rows lose roughly 300 further ADCs whose municipalities lack a season-matched DGSIAP anchor. The common-sample tables below additionally intersect all models onto one shared ADC set and confirm that none of these differences drives the model ranking.

Table A.2: Accuracy metrics for maize yield predictions vs. INEGI 2022 census, ADC level — fall-winter (O-I) season. Corrected rows use ex-ante agricultural-land weights for the municipal anchor, and the anchor is the DGSIAP municipal maize yield for the fall-winter season only, matching the census target scored here. All rows are evaluated on the ADCs for which the AEF Hist Ensemble is defined (at least one cropland pixel; see the sample accounting in the appendix). RMSE in t/ha.

Model	N	R^2	Between R^2	Within R^2	RMSE
Landsat-derived features					
NDVI Raw	7,275	0.355	0.527	−1.161	3.645
NDVI Corr.	6,258	0.683	−0.664	−1.374	2.628
NDVI Shrink	7,275	0.366	0.509	−0.547	3.612
AEF-derived features					
AEF mean Raw	7,686	0.644	0.587	−0.453	2.678
AEF mean Corr.	6,647	0.725	−0.707	−0.537	2.421
AEF mean Shrink	7,686	0.650	0.592	−0.174	2.655
Agg-NN Raw	7,686	0.499	0.482	−2.535	3.175
Agg-NN Corr.	6,647	0.609	−0.758	−2.928	2.889
Agg-NN Shrink	7,686	0.558	0.522	−1.212	2.983
AEF Hist Raw	7,686	0.687	0.673	−0.635	2.510
AEF Hist Corr.	6,647	0.715	−0.651	−0.740	2.465
AEF Hist Shrink	7,686	0.691	0.666	−0.244	2.493
AEF Hist Ens. Raw	7,686	0.623	0.661	−0.337	2.754
AEF Hist Ens. Corr.	6,647	0.735	−0.643	−0.421	2.378
AEF Hist Ens. Shrink	7,686	0.620	0.646	−0.096	2.766
Benchmark					
DGSIAP Mun. Avg.	6,647	0.758	−0.605	0.000	2.271

Table A.3: ADC-level yield prediction accuracy for non-maize crops vs. INEGI 2022 census. RMSE in t/ha.

Crop	Model	N	R^2	Between R^2	Within R^2	RMSE
Sorghum	AEF mean	7,095	0.294	0.344	0.007	1.931
Sorghum	AEF mean Corr.	6,558	0.325	0.384	0.037	1.863
Sorghum	AEF Hist	6,984	0.433	0.531	0.097	1.731
Sorghum	AEF Hist Corr.	6,454	0.373	0.419	0.111	1.795
Sorghum	AEF Hist Ens.	6,984	0.461	0.563	0.143	1.687
Sorghum	AEF Hist Ens. Corr.	6,454	0.392	0.420	0.157	1.768
Sorghum	AEF Hist Ens. Shrink	6,984	0.460	0.563	0.140	1.689
Sorghum	DGSIAP Mun. Avg.	11,267	0.372	0.374	−0.000	1.661
Sugarcane	AEF mean	5,759	0.256	0.172	−0.327	23.023
Sugarcane	AEF mean Corr.	5,372	0.196	0.393	−0.313	22.311
Sugarcane	AEF Hist	5,064	0.178	0.346	−0.603	20.825
Sugarcane	AEF Hist Corr.	4,792	−0.127	0.341	−0.595	22.332
Sugarcane	AEF Hist Ens.	5,064	0.278	0.373	−0.360	19.523
Sugarcane	AEF Hist Ens. Corr.	4,792	−0.025	0.353	−0.355	21.304
Sugarcane	AEF Hist Ens. Shrink	5,064	0.332	0.373	−0.150	18.777
Sugarcane	DGSIAP Mun. Avg.	8,893	0.164	0.341	0.000	20.814
Wheat	AEF mean	3,368	−0.329	−1.487	−1.778	2.827
Wheat	AEF mean Corr.	3,200	0.351	−0.254	−1.564	1.957
Wheat	AEF Hist	3,344	0.468	0.097	−0.572	1.789
Wheat	AEF Hist Corr.	3,178	0.527	−0.090	−0.540	1.672
Wheat	AEF Hist Ens.	3,344	0.417	−0.074	−0.561	1.872
Wheat	AEF Hist Ens. Corr.	3,178	0.530	−0.096	−0.541	1.666
Wheat	AEF Hist Ens. Shrink	3,344	0.448	−0.074	−0.272	1.821
Wheat	DGSIAP Mun. Avg.	5,912	0.624	−0.176	0.000	1.466
Avocados	AEF mean	6,601	−0.972	−1.174	−0.362	6.674
Avocados	AEF mean Corr.	4,056	−0.634	−0.599	−0.309	5.632
Avocados	AEF Hist	5,144	−1.142	−1.004	−0.434	7.008
Avocados	AEF Hist Corr.	3,352	−0.640	−0.530	−0.360	5.500
Avocados	AEF Hist Ens.	5,144	−0.633	−0.695	−0.201	6.118
Avocados	AEF Hist Ens. Corr.	3,352	−0.554	−0.519	−0.182	5.355
Avocados	AEF Hist Ens. Shrink	5,144	−0.599	−0.695	−0.096	6.055
Avocados	DGSIAP Mun. Avg.	5,753	−0.525	−0.605	0.000	5.430

Table A.4: Municipality-level yield prediction results for non-maize crops, 2022. ADC-level AEF Hist Ens. Shrink predictions are aggregated to the municipality level weighting each ADC by its *ex-ante* agricultural-land area, as in Table 6; no census information enters the aggregation. RMSE in t/ha.

Panel A: vs. INEGI Census				
Crop	Model	N	R^2	RMSE
Sorghum	AEF Hist Ens. Shrink (agg.)	508	0.565	1.332
Sorghum	DGSIAP Mun. Avg.	373	0.411	1.513
Sugarcane	AEF Hist Ens. Shrink (agg.)	409	0.399	22.928
Sugarcane	DGSIAP Mun. Avg.	306	0.391	21.129
Wheat	AEF Hist Ens. Shrink (agg.)	380	−0.070	2.241
Wheat	DGSIAP Mun. Avg.	323	−0.167	2.404
Avocados	AEF Hist Ens. Shrink (agg.)	867	−0.686	5.805
Avocados	DGSIAP Mun. Avg.	442	−0.356	4.891
Panel B: vs. DGSIAP Municipal Estimates				
Crop	Model	N	R^2	RMSE
Sorghum	AEF Hist Ens. Shrink (agg.)	367	0.642	1.130
Sorghum	INEGI Census (agg.)	373	0.351	1.513
Sugarcane	AEF Hist Ens. Shrink (agg.)	275	0.747	12.295
Sugarcane	INEGI Census (agg.)	306	0.320	21.129
Wheat	AEF Hist Ens. Shrink (agg.)	313	0.702	1.560
Wheat	INEGI Census (agg.)	323	0.358	2.404
Avocados	AEF Hist Ens. Shrink (agg.)	416	0.631	2.728
Avocados	INEGI Census (agg.)	442	−0.215	4.891

Table A.5: Common-sample accuracy for maize yield predictions vs. INEGI 2022 census, ADC level — Combined season. Every model is evaluated on the *same* 57,421 ADCs (the intersection of ADCs for which all Landsat- and AEF-derived models produce a prediction and a DGSIAP municipal yield for all growing seasons exists), so cross-model differences are not driven by sample composition. RMSE in t/ha.

Model	N	R^2	Between R^2	Within R^2	RMSE
Landsat-derived features					
NDVI Raw	57,421	0.304	0.523	−0.038	2.532
NDVI Shrink	57,421	0.334	0.523	0.077	2.478
AEF-derived features					
AEF mean Raw	57,421	0.487	0.643	0.025	2.174
AEF mean Shrink	57,421	0.511	0.643	0.119	2.123
Agg-NN Raw	57,421	0.418	0.689	−0.413	2.315
Agg-NN Shrink	57,421	0.502	0.689	−0.089	2.143
AEF Hist Raw	57,421	0.605	0.761	0.081	1.908
AEF Hist Shrink	57,421	0.633	0.761	0.190	1.840
AEF Hist Ens. Raw	57,421	0.592	0.751	0.171	1.938
AEF Hist Ens. Shrink	57,421	0.607	0.751	0.227	1.904
Benchmark					
DGSIAP Mun. Avg.	57,421	0.491	0.574	0.000	2.166

Table A.6: Common-sample accuracy for maize yield predictions vs. INEGI 2022 census, ADC level — spring-summer (P-V) season. Every model is evaluated on the *same* 55,817 ADCs (the intersection of ADCs for which all Landsat- and AEF-derived models produce a prediction and a DGSIAP municipal yield for the spring-summer season only exists), so cross-model differences are not driven by sample composition. RMSE in t/ha.

Model	N	R^2	Between R^2	Within R^2	RMSE
Landsat-derived features					
NDVI Raw	55,817	0.314	0.527	−0.024	2.444
NDVI Shrink	55,817	0.344	0.527	0.086	2.390
AEF-derived features					
AEF mean Raw	55,817	0.475	0.633	0.033	2.139
AEF mean Shrink	55,817	0.499	0.633	0.124	2.088
Agg-NN Raw	55,817	0.411	0.680	−0.390	2.265
Agg-NN Shrink	55,817	0.497	0.680	−0.076	2.093
AEF Hist Raw	55,817	0.607	0.761	0.100	1.850
AEF Hist Shrink	55,817	0.635	0.761	0.202	1.783
AEF Hist Ens. Raw	55,817	0.598	0.753	0.183	1.872
AEF Hist Ens. Shrink	55,817	0.612	0.753	0.235	1.838
Benchmark					
DGSIAP Mun. Avg.	55,817	0.408	0.557	0.000	2.271

CRediT authorship contribution statement

Joel Ferguson: Conceptualization, Methodology, Software, Investigation, Writing – original draft, review & editing.
Kangogo Kibarar: Conceptualization, Investigation, Validation, Visualization, Writing – review & editing.
James E. Sayre: Conceptualization, Methodology, Data acquisition and curation, Formal analysis, Software, Validation, Visualization, Writing – original draft, review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Acknowledgements

We greatly appreciate the invaluable assistance and hospitality of Natalia Volkow and the INEGI microdata laboratory staff, especially Lidia Hernández Hernández, who provided us the microdata used in this project. This project is based on information processed under INEGI microdata laboratory project LM2304. The conclusions and opinions in this project are our own and do not form part of the official statistics of INEGI nor that of the National System of Statistical and Geographical Information (SNIEG). We thank audiences at AGU, TWEEDS, UC Berkeley, UC Davis, and the World Bank for their feedback on our work.

Declaration of generative AI and AI-assisted technologies in the manuscript preparation process

During the preparation of this work the authors used Claude (Anthropic) in order to assist with writing and refactoring the analysis code, generating figures from the underlying data with reproducible code, and copy-editing the text. After using this tool, the authors reviewed and edited the code and content as needed and take full responsibility for the content of the published article.

References

Ahmad, A., Liew, A.X., Venturini, F., Kalogeras, A., Candiani, A., Di Benedetto, G., Ajibola, S., Cartujo, P., Romero, P., Lykoudi, A., et al., 2024. AI can empower agriculture for global food security: challenges and prospects in developing nations. Frontiers in Artificial Intelligence 7, 1328530.

Brown, C.F., Kazmierski, M.R., Pasquarella, V.J., Rucklidge, W.J., Zhang, C., Shelhamer, E., Lahera, E., Wiles, O., Ilyushchenko, S., Zhang, L.L., Alj, S., Schechter, E., Askay, S., Guinan, O., Moore, R., Boukouvalas, A., Kohli, P., 2025. AlphaEarth Foundations: An embedding field model for accurate and efficient global mapping from sparse label data. arXiv preprint arXiv:2507.22291.

Burke, M., Driscoll, A., Lobell, D.B., Ermon, S., 2021. Using satellite imagery to understand and promote sustainable development. Science 371, eabe8628.

Carletto, C., Jolliffe, D., Banerjee, R., 2013. The emperor has no data! agricultural statistics in sub-Saharan Africa. World Bank Working Paper 565.

Cong, Y., Khanna, S., Meng, C., Liu, P., Rozi, E., He, Y., Burke, M., Lobell, D., Ermon, S., 2022. SatMAE: Pre-training transformers for temporal and multi-spectral satellite imagery, in: Advances in Neural Information Processing Systems, pp. 197–211.

Dado, W.T., Deines, J.M., Patel, R., Liang, S.Z., Lobell, D.B., 2020. High-resolution soybean yield mapping across the US Midwest using subfield harvester data. Remote Sensing 12, 3471. doi:10.3390/rs12213471.

Deines, J.M., Patel, R., Liang, S.Z., Dado, W., Lobell, D.B., 2021. A million kernels of truth: Insights into scalable satellite maize yield mapping and yield gap analysis from an extensive ground dataset in the US Corn Belt. Remote Sensing of Environment 253, 112174. doi:10.1016/j.rse.2020.112174.

Feng, Z., Atzberger, C., Jaffer, S., Knezevic, J., Sormunen, S., Young, R., Lisaius, M.C., Immitzer, M., Jackson, T., Ball, J., Coomes, D.A., Madhavapeddy, A., Blake, A., Keshav, S., 2026. TESSERA: Temporal embeddings of surface spectra for Earth representation and analysis. arXiv preprint arXiv:2506.20380 (accepted at CVPR 2026).

Ferreira, L.B., Martins, V.S., Aires, U.R.V., Wijewardane, N., Zhang, X., Samiappan, S., 2025. FieldSeg: A scalable agricultural field extraction framework based on the Segment Anything Model and 10-m Sentinel-2 imagery. Computers and Electronics in Agriculture 232, 110086. doi:10.1016/j.compag.2025.110086.

Friedman, J.H., 2001. Greedy function approximation: A gradient boosting machine. The Annals of Statistics 29, 1189–1232. doi:10.1214/aos/1013203451.

Gao, F., Anderson, M.C., Zhang, X., Yang, Z., Alfieri, J.G., Kustas, W.P., Mueller, R., Johnson, D.M., Prueger, J.H., 2017. Toward mapping crop progress at field scales through fusion of Landsat and MODIS imagery. Remote Sensing of Environment 188, 9–25. doi:10.1016/j.rse.2016.11.004.

Gao, F., Zhang, X., 2021. Mapping crop phenology in near real-time using satellite remote sensing: Challenges and opportunities. Journal of Remote Sensing 2021, 1–14. doi:10.34133/2021/8379391.

Gardeazábal-Monsalve, A., Cabello Cortés, J.A., Fonteyne, S., Jiménez Gomez, B., Leal Gonzalez, A.J., Núñez Jiménez, D., Odjo, S., Pacheco Rodriguez, G.M., Ramírez Ortega, M.d.L., Van Loon, J., et al., 2025. A decade of on-farm data about improved cereal and legume cropping in Mexico. Scientific Data 12, 1873.

Hultgren, A., Carleton, T., Delgado, M., Gergel, D.R., Greenstone, M., Houser, T., Hsiang, S., Jina, A., Kopp, R.E., Malevich, S.B., et al., 2025. Impacts of climate change on global agriculture accounting for adaptation. Nature 642, 644–652.

Joshi, A., Pradhan, B., Gite, S., Chakraborty, S., 2023. Remote-sensing data and deep-learning techniques in crop mapping and yield prediction: A systematic review. Remote Sensing 15, 2014.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., Liu, T.Y., 2017. LightGBM: A highly efficient gradient boosting decision tree, in: Advances in Neural Information Processing Systems, pp. 3146–3154.

Lobell, D.B., Burke, M.B., Tebaldi, C., Mastrandrea, M.D., Falcon, W.P., Naylor, R.L., 2008. Prioritizing climate change adaptation needs for food security in 2030. Science 319, 607–610.

Ma, Y., Liang, S.Z., Myers, D.B., Swatantran, A., Lobell, D.B., 2024. Subfield-level crop yield mapping without ground truth data: A scale transfer framework. Remote Sensing of Environment 315, 114427.

Ma, Y., Shen, Y., Swatantran, A., Lobell, D.B., 2026. Harvesting AlphaEarth: Benchmarking the geospatial foundation model for agricultural downstream tasks. International Journal of Applied Earth Observation and Geoinformation 149, 105258.

Mai, G., Cundy, C., Choi, K., Hu, Y., Lao, N., Ermon, S., 2022. Towards a foundation model for geospatial artificial intelligence (vision paper), in: Proceedings of the 30th International Conference on Advances in Geographic Information Systems, pp. 1–4.

Mai, G., Huang, W., Sun, J., Song, S., Mishra, D., Liu, N., Gao, S., Liu, T., Cong, G., Hu, Y., et al., 2023. On the opportunities and challenges of foundation models for geospatial artificial intelligence. arXiv preprint arXiv:2304.06798.

Metzger, N., Daudt, R.C., Tuia, D., Schindler, K., 2024. High-resolution population maps derived from Sentinel-1 and Sentinel-2. Remote Sensing of Environment 314, 114383. doi:10.1016/j.rse.2024.114383.

Metzger, N., Vargas-Muñoz, J.E., Daudt, R.C., Kellenberger, B., Whelan, T.T.T., Ofli, F., Imran, M., Schindler, K., Tuia, D., 2022. Fine-grained population mapping from coarse census counts and open geodata. Scientific Reports 12, 20085. doi:10.1038/s41598-022-24495-w.

Muruganantham, P., Wibowo, S., Grandhi, S., Samrat, N.H., Islam, N., 2022. A systematic literature review on crop yield prediction with deep learning and remote sensing. Remote Sensing 14, 1990.

Nejadshamsi, S., Zhang, Y., Porth, B., Zaki, S., Porth, L., Khoshdel, V., 2026. FARM: Crop yield prediction via regression on Prithvi's encoder for satellite sensing. AgriEngineering 8, 2. doi:10.3390/agriengineering8010002.

Paudel, D., Marcos, D., de Wit, A., Boogaard, H., Athanasiadis, I.N., 2023. A weakly supervised framework for high-resolution crop yield forecasts. Environmental Research Letters 18, 094062. doi:10.1088/1748-9326/acf50e.

Reed, C.J., Gupta, R., Li, S., Brockman, S., Funk, C., Clipp, B., Keutzer, K., Candido, S., Uyttendaele, M., Darrell, T., 2023. Scale-MAE: A scale-aware masked autoencoder for multiscale geospatial representation learning, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 4088–4099.

Szwarcman, D., Roy, S., Fraccaro, P., Gíslason, T.E., Blumenstiel, B., Ghosal, R., Oliveira, P.H.d., Almeida, J.L.d.S., Sedona, R., Kang, Y., Chakraborty, S., Wang, S., Kumar, A., Truong, M., Godwin, D., Lee, H., Hsu, C.Y., Asanjan, A.A., Mujeci, B., Keenan, T., Arevalo, P., Li, W., Alemohammad, H., Olofsson, P., Hain, C., Kennedy, R., Zadrozny, B., Watson, C., Maskey, M., Ramachandran, R., Moreno, J.B., 2024. Prithvi-EO-2.0: A versatile multi-temporal foundation model for Earth Observation applications. arXiv preprint arXiv:2412.02732.

Tripathy, P., Baylis, K., Wu, K., Watson, J., Jiang, R., 2026. Zero-shot inference strategies for smallholder (<0.1 ha) agriculture field delineation with the Segment Anything foundation model. Science of Remote Sensing 13, 100425. doi:10.1016/j.srs.2026.100425.

Wu, K., Zhang, Y., Ru, L., Dang, B., Lao, J., Yu, L., Luo, J., Zhu, Z., Sun, Y., Zhang, J., Zhu, Q., Wang, J., Yang, M., Chen, J., Zhang, Y., Li, Y., 2025. A semantic-enhanced multi-modal remote sensing foundation model for Earth observation. Nature Machine Intelligence doi:10.1038/s42256-025-01078-8.

Xiong, Z., Wang, Y., Zhang, F., Stewart, A.J., Hanna, J., Borth, D., Papoutsis, I., Le Saux, B., Camps-Valls, G., Zhu, X.X., 2024. Neural plasticity-inspired multimodal foundation model for Earth observation. arXiv preprint arXiv:2403.15356.

Yan, J., Gu, X., Chen, Y., 2025. CropSTS: A remote sensing foundation model for cropland classification with decoupled spatiotemporal attention. Remote Sensing 17, 2481. doi:10.3390/rs17142481.

Yang, L., Chen, N., Yue, J., Liu, Y., Ma, J., Ghamisi, P., Plaza, A., Fang, L., 2025. Survey of multimodal geospatial foundation models: Techniques, applications, and challenges. arXiv preprint arXiv:2510.22964.

Yang, Z., Diao, C., Gao, F., 2023. Towards scalable within-season crop mapping with phenology normalization and deep learning. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 16, 1390–1402. doi:10.1109/JSTARS.2023.3237500.

You, J., Li, X., Low, M., Lobell, D., Ermon, S., 2017. Deep Gaussian process for crop yield prediction based on remote sensing data, in: Proceedings of the AAAI Conference on Artificial Intelligence, pp. 4559–4565.