Physics Guided Neural Networks for Time-Aware Fairness: An Application in Crop Yield Prediction

Erhu He[1]*, Yiqun Xie[2]*, Licheng Liu[3], Weiye Chen[2], Zhenong Jin[3], Xiaowei Jia[1]

[1]Department of Computer Science, University of Pittsburgh
[2]Department of Geographical Sciences, University of Maryland
[3]Department of Bioproducts and Biosystems Engineering, University of Minnesota
erh108@pitt.edu, xie@umd.edu, lichengl@umn.edu, weiyec@umd.edu, jinzn@umn.edu, xiaowei@pitt.edu

Abstract

This paper proposes a physics-guided neural network model to predict crop yield and maintain the fairness over space. Failures to preserve the spatial fairness in predicted maps of crop yields can result in biased policies and intervention strategies in the distribution of assistance or subsidies in supporting individuals at risk. Existing methods for fairness enforcement are not designed for capturing the complex physical processes that underlie the crop growing process, and thus are unable to produce good predictions over large regions under different weather conditions and soil properties. More importantly, the fairness is often degraded when existing methods are applied to different years due to the change of weather conditions and farming practices. To address these issues, we propose a physics-guided neural network model, which leverages the physical knowledge from existing physics-based models to guide the extraction of representative physical information and discover the temporal data shift across years. In particular, we use a reweighting strategy to discover the relationship between training years and testing years using the physics-aware representation. Then the physics-guided neural network will be refined via a bi-level optimization process based on the reweighted fairness objective. The proposed method has been evaluated using real county-level crop yield data and simulated data produced by a physics-based model. The results demonstrate that this method can significantly improve the predictive performance and preserve the spatial fairness when generalized to different years.

Introduction

The global food system has been threatened by many rising challenges, such as the population explosion, sub-optimal or even destructive farming practices, climate change, and loss of productive land due to urbanization (Ortiz et al. 2008; d'Amour et al. 2017; Bebber, Holmes, and Gurr 2014; Jia et al. 2019). Given the limited land and water resources available for crop production, ensuring food security requires effective use of existing farmland by increasing crop productivity through sustainable farming practices. The importance and urgency of the problem have also led to major national and international efforts to monitor crops at large scales, including G20's GEOGLAM (Singh Parihar et al. 2012) and NASA Harvest (Whitcraft, Becker-Reshef, and Justice 2020). Importantly, resulting products are also used to inform critical actions (e.g., distribution of subsidies (Bock, Kirkendall et al. 2018; Bailey and Boryan 2010; Boryan et al. 2011)) to mitigate natural disturbance-incurred food shortage, which is necessary for continued sustainability and stability.

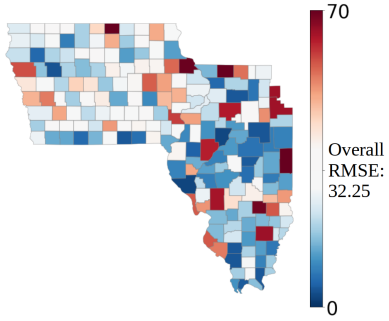


Figure 1: Example of spatial bias on crop yield prediction.

Physics-based crop models have been developed and widely used to simulate different components in the crop growing process (Grant et al. 2010; Zhou et al. 2021; Jones et al. 2003; Srinivasan, Zhang, and Arnold 2010). Even though these models are based on known physical laws that govern relationships between input and output variables (e.g., mass and energy conservation laws), most physics-based models are necessarily approximations of reality due to incomplete knowledge of certain processes or omission of processes to maintain computational efficiency. Moreover, running these models is often extremely time intensive due to the needs to solve hundreds of ordinary and partial differential equations that are used to describe the complex interactions among physical processes (Beven 2006). For example, the Ecosys model (Zhou et al. 2021) used in this paper takes about 4.15 days for training for 199 counties over the years 2000-2020. Although machine learning (ML) models (Fan et al. 2022) can significantly speed up crop yield predictions, existing products are largely subject to unfairness across locations due to the data variability across space and the ignorance of fairness during the training process. For example, Fig. 1 shows the spatial distribution of the RMSE

*These authors contributed equally.

values obtained by a deep learning model for real-world crop yield prediction in 199 counties in the states of Illinois and Iowa. As we can see, the differences between overall and local results show that prediction accuracy in some regions can be easily compromised to pursue better results at other places. The unequal prediction accuracy may cause concerns about inequality in many socioeconomic decision makings. For example, higher risks associated with larger prediction uncertainty can reduce farmers' benefits through crop insurance or subsidy-based programmes (Benami et al. 2021).

Various methods have been developed to enforce fairness in machine learning models, and they can be broadly classified into several categories, such as bi-level learning (Xie et al. 2022; He et al. 2022), regularization (Zafar et al. 2017; Yan and Howe 2019; Kamishima, Akaho, and Sakuma 2011; Serna et al. 2020), sensitive category de-correlation (Sweeney and Najafian 2020; Zhang and Davidson 2021; Alasadi, Al Hilli, and Singh 2019), data collection/filtering strategies (Jo and Gebru 2020; Yang et al. 2020; Steed and Caliskan 2021), and more (e.g., a recent survey (Mehrabi et al. 2021)). However, these fairness-preserving methods are faced with two major challenges when used in crop yield prediction. First, they are not designed to model underlying physical processes, the complexity of which can vary across space due to the variation of weather conditions and soil properties. For example, crop yield is very sensitive to soil moisture, which is highly variable over the landscape due to changes in precipitation, local topography and water table depth. And water table interacts with crop roots to determine not only crop water uptake but also nutrient supplies that are essential for crop production. Hence, standard ML models may not fully capture key physical variables and processes, and thus perform differently across space. Existing heterogeneity-aware learning methods (Xie et al. 2021; Gupta et al. 2021) can adapt over space but do not account for fairness. Another major issue is the temporal data distribution shift across years due to changes in weather conditions and farming practices. As a result, a fairness-enforced model learned from training years may fail to preserve the fairness in target testing years.

To address these issues, we develop a physics-guided attention network (PG-AN), which leverages the physical knowledge from existing physics-based models to guide the extraction of representative physical information and discover the distribution shift across years. The PG-AN model introduces threefold benefits. First, as inspired by prior work (Jia et al. 2021b,c), the predictive performance can be improved if key physical variables involved in the crop growth can be extracted from high-dimensional raw data. Second, the representation learned by the PG-AN model can facilitate addressing the temporal data shift. For example, rather than directly operating on the raw data, methods for addressing domain shift (e.g., the reweighting strategy (Freedman and Berk 2008)) can better estimate the gap and similarity between training and testing samples based on the most relevant information for crop yield. Third, the awareness of physical knowledge can contribute to the overall spatial fairness because the model has a higher chance at learning a uniform mapping from the physics-aware representation to the crop yield, which performs well over all the spatial regions.

In particular, we use the PG-AN model to embed key physical variables involved in the carbon cycle of the crop growing process with an additional constraint of mass conservation. Then we use the extracted physics-aware embeddings to reweight training samples in refining the PG-AN model so as to reduce the distribution gap with the target years. Moreover, the obtained sample weights are used to modify the fairness objective, which is enforced through a bi-level optimization in the refinement process. The model with the reweighted fairness objective stands a higher chance at preserving the fairness in the target years.

We evaluate the proposed method using real corn yield data over a 21-year period in Iowa and Illinois, two leading states of corn production in the United States. The results demonstrate the fairness improvement achieved by incorporating the physical information and the bi-level fairness-driven refinement. Moreover, the integration of physics into the PG-AN model can significantly improve the overall predictive performance and maintain the consistency of known physical relationships. The experiments are also conducted using different spatial partitionings and different target years, which confirms the robustness and stability.

Problem Formulation and Preliminaries

The objective is to predict the county-level yield for corn in target years. For each county i, we are provided with input features within each year t, as $\mathbf{X}_{i,t} = \{\mathbf{x}^1_{i,t}, \mathbf{x}^2_{i,t}, ..., \mathbf{x}^D_{i,t}\}$, which are available at daily scales, i.e., $D = 365$ in a non-leap year. The daily features $\mathbf{x}^d_{i,t}$ include weather drivers (e.g., precipitation, solar radiation), and soil and crop properties. The feature values are obtained as the average of the variable values from a set of randomly sampled farm locations in each county. More details can be found in the Experiment Section. Additionally, we have the access to the crop yield labels $\mathbf{Y} = \{y_{i,t}\}$ from agricultural surveys in the training years $\mathcal{R}$. In the target testing years $\mathcal{T}$, we only have the input features but do not have the crop yield labels in the training process.

In addition to the real crop yield dataset, we also run the physics-based Ecosys model (Zhou et al. 2021) to simulate crop yield. Here we use $\mathcal{S}$ to represent the set of locations and years (i, t) for which we have the simulated crop yield. Another benefit of the physics-based model is that it can also simulate some intermediate physical variables in the crop growing process, such as variables involved in carbon and nitrogen cycling. It is noteworthy that physics-based models are often biased as they are necessarily approximations of reality due to incomplete knowledge or excessive complexity in modeling underlying processes. Hence, the simulated data can only be used for weak supervision.

Attention-based crop yield predictive model: The predictive model $\mathcal{F}_{\mathbf{\Theta}}(\mathbf{x}_{i,t})$ used in this work is based on an LSTM-Attention network. Here $\mathbf{\Theta}$ represents all the parameters in this network. Specifically, we first use an LSTM network to extract hidden representations at every time step (i.e., each date in a year), as $\mathbf{h}^{d=1:D}_{i,t} = \text{LSTM}(\mathbf{x}^{d=1:D}_{i,t})$.

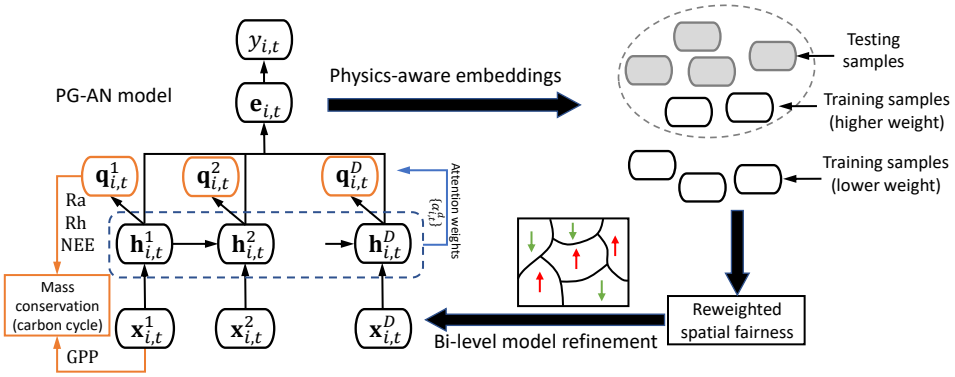


Figure 2: The overall flow of the proposed method. An LSTM-Attention network is used as a base model.

Then we create attention weights for each time step from its corresponding hidden representation via a linear transformation and a softmax function, as follows:

$$\alpha_{i,t}^{d} = \frac{\exp(\mathbf{w}_\alpha \cdot \mathbf{h}_{i,t}^{d} + b_\alpha)}{\sum_{d'} \exp(\mathbf{w}_\alpha \cdot \mathbf{h}_{i,t}^{d'} + b_\alpha)}, \tag{1}$$

where $\mathbf{w}_\alpha \in \mathbb{R}^{D_h}$ and $\mathbf{b}_\alpha \in \mathbb{R}^1$ are attention model parameters. Hereinafter we use $\{\mathbf{w}_*, \mathbf{W}_*, b_*, \mathbf{b}_*\}$ to represent model parameters.

The embedding for each county i in year t can be obtained by the weighted mean over all the time steps using the attention weights, as:

$$\mathbf{e}_{i,t} = \sum_{d} \alpha_{i,t}^{d} \mathbf{h}_{i,t}^{d}. \tag{2}$$

Finally, the model outputs the predicted yield value of the county i in year t as:

$$\mathcal{F}_{\mathbf{\Theta}}(\mathbf{x}_{i,t}) = \mathbf{w}_y \mathbf{e}_{i,t} + b_y. \tag{3}$$

The model can be trained by minimizing a mean squared error-based loss function, as follows:

$$\min_{\mathbf{\Theta}} \mathcal{L}_{sup} = \frac{\sum_{t\in\mathcal{R}} \sum_{i} (\mathcal{F}_{\mathbf{\Theta}}(\mathbf{x}_{i,t}) - y_{i,t})^2}{N|\mathcal{R}|}, \tag{4}$$

where N is the total number of counties.

Spatial fairness: Here we introduce the spatial fairness measure M_{fair}(Xie et al. 2022), which is defined on a spatial partitioning $\mathcal{P}$. The partitioning $\mathcal{P}$ splits an input space into multiple partitions, as $\mathcal{P} = \{p \,|\, p \in \mathcal{P}\}$. The objective of fairness over a spatial partitioning $\mathcal{P}$ is to guarantee that the model's performance is balanced across all the space partitions p that are contained in $\mathcal{P}$. First, we consider a metric $M_{\mathcal{F}}$ used to evaluate the performance of a model $\mathcal{F}$, e.g., root mean squared error (RMSE). Another important variable for the fairness definition is $E_{\mathcal{P}}$, which quantifies the mean model performance over all the partitions. $E_{\mathcal{P}}$ is implemented as the overall performance of a base model $\mathcal{F}_{\mathbf{\Theta}_0}$ over all the partitions, where parameters $\mathbf{\Theta}_0$ are trained without any consideration of spatial fairness. More formally, this can be represented as:

$$E_{\mathcal{P}} = M_{\mathcal{F}}(\mathcal{F}_{\mathbf{\Theta}_0}, \{\cup p \,|\, p \in \mathcal{P}\}) \tag{5}$$

Intuitively, if the model performance $M_{\mathcal{F}}(\mathcal{F}_{\mathbf{\Theta}}, p)$ on a specific partition p deviates significantly from the overall mean performance $E_{\mathcal{P}}$, the model $\mathcal{F}_{\mathbf{\Theta}}$ tends to be unfair across partitions. The fairness is formally defined as

$$M_{fair}(\mathcal{F}_{\mathbf{\Theta}}, M_{\mathcal{F}}, \mathcal{P}) = \sum_{p\in\mathcal{P}} \frac{d(M_{\mathcal{F}}(\mathcal{F}_{\mathbf{\Theta}}, p), E_{\mathcal{P}})}{|\mathcal{P}|} \tag{6}$$

where $d(\cdot,\cdot)$ is a distance measure, e.g., the absolute distance in our test.

Sample reweighting strategy: Sample reweighting strategy has been explored to reduce the gap of input space between the source and target domains (Bickel, Brückner, and Scheffer 2007; Freedman and Berk 2008). In our problem, a classifier $\mathcal{G}$ is trained to distinguish between source/training and target/testing years. The classifier $\mathcal{G}$ is implemented as a four-layer fully-connected network. Its output is in the range of [0,1], and is closer to 1 if it predicts the data to be more likely from the target years and otherwise is closer to 0. Then the weight of each sample (e.g., county i in each year t) is estimated as:

$$w_{i,t} = \frac{\mathcal{G}(\mathbf{e}_{i,t})}{1 - \mathcal{G}(\mathbf{e}_{i,t})} \tag{7}$$

After gathering the estimated sample weights, we normalize them to the range of [γ,1], where γ is a small value, e.g., 0.1 in our test, which is used to ensure that all the samples are involved in the training process. We represent the normalized weights as $\bar{w}$.

The obtained weights can then be used in the training loss function to alleviate the temporal data shift in the training process, as follows:

$$\min_{\mathbf{\Theta}} \mathcal{L}_{rew} = \frac{\sum_{i=1}^{N} \sum_{t=1}^{T} \bar{w}_{i,t} (\mathcal{F}_{\mathbf{\Theta}}(\mathbf{x}_{i,t}) - y_{i,t})^2}{N|\mathcal{R}|} \tag{8}$$

Proposed Method

The overall flow of the proposed method is shown in Fig. 2. We first introduce the proposed PG-AN model and the enhanced reweighting strategy using the PG-AN representation. Then we discuss the refinement process for the PG-AN

model to further enhance the spatial fairness with the awareness of temporal data shift. Our implementation is released[1].

Physics-guided Sample Reweighting

We first build the PG-AN model to embed key variables involved in the carbon cycle and improve the prediction of crop yield (Fig. 2). During the crop growing process, carbon is cycled through the atmosphere, crops and the soil. Carbon makes a major contribution to soil fertility and soil's capacity to retain water (Zhou et al. 2021). Carbon is absorbed by crops in the form of carbon dioxide, which contributes to the growth of crops. While the crops grow up, their produced roots and leaves also affect the soil carbon storage.

Although most variables in the carbon cycle are not observable, they can be simulated by existing physics-based models based on known physical theories. In this work, we use the physics-based Ecosys model (Zhou et al. 2021) to simulate three key variables in the carbon cycle, ecosystem autotrophic respiration (Ra), ecosystem heterotrophic respiration (Rh), and net ecosystem exchange (NEE). The entire carbon cycle can be captured by a mass conservation relation, as $-\text{NEE} = \text{GPP} - \text{Ra} - \text{Rh}$, where GPP represents the gross primary production, and can be estimated from remote sensing. The estimated GPP values are available over large regions and used as input to the predictive model.

Given the hidden representation $\mathbf{h}_{i,t}^d$ extracted by the LSTM-Attention model on each date d, we predict the physical variables Ra, Rh, and NEE using another transformation $\hat{\mathbf{q}}_{i,t}^d = f(\mathbf{h}_{i,t}^d)$, where $\hat{\mathbf{q}}$ represents the predicted values of [Ra,Rh,NEE] on the date d, and $f(\cdot)$ can be implemented as a fully connected network. By applying the model on the simulated data, we can compare the predicted $\hat{\mathbf{q}}$ and the simulated values $\mathbf{q}$ in each year, as follows:

$$\text{Diff-sim}_{i,t} = \sum_d ||\hat{q}_{i,t}^d - q_{i,t}^d||^2 \tag{9}$$

Given the GPP values, we also consider a penalty for violating the carbon mass conservation, as follows:

$$\text{MC}_{i,t} = \sum_d (\text{GPP}_{i,t}^d - \text{Ra}_{i,t}^d - \text{Rh}_{i,t}^d + \text{NEE}_{i,t}^d)^2 \tag{10}$$

We then combine Diff-sim and MC to define a physical loss. Here Diff-sim can be measured only on the simulated data. The mass conservation MC can be measured on both simulated data and real data (both $\mathcal{R}$ and $\mathcal{T}$) using the predicted Ra, Rh, and NEE. The physical loss can be expressed as:

$$\begin{aligned}\mathcal{L}_{phy} = \beta_1 \frac{\sum_{(i,t)\in\mathcal{S}} \text{Diff-sim}_{i,t}}{|\mathcal{S}|} \\ + \beta_2 \left(\frac{\sum_{t\in\mathcal{R}\cup\mathcal{T}} \sum_i \text{MC}_{i,t}^2}{|\mathcal{R}\cup\mathcal{T}|N} + \frac{\sum_{(i,t)\in\mathcal{S}} \text{MC}_{i,t}^2}{|\mathcal{S}|}\right),\end{aligned} \tag{11}$$

where β_1 and β_2 are model hyper-parameters.

Finally, we optimize the model combining the supervised loss (Eq. 4) and the physical loss (Eq. 11), as follows:

$$\mathcal{L}_{\text{PG-AN}} = \mathcal{L}_{sup} + \mathcal{L}_{phy} \tag{12}$$

We obtain the model $\mathcal{F}$ by minimizing the loss $\mathcal{L}_{\text{PG-AN}}$. We will then use the obtained embeddings $\mathbf{e}$ from this PG-AN model to estimate the weights $\{\bar{w}_{i,t}\}$ following Eq. 7 and the normalization.

Fairness-driven Model Refinement

After collecting the normalized weights $\bar{w}_{i,t}$, we refine the PG-AN model $\mathcal{F}$ to alleviate the temporal domain shift while preserving the spatial fairness. Note that the direct fine-tuning using the preliminary reweighted loss function (Eq. 8) only reduces the temporal gap but may impair the spatial fairness. Hence, we propose a bi-level fairness-driven refining strategy for the PG-AN model that takes into account both the temporal data shift and the spatial fairness.

First, we modify the original fairness objective (Eq. 6) by considering the similarity to the target dataset $\mathcal{T}$ based on the obtained sample weights $\bar{w}_{i,t}$. Each partition p contains training samples from multiple locations and multiple years. We will increase the weight for each sample (i,t) if the corresponding weight $\bar{w}_{i,t}$ is higher. This will be reflected in the performance measure $M_{\mathcal{F}}(\mathcal{F}_{\mathbf{\Theta}}, p)$ and the overall mean performance $E_{\mathcal{P}}$ in the fairness definition (Eq. 6). In this work, we use the predictive RMSE as the performance metric, and the weighted performance on each partition p and the weighted overall performance can be computed as:

$$\begin{aligned}\tilde{M}_{\mathcal{F}}(\mathcal{F}_{\mathbf{\Theta}}, p) &= \sqrt{\frac{\sum_{i\in p, t\in\mathcal{R}} \bar{w}_{i,t} (\mathcal{F}_{\mathbf{\Theta}}(\mathbf{x}_{i,t}) - y_{i,t})^2}{\sum_{i\in p, t\in\mathcal{R}} \bar{w}_{i,t}}} \\ \tilde{E}_{\mathcal{P}} &= \sqrt{\frac{\sum_{i\in\mathcal{P}, t\in\mathcal{R}} \bar{w}_{i,t} (\mathcal{F}_{\mathbf{\Theta}_0}(\mathbf{x}_{i,t}) - y_{i,t})^2}{\sum_{i\in\mathcal{P}, t\in\mathcal{R}} \bar{w}_{i,t}}}\end{aligned} \tag{13}$$

Then we use the weighted performance measure to redefine the spatial fairness, as follows:

$$\tilde{M}_{fair} = \sum_{p\in\mathcal{P}} \frac{d(\tilde{M}_{\mathcal{F}}(\mathcal{F}_{\mathbf{\Theta}}, p), \tilde{E}_{\mathcal{P}})}{|\mathcal{P}|} \tag{14}$$

A traditional way to incorporate the fairness objective (e.g., Eq. (14)) is to include it as an additional term in the loss function, e.g., $\mathcal{L} = \mathcal{L}_{sup} + \lambda \cdot \tilde{M}_{fair}$, where λ is a scaling factor or weight. This regularization-based formulation has several limitations when used for spatial-fairness enforcement. In particular, deep learning training often uses mini-batches due to data size, but it is difficult for each mini-batch to contain representative samples from all partitions $\{p \,|\, \forall p \in \mathcal{P}\}$ when calculating $\tilde{M}_{fair}$. More importantly, the regularization brings direct competition between $\mathcal{L}_{sup}$ and $\tilde{M}_{fair}$, and thus may lead to limited fairness improvement while preserving similar overall performance. Also, the regularization term requires another scaling factor λ, the choice of which directly impacts the final output and varies from problem to problem.

To mitigate these concerns, we propose to disentangle $\mathcal{L}_{sup}$ and $\tilde{M}_{fair}$ via a bi-level model refinement of the PG-AN model. Specifically, there are two levels of decision-making in this model refinement process:

[1] https://github.com/ai-spatial/PG-AN

Global referee: A referee evaluates the spatial fairness before each epoch using the metric $\tilde{M}_{\mathcal{F}}$ (e.g., RMSE) and Eq. (14). The evaluation is performed on all partitions $p \in \mathcal{P}$, guaranteeing the representativeness. Based on the deviation $d(\tilde{M}_{\mathcal{F}}(\mathcal{F}_{\mathbf{\Theta}}, p), \tilde{E}_{\mathcal{P}})$ of a particular partition p, we first create a revised learning rate $\eta'_p = \max((\tilde{M}_{\mathcal{F}}(\mathcal{F}_{\mathbf{\Theta}}, p) - \tilde{E}_{\mathcal{P}}), 0)$, which aims to increase the learning rate for partitions with larger RMSE values. Then we normalize the learning rate as $\eta_p = \frac{\eta'_p - \eta'_{min}}{\eta'_{max} - \eta'_{min}} \cdot \eta_{init}$, where η_{init} is the learning rate used to train the base model, $\eta'_{min} = \arg\min_{\eta'_p}\{\eta'_p \,|\, \eta'_p > 0, \forall p \in \mathcal{P}\}$, and $\eta'_{max} = \arg\max_{\eta'_p}\{\eta'_p \,|\, \forall p \in \mathcal{P}\}$ (Xie et al. 2022).

The intuition is that, if a partition's performance measure (e.g., RMSE) is worse than the expectation $\tilde{E}_p$, its learning rate η_p will be increased relative to other partitions. This increase ensures that the prediction loss for that partition has a higher impact during parameter updates in this epoch. In contrast, if a partition's performance is the same or better than the expectation, its η_p will be set to 0 to prioritize the worse-performing partitions. Positive learning rates after the update are then normalized back to the range $[0, \eta_{init}]$ to maintain gradient stability. This bi-level design also help get rid of the additional scaling factor to combine the prediction and fairness losses.

Model update: Using learning rates $\{\eta_p\}$ assigned by the referee, we perform regular training with the prediction loss $\mathcal{L}_{sup}$ over data in all individual partitions $p \in \mathcal{P}$ in mini-batches. Note that each partition p contains multiple data samples. When we iterate over each sample (i, t), we also use the obtained sample weights in updating the model parameters. Given the gradient $\Delta_{i,t}$ for each sample (i, t), the aggregated gradient descent can be expressed as:

$$\mathbf{\Theta}^{new} \leftarrow \mathbf{\Theta} - \sum_{p \in \mathcal{P}} \eta_p \sum_{i \in p} \bar{w}_{i,t} \Delta_{i,t} \tag{15}$$

Experiments

Dataset

We use the corn yield data in Illinois and Iowa from the years 2000-2020 provided by USDA National Agricultural Statistics Service (NASS) [2]. In particular, there are in total 199 counties in our study region (100 counties in Illinois and 99 counties in Iowa). The corn yield data (in gCm^{-2}) are available for each county each year. The input features have 19 dimensions, including NLDAS-2 climate data (Xia et al. 2012), 0-30cm gSSURGO soil properties [3], crop type information, the 250m Soil Adjusted Near-Infrared Reflectance of vegetation (SANIRv) based daily GPP product (Jiang et al. 2021), and calendar year. Moreover, we use the physics-based Ecosys model (Zhou et al. 2021) to simulate Ra, Rh, NEE, and crop yield for 10,335 samples from the years 2001-2018.

In our experiments, we consider two major use cases for yield prediction, data reanalysis and future prediction, and hence, two testing scenarios are applied, using the years 2005-2006 and the last two years 2019-2020 as target testing years, respectively. In each testing scenario, the remaining years are used for model training. We also consider two different spatial partitionings. The first partitioning $\mathcal{P}_{199}$ treats each county as a spatial partition, and there are totally 199 partitions. The second partitioning $\mathcal{P}_{30}$ merges neighboring 6-10 counties as a partition, and contains in total 30 partitions. The number of counties in each partition varies across different partitions as we need to ensure each partition is continuous over space.

Experiment Design

We aim to answer several questions in our experiments:

1. Can the proposed method outperform existing methods given the temporal data shift? The proposed method is compared against multiple baselines, including the standard LSTM-Attention networks (LSTM-Attn), the adversarial domain adaptation methods (DA) (Ganin et al. 2016), the adversarial discriminating-based learning for preserving fairness (ADL) (Alasadi, Al Hilli, and Singh 2019), regularization-based fairness enforcement method (REG) (Kamishima, Akaho, and Sakuma 2011; Yan and Howe 2019), REG with the reweighting strategy (REG^{rew}), and self-training-based fairness enforcement method (Self-training) (An et al. 2022). All these methods use the base LSTM-Attn model but adopt different strategies for preserving fairness or addressing the temporal data shift. Amongst these methods, ADL and REG consider the fairness objective, DA considers the temporal data shift, and REG^{rew} and Self-training consider both. We also compare with two methods that leverage simulated data for enhancing the LSTM-Attn model. As inspired by the prior work (Read et al. 2019; Jia et al. 2021a), the first method SIM-ptr pre-trains the LSTM-Attn model using simulated yield data and then fine-tunes it using real data. The second method SIM-inp is trained using simulated data to predict Ra, Rh, and NEE, and then use them as additional input features. We also implement the SIM-inp method with the bi-level refinement (SIM-inp^{ref}). Finally, we evaluate two versions of the proposed method PG-AN (without using the bi-level refinement) and PG-AN^{ref} (using the bi-level refinement). For each method, we measure the predictive RMSE and the spatial fairness (Eq. 6 using the mean absolute distance) under two different partitionings $\mathcal{P}_{199}$ and $\mathcal{P}_{30}$.

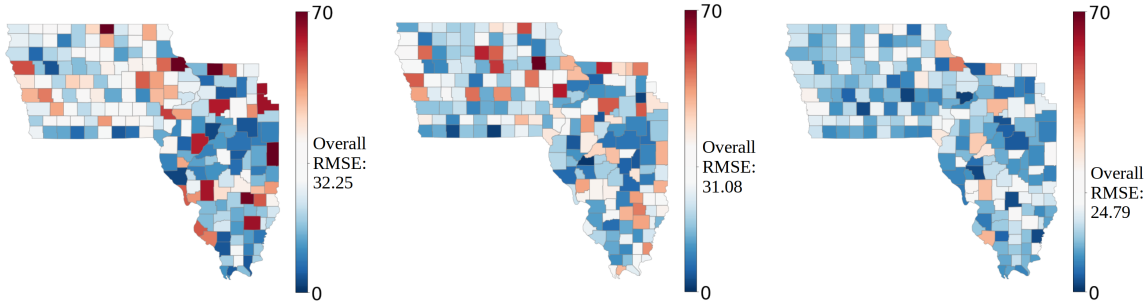
2. How will the performance change by adding sample weights and different levels of physical information? We compare the performance of LSTM-Attn, LSTM-Attn + sample weights (LSTM-Attn^{rew}), LSTM-Attn + sample weights + pre-training using simulated yield (LSTM-Attn$^{rew+ptr}$), LSTM-Attn + sample weights + pre-training using simulated yield, Ra, Rh, NEE, and the mass conservation on these simulated variables and GPP (LSTM-Attn$^{rew+phy}$), and LSTM-Attn + sample weights + training using simulated yield, Ra, Rh, NEE, the mass conservation on both simulated data and predicted values in real data, and real yield data (the proposed PG-AN model). We will also report the performance and fairness for each model either using or without using the bi-level fairness refinement.

[2] https://quickstats.nass.usda.gov/

[3] https://gdg.sc.egov.usda.gov/

Method	Testing scenario 2019-2020				Testing scenario 2005-2006			
	Partitioning $\mathcal{P}_{30}$		Partitioning $\mathcal{P}_{199}$		Partitioning $\mathcal{P}_{30}$		Partitioning $\mathcal{P}_{199}$	
	RMSE	Fairness	RMSE	Fairness	RMSE	Fairness	RMSE	Fairness
LSTM-Attn	37.4284	11.0158	37.4284	16.8612	32.2486	8.9010	32.2486	13.6076
DA	38.0840	11.0802	38.0840	16.8274	32.2888	9.1424	32.0610	13.9750
ADL	38.6144	10.9396	38.2536	16.7950	32.2870	9.0022	32.1376	13.6252
REG	37.6738	10.9102	38.5752	16.7746	31.6602	8.9202	31.3974	13.5626
REG^{rew}	36.2342	10.4966	36.5012	16.2694	29.5366	8.6024	30.1106	13.0416
Self-training	35.6784	10.3912	35.9520	16.1510	31.0714	8.6522	31.0758	12.9724
SIM-ptr	36.0920	10.5758	36.0920	16.1400	30.8404	8.6258	30.8404	12.7468
SIM-inp	34.3598	9.8968	34.3598	15.9064	30.6056	7.8356	30.6056	12.6990
SIM-inp^{ref}	33.9332	9.5888	33.9892	15.4732	30.0814	7.3696	31.0480	12.1536
PG-AN	30.3688	7.8064	**30.3688**	13.6370	**24.7858**	6.6092	**24.7858**	10.2498
PG-AN^{ref}	**29.9558**	**7.2682**	30.9058	**12.5252**	25.7476	**5.7254**	25.3546	**9.8554**

Table 1: The fairness and overall RMSE with two different partitionings for two testing scenarios.



(a) (b) (c)

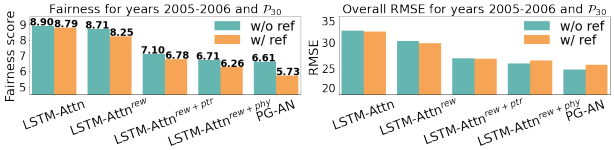
Figure 3: The distributions of predictive RMSE in 199 counties by three models for the testing years 2005-2006 and partitioning $\mathcal{P}_{199}$. (a): The LSTM-Attn model. (b): The Self-training model. (c): The proposed PG-AN model.

3. Can the bi-level fairness-driven refinement outperform other fairness enforcement methods? We will incorporate the same level of physical information and sample weights for the REG and ADL methods to create two baselines PG-AN^{REG} and PG-AN^{ADL}. We then compare their performance with the proposed PG-AN^{ref} method.

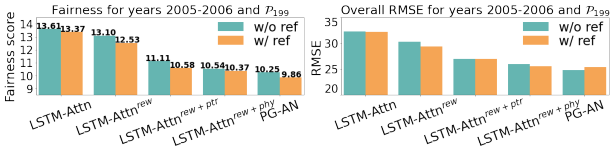
4. Can we interpret the distribution sample weights? We will study the distribution of the learned sample weights over different counties and different years.

Results

Performance comparison: Table 1 reports the performance for the proposed method and other baselines using different testing years and different spatial partitionings. It can be seen that the proposed methods (PG-AN and PG-AN^{ref}) outperform other methods by a decent margin in terms of both predictive RMSE and fairness measures. We also have several observations: (1) Compared to the base model LSTM-Attn, existing fairness enforcement methods (ADL, REG) only slightly improve the fairness in some testing cases, and can even lead to degraded fairness when tested in the years 2005-2006. This is because they do not consider the temporal data shift across years. (2) The DA method generally produces worse performance compared to LSTM-Attn because it cannot extract informative embeddings for enforcing invariance in the adversarial learning process. (3) The methods using the simulated data (SIM-ptr, SIM-inp, and SIM-inp^{ref}) perform better than the base LSTM-Attn model and most of other baselines, which confirms the effectiveness of incorporating the simulated data. Moreover, SIM-inp performs better than SIM-ptr because



(a) Fairness and predictive RMSE on $\mathcal{P}_{30}$.



(b) Fairness and predictive RMSE on $\mathcal{P}_{199}$.

Figure 4: The performance change for the testing years 2005-2006. A higher fairness score indicates larger mean absolute distance values and worse fairness performance.

Method	Testing scenario 2019-2020				Testing scenario 2005-2006			
	Partitioning $\mathcal{P}_{30}$		Partitioning $\mathcal{P}_{199}$		Partitioning $\mathcal{P}_{30}$		Partitioning $\mathcal{P}_{199}$	
	RMSE	Fairness	RMSE	Fairness	RMSE	Fairness	RMSE	Fairness
PG-AN	30.3688	7.8064	**30.3688**	13.6370	24.7858	6.6092	**24.7858**	10.2498
PG-AN^{ADL}	30.5730	7.7638	30.8136	13.5244	25.9296	6.3658	25.4034	10.2104
PG-AN^{REG}	**29.2328**	7.7154	31.5000	13.5916	**24.5180**	6.3400	25.4384	10.2210
PG-AN^{ref}	29.9558	**7.2682**	30.9058	**12.5252**	25.7476	**5.7254**	25.3546	**9.8554**

Table 2: Comparison between the bi-level refinement and other fairness enforcement methods for refining the PG-AN model.

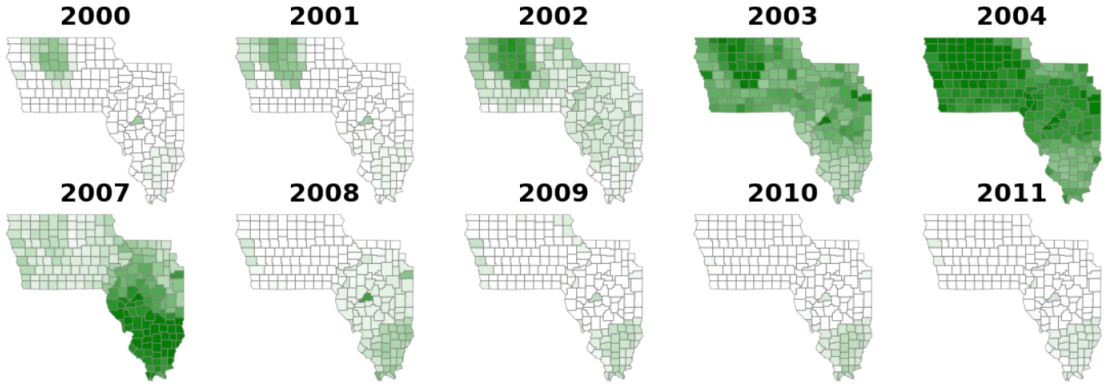


Figure 5: Training sample weights when we use 2005-2006 as testing years (green: higher weights; white: lower weights).

it captures the intermediate physical variables in the carbon cycle. (4) The comparisons between SIM-inp and SIM-inp^{ref} and between PG-AN and PG-AN^{ref} show the effectiveness of the bi-level refinement in enhancing the fairness.

Fig. 3 also shows the distributions of RMSE for each partition in $\mathcal{P}_{199}$ (i.e., each county) for the testing years 2005-2006 by the base LSTM-Attn model, the Self-training model, and the proposed PG-AN model. It can be clearly seen that the proposed method can effectively reduce the RMSE for those counties that are poorly modeled by the LSTM-Attn method and the Self-training method. Also, the overall RMSE gets significantly improved.

Ablation study: Fig. 4 shows that the model performance and spatial fairness get improved as we incorporate sample weights and more physical information. The PG-AN model performs better than LSTM-Attn$^{rew+phy}$ due to the gap between simulated and real data. Also, the bi-level refinement can always improve the spatial fairness for each model while maintaining a similar level of overall performance.

Effectiveness of bi-level training: Table 2 shows that the PG-AN model with the bi-level refinement achieves the best fairness without compromising the predictive RMSE performance. This is because the bi-level refinement mitigates the direct competition between predictive performance and spatial fairness, and avoids the selection of hyper-parameters.

Sample weights over space and time: Fig. 5 shows the sample weights for each county over multiple training years 2000-2011. We can see that the years that are closer to the testing years 2005-2006 generally have higher sample weights. Also, it shows the variability across space. Some counties in the testing years can be better predicted using the knowledge transferred from previous years, and the testing data are more similar to latter years for some other counties.

Conclusions

In this paper, we introduce a new method for predicting crop yield while maintaining spatial fairness. The proposed PG-AN model extracts the physics-aware representation, which is then used to discover the temporal data shift using a sample reweighting strategy. Finally, the PG-AN model is refined through a bi-level optimization process based on the reweighted fairness objective. The evaluations on the real corn yield dataset provided by NASS demonstrate that our proposed method outperforms a diverse set of baselines for enforcing fairness and addressing temporal data shift. Also, it is shown that the incorporation of sample weights and physical information can greatly improve both predictive performance and spatial fairness. The bi-level optimization also brings a larger fairness improvement compared with other fairness enforcement methods for model refinement.

Although the proposed method is developed and evaluated in the context of agricultural monitoring, it is generally applicable to many real applications of great societal relevance. For example, flood mapping is critical to inform timely actions (e.g., construction of barriers, sending emergency workers) and estimate insurance for surrounding areas, but existing streamflow prediction methods often lead to spatial biases due to the variability of soil conditions, catchment characteristics, and data quantity over space. Future work will also be pursued to consider other spatial partitionings, such as United States agricultural districts. We will also explore the interpretation of spatial biases for different partitions, e.g., which physical factors in certain regions lead to increased complexity for predicting crop yield.

Acknowledgements

This work was supported by NSF awards 2147195, 2105133, and 2126474, NASA award 80NSSC22K1164, the USGS awards G21AC10207, G21AC10564, and G22AC00266, Google's AI for Social Good Impact Scholars program, the DRI award at the University of Maryland, and CRC at the University of Pittsburgh.

References

Alasadi, J.; Al Hilli, A.; and Singh, V. K. 2019. Toward fairness in face matching algorithms. In *Proceedings of the 1st International Workshop on Fairness, Accountability, and Transparency in MultiMedia*, 19–25.

An, B.; Che, Z.; Ding, M.; and Huang, F. 2022. Transferring Fairness under Distribution Shifts via Fair Consistency Regularization. *arXiv preprint arXiv:2206.12796*.

Bailey, J. T.; and Boryan, C. G. 2010. Remote sensing applications in agriculture at the USDA National Agricultural Statistics Service. Technical report, Research and Development Division, USDA, NASS, Fairfax, VA.

Bebber, D. P.; Holmes, T.; and Gurr, S. J. 2014. The global spread of crop pests and pathogens. *Global Ecology and Biogeography*, 23(12): 1398–1407.

Benami, E.; Jin, Z.; Carter, M. R.; Ghosh, A.; Hijmans, R. J.; Hobbs, A.; Kenduiywo, B.; and Lobell, D. B. 2021. Uniting remote sensing, crop modelling and economics for agricultural risk management. *Nature Reviews Earth & Environment*, 2(2): 140–159.

Beven, K. 2006. A manifesto for the equifinality thesis. *Journal of hydrology*, 320(1-2): 18–36.

Bickel, S.; Brückner, M.; and Scheffer, T. 2007. Discriminative learning for differing training and test distributions. In *Proceedings of the 24th international conference on Machine learning*, 81–88.

Bock, M. E.; Kirkendall, N. J.; et al. 2018. *Improving crop estimates by integrating multiple data sources*. National Academies Press.

Boryan, C.; Yang, Z.; Mueller, R.; and Craig, M. 2011. Monitoring US agriculture: the US department of agriculture, national agricultural statistics service, cropland data layer program. *Geocarto International*, 26(5): 341–358.

d'Amour, C. B.; Reitsma, F.; Baiocchi, G.; Barthel, S.; Güneralp, B.; Erb, K.-H.; Haberl, H.; Creutzig, F.; and Seto, K. C. 2017. Future urban land expansion and implications for global croplands. *Proceedings of the National Academy of Sciences*, 114(34): 8939–8944.

Fan, J.; Bai, J.; Li, Z.; Ortiz-Bobea, A.; and Gomes, C. P. 2022. A GNN-RNN approach for harnessing geospatial and temporal information: application to crop yield prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 11873–11881.

Freedman, D. A.; and Berk, R. A. 2008. Weighting regressions by propensity scores. *Evaluation review*, 32(4): 392–409.

Ganin, Y.; Ustinova, E.; Ajakan, H.; Germain, P.; Larochelle, H.; Laviolette, F.; Marchand, M.; and Lempitsky, V. 2016. Domain-adversarial training of neural networks. *The Journal of Machine Learning Research*, 17(1): 2096–2030.

Grant, R.; Black, T. A.; Jassal, R. S.; and Bruemmer, C. 2010. Changes in net ecosystem productivity and greenhouse gas exchange with fertilization of Douglas fir: Mathematical modeling in ecosys. *Journal of Geophysical Research: Biogeosciences*, 115(G4).

Gupta, J.; Molnar, C.; Xie, Y.; Knight, J.; and Shekhar, S. 2021. Spatial variability aware deep neural networks (svann): A general approach. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 12(6): 1–21.

He, E.; Xie, Y.; Jia, X.; Chen, W.; Bao, H.; Zhou, X.; Jiang, Z.; Ghosh, R.; and Ravirathinam, P. 2022. Sailing in the location-based fairness-bias sphere. In *Proceedings of the 30th International Conference on Advances in Geographic Information Systems*, 1–10.

Jia, X.; Khandelwal, A.; Mulla, D. J.; Pardey, P. G.; and Kumar, V. 2019. Bringing automated, remote-sensed, machine learning methods to monitoring crop landscapes at scale. *Agricultural Economics*, 50: 41–50.

Jia, X.; Willard, J.; Karpatne, A.; Read, J. S.; Zwart, J. A.; Steinbach, M.; and Kumar, V. 2021a. Physics-guided machine learning for scientific discovery: An application in simulating lake temperature profiles. *ACM/IMS Transactions on Data Science*, 2(3): 1–26.

Jia, X.; Zwart, J.; Sadler, J.; Appling, A.; Oliver, S.; Markstrom, S.; Willard, J.; Xu, S.; Steinbach, M.; Read, J.; et al. 2021b. Physics-Guided Recurrent Graph Model for Predicting Flow and Temperature in River Networks. In *Proceedings of the 2021 SIAM International Conference on Data Mining (SDM)*, 612–620. SIAM.

Jia, X.; Zwart, J.; Sadler, J.; Appling, A.; Oliver, S.; Markstrom, S.; Willard, J.; Xu, S.; Steinbach, M.; Read, J.; et al. 2021c. Physics-guided recurrent graph model for predicting flow and temperature in river networks. In *Proceedings of the 2021 SIAM International Conference on Data Mining (SDM)*, 612–620. SIAM.

Jiang, C.; Guan, K.; Wu, G.; Peng, B.; and Wang, S. 2021. A daily, 250 m and real-time gross primary productivity product (2000–present) covering the contiguous United States. *Earth System Science Data*, 13(2): 281–298.

Jo, E. S.; and Gebru, T. 2020. Lessons from archives: Strategies for collecting sociocultural data in machine learning. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 306–316.

Jones, J. W.; Hoogenboom, G.; Porter, C. H.; Boote, K. J.; Batchelor, W. D.; Hunt, L.; Wilkens, P. W.; Singh, U.; Gijsman, A. J.; and Ritchie, J. T. 2003. The DSSAT cropping system model. *European journal of agronomy*, 18(3-4): 235–265.

Kamishima, T.; Akaho, S.; and Sakuma, J. 2011. Fairness-aware learning through regularization approach. In *2011 IEEE 11th International Conference on Data Mining Workshops*, 643–650. IEEE.

Mehrabi, N.; Morstatter, F.; Saxena, N.; Lerman, K.; and Galstyan, A. 2021. A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*, 54(6): 1–35.

Ortiz, R.; Sayre, K. D.; Govaerts, B.; Gupta, R.; Subbarao, G.; Ban, T.; Hodson, D.; Dixon, J. M.; Ortiz-Monasterio, J. I.; and Reynolds, M. 2008. Climate change: can wheat beat the heat? *Agriculture, Ecosystems & Environment*, 126(1-2): 46–58.

Read, J. S.; Jia, X.; Willard, J.; Appling, A. P.; Zwart, J. A.; Oliver, S. K.; Karpatne, A.; Hansen, G. J.; Hanson, P. C.; Watkins, W.; et al. 2019. Process-guided deep learning predictions of lake water temperature. *Water Resources Research*, 55(11): 9173–9190.

Serna, I.; Morales, A.; Fierrez, J.; Cebrian, M.; Obradovich, N.; and Rahwan, I. 2020. Sensitiveloss: Improving accuracy and fairness of face representations with discrimination-aware deep learning. *arXiv preprint arXiv:2004.11246*.

Singh Parihar, J.; Justice, C.; Soares, J.; Leo, O.; Kosuth, P.; Jarvis, I.; Williams, D.; Bingfang, W.; Latham, J.; and Becker-Reshef, I. 2012. GEO-GLAM: A GEOSS-G20 initiative on global agricultural monitoring. In *39th COSPAR Scientific Assembly*, volume 39, 1451.

Srinivasan, R.; Zhang, X.; and Arnold, J. 2010. SWAT ungauged: hydrological budget and crop yield predictions in the Upper Mississippi River Basin. *Transactions of the ASABE*, 53(5): 1533–1546.

Steed, R.; and Caliskan, A. 2021. Image representations learned with unsupervised pre-training contain human-like biases. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 701–713.

Sweeney, C.; and Najafian, M. 2020. Reducing sentiment polarity for demographic attributes in word embeddings using adversarial learning. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 359–368.

Whitcraft, A. K.; Becker-Reshef, I.; and Justice, C. O. 2020. NASA Harvest (ing) Earth Observations for Informed Agricultural Decisions. In *IGARSS 2020-2020 IEEE International Geoscience and Remote Sensing Symposium*, 3706–3708. IEEE.

Xia, Y.; Mitchell, K.; Ek, M.; Cosgrove, B.; Sheffield, J.; Luo, L.; Alonge, C.; Wei, H.; Meng, J.; Livneh, B.; et al. 2012. Continental-scale water and energy flux analysis and validation for North American Land Data Assimilation System project phase 2 (NLDAS-2): 2. Validation of model-simulated streamflow. *Journal of Geophysical Research: Atmospheres*, 117(D3).

Xie, Y.; He, E.; Jia, X.; Bao, H.; Zhou, X.; Ghosh, R.; and Ravirathinam, P. 2021. A statistically-guided deep network transformation and moderation framework for data with spatial heterogeneity. In *2021 IEEE International Conference on Data Mining (ICDM)*, 767–776. IEEE.

Xie, Y.; He, E.; Jia, X.; Chen, W.; Skakun, S.; Bao, H.; Jiang, Z.; Ghosh, R.; and Ravirathinam, P. 2022. Fairness by "Where": A Statistically-Robust and Model-Agnostic Bi-Level Learning Framework. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Yan, A.; and Howe, B. 2019. Fairst: Equitable spatial and temporal demand prediction for new mobility systems. In *Proceedings of the 27th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, 552–555.

Yang, K.; Qinami, K.; Fei-Fei, L.; Deng, J.; and Russakovsky, O. 2020. Towards fairer datasets: Filtering and balancing the distribution of the people subtree in the imagenet hierarchy. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 547–558.

Zafar, M. B.; Valera, I.; Gomez Rodriguez, M.; and Gummadi, K. P. 2017. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *Proceedings of the 26th international conference on world wide web*, 1171–1180.

Zhang, H.; and Davidson, I. 2021. Towards Fair Deep Anomaly Detection. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 138–148.

Zhou, W.; Guan, K.; Peng, B.; Tang, J.; Jin, Z.; Jiang, C.; Grant, R.; and Mezbahuddin, S. 2021. Quantifying carbon budget, crop yields and their responses to environmental variability using the ecosys model for US Midwestern agroecosystems. *Agricultural and Forest Meteorology*, 307: 108521.