SimFair: Physics-Guided Fairness-Aware Learning with Simulation Models

Zhihao Wang[1], Yiqun Xie[1]*, Zhili Li[1], Xiaowei Jia[2], Zhe Jiang[3], Aolin Jia[1], Shuo Xu[1]

[1]University of Maryland
[2]University of Pittsburgh
[3]University of Florida
{zhwang1, xie, lizhili}@umd.edu, xiaowei@pitt.edu, zhe.jiang@ufl.edu, {aolin, shuoxu98}@umd.edu

Abstract

Fairness-awareness has emerged as an essential building block for the responsible use of artificial intelligence in real applications. In many cases, inequity in performance is due to the change in distribution over different regions. While techniques have been developed to improve the transferability of fairness, a solution to the problem is not always feasible with no samples from the new regions, which is a bottleneck for pure data-driven attempts. Fortunately, physics-based mechanistic models have been studied for many problems with major social impacts. We propose SimFair, a physics-guided fairness-aware learning framework, which bridges the data limitation by integrating physical-rule-based simulation and inverse modeling into the training design. Using temperature prediction as an example, we demonstrate the effectiveness of the proposed SimFair in fairness preservation.

Introduction

As the use of artificial intelligence (AI) expands to more and more traditional domains, the bias in predictions made by AI has also raised broad concerns in recent years. To facilitate the responsible use of AI, fairness-aware learning has emerged as an essential component in AI's deployment in societal applications. In this study, we focus on learning-based mapping applications, where it is important to evaluate fairness over locations. Such maps are often used to inform critical decision-making in major social sectors, such as food, energy, water, public safety, etc.

In these applications, especially at large scales, inequity in performance is often caused by changes in distribution over different regions (Xie et al. 2021; Goodchild and Li 2021). One of the major bottlenecks is the unavailability of ground truth data in test regions. With no labels from the test area (e.g., when applying models trained in one state to another), it is very difficult to know how to obtain fairness over new locations in the test area. This is more challenging than transferring the overall prediction performance (e.g., measured by RMSE), which only needs to consider $f : \mathbf{X} \rightarrow \mathbf{Y}$ for the whole dataset. In the fairness-driven scenario, we also need to understand how the errors may vary over locations in

a different region, which often does not follow the same pattern as the training region (e.g., the number of locations may vary; data distribution may vary). Finally, the training and test areas often have completely different sets of locations, making the groups used in the fairness evaluation nonstationary as well.

In this paper, we use the temperature prediction problem as a concrete example. Air and surface temperatures are two key variables for estimating the Earth's energy budget, which connects to a diverse range of social applications, such as solar power, agriculture, climate change, global warming, ecosystem dynamics, and urban heat islands (Kim and Entekhabi 1998; Peng et al. 2014; Wang et al. 2023; Li et al. 2022b). For example, temperature-related variables help estimate solar energy potential or predict the risks of floods or droughts at different locations. The results may affect resource allocation decisions such as subsidies, promotions, or insurance. Practically, satellite remote sensing is the only approach to measuring these variables at the spatial and temporal resolution needed for most applications (Liang 2001). Due to the large volume of satellite data, machine learning methods have become increasingly popular choices in predicting temperature-related variables (Deo and Şahin 2017; Wang et al. 2021). However, fairness has yet to be considered. Due to the social impact, it is important to ensure fairness among different places in the prediction map.

Given passive microwave and multi-spectral optical remote sensing imagery, the goal of the paper is to predict temperature while maintaining fairness among prediction performance over locations. In particular, we aim to improve the fairness of predictions in new test areas.

Recent studies have developed various approaches for fairness improvement. On the data side, fairness-driven collection methods and filtering strategies were proposed to reduce bias caused by data issues such as imbalance (Jo and Gebru 2020; Yang et al. 2020; Steed and Caliskan 2021). The methods are more suitable for domains where ground-truth data are reasonably easy to obtain. However, for most remote sensing problems, it is resource-intensive and time-consuming to collect new ground-truth samples (e.g., field surveys, sensor installation, and monitoring stations). Many formulations explored decorrelating the feature learning process with sensitive attributes, which revealed information such as races and genders should not be discrim-

inated in prediction (Zhang and Davidson 2021; Alasadi, Al Hilli, and Singh 2019; Sweeney and Najafian 2020). For example, adversarial learning is a popular design choice in learning group-invariant features. The use of regularization terms is another common approach to reduce bias risks, where fairness loss is used to penalize biased predictions (Yan and Howe 2019; Serna et al. 2020; Zafar et al. 2017). These methods, however, are not suitable for fair learning here between spatial regions as they require a fixed set of groups such as different genders, whereas the groups represented by locations vary between different regions. There have also been studies for the time-series or online setups (Zhao et al. 2022; Bickel, Brückner, and Scheffer 2007; An et al. 2022). They aim to maintain fairness as new samples come in by sample reweighting, meta-learning, etc. Similarly, these methods focus on fixed groups and are designed for dynamic changes in time series. They may also require additional ground-truth samples for finetuning. Location-based fairness was also recently explored (Xie et al. 2022; He et al. 2022, 2023), which reduced the statistical sensitivity in fairness evaluation for regression and classification tasks. However, it also requires training and test data from the same region. Finally, all the above methods are purely data-driven, and their transferability is limited when no labels are available in a new region.

To address the limitations, we propose SimFair, a physics-guided fairness-aware learning approach, which uses simulations from mechanistic models to improve fairness in test regions. To the best of our knowledge, this is the first work that integrates physics-based simulation (mechanistic) models into fairness-aware learning. Our contributions include:

- We present an inverse-modeling based design to integrate physics-based simulation models into the training process, which are often incompatible with the learning objectives in remote sensing problems.

- We propose a training strategy with dual-fairness consistency to improve fairness over new test locations.

- We incorporate physical-rule-based constraints to further improve the prediction performance.

- We integrate SimFair with different simulation models and real-world datasets for temperature prediction.

Through experiments, we demonstrate that the inverse modeling is robust, and SimFair can greatly improve fairness over new locations in test regions.

Problem Definition

Definition 1 (Spatiotemporal (ST) domain) *Given a geographic space $S = \{s_1, s_2, ...\}$ and a time-period $T = \{t_1, t_2, ...\}$, a ST-domain $\mathcal{D}$ is a contiguous subspace in $S \times T$. For example, $\mathcal{D}$ can represent a contiguous geographic area (e.g., a county) over a month.*

Definition 2 (Location-based fairness measure) *It evaluates **prediction quality parity**, one of the standard definitions of fairness (Du et al. 2020), over a set of locations in a geographic region. Denote $\mathcal{F}$ as a prediction model; $\mathcal{L}_p$ as the measure of prediction errors (e.g., RMSE); X and Y as test*

features and labels, respectively; and $\boldsymbol{x}_i \in X$ and $\boldsymbol{y}_i \in Y$ as features and labels for location $s_i \in S$, respectively. Here the location-based fairness $\mathcal{L}_f$ is defined as:

$$\mathcal{L}_f = \frac{1}{|S|} \sum_{s_i \in S} \left| \mathcal{L}_p(\mathcal{F}(\boldsymbol{x}_i), \boldsymbol{y}_i) - \overline{\mathcal{L}_p}(\mathcal{F}(X), Y) \right| \quad (1)$$

$\mathcal{L}_f$ evaluates the deviation of prediction performance from the global performance (i.e., a scalar obtained using entire test data X and Y). A smaller $\mathcal{L}_f$ means the overall deviation is smaller, and thus the model is fairer over the locations.

Formulation of location-based fair learning. Given training samples X and Y from a ST-domain $\mathcal{D}$, and test features X' from a new ST-domain $\mathcal{D}'$, we aim to learn a (location-based) fairness-aware model from $\mathcal{D}$, which performs well in $\mathcal{D}'$ and, more importantly, offers fairer solution quality over locations in $\mathcal{D}'$.

A key characteristic of the problem is that the groups (i.e., locations $s \in S \in \mathcal{D}$) being considered are not prefixed and can be highly dynamic. From one ST-domain to another, the locations being considered can be completely different (e.g., from one state to another). This makes it difficult to connect the learning objectives from the training domain $\mathcal{D}$ to the target domain $\mathcal{D}'$. Making the problem more challenging, only the features X' are available from the new domain, and no label is available. In essence, we need to build a fairness-aware model under distribution-shifts, different groups for fairness evaluation, and unknown labels.

Method

We propose SimFair, a physical-simulation-guided learning framework to improve the fairness-awareness of models for new ST-domains. To be concrete, we use temperature prediction as an example to illustrate the design. In this section, we first provide brief overviews of two physics-based models we use, and then discuss the new SimFair framework.

Physics-based Mechanistic Models

Physics-Model 1 (PM1): The Community Microwave Emission Model (CMEM), as a subset of global operating systems at the European Centre, estimates low-frequency passive microwave brightness temperature (BT) (Kerr et al. 2010; Wigneron et al. 2017). In the simulation process (Fig. 1(b)), CMEM computes the Top-of-Atmosphere (TOA) BTs $T_{Btov,p,\theta}$ over vegetation layers for each polarisation direction p and incidence angle θ by summing the soil effective temperature T_{eff}, vegetation temperature T_{Bveg}, and atmospheric components T_{Bad} and T_{Bau} (identical for high-altitude satellites). The overall physical process can be expressed as:

$$T_{Btov,p,\theta} = (1 - \boldsymbol{r}_{r,p,\theta})T_{eff} \cdot \exp(-\tau_{veg,p,\theta})$$
$$+ T_{Bveg,p,\theta}(1 + \boldsymbol{r}_{r,p,\theta} \cdot \exp(-\tau_{veg,p,\theta})) \quad (2)$$
$$+ T_{Bad,p,\theta} \cdot \boldsymbol{r}_{r,p,\theta} \cdot \exp(-2\tau_{veg,p,\theta})$$

where $\boldsymbol{r}$ is soil surface reflectivity and τ is optical depth.

(a) Overall framework of SimFair
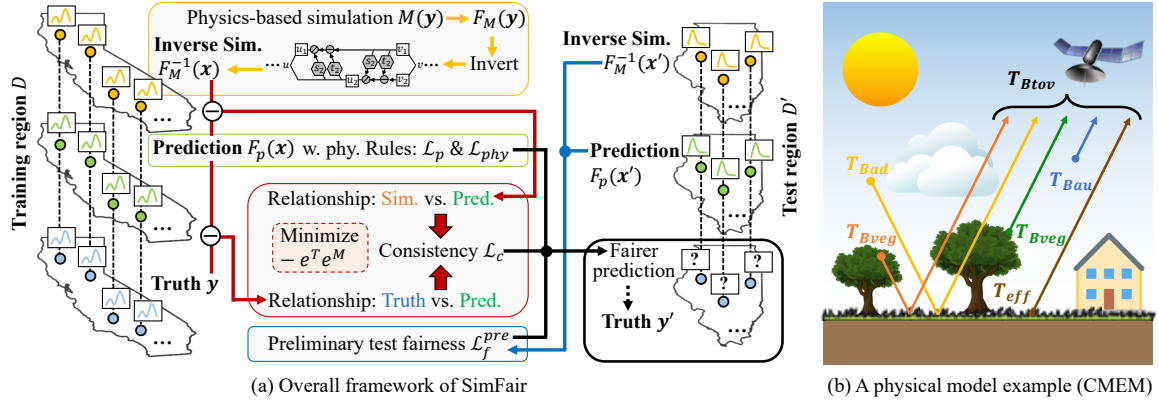
(b) A physical model example (CMEM)

Figure 1: An illustration of simulation-model-guided fairness-aware learning.

Physics-Model 2 (PM2): The MODerate resolution atmospheric TRANsmission (MODTRAN) model has been used worldwide to analyze, estimate, and predict the optical characteristics of the atmosphere based on the radiation transport physics (Berk et al. 2008, 2014). In remote sensing, TOA radiance, observed by satellites, is the mixed BT that is emitted, reflected, and transmitted by atmosphere and surface objects. MODTRAN simulation process is governed by:

$$R_i(\theta) = (\varepsilon_i B_i(T_s) + (1 - \varepsilon_i) R_{i\downarrow}) \tau_i(\theta) + R_{i\uparrow}(\theta) \quad (3)$$

where $R_i(\theta)$ is the TOA radiance captured by a certain range of wavelength (i.e., a satellite band) i at a viewing zenith angle θ; $R_{i\downarrow}$ and $R_{i\uparrow}$ represent the downward and upward atmospheric thermal radiance, respectively; ε is land surface emissivity; τ is the atmospheric transmittance; and $B_i(T_s)$ denotes the Planck radiance at land surface temperature.

SimFair: Simulation-Enabled Fair Learning

The overall framework of SimFair is illustrated in Fig. 1. Intuitively, we aim to learn the relationships between the data- and simulation-based predictions, and leverage these relationships to approximate fairness in a new test area. SimFair has four components: (1) inverse learning of the simulation models, which aligns the mechanistic model with a deep learning model; (2) preliminary test fairness, which weakly estimates fairness in the test region using simulations but by itself is insufficient to improve fairness; (3) a dual-fairness consistency, which tries to minimize the gap between data- and simulation-based fairness; and (4) physical rules, which are used as soft constraints to improve generalizability.

Inverse Modeling for Learning. In physics-based modeling, the processes are not necessarily derived from a direction that aligns with the one we use in prediction tasks. For example, in temperature simulation for passive remote sensing (PM1), the real physical process starts from the air or surface temperature, where radiance travels through the air – being absorbed, reflected/deflected, emitted, or transmitted by vegetation, built-ups, atmospheric particles, etc., – and finally reaches the spectral sensor from the satellites and recorded as signal values. This process can be described as $\mathbf{X} = \mathcal{M}(\mathbf{Y})$ where $\mathbf{X}$ represents satellite signals, $\mathbf{Y}$ is

the temperature, and $\mathcal{M}$ is the mechanistic model. However, in real-world applications, it often goes in the opposite direction, where users predict the temperature (i.e., $\mathbf{Y}$) using satellite readings $\mathbf{X}$. Having consistent directions is important for the use of simulation models in guiding data-driven approaches, because for each observation $\mathbf{x}_i$ we need to know the corresponding simulated value $\mathbf{y}_i = \mathcal{M}^{-1}(\mathbf{x}_i)$ (e.g., temperature) to extract useful information. Unfortunately, it is often very difficult to directly find the inverse of a mechanistic model due to the complexity of the physical process. For example, there are no known inversions of the mechanistic models PM1 and PM2 used here.

To address this issue, we first use bijector-based invertible networks (Kobyzev, Prince, and Brubaker 2020; Kingma et al. 2016; Dinh, Sohl-Dickstein, and Bengio 2016) to approximate the inverses of physics-based models; the structures were often used in normalizing flows for the estimation of complex statistical distributions and random sampling. While the direction can also be reversed in vanilla neural networks by swapping $\mathbf{X}$ and $\mathbf{Y}$, we use the invertible design for three major reasons:

- In physical processes $\mathbf{X} = \mathcal{M}(\mathbf{Y})$, many physics constraints can only be used on the variables in $\mathbf{X}$ (the constraints are built into the loss later in a neural network). There is no problem if we train an invertible network using direction $\mathbf{X} = \mathcal{F}(\mathbf{Y})$ and then inverse it. However, if we simply use a data swap $\mathbf{Y} = \mathcal{F}(\mathbf{X})$, we can no longer apply the constraints, as $\mathbf{X}$ are fixed inputs for training instead of outputs.

- The invertible structure naturally provides extra regularization, as the learned weights need to work simultaneously for both directions, improving prediction quality during the test (evaluated later in experiments).

- When $\mathbf{x}_i$ and $\mathbf{y}_i$ have different lengths, the invertible structure can be naturally extended with normalizing flow to quantify the uncertainty from fewer to more variables.

In the application context, we denote $\mathbf{X}$ as satellite signals, $\mathbf{Y}$ as the prediction target (e.g., temperature), $\mathcal{M}$ as the mechanistic model, $\mathcal{F}_{\mathcal{M}}(\cdot; \boldsymbol{\Theta})$ as an invertible neural network, $\mathcal{F}_{\mathcal{M}}^{-1}(\cdot; \boldsymbol{\Theta})$ as its inverse, and $\mathcal{L}$ as a loss function (e.g.,

RMSE). The inverse approximation is given by:

$$\mathcal{F}_{\mathcal{M}}^{-1}(\mathbf{X}; \Theta^* = \arg\min_{\Theta} \mathcal{L}(\mathcal{M}(\mathbf{Y}), \mathcal{F}_{\mathcal{M}}(\mathbf{Y}, \Theta))) \quad (4)$$

The bijector-based invertible layers present a great fit for the inverse approximation because (1) while complex, a physics-based mechanistic model describes a single function, i.e., all simulated labels $\mathcal{M}(\mathbf{X})$ follow the same distribution $P(\mathcal{M}(\mathbf{X})\,|\,\mathbf{X})$. Thus, $\mathcal{F}_{\mathcal{M}}(\cdot; \Theta)$ can effectively approximate $\mathcal{M}(\mathbf{X})$ given the capability of deep neural networks to universally approximate continuous functions. (2) Bijectors use mathematically exact inversion, enabling us to create a highly accurate approximation of the inverse of $\mathcal{M}(\mathbf{X})$. Specifically, we use the following formulation (Dinh, Sohl-Dickstein, and Bengio 2016):

$$
\begin{cases}
\mathbf{v}_1 &= \mathbf{u}_1 \odot \exp\left(s_2(\mathbf{u}_2)\right) + t_2(\mathbf{u}_2) \\
\mathbf{v}_2 &= \mathbf{u}_2 \odot \exp\left(s_1(\mathbf{v}_1)\right) + t_1(\mathbf{v}_1)
\end{cases}
$$
$$
\Longleftrightarrow
\begin{cases}
\mathbf{u}_2 &= \left(\mathbf{v}_2 - t_1(\mathbf{v}_1)\right) \odot \exp\left(-s_1(\mathbf{v}_1)\right) \\
\mathbf{u}_1 &= \left(\mathbf{v}_1 - t_2(\mathbf{u}_2)\right) \odot \exp\left(-s_2(\mathbf{u}_2)\right)
\end{cases}
$$
$$(5)$$

where $\mathbf{u}$ and $\mathbf{v}$ are the input and output of an invertible layer, respectively; $\mathbf{u} = [\mathbf{u}_1, \mathbf{u}_2]$ and $\mathbf{v} = [\mathbf{v}_1, \mathbf{v}_2]$; $s_1(\cdot)$, $t_1(\cdot)$, $s_2(\cdot)$, and $t_2(\cdot)$ are learnable functions. Note that both the input and output have the same length, which is a property of the bijector needed to make inversions.

Using a chain of bijectors as network layers, we construct the invertible network $\mathcal{F}_{\mathcal{M}}$ to approximate the inverse of $\mathcal{M}$. As the parameters we are interested in are a subset of those in the complete mechanistic models, we select the most related ones to define the original input and final output of the chain of bijectors $\mathcal{F}_{\mathcal{M}}$, which also allows us to make their lengths equivalent. Through experiments, we found that this led to approximations with higher precision for both directions of $\mathcal{F}_{\mathcal{M}}$ (i.e., $\hat{\mathbf{X}} = \mathcal{F}_{\mathcal{M}}(\mathbf{Y})$ and $\hat{\mathbf{Y}} = \mathcal{F}_{\mathcal{M}}^{-1}(\mathbf{X})$), compared to the formulations where $\mathbf{X}$ and $\mathbf{Y}$ had different lengths (a random vector $\mathbf{z}$ needs to be added to the shorter one in this case, which is also a more flexible option).

Preliminary Fairness on Test Region. This is the first component of SimFair. As shown in Fig. 1, we aim to approximate the fairness between data samples at different locations in the test region $\mathcal{D}'$ using relationships between simulation- and learning-based predictions. Here the results from the inverse simulation model $\hat{\mathbf{Y}}^{\mathcal{M}}$, obtained through the invertible network $\mathcal{F}_{\mathcal{M}}^{-1}(\mathbf{X})$, provide us a preliminary peek into the labels $\mathbf{Y}$ from the test region $\mathcal{D}'$.

Here we need to emphasize that as there is no guarantee about the distances between $\hat{\mathbf{Y}}^{\mathcal{M}}$ and real labels $\mathbf{Y}$ (or the variance of the distances), fairness scores evaluated using Eq. (1) with $\hat{\mathbf{Y}}^{\mathcal{M}}$ as the truth are not directly representative of the true fairness. Thus, the goal of this part is only to create a "preliminary fairness" as a preparation step for the dual-fairness consistency module in the next section, where new designs will be used to bridge the gap.

With that clarified, the preliminary fairness loss on $\mathcal{D}'$ is:

$$\mathcal{L}_f^{pre} = \frac{1}{|D'|} \sum_{\mathbf{x}_i \in D'} \left| \mathcal{L}_p(\mathcal{F}_p(\mathbf{x}_i), \hat{\mathbf{y}}_i^{\mathcal{M}}) - \overline{\mathcal{L}_p}(\mathcal{F}_p(\mathbf{X}), \hat{\mathbf{Y}}^{\mathcal{M}}) \right| \quad (6)$$

(a) No consideration of the triplet relationships

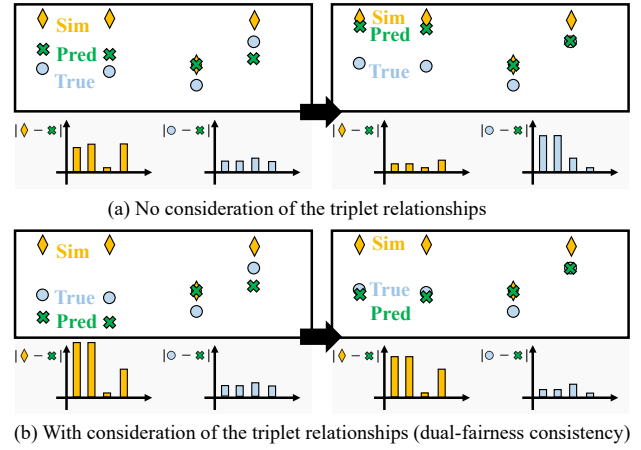(b) With consideration of the triplet relationships (dual-fairness consistency)

Figure 2: Illustrative example of dual-fairness consistency.

where $\mathbf{x}_i$ represents the data point at location s_i, $\hat{\mathbf{Y}}^{\mathcal{M}} = \mathcal{F}_{\mathcal{M}}^{-1}(\mathbf{X})$, $\mathcal{L}_p$ is the prediction loss, and $\mathcal{F}_p$ is the prediction model, which is used to predict the real values rather than approximate the simulation model $\mathcal{M}$.

Dual-Fairness Consistency. The dual-fairness consistency module aims to reduce the gap between the preliminary fairness loss $\mathcal{L}_f^{pre}$ and the real fairness loss $\mathcal{L}_f^{real}$ (not evaluable in training) for the test region $\mathcal{D}'$. It achieves this by learning and governing the triplet relationships among the following in the training data:

- Physical simulations $\hat{\mathbf{Y}}^{\mathcal{M}}$ (inversely approximated);
- Deep neural network predictions $\hat{\mathbf{Y}}$; and
- True labels $\mathbf{Y}$.

While we do not know the relationships between the simulation results $\hat{\mathbf{Y}}^{\mathcal{M}}$ and true labels $\mathbf{Y}$ in the test region $\mathcal{D}'$, we can find a solution $\mathcal{F}_p$ whose predictions $\hat{\mathbf{Y}}$ have a similar relationship with simulated-based $\hat{\mathbf{Y}}^{\mathcal{M}}$ and true $\mathbf{Y}$. Specifically, for the triplet relationship, our desired property is:

Definition 3 (Dual-fairness consistency) *Denote $\mathbf{e} = \mathbf{Y} - \hat{\mathbf{Y}}$ and $\mathbf{e}_{\mathcal{M}} = \hat{\mathbf{Y}}^{\mathcal{M}} - \hat{\mathbf{Y}}$, which represent the differences between the true labels and the predictions, and the simulated labels and the predictions, respectively. Dual-fairness refers to the fairness evaluation defined using true labels (Eq. (1)) and simulation labels (Eq. (6); here for training data in $\mathcal{D}$), respectively. To make the two fairness results more consistent, we aim to align the direction of the predicted labels $\hat{\mathbf{Y}}$ with respect to the true labels $\mathbf{Y}$ and simulation labels $\hat{\mathbf{Y}}^{\mathcal{M}}$:*

$$
\begin{cases}
\mathbf{e}_i \geq 0, & \text{if } (\mathbf{e}_{\mathcal{M}})_i \geq 0; \\
\mathbf{e}_i < 0, & \text{otherwise.}
\end{cases}
\quad (7)
$$

where i denotes the ith data point.

Fig. 2 shows the high-level idea with an illustrative example. As we can see, the relationships, i.e., $\mathbf{e}$ and $\mathbf{e}_{\mathcal{M}}$, are often not aligned in Fig. 2 (a), which does not consider the

consistency. As a result, improving fairness w.r.t. the simulation results leads to a less fair result. In contrast, with the consistency, the fairness improvement w.r.t. simulation data is more likely to lead to improvements w.r.t. the true data. Intuitively, when the directions represented by $\mathbf{e}_i$ and $(\mathbf{e}_{\mathcal{M}})_i$ are aligned, reducing the distance between a prediction $\hat{\mathbf{Y}}_i$ and a simulation label $\hat{\mathbf{Y}}_i^{\mathcal{M}}$ will accordingly reduce the distance between $\hat{\mathbf{Y}}_i$ and $\mathbf{y}_i$. Note that since both $\mathbf{Y}$ and $\hat{\mathbf{Y}}^{\mathcal{M}}$ are fixed inputs at this stage and they are not identical, it is impossible to make $\mathbf{e} = \mathbf{e}_{\mathcal{M}}$. Instead, our focus here is to promote a solution $\mathcal{F}_p$ if it maintains similar directional relationships between $\mathbf{e}$ and $\mathbf{e}_{\mathcal{M}}$. This is important for reducing the fairness loss, as we are trying to re-balance the prediction losses among points at different locations while keeping a similar global prediction loss (Eq. (1) or (6)). In other words, the fairness loss moves a prediction closer to the true label if the loss is worse than the global mean loss, and farther otherwise. Based on the dual-fairness consistency, we have the consistency loss as:

$$
\begin{aligned}
\mathcal{L}_c &= -\mathbf{e}^T \mathbf{e}_{\mathcal{M}} \\
&= - \sum_{(\mathbf{x}_i, \mathbf{y}_i) \in \mathcal{D}} \left(\mathbf{y}_i - \mathcal{F}_p(\mathbf{x}_i)\right) \cdot \left(\mathcal{F}_{\mathcal{M}}^{-1}(\mathbf{X}) - \mathcal{F}_p(\mathbf{x}_i)\right)
\end{aligned}
\tag{8}
$$

$\mathcal{L}_c$ allows gradients based on the preliminary test fairness loss $\mathcal{L}_f^{pre}$ to be more reflective of the true fairness loss.

Improvements with Physics-guided Predictions. We incorporate physical constraints from the mechanistic models as part of the loss functions to reduce the overfitting of the prediction model $\mathcal{F}_p$ (Jia et al. 2021; Chen et al. 2023), which accordingly makes it generalize better to the test region $\mathcal{D}'$. As physical models rely on different assumptions, we use two different constraints for the two physical models (i.e., PM-1 and PM-2). The physical rule used for PM-1 is the Rayleigh-Jeans Law of radiation, which states the radiance emitted by a gray body (e.g., trees, rocks) is less than a black body with unity emissivity. The loss $\mathcal{L}_{phy}^{PM1}$ is then:

$$
\mathbf{1}^T [\max\left(0, \mathbf{X} - \boldsymbol{\varepsilon} \otimes \mathcal{F}_p(\mathbf{X})\right) + \min\left(0, \mathbf{X} - \boldsymbol{\eta} \otimes \mathcal{F}_p(\mathbf{X})\right)]
\tag{9}
$$

where $\boldsymbol{\varepsilon} = \min(\mathbf{1}, \mathcal{F}_{\mathcal{M}}(\mathbf{y}_i) \otimes (\mathcal{F}_{\mathcal{M}}^{-1}(\mathbf{x}_i)))^{-1}$, $\boldsymbol{\eta} = \max(\mathbf{0}, \mathcal{F}_{\mathcal{M}}(\mathbf{y}_i) \otimes (\mathcal{F}_{\mathcal{M}}^{-1}(\mathbf{x}_i)))^{-1}$, and $\otimes$ is the Hadamard product.

For PM-2, the model output (temperature) is bounded by a well-known principle, the surface energy balance equation. Specifically, in the solar-earth energy exchange system, the overall energy is balanced by solar downward shortwave $R_{S\downarrow}$ and longwave $R_{L\downarrow}$, surface upward shortwave $R_{S\uparrow}$ and longwave $R_{L\uparrow}$, and net radiance R_N. The balance of energy at the surface is also related and can be expressed as the combination of upward surface sensible heat flux H_S, upward surface latent heat flux H_L, and downward ground heat flux H_G. This leads to the following loss $\mathcal{L}_{phy}^{PM2}$:

$$
-\varepsilon\sigma\mathcal{F}_p(\mathbf{X})^4 + R_{S\downarrow} - R_{S\uparrow} + \varepsilon R_{L\downarrow} - (H_S + H_L + H_G)
\tag{10}
$$

where ε is the surface emissivity, and σ is the Stefan-Boltzmann contant. Finally, the overall loss is $\mathcal{L} = \mathcal{L}_p + \mathcal{L}_f^{pre} + \mathcal{L}_c + \mathcal{L}_{phy}$, where $\mathcal{L}_{phy}$ is selected based on the physical model used (e.g., PM-1), $\mathcal{L}_p$ we use in the paper is mean
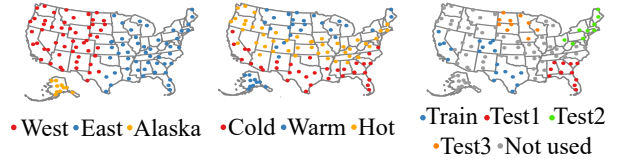

•West •East •Alaska   •Cold •Warm •Hot   •Train •Test1 •Test2
•Test3 •Not used

Figure 3: Spatial distributions of training and testing data.

squared loss $\mathcal{L}_p = ||\mathcal{F}_p(\mathbf{X}) - \mathbf{Y}||_2^2/|\mathcal{D}|$, and all losses are normalized based on the number of samples.

Deep Networks

We implemented SimFair using two types of networks: (1) a fully-connected neural network, FNN, which uses observed signals from satellite snapshots to make predictions; and (2) a long-short-term-memory (LSTM) model that uses a time-series-based input. Our invertible network uses a chain of 7 bijector layers. We use root-mean-squared-errors (RMSE) as the loss function and the Adam optimizer with an initial learning rate of 10^{-2}. More details are in the Appendix.

Experiments

In-Situ and Remote Sensing Datasets

We use three real datasets for evaluation: AT1, AT2 and LST (detailed in the following paragraphs). As AT1 contains the largest number of high-quality stations (122), we use it to evaluate the models' ability to promote fairness in test regions, that contain different locations from the training region. Additionally, we include two smaller datasets AT2 and LST, which are used to evaluate if fairness learned among the 7 locations during a period can be transferred to a new period (i.e., the same set of locations over different periods).

AT1: USCRN-CMEM air temperature data. We collected the ground truth from the entire USCRN stations (200+) in 2014. All station measurements were carefully examined, and only the dates with high-quality measurements were used. Satellite observations were collected from the AMSR2 satellite with two observations per day. CMEM (PM1) model inputs were collected from ERA5 hourly dataset, including soil temperature, volumetric soil water layer, etc. Other surface and atmospheric datasets were simulated using ecoClimate. We used three types of space partitionings to create train-test splits (Fig. 3) with different geographic regions, temperature zones, and random local states.

AT2: SURFRAD-CMEM air temperature data. Different from AT1, the ground truth in AT2 was collected from a well-known and high-quality network, SURFRAD, which measured surface conditions and energy at minute scales. As discussed, we separated the data using two temporal splits: (1) first 8 months as training and last 4 months as test; and (2) first 4 months as training and last 8 months as test.

LST: SURFRAD-MODTRAN land surface temperature data. We collected the surface temperature and four radiance measurements in SURFRAD stations from 2013 to 2020. Satellite observations were collected from Landsat images. MODTRAN (PM2) inputs were collected from

		Train-Test: West-East			Train-Test: East-West			Train-Test: East-Alaska		
	Model	RMSE	Corr.	Fairness	RMSE	Corr.	Fairness	RMSE	Corr.	Fairness
FNN	BaseNet	6.75	0.83	4.49 (±0.81)	27.56	0.34	13.65 (±2.48)	52.03	0.19	44.42 (±3.37)
	Sim	6.45	0.86	4.69 (±0.82)	20.33	0.44	11.93 (±2.19)	43.74	0.29	38.84 (±5.69)
	SimPhy	7.19	0.84	5.56 (±0.4)	17.78	0.48	10.79 (±1.98)	45.52	0.29	38.84 (±3.4)
	RegFair	7.22	0.8	4.97 (±0.69)	25.35	0.37	12.36 (±2.42)	38.5	0.06	29.73 (±6.78)
	Self-Reg	6.35	0.84	4.27 (±0.7)	31.97	0.31	16.48 (±2.42)	38.01	0.06	28.95 (±4.15)
	SimFair	3.07	0.97	2.04 (±0.19)	3.11	0.96	**1.94 (±0.03)**	6.23	0.84	**4.25 (±0.78)**
	SimFair-P	2.88	0.97	**1.89 (±0.06)**	3.13	0.96	1.96 (±0.05)	6.29	0.81	4.45 (±0.51)
LSTM	BaseNet	4.22	0.93	2.66 (±0.14)	4.02	0.97	2.45 (±0.16)	11.93	0.8	5.14 (±0.31)
	Sim	3.89	0.95	2.43 (±0.15)	3.3	0.97	2.21 (±0.40)	13.32	0.85	5.25 (±0.49)
	SimPhy	4.46	0.95	2.69 (±0.17)	3.23	0.97	2.04 (±0.17)	12.27	0.88	4.82 (±0.28)
	RegFair	4.17	0.94	2.66 (±0.22)	4.03	0.96	2.59 (±0.58)	12.16	0.81	5.03(±0.4)
	Self-Reg	4.10	0.94	2.57 (±0.26)	3.85	0.96	2.41 (±0.16)	11.24	0.84	4.68(±0.41)
	SimFair	3.46	0.96	2.21 (±0.11)	3.22	0.98	**1.91 (±0.11)**	11.05	0.86	4.55(±0.27)
	SimFair-P	3.35	0.96	**2.12(±0.11)**	3.24	0.97	1.99 (±0.17)	10.52	0.89	**4.15(±0.23)**

Table 1: AT1: Fairness results on temperature prediction (split by geographic regions in Fig. 3(a)).

NCEP Reanalysis and ASTER Global Emissivity products. We split train-test by: (1) first 5 years as training and last 3 as test; and (2) first 3 years as train and last 5 as test.

Results and Analysis

For the three datasets (AT1, AT2, LST), we evaluate the following methods with the same fairness definition in Eq. (1):

- **BaseNet:** This is the baseline neural network, i.e., FNN or LSTM, without additional fairness consideration.
- **Sim:** BaseNet that uses the physics-based simulation data in pre-training (Jia et al. 2021; Li et al. 2022a). This provides a more generalizable initialization of the model.
- **SimPhy:** This approach uses both simulation-based pre-training as well as physical constraints in loss design to regularize the training and improve generalizability to test samples from different regions (Willard et al. 2020).
- **RegFair:** This is the regularization-based fairness-aware learning (Serna et al. 2020; Yan and Howe 2019), which includes additional fairness-related loss to learn a fairer model on the training dataset.
- **Self-Reg:** A self-training based fair learning framework, which uses predicted labels on the test data to create a pseudo-fairness-loss to adapt to the test area. The predicted labels are dynamically updated during training.
- **SimFair:** Proposed approach (no physical constraints).
- **SimFair-P:** Complete version with physical constraints.

Quality of inverse approximation. Fig. 4 shows the results of inverse approximations for the physical model, where the inversion is necessary since the direction of simulation is often opposite to that of a prediction task. Here we use the CMEM model as an example, which simulates the process from the temperature to many different bands observed by the satellite (i.e., $\mathbf{Y}$ to $\mathbf{X}$). Fig. 4 (a) directly swaps the inputs and outputs of the physical model when training the network, whereas (b) uses the invertible network for the approximation. We can see that the regularization effects from the inversion can effectively reduce the RMSE and improve the approximation quality. For the original physical

(a) X, Y data swap appro   (b) $\mathcal{F}_{\mathcal{M}}^{-1}$: inverse approx.
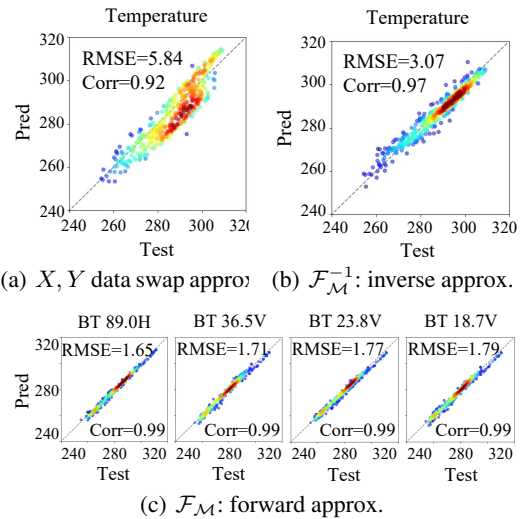

(c) $\mathcal{F}_{\mathcal{M}}$: forward approx.

Figure 4: Example approximation results on CMEM.

model direction, Fig. 4 (c) includes examples of four approximated satellite bands using the invertible network to demonstrate that it works well in both directions.

Fairness results on AT1. The prediction performance and fairness results are shown in Tables 1 to 3, where each table corresponds to a different type of non-overlapping partitioning for training and testing. We show the results of both FNN and LSTM in Table 1 and keep FNN results in Table 2 and Table 3 as their trends are very similar. All results are aggregated over 5 runs. We use three metrics: RMSE, correlation coefficient (Corr.), and fairness (Eq. (1)). **Geographic-region partitions:** As shown in Table 1, the overall trend is that the two variants of SimFair consistently obtained the best fairness results for all three train-test splits. Comparing different splits, SimFair methods have more consistent fairness results, whereas the other methods tend to perform better for the East-West split but worse for the other two splits. It is interesting to note that the prediction perfor-

Model	Train-Test: Hot-Cold			Train-Test: Cold-Hot			Train-Test: Hot-Warm		
	RMSE	Corr.	Fairness	RMSE	Corr.	Fairness	RMSE	Corr.	Fairness
BaseNet	19.7	0.56	12.36(±1.78)	35.95	0.25	21.87(±2.92)	17.5	0.61	12.1(±0.88)
Sim	20.26	0.5	13.71(±1.86)	35.69	0.18	22.19(±1.43)	15.86	0.58	11.54(±1.19)
SimPhy	17.84	0.59	11.28(±1.63)	35.37	0.2	22.32(±3.16)	16.35	0.59	11.83(±1.)
RegFair	19.42	0.57	11.9(±0.29)	35.48	0.23	21.69(±0.8)	16.95	0.63	11.86(±1.32)
Self-Reg	19.32	0.58	12.(±0.55)	36.11	0.24	21.81(±0.61)	16.27	0.65	11.03(±0.54)
SimFair	11.97	0.88	4.8(±1.24)	9.25	0.78	3.61(±0.56)	5.77	0.9	3.62(±0.44)
SimFair-P	12.37	0.91	**4.43(±0.78)**	9.42	0.78	**3.46(±0.22)**	5.62	0.89	**3.56(±0.21)**

Table 2: AT1: Fairness results on temperature prediction (split by temperature zones in Fig. 3(b)).

Model	Train-Test: Train-Test1			Train-Test: Train-Test2			Train-Test: Train-Test3		
	RMSE	Corr.	Fairness	RMSE	Corr.	Fairness	RMSE	Corr.	Fairness
BaseNet	24.02	0.13	14.22(±0.98)	28.58	0.51	19.48(±3.16)	27.87	0.56	15.68(±0.83)
Sim	22.72	0.25	14.24(±1.63)	29.71	0.41	19.97(±2.98)	22.18	0.54	11.59(±2.59)
SimPhy	22.86	0.25	14.02(±2.59)	28.21	0.43	19.35(±2.94)	21.64	0.60	10.88(±0.98)
RegFair	23.71	0.14	14.22(±0.62)	30.66	0.47	20.87(±2.83)	26.77	0.57	15.30(±2.26)
Self-Reg	25.58	0.13	15.57(±1.65)	28.70	0.49	19.26(±2.41)	26.55	0.56	14.79(±1.52)
SimFair	8.42	0.82	5.37(±0.52)	6.55	0.90	3.52(±0.30)	10.01	0.92	5.06(±0.38)
SimFair-P	7.52	0.86	**4.86(±0.31)**	5.94	0.90	**3.26(±0.28)**	9.64	0.91	**4.90(±0.49)**

Table 3: AT1: Fairness results on temperature prediction (split by random state groups in Fig. 3(c)).

(a) AT2 8/4 month split.
(b) AT2 4/8 month split..
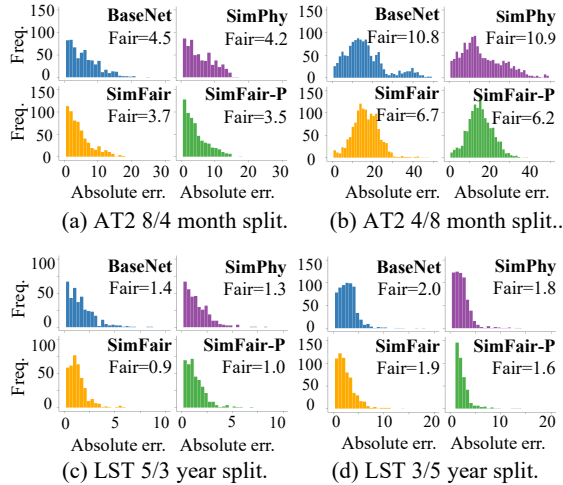(c) LST 5/3 year split.
(d) LST 3/5 year split.

Figure 5: Distributions of absolute errors in AT2 & LST

mance (RMSE) of SimFair also tends to be much better than the other baseline approaches. This is potentially due to the complimentary regularization effects brought by the deeper integration between the deep network and the simulation model using the dual-fairness consistency. **Temperature-zone and state-based partitions:** Comparison results in Table 2 are similar to the previous partitioning, where SimFair continues to show the best performances in both fairness and prediction quality. It is worth noting that the performance is better in cold-to-hot than hot-to-cold scenarios. The reason may be that temperature in colder regions is more stable and contains a narrower distribution, whereas it becomes more dynamic in hotter regions. Table 3 demonstrates that SimFair is able to obtain fairer results in more local regions with a smaller amount of training data.

Fairness results on AT2 and LST. We include additional results to see how well the methods can attach location-based fairness to the same set of locations over different periods. Specifically, Fig. 5 shows the absolute error distributions for the AT2 and LST datasets under various time splits for training and testing: (a) 8 and 4 months; (b) 4 and 8 months; (c) 5 and 3 years; and (d) 3 and 5 years. For AT2, SimFair methods can reduce the variation of the prediction performance, and the 8/4-month split is easier for the methods. Compared to the spatial tasks of AT1 in Table 1, the task here is overall easier based on the performance, as at least the groups (i.e., locations) used in fairness evaluation remain the same. For LST, while BaseNet already performs well, SimFair methods are still able to further improve fairness scores. **AT2/LST: Effects of physics models.** For the results of the two physics-based models, CMEM for AT2 and MODTRAN for LST, SimFair methods perform well with both, showing that the general framework can potentially fit different types of simulations. Comparing CMEM and MODTRAN, the level of improvement is similar.

Conclusions

We proposed a SimFair framework to integrate physical simulation models into fairness-aware learning with inverse physical approximations, a dual-fairness consistency module, and physical constraints to promote fairer solutions. Our results on various simulation models and real datasets show SimFair can effectively improve fairness while keeping a similar (and sometimes better due to potential regularization effects) global performance as the baseline methods. Our future work will expand this to broader application domains and more knowledge- or rule-based simulation models.

Acknowledgments

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*.

Berk, A.; Acharya, P. K.; Bernstein, L. S.; Anderson, G. P.; Lewis, P.; Chetwynd, J. H.; and Hoke, M. L. 2008. Band model method for modeling atmospheric propagation at arbitrarily fine spectral resolution. US Patent 7,433,806.

Berk, A.; Conforti, P.; Kennett, R.; Perkins, T.; Hawes, F.; and Van Den Bosch, J. 2014. MODTRAN® 6: A major upgrade of the MODTRAN® radiative transfer code. In *2014 6th Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing (WHISPERS)*, 1–4. IEEE.

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.

Chen, S.; Xie, Y.; Li, X.; Liang, X.; and Jia, X. 2023. Physics-Guided Meta-Learning Method in Baseflow Prediction over Large Regions. In *Proceedings of the 2023 SIAM International Conference on Data Mining (SDM)*, 217–225. SIAM.

Deo, R. C.; and Şahin, M. 2017. Forecasting long-term global solar radiation with an ANN algorithm coupled with satellite-derived (MODIS) land surface temperature (LST) for regional locations in Queensland. *Renewable and Sustainable Energy Reviews*, 72: 828–848.

Dinh, L.; Sohl-Dickstein, J.; and Bengio, S. 2016. Density estimation using real nvp. *arXiv preprint arXiv:1605.08803*.

Du, M.; Yang, F.; Zou, N.; and Hu, X. 2020. Fairness in deep learning: A computational perspective. *IEEE Intelligent Systems*, 36(4): 25–34.

Goodchild, M. F.; and Li, W. 2021. Replication across space and time must be weak in the social and environmental sciences. *Proceedings of the National Academy of Sciences*, 118(35): e2015759118.

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.

He, E.; Xie, Y.; Liu, L.; Chen, W.; Jin, Z.; and Jia, X. 2023. Physics guided neural networks for time-aware fairness: an application in crop yield prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 14223–14231.

Jia, X.; Xie, Y.; Li, S.; Chen, S.; Zwart, J.; Sadler, J.; Appling, A.; Oliver, S.; and Read, J. 2021. Physics-guided machine learning from simulation data: An application in modeling lake and river systems. In *2021 IEEE International Conference on Data Mining (ICDM)*, 270–279. IEEE.

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.

Kerr, Y. H.; Waldteufel, P.; Wigneron, J.-P.; Delwart, S.; Cabot, F.; Boutin, J.; Escorihuela, M.-J.; Font, J.; Reul, N.; Gruhier, C.; et al. 2010. The SMOS mission: New tool for monitoring key elements ofthe global water cycle. *Proceedings of the IEEE*, 98(5): 666–687.

Kim, C.; and Entekhabi, D. 1998. Feedbacks in the land-surface and mixed-layer energy budgets. *Boundary-Layer Meteorology*, 88(1): 1–21.

Kingma, D. P.; Salimans, T.; Jozefowicz, R.; Chen, X.; Sutskever, I.; and Welling, M. 2016. Improved variational inference with inverse autoregressive flow. *Advances in neural information processing systems*, 29.

Kobyzev, I.; Prince, S. J.; and Brubaker, M. A. 2020. Normalizing flows: An introduction and review of current methods. *IEEE transactions on pattern analysis and machine intelligence*, 43(11): 3964–3979.

Li, R.; Wang, D.; Liang, S.; Jia, A.; and Wang, Z. 2022a. Estimating global downward shortwave radiation from VIIRS data using a transfer-learning neural network. *Remote Sensing of Environment*, 274: 112999.

Li, Y.; Liu, Y.; Bohrer, G.; Cai, Y.; Wilson, A.; Hu, T.; Wang, Z.; and Zhao, K. 2022b. Impacts of forest loss on local climate across the conterminous United States: Evidence from satellite time-series observations. *Science of the Total Environment*, 802: 149651.

Liang, S. 2001. An optimization algorithm for separating land surface temperature and emissivity from multispectral thermal infrared imagery. *IEEE Transactions on geoscience and remote sensing*, 39(2): 264–274.

Peng, S.-S.; Piao, S.; Zeng, Z.; Ciais, P.; Zhou, L.; Li, L. Z.; Myneni, R. B.; Yin, Y.; and Zeng, H. 2014. Afforestation in China cools local land surface temperature. *Proceedings of the National Academy of Sciences*, 111(8): 2915–2919.

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*.

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.

Wang, H.; Mao, K.; Yuan, Z.; Shi, J.; Cao, M.; Qin, Z.; Duan, S.; and Tang, B. 2021. A method for land surface temperature retrieval based on model-data-knowledge-driven and deep learning. *Remote Sensing of Environment*, 265: 112665.

Wang, Z.; Xie, Y.; Jia, X.; Ma, L.; and Hurtt, G. 2023. High-Fidelity Deep Approximation of Ecosystem Simulation over Long-Term at Large Scale. In *Proceedings of the 31st ACM International Conference on Advances in Geographic Information Systems*, 1–10.

Wigneron, J.-P.; Jackson, T.; O'neill, P.; De Lannoy, G.; de Rosnay, P.; Walker, J.; Ferrazzoli, P.; Mironov, V.; Bircher, S.; Grant, J.; et al. 2017. Modelling the passive microwave signature from land surfaces: A review of recent results and application to the L-band SMOS & SMAP soil moisture retrieval algorithms. *Remote Sensing of Environment*, 192: 238–262.

Willard, J.; Jia, X.; Xu, S.; Steinbach, M.; and Kumar, V. 2020. Integrating physics-based modeling with machine learning: A survey. *arXiv preprint arXiv:2003.04919*.

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.

Zhao, C.; Mi, F.; Wu, X.; Jiang, K.; Khan, L.; and Chen, F. 2022. Adaptive Fairness-Aware Online Meta-Learning for Changing Environments. In *Proceedings of the 28th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. ACM.