

Article

Leveraging Remote Sensing Data for Yield Prediction with Deep Transfer Learning

Florian Huber ^{*,†}, Alvin Inderka [†]  and Volker Steinhage ^{*} 

Department of Computer Science IV, University of Bonn, 53121 Bonn, Germany; s6alinde@uni-bonn.de

^{*} Correspondence: huber@cs.uni-bonn.de (F.H.); steinhage@cs.uni-bonn.de (V.S.)[†] These authors contributed equally to this work.

Abstract: Remote sensing data represent one of the most important sources for automatized yield prediction. High temporal and spatial resolution, historical record availability, reliability, and low cost are key factors in predicting yields around the world. Yield prediction as a machine learning task is challenging, as reliable ground truth data are difficult to obtain, especially since new data points can only be acquired once a year during harvest. Factors that influence annual yields are plentiful, and data acquisition can be expensive, as crop-related data often need to be captured by experts or specialized sensors. A solution to both problems can be provided by deep transfer learning based on remote sensing data. Satellite images are free of charge, and transfer learning allows recognition of yield-related patterns within countries where data are plentiful and transfers the knowledge to other domains, thus limiting the number of ground truth observations needed. Within this study, we examine the use of transfer learning for yield prediction, where the data preprocessing towards histograms is unique. We present a deep transfer learning framework for yield prediction and demonstrate its successful application to transfer knowledge gained from US soybean yield prediction to soybean yield prediction within Argentina. We perform a temporal alignment of the two domains and improve transfer learning by applying several transfer learning techniques, such as L²-SP, BSS, and layer freezing, to overcome catastrophic forgetting and negative transfer problems. Lastly, we exploit spatio-temporal patterns within the data by applying a Gaussian process. We are able to improve the performance of soybean yield prediction in Argentina by a total of 19% in terms of RMSE and 39% in terms of R² compared to predictions without transfer learning and Gaussian processes. This proof of concept for advanced transfer learning techniques for yield prediction and remote sensing data in the form of histograms can enable successful yield prediction, especially in emerging and developing countries, where reliable data are usually limited.

Keywords: remote sensing; yield prediction; deep learning; transfer learning; regularization; Gaussian process



Citation: Huber, F.; Inderka, A.; Steinhage, V. Leveraging Remote Sensing Data for Yield Prediction with Deep Transfer Learning. *Sensors* **2024**, *24*, 770. <https://doi.org/10.3390/s24030770>

Academic Editor: Baocheng Zhang

Received: 22 December 2023

Revised: 10 January 2024

Accepted: 22 January 2024

Published: 24 January 2024



Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Yield prediction is a challenging machine learning task. The relations determining the yearly yields are complex and depend on a variety of different factors. To predict yields, we need to measure plant condition and climatic states over an extended period of time. This monitoring is usually very costly and is dependent on domain experts or expensive sensors. A cost-effective solution is provided by using remote sensing data. Utilizing satellite images provided by huge organizations like NASA is free of cost and gives an extensive coverage of the phenological conditions of plants and the climatic state nearly worldwide. The second difficulty for yield prediction is the availability of historic ground truth yield data for training machine learning models. New data points can only be acquired once a year, so to train yield prediction models, we are reliant on historical databases that capture former yield quantities. Although databases like this are often available in developed countries, data are increasingly hard to come by for others. The aim

of this study is to remove the gap in performance for yield prediction between countries with many data available and countries with few data available, especially since countries with fewer options to capture data are often also lacking other resources and could therefore greatly benefit from increased planning capabilities, which naturally arise from better yield prediction. As a solution to this problem, we leverage the concept of transfer learning to limit the number of ground truth yields needed for accurate predictions. The main idea is to use deep learning to adapt to the underlying relations that determine the yearly yields from remote sensing images and ground truth data in countries where data are plentiful, and to transfer this knowledge to yield prediction in countries with limited data. The difficulty of knowledge transfer for deep learning is determined by how close the two domains chosen to transfer knowledge are to each other. In this study, we explore the transfer of knowledge from the prediction of soybean yields in the US to the prediction of soybean yields in Argentina. The main difference between the two domains comes from the location of the two countries in different hemispheres of the globe and the two countries showing different climatic conditions and growth patterns for soy fields. We prepare the two datasets for transfer learning by performing a spatial-temporal alignment. Additionally, we adapt to common transfer learning problems such as catastrophic forgetting and negative transfer by applying the state-of-the-art regularization methods L^2 -SP and Batch Spectral Shrinking (BSS). Furthermore, we find that a Gaussian process attached to our convolutional neural network (CNN) improves the accuracy of the yield prediction even further.

1.1. Related Work

The related work that influences our research can be categorized into three main groups. First, the general developments in transfer learning, as it is mostly used for computer vision tasks. Second, the advances of machine learning for yield prediction. And third, the advances of transfer learning for remote sensing applications.

1.1.1. General Transfer Learning

Transfer learning is a topic of increasing popularity in different research areas. Most commonly, computer vision tasks are solved with the help of deep learning networks that are pretrained on huge task-agnostic datasets, before being fine-tuned to solve specific problems. Plested and Gedeon [1] give an in-depth survey of the state of transfer learning. Due to the versatile applicability of CNNs, the application of transfer learning methods to CNNs represents a well-known study objective. In image classification, transfer learning methods achieve significant success [2]. In this work, we examine whether this success can be translated into yield prediction while we follow the insights of the community. Huh et al. [3] found out that a larger dataset for pretraining results in a better model performance. Furthermore, the number of layers that should be transferred during the training process depends on the similarity between the target and the source dataset [4].

A common problem in transfer learning is the negative transfer of knowledge [5]. Negative transfer occurs when the source dataset used for pretraining and the target dataset are not well related, and transfer learning has a negative impact on the model accuracy. According to Plested and Gedeon [1], greater similarity between the domains of transfer learning improves the ability to transfer knowledge between domains. If the domains are not well aligned, the key to overcoming negative transfer is regularization which restricts the amount of knowledge that can be lost during the fine-tuning step of transfer learning. The L^2 -SP regularization method achieves success in dealing with negative transfer through an L^2 regularization with the origin parameters of the more general model [6]. In our use case, this means that features that are extracted from the patterns of remote sensing data can be preserved during the transfer learning. Similar results can be achieved with DELTA regularization [7], following the idea of only altering CNN channels that are not already useful for the target task. Batch Spectral Shrinkage (BSS) is another regularization method that often successfully eliminates negative transfer by suppressing non-transferable spectral components [8]. Chen et al. [8] report that BSS will

never negatively affect performance in a given dataset and it is therefore also considered within our work to stabilize transfer learning.

Our work is the first to examine the extended use of regularization and transfer learning on hyperspectral remote sensing data presented as histograms, investigating how ideas developed for classical computer vision can help us with remote sensing applications.

1.1.2. Remote Sensing in Yield Prediction

More and more studies are addressing the investigation of yield prediction using machine learning [9,10]. In traditional machine learning, random forests, neural networks, and gradient boosting trees represent the most common methods of yield prediction [10]. Deep learning methods, especially CNNs and LSTMs, provide better feature extraction and generally lead to higher performance [9]. The basis for the data provided is mainly satellite imagery, which in many cases is enhanced with additional features [9]. You et al. [11] propose a novel and widely adopted dimensionality reduction technique to prepare irregularly shaped hyperspectral remote sensing images for regression tasks with deep learning. Remote sensing data from the Moderate-Resolution Imaging Spectroradiometers (MODIS) installed upon the NASA Terra and Aqua satellites are processed toward histograms, and both a CNN and an LSTM network are combined with a deep Gaussian process to perform a soybean crop forecast. The success of the deep Gaussian process built on an LSTM is confirmed by Kaneko et al. [12] for even sparser datasets, again using MODIS observations as input. Moreover, an LSTM trained in several countries without a Gaussian process results in a similar performance in predicting the districts of individual countries as the same model trained only in the country of prediction [12], showing the capabilities of the Gaussian process. Wang et al. [13] use, again, the same MODIS observations and data preprocessing to train an LSTM network to perform simple transfer learning in the form of pre-initialization of weights by a model trained on Argentina to predict soybean yields in Brazil. Similarly, Khaki et al. [14] use a CNN with a common backbone to simultaneously predict corn and soybean yields in the US. Regarding gradient boosting trees, the work of Huber et al. [15] shows on-par performance when compared with deep learning methods to predict soybean yield in the US based on MODIS data, while highlighting the improved explainability of gradient boosting trees and the reduced resources needed to train models. Gradient boosting trees is used along with a comparison including deep learning by Desloires et al. [16] to predict corn yield in the US based on Sentinel-2 satellite images.

The related work clearly suggests to us to also present MODIS observations in the form of histograms to our deep learning architecture. The data sources we consider for our study are also in line with other research in the field, including multiple wavelengths of surface reflections and surface temperature data from the MODIS satellites [10,11,13]. However, none of the previous works aimed to transfer knowledge between countries in different hemispheres of the globe. As different climatic conditions are challenging, this allows us to use the most extensive source of historic yield data in the US [17] to help with yield prediction all around the world.

1.1.3. Transfer Learning for Remote Sensing Applications

Transfer learning for remote sensing applications is a widely investigated topic. A recent extensive survey on transfer learning on remote sensing data is done by Ma et al. [18]. Most commonly, the task of land use/land cover classification is improved by starting from pretrained models. Dastour and Hassan [19] give an in-depth analysis of different deep learning architectures for this task. In general, all CNNs tested are pretrained on the well-known ImageNet dataset [20] and applied to land cover classification on Sentinel-2A images. ResNet50 [21] was found to be the best architecture for transfer learning in this scenario. Li et al. [22] show that a ResNet architecture again works best on the HSRRS dataset for scene classification in urban built-up areas after being pretrained on the ImageNet dataset. Similar results are shown in the work of Alem and Kumar [23] where a ResNet50 architecture pretrained on ImageNet data shows high accuracy for land cover classification

on remote sensing data. Tseng et al. [24] use the pretrained ResNet architecture inside Faster R-CNN as a backbone for rice seedling detection in RGB images outperforming a support vector machine. Chen et al. [25] use the same pretrained architecture, the Faster R-CNN network. First, features are learned on the ImageNet dataset before the network is fine-tuned to detect objects in high-resolution satellite images. Hilal et al. [26] extend the idea of a pretrained ResNet50 for land cover classification by including discrete local binary patterns to the ResNet features for the final classification. In very recent advances, Ma et al. [27] improve the idea of Domain-Adversarial Neural Networks (DANNs) [28], where transfer learning is extended by projecting the input features of each domain into a common subspace. Their proposed partial DANN (PDANN) applies weights to the source samples according to their estimated yield distribution in the target domain and is used to improve the transfer learning of soybean and corn yields between different regions in the US.

Our work is quite distinguishable from the approaches described above. As our input data are processed to be represented by histograms and include multiple hyperspectral image channels and our target is not a classification but a regression, it is not possible to use huge image datasets like ImageNet as a source domain for transfer learning. Even in our source domain, we have relatively small numbers of training data when compared with modern image classification tasks. We need to overcome this issue by using deliberate regularization methods during knowledge transfer.

2. Materials

In this Section, we describe the two sources of data used within the study. We obtained ground truth yields for both countries from the relevant agricultural departments and used remote sensing data provided by the NASA Moderate-Resolution Imaging Spectroradiometer (MODIS) satellites for prediction.

2.1. Yield Data

Soybean yield data are taken from the US Department of Agriculture (USDA) for the US during the period 2010 to 2020, inclusive [17]. In Argentina, yield data are taken from the Ministerio de Agricultura [29] over the 2010/11 to 2020/2021 seasons, inclusive. This leaves us with a total of 8465 data points in the US and 1542 data points in Argentina. An overview of the yield data can be seen in Figure 1. Generally, we see that the yield data in the US and Argentina show similar patterns, with the relative yields in Argentina consistently lower than in the US. Both datasets show a positive linear trend that indicates higher soybean yields in recent history. This is caused mainly by technological innovations that lead to more efficient agriculture [30].

As described in Section 1.1.1, greater similarity between the domains of transfer learning improves the ability to transfer knowledge between domains. Here, the data for the US denote the source domain, and the data for Argentina denote the target domain. In order to keep the similarity of the data as high as possible, the same data cleansing procedure is used for both domains. In our case, data cleansing serves to ensure a sufficiently large input dataset. First, we select the regions of each country where soybeans are grown. This is the northeastern part of the US and the north of Argentina. Then, on the one hand, data points that cannot be assigned to a county due to missing information and counties with no yield specified or with a yield per hectare of zero are removed. On the other hand, all counties that have fewer than 2000 pixels of crop mask are removed. This is necessary to ensure a significant amount of soybean cropland for every data point and to reduce the number of noise from other sources. Figure 2 shows the counties used and discarded as a result in northeast Argentina and the eastern part of the US, leaving us with the main agricultural producers.

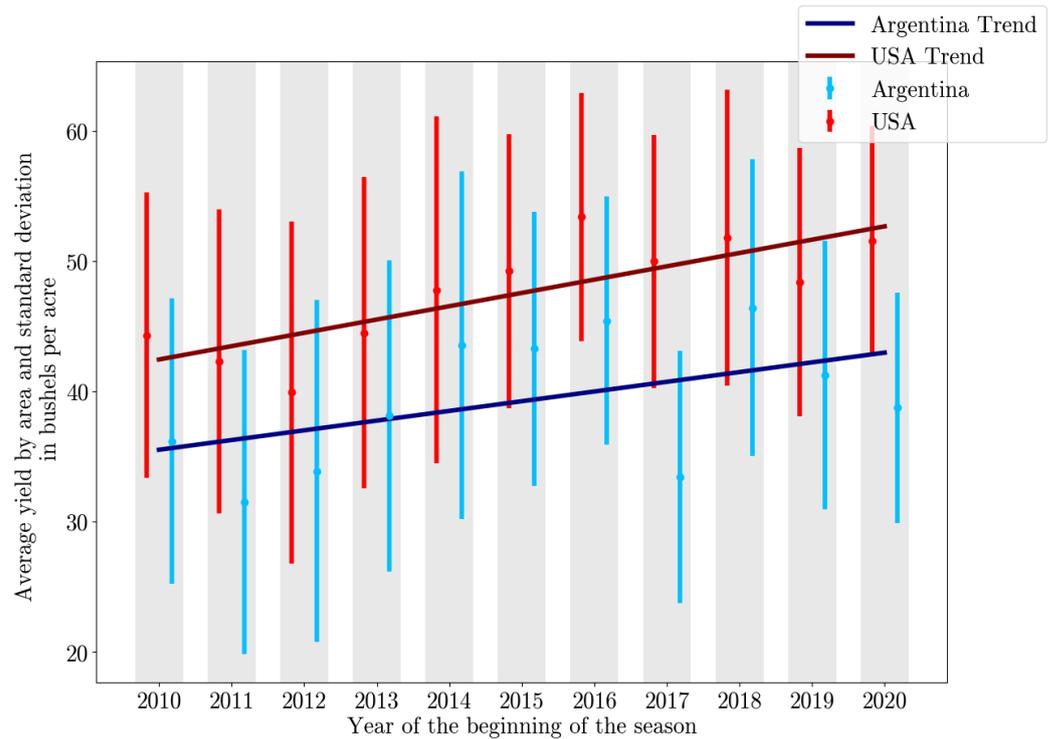


Figure 1. Visualization of the ground truth yield data in the US (red) and Argentina (blue). Average yields are indicated by the points and standard deviation is shown by the relevant lines, both in bushels per acre. The dark red and blue lines indicate the linear trend in the data.

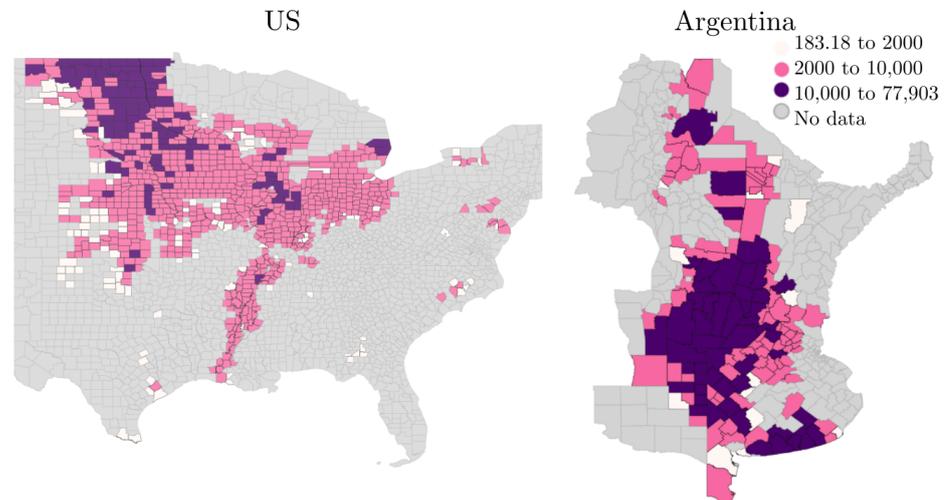


Figure 2. Average number of cropland pixels over the 11 years from 2010 to 2020 in the eastern part of the US and the northeast of Argentina. The pink to purple areas show the counties used. No cropland pixels or yield data were found in the gray areas, and too few in the white areas.

2.2. Remote Sensing Data

The use of satellite imagery for yield prediction is one of the main sources of reliable and cost-effective yield prediction. The MODIS satellite images we used were acquired on NASA's Terra and Aqua satellites and are listed in the Google Earth Engine catalog. We used three different types of remote sensing data.

The surface reflections of light are captured in a range starting from the visible spectrum (459–479 nm) to the invisible infrared spectrum (2105–2155 nm) and are captured by 7 different channels. Each channel is referred to as a band in the context of remote sensing images. Specifically, we used version 6.1 of the MOD09A1 dataset [31] for surface

reflectance corrected for atmospheric conditions. The final product is an 8-day composite of the images taken by the satellites with a resolution of 500 m by 500 m.

To monitor climatic conditions, we use a remote sensing dataset consisting of measured surface temperatures during day- and nighttime. The dataset used is version 6.1 of the MYD11A2 dataset [32]. The original resolution of 1 km by 1 km is resampled to match the resolution of 500 m by 500 m of the surface reflection data.

A cropland mask is used to filter out pixels that are not related to crop growths. Version 6 of the MCD12Q1 dataset [33] offers a yearly classification of several types of land cover with a resolution of 500 m by 500 m. Areas consisting of more than 60% cultivated fields are classified as croplands and will be the focus of our study.

Lastly, we crop the satellite images for the individual counties. Therefore, we require district assignments labeled with the county names. In the United States, the TIGER dataset from the United States Census Bureau [34] is used for this purpose. In Argentina, we use the county (Departamentos in Argentina) mapping used by Wang et al. [13], which contains all counties relevant to soybean cultivation.

3. Methods

We use multiple different tools to achieve the best accuracy for yield prediction on a smaller dataset. To allow for successful transfer learning, we align our two data sources first, before we apply multiple regularization techniques, and subsequently append a Gaussian process to recover the information lost due to histogramization. An overview of the process is shown in Figure 3.

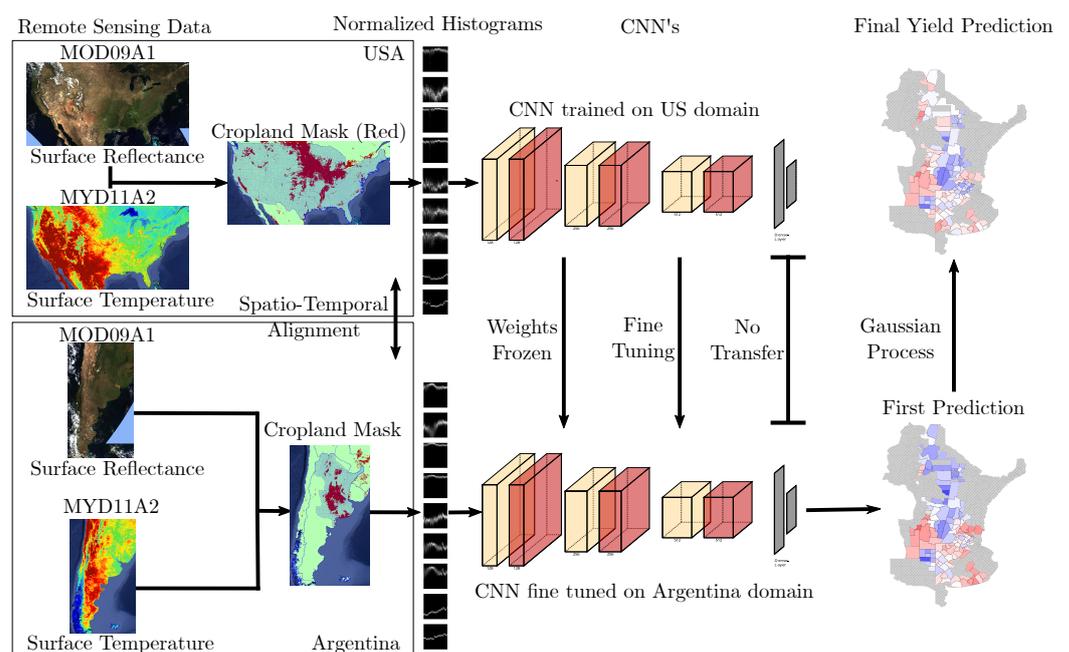


Figure 3. Overview of our approach for leveraging remote sensing data for yield prediction with deep transfer learning. The two boxes on the left side show the kind of input data that we use, consisting of surface reflectance data, surface temperature data, and the cropland mask for each of the two domains (Section 2). The data is spatio-temporally aligned as preparation for transfer learning (Section 3.1), before being processed toward normalized histograms (Section 3.2). The middle section of the figure shows some of the transfer learning techniques used: frozen weights, fine-tuning with regularization, and the initialization of the dense layer at the end of the network (Section 3.2). Lastly, on the right-hand side, we show an example of refining the prediction results by applying a Gaussian process (Section 3.3).

3.1. Spatio-Temporal Alignment

The geographical distance between the US and Argentina accounts for some differences in soybean cultivation in both countries. Specifically, we need to adjust for different crop growth cycles. In the US, we monitor the growing conditions from March 23 to December 4. In Argentina, due to different climatic conditions, we monitor the crop growth cycle from November 26 to August 9. Both periods extend beyond harvest time, so that we can leverage the capabilities of our models to determine important input data themselves. As soybean yield prediction is especially valuable during the early stages of soybean growth, we examine a second in-season forecast to test our models. For this case, the time period only contains the first 14 of the 8-day intervals. The time periods used, as well as the underlying crop calendars [35], can be seen in Figure 4.

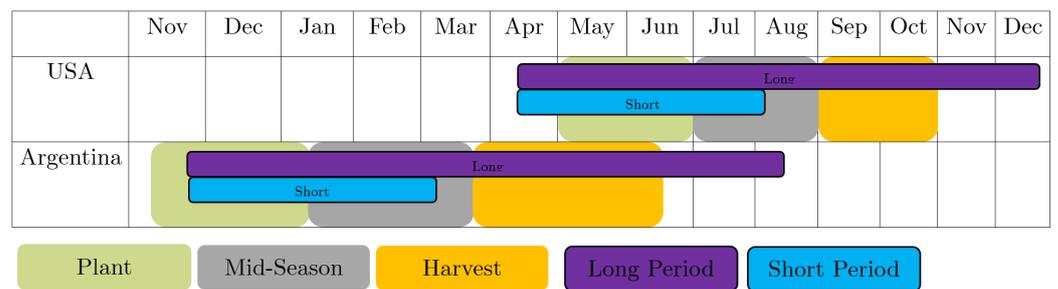


Figure 4. Crop calendar of Argentina and USA (data from USDA [35]). The data of the long period are shown in purple. The data of the short period are shown in blue.

After data cleansing, a histogramization is performed, as was first done by You et al. [11]. For each band and each satellite image, the pixels masked as cropland within a county are summed for each record. Subsequently, these sums are transformed into a normalized histogram with 32 bins of the frequency distributions of the records in the vertical and the associated time in the horizontal. The resulting data point consists of 9 times 34 histograms, a county name, and a season year.

3.2. Model Design and Transfer Learning Techniques

At the center of our predictive model is a CNN, as was first used for yield prediction by You et al. [11]. As a baseline for our implementation, we started from the work of Tseng [36]. The CNN consists of 6 convolutional layers and the following dense layer. As a result of histogramization of the input data, we need an alternative solution to the usually used pooling layers in conjunction with the convolutional layers. The solution is convolutional layers with a stride of 2 to prevent locational invariance due to pooling. In addition, we use dropout layers, early stopping, and batch normalization when training the CNN. Primarily, to apply transfer learning, this CNN is trained in the US domain and will be referred to as the US model in the following. The most natural way to apply transfer learning is to fine-tune the US model in the Argentine domain. This corresponds to initializing the weights of the CNN for the Argentine domain with the weights of the US model. By assuming a similar distribution of the US domain and the Argentine domain, the starting point of the Argentine model will thus be chosen closer to the desired model, and the hypothesis space will be constrained by limiting the hyperparameters of the training. Furthermore, since it is known through extensive research that later layers have greater domain-specific significance while early layers extract general properties, it may be useful to distinguish the approach based on layer depth [1]. The first way to treat layers differently is to freeze some of them. The frozen layers keep their weights unchanged during fine-tuning and are effectively not trained directly on the new domain. In our case, this means that the frozen layers are trained only on the larger US domain to contribute to a prediction on the Argentine domain. Due to the high generalization of the front layers, this procedure is applied to the front convolutional layers and thus adopts the basic framework of the US model for feature extraction. This serves, in particular, to avoid

catastrophic forgetting [37]. Catastrophic forgetting is one of the two major problems in transfer learning and is the tendency of neural networks to abruptly lose the knowledge of previous tasks when learning a new task [38]. Later convolutional layers are initialized with the weights of the US model but retain their ability to adapt to the Argentine domain. To name this retraining on top of the source domain, we overload the term fine-tuning, which also refers to the complete retraining of the net on the Argentine domain on top of the US model. Last, we use the ability to completely retrain layers independent of the US model so that they can fully adapt to the Argentine domain. This complete retraining is used in the dense layers to account for their strong specialization to the target domain associated with their depth.

Transfer-specific regularization methods are suitable to further limit catastrophic forgetting and to avoid negative transfer. Regularization influences CNN training by an additional summand Ω to the loss. The widely known L^2 regularization, also known as weight decay, penalizes large weights in the CNN. In this process, as seen in Equation 1, the vector w of weights of the CNN is normalized, squared, and parameterized by α to set the regularization strength:

$$\Omega(w) = \frac{\alpha}{2} \|w\|_2^2. \quad (1)$$

Based on this, Xuhong et al. [6] modifies the starting point of this regularization so that instead of a deviation of the weights from zero, a deviation of the weights from the US model is penalized. Thus, the hypothesis space is again restricted to the surroundings of the US model, making catastrophic forgetting less likely. Since for the deviation of the target model from the source model an equal structure of both networks is required, Xuhong et al. [6] extends the model with the option to regularize non-transferable and especially newly added parts of the model with the L^2 regularization. Equation (2) shows the corresponding formula, where α and β set the strengths of the regularization of the constant and newly added parts:

$$\Omega(w) = \frac{\alpha}{2} \|w_S - w_S^0\|_2^2 + \frac{\beta}{2} \|w_{\mathcal{S}}\|_2^2, \quad (2)$$

where w_S is a weight vector of the constant structures of the target model, w_S^0 is a weight vector of the constant structures of the source model, and $w_{\mathcal{S}}$ is the weight vector of the newly added structures of the target model. We exploit the application of L^2 regularization for greater adaptability of the dense layer, so that dense layers are effectively regularized with L^2 regularization even when L^2 -SP regularization is applied. This means in particular that the vectors of the constant structures are equal to the convolutive layers and the newly added structures are equal to the complete reinitialization of the dense layers.

Negative transfer refers to a loss of performance due to the transfer of knowledge and occurs due to a lack of transferability of some features caused by differences in the domains [1,8]. To reduce negative transfer, we use the Batch Spectral Shrinkage (BSS) regularization method. Chen et al. [8] observed that large datasets lead to highly generalized models, while at the same time these models suppress small singular values of the extracted features. To exploit this behavior of generalized models, Chen et al. [8] used artificial suppression of small singular values. Specifically, they introduced the concept of relative angles between domains that can be used to measure the transferability of eigenvectors in the weight matrices. Subsequently, they found out that in later layers of the network only eigenvectors corresponding to relatively large eigenvalues produce small angles and hence are well suited for knowledge transfer. Here, the feature maps f_i of the input x_i are vectorized and aggregated into a feature matrix $F = [f_1 \dots f_b]$ per batch of size b . Subsequently, the singular value decomposition is applied to this feature matrix,

and the k smallest singular values are used to suppress the associated poorly transferable eigenvectors. The BSS regularization is thus given by

$$\Omega(F) = L_{bss}(F) = \eta \sum_{i=1}^k \sigma_{-i}^2,$$

where η as a hyperparameter specifies the strength of the regularization and k specifies the number of smallest singular values σ_{-i} to be penalized. k is set to 1 according to Chen et al. [8] and the regularization is completely adjusted via η . The use of BSS and L²-SP has the advantage that both regularization methods can be applied simultaneously [8].

3.3. Deep Gaussian Process

To extend the CNN to include spatio-temporal features, You et al. [11] propose the use of a Gaussian process on top of the CNN. For a data point x , the deep Gaussian process uses the input of the last dense layer of the CNN $h(x)$ and a Gaussian process $f(x) \sim \mathcal{GP}(0, k(x, x'))$ to make a harvest prediction:

$$y(x) = f(x) + h(x)^T \beta$$

$f(x) \sim \mathcal{GP}(0, k(x, x'))$ is distributed from a Gaussian process with zero mean and covariance defined by the squared exponential kernel

$$k(x, x') = \sigma^2 \exp\left(\frac{\|g_{loc} - g'_{loc}\|_2^2}{2r_{loc}^2} - \frac{\|g_{year} - g'_{year}\|_2^2}{2r_{year}^2}\right) \sigma_e^2 \delta_{g, g'}$$

where g_{loc} and g_{year} indicate the spatial and temporal data of the associated data point x and $\delta_{g, g'}$ is the Kronecker delta as noise factor over $g = (g_{loc}, g_{year})$ parameterized by σ_e . σ , σ_e , r_{loc} , and r_{year} are hyperparameters. In summary, $f(x)$ distributed from a Gaussian process is a collection of random variables with a joint Gaussian distribution. The next summand for the prediction is given by a set of basis functions $h(\cdot)$, corresponding to the last layer, and a random variable β . β is distributed from a normal distribution with the weight vector of the last layer as mean and $\sigma_b I$ as variance with the hyperparameter σ_b .

The choice of β and $h(x)$ results in a distribution around the CNN prediction. The second summand then results from the covariance of the training data with the test data formed over the RBF kernel multiplied by the differences of the prediction $h(x)^T \beta$ on the training data and the ground truths of the training data, which are weighted again with the help of their covariance. This forms $f(x) \sim \mathcal{GP}(0, k(x, x'))$, which reduces the error values depending on the local and temporal relationships of the underlying data. For a deeper insight into the calculations, we recommend a look at the implementation of Tseng [36] and the explanations of Williams and Rasmussen [39].

3.4. Performance Metrics and Hyperparameter Tuning

Two main metrics are used to quantify the performance of the prediction. We measure the overall error by applying the Root Mean Square Error (RMSE), calculated as follows:

$$\text{RMSE}(\hat{\mathbf{y}}, \mathbf{y}) = \sqrt{\text{MSE}(\hat{\mathbf{y}}, \mathbf{y})} = \sqrt{\frac{1}{n} \sum_{i=1}^n (\hat{y}_i - y_i)^2}. \quad (3)$$

Here, \mathbf{y} represents the ground truth values and $\hat{\mathbf{y}}$ the prediction values. Furthermore, we measure the variability in the target variable explained by the models with the coefficient of determination (R^2):

$$R^2(\hat{\mathbf{y}}, \mathbf{y}) = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2}, \quad (4)$$

where \bar{y} is the mean of the ground truth values y . All models are evaluated over four runs each to account for the random factors of the model.

The models are trained from season 2010, respectively, 2010/2011 until one season before the prediction. The hyperparameters of the model are determined using the Optuna hyperparameter optimization framework [40]. Within Optuna, the number of adjusted hyperparameters is kept low to save computational resources and is estimated using a tree-structured parzen estimator algorithm. The search space of the hyperparameters was determined using Plested and Gedeon [1]. The remaining hyperparameters, such as those of the Gaussian process, are taken from Tseng [36]. Due to the required US data for information-breach-free application of transfer learning, the hyperparameters of the US model are derived using the three seasons from 2012 to 2014, resulting in an average number of 2457 training data points during hyperparameter tuning. For the Argentine model and the transfer learning parameters shown in Table 1, the seasons 2015/16 to 2017/18 are used to estimate the hyperparameters, resulting in an average number of 776 training data points during this step. The test data contain the years from 2018/19 to 2020/21 and contain 1268 training data points on average for Argentina and 7008 data points on average for the US. For every seasonal forecast in Argentina, we first train the US model with data from 2010 up to the previous season. This ensures that we have no information breach regarding temporal dependencies caused by transfer learning but always have the same data available that would be accessible in a real-world scenario.

Table 1. Hyperparameter configuration for transfer learning-related hyperparameter. Hyperparameters are either tuned via the tree parzen estimator (TPE) within the Optuna Framework or estimated empirically. The optimal hyperparameters are rounded and refer to the application of all methods in the long period, as shown in the last row of Table 2.

| Hyperparameter | Tuning Method | Optimal Parameter |
|-----------------------------|---------------|--------------------------|
| Number of frozen layer | Empirical | 4 \in [0, 6] |
| US init. of weight | Empirical | True \in [True, False] |
| L ² -SP α | TPE | 0.23 \in [0, 1] |
| L ² -SP β | TPE | 0.77 \in [0, 1] |
| BSS η | TPE | 0.07 \in [0, 1] |
| BSS k | Empirical | 1 \in [0, 1] |

Table 2. Average RMSE in bu/ac and R² as fraction of 1, of different model configurations for the long period. The long period contains satellite images beyond the harvest time and thus includes the full crop growth cycle. Bold numbers indicate the best scores. A full display of the results is shown in Appendix A.

| | RMSE ↓ | R ² ↑ | RMSE + GP ↓ | R ² + GP ↑ |
|--|--------|------------------|-------------|-----------------------|
| USA: CNN | 5.94 | 0.583 | 6.81 | 0.439 |
| Argentina without transfer | 7.47 | 0.442 | 6.76 | 0.554 |
| Argentina + freezing | 6.94 | 0.526 | 6.85 | 0.547 |
| Argentina + freezing and L ² -SP | 6.80 | 0.550 | 6.25 | 0.618 |
| Argentina + freezing, L ² and BSS | 7.05 | 0.516 | 6.43 | 0.593 |
| Argentina + freezing, L ² -SP and BSS | 7.07 | 0.511 | 6.31 | 0.608 |

4. Results

To analyze the capabilities of transfer learning for yield prediction, we analyze the accuracy of yield prediction models built with different configurations of regularization techniques and use of the Gaussian process. First, we examine the models over the long

period, including the full crop growth cycle, before conducting experiments on the performance of an in-season prediction.

4.1. Full Growth Cycle Prediction

We compare the application of different transfer learning methods on the long period covering the full crop growth cycle, with a summary of the results being shown in Table 2. Additionally, a full overview of the results broken down to the performance each year is displayed in Appendix A. To account for the influence of random parameters, all values are determined over four runs each. For the Argentine models, the application of the Gaussian process improves the accuracy in all cases by up to 10% in terms of RMSE and by up to 25% in terms of R^2 . Transfer learning approaches with regularization improve the results further. First, we examine the application of an initialization with the US model for all six layers with a freezing of the parameters of the first four layers. We see initial improvements compared to a model without any transfer learning applied that diminish after the concatenation of the Gaussian process. Incorporating the regularization techniques explained in Section 3.2 stabilizes the training and subsequently gives the best results. The overall best configuration is an initialization with the weights of the US model, a freezing of the first four layers, a fine-tuning using L^2 -SP regularization, and lastly the application of the Gaussian process. Using BSS as an additional regularization slightly decreases the average accuracy but stabilizes the training process, as can be seen in Figure 5.

Figure 5 shows the distribution of the RMSE in bu/ac for each method. The first two boxplots show the basic drop in RMSE when the Gaussian process is applied. The US initialization with freezing subsequently causes a wide dispersion of the error values. While BSS places the center of these scattered error values at a low level, L^2 -SP causes the values to be centered at a lower value, indicating a lower average error. The simultaneous use of BSS and L^2 -SP further reduces the scattering of error values, indicating that the predictions provide greater reliability.

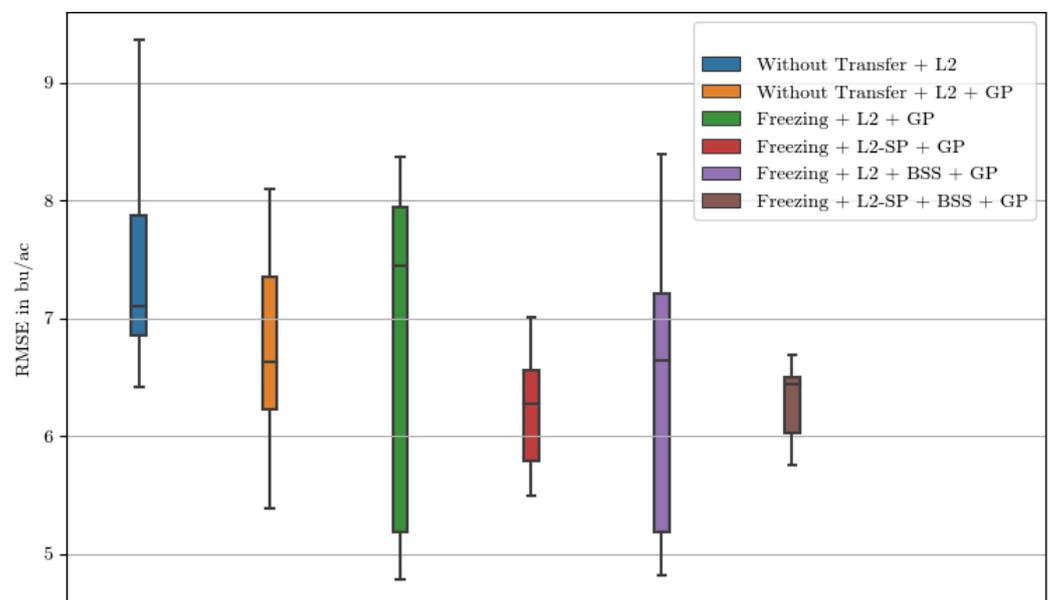


Figure 5. Distributions of the averaged district errors of all runs in all test years as boxplots. In this case, freezing implies the initialization of the 6 convolutional layers by the US model without further fine-tuning of the first 4 layers.

4.2. In Season Prediction

As described in Figure 4, we examine a second shorter time frame of available information for our prediction models. A model capable of inferring the estimated yield long before harvest has a very high value for crop management. Table 3 shows a summary of the prediction results in the short period. The full breakdown of the results for every

individual year can be found in Appendix B. Here, the satellite images no longer include the harvest and end before the first harvests begin. Training, evaluation, and testing are performed exactly as in the long period. We used only the first 14 satellite images, instead of 32 as done previously. This is reflected in a reduced performance compared to the long period. At the same time, it can be seen that transfer learning methods, especially with simultaneous application of the Gaussian process, result in an even greater increase in performance.

Table 3. Average RMSE in bu/ac and R^2 as fraction of 1, of different model configurations for the short time period. Satellite image coverage ends before the start of the harvests. Bold numbers indicate the best scores. A full display of the results is shown in Appendix B.

| | RMSE ↓ | R^2 ↑ | RMSE + GP ↓ | R^2 + GP ↑ |
|---|--------|---------|-------------|--------------|
| USA: CNN | 7.12 | 0.394 | 7.00 | 0.414 |
| Argentina without transfer | 9.36 | 0.140 | 8.37 | 0.314 |
| Argentina + freezing, L^2 -SP and BSS | 8.20 | 0.349 | 6.92 | 0.537 |

5. Discussion

With our experiments, we are able to confirm the capabilities of CNNs for yield prediction when they are trained on MODIS observations that are preprocessed to resemble histograms. The approach is tested within the literature for several yield prediction scenarios, as described in Section 1.1. The addition of the Gaussian process for predicting US yields is beneficial for the short prediction period, while it decreases accuracy for the long prediction period. For the Argentine yield prediction without transfer learning, the Gaussian process improves the results by ca. 10% in terms of RMSE for the long period and the short period. Similar results are discovered by Kaneko et al. [12] where a Gaussian process improves the results of predictions based on small datasets up to a level similar to that which can be observed when training with more data. The same effects are shown when examining the first step of transfer learning for Argentina, where the CNN weights are initialized with the US model's weights and the first four layers are frozen. Although the improvement without a Gaussian process is about 7% in terms of RMSE, the results are worse when the Gaussian process is included. Obtaining worse results with a machine learning model after application of knowledge transfer, like we have in the case of our model when the Gaussian process is included, can be described as negative transfer and is well anticipated in the literature (Section 2.1). As yield prediction is a research area that suffers from data scarcity, the selection of a source domain for pretraining is prone to a known tradeoff. On the one hand, more data for pretraining improves the results, while on the other hand, less similarity between the domains endangers the knowledge transfer. Our experimental setup tends to emphasize the amount of training data for pretraining over the similarity of the domains, as in real-world yield prediction applications it is often not possible to produce more training data close to the target domain. Despite this choice, the first transfer experiments indicate that features extracted from remote sensing data can be transferred similarly to those obtained in many computer vision tasks. This claim is supported by the fact that we were able to freeze the first four layers in our CNN which commonly condense high-level features from the data and improve the model accuracy without the Gaussian process.

The negative transfer that is prevalent when we evaluate the models with the Gaussian process can be addressed by regularization techniques. Our results support the claim that regularization techniques designed for commonly used image features have similar effects on remote sensing data presented as histograms. The L^2 -SP regularization together with the Gaussian process gives the best results in terms of average RMSE and R^2 , removing the negative transfer that occurred without regularization. As indicated by Chen et al. [8], the inclusion of BSS gives us a small decrease in average performance by 2.8% in terms of RMSE but stabilizes the knowledge transfer. As can be seen in Figure 5, the worst error

values are closer to the average RMSE than in any other constellation, which makes us recommend the combination of layer freezing, L^2 -SP, and BSS for transfer learning tasks, including remote sensing data represented by histograms. The same constellation also works well when considering the short prediction period (Table 3).

Regarding the specific transfer of knowledge between the US and Argentina, we can analyze Figure 6, providing an overview of the distribution of errors in Argentina after applying different procedures for the long prediction period. The graphics shown are taken from the first run of 2019 as an example. The area shown includes soybean acres in northeastern Argentina. Training or test data are missing for the gray areas. On the basis of the lower saturation of the districts, it can be seen that the Gaussian process reduces the error values. The application of transfer learning leads to low saturation in many districts in advance and a more balanced distribution of overestimates and underestimates. At the same time, especially in transfer learning, some districts with a worse forecast are visible. We need to assume that the relations learned in the US domain do not carry over to these districts, due to local dependencies or different crop management techniques. All in all, we find the regularized transfer learning to be beneficial and generally reduce the error on our datasets.

Within the wider literature context, the first parallel between our work and related work also considering yield prediction with remote sensing data is the use of MODIS satellite data as a primary data source, as is done, for example, by [11,13,14]. While all those works report good results, it is worth mentioning that alternative data sources exist. The authors of [41] also use the Sentinel-2 satellite to achieve state-of-the-art results, exploiting the fact that the images are available at a higher resolution. Furthermore, the literature is beginning to investigate the use of non-fixed time steps during histogram creation as input for machine learning procedures [16], which may be useful to increase alignment between the two domains used for transfer learning in the future. When considering the literature for yield prediction in general, our results improve on state-of-the-art performance, as we can deduce from our comparison to successfully deployed deep learning architectures [11] trained and evaluated in our data. This increase in performance comes from enabling the US, as the biggest repository for ground truth yield data, as the source domain for transfer learning, even for a country in a different hemisphere. This builds on the work of [13], where transfer learning is first used to improve yield prediction via transfer learning, but the two countries used, i.e., Argentina as source and Brazil as target, are much more closely related than in our case.

However, it is important to mention that the improvements come at the price of many additional hyperparameters that have to be tuned. Tuning hyperparameters in a deep learning context is always difficult, since the impact of a hyperparameter can mostly only be observed after a significant number of computations. This makes it so that research often turns to empirical values or educated guesses. Regularization and transfer learning include an additional six important hyperparameters to adjust: L^2 -SP, BSS, the number of frozen layers, and the initialization of the non-frozen layers with parameters of the source model, as can be seen in Table 1. In our experiments, the number of four frozen layers indicates that many features learned from remote sensing data in our target domain can be directly transferred to the new task. The high transferability is also indicated by the advantageous initialization with the source weights for the non-frozen layers. The L^2 -SP hyperparameter α with a value of 0.23 quantifies the punishment for altering the weights of the source models. This value being relatively low means that the non-frozen layers must be able to be highly adjustable to the new task, hinting that the yield-related patterns utilizing the frozen features of the first layers are quite different for both our yield prediction tasks. The high L^2 -SP hyperparameter β is the standard L^2 punishment for high weights suppressing overfitting. The BSS hyperparameters η and k and their respective values 0.07 and 1 indicate that the strength of the regularization is relatively low and the smallest singular value is penalized. For all these hyperparameters, small adjustments can

alter the models' performance, increasing the risk of a bad model due to careless handling of the hyperparameters compared to a simpler model without transfer learning.

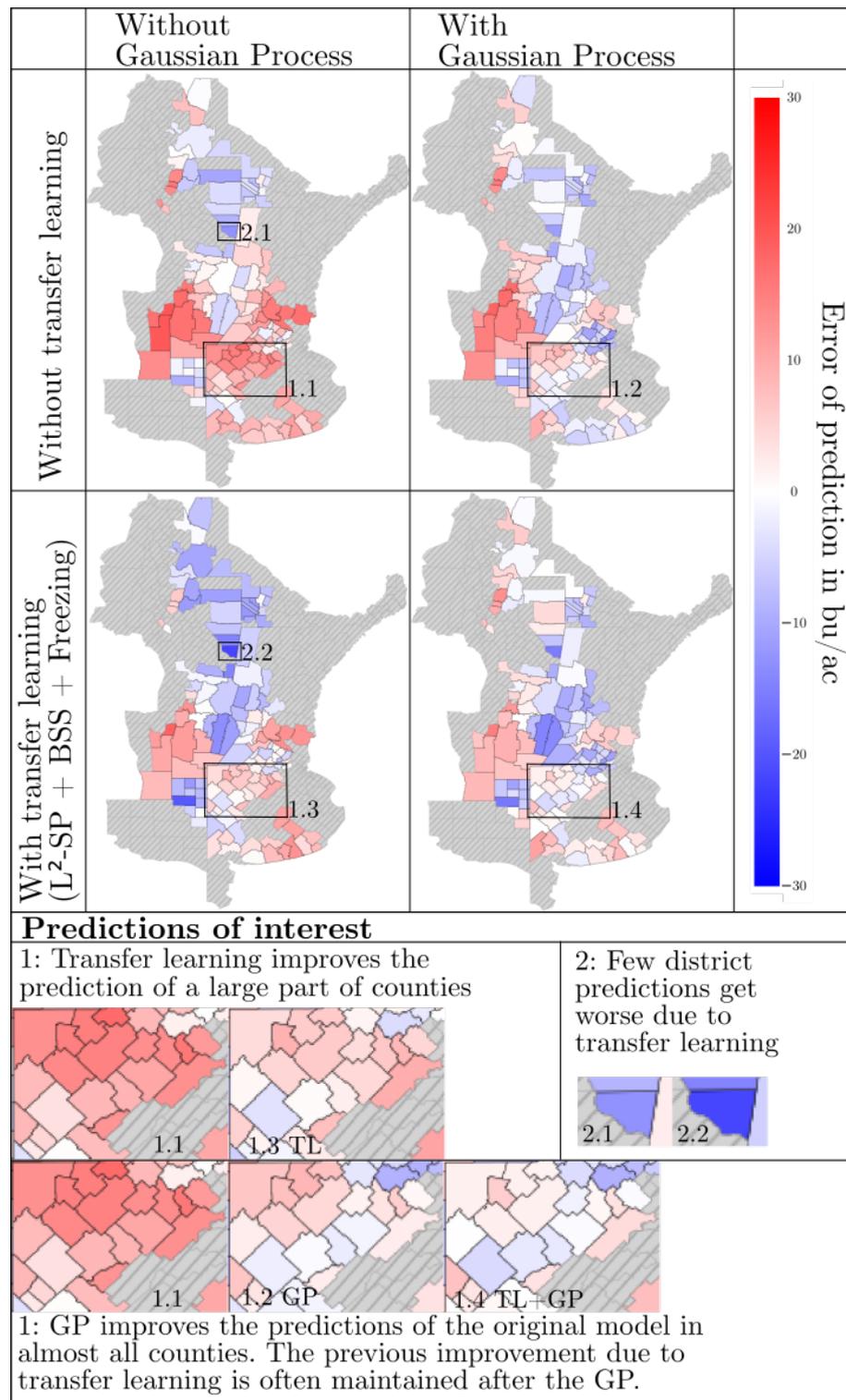


Figure 6. Comparison of different results in the first run of the long period of 2019. Districts without ground truth are grayed out. An overestimate is indicated in red and an underestimate is indicated in blue. The lower the saturation, the better the prediction.

6. Conclusions

Overall, our proof of concept shows that transfer learning can lead to improved crop prediction using CNNs, in particular through a joint application of several methods with careful determination of the hyperparameters. In addition to the usual procedures of initializing the weights by a model trained on a larger dataset and fixing these weights, transfer-specific regularization methods with simultaneous application of the Gaussian process particularly lead to improved prediction. For prediction over the full crop growth cycle, performance was improved by an RMSE of 0.51 bu/ac compared to the CNN with Gaussian process and without transfer learning by using a US initialization of the weights, fixing four convolutional layers, and applying L^2 -SP. The same procedure results in a performance increase of 1.35 bu/ac for the in-season prediction. The BSS regularization method is not directly beneficial to model performance, as is the case for L^2 -SP, but it helps to stabilize the training process and prevents a particular bad prediction. We see that we can utilize transfer learning methods designed for other deep learning tasks, such as image classification, for yield prediction with remote sensing data in the form of histograms as input. Future work includes investigations in other countries, especially with even smaller yield databases, to further confirm these methods. In general, our approach for leveraging remote sensing data for yield prediction with deep transfer learning shows that:

- Spatio-temporal alignment can be performed even between two varying remote sensing data sources to allow for transfer learning.
- The capabilities of transfer-specific regularization methods L^2 -SP and BSS together with Gaussian processes for transfer learning translate to the context of yield prediction and hyperspectral remote sensing data in the form of histograms.
- Regularized transfer learning can improve yield predictions in regions where fewer data are available and should be considered as an alternative to state-of-the-art approaches, especially for smaller study areas.

Author Contributions: Conceptualization, F.H., A.I. and V.S.; Funding acquisition, V.S.; Methodology, F.H., A.I. and V.S.; Project administration, V.S.; Software, F.H. and A.I.; Supervision, V.S.; Validation, F.H. and A.I.; Visualization, F.H. and A.I.; Writing—original draft, F.H., A.I. and V.S.; Writing—review and editing, F.H., A.I. and V.S. All authors have read and agreed to the published version of the manuscript.

Funding: This work was partially carried out within the project “Artificial Intelligence for innovative Yield Prediction of Grapevine” (KI-iRepro). The project is supported by funds of the Federal Ministry of Food and Agriculture (BMEL) based on a decision of the Parliament of the Federal Republic of Germany. The Federal Office for Agriculture and Food (BLE) provides coordinating support for artificial intelligence (AI) in agriculture as funding organization, grant number FKZ 28DK128B20.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: All data used within this research are publicly available and cited accordingly.

Acknowledgments: We thank Timm Haucke and Frank Schindler for fruitful discussions and the proofreading of our manuscript.

Conflicts of Interest: The authors declare no conflicts of interest. During the preparation of this work, the authors used Writefull Premium and DeepL in order to improve readability. After using this tool/service, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

Appendix A

Table A1. Yearly average RMSE in bu/ac and R^2 as fraction of 1, of different model configurations for the long time period. The long period contains satellite images beyond the harvest time and therefore includes the full crop growth cycle. Each year is averaged over 4 runs and the overall average is averaged over each individual run.

| USA: CNN | | | | |
|---|------|-------|-----------|------------|
| Year | RMSE | R^2 | RMSE + GP | $R^2 + GP$ |
| 2015 | 5.57 | 0.639 | 6.89 | 0.290 |
| 2016 | 6.64 | 0.375 | 8.47 | −0.071 |
| 2017 | 4.85 | 0.737 | 5.12 | 0.801 |
| 2018 | 6.77 | 0.567 | 7.04 | 0.479 |
| 2019 | 5.24 | 0.635 | 6.04 | 0.545 |
| 2020 | 6.58 | 0.545 | 7.28 | 0.480 |
| AVG | 5.94 | 0.583 | 6.81 | 0.439 |
| Argentina without Transfer | | | | |
| Year | RMSE | R^2 | RMSE + GP | $R^2 + GP$ |
| 2018 | 7.44 | 0.567 | 7.65 | 0.543 |
| 2019 | 8.15 | 0.360 | 6.70 | 0.572 |
| 2020 | 6.83 | 0.400 | 5.91 | 0.549 |
| AVG | 7.47 | 0.442 | 6.76 | 0.439 |
| Argentina + Freezing | | | | |
| Year | RMSE | R^2 | RMSE + GP | $R^2 + GP$ |
| 2018 | 7.54 | 0.557 | 8.13 | 0.487 |
| 2019 | 7.15 | 0.513 | 7.48 | 0.468 |
| 2020 | 6.13 | 0.507 | 4.95 | 0.685 |
| AVG | 6.94 | 0.526 | 6.85 | 0.547 |
| Argentina + Freezing, L^2 -SP | | | | |
| Year | RMSE | R^2 | RMSE + GP | $R^2 + GP$ |
| 2018 | 7.55 | 0.552 | 6.80 | 0.640 |
| 2019 | 7.23 | 0.503 | 6.30 | 0.623 |
| 2020 | 5.62 | 0.594 | 5.64 | 0.592 |
| AVG | 6.80 | 0.550 | 6.25 | 0.618 |
| Argentina + Freezing, L^2 and BSS | | | | |
| Year | RMSE | R^2 | RMSE + GP | $R^2 + GP$ |
| 2018 | 7.82 | 0.523 | 6.71 | 0.650 |
| 2019 | 7.83 | 0.414 | 7.62 | 0.445 |
| 2020 | 5.51 | 0.610 | 4.96 | 0.684 |
| AVG | 7.05 | 0.516 | 6.43 | 0.593 |
| Argentina + Freezing, L^2 -SP and BSS | | | | |
| Year | RMSE | R^2 | RMSE + GP | $R^2 + GP$ |
| 2018 | 7.61 | 0.548 | 6.58 | 0.664 |
| 2019 | 7.55 | 0.459 | 6.44 | 0.606 |
| 2020 | 6.06 | 0.527 | 5.90 | 0.553 |
| AVG | 7.07 | 0.511 | 6.31 | 0.608 |

Appendix B

Table A2. Yearly average RMSE in bu/ac and R^2 as fraction of 1, of different model configurations for the short time period. Satellite image coverage ends before the start of harvests. Each year is averaged over 4 runs and the overall average is averaged over each individual run.

| USA: CNN | | | | |
|--|-------|-------|-----------|------------|
| Year | RMSE | R^2 | RMSE + GP | $R^2 + GP$ |
| 2015 | 6.75 | 0.471 | 6.58 | 0.541 |
| 2016 | 8.33 | 0.021 | 8.35 | 0.018 |
| 2017 | 6.15 | 0.580 | 6.02 | 0.590 |
| 2018 | 7.69 | 0.445 | 7.52 | 0.429 |
| 2019 | 7.32 | 0.289 | 7.11 | 0.381 |
| 2020 | 6.504 | 0.556 | 6.38 | 0.586 |
| AVG | 7.12 | 0.394 | 7.00 | 0.414 |
| Argentina without Transfer | | | | |
| Year | RMSE | R^2 | RMSE + GP | $R^2 + GP$ |
| 2018 | 10.73 | 0.086 | 9.08 | 0.356 |
| 2019 | 9.23 | 0.189 | 8.73 | 0.273 |
| 2020 | 8.12 | 0.146 | 7.30 | 0.312 |
| AVG | 9.36 | 0.140 | 8.37 | 0.314 |
| Argentina + Freezing, BSS and L ² -SP | | | | |
| Year | RMSE | R^2 | RMSE + GP | $R^2 + GP$ |
| 2018 | 9.53 | 0.293 | 8.49 | 0.440 |
| 2019 | 8.43 | 0.321 | 6.45 | 0.604 |
| 2020 | 6.64 | 0.434 | 5.81 | 0.567 |
| AVG | 8.20 | 0.349 | 6.92 | 0.537 |

References

1. Plested, J.; Gedeon, T. Deep transfer learning for image classification: A survey. *arXiv* **2022**, arXiv:2205.09904.
2. Sharif Razavian, A.; Azizpour, H.; Sullivan, J.; Carlsson, S. CNN features off-the-shelf: An astounding baseline for recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Columbus, OH, USA, 23–28 June 2014; pp. 806–813.
3. Huh, M.; Agrawal, P.; Efros, A.A. What makes ImageNet good for transfer learning? *arXiv* **2016**, arXiv:1608.08614.
4. Chu, B.; Madhavan, V.; Beijbom, O.; Hoffman, J.; Darrell, T. Best practices for fine-tuning visual classifiers to new domains. In Proceedings of the Computer Vision—ECCV 2016 Workshops, Amsterdam, The Netherlands, 8–10 and 15–16 October 2016; pp. 435–442.
5. Pan, S.J.; Yang, Q. A survey on transfer learning. *IEEE Trans. Knowl. Data Eng.* **2009**, *22*, 1345–1359. [[CrossRef](#)]
6. Xuhong, L.; Grandvalet, Y.; Davoine, F. Explicit inductive bias for transfer learning with convolutional networks. In Proceedings of the International Conference on Machine Learning, PMLR, Stockholm, Sweden, 10–15 July 2018; pp. 2825–2834.
7. Li, X.; Xiong, H.; Wang, H.; Rao, Y.; Liu, L.; Chen, Z.; Huan, J. Delta: Deep learning transfer using feature map with attention for convolutional networks. *arXiv* **2019**, arXiv:1901.09229.
8. Chen, X.; Wang, S.; Fu, B.; Long, M.; Wang, J. Catastrophic forgetting meets negative transfer: Batch spectral shrinkage for safe transfer learning. In Proceedings of the Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, Vancouver, BC, Canada, 8–14 December 2019; Volume 32.
9. Muruganatham, P.; Wibowo, S.; Grandhi, S.; Samrat, N.H.; Islam, N. A systematic literature review on crop yield prediction with deep learning and remote sensing. *Remote Sens.* **2022**, *14*, 1990. [[CrossRef](#)]
10. Van Klompenburg, T.; Kassahun, A.; Catal, C. Crop yield prediction using machine learning: A systematic literature review. *Comput. Electron. Agric.* **2020**, *177*, 105709. [[CrossRef](#)]
11. You, J.; Li, X.; Low, M.; Lobell, D.; Ermon, S. Deep Gaussian Process for Crop Yield Prediction Based on Remote Sensing Data. In Proceedings of the AAAI Conference on Artificial Intelligence, San Francisco, CA, USA, 4–9 February 2017; Volume 31.
12. Kaneko, A.; Kennedy, T.; Mei, L.; Sintek, C.; Burke, M.; Ermon, S.; Lobell, D. Deep learning for crop yield prediction in Africa. In Proceedings of the ICML Workshop on Artificial Intelligence for Social Good, Long Beach, CA, USA, 9–15 June 2019; pp. 33–37.
13. Wang, A.X.; Tran, C.; Desai, N.; Lobell, D.; Ermon, S. Deep transfer learning for crop yield prediction with remote sensing data. In Proceedings of the Proceedings of the 1st ACM SIGCAS Conference on Computing and Sustainable Societies, Menlo Park and San Jose, CA, USA, 20–22 June 2018; pp. 1–5.
14. Khaki, S.; Pham, H.; Wang, L. Simultaneous corn and soybean yield prediction from remote sensing data using deep transfer learning. *Sci. Rep.* **2021**, *11*, 11132. [[CrossRef](#)] [[PubMed](#)]

15. Huber, F.; Yushchenko, A.; Stratmann, B.; Steinhage, V. Extreme Gradient Boosting for yield estimation compared with Deep Learning approaches. *Comput. Electron. Agric.* **2022**, *202*, 107346. [CrossRef]
16. Desloires, J.; Ienco, D.; Botrel, A. Out-of-year corn yield prediction at field-scale using Sentinel-2 satellite imagery and machine learning methods. *Comput. Electron. Agric.* **2023**, *209*, 107807. [CrossRef]
17. USDA. USDA NASS Quick Stats Database. 2022. Available online: <https://quickstats.nass.usda.gov/> (accessed on 1 October 2022).
18. Ma, Y.; Chen, S.; Ermon, S.; Lobell, D.B. Transfer learning in environmental remote sensing. *Remote Sens. Environ.* **2024**, *301*, 113924. [CrossRef]
19. Dastour, H.; Hassan, Q.K. A Comparison of Deep Transfer Learning Methods for Land Use and Land Cover Classification. *Sustainability* **2023**, *15*, 7854. [CrossRef]
20. Deng, J.; Dong, W.; Socher, R.; Li, L.J.; Li, K.; Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009; pp. 248–255.
21. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
22. Li, W.; Wang, Z.; Wang, Y.; Wu, J.; Wang, J.; Jia, Y.; Gui, G. Classification of high-spatial-resolution remote sensing scenes method using transfer learning and deep convolutional neural network. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2020**, *13*, 1986–1995. [CrossRef]
23. Alem, A.; Kumar, S. Transfer learning models for land cover and land use classification in remote sensing image. *Appl. Artif. Intell.* **2022**, *36*, 2014192. [CrossRef]
24. Tseng, H.H.; Yang, M.D.; Saminathan, R.; Hsu, Y.C.; Yang, C.Y.; Wu, D.H. Rice seedling detection in UAV images using transfer learning and machine learning. *Remote Sens.* **2022**, *14*, 2837. [CrossRef]
25. Chen, J.; Sun, J.; Li, Y.; Hou, C. Object detection in remote sensing images based on deep transfer learning. *Multimed. Tools Appl.* **2022**, *81*, 12093–12109. [CrossRef]
26. Hilal, A.; Al-Wesabi, F.; Alzahrani, K.; Al Duhayyim, M.; Hamza, M.; Rizwanullah, M.; Díaz, V. Deep Transfer Learning based Fusion Model for Environmental Remote Sensing Image Classification Model. *Eur. J. Remote Sens.* **2022**, *55*, 1–12. [CrossRef]
27. Ma, Y.; Yang, Z.; Huang, Q.; Zhang, Z. Improving the Transferability of Deep Learning Models for Crop Yield Prediction: A Partial Domain Adaptation Approach. *Remote Sens.* **2023**, *15*, 4562. [CrossRef]
28. Ganin, Y.; Ustinova, E.; Ajakan, H.; Germain, P.; Larochelle, H.; Laviolette, F.; March, M.; Lempitsky, V. Domain-adversarial training of neural networks. *J. Mach. Learn. Res.* **2016**, *17*, 1–35.
29. Ministerio de Agricultura. Soja—Serie Siembra, Cosecha, Producción, Rendimiento. 2023. Available online: <https://datosestimaciones.magyp.gob.ar/reportes.php?reporte=Estimaciones> (accessed on 9 May 2023).
30. Chambers, R.G.; Pieralli, S. The sources of measured US agricultural productivity growth: Weather, technological change, and adaptation. *Am. J. Agric. Econ.* **2020**, *102*, 1198–1226. [CrossRef]
31. Vermote, E.. MODIS/Terra Surface Reflectance 8-Day L3 Global 500 m SIN Grid V061 [Data Set]. NASA EOSDIS Land Processes DAAC. 2021. Available online: <https://lpdaac.usgs.gov/products/mod09a1v061/> (accessed on 26 October 2022).
32. Wan, Z.; Hook, S.; Hulley, G.. MODIS/Aqua Land Surface Temperature/Emissivity 8-Day L3 Global 1km SIN Grid V061 [Data Set]. NASA EOSDIS Land Processes DAAC. 2021. Available online: <https://lpdaac.usgs.gov/products/myd11a2v061/> (accessed on 11 November 2022).
33. Friedl, M.; Sulla-Menashe, D. MCD12Q1 MODIS/Terra+Aqua Land Cover Type Yearly L3 Global 500 m SIN Grid V006 [Data Set]. NASA EOSDIS Land Processes DAAC. 2019. Available online: <https://lpdaac.usgs.gov/products/mcd12q1v006/> (accessed on 11 November 2022).
34. Bureau, U.C. Tiger: US Census Counties. 2018. Available online: https://developers.google.com/earth-engine/datasets/catalog/TIGER_2018_Counties/ (accessed on 26 October 2022).
35. USDA. Country Summary. 2023. Available online: <https://ipad.fas.usda.gov/countrysummary/> (accessed on 21 February 2023).
36. Tseng, G. Pycrop-Yield-Prediction. 2022. Available online: <https://github.com/gabrieltseng/pycrop-yield-prediction> (accessed on 30 July 2022).
37. Iman, M.; Rasheed, K.; Arabnia, H.R. A review of deep transfer learning and recent advancements. *arXiv* **2022**, arXiv:2201.09679.
38. Kirkpatrick, J.; Pascanu, R.; Rabinowitz, N.; Veness, J.; Desjardins, G.; Rusu, A.A.; Milan, K.; Quan, J.; Ramalho, T.; Grabska-Barwinska, A.; et al. Overcoming catastrophic forgetting in neural networks. *Proc. Natl. Acad. Sci. USA* **2017**, *114*, 3521–3526. [CrossRef]
39. Williams, C.K.; Rasmussen, C.E. *Gaussian Processes for Machine Learning*; MIT Press: Cambridge, MA, USA, 2006; Volume 2.
40. Akiba, T.; Sano, S.; Yanase, T.; Ohta, T.; Koyama, M. Optuna: A Next-generation Hyperparameter Optimization Framework. In Proceedings of the 25rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Anchorage, AK, USA, 4–8 August 2019.
41. Fernandez-Beltran, R.; Baidar, T.; Kang, J.; Pla, F. Rice-yield prediction with multi-temporal sentinel-2 data and 3D CNN: A case study in Nepal. *Remote Sens.* **2021**, *13*, 1391. [CrossRef]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.