

# A WEAKLY SUPERVISED FRAMEWORK FOR HIGH-RESOLUTION CROP YIELD FORECASTS

**Dilli R. Paudel, Diego M. Gonzalez, Allard de Wit, Hendrik Boogaard & Ioannis N. Athanasiadis**

Wageningen University and Research

PO Box 47, 6700 AA Wageningen, The Netherlands.

{dilli.paudel, diego.marcos, allard.dewit,  
hendrik.boogaard, ioannis.athanasiadis}@wur.nl

## ABSTRACT

Predictor inputs and label data for crop yield forecasting are not always available at the same spatial resolution. We propose a deep learning framework that uses high resolution inputs and low resolution labels to produce crop yield forecasts for both spatial levels. The forecasting model is calibrated by weak supervision from low resolution crop area and yield statistics. We evaluated the framework by disaggregating regional yields in Europe from parent statistical regions to sub-regions for five countries (Germany, Spain, France, Hungary, Italy) and two crops (soft wheat and potatoes). Performance of weakly supervised models was compared with linear trend models and Gradient-Boosted Decision Trees (GBDT). Higher resolution crop yield forecasts are useful to policymakers and other stakeholders. Weakly supervised deep learning methods provide a way to produce such forecasts even in the absence of high resolution yield data.

## 1 INTRODUCTION

Predictor inputs and label data for crop yield forecasting are often not available at the same spatial resolution. Label data, such as yield statistics, are published at regional and national level. Weather inputs are available at grid-level (EC-JRC, 2022; Thornton et al., 2020) and soil and remote sensing data at sub-kilometer resolutions (ESDAC, 2021; Poggio et al., 2021; Copernicus-ESA, 2022). Common statistical and machine learning methods require both inputs and labels at the same spatial level. Therefore, predictor inputs are usually aggregated to the level of yield data. Deep learning methods can handle input and label data at two spatial levels, limiting the spatial aggregation required for input data. Neural network architectures can be trained using high resolution inputs and low resolution yield data to produce crop yield forecasts for both spatial levels.

Many studies have used deep learning for crop yield forecasting (Fan et al., 2021; Shahhosseini et al., 2021; Wolanin et al., 2020; Khaki et al., 2020), but they do not disaggregate yields to high resolutions. Other methods of disaggregating crop yields exist, for example, area-to-point kriging (Brus et al., 2018; Steinbuch et al., 2020) and spatial allocation based on cross-entropy method (You et al., 2014) or remote sensing indicators (Shirsath et al., 2020). We draw inspiration from Jacobs et al. (2018), who trained a convolutional neural network and an aggregation layer to predict pixel-level population density from high resolution satellite images and low resolution density statistics. To our knowledge, weakly supervised methods have not been used to disaggregate crop yields to high resolutions.

We propose a weakly supervised deep learning framework that uses high resolution inputs and low resolution labels to produce crop yield forecasts for both spatial resolutions. Predictor inputs come from NUTS3 regional level and yield and crop area statistics from NUTS2 level, with NUTS2 and NUTS3 regions representing the low and high spatial resolutions. NUTS (Nomenclature of Territorial Units for Statistics) is a hierarchical system of dividing the territory of the European Union for statistics and policy (Eurostat, 2016). Our approach is weakly supervised because models to produce NUTS3 forecasts are trained using low resolution NUTS2 labels. Our objective is to build and evaluate crop yield forecasting models that can produce high resolution forecasts even when high resolution yield data are not available. This objective can be viewed from two perspectives.

First, we assess the performance of weak supervised models in disaggregating crop yields from low to high resolution. Second, we evaluate the quality of low resolution yield forecasts produced using high resolution inputs. Our analysis includes two crops (soft wheat and potatoes) and five countries: Germany (DE), Spain (ES), France (FR), Hungary (HU) and Italy (IT).

The rest of the paper is structured as follows: Section 2 describes data and methods; Section 3 presents the results; and Section 4 discusses our findings and outlines directions for future work.

## 2 METHODS

Our weakly supervised deep learning framework uses high resolution predictor inputs and low resolution labels to produce crop area and yield forecasts for both spatial resolutions. To evaluate these yield forecasts, we compared performance with linear trend models and Gradient-Boosted Decision Trees. To assess their usefulness to policymakers and other stakeholders, we analyzed the spatial variability of high resolution forecasts.

### 2.1 DATA

Data sources used in this paper are summarized in Table A.1. Most of our data came from the MARS Crop Yield Forecasting System of the European Commission’s Joint Research Centre (MARSWiki, 2021) and covered two crops and five countries: soft wheat (DE, ES, FR, IT) and potatoes (DE, FR, HU, IT). Data from all countries was combined to build one prediction model per crop. Seasonal data included outputs of the WOFOST crop model (Van Diepen et al., 1989; Supit et al., 1994; De Wit et al., 2019), weather variables and remote sensing indicators aggregated to NUTS3. The yield trend was captured using yield values of five previous years at NUTS2. Static differences among regions were captured by soil water holding capacity and agro-environmental features, such as elevation, slope, field sizes, irrigated area (Paudel et al., 2022). In addition, agro-environmental zones and countries were added as categorical variables to account for other agro-climatic and administrative differences. Reported yield and crop area statistics at NUTS2 served as labels. In most cases, we had data from 1999 to 2018. The most recent 30% of the years were allocated to the test set. From the remaining 70% training years, 5 most recent years were used for a sliding-window 5-fold validation (Figure A.1).

### 2.2 THE WEAKLY SUPERVISED FRAMEWORK

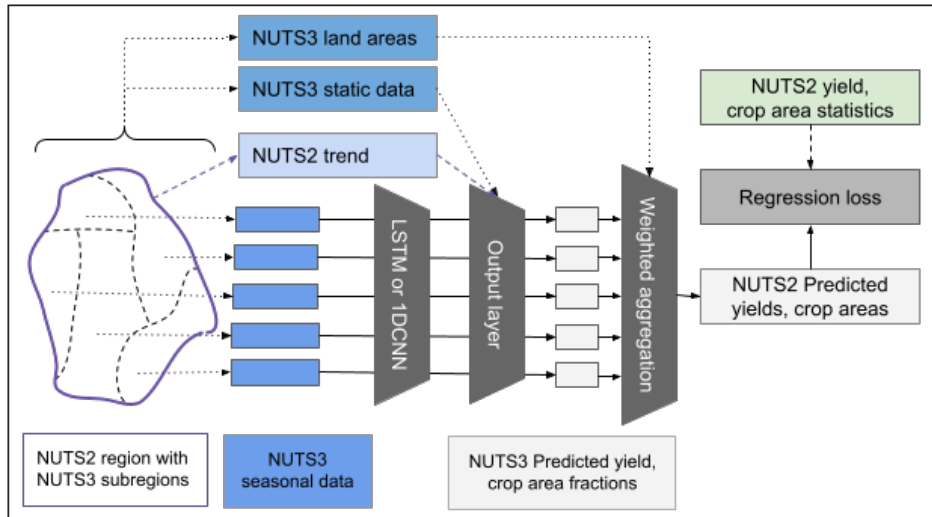


Figure 1: Weakly supervised framework to produce high resolution crop yield forecasts.

We selected two neural network architectures, namely long short-term memory (LSTM) networks and 1-dimensional convolutional neural networks (1DCNN), that can learn from seasonal time series of predictors. Both LSTMs and 1DCNNs have been used in literature to learn features from sequential data (You et al., 2017; Khaki et al., 2020). Seasonal data at NUTS level 3, including crop model outputs, weather and remote sensing indicators, was processed by LSTM or 1DCNN. Features from the LSTM or 1DCNN layers, together with static agro-environmental data and yield trend features (based on NUTS2 yields), were passed to an output layer (Figure 1). An aggregation layer computed crop area weights and aggregated the forecasts to low resolution (NUTS2). The framework was supervised with NUTS2 yields and crop areas. Data from all NUTS3 regions within a NUTS2 region formed a batch to enable aggregation of NUTS3 forecasts.

The weakly supervised model (WS model) produced NUTS3 yield forecasts and crop area fractions. We believe remote sensing inputs can predict crop area fractions (crop area/land area), but not the absolute crop areas. Predicted crop area fractions were multiplied with land areas to produce NUTS3 crop areas, which were used to calculate crop area weights for the aggregation layer.

### 2.3 EVALUATION

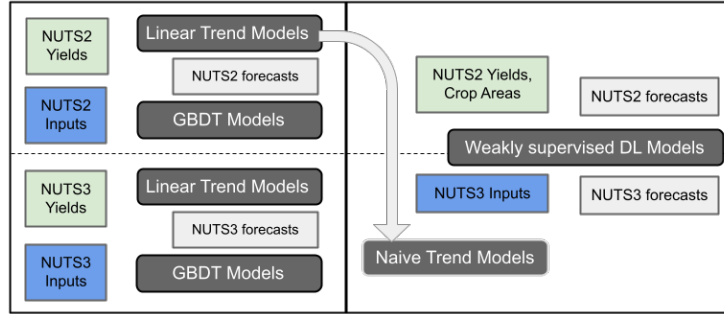


Figure 2: Evaluation framework.

Forecasts from the WS model were evaluated at both spatial resolutions. The quality of disaggregated NUTS3 yields was assessed by comparing with two sets of models. The first set was a naive trend model that acted as our baseline for disaggregation. The naive trend model predicted the NUTS2 yield trend as NUTS3 forecasts for all NUTS3 sub-regions. The second set included the NUTS3 trend model and Gradient-Boosted Decision Trees (GBDT) model (Figure 2). Unlike the naive trend and WS models, the NUTS3 trend and GBDT models had access to NUTS3 yield data. The NUTS3 GBDT and WS models used the same predictor inputs, except yield trend features: GBDT had access to NUTS3 trend; WS only had NUTS2 trend. Forecasts were made 60 days before harvest. At NUTS2, WS model forecasts were compared with the NUTS2 trend model and NUTS2 GBDT model. All trend models were calibrated with yield values of five previous years.

Models were compared based on box plots of prediction residuals (i.e., predicted yield - reported yield) and normalized root mean squared errors (NRMSE), normalized by average yield of the test set. The significance of model performance was evaluated by running Wilcoxon signed-rank test on the prediction residuals of different models.

In addition to error comparisons, we also analyzed the spatial variability of NUTS3 yield forecasts from disaggregation models (naive trend and WS model). A significant part of yield variability is explained using the yield trend attributed to factors such as technological improvements (see Lecerf et al. (2019)). We expected the NUTS2 trend features to make the model more accurate, but suppress spatial variability among NUTS3 regions. Therefore, we ran a version of the WS model without NUTS2 trend to learn spatial differences. Spatial variability of NUTS3 yield forecasts was analyzed for soft wheat (FR). We selected regions based on maximum acreage and years (2016 and 2017) based on significant yield losses reported in the north of FR in 2016 (see Ben-Ari et al. (2018)).

### 3 RESULTS

In this section, we report results from the LSTM version of the WS model because of its superior validation set performance over the IDCNN one (Figure A.2).

#### 3.1 EVALUATION OF HIGH RESOLUTION YIELD FORECASTS

The WS models were statistically similar to NUTS3 GBDT models (p-values: 0.67 for soft wheat and 0.265 for potatoes) (Table A.2), and significantly better than the naive trend models (p-values near zero). For soft wheat, the WS model was also better than the NUTS3 trend model (p-value near zero). For soft wheat, interquartile range was smaller and per-country NRMSEs were mostly lower compared to both trend models (Figure 3). Disaggregation worked less well for potatoes, but the performance was still better than the naive trend model (Table A.2, Figure A.3).

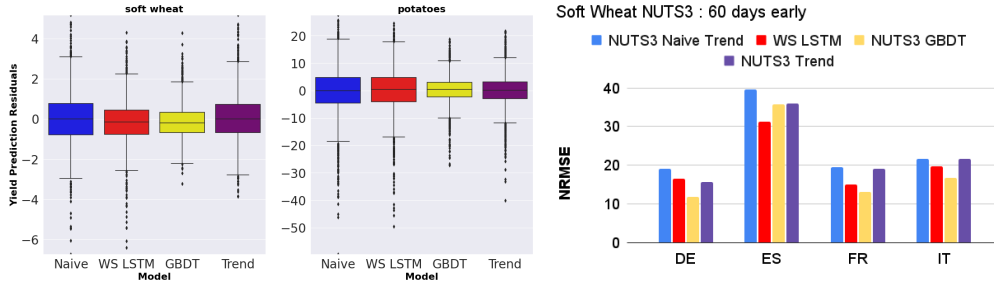


Figure 3: *Left*: Boxplots of NUTS3 prediction residuals. *Right*: NRMSEs for soft wheat.

#### 3.2 EVALUATION OF LOW RESOLUTION YIELD FORECASTS

The WS models produced NUTS2 forecasts that were similar compared to NUTS2 GBDT models (p-values: 0.083 for soft wheat and 0.455 for potatoes) (Table A.3), and better than the NUTS2 trend models (p-values: 0.00095 for soft wheat and 0.000002 for potatoes). Box plots and per-country NRMSEs again showed that residuals and errors were smaller for soft wheat than for potatoes (Figure 4, A.3). The WS models had smaller median residuals compared to GBDT models for both crops (Table A.3), indicating some value in using NUTS3 inputs to forecast at NUTS2.

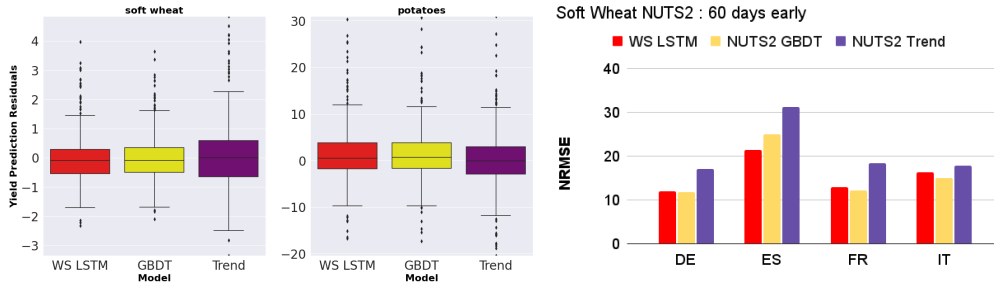


Figure 4: *Left*: Boxplots of NUTS2 prediction residuals. *Right*: NRMSEs for soft wheat.

#### 3.3 SPATIAL VARIABILITY OF HIGH RESOLUTION FORECASTS

As mentioned before, the north of France had significant yield losses for soft wheat in 2016. The naive trend model predicted higher yields, with an average prediction residual of 2.865. The WS model predicted slightly lower (average residual: 2.21) but still higher than the reported yields (Figure 5, *top*). In 2017, the naive trend model was heavily influenced by the 2016 yields (especially in the middle: FRI3, FRB0, FRC1), while the WS model predicted values similar to the reported

yields (Figure 5, *bottom*). As expected, the WS model with trend did not capture yield variability within NUTS2 regions. The WS No Trend model captured such differences better, but generally underestimated the yields: the average prediction residual was -0.57 compared to -0.028 for the WS model using trend. Combining information from both versions of the WS model provided more accurate estimates of yields as well as spatial differences among NUTS3 regions.

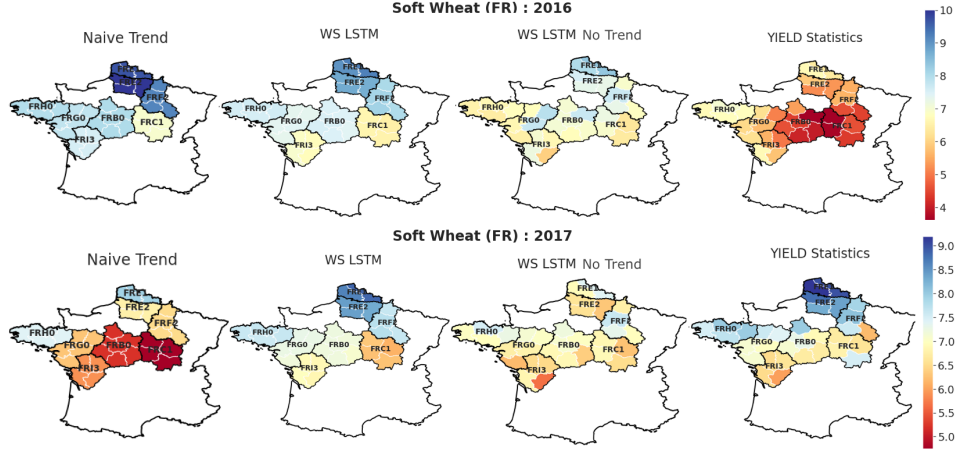


Figure 5: Spatial variability of soft wheat yields and forecasts. *Top*: FR 2016. *Bottom*: FR 2017

## 4 DISCUSSION

Data on crop yield predictors will be increasingly available at higher resolutions. Yield data may not be available due to many reasons, including privacy concerns. When there is an imbalance between spatial resolutions of inputs and yields, weak supervised methods provide a solution. Our results showed that weakly supervised models were able to produce reliable NUTS3 yield forecasts, especially for soft wheat, without using NUTS3 yields. The no-trend version of the model also captured some of NUTS3-level yield variability. Our approach will continue to work when yield data is available at low resolution for some regions and at high resolution for others. Similarly, high resolution crop areas, when available, will further improve the quality of yield forecasts.

In this paper, we have scratched the surface of high resolution crop yield forecasting without strong supervision. We see three areas that need further research to gauge the ability of weak supervision to produce reliable forecasts. First, the scale differences that can be handled by weak supervision needs investigation. For example, weak supervision from NUTS3 yields may produce good quality forecasts for 25km grids, but not 1km grids. Second, predictor inputs must be suitable to capture yield variability at selected resolutions. Weather variables may influence NUTS2 and NUTS3 yields, but become less relevant at farm level. Third, we experimented with standard neural network architectures. Future work could investigate other architectures that are more suitable for weak supervision. Data size and quality will always play a role due to the data-driven nature of neural networks.

High resolution crop yield forecasts provide useful information to policymakers and other stakeholders for local analysis and monitoring. We have shown that weakly supervised methods can produce such forecasts in the absence of high resolution labels.

## 5 ACKNOWLEDGMENTS

This work was partially supported by the European Union’s Horizon 2020 research and innovation programme under grant agreement No. 825355 (CYBELE). We would like to thank S. Niemeyer from the European Commission’s Joint Research Centre (JRC) for permission to use MARS Crop Yield Forecasting System data. We also would like to thank M. van der Velde, L. Nisini and I. Cerrani from JRC for sharing the Eurostat regional yield statistics and crop areas.

## REFERENCES

- Richard G Allen, Luis S Pereira, Dirk Raes, Martin Smith, et al. Crop evapotranspiration – guidelines for computing crop water requirements. In *Irrigation and Drainage, paper 56*. FAO, Rome, 1998. ISBN 92-5-104219-5.
- Tamara Ben-Ari, Julien Boé, Philippe Ciais, Remi Lecerf, Marijn Van der Velde, and David Makowski. Causes and implications of the unforeseen 2016 extreme yield loss in the breadbasket of France. *Nature communications*, 9(1):1–10, 2018. doi: 10.1038/s41467-018-04087-x.
- DJ Brus, H Boogaard, T Ceccarelli, TG Orton, S Traore, and M Zhang. Geostatistical disaggregation of polygon maps of average crop yields by area-to-point kriging. *European Journal of Agronomy*, 97:48–59, 2018.
- Copernicus-ESA. Sentinel Earth Observation Data, 2022. <https://scihub.copernicus.eu/>, Last accessed: Feb 22, 2022.
- Copernicus GLS. Fraction of Absorbed Photosynthetically Active Radiation, 2020. <https://land.copernicus.eu/global/products/fapar>, Last accessed: Oct 19, 2020.
- DE-RegionalStatistiks. Regionaldatenbank deutschland, 2020. <https://www.regionalstatistik.de/genesis/online/data>, Last accessed: May 11, 2020.
- Allard De Wit, Hendrik Boogaard, Davide Fumagalli, Sander Janssen, Rob Knapen, Daniel van Kraalingen, Iwan Supit, Raymond van der Wijngaart, and Kees van Diepen. 25 years of the WOFOST cropping systems model. *Agricultural Systems*, 168:154–167, 2019. doi: 10.1016/j.agry.2018.06.018.
- EC-JRC. JRC Agri4Cast Data Portal, 2022. <https://agri4cast.jrc.ec.europa.eu/DataPortal/Index.aspx>, Last accessed: Feb 11, 2022.
- ESDAC. European soil database, 2021. <https://esdac.jrc.ec.europa.eu/resource-type/datasets>, Last accessed: April 28, 2021.
- Eurostat. Nomenclature of territorial units for statistics, 2016. <https://ec.europa.eu/eurostat/web/nuts/background>, Last accessed: May 11, 2020.
- Eurostat. Eurostat - Agricultural Production - crops, 2021. [https://ec.europa.eu/eurostat/statistics-explained/index.php/Agricultural\\_production\\_-\\_crops](https://ec.europa.eu/eurostat/statistics-explained/index.php/Agricultural_production_-_crops), Last accessed: May 11, 2021.
- Joshua Fan, Junwen Bai, Zhiyun Li, Ariel Ortiz-Bobea, and Carla P. Gomes. A GNN-RNN approach for harnessing geospatial and temporal information: Application to crop yield prediction. *CoRR*, abs/2111.08900, 2021. doi: 10.48550/arXiv.2111.08900. URL <https://arxiv.org/abs/2111.08900>.
- FR-Agrete. Agreste Web Data Portal, 2020. <https://agreste.agriculture.gouv.fr/agreste-web/>, Last accessed: May 11, 2020.
- Nathan Jacobs, Adam Kraft, Muhammad Usman Rafique, and Ranti Dev Sharma. A weakly supervised approach for estimating spatial density functions from high-resolution satellite imagery. In *Proceedings of the 26th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pp. 33–42, 2018. doi: 10.1145/3274895.3274934.
- Saeed Khaki, Lizhi Wang, and Sotirios V Archontoulis. A CNN-RNN framework for crop yield prediction. *Frontiers in Plant Science*, 10:1750, 2020. doi: 10.3389/fpls.2019.01750.
- Rémi Lecerf, Andrej Ceglar, Raúl López-Lozano, Marijn Van Der Velde, and Bettina Baruth. Assessing the information in crop model and meteorological indicators to forecast crop yield over Europe. *Agricultural Systems*, 168:191–202, 2019. doi: 10.1016/j.agry.2018.03.002.
- Myroslava Lesiv, Juan Carlos Laso Bayas, Linda See, Martina Duerauer, Domian Dahlia, Neal Durando, Rubul Hazarika, Parag Kumar Sahariah, Mar’yana Vakolyuk, Volodymyr Blyshchuk, et al. Estimating the global distribution of field size using crowdsourcing. *Global Change Biology*, 25(1):174–186, 2019. doi: 10.1111/gcb.14492.

- MARSWiki. MARS Crop Yield Forecasting System, 2021. [https://marswiki.jrc.ec.europa.eu/agri4castwiki/index.php/Welcome\\_to\\_WikiMCYFS](https://marswiki.jrc.ec.europa.eu/agri4castwiki/index.php/Welcome_to_WikiMCYFS), Last accessed: May 11, 2021.
- Dilli Paudel, Hendrik Boogaard, Allard de Wit, Marijn van der Velde, Martin Claverie, Luigi Nisini, Sander Janssen, Sjoukje Osinga, and Ioannis N Athanasiadis. Machine learning for regional crop yield forecasting in Europe. *Field Crops Research*, 276:108377, 2022. doi: 10.1016/j.fcr.2021.108377.
- F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.
- Laura Poggio, Luis M De Sousa, Niels H Batjes, Gerard Heuvelink, Bas Kempen, Eloi Ribeiro, and David Rossiter. Soilgrids 2.0: producing soil information for the globe with quantified spatial uncertainty. *Soil*, 7(1):217–240, 2021. doi: 10.5194/soil-7-217-2021.
- Scikit-optimize Contributors. Scikit-optimize: Sequential model-based optimization, 2021. [https://scikit-optimize.github.io/stable/getting\\_started.html](https://scikit-optimize.github.io/stable/getting_started.html), Last accessed: Sept 20, 2021.
- Mohsen Shahhosseini, Guiping Hu, Saeed Khaki, and Sotirios V Archontoulis. Corn yield prediction with ensemble CNN-DNN. *Frontiers in Plant Science*, 12, 2021. doi: 10.3389/fpls.2021.709008.
- Paresh B Shirsath, Vinay Kumar Sehgal, and Pramod K Aggarwal. Downscaling regional crop yields to local scale using remote sensing. *Agriculture*, 10(3):58, 2020. doi: 10.3390/agriculture10030058.
- Luc Steinbuch, Thomas G Orton, and Dick J Brus. Model-based geostatistics from a bayesian perspective: Investigating area-to-point kriging with small data sets. *Mathematical Geosciences*, 52(3):397–423, 2020. doi: 10.1007/s11004-019-09840-6.
- I Supit, AA Hooijer, and CA Van Diepen. System description of the WOFOST 6.0 crop simulation model implemented in CGMS. vol. 1. theory and algorithms. In *EUR Publication No. 15959 EN*, pp. 146. Office for Official Publications of the European Communities, Luxembourg, 1994.
- M.M. Thornton, Y. Wei, P.E. Thornton, R. Shrestha, S. Kao, and B.E. Wilson. Daymet: Station-Level Inputs and Cross-Validation Result for North America, Version 4, 2020. URL [https://daac.ornl.gov/cgi-bin/dsvviewer.pl?ds\\_id=1850](https://daac.ornl.gov/cgi-bin/dsvviewer.pl?ds_id=1850).
- USGS-EROS. USGS EROS Archive - Digital Elevation - Global 30 Arc-Second Elevation (GTOPO30), 2021. <https://www.usgs.gov/centers/eros/data>, Last accessed: May 11, 2021.
- CA Van Diepen, J Wolf, H Van Keulen, and C Rappoldt. WOFOST: a simulation model of crop production. *Soil Use and Management*, 5(1):16–24, 1989. doi: 10.1111/j.1475-2743.1989.tb00755.x.
- Aleksandra Wolanin, Gonzalo Mateo-García, Gustau Camps-Valls, Luis Gómez-Chova, Michele Meroni, Gregory Duveiller, You Liangzhi, and Luis Guanter. Estimating and understanding crop yields with explainable deep learning in the Indian Wheat Belt. *Environmental Research Letters*, 15(2):024019, 2020. doi: 10.1088/1748-9326/ab68ac.
- Jiaxuan You, Xiaocheng Li, Melvin Low, David Lobell, and Stefano Ermon. Deep Gaussian process for crop yield prediction based on remote sensing data. In *Thirty-First AAAI Conference on Artificial Intelligence*, 2017. <https://www.aaai.org/ocs/index.php/AAAI/AAAI17/paper/view/14435/14067>, Last accessed: May 11, 2020.
- Liangzhi You, Stanley Wood, Ulrike Wood-Sichra, and Wenbin Wu. Generating global crop distribution maps: From census to grid. *Agricultural Systems*, 127:53–60, 2014. doi: 10.1016/j.agsy.2014.01.002.

## A APPENDIX

### A.1 DATA SOURCES

Most of our data comes from the MARS Crop Yield Forecasting System (MCYFS) of European Commission’s Joint Research Centre (MARSWiki, 2021). Other sources are indicated in Table A.1.

Table A.1: Data sources summary

DATA	INDICATORS, SOURCES
WOFOST crop model outputs	Water-limited dry weight biomass ( $kg\ ha^{-1}$ ), Water-limited dry weight storage organs ( $kg\ ha^{-1}$ ), Water-limited leaf area divided by surface area ( $m^2\ m^{-2}$ ), Development stage (0 – 200), root-zone soil moisture as % of soil water holding capacity, sum of water limited transpiration ( $cm$ ). <b>Source:</b> MCYFS. See Lecerf et al. (2019).
Meteo	Maximum, minimum, average daily air temperature ( $^{\circ}C$ ), sum of daily precipitation (PREC) ( $mm$ ), sum of daily evapotranspiration of short vegetation (ET0) (Penman-Monteith, Allen et al. (1998)) ( $mm$ ), climate water balance = (PREC - ET0). <b>Source:</b> MCYFS. See Lecerf et al. (2019).
Remote Sensing	Fraction of Absorbed Photosynthetically Active Radiation (Smoothed) (FAPAR). <b>Source:</b> MCYFS. See Copernicus GLS (2020).
Crop Areas	Absolute crop areas ( $ha$ ). Fraction of parent regions’s crop area. <b>Source:</b> Eurostat (Eurostat, 2021) and MCYFS (EC-JRC, 2022).
Irrigated area	Irrigated total area and irrigated crop-specific area. <b>Source:</b> EC-JRC (2022).
Elevation, slope	Average and standard deviation. <b>Source:</b> USGS-EROS (2021).
Soil	SM_WHC (water holding capacity). <b>Source:</b> MCYFS. See Lecerf et al. (2019).
Field Size	Average and standard deviation. <b>Source:</b> Lesiv et al. (2019).
Yield	Yield at NUTS3 level. <b>Source:</b> FR-Agreste (2020); DE-RegionalStatistiks (2020); Eurostat (2021); EC-JRC (2022).

### A.2 WEAKLY SUPERVISED MODEL

The weakly supervised model was supervised using NUTS2 crop areas and yields. The combined loss was the sum of the two losses (crop area loss and yield loss) normalized by the training set average of the corresponding labels. The hyperparameters learning rate and L2-penalty lambda were optimized using custom sliding validation (Figure A.1; Paudel et al. (2022)). After optimizing hyperparameters, the model was trained on the entire validation set (no 5-fold) with early stopping: training stopped after validation error increased for two successive epochs. The optimized hyperparameters and early stopping epochs were used to evaluate the model on the test set.

#### A.2.1 1DCNN ARCHITECTURE

CNN Layer1 : Conv1d(11, 16, kernel\_size=(3,), stride=(1,), padding=(1,))  
 CNN Layer 2: Conv1d(16, 32, kernel\_size=(3,), stride=(2,), padding=(1,))  
 CNN Layer 3: Conv1d(32, 8, kernel\_size=(3,), stride=(2,), padding=(1,))  
 Batch Normalization, ReLU Activation and Dropout(p=0.1) were added after each CNN layer.  
 Output Layer: Linear(in\_features=100, out\_features=2, bias=True)

#### A.2.2 LSTM ARCHITECTURE

LSTM Layer : LSTM(11, 64, batch\_first=True)  
 Output Layer : Linear(in\_features=100, out\_features=2, bias=True)

Both LSTM and 1DCNN were implemented using pytorch (<https://pytorch.org/>).

### A.3 GBDT MODEL

Input data and training, validation and test splits were identical between the GBDT model and the weakly supervised model, except for trend features (NUTS3 vs NUTS2). The GBDT model used in this paper has some differences compared to Paudel et al. (2022). First, a combined model was built for four countries. Second, agro-environmental zones and countries were added as categorical features.

The GBDT model is based on `GradientBoostingRegressor()` from `scikit-learn` (Pedregosa et al., 2011). Hyperparameters including GBDT parameters, feature selector and number of features optimized using `BayesSearchCV` from `scikit-optimize` package (Scikit-optimize Contributors, 2021). Feature selectors included a Random Forests model (`SelectFromModel`) and Recursive Feature Elimination using a Lasso Regression model.

### A.4 TRAINING, VALIDATION AND TEST SPLITS

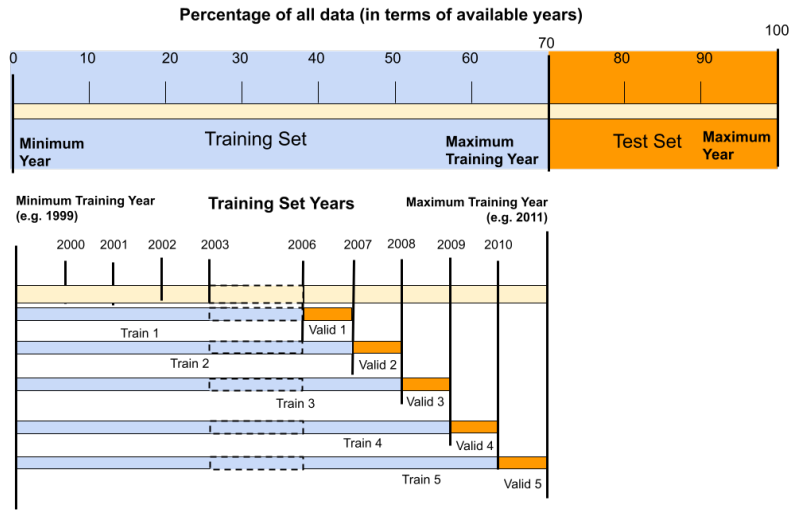


Figure A.1: Training, validation and test splits. Reproduced from Paudel et al. (2022)

### A.5 SOFTWARE IMPLEMENTATION

Software implementation of the weakly supervised framework can be accessed here:  
<https://github.com/BigDataWUR/MLforCropYieldForecasting/tree/weaksup>

### A.6 VALIDATION SET RESULTS FOR ARCHITECTURE SELECTION

The decision to use combined data from four countries was based on validation set performance of strongly supervised NUTS3 models for soft wheat. Since NRMSEs were similar for both cases, we chose to use combined data because the larger data size would limit overfitting issues. CV comparisons for weak supervision using 1DCNN and LSTM showed that LSTM had lower NRMSEs on the validation set (Figure A.2).

### A.7 SUPPLEMENTARY RESULTS

Wilcoxon signed rank test was run using prediction residuals (predicted yield - reported yield).

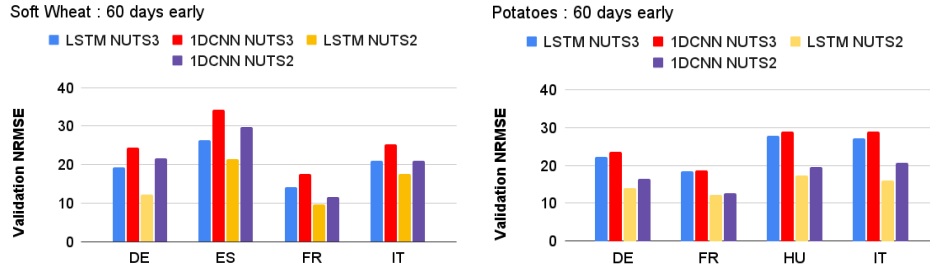


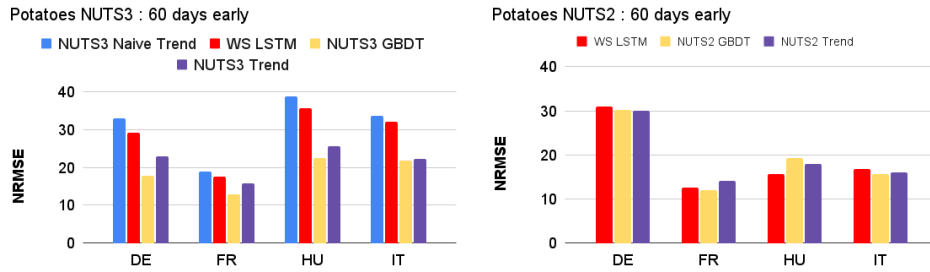
Figure A.2: Validation NRMSEs of weakly supervised NUTS3 and NUTS2 forecasts.

Table A.2: Wilcoxon-signed rank test (NUTS3)

MODEL	P-VALUE	MEDIAN
<b>Soft wheat</b>		
WS LSTM	—	-0.144
Naive Trend	0	0
NUTS3 GBDT	0.67	-0.193
NUTS3 Trend	0	0.01
<b>Potatoes</b>		
WS LSTM	—	0.447
Naive Trend	0.00002	-0.092
NUTS3 GBDT	0.265	0.397
NUTS3 Trend	0.07856	0.039

Table A.3: Wilcoxon-signed rank test (NUTS2)

MODEL	P-VALUE	MEDIAN
<b>Soft wheat</b>		
WS LSTM	—	-0.089
NUTS2 GBDT	0.08346	-0.0942
NUTS2 Trend	0.00095	-0.005
<b>Potatoes</b>		
WS LSTM	—	0.494
NUTS2 GBDT	0.455	0.664
NUTS2 Trend	0.000002	-0.1

Figure A.3: Per-country NRMSEs for potatoes. *Left*: NUTS3. *Right*: NUTS2.