Integrating processed-based models and machine learning for crop yield prediction

Michiel G.J. Kallenberg [*][1] **Bernardo Maestrini** [*][1] **Ron van Bree** [1] **Paul Ravensbergen** [1] **Christos Pylianidis** [1] **Frits van Evert** [1] **Ioannis N. Athanasiadis** [1]

Abstract

Crop yield prediction typically involves the utilization of either *theory-driven* process-based crop growth models, which have proven to be difficult to calibrate for local conditions, or *data-driven* machine learning methods, which are known to require large datasets. In this work we investigate potato yield prediction using a hybrid meta-modeling approach. A crop growth model is employed to generate synthetic data for (pre)training a convolutional neural net, which is then fine-tuned with observational data. When applied in silico, our meta-modeling approach yields better predictions than a baseline comprising a purely data-driven approach. When tested on real-world data from field trials (n=303) and commercial fields (n=77), the meta-modeling approach yields competitive results with respect to the crop growth model. In the latter set, however, both models perform worse than a simple linear regression with a hand-picked feature set and dedicated preprocessing designed by domain experts. Our findings indicate the potential of meta-modeling for accurate crop yield prediction; however, further advancements and validation using extensive real-world datasets is recommended to solidify its practical effectiveness.

1. Introduction

Crop yield prediction plays an important role in ensuring food security, optimizing agricultural practices, and managing risks in the farming industry. By accurately estimating the expected crop yields, stakeholders ranging from farmers and policymakers to traders and researchers can make informed decisions and take proactive measures. In a world facing increasing population growth and climate uncertainties, the ability to predict crop yields has become indispensable for addressing global food challenges and fostering sustainable agricultural practices.

In crop yield prediction, *theory-driven* process-based crop models and *data-driven* machine learning models are usually seen as opposite approaches (Leng & Hall, 2020; Paudel et al., 2022; Maestrini et al., 2022). In this work, we investigate the use of a meta-modeling approach to combine the best of these two worlds. The integration of the two approaches may help overcome their individual approach limitations, for process-based models the limited processes included in the model (van Ittersum et al., 2013), and their difficulty in adapting to local conditions, and for data-driven models the need for large, variable and orthogonal datasets, that are often unavailable in the agricultural context.

One common method of integrating data-driven and process-based approaches is through the development of a metamodel (Wallach et al., 2019; Pylianidis et al., 2022). The workflow to create a metamodel is typically the following: a synthetic dataset is created using a process-based model fed with different inputs representative for the conditions where the model will be used, and then a data-driven model is fit on the synthetic dataset (Razavi et al., 2012; Pylianidis et al., 2022). Usually only a subset of the input (e.g. weather, management, soil texture) and a subset of the output (e.g. yield, or yield and leached N) is retained in the metamodel.

To align model predictions with local conditions, crop yield prediction using process-based crop growth models approaches typically involves a tedious calibration process (Seidel et al., 2018). Calibration of these models can be challenging due to the complexity of cropping systems, spatial and temporal variability, limited data availability, and parameter sensitivity. Calibrating a data-driven method is often considered easier compared to calibrating a process-based model. For data-driven models transfer learning techniques, such as fine-tuning, can be leveraged to adapt a generic model to local conditions.

[*]Equal contribution [1]Wageningen University & Research, the Netherlands. Correspondence to: Michiel Kallenberg <michiel.kallenberg@wur.nl>.

Accepted after peer-review at the 1st workshop on Synergy of Scientific and Machine Learning Modeling, SynS & ML ICML, Honolulu, Hawaii, USA. July, 2023. Copyright 2023 by the author(s).

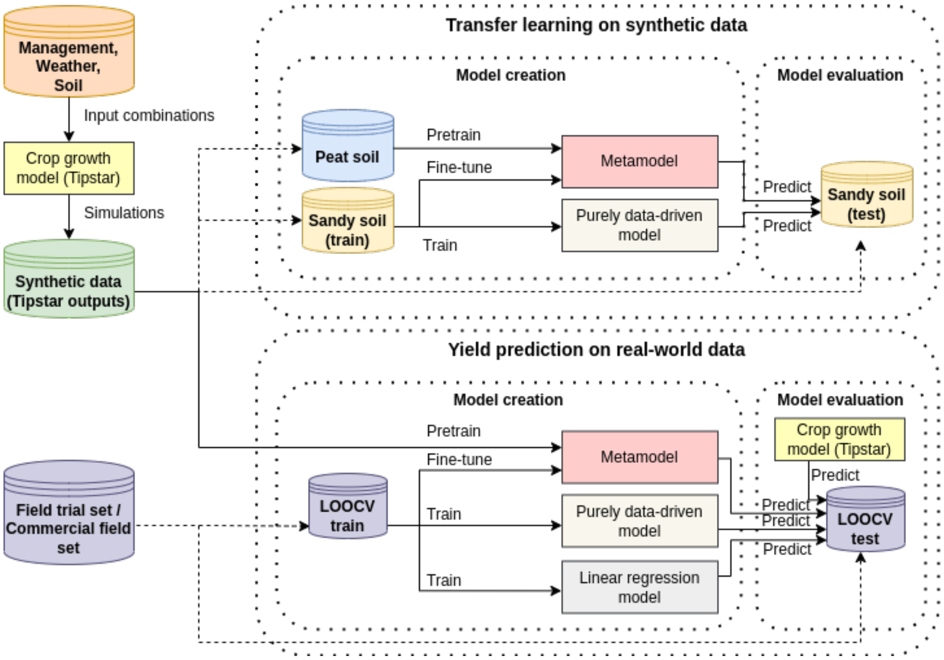

Figure 1. Schematic overview of experimental setup, data flow and model development

In this work, we compare three methods for predicting potato yield, quantified as fresh tuber weight at harvest time: (1) a process based crop growth model, in our case Tipstar (Jansen, 2008) (2) a (fine-tuned) metamodel, pre-trained with synthetic data generated by Tipstar, and (3) a purely data-driven model, not (pre)trained with Tipstar.

2. Materials and methods

A schematic overview of the experimental setup, data flow and model development is shown in Figure 1.

2.1. Models

2.1.1. Process-based crop growth model

Tipstar (Jansen, 2008) is a crop growth model that simulates potato growth under water and nitrogen limited conditions with daily time steps. The input to the model are crop management (planting, fertilization, and irrigation), weather (solar radiation, rainfall, maximum and minimum air temperature, wind), and soil characteristics (texture, organic matter percentage, van Genuchten parameters (van Genuchten, 1980)). The model was not calibrated for this specific setup (i.e., no model parameters were adjusted to fit the data of this study), but we re-used default parameters that were derived from historical data from the same country (i.e., the Netherlands).

2.1.2. Metamodel

We developed a metamodel of Tipstar by creating a synthetic dataset of 86,000 entries, which was then used to pretrain a convolutional neural network. The synthetic dataset was created by taking the full factorial of the model inputs weather (7 locations, 32 years) and soil (32 soil types (Wosten et al., 2013)). For each of the resulting 7,168 combinations we generated 12 simulations using random sampling while varying the following model inputs: nitrogen fertilization, sowing date, irrigation, maximum rooting depth, and cultivar earliness.

The metamodel was constructed as a multi-stream (n=3) convolutional neural network. The first stream processed the temporal data (i.e. solar radiation, rainfall, maximum and minimum air temperature, cumulative irrigation and cumulative nitrogen input) with twice a 1D convolution (number of filters 20 and 7, kernel sizes 3 and 2) followed by an average pooling (5 and 5). The second stream processes the scalars (i.e. maximum rooting depth, sowing day of year and cultivar earliness) with two dense layers (20, and 20). The (optional) third stream processes the soil characteristics (i.e., clay, loam, organic matter percentage, soil moisture at saturation, and the Van Genuchten parameters: α, λ and n (van Genuchten, 1980)) with a 1D convolution (number of filers 5, kernel size 5) followed by an average pooling (24). The three streams are all flattened and concatenated, and subsequently fed to three dense layers (25, 5 and 1).

2.1.3. Data-driven model

As a baseline for our model comparisons we trained a purely data-driven model. This model was not pretrained with synthetic data, but was trained with observational data. The model had the same architecture as the metamodel.

(Pre)training of the metamodel and the data-driven model was done with mean squared error (MSE) as a loss function. We used the ADAM optimizer with an initial learning rate of 0.001. Training was monitored on an hold-out validation set. To prevent overfitting, we used early stopping (min_delta=0.001, patience=20), and learning rate reduction (factor=0.5, min_delta=0.001, patience=10).

2.2. Experiments

We conducted two experiments to investigate the merits of our meta-modeling approach.

2.2.1. Transfer learning on synthetic data

The first experiment served as a proof-of-concept of our transfer learning approach, and was done with synthetic data only. We selected ”soil” as our domain of interest, since, in general, potato yield depends substantially on the characteristics of the soil (Porter et al., 1999). We divided the synthetic dataset in two subsets; the first set (i.e., source domain) contained the simulations on peat soils (category 2xx in BOFEK 2012 (Wosten et al., 2013)) and the second (i.e., target domain) the sandy soils (category 3xx in BOFEK 2012 (Wosten et al., 2013)). We confirmed that yields differed between the two soil types, which was found to be partially caused by a different yield response to water availability.

A metamodel (excluding the soil input stream) was pretrained on the peat soil dataset (n=10k), and then fine-tuned with the sandy soil dataset. For the fine-tuning we chose to freeze all layers of the network, except for the last two layers, as this approach yielded slightly superior results compared to alternatives such as training all layers, and training the first layers only. For comparison, we trained a data-driven model with the same architecture of the metamodel. The data-driven model was exclusively trained with the sandy soil dataset. Both models were evaluated on a hold-out set (n=17k) with sandy soils.

Transfer learning is especially instrumental in cases where there is a limited availability of data from the target domain. We simulated such a scenario by limiting the size of the dataset available for fine-tuning; we evaluated three different levels of size n=50, 200, and 1000 respectively.

2.2.2. Yield prediction on real-world data

The second experiment was done with two real-world datasets; one from field trials and one from commercial fields. The field trial dataset (n=303, 1994-2003) came from a collection of 36 potato experiments carried out by Wageningen University and Research. The factors that were varied in the experiments were cultivar, nitrogen fertilization rates and timings, irrigation rates and timings, sowing date, and planting density. The field trials were conducted under well-controlled conditions, with e.g. effective pest management. The commercial field dataset (n=77, 2015-2020) was collected from an arable farm located near Wageningen University and Research. As the primary focus of the commercial cultivation was profit rather than scientific research, crop management was carried out based on standard practices aiming to achieve an optimal balance between yield and costs.

A metamodel was pretrained on the synthetic dataset, and subsequently fine-tuned with the observational dataset. To prevent data leakage, we excluded the years in which the experiments of the potato trials dataset were performed from the synthetic dataset. For comparison, we trained a data-driven model with the same architecture of the metamodel. Because the observational dataset was relatively modest in size, we used leave-one-out cross-validation for model development and testing. As weather is a major determinant of final yield, we used weather years as criteria for the splits. To assess the stability of the training, we replicated the trainings three times using different random seeds.

As an additional baseline, representing one of the most commonly used data-driven approaches in yield prediction, we trained a simple linear regression model with a hand-picked feature set (i.e. earliness, sowing date, precipitation, and average daily temperature) selected by domain experts. The time series variables (i.e., precipitation, and average daily temperature) were preprocessed by averaging over the period from May-August.

3. Results

We (pre)trained a metamodel for potato yield prediction, using synthetic data obtained from the process-based crop growth model Tipstar. The obtained metamodel was able to reproduce Tipstar with an RMSE of 5.1 fresh tonne/ha and $r = 0.95$ on a hold-out set (n=13,000) (see Fig.2).

3.1. Transfer learning on synthetic data

In our first experiment we investigated the merits of incorporating fine-tuning in the training process of our metamodel, to cover for domain shifts. As an alternative approach we included a purely data-driven baseline that was not pretrained with the crop growth model.

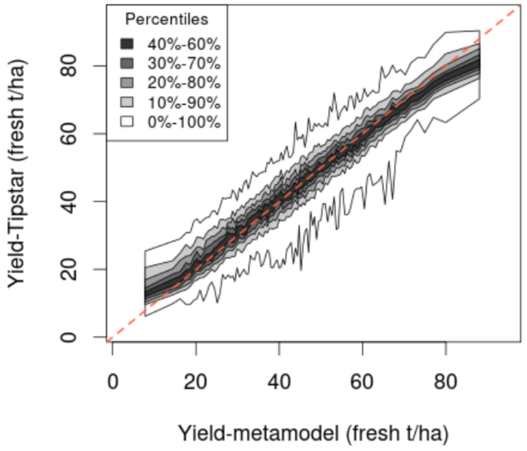

Figure 2. Fitness of the metamodel on the (hold-out) synthetic dataset (n=13,000; RMSE=5.1, black $r = 0.95$)

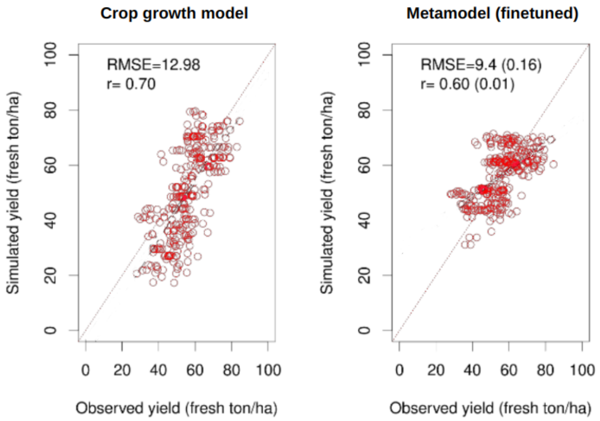

Figure 3. Fitness of the crop growth model and the metamodel on the field trial set (n=303). The number in parentheses represents the standard deviation of the three replicates of the metamodel trainings.

Table 1 reports on the performance of the models as function of the fine tuning set size. When target domain data availability is limited, best results are obtained with the fine-tuned metamodel. These results indicate that (1) fine-tuning improves the performance of a pretrained metamodel, and (2) pretraining with synthetic data obtained from a processed based crop model is effective, even when that data is collected from a source domain that differs from the target domain.

3.2. Yield prediction on real-world data

In our second experiment we evaluated our metamodel approach on real-world data.

Figures 3 and 4 show scatter plots of the predicted and the observed yields for both the crop growth model and the metamodel on the field trial set (n=303), and the commercial field set (n=77), respectively. In general, predictions of both models are better in the field trial set than in the commercial field set. While for the field trial set the predictions of the crop growth model are, on average, in line with the observed yields, for the commercial fields the crop growth model systematically underestimates the yield. This emphasizes the necessity of calibrating a crop growth model prior to its application to real-world data. The (fine-tuned) metamodel, on the contrary, exhibits no systematic bias, in either sets. Yet, the metamodel, has a relatively high scatter.

As a reference, we trained a purely data-driven model (i.e., we did not use the crop growth model for pretraining), with the same architecture of the metamodel. With an RMSE=15.80, and r=0.02 on the field trial set, and RMSE=14.10, and r=0.35 on the commercial set, the purely data-driven model showed poor performance. Arguably the complexity of the model was excessive considering the limited number of available training samples. We also trained a significantly less complex model, specifically a linear regressor, for which the feature set was hand picked by domain experts, with dedicated preprocessing. This model yielded an RMSE=8.93, and r=0.64 on the field trial set, and RMSE=9.75, and r=0.72 on the commercial field set, outperforming both the crop growth model and the metamodel on the latter dataset.

Table 1. Model performance as function of fine-tuning set size

METRIC	MODEL	FINE-TUNING SET SIZE			
		0^a	50	200	1,000
r	METAMODEL	*0.884*	0.899	0.905	0.908
	DATA-DRIVEN^b	N/A	0.767	0.790	0.902
RMSE	METAMODEL	*10.3*	7.6	7.4	7.2
	DATA-DRIVEN^b	N/A	11.2	11.0	7.5

[a] *Exclusively pretrained (on source domain)*
[b] *Exclusively trained on fine-tuning set*

Table 2. Model performance on real-world datasets

METRIC	MODEL	TRIAL	COMMERCIAL
r	CROP GROWTH MODEL	0.70	0.63
	METAMODEL	0.60	0.39
	DATA-DRIVEN	0.02	0.35
	LINEAR REGRESSION	0.64	0.72
RMSE	CROP GROWTH MODEL	12.98	24.90
	METAMODEL	9.40	14.10
	DATA-DRIVEN	15.80	14.10
	LINEAR REGRESSION	8.93	9.75

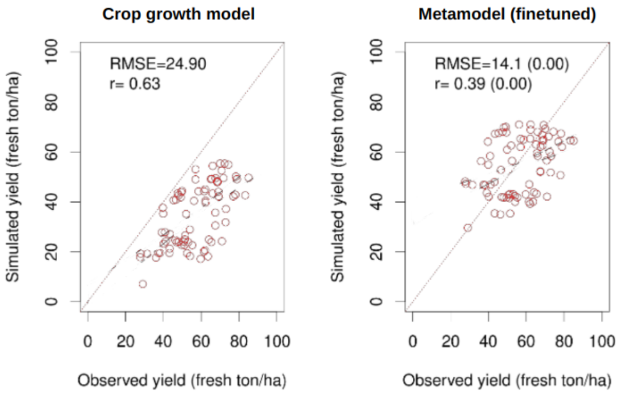

Figure 4. Fitness of the crop growth model and the metamodel on the commercial field set (n=77). The number in parentheses represents the standard deviation of the three replicates of the metamodel trainings.

4. Discussion and conclusion

In this work, we investigated potato yield prediction using a meta-modeling approach that integrates a *theory-driven* process-based crop growth model into a *data-driven* machine learning method. We found that, in silico, our metamodel yields better predictions than a purely data-driven approach. When tested on real-world data, comprising a field trial set and a commercial field set, the metamodel yields competitive results with respect to the crop growth model. In the latter set, which was characterized by a small sample size, however, a simple linear regression model with a hand-picked feature set, representing one of the most commonly used data-driven approaches in yield prediction, outperformed both the crop growth model and the metamodeling approach.

Our results indicate a benefit of pretraining a machine learning model with synthetic data obtained from a crop growth model, rather than using real-world data only. In our in silico experiments, pretrained models fine-tuned with only 50 data points had similar performance to data-driven models using 1,000 data points. This is important for data poor domains as agriculture. Also, training data exclusively obtained from a real-world setting may lack important contrasts in the input space, as, for example, especially in commercial settings, management practices are typically standardised. In this context, employing synthetic data may be seen as a data augmentation strategy that prevents the model from exploiting spurious features, which is especially relevant in models with high complexity. Model over-complexity combined with a small, low-contrast training set, may be one of the reasons why our purely data driven model did not perform well in the real-world datasets. Pretraining our model with synthetic data turned out to be a successfully strategy. It should be noted though that a simple approach, using an expert designed, regression model may still be preferred in a real-world setting.

We conclude that meta-modeling has potential for accurate crop yield prediction, yet further development and validation on large real-world datasets is recommended.

5. Impact

Predicting crop yields is crucial for ensuring food security, optimizing agricultural practices, and managing risks in the farming industry. In crop yield prediction, theory-driven process-based crop models and data-driven machine learning models are usually seen as opposite approaches. In this work, we explore the utilization of a meta-modeling approach that combines the strengths of both these approaches. By integrating these two methods, we aim to overcome their respective limitations. Process-based models often suffer from a restricted set of included processes, while data-driven models rely on large, diverse, and independent datasets that are often scarce in the agricultural domain. Our findings suggest the potential of meta-modeling for accurate crop yield prediction, and may as such facilitate the uptake of machine learning in agriculture.

Acknowledgements

We are grateful to three anonymous reviewers for their insightful comments. This work was partially supported by the Dutch Ministry of Agriculture, Nature and Food Quality project Data driven and High Tech (KB-38-001-002), the Wageningen University and Research Investment Theme *Digital Twins*, and the European Union Horizon Research and Innovation program (Grant #101070496, Smart Droplets).

References

Jansen, D. *Beschrijving van TIPSTAR : hét simulatiemodel voor groei en productie van zetmeelaardappelen*. Number 547 in Nota / Plant Research International. Plant Research International, Netherlands, 2008.

Leng, G. and Hall, J. W. Predicting spatial and temporal variability in crop yields: an inter-comparison of machine learning, regression and process-based models. *Environmental Research Letters*, 15(4):044027, 2020. doi: 10.1088/1748-9326/ab7b24.

Maestrini, B., Mimić, G., van Oort, P. A. J., Jindo, K., Brdar, S., Athanasiadis, I. N., and van Evert, F. K. Mixing process-based and data-driven approaches in yield prediction. *European journal of agronomy: the journal of the*

European Society for Agronomy, 139:126569, 2022. doi: 10.1016/j.eja.2022.126569.

Paudel, D., Boogaard, H., de Wit, A., van der Velde, M., Claverie, M., Nisini, L., Janssen, S., Osinga, S., and Athanasiadis, I. N. Machine learning for regional crop yield forecasting in Europe. *Field crops research*, 276: 108377, 2022. doi: 10.1016/j.fcr.2021.108377.

Porter, G. A., Bradbury, W. B., Sisson, J. A., Opena, G. B., and McBurnie, J. C. Soil management and supplemental irrigation effects on potato: I. soil properties, tuber yield, and quality. *Agronomy Journal*, 91(3):416–425, 1999. doi: 10.2134/agronj1999.00021962009100030010x.

Pylianidis, C., Snow, V., Overweg, H., Osinga, S., Kean, J., and Athanasiadis, I. N. Simulation-assisted machine learning for operational digital twins. *Environmental Modelling and Software*, 148:105274, 2022. doi: 10.1016/j.envsoft.2021.105274.

Razavi, S., Tolson, B. A., and Burn, D. H. Review of surrogate modeling in water resources. *Water Resources Research*, 48(7):W07401, 2012. doi: 10.1029/2011WR011527.

Seidel, S. J., Palosuo, T., Thorburn, P., and Wallach, D. Towards improved calibration of crop models–where are we now and where should we go? *European Journal of Agronomy*, 94:25–35, 2018.

van Genuchten, M. T. A closed-form equation for predicting the hydraulic conductivity of unsaturated soils. *Soil Science Society of America Journal*, 44(5):892–898, 1980. doi: 10.2136/sssaj1980.03615995004400050002x.

van Ittersum, M. K., Cassman, K. G., Grassini, P., Wolf, J., Tittonell, P., and Hochman, Z. Yield gap analysis with local to global relevance—a review. *Field Crops Research*, 143:4–17, 2013. doi: 10.1016/j.fcr.2012.09.009.

Wallach, D., Makowski, D., Jones, J. W., and Brun, F. Chapter 11 - Metamodeling. In Wallach, D., Makowski, D., Jones, J. W., and Brun, F. (eds.), *Working with Dynamic Crop Models (Third Edition)*, pp. 407–423. Academic Press, 3rd edition, 2019. doi: 10.1016/B978-0-12-811756-9.00011-3.

Wosten, J., de Vries, F., Hoogland, T., Massop, H., Veldhuizen, A., Vroon, H., Wesseling, J., Heijkers, J., and Bolman, A. BOFEK2012, de nieuwe bodemfysische schematisatie van Nederland. Technical report, Alterra, 2013.