# Location-specific vs location-agnostic machine learning metamodels for predicting pasture nitrogen response rate

Pattern Recognition. ICPR International Workshops and Challenges.

Pylianidis, Christos; Snow, Val; Holzworth, Dean; Bryant, Jeremy; Athanasiadis, Ioannis N.

https://doi.org/10.1007/978-3-030-68780-9_5

# Location-Specific vs Location-Agnostic Machine Learning Metamodels for Predicting Pasture Nitrogen Response Rate

Christos Pylianidis[1(✉)], Val Snow[2], Dean Holzworth[2,3], Jeremy Bryant[2], and Ioannis N. Athanasiadis[1]

[1] Wageningen University, Wageningen, Netherlands
{christos.pylianidis,ioannis.athanasiadis}@wur.nl
[2] AgResearch, Christchurch, New Zealand
{Val.Snow,Dean.Holzworth,Jeremy.Bryant}@agresearch.co.nz
[3] CSIRO, Brisbane, Australia

**Abstract.** In this work we compare the performance of a location-specific and a location-agnostic machine learning metamodel for crop nitrogen response rate prediction. We conduct a case study for grass-only pasture in several locations in New Zealand. We generate a large dataset of APSIM simulation outputs and train machine learning models based on that data. Initially, we examine how the models perform at the location where the location-specific model was trained. We then perform the *Mann–Whitney U test* to see if the difference in the predictions of the two models (i.e. location-specific and location-agnostic) is significant. We expand this procedure to other locations to investigate the generalization capability of the models. We find that there is no statistically significant difference in the predictions of the two models. This is both interesting and useful because the location-agnostic model generalizes better than the location-specific model which means that it can be applied to virgin sites with similar confidence to experienced sites.

**Keywords:** Machine learning · Process-based simulation · APSIM · Metamodels

## 1 Introduction

Environmental data are growing in an unprecedented way [8]. Many domains of Environmental Research utilize those data and combine them with Machine Learning (ML) techniques [7] to enable understanding. However, there are domains like grassland-based primary production systems where certain areas (e.g. pasture production, nitrogen leaching) have limited, low quality data, making them poor candidates for ML applications. In such areas, dynamic models are deployed to seek causality and make predictions based on first principles but sometimes they need data that is not available.


A. Del Bimbo et al. (Eds.): ICPR 2020 Workshops, LNCS 12666, pp. 45–54, 2021.
https://doi.org/10.1007/978-3-030-68780-9_5

ML has been used in a complementary way with dynamic models to summarize them and capture their embedded knowledge. The resulting ML models are also known as metamodels, surrogate models or emulators. The knowledge summarization is achieved by training ML models using the output of dynamic model simulations. Advantages of this technique include the reduction in need of observation data [1], the use of fewer inputs [10] and faster computation times [13] for large scale systems than the dynamic models. The paradigm of summarizing dynamic models is applied in several disciplines from physics [2] to hydrology [14].

Dynamic model summarization has also been studied in agriculture [11]. Several studies have examined the application of ML surrogate models for sensitivity analysis [4], the performance of different ML algorithms for crop model summarization [12] and the amount of data needed for accurate predictions [12]. In these works, the authors trained ML models in generated datasets to examine how well the models can generalize, using either one or all the available locations, and not testing in other locations. However, the generalization capability of a model over multiple locations does not mean that it performs better than a model specifically trained for that location. Since there are cases where the interest lies in absolute performance or generalizability of the summarization model it would be compelling to investigate how location-specific and location-agnostic models compare in those aspects.

The purpose of this work is to investigate the performance difference of location-specific and location-agnostic ML metamodels using a case study approach. To achieve this goal, we first generate a large dataset across several locations using a crop simulation framework. Second, we aggregate the generated data and train a ML model using all the available locations, and a second ML model using only one location. Next, we test the ML models on a dataset comprised of samples of the latter location. We compare the results using statistical metrics, and examine if they are statistically different using the *Mann–Whitney U test* [9] which has been used for comparing ML models in other works [5]. Finally, we investigate the trade-off between model performance and generalizability by testing the models in the rest of the locations of our dataset.

## 2 Materials and Methods

### 2.1 Case Study, Data Description

A case study was performed to predict the grass-only pasture nitrogen response rate in different locations in New Zealand. The application of nitrogen along with environmental factors such as temperature and time of year greatly affects pasture growth [3] so it is important to know the nitrogen response rate. Our dataset consisted of grass pasture growth simulations performed with the APSIM modeling and simulation framework [6]. A hyperspace of parameters was created and put to the simulator. The simulation parameters for APSIM included daily historical weather data from eight locations in New Zealand and management treatment options which can be seen in Table 1. The cross-product of those parameters

was used to create a hyperspace of input combinations for APSIM. The total number of simulations was 1,658,880 which should have yielded 1,382,400 nitrogen response rates. However, the input combinations included application of fertilizer at times when pasture growth was near zero because of dry soil conditions or cold temperatures. These were excluded from the analysis as the calculated N response rate was known to be unreliable. In total there were 1,036,800 response rates available for further analysis. Our target was to predict the 3-month nitrogen response rate – the additional pasture dry matter grown in the three months after fertilizer application over that from a non-fertilizer control divided by the kg of nitrogen in the fertilizer applied. The outputs of APSIM consisted of the nitrogen response rate, biophysical variables related to fertilizer concentration in grass and moisture in soil.

**Table 1.** The simulation parameters of APSIM. The cross-product of those parameters was used to create a hyperspace of input combinations.

| | Simulation parameters |
|---|---|
| Location | Weather from eight sites spanning the country |
| Soil water | 42 or 77 mm of plant-available water stored to 600 mm deep |
| Soil fertility | Carbon concentration in the top 75 mm of 2, 4, or 6% |
| Irrigation | Irrigated with a centre-pivot or dryland |
| Fertilizer year | All years from 1979 to 2018 |
| Fertilizer month | All months of the year |
| Fertilizer day | 5th, 15th and 25th of the month |
| Fertilizer rate | 0 (control), 20, 40, 60, 80 and 100 kg N /ha |

### 2.2 Data Preprocessing

The generated data were preprocessed to formulate a regression problem where the target variable was the nitrogen response rate and the inputs were the weather, some treatment options regarding the fertilizer and irrigation, and some biophysical variables. The generated data were aggregated from a daily to a simulation basis, to imbue memory to the data. First, the data were split into training and test sets to avoid information leakage during the latter stages of processing. The split happened based on the year, taking one year to the test set every five years and the rest to the training set. The resulting percentage of training and test samples was 80/20%. Second, from the generated daily data only the samples in a window of 28 days before fertilization were kept. This range was selected because grass pasture is known to not be affected by past conditions further than this window provided it is not under- or over-grazed. Also, weather data after the first fertilization was not considered because preliminary work has shown that it is not needed to achieve meaningful results. Third, only the variables related to the weather, simulation parameters, nitrogen response

rate and to some of the biophysical variables were preserved which were considered to be likely drivers, based on expert knowledge of the nitrogen response rate. Fourth, the weather and biophysical variables were aggregated using their weekly mean values. Finally, the aforementioned steps were repeated once to form an aggregated dataset containing all the locations, and once for each of the eight locations contained in our dataset. The output of those steps was an aggregated dataset (training set) for the location-specific model, an aggregated dataset (training set) for the location-agnostic model, and an aggregated dataset (test set) for each location.

### 2.3 Machine Learning Pipeline

The aggregated datasets were then passed to the ML stage. In this stage, the training and test data were standardized using the same data transformer to keep the same mean for both transformations. To clarify further, each test set was using the scaler of the location-agnostic model and the location-specific model so that each model can have a version of the test set according to the mean of its training set. Categorical variables were converted to ordinal by substituting them with numbers. Then, hyperparameter optimization was performed to the Random Forest algorithm using gridsearch with 5-fold cross-validation. The gridsearch parameters were *n_estimators* {200, 300, 400, 500}, *max_depth* {3, 5, 7, 11}, *min_samples_split* {2, 3, 4, 8, 16}, *min_samples_leaf* {1, 2, 4, 8, 16} and *max_features* {0.33, sqrt, None}. The out-of-bag score was used for the building of the Random Forest trees. No feature selection was performed because the number of features was small (64) compared to the size of the training datasets (1,044,060 and 130,095 samples for the multiple and single locations correspondingly). After training, the optimized models of the location-agnostic and location-specific models were tested using the test set of location *Waiotu* where the location-specific model was trained. The pipeline of the ML stage is shown in Fig. 1.

### 2.4 Evaluation

The performance of the location-specific and location-agnostic models was first evaluated by comparing error metrics (MAE, RMSE, $R^2$) of their results on the test set. Then, the *Mann–Whitney U test* was performed on the models' results on the test set to see if the differences were significant. The *Mann–Whitney U test* examines if the distributions of the populations of two groups are equal and it was preferred among other statistical tests because first it is non-parametric, second it assumes that the pairs in the samples do not come from the same populations and third that the observations are ordinal, all of which fit our problem. Consequently, error metrics and the *Mann–Whitney U test* were calculated for the rest of the locations to test the models' generalizability.

### 2.5 Implementation

The data preprocessing stage was developed utilizing the Apache Spark framework. The ML models were developed using the *scikit-learn* library in Python. The experiments took place in a Databricks node consisting of 96 cores and 384 GB of RAM to speed up procedures through parallelization.

## 3 Results

The hyperparameter tuning procedure selected the following parameters for both models: *n_estimators* 400, *max_depth* 11, *min_samples_split* 2, *min_samples_leaf* 1, *max_features* 0.33. The results of the ML models on the training and test sets are shown in Fig. 2, along with the distributions of the simulation and the model predictions. We observe that the angle between the identity and regression lines on the test set is smaller for the location-specific model which means that it fits better the location-specific test data. The data points on the test set of the location-agnostic model are more dispersed. Also, we notice that the distributions of the location-specific and location-agnostic model predictions on the test set appear to be similar. The mean and variance of the distributions appear to be close as it can be seen in Table 2.

Regarding the error metrics, in Table 3 we observe the Mean Average Error, Root Mean Square Error and coefficient of determination ($R^2$) for both models on the test set of each location. For the location where the location-specific model was trained (Waiotu), we observe that the location-specific model performs better than the location-agnostic model. For the rest of the locations, the location-agnostic model outperforms the location-specific one.

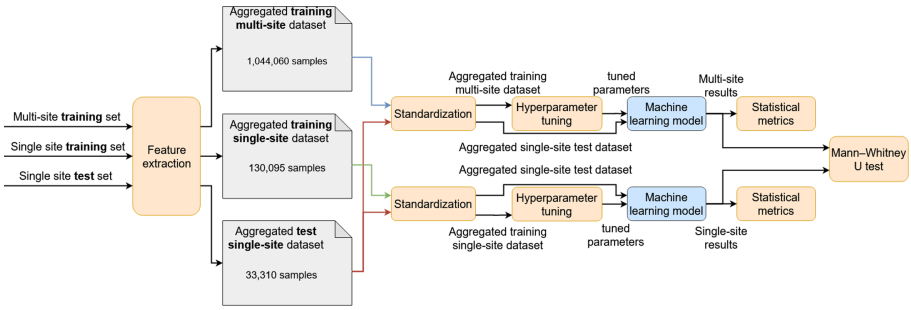


**Fig. 1.** The pipeline for the training and testing of the models on the location where the location-specific model was trained. At the end there is also the evaluation stage. The process starts by taking the training and test datasets from the preprocessing stage. It has to be explicitly noted that hyperparameter tuning was performed only on the training set. More specifically the test set was the same for both models but it was standardized for each model individually to preserve the same mean which was used for each training set.

**Table 2.** The distribution characteristics of the two models for the test set predictions on the location where the location-specific model was trained.

| | Location-specific | Location-agnostic |
|---|---|---|
| Mean | 18.47 | 18.44 |
| Variance | 44.73 | 40.99 |
| Skewness | 0.11 | 0.18 |
| Kurtosis | −0.90 | −0.82 |

In Table 3 we also observe the results of the *Mann–Whitney U test* for each location. For the location where the location-specific model was trained (*Waiotu*) we see that there is no statistically significant difference between the models. The same applies to the location *Ruakura*.

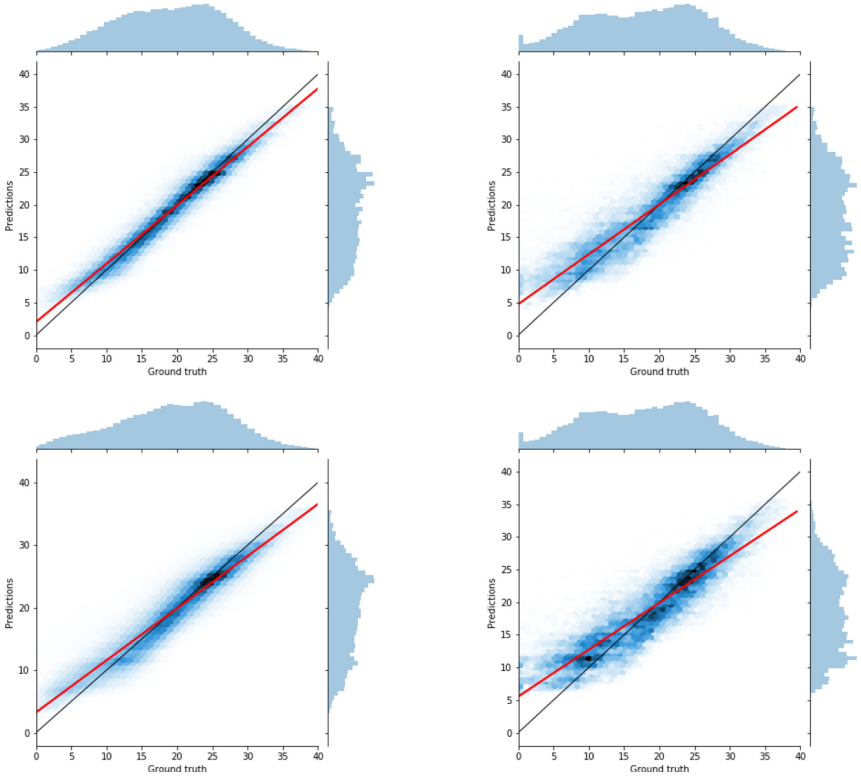


**Fig. 2.** The results of the **location-specific model (top row)** and **location-agnostic model (bottom row)** for the **training (left)** and **test (right)** sets. The test set is common for both models and contains data from the location where the location-specific model was trained (*Waiotu*). On the vertical axes are the predictions of the model and on the horizontal axes the simulated values. On top and right of the plots are the distributions of the simulated and predicted values correspondingly. The black lines are the identity lines. The red lines are the regression of Prediction on Ground truth. Darker spots indicate that more predictions fall on the same area. (Color figure online)

**Table 3.** The error metrics of the location-specific and site agnostic models on the different locations. On the first row are the locations existing in our dataset. *Waiotu* is the location where the site specific model was trained. On the second row are the mean absolute error (MAE), root mean squared error (RMSE) and coefficient of determination ($R^2$), for each model and location. The blue and red colors indicate the models with the highest and lowest performance correspondingly, for each location and error metric. On the third row, statistically significant difference on *Mann–Whitney U test* between the predictions of the two models is denoted with as asterisk.

| | | **Waiotu** | **Ruakura** | **Wairoa** | **Marton** | **Mahana** | **Kokatahi** | **Lincoln** | **Wyndham** |
|---|---|---|---|---|---|---|---|---|---|
| Location-specific | MAE | 2.37 | 2.72 | 2.92 | 3.27 | 3.44 | 4.36 | 4.96 | 5.62 |
| | RMSE | 3.19 | 3.62 | 4.03 | 4.41 | 4.2 | 5.81 | 6.63 | 7.29 |
| | $R^2$ | 0.85 | 0.78 | 0.68 | 0.66 | 0.66 | 0.5 | 0.41 | 0.38 |
| Location-agnostic | MAE | 2.71 | 2.13 | 2.71 | 2.06 | 2.29 | 2.56 | 2.88 | 2.31 |
| | RMSE | 3.55 | 2.95 | 3.91 | 2.83 | 3.04 | 3.33 | 4.08 | 3.06 |
| | $R^2$ | 0.81 | 0.85 | 0.7 | 0.86 | 0.82 | 0.83 | 0.78 | 0.89 |
| MannWhitney U test | | | | * | * | * | * | * | * |

## 4 Discussion

The results showed slightly better error metrics for the location-specific model over the location-agnostic model for *Waiotu*. The reason may be that the location-specific model learns the local conditions better since they are only from this location and fewer than those included in the training of the site-agnostic model. For the rest of the locations, the location-agnostic model performs better because it was trained with more data, which also included these locations and as a result, it can generalize better. An interesting finding is that the errors of the location-specific model increase as we move further away from *Waiotu*, as shown in Fig. 3. The locations can be seen in Fig. 4. This finding indicates that the further away a prediction is made from the training location, the higher the error will be for a location-specific model. On the other hand, the location-agnostic model is not affected since it was trained in a larger dataset which included data from those locations.

Another finding was that there was no statistical difference between the predictions of the two models for *Waiotu*. The location-specific model may perform better but it seems that the gain is marginal and is lost when moving to other locations. The second location with no statistical difference between the models' predictions is *Ruakura*. We assume that this happens because *Ruakura* and *Waiotu* are close to each other and as a result, environmental factors do not vary substantially between those locations.

We deduct that there seems to be a trade-off between accuracy and generalization performance. The location-specific model is trained on a smaller dataset and overfits the data. As a result it performs better for *Waiotu* but the location-agnostic model generalizes better. In our opinion, the decision for which model to deploy depends on the use. We emphasize though that the performance difference in this case study is not dramatic for *Waiotu*. On the other hand, the generalization performance is evident especially as we move further away from the location where the location-specific model was trained.

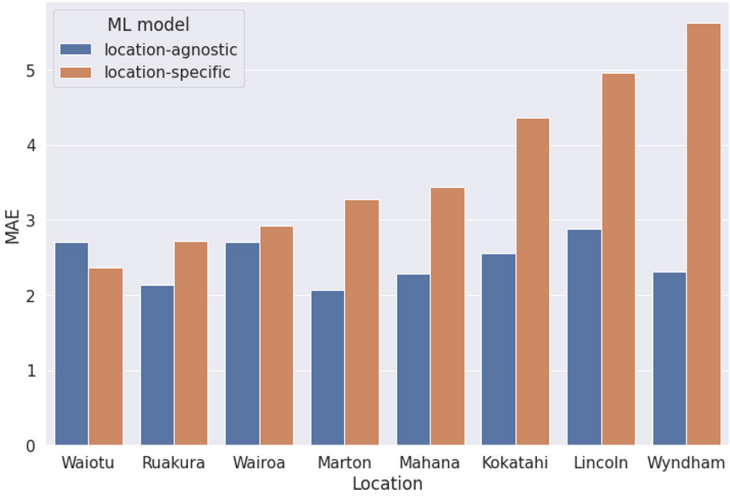


**Fig. 3.** The mean absolute error (MAE) of the location-specific and agnostic models for all the locations in our dataset. On the vertical axis is the error and on the horizontal the locations. The orange and blue colors indicate the results of the location-specific and agnostic models respectively. (Color figure online)


**Fig. 4.** The locations in New Zealand which were included in our dataset. On the top right is *Waiotu* which was used to train the location-specific model. Right next to *Waiotu* is *Ruakura*. The rest of the locations are further away.

## 5 Limitations

A limitation of our study regarding the performance comparison of the ML models is that the location-agnostic model was trained using data from all the locations. As a result we did not test how the models would perform in a location that would be new to both of them.

Another limitation is that the performance of both models was affected by the way we partitioned years into the training/test split. That is due to seasonality in the generated data, which was not taken into account when performing the split.

## 6 Conclusion and Future Work

In this work, we examined the performance difference between a location-specific and a location-agnostic metamodel using error metrics and the *Mann–Whitney U test*. We tested the models in different locations including the location where the location-specific model was trained. We found that the location-specific model performs better for the location where it was trained, although not in a statistically significant way. Also, the error metrics in other locations showed that the location-agnostic model generalizes better.

Future work could include the setup of the methodology in a way to test location-specific models for all the available locations to examine if the results will be the same. Also, a location could be left out of both training sets to allow testing in a new location for both models. Besides, different machine learning algorithms could be deployed and tuned even further. The performance of the models could also be improved by adding complex features and features based on agronomic knowledge.

**Acknowledgements.** This work has been partially supported by the European Union Horizon 2020 Research and Innovation programme (Grant #810775, Dragon); the Wageningen University and Research Investment Programme "Digital Twins" and AgResearch Strategic Science Investment Fund (SSIF) under "Emulation of pasture growth response to nitrogen application".

## References

1. Albert, A.T., Rhoades, A., Ganguly, S., Feldman, D., Jones, A.D., Prabhat, M.: Towards generative deep learning emulators for fast hydroclimate simulations. In: AGU Fall Meeting Abstracts, vol. 2018, pp. IN21C-0723, December 2018
2. Garrido Torres, J.A., Jennings, P.C., Hansen, M.H., Boes, J.R., Bligaard, T.: Low-Scaling algorithm for nudged elastic band calculations using a surrogate machine learning model. Phys. Rev. Lett. **122**(15), 156001 (2019). https://doi.org/10.1103/PhysRevLett.122.156001
3. Gillingham, A.G., Morton, J.D., Gray, M.H.: Pasture responses to phosphorus and nitrogen fertilisers on east coast hill country: 2. Clover and grass production from easy slopes. N. Z. J. Agric. Res. **51**(2), 85–97 (2008). https://doi.org/10.1080/00288230809510438

4. Gladish, D.W., Darnell, R., Thorburn, P.J., Haldankar, B.: Emulated multivariate global sensitivity analysis for complex computer models applied to agricultural simulators. J. Agric. Biol. Environ. Stat. **24**(1), 130–153 (2018). https://doi.org/10.1007/s13253-018-00346-y
5. Goetz, J.N., Brenning, A., Petschko, H., Leopold, P.: Evaluating machine learning and statistical prediction techniques for landslide susceptibility modeling. Comput. Geosci. **81**, 1–11 (2015). https://doi.org/10.1016/j.cageo.2015.04.007
6. Holzworth, D.P., et al.: APSIM - evolution towards a new generation of agricultural systems simulation. Environ. Model. Softw. **62**, 327–350 (2014). https://doi.org/10.1016/j.envsoft.2014.07.009
7. Lima, A.R., Cannon, A.J., Hsieh, W.W.: Nonlinear regression in environmental sciences using extreme learning machines: a comparative evaluation. Environ. Model. Softw. **73**, 175–188 (2015). https://doi.org/10.1016/j.envsoft.2015.08.002
8. Lokers, R., Knapen, R., Janssen, S., van Randen, Y., Jansen, J.: Analysis of big data technologies for use in agro-environmental science. Environ. Model. Softw. **84**, 494–504 (2016). https://doi.org/10.1016/j.envsoft.2016.07.017
9. Mann, H.B., Whitney, D.R.: On a test of whether one of two random variables is stochastically larger than the other. Ann. Math. Stat. **18**(1), 50–60 (1947). https://doi.org/10.1214/aoms/1177730491
10. Ramanantenasoa, M.M.J., Génermont, S., Gilliot, J.M., Bedos, C., Makowski, D.: Meta-modeling methods for estimating ammonia volatilization from nitrogen fertilizer and manure applications. J. Environ. Manage. **236**, 195–205 (2019). https://doi.org/10.1016/j.jenvman.2019.01.066
11. Ramankutty, P., Ryan, M., Lawes, R., Speijers, J., Renton, M.: Statistical emulators of a plant growth simulation model. Clim. Res. **55**(3), 253–265 (2013). https://doi.org/10.3354/cr01138
12. Shahhosseini, M., Martinez-Feria, R.A., Hu, G., Archontoulis, S.V.: Maize yield and nitrate loss prediction with machine learning algorithms. Environ. Res. Lett. **14**(12), p. 124026, December 2019. https://doi.org/10.1088/1748-9326/ab5268
13. Weber, T., Corotan, A., Hutchinson, B., Kravitz, B., Link, R.: Technical note: deep learning for creating surrogate models of precipitation in earth system models. Atmos. Chem. Phys. **20**(4), 2303–2317 (2020). https://doi.org/10.5194/acp-20-2303-2020
14. Zhang, R., Zen, R., Xing, J., Arsa, D.M.S., Saha, A., Bressan, S.: Hydrological process surrogate modelling and simulation with neural networks. In: Lauw, H.W., Wong, R.C.-W., Ntoulas, A., Lim, E.-P., Ng, S.-K., Pan, S.J. (eds.) PAKDD 2020. LNCS (LNAI), vol. 12085, pp. 449–461. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-47436-2_34