Hybrid Machine Learning and process-based modelling approaches for climate adaptation strategies

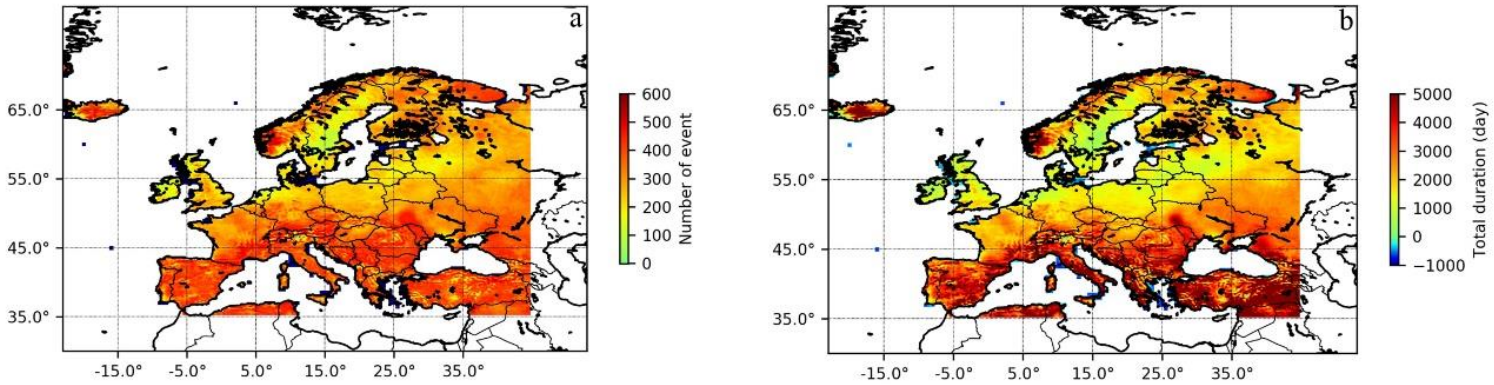
George van Voorn (*Biometris / applied mathematics & statistics*)

Data driven discoveries for climate change, November 6, 2024

WAGENINGEN
UNIVERSITY & RESEARCH

Climate change impacts agri-food value chain

- Climate change is happening
- Crops, post-harvest products, product lines, etc. affected by changes in temperature, precipitation, extreme events, compound events
- Quantitative assessment for climate adaptation



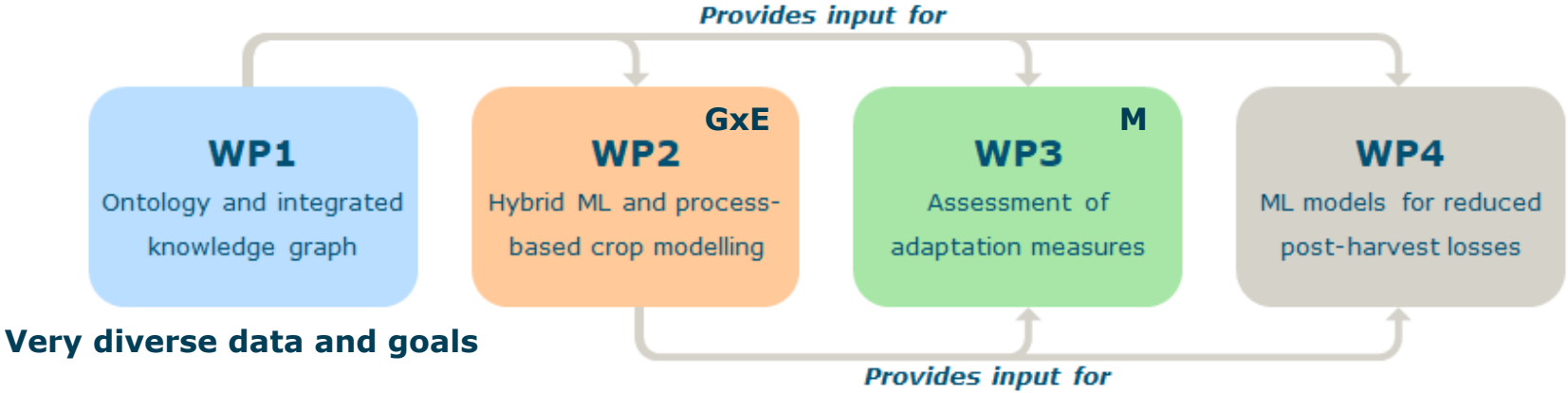
Current and future heatwaves (Samuel Sutanto, HPC: /lustre/nobackup/WUR/ESG/sutan001/Heatwaves/*.nc)

Hybrid machine learning: what and why?

- ML: use big data to train models for forecasting impacts
- But... Data contain ‘pollution’, ML is black box, over-fitting and equifinality, and importantly: future conditions (we do not have the data...)

- Hybrid ML uses process-based knowledge to augment data-driven modelling and (if possible) extrapolate beyond data domain
- ***Goal***: Explore utility of hybrid ML for quantitative climate impact assessment

D3-C2 project set-up



- Team:
  - WU: PSG, ESG, SSG;
  - WR: WECR, WENR, WFBR, WFSR, WPR

WP1: Data sharing

- Data reuse is often ad hoc and laborious
- ***Goal:*** improve (re)usability of data for (hybrid) ML, particularly for combining data from different sources
- Explored options:
    - README files (human-readable)
    - Croissant files (computer-readable)
    - Knowledge graph-based access layer

Human vs machine readable

AN.SOL

Type: cr:FileObject
Id: AN.SOL
Description: Data from Harvard Dataverse
ContentSize: 1412 B
ContentUrl:
EncodingFormat:
CiteAs:
License:
DataType: https://schema.org/URL, https://...AFFAIR.../Soil, https://...GloSIS.../SoilInfo

Comment:
Source:

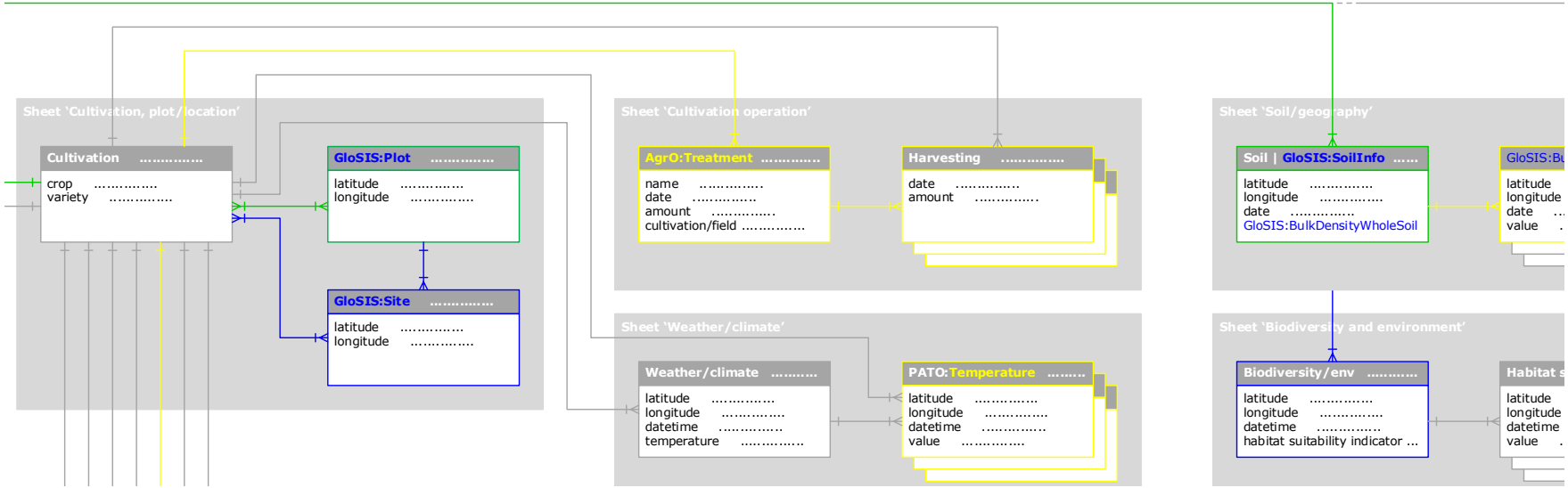
Columns:

Name	Datatype	Unit	Range	Resolution	Annotation (same as)
SCS FAMILY					soil type | glosis_pr:soilTypeProperty
LAT		°			Latitude
LONG		°			Longitude
SLB					soil layer base depth | glosis_cm:SoilDepth
SLLL					lower limit
SDUL					drained upper limit
SSAT					saturation | glosis_lh:BaseSaturation
SRGF					root growth factor, soil only


```
{
 "@context": {
  "@language": "en",
  "@vocab": "https://schema.org/"
 },

 "name": "AN.SOL",

 "@type": "cr:FileObject",
 "@id": "AN.SOL",
 "description": "Data from Harvard Dataverse",
 "contentSize": "1412 B",
 "contentUrl": "",
 "encodingFormat": "",
 "citeAs": "",
 "license": "",
 "url": "",
 "dataType": [https://schema.org/URL, https://...AFFAIR.../Soil,
http://...GloSIS.../SoilInfo],
 "field": [
  {
   "@id": "AN.SOL/SCS FAMILY",
   "@type": "cr:Field",
   "dataType": [https://schema.org/URL, https://...AFFAIR.../soil type,
https://...AFFAIR.../glosis_pr:soilTypeProperty]
  },

...
```


A lot is still empty…

Knowledge graph based



- Ontologies (including naming) and representations differ per data sets
- "Draw in" data sets into encompassing framework *without* the need for final decisions on definitions, relationships, naming, etc.

Collected data

- Aberdeen, Idaho, potato variety trials
  USDA ARS - Agricultural Research Service
- Data from global field experiments for potato simulations
  Harvard Dataverse, ODjAR (Open Data Journal for Agricultural Research) Dataverse
- Potato, Unifarm WP2
- Web Soil Survey (in ArcGIS format) USDA
- Data_sharing.xlsx This project
- Indicators Global-Detector V17.docx This project
- Crop growth time series WP2 ETH Zürich
- Climate data to run WOFOST in India (the whole country and pilot sites) and Wageningen, NL This project

WP1 discussion

- All three approaches found useful for researchers
- Preparation of different data sets in this project still laborious: a lot of aspects are missing, non-matching ontologies,...

Under construction

- ***Outlook***: methods for (semi-)automatic processing, e.g. Python tooling for croissant files (or ChatGTP?)

WP2: GxE crop modelling

- ***Climate adaptation strategy***: different varieties of crops
- Crop target traits for breeders such as yield result from complicated genetic-by-environment interactions (GxE)

- ***Goal***: models that capture essential GxE to forecast these traits in new environments, with focus on secondary traits underlying yield
- Main focus: wheat, potato

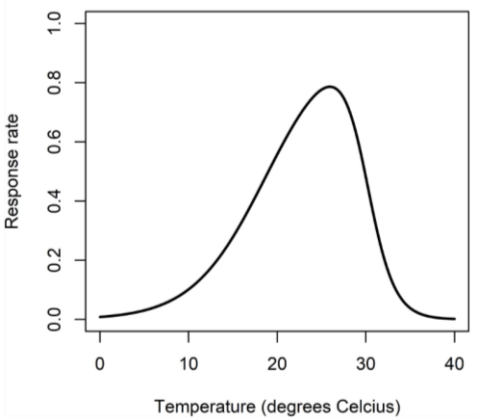
Issues and approaches

- Data on GxE fragmented; augmented by process-based modelling
- But... Temperature response functions (TRFs) in crop growth models (CGMs):
    - Large diversity in TRFs: major source of uncertainty
    - ‘Broken-stick’; ML fitting requires continuous-smooth functions

- Adaptation of TRF
- Low-complexity dynamic models with key environmental factors, used in
    - Bayesian approach for fitting of multiple genotypes / environments
    - PIML (Process Informed Machine Learning)

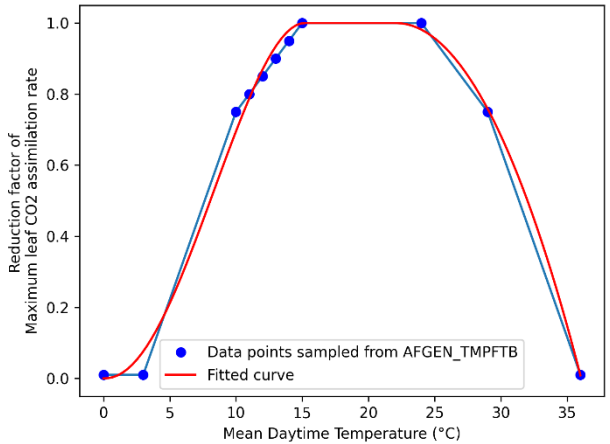
Temperature response

- In context of low-complexity, dynamic models (wheat data)



Van Voorn et al. (2023) Frontiers

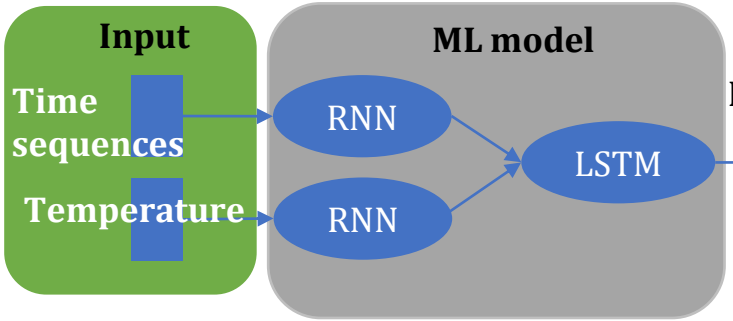
- In CGM (Wofost-Potato)



Liu, Van Voorn, De Wit et al. (Subm.)

WAGENINGEN
UNIVERSITY & RESEARCH

Physics-Informed Machine Learning



$$\text{data loss} = \frac{1}{T_n}\sum_{t=0}^{T_n}(Y(t) - Y_n(t))^2$$

$$+$$

$$\text{physics loss} = \frac{1}{T_m}\sum_{t=0}^{T_m}\left(\frac{dY_m(t)}{dt} - \frac{\partial \tilde{Y}_m(t)}{\partial t}\right)^2$$

Secondary trait (e.g., biomass)

Rate parameters

Environmental exogenous input (e.g., soil, weather)

$$\frac{dY_{sij}(t)}{dt} = f(Y_s(t), \theta, \omega(t))$$

$$\theta_{ij}^{p} = \mu^{p} + \alpha_i^{p} + \beta_j^{p} + (\alpha \cdot \beta)_{ij}^{p}$$

Genotype × Trial effect

Mean effect

Genotype effect

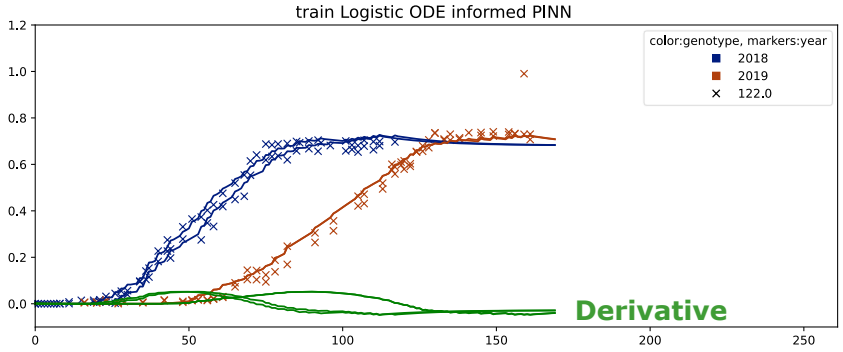
Trial effect

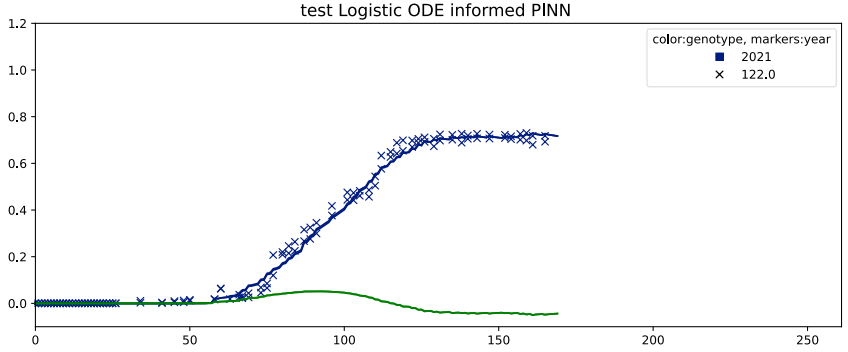
WP case 1: wheat

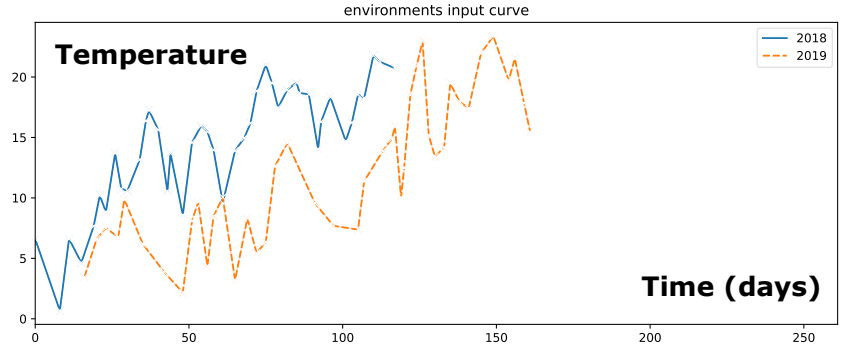
- Experimental / observational data for wheat:
  - Medium time-resolution, non-destructive wheat trait observations like height, leaf area, spikelet, ...
  - Sensor data from several soil sensors with 10 cm depth intervals

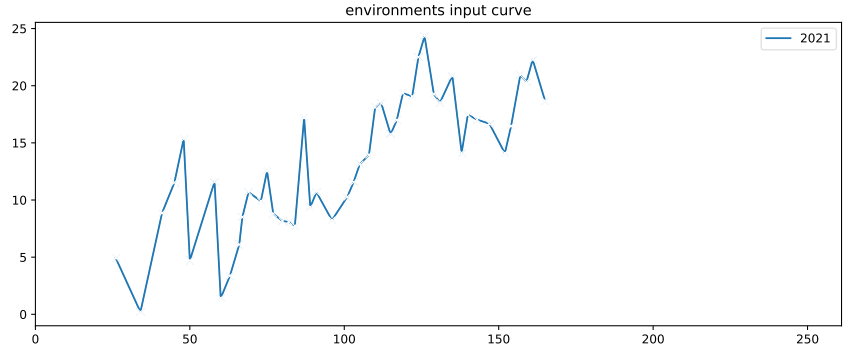
Courtesy of Olivia Zumsteg, Lukas Roth (ETH Zürich)
https://kp.ethz.ch/infrastructure/FIP.html

Results with PIML









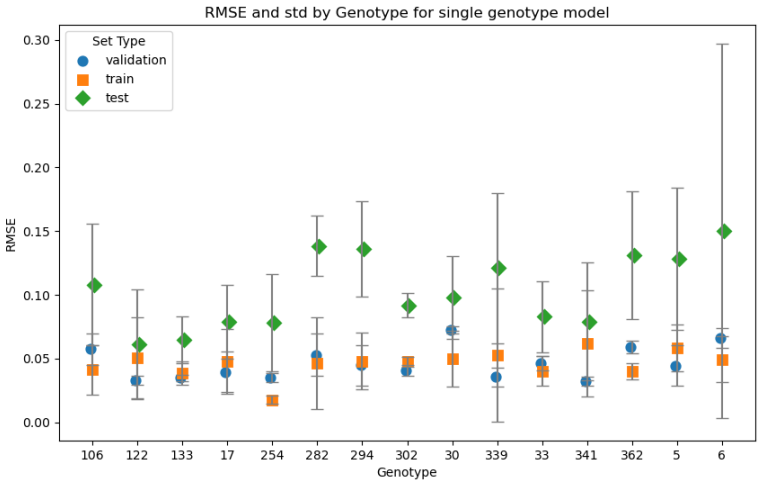
WAGENINGEN
UNIVERSITY & RESEARCH

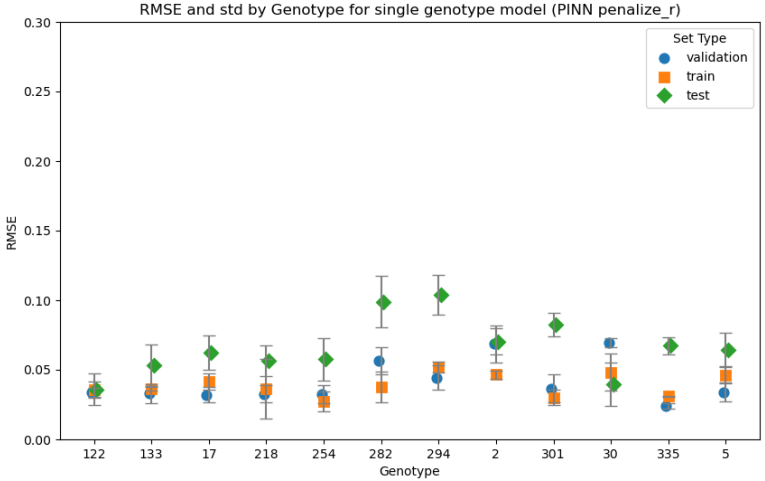
Comparison between models

Model	Train RMSE	Validation RMSE	Test RMSE	Train std	Validation std	Test std
RNN+LSTM	0.046	0.045	0.101	0.022	0.008	0.045
Random forest	0.021	0.178	0.152	0.003	0.001	0.002
PINN	0.041	0.044	0.078	0.007	0.006	0.028
PINN (penalize_r)	0.041	0.045	**0.073***	0.007	0.007	0.019
Logistic ODE	0.068	0.067	0.099	*NA*	*NA*	*NA*

Covering 19 wheat genotypes from multiple seasons (data from ETH Zürich)

ML vs PIML (for single genotypes)

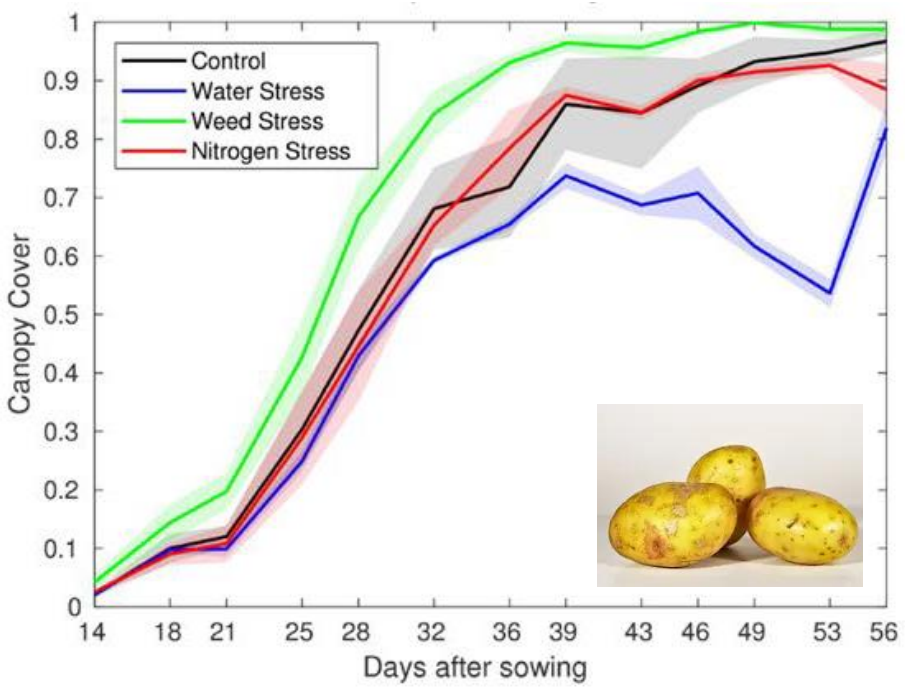




- PIML improves prediction performance in test set and reduces standard deviation

WP2 case 2: potato

- Different stresses
- ***Goal***: model for classification of potato stresses based on canopy reflectance (such as nitrogen deficiency, poor tuber emergence, presence of weeds)
- Hybrid ML approach combines process model (Tipstar) with LSTM NN



WAGENINGEN UNIVERSITY & RESEARCH

Bernardo Maestrini, Jian Liu, a.o.

Hybrid model workflow

Crop growth model (Tipstar) → PRO4SAIL simulate canopy1 spectra →

Synthetic dataset

Run #	Stress	date	Clay cont.	Planting density	N Fert KgN/ ha	Irr. mm	NDVI
1	xx						

Management, soil, weather, canopy, NDVI

Model

↓

Stress label (no stress, N stress, poor Emergence, weeds)

↓

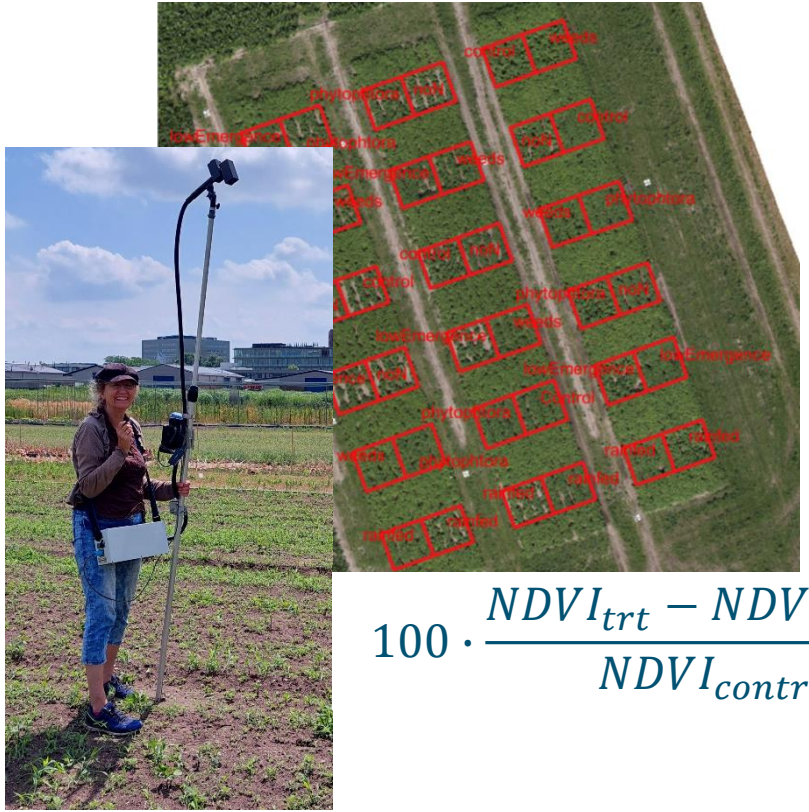
AI MODEL trained on synthetic data
Disturbance = function(mgt, weather, reflectance)

↓

Collect field data
Validation on field data 2023 and 2024

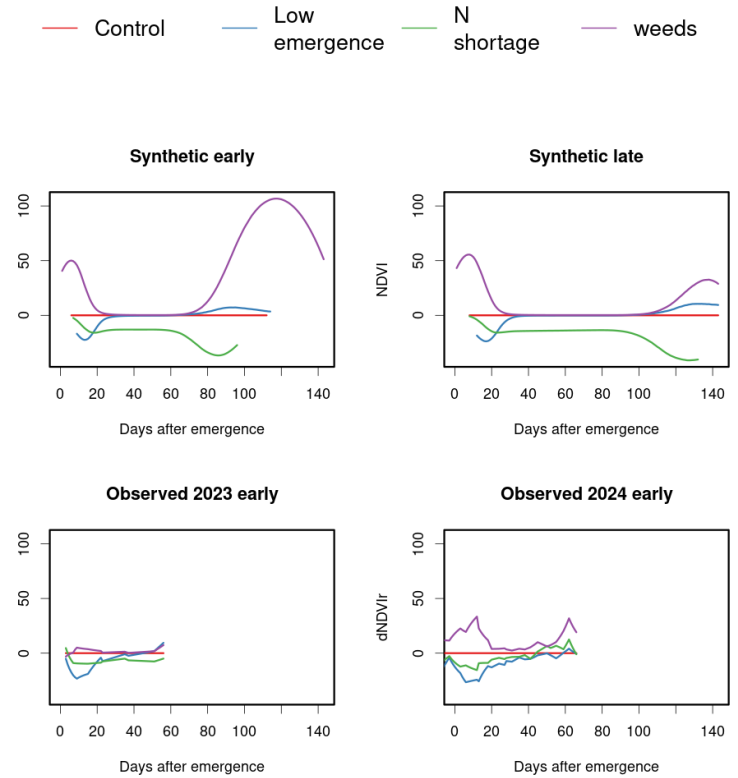
WAGENINGEN UNIVERSITY & RESEARCH

Validation experiment



$$NDVI$$

$$100 \cdot \frac{NDVI_{trt} - NDVI_{control}}{NDVI_{control}}$$



WAGENINGEN UNIVERSITY & RESEARCH

Prediction accuracy

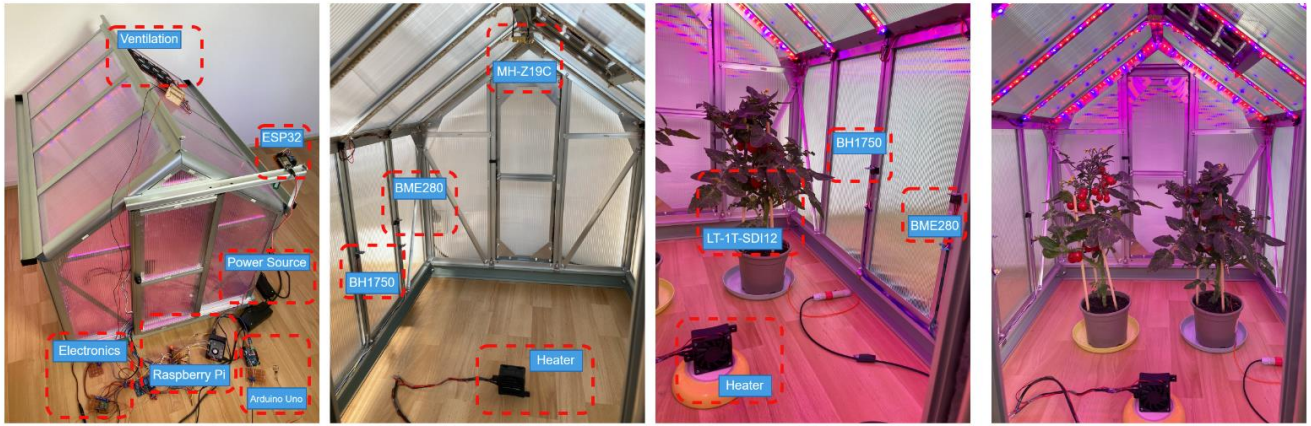
Year	Cultivar	Weeds	N shortage	Low emergence	Control	Average
2023	Early	0.80	0.80	0.50	0.5	0.65
2024	Early	0.88	0.67	0.79	0.5	0.71
2023	Late	0.90	1.00	0.80	0.5	0.80
2024	Late	0.71	0.79	0.62	0.5	0.66
Average		0.82	0.82	0.68	0.5	0.70

- Low emergence proved to be the hardest to predict because as soon as the canopy closes there is no difference with control

WP2 case 3: greenhouse

- ***Goal***: develop a data-driven model (GRU, LSTM) for forecasting growth and final yield with limited sensing data in greenhouse; https://github.com/patatasal/WUR_INF_CropModel
- Automatic calibration for decreasing residuals between simulation model and data; GitHub - EfraimManurung/mini-greenhouse-model

Mini-greenhouse for proof-of-principle



(a) Mini-greenhouse look from the outside with components.

(b) Mini-greenhouse look from the inside (left-side) with the sensors and heater.

(c) Mini-greenhouse look from the inside (right-side) with the tomato plants, sensors and heater.

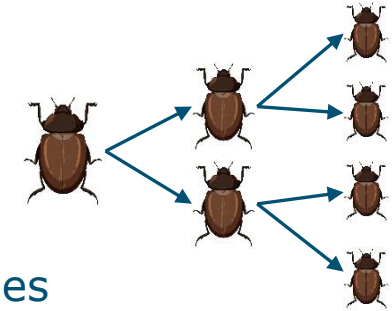
(d) Toplights on from inside.

WP2 discussion

- PIML not (yet) commonly used in crop modelling; relatively short time spans compared to e.g. economic time series used for LSTM / GRU models
- PIML seems to be able to improve performance compared to solely ML or process-based models
- ***Outlook***:
    - Exploring options including alternative ODE models and alternative physics-based ML (e.g. neural ODEs)
    - Include (more) genotypic variation in PIML
    - As part of controller loop in greenhouse
- Critical issue of missing data (*the future*) reduced but remains...

WP3: Pest management

- Climate change affects pests and diseases, e.g.
  - Increased no. of generations
  - Increased survival over winters
  - Desynchronization of pests and their natural enemies



- ***Climate adaptation strategy***: pest management

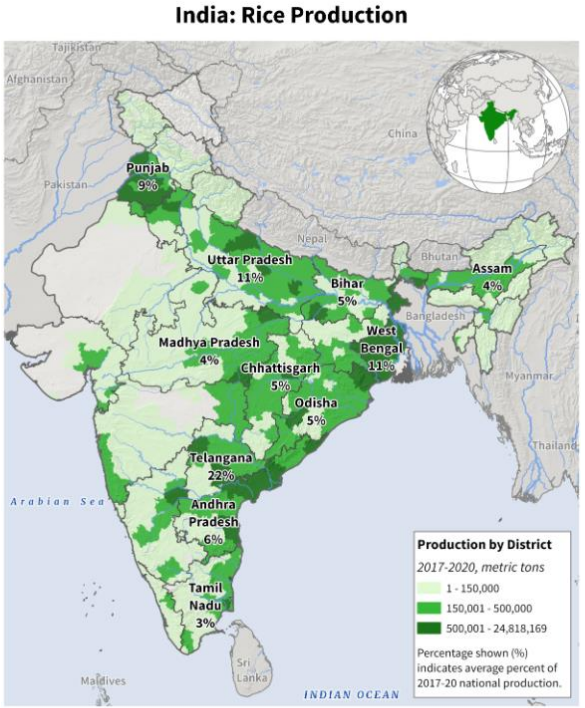
- Crop modelling typically does not account for pests and disease
- ML approach to forecasts outbreaks

Case study: yellow stem borer on rice in India

- India is 2nd largest rice producer (from 53.6**M** tons in 1980 to 130**M** tons in 2023)
- Insects cause 25% rice production loss = 30**B** USD
- 20% of pesticide use in India

‘dead heart’ symptom

‘white ear’ symptom



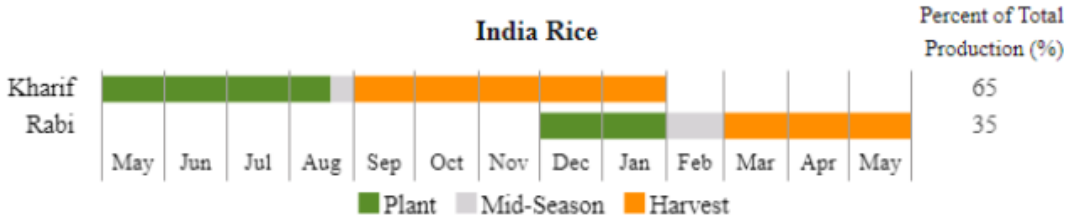
USDA Foreign Agricultural Service
U.S. DEPARTMENT OF AGRICULTURE

Source: India Ministry of Agriculture, Directorate of Economics and Statistics, Market Year 2017/18 - 2019/2020 data by districts

WAGENINGEN
UNIVERSITY & RESEARCH

Data sources

- Field observations by IRRI (International Rice Research Institute)
- Rajendranagar region

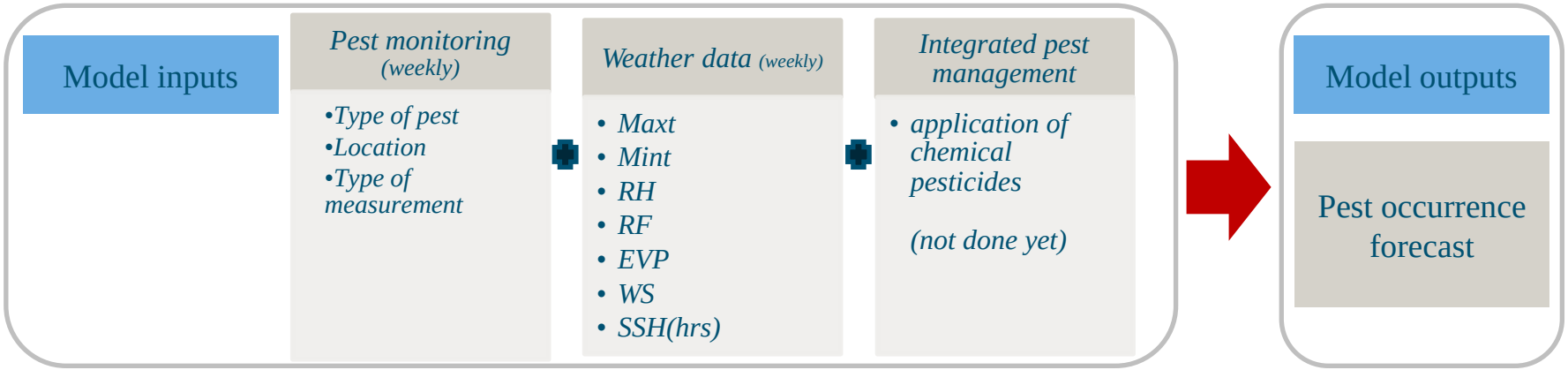


Yellow stem borer

https://www.irri.org/where-we-work/countries/india

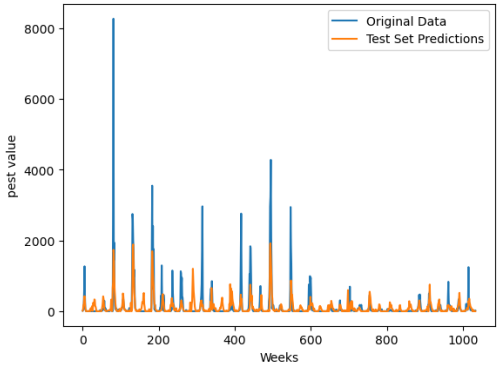
Machine Learning model

- Long Short Term Memory-based model

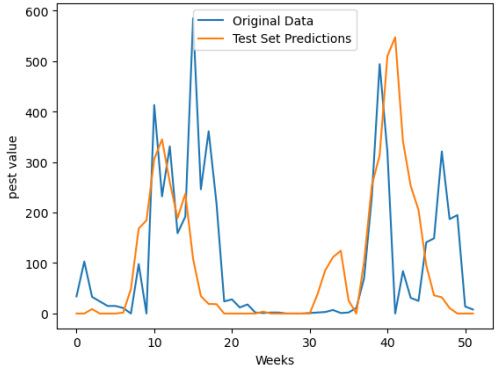


Model results

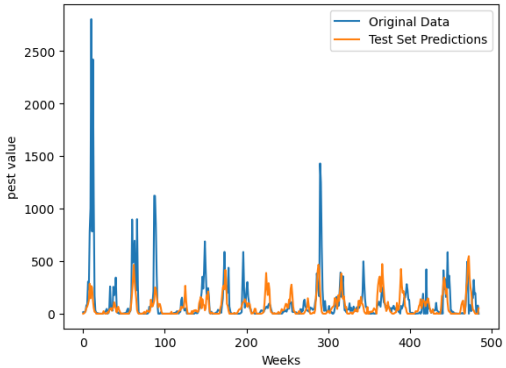
Training dataset (1974-2000)



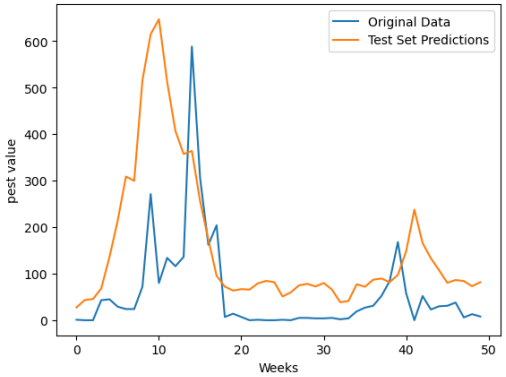
Validation (2010)



Testing dataset (2001-2009)



Validation (2011)



WAGENINGEN
UNIVERSITY & RESEARCH

WP3 discussion

- ML model seems to capture trends linked to weather variables
- Could be useful for decision makers / farmers in forecasting when to act
- ***Outlook***:
    - Forecast 2014-2100 under SSP1-2.6 and SSP5-8.5 as inputs
- Uncertainty in climate scenarios
- Uncertainty in pest-climate interactions
- Process-based knowledge not yet included; e.g., models of spreading (partial differential equations)

WP4: Post-harvest

- Changed conditions affect shelf life of post-harvest products, resulting in losses
- ***Climate adaptation strategy***: optimize conditions for shelf life of products

- Experiment: tomatoes in a tray of 4 with different **light** intensity levels and different levels of **EC** (electrical conductance of fertilizer solution)
- Tomatoes are harvested pink (0) or red (1)
- Their quality is scored as OVQ (Overall Visual Quality)

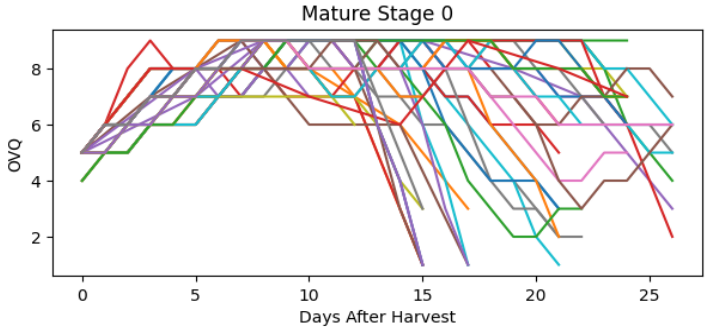
Modelling approach

- ***Goal***: predict the OVQ (Overall Visual Quality) for the post-harvest stage

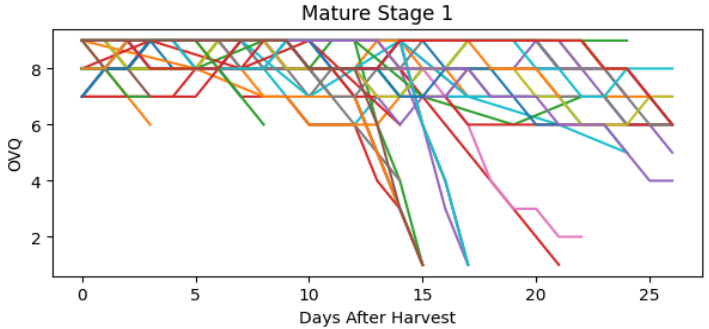
- Data points are removed after tomatoes in a tray drop below OVQ threshold
- Three models:
    - Baseline: Linear mixed model
    - Mixed Effects Random Forest model
    - PIML

Data

Block (repeat) = [10.3, 10.5]
Mature stage = [0=pink, 1=red]
Light = [100, 200, 300]
EC = [2, 7]
Pool (repeat within block) = [1, 2, 3, 4]
Days after harvest = [1..25]
OVQ = [1..9]



Pink harvested



Red harvested

WAGENINGEN
UNIVERSITY & RESEARCH

Mixed effects Random Forest (MERF) model

	ID	EC	light	mature_stage	days_after_harvest	block	ovq_at_zero	DMC	Soluble Carbohydrates g/100gDW	pool	ovq	vit_c
0	0	2.0	100.0	0.0	0	10.3	5.0	0.017817	35.484014	1.0	5.0	10.683660
1	0	2.0	100.0	0.0	1	10.3	5.0	0.023774	34.086132	1.0	6.0	11.904849
2	0	2.0	100.0	0.0	2	10.3	5.0	0.029731	32.688251	1.0	7.0	13.126038
3	0	2.0	100.0	0.0	3	10.3	5.0	0.035688	31.290369	1.0	8.0	14.347228
4	0	2.0	100.0	0.0	4	10.3	5.0	0.041646	29.892488	1.0	8.0	15.568417
...	...	...	...	...	...	...	...	...	...	...	...	...
1884	95	7.0	300.0	1.0	11	10.5	9.0	0.053505	36.187424	4.0	8.0	20.244179
1885	95	7.0	300.0	1.0	12	10.5	9.0	0.052636	35.562389	4.0	8.0	20.774807
1886	95	7.0	300.0	1.0	13	10.5	9.0	0.051768	34.937353	4.0	6.0	21.305434
1887	95	7.0	300.0	1.0	14	10.5	9.0	0.050899	34.312318	4.0	3.0	21.836062
1888	95	7.0	300.0	1.0	15	10.5	9.0	0.049920	33.695435	4.0	1.0	21.386564

1889 rows × 12 columns

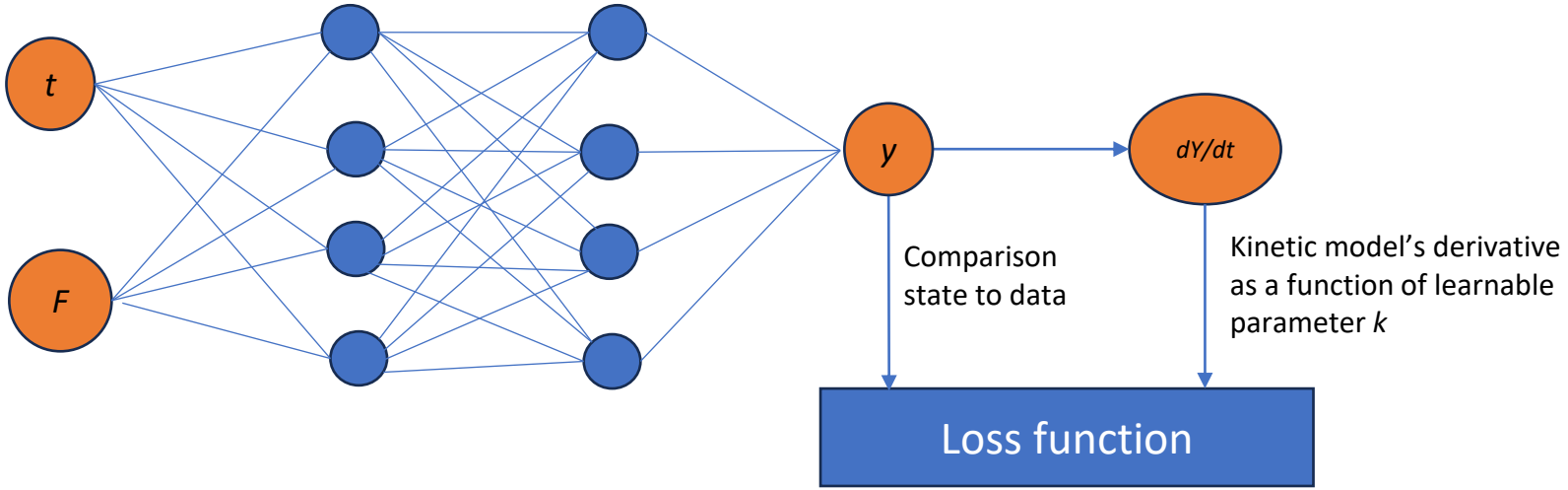
- **Fixed effects**
- **Cluster index**; random effect, regularization across clusters = trays
- **Target variable** **(overall visual quality)**

$$y = f(X) + b_i Z + e$$

- y is the target variable. Assume regression for now, e.g. continuously varying scalar value
- X is the fixed effect features. Assume p dimensional
- f(.) is the random forest
- Z is the random effect features. Assume q dimensional.
- e is iid noise ~N(0, sigma_e²)
- i is the cluster index. Assume k clusters in the training.
- bi is the random effect coefficients. They are different per cluster i but are assumed to be drawn from the same distribution ~N(0, Sigma_b) where Sigma_b is learned from the data.

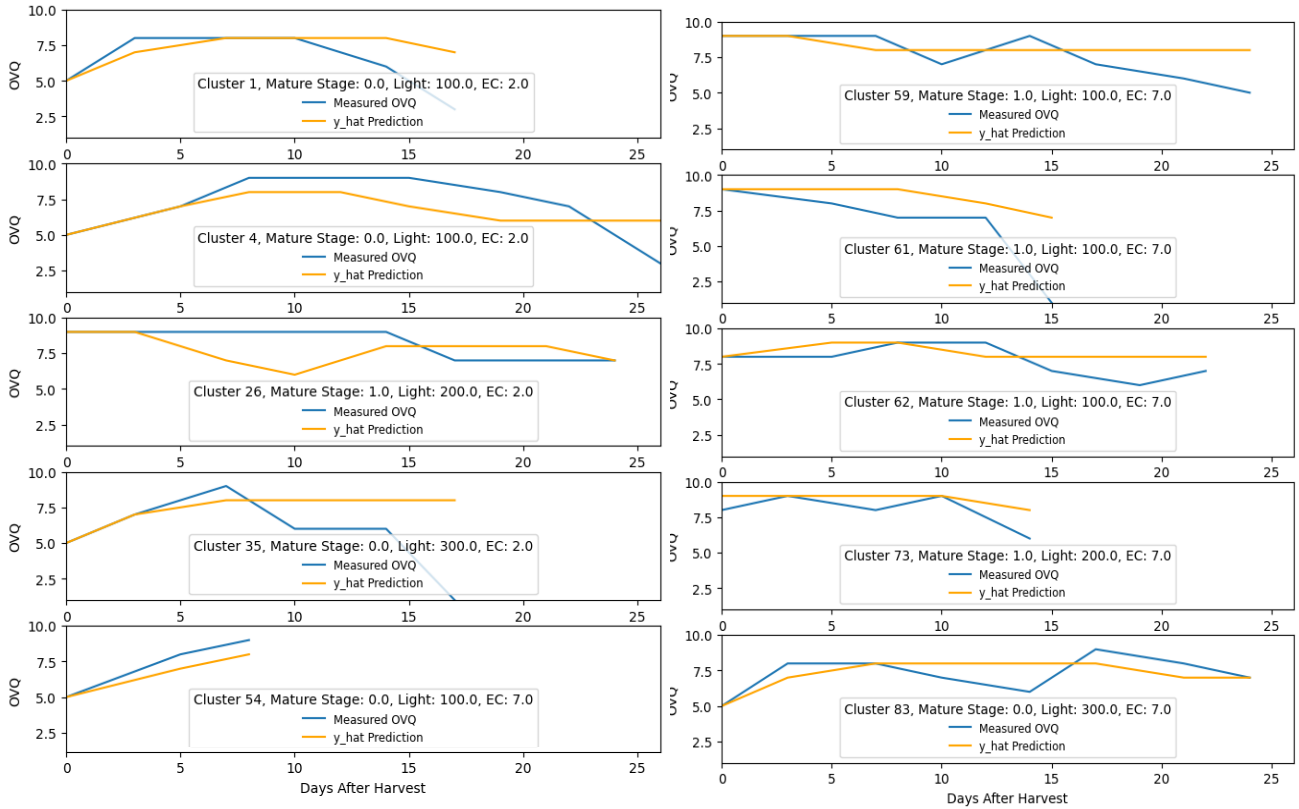
Physics Informed ML

$$C = C_0 \exp(-kt)$$



Results

- LME MAE on **validation set**: 1.16
- MERF MAE on **validation set**: 0.91
- MERF MAE on **test set**: 1.16
- PINNs on **test set**: 1.13



Results MERF on the 10 test tomato baskets

WP4 discussion

- MERF best performance, though average error is not fantastic
- Exponential decay is (mathematically) a linear process, statistically tailored set-up
- ***Outlook***:
    - Involving more data; experiments in which fruits are not removed to mimic a natural decaying process
    - Forecasting the next time points rather than the whole trajectory

Some overall discussion and conclusions

- Hybrid ML methods may be useful in assessing climate adaptation measures
- Data commonly not in (re)usable formats; work needed to make it machine readable
- Crop GxE models seem to improve with hybrid methods
- Pest management model currently lacks physical knowledge of insect-climate interactions; method may nevertheless be useful for decision makers
- Hybrid shelf-life models do not outperform statistical methods
- Of the three cases, GxE modelling is the most promising for exploring further methods

Note: AI accounts for $0.5 - 1\%$ of energy consumption… And rising

Thank you!

George.vanVoorn@wur.nl

Special thanks to Yingjie Shao

And thanks to all colleagues who participated