Assessment of hybrid models developed for the retrieval of vegetation traits from CHIME L2A data

Jochem Verrelst, José Luis García Soria, Juan Pablo Rivera-Caicedo, Katja Berger, Gabriele Candiani, Cinzia Panigada, Giulia Tagliabue & co-authors

Vegetation retrieval models

Hybrid models were developed for vegetation properties retrieval within CHIME E2E and parallel scientific studies. All models are based on **GPR**: **(1) competitive, and (2) provision of associated uncertainty estimates**

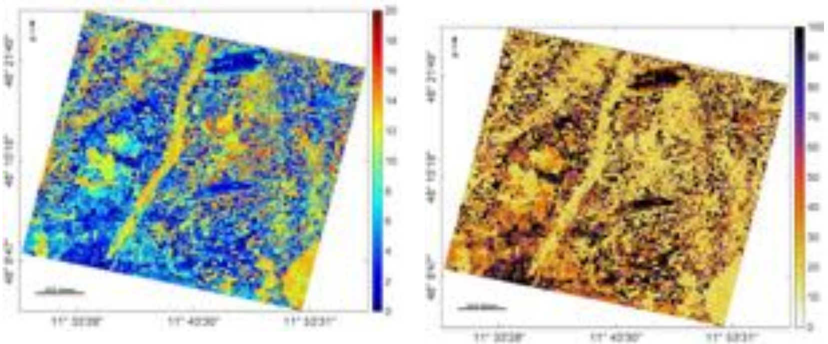
Need for a critical review of:

1. All developed hybrid models (>10 variables)
2. Evaluation of alternative machine learning regression algorithms (MLRAs) **given provision of uncertainty estimates**
3. Models' robustness under noise scenarios

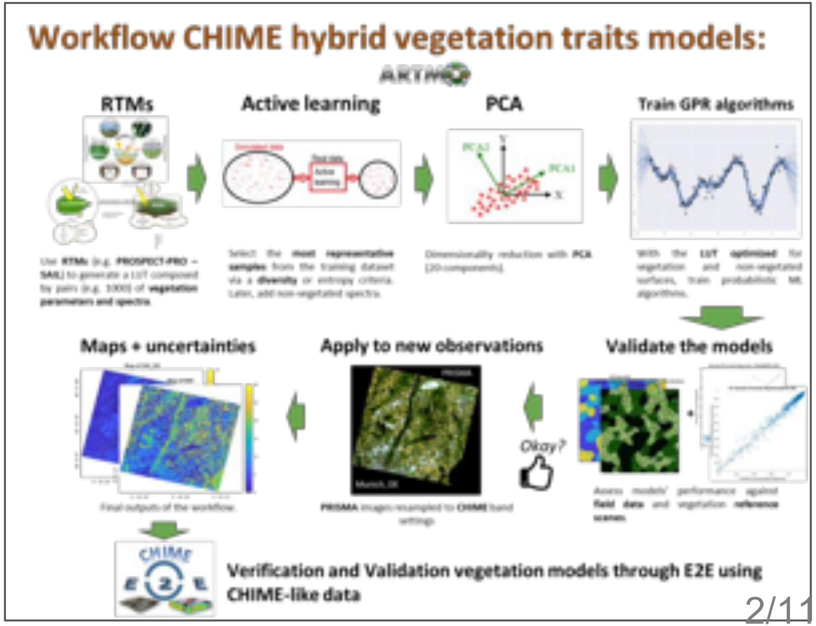
We evaluated hybrid models as developed by:

- E2E studies: v.1.9 models
- CNR-IREA: GRO18 models
- Milano-Bicocca-UNIMIB: JDS20 models

Canopy Nitrogen Content (CNC) | **Uncertainty**

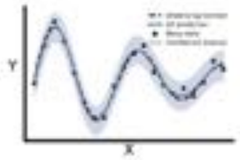


Verrelst, et al., 2021. Mapping landscape canopy nitrogen content from space using PRISMA data. ISPRS

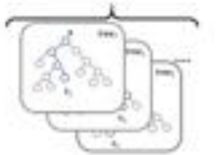


Powerful machine learning regression algorithms for hybrid models

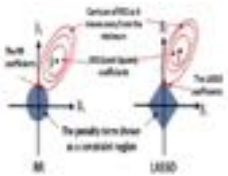
- **Gaussian process regression (GPR).** GPR is based on Gaussian processes (GPs), which generalize Gaussian probability distributions in a function's space. GPs provide a natural way of assessing the **uncertainty of the predictions** through the **predictive variance** (error bars).



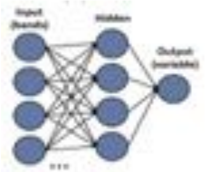
- **Quantile Regression Forests (QRF).** Random Forests is a specific type of bagging trees that constructs a collection of decision trees with controlled variance. In QRF, the distribution of responses is taken, allowing **the estimation of the full prediction interval, i.e. uncertainties.**

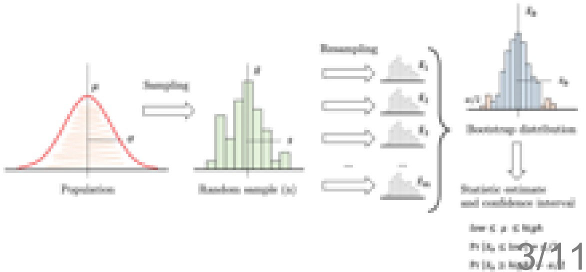


- **Kernel ridge regression (KRR).** **KRR combines RR with the kernel trick.** It thus learns a linear function in the space induced by the respective kernel and the data. **Uncertainty can be calculated through bootstrapping!**

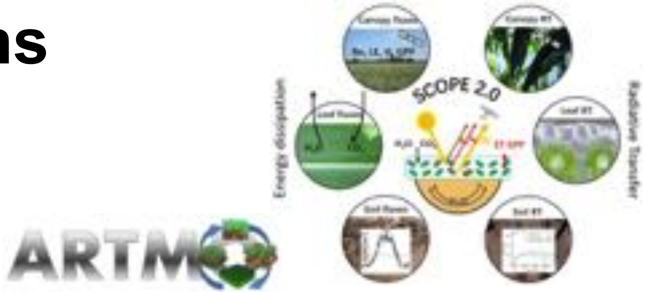


- **Artificial neural networks (ANN)... is basically a pointwise nonlinear function (e.g., a sigmoid or Gaussian function) applied to the output of a linear regression.** The most common structure is a feed-forward ANN. **Uncertainty can be calculated through bootstrapping!**





Experimental Setup: SCOPE simulations

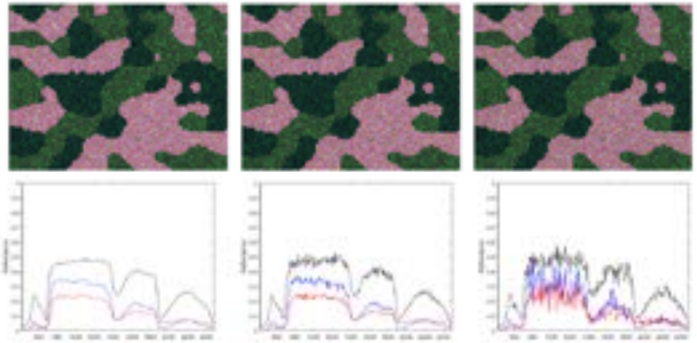


Simulated scene:

- 300 x 300 pixels; 295 CHIME-like bands
- 3 classes with random sampling: low, medium, high
- broad ranges: evaluation if generic
- Input images of vegetation variables
- Gaussian noise levels added: 5%, 10%, 20%, 50%

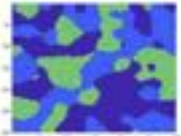
Input Variable	Acronym	Small Min	Small Max	Medium Min	Medium Max	High Min	High Max
Leaf Variables: PROSPECT-PRO							
Leaf Chlorophyll Content [μg/cm^2]	Cab / LCC	5	20	20	50	50	80
Leaf Water Content [g/cm^2]	Cw / LWC	0.003	0.005	0.005	0.015	0.015	0.03
Leaf Dry Matter Content [g/cm^2]	Cdm /LDMC	0.0005	0.002	0.002	0.006	0.006	0.01
Leaf Protein Content [g/cm^2]	Cp	0.0001	0.001	0.001	0.0015	0.0015	0.002
Leaf mesophyll structure parameter	N	1	2.7	1	2.7	1	2.7
Leaf Carbon Based Constituents [g/cm^2]	CBC	0.0004	0.001	0.001	0.0045	0.0045	0.0080
Leaf Carotenoids Content [μg/cm^2]	Cxc	0	20	0	20	0	20
Canopy Variables: 4SAIL							
Leaf Area Index [m^2/m^2]	LAI	0.1	1.5	1.5	4.0	4.0	8
Leaf Inclination determination a [-]	LIDFa	-1	1	-1	1	-1	1
Leaf Inclination determination b [-]	LIDFb	-1	1	-1	1	-1	1
Soil Variables: BSM							
Soil Moisture Content [%]	SMC	5	55	5	55	5	55
Model parameter for Soil Brightness [-]	B	0	0.9	0	0.9	0	0.9
Model parameter 'lat' [-]	BSMlat	20	40	20	40	20	40
Model parameter 'lon' [-]	BSMlon	45	65	45	65	45	65
Illumination and Observation Angles							
Relative Azimuth Angle [deg]	RAA	0	180	0	180	0	180
Observer Zenith Angle [deg]	OZA	0	25	0	25	0	25
Sun Zenith Angle [deg]	SZA	0	80	0	80	0	80

RGB output scene with noise added:

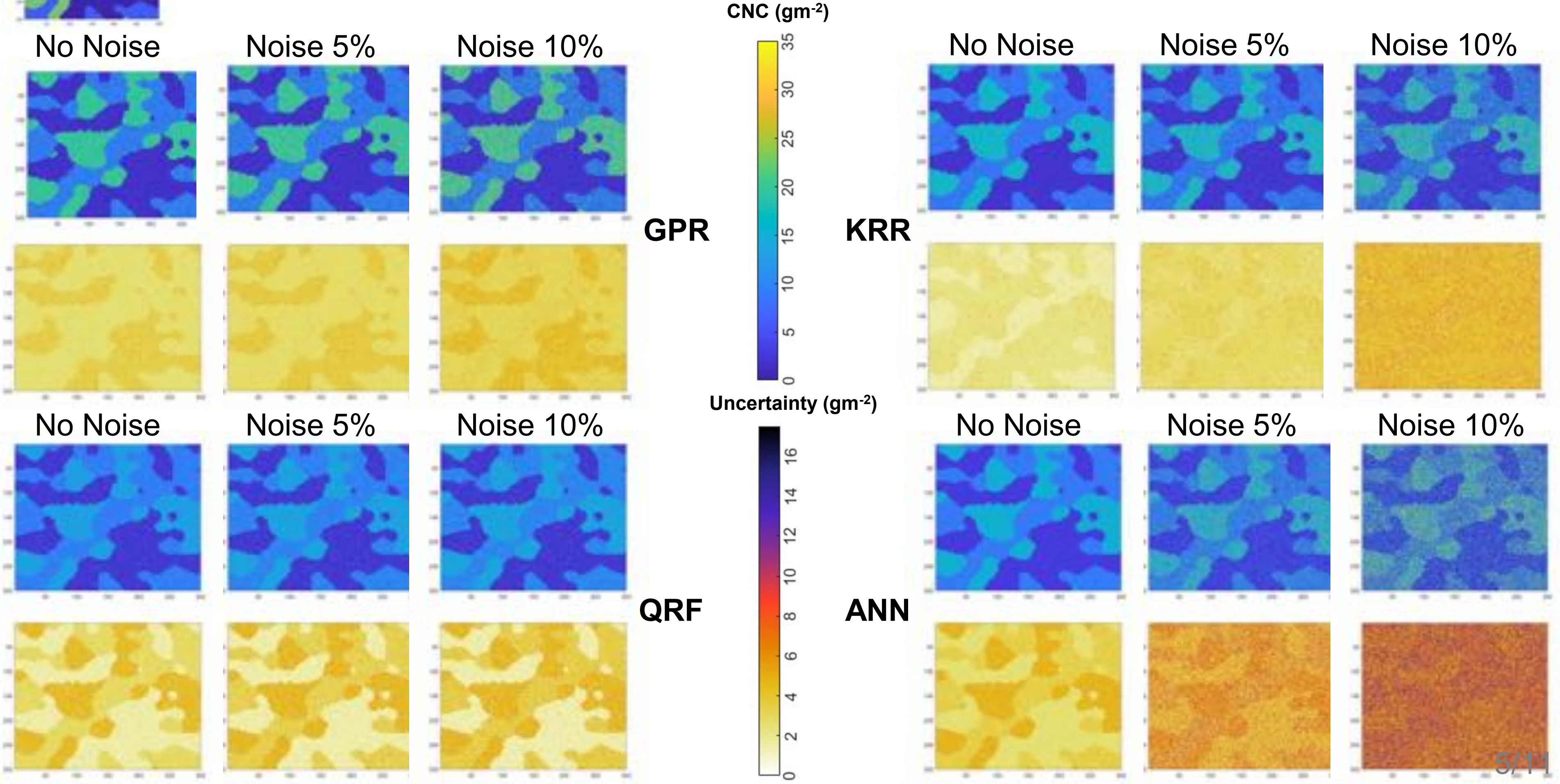


E2E Priority variables:

- specific leaf area (SLA)
- leaf nitrogen content (LNC)
- canopy nitrogen content (CNC)

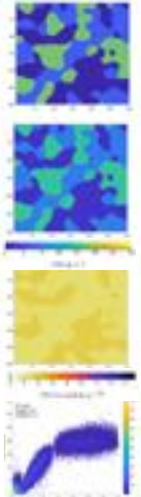


Canopy Nitrogen Content (CNC): E2E v.1.9



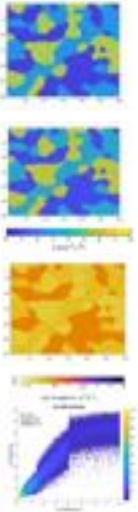
Canopy nitrogen content [g/m²]: CNC (E2E v.1.9)

Method	Statistic	Noise level 0%	5%	10%	20%	50%
GPR	R^2	**0.86**	**0.81**	**0.69**	0.45	0.16
	$RMSE$	3.92	4.32	5.38	7.98	15.90
	$NRMSE$	0.11	0.12	0.15	0.23	0.45
KRR	R^2	0.84	0.71	0.49	0.22	0.05
	$RMSE$	5.01	5.64	7.01	10.10	19.75
	$NRMSE$	0.14	0.16	0.19	0.28	0.55
QRF	R^2	0.66	0.62	0.58	**0.53**	**0.44**
	$RMSE$	6.67	6.61	6.69	6.96	7.38
	$NRMSE$	0.19	0.18	0.19	0.19	0.21
NN	R^2	0.84	0.42	0.18	0.06	0.01
	$RMSE$	5.53	7.37	10.06	13.11	15.49
	$NRMSE$	0.15	0.20	0.28	0.36	0.43



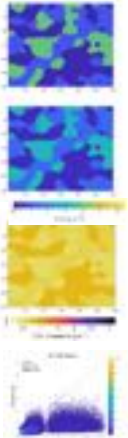
Leaf area index [m²/m²]: LAI (E2E v.1.9)

Method	Statistic	Noise level 0%	5%	10%	20%	50%
GPR	R^2	**0.87**	**0.83**	**0.75**	0.64	0.43
	$RMSE$	0.85	0.95	1.13	1.43	1.95
	$NRMSE$	0.11	0.12	0.14	0.18	0.24
KRR	R^2	0.80	0.79	0.74	0.59	0.24
	$RMSE$	1.01	1.06	1.20	1.63	3.11
	$NRMSE$	0.13	0.13	0.15	0.20	0.39
QRF	R^2	0.77	0.74	0.71	**0.67**	**0.60**
	$RMSE$	1.07	1.14	1.20	1.29	1.44
	$NRMSE$	0.13	0.14	0.15	0.16	0.18
NN	R^2	0.83	0.75	0.59	0.33	0.08
	$RMSE$	0.96	1.18	1.59	2.31	3.34
	$NRMSE$	0.12	0.15	0.20	0.29	0.42



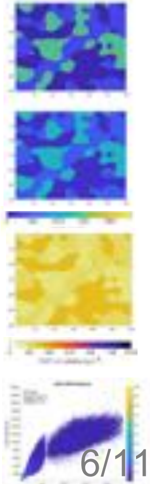
Canopy chlorophyll content [g/m²]: CCC (E2E v.1.9)

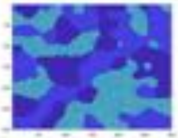
Method	Statistic	Noise level 0%	5%	10%	20%	50%
GPR	R^2	**0.87**	**0.83**	**0.76**	0.63	0.38
	$RMSE$	0.69	0.80	0.99	1.29	1.72
	$NRMSE$	0.11	0.13	0.15	0.20	0.27
KRR	R^2	0.82	0.80	0.73	0.56	0.21
	$RMSE$	0.80	0.83	0.89	1.09	1.83
	$NRMSE$	0.13	0.13	0.14	0.17	0.29
QRF	R^2	0.74	0.73	0.72	**0.68**	**0.56**
	$RMSE$	0.93	0.95	0.98	1.03	1.15
	$NRMSE$	0.15	0.15	0.15	0.16	0.18
NN	R^2	0.85	0.77	0.61	0.34	0.08
	$RMSE$	0.64	0.80	1.12	1.76	2.83
	$NRMSE$	0.10	0.12	0.17	0.28	0.44



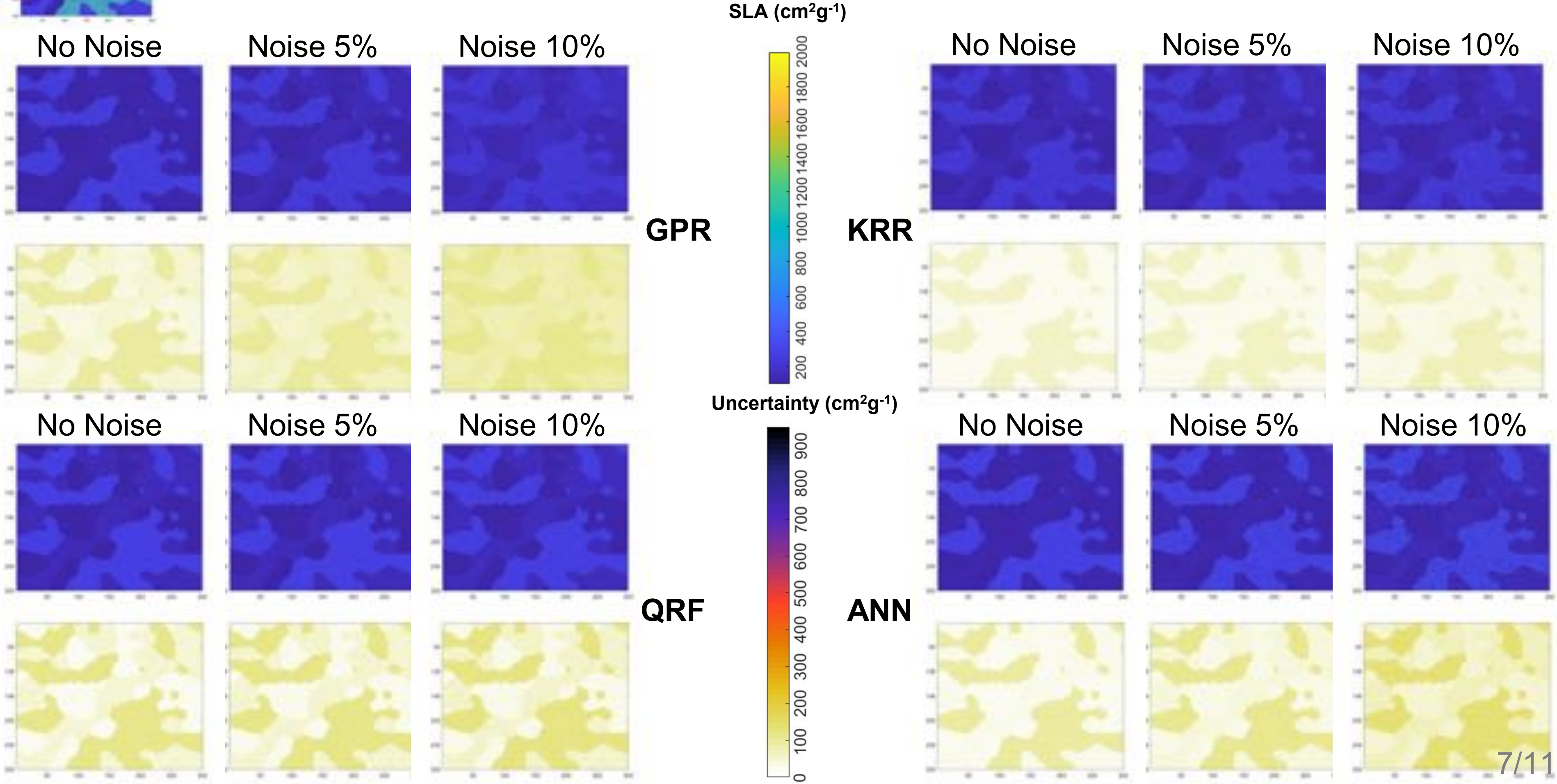
Canopy water content [g/m²]: CWC (E2E v.1.9)

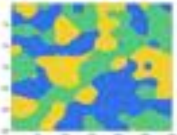
Method	Statistic	Noise level 0%	5%	10%	20%	50%
GPR	R^2	**0.89**	0.79	0.66	0.50	0.26
	$RMSE$	229.31	307.98	391.81	477.00	553.43
	$NRMSE$	0.10	0.13	0.16	0.20	0.23
KRR	R^2	0.87	**0.85**	**0.80**	0.65	0.27
	$RMSE$	236.69	246.12	272.43	351.25	636.84
	$NRMSE$	0.10	0.10	0.11	0.15	0.27
QRF	R^2	0.82	0.80	0.79	**0.77**	**0.72**
	$RMSE$	285.99	291.56	297.39	306.58	329.70
	$NRMSE$	0.12	0.12	0.12	0.13	0.14
NN	R^2	0.87	0.83	0.75	0.54	0.18
	$RMSE$	239.31	257.60	300.28	402.36	620.41
	$NRMSE$	0.10	0.11	0.13	0.17	0.26



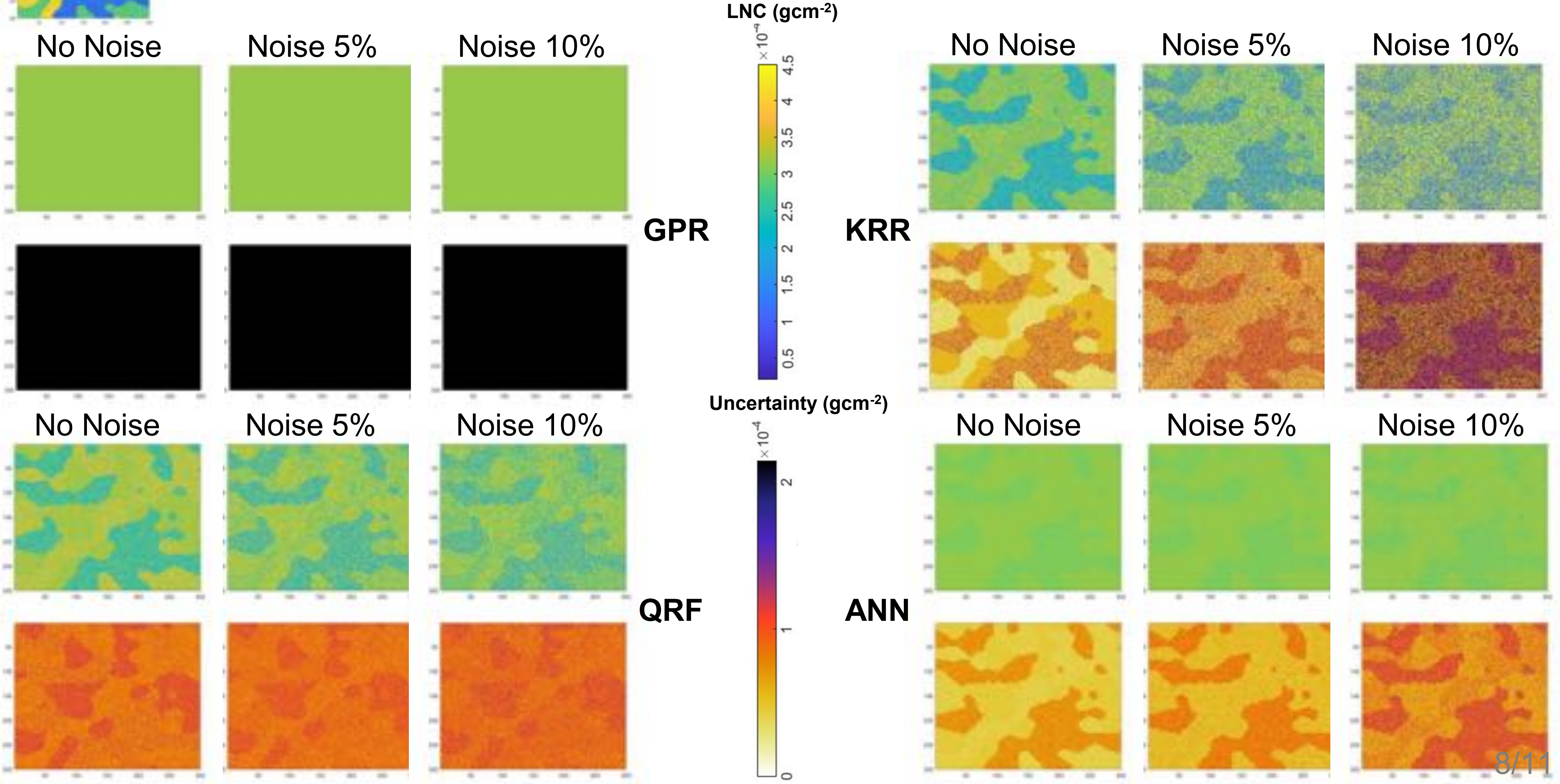


Specific Leaf Area (SLA) (E2E v.1.9)





Leaf Nitrogen Content (LNC) E2E v.1.9



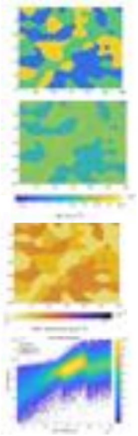
Leaf level variables

Specific leaf area (SLA [cm²/g]): E2E v.1.9

Method	Statistic	Noise level 0%	5%	10%	20%	50%
GPR	R^2	0.60	0.48	0.27	0.06	0.00
	$RMSE$	477.18	481.45	487.93	499.83	511.29
	$NRMSE$	0.24	0.24	0.24	0.25	0.26
KRR	R^2	0.47	0.43	0.34	0.19	0.07
	$RMSE$	490.59	491.23	492.79	497.28	501.82
	$NRMSE$	0.25	0.25	0.25	0.25	0.25
QRF	R^2	0.63	**0.61**	**0.57**	**0.42**	**0.05**
	$RMSE$	466.97	465.46	468.08	476.49	503.56
	$NRMSE$	0.23	0.23	0.23	0.24	0.25
NN	R^2	**0.70**	0.56	0.38	0.18	0.04
	$RMSE$	464.01	463.24	463.90	471.67	490.47
	$NRMSE$	0.23	0.23	0.23	0.24	0.25

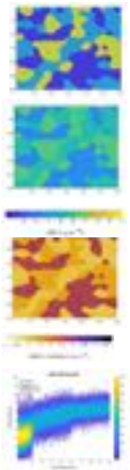
Leaf nitrogen content (LNC [g/cm²]): E2E v.1.9

Method	Statistic	Noise level 0%	5%	10%	20%	50%
GPR	R^2	**0.44**	**0.28**	**0.20**	**0.13**	**0.05**
	$RMSE$	0.00	0.00	0.00	0.00	0.00
	$NRMSE$	0.29	0.29	0.29	0.29	0.29
KRR	R^2	0.42	0.07	0.02	0.01	0.00
	$RMSE$	0.00	0.00	0.00	0.00	0.00
	$NRMSE$	0.21	0.35	0.51	0.85	1.91
QRF	R^2	0.32	0.14	0.09	0.04	0.00
	$RMSE$	0.00	0.00	0.00	0.00	0.00
	$NRMSE$	0.24	0.27	0.28	0.29	0.31
NN	R^2	0.29	0.09	0.04	0.02	0.01
	$RMSE$	0.00	0.00	0.00	0.00	0.00
	$NRMSE$	0.28	0.28	0.28	0.29	0.30



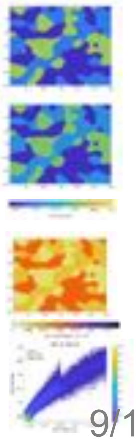
Leaf mass per area (LMA [g/cm²]): JDS20

Method	Statistic	Noise level 0%	5%	10%	20%	50%
GPR	R^2	0.36	0.12	0.05	0.01	0.00
	$RMSE$	0.00	0.00	0.00	0.00	0.00
	$NRMSE$	0.30	0.30	0.30	0.30	0.30
KRR	R^2	**0.71**	**0.53**	0.26	0.05	0.00
	$RMSE$	0.00	0.00	0.00	0.00	0.00
	$NRMSE$	0.23	0.25	0.28	0.38	0.80
QRF	R^2	0.51	0.43	**0.29**	**0.11**	0.00
	$RMSE$	0.00	0.00	0.00	0.00	0.00
	$NRMSE$	0.25	0.25	0.26	0.27	0.30
NN	R^2	0.10	0.03	0.01	0.00	0.00
	$RMSE$	0.00	0.00	0.00	0.00	0.00
	$NRMSE$	0.31	0.31	0.31	0.32	0.32



Leaf water content (LWC [cm]): E2E v.1.9

Method	Statistic	Noise level 0%	5%	10%	20%	50%
GPR	R^2	**0.91**	**0.87**	0.80	0.66	0.37
	$RMSE$	0.00	0.00	0.00	0.00	0.01
	$NRMSE$	0.08	0.10	0.12	0.16	0.23
KRR	R^2	0.86	0.84	0.80	0.66	0.30
	$RMSE$	0.00	0.00	0.00	0.00	0.01
	$NRMSE$	0.13	0.13	0.14	0.17	0.27
QRF	R^2	0.82	0.82	**0.81**	**0.77**	**0.68**
	$RMSE$	0.00	0.00	0.00	0.00	0.00
	$NRMSE$	0.12	0.12	0.12	0.13	0.15
NN	R^2	0.79	0.70	0.54	0.30	0.07
	$RMSE$	0.00	0.00	0.01	0.01	0.01
	$NRMSE$	0.14	0.16	0.19	0.26	0.36



Overview of the model results

R^2 (GPR: bolded if best result of all MLRAs)

	Noise level				
Variable	0%	5%	10%	20%	50%
E2E v.1.9:					
SLA	0.55	0.48	0.27	0.05	0.00
LNC	**0.44**	**0.28**	**0.20**	**0.13**	**0.05**
LAI	**0.87**	**0.83**	**0.75**	0.64	0.43
CNC	**0.86**	**0.81**	**0.69**	0.45	0.16
LCC	**0.91**	**0.87**	**0.81**	**0.68**	0.4
LWC	**0.91**	**0.87**	0.8	0.66	0.37
CCC	**0.87**	**0.83**	0.76	0.63	0.38
CWC	**0.89**	0.79	0.66	0.50	0.26
FVC	**0.90**	**0.90**	**0.88**	0.80	0.50
FAPAR	0.96	0.95	0.92	0.85	0.61
ESA-IT-GRO18:					
LNC	0.16	0.14	0.11	0.06	0.01
CNC	0.05	0.02	0	0	0
LCC	0.69	0.65	0.58	0.41	0.16
CCC	0.79	0.62	0.42	0.21	0.05
ESA-IT-JDS20:					
LMA	0.15	0.15	0.15	0.13	0.08
LNC	0.36	0.12	0.05	0.01	0.00
LAI	**0.64**	0.32	0.12	0.03	0.00
CNC	**0.66**	0.29	0.13	0.05	0.01
LCC	0.56	0.49	0.41	0.30	0.15
CCC	**0.57**	**0.56**	**0.53**	0.43	0.16
LWC	0.88	0.84	0.72	0.48	0.16
CWC	0.80	0.48	0.34	0.22	0.09

NRMSE (%) for 0% noise (best result bolded)

Variable	GPR	KRR	QRF	NN
E2E v.1.9:				
SLA	24%	25%	**23%**	**23%**
LNC	29%	**21%**	24%	28%
LAI	**11%**	13%	13%	12%
CNC	**11%**	14%	19%	15%
LCC	**10%**	18%	16%	14%
LWC	**8%**	13%	12%	14%
CCC	11%	13%	15%	**10%**
CWC	**10%**	**10%**	12%	**10%**
FVC	12%	**10%**	18%	14%
FAPAR	8%	9%	13%	**5%**
ESA-IT-GRO18 models:				
LNC	35%	**30%**	38%	40%
CNC	29%	**28%**	31%	27%
LCC	19%	13%	25%	17%
CCC	19%	24%	21%	**19%**
ESA-IT-JDS20 models:				
LMA	27%	**21%**	23%	25%
LNC	30%	**23%**	25%	31%
LAI	**17%**	19%	25%	30%
CNC	**21%**	**21%**	24%	26%
LCC	21%	30%	22%	**15%**
CCC	**21%**	23%	22%	25%
LWC	17%	14%	**12%**	15%
CWC	16%	**9%**	11%	18%
# Total best models	8	9	2	6
# Total best models <16% NRMSE	5	3	1	4
# Total models <16% NRMSE	8	10	7	9

Conclusions & recommendations

- GPR was evaluated as the best performing, but when fails, **KRR** is most promising alternative (in case of low noise levels). **QRF** appears as first choice as backup model (most robust in presence of noise). Note: *The E2E-L2BV module needs to be adapted for running these models.*

- **Priority variables:** Promising results for canopy nitrogen content (CNC). Also for other canopy variables (E2E v.1.9).

- Retrieval of the **priority leaf variables needs improvements** through tuning at training stages *(gains can be expected when focusing the training data to specific spectral domains, i.e., where variables are most sensitive)* and providing an **optimized training database** in terms of variable sampling (distribution and ranges).

- The available data sets (E2E v.1.9, GRO18, JDS20) were used separately to train different algorithms for each variable. **Merging the data sets into one pool** could lead to more robust models.

- A next step is to assess the robustness over varying land covers and differing sun-target-geometries: **multiple PRISMA images.**

- The hybrid workflows could be easily adapted towards highly demanded variables, such as carbon content, fresh biomass, or non-photosynthetic (dry) biomass, potentially providing first generic models of these traits within the E2E chain.