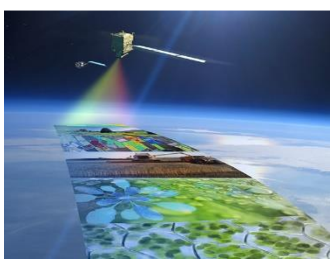
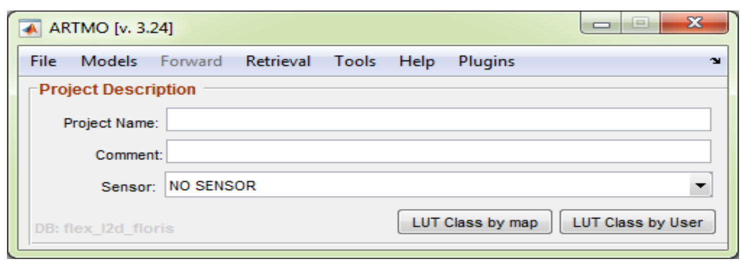
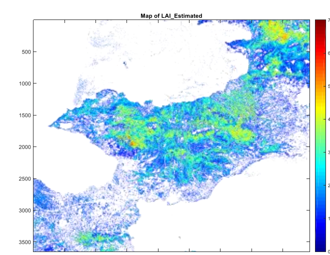
Automated vegetation properties mapping:

Application of S2 vegetation loss quantification







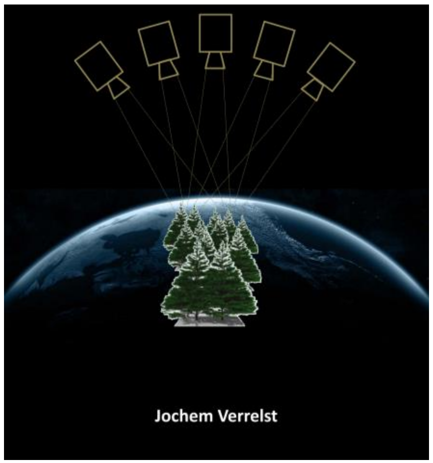
Jochem Verrelst

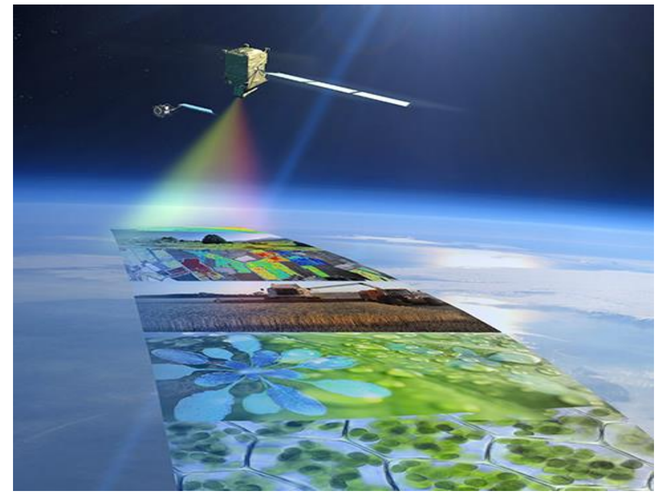
PECS, Bratislava, 20/09/2018

Me:

Wageningen (Nl)
(2005-2010)

Valencia (Es)
(2010-now)





Hyperspectral vegetation properties mapping

FLEX:
Sun-induced fluorescence mapping

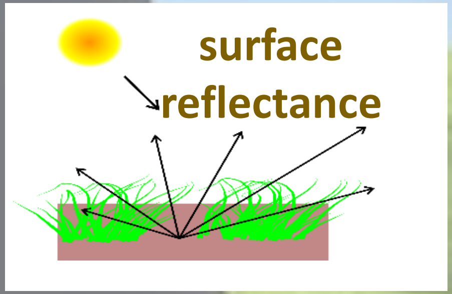
Quantification of vegetation properties from optical data

How to quantify vegetation productivity?

Sentinel 3

Operational

- Global coverage
- ☐ 300 m, 🕘

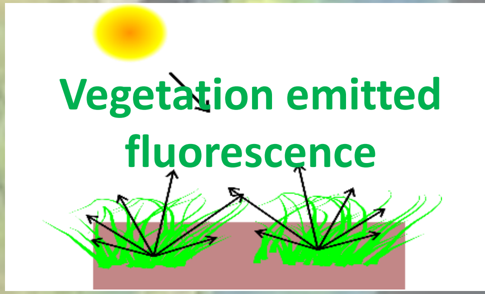


Greenness indicators
Stress?
Potential photosynthesis

8th Earth Explorer: FLEX

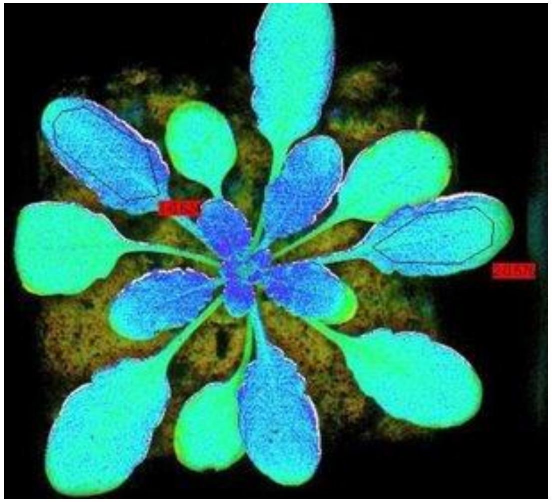
Launch: 2022

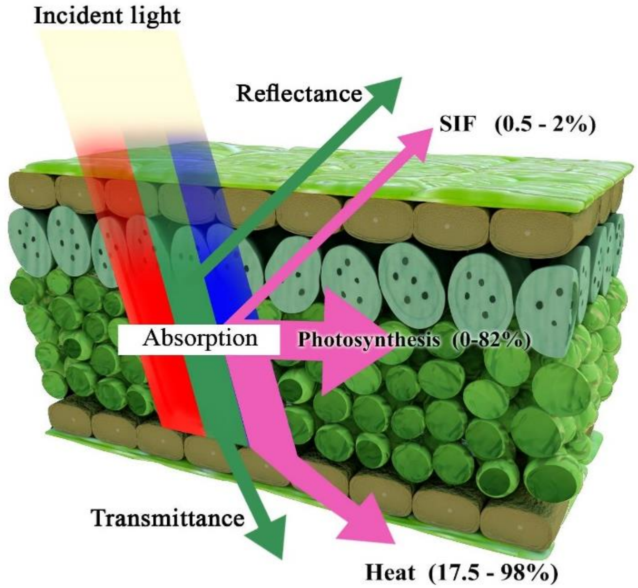
- Global coverage
- ☐ 300 m, 🕘

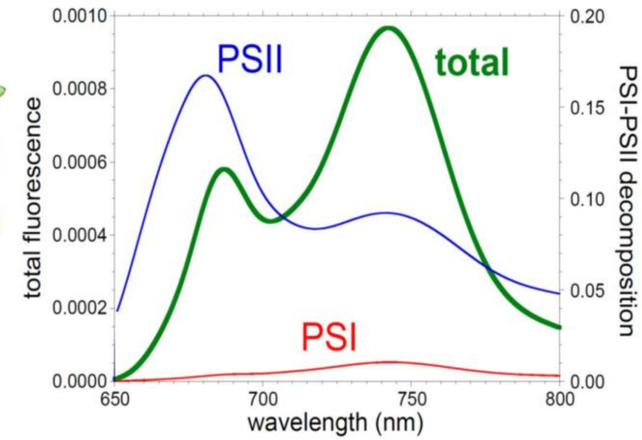


A signal coming directly from the plant
Actual photosynthesis

What is sun-induced chlorophyll fluorescence (SIF)?

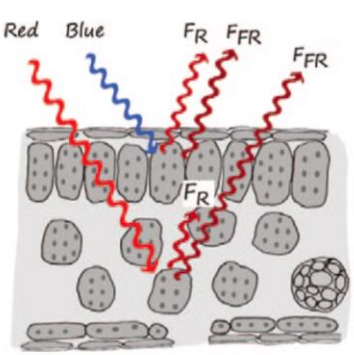




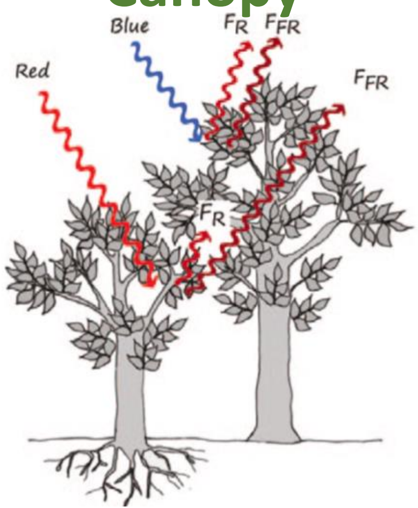


(re-)absorption and scattering mechanisms

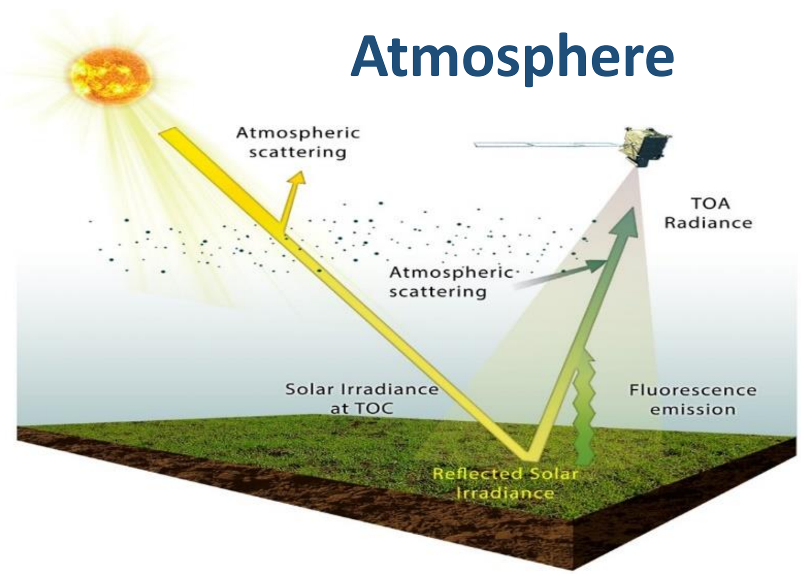
Leaf

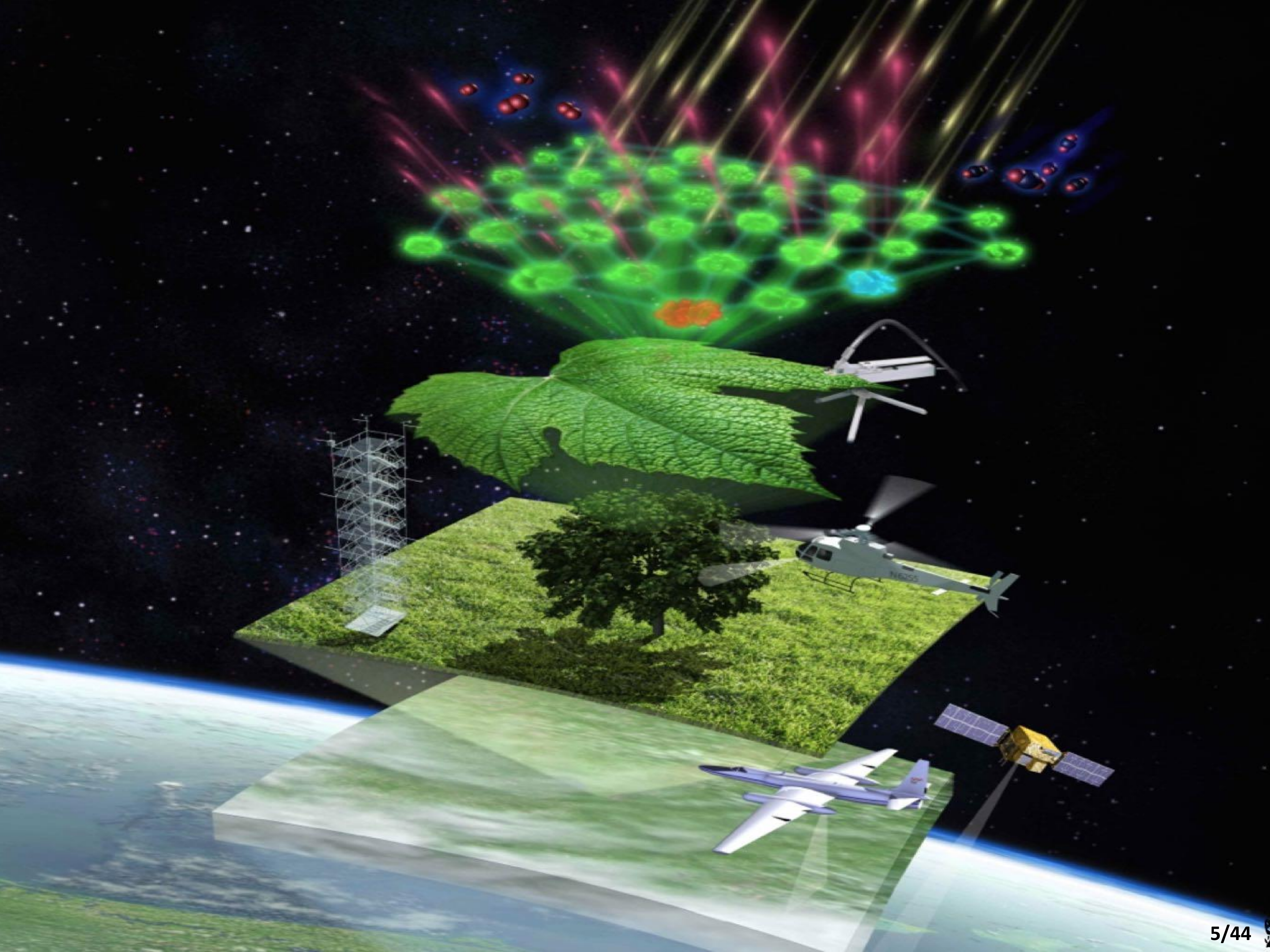


Canopy



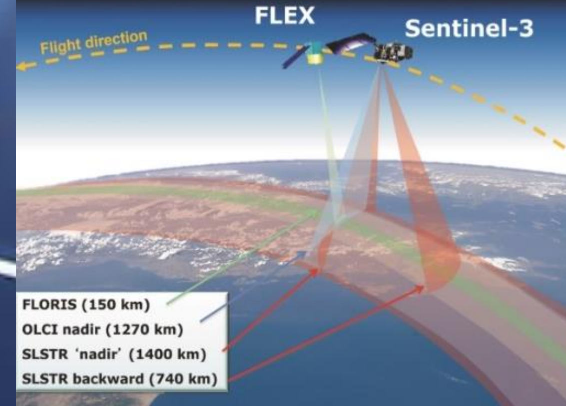
Atmosphere





flex

FLEX aims to quantify **actual photosynthetic activity** of terrestrial ecosystems from space, accounting for **vegetation health** status and **stress** conditions.

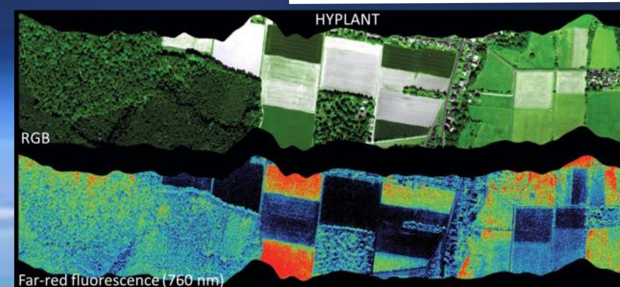


Space

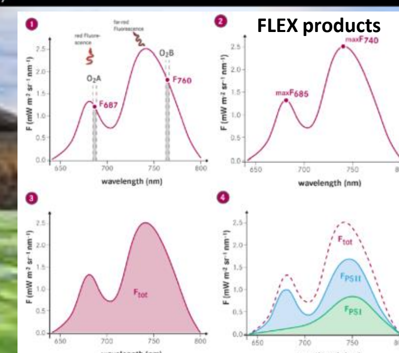
?

State of the art

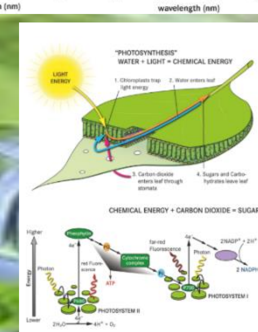
Airborne



Canopy



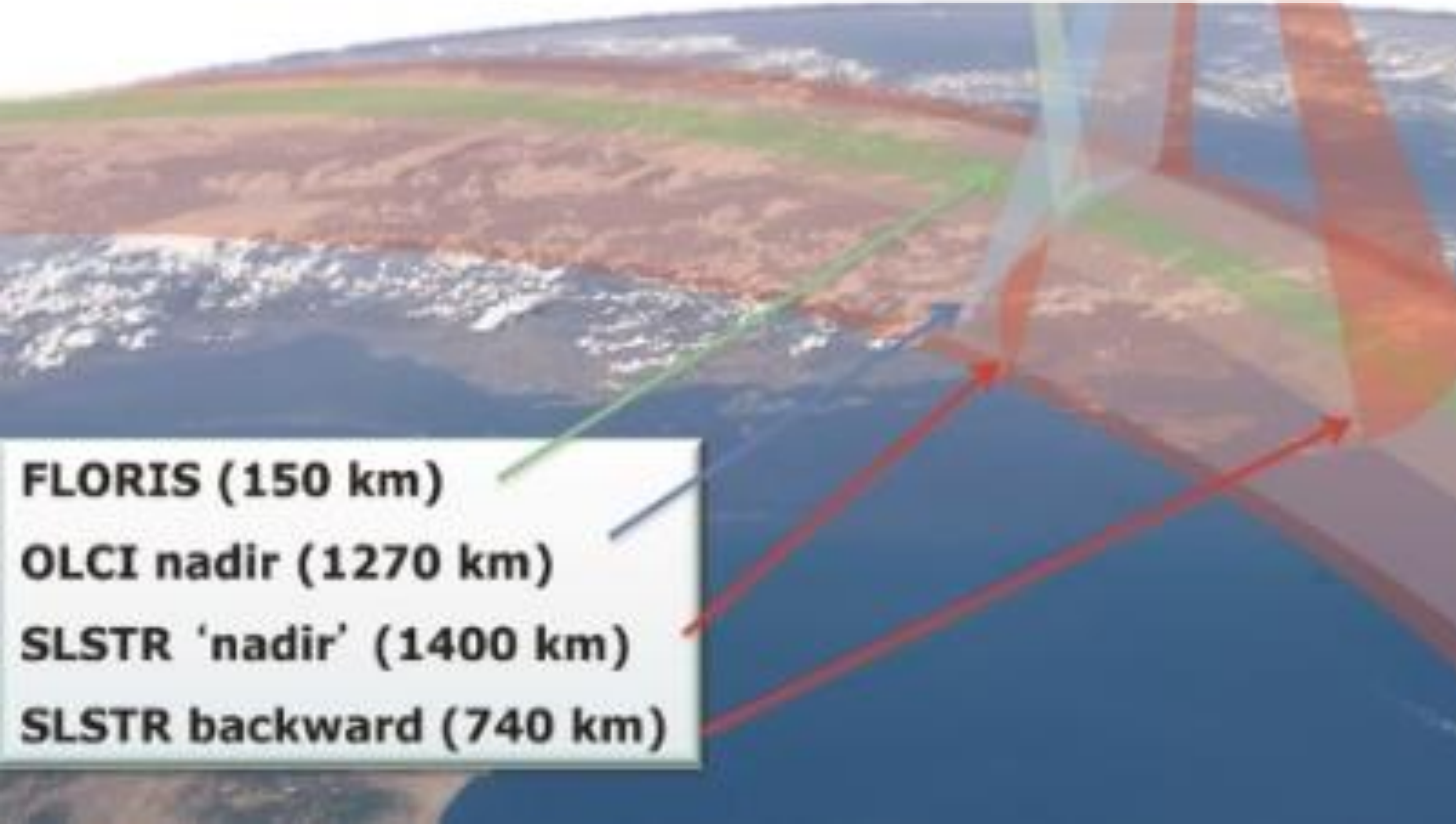
Leaf



Cell

Tandem mission concept driven by synergy:

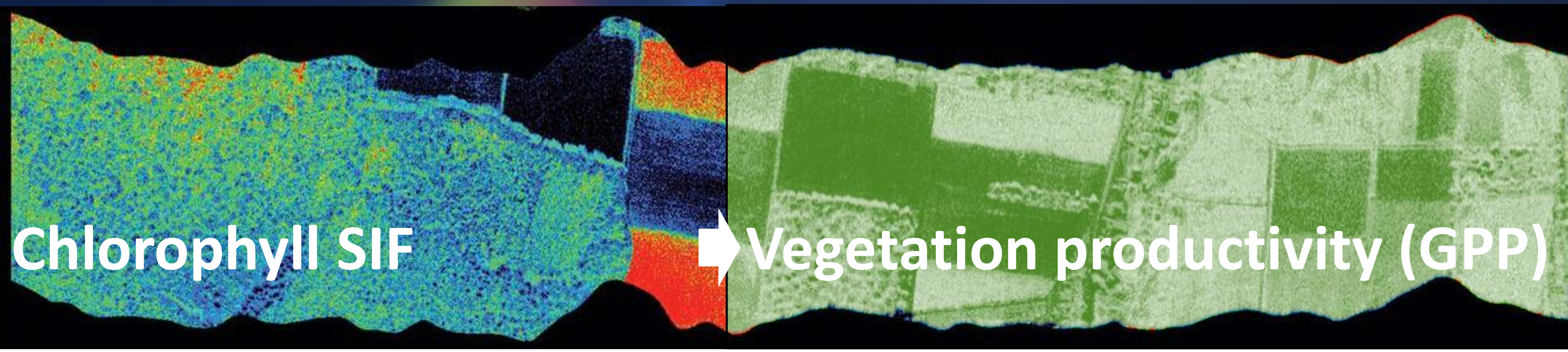
- S3 OLCI & SLSTR used for FLEX atmospheric correction
- Synergy of S3 OLCI and FLEX-FLORIS for improved biophysical parameter retrieval.
- S3 & FLEX products used as inputs in photosynthesis model (CO2 assimilation)

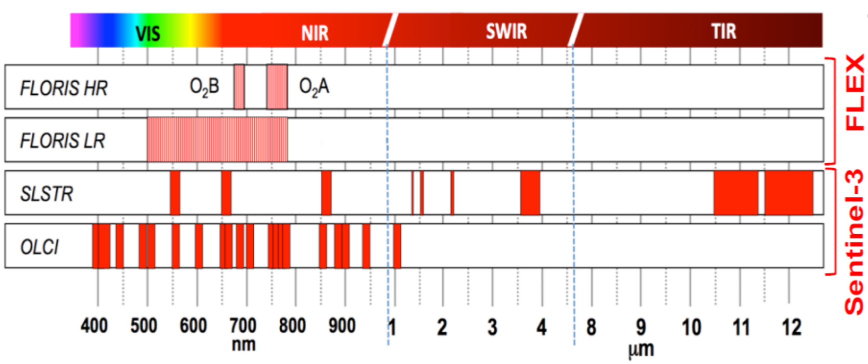


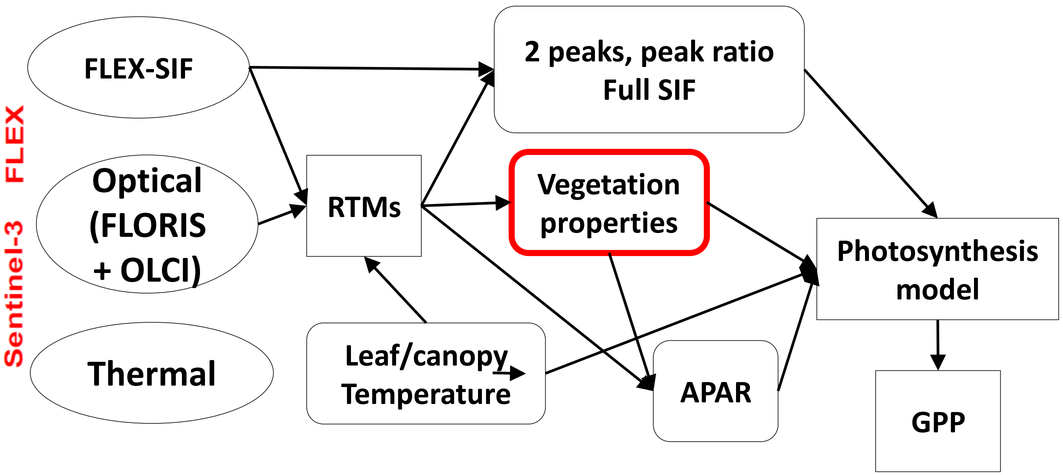
Mission characteristics:

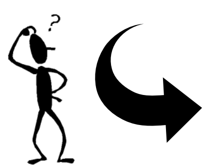
Tandem with Sentinel-3	
Sun-synchronous orbit	
Swath (km)	150
Spatial resolution (km)	0.3
Revisit time (days)	<27
Equatorial crossing	10:00

Synergistic approach to quantify photosynthesis





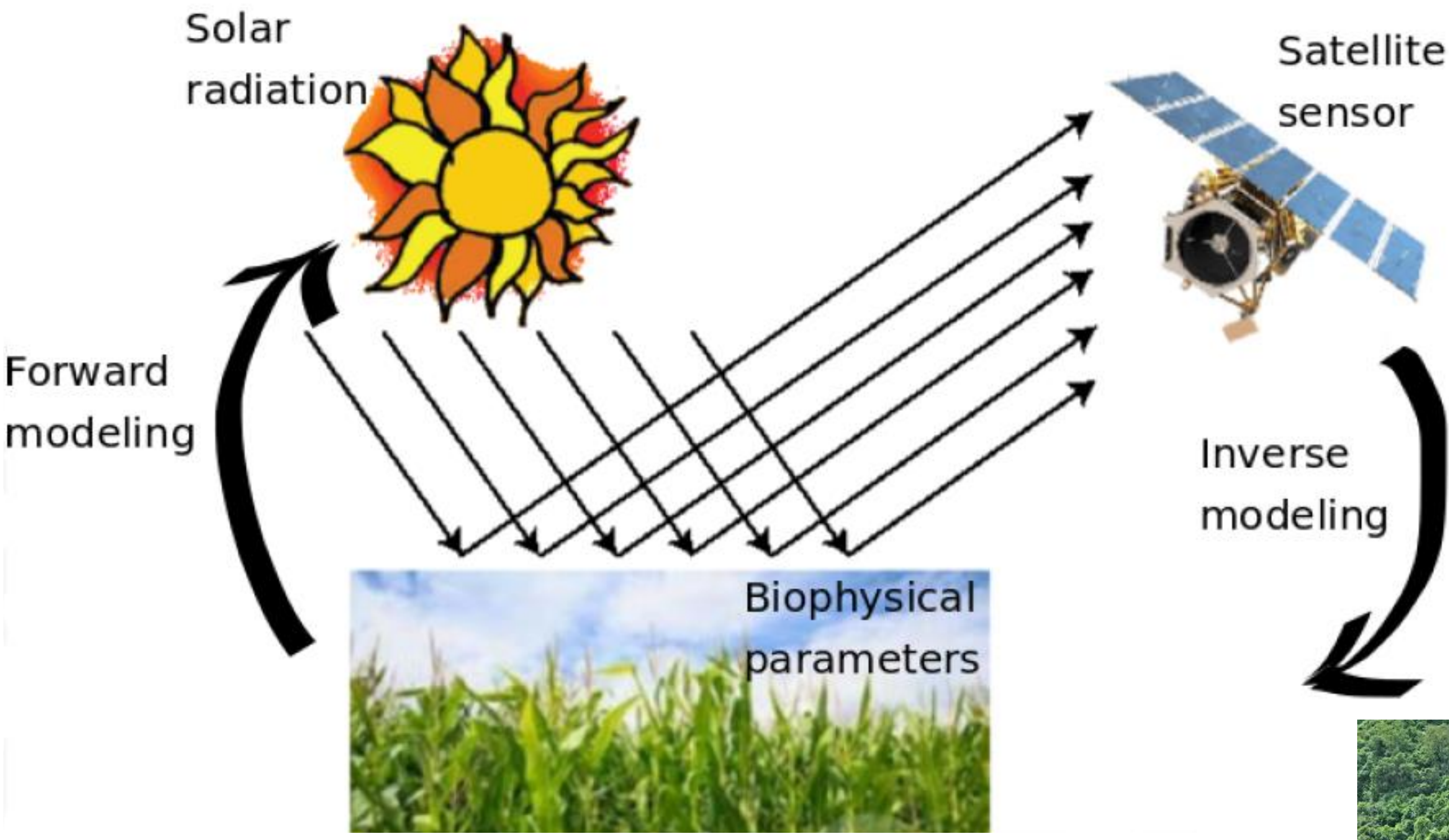




Assimilation approach.

How to quantify vegetation properties?

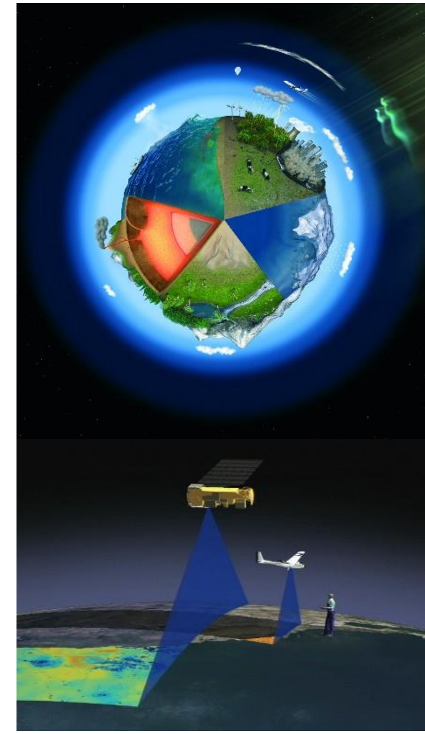
Today we will learn:
Semi-automated mapping of vegetation properties from optical RS data



- *What are biophysical parameters?*
- *Why is it important to quantify them?*

The problem:

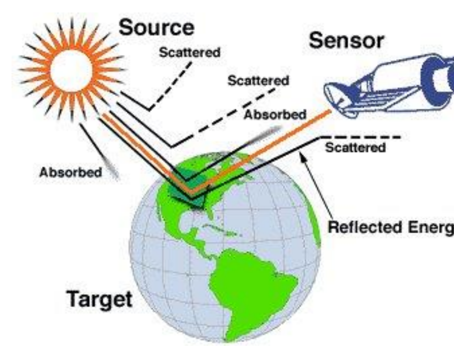
Biophysical parameter retrieval is an essential step in modeling the processes occurring on Earth and the interactions with the atmosphere.



The analysis can be done at **local** or **global** scales by looking at bio-geo-chemical cycles, atmospheric situations, ocean/river/ice states, and vegetation dynamics.

Main parameters: crop yield, biomass, leaf area coverage, chlorophyll content, fraction vegetation cover, GPP,….

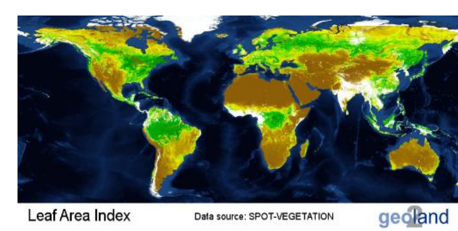
Land/vegetation parameters cannot be estimated directly from optical RS data. A model is required!



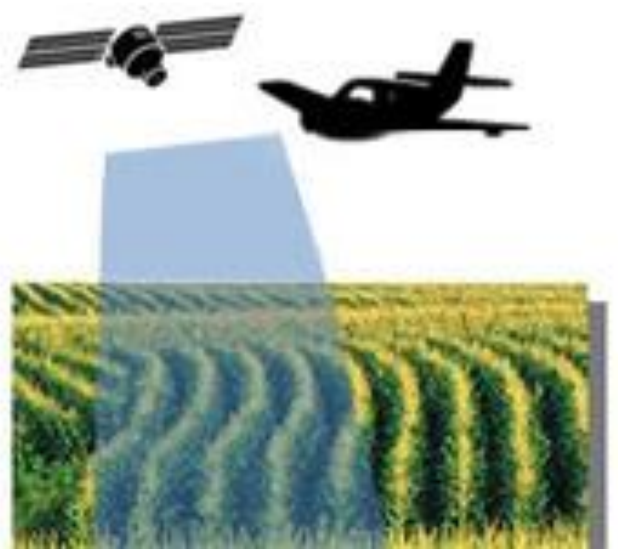
The objective: Transform measurements into biophysical parameter estimates.

The data:

- **Input data:** satellite/airborne spectra, in situ (field) radiometers, or simulated spectra by RTMs
- **Output results:** estimation of a biophysical parameter

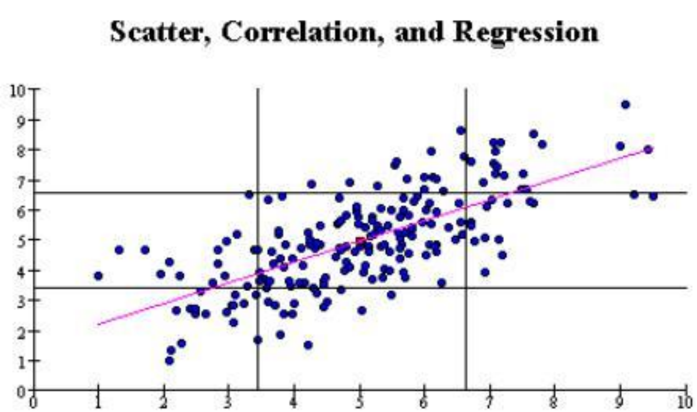


Introduction retrieval biophysical parameters

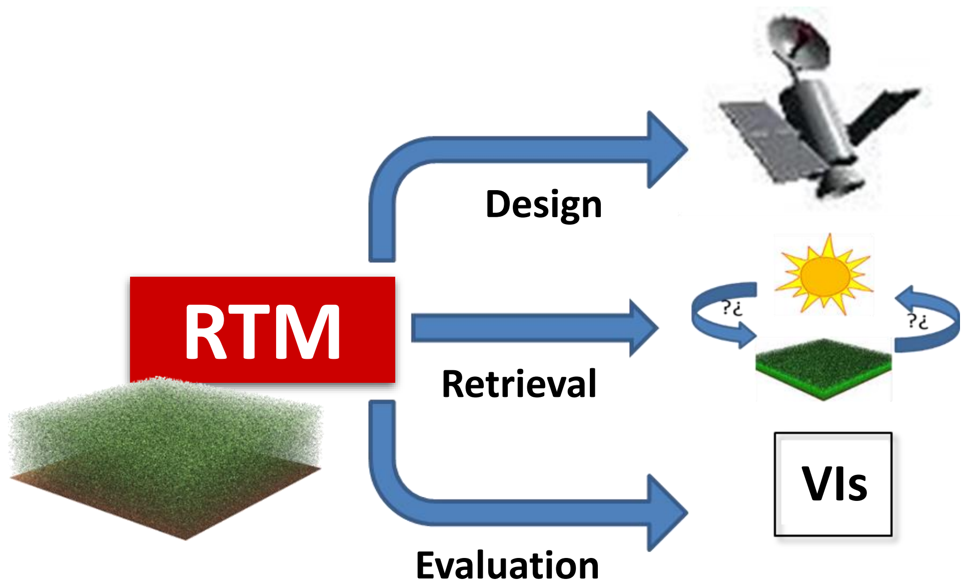


Retrieval of biophysical parameters from Remote Sensing (RS) data **always occurs through a model**, e.g. through statistical models or through inversion of physically-based **radiative transfer models (RTM)**.

Statistical approaches



Physically based RTM approaches

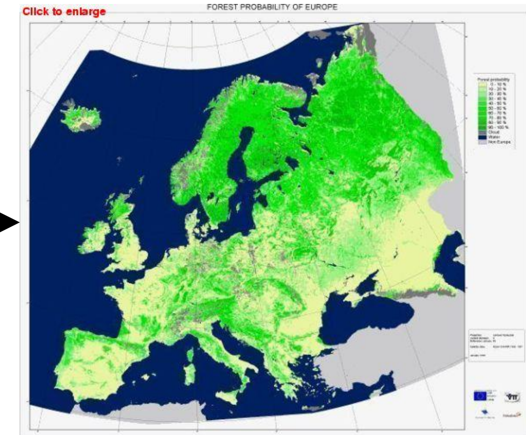


Retrieval of (continuous) vegetation properties

Remote sensing image

Model

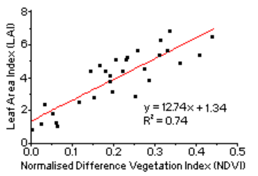
Map of a vegetation property



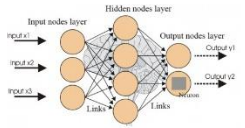
1. **Statistical models**
   1. Parametric regression models
   2. Nonparametric regression models
      1. Linear
      2. Nonlinear
2. **Inversion of physically based radiative transfer models**
   1. Numerical optimization
   2. Lookup-table (LUT)-based inversion

Taxonomy of retrieval methods, three main families:

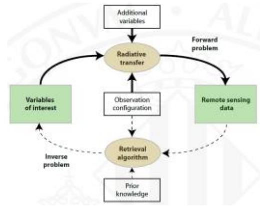
1. ***Statistical:*** **parametric and non-parametric:**
   - **Parametric** models rely on ***some*** **physical knowledge** of the problem and build explicit parametrized expressions that relate a few spectral bands with the biophysical parameter(s) of interest.



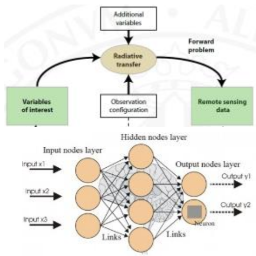
   - **Non-parametric** models are ***data-driven*** **models**. They are adjusted to predict a variable of interest using a training dataset of input-output data pairs.



2. ***Physical***: **try to reverse RTMs.**
   - Physically based algorithms are applications of physical laws establishing photon interaction ***cause–effect relationships***. Model variables are inferred based on specific knowledge, typically obtained with **radiative transfer functions**.



3. ***Hybrid***:
   - A hybrid-method **combines elements of nonparametric statistics and physically based methods**. Hybrid models rely on the generic properties of physically based methods combined with the flexibility and computational efficiency of nonparametric nonlinear regression methods.



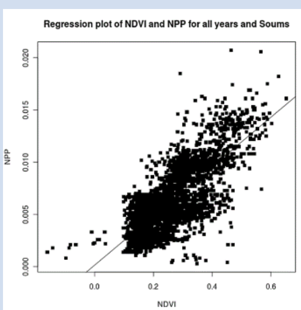
Retrieval families

Parametric regression

Spectral relationships that are sensitive to specific vegetation properties

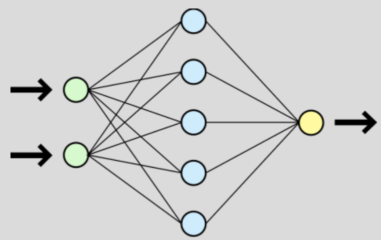
$$NDVI = \frac{(\rho_{NIR} - \rho_{RED})}{(\rho_{NIR} + \rho_{RED})}$$

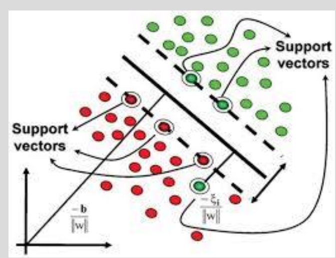
Normalized Difference Vegetation Index



Non-parametric regression

Advanced techniques that search for relationships between spectral data and biophysical variables

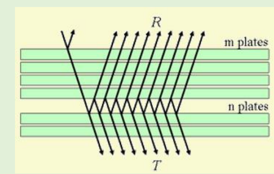


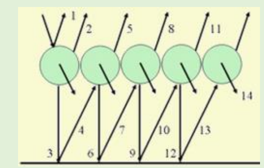


RTM inversion

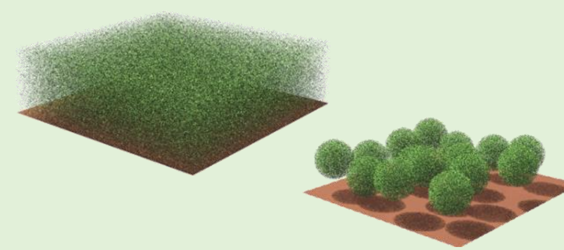
Models that simulate interactions between vegetation and radiation

leaf

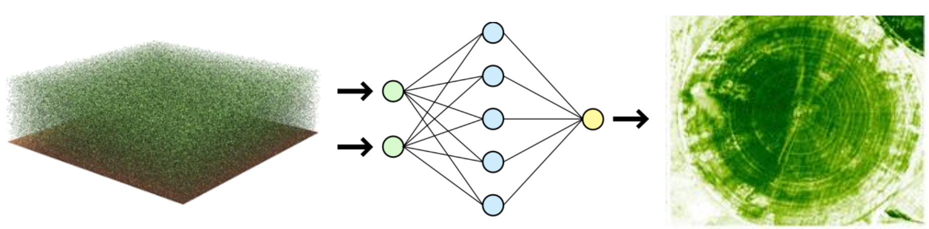




canopy



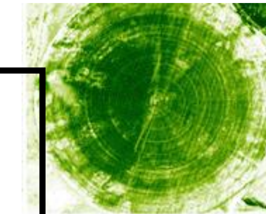
Methods of these different families can be combined: *hybrid methods*



Statistical interpretation of RS

Remote Sensing Data ⟷ **Statistical relationship** ⟷ **Variable of Interest**

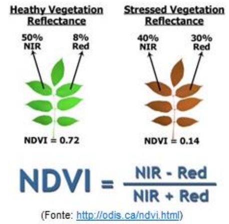
- **Parametric regression**
- **Non-parametric regression**



- Simple statistical relationships (VIs) constitute the **BULK of RS analysis**.
- These analyses allow to determine **IF** there is a relationship, **not WHY** there is a relationship.
- Linear methods such as VIs are **useful indicators** of biophysical (e.g. structure) or biochemical (e.g. chlorophyll) parameters, however in natural, complex environments indices are **confounded** by additional abiotic and biotic factors.
- **VIs lack generality** for estimating biophysical parameters.
- Apart from VIs a large number of powerful **alternative statistical retrieval** methods exists (e.g. non-parametric regression methods).

Parametric regression

Parametric regression assume an explicit model for retrieval

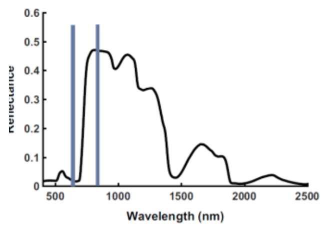


$$PRI = \frac{(\rho 570 - \rho 531)}{(\rho 570 + \rho 531)}$$

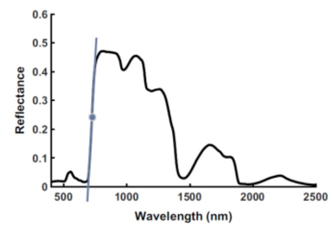
- **Discrete band methods(VIs):**
  - 2-band: SR, NDVI, PRI, OSAVI
  - 3-band: TVI, MCARI, SIPI
  - 4-band: TCARI/OSAVI

$$TCARI / OSAVI = 3 \cdot \left[\left(\rho_{\mu 700} - \rho_{\mu 670} \right) - 0.2 \cdot \left(\rho_{\mu 700} - \rho_{\mu 550} \right) \cdot \right] \left(\rho_{\mu 700} - \rho_{\mu 670} \right) \left(1 + 0.16 \right) \cdot \left(\rho_{\mu 800} - \rho_{\mu 670} \right) / \left(\rho_{\mu 800} - \rho_{\mu 670} + 0.16 \right) \qquad (2)$$

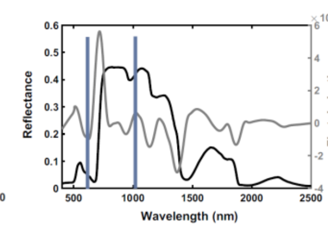
- **Shape-based methods:**
  - Red-edge position (REP)
  - Derivative/Integral indices
  - Continuum removal
  - wavelet



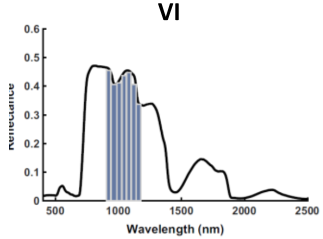
VI



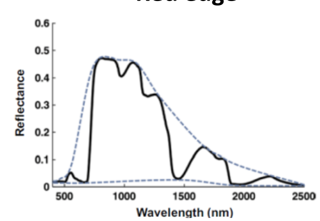
Red edge



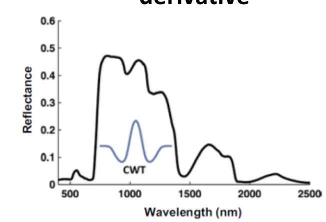
derivative



integral

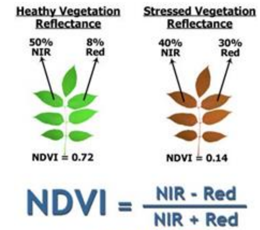


Continuum removal



wavelet

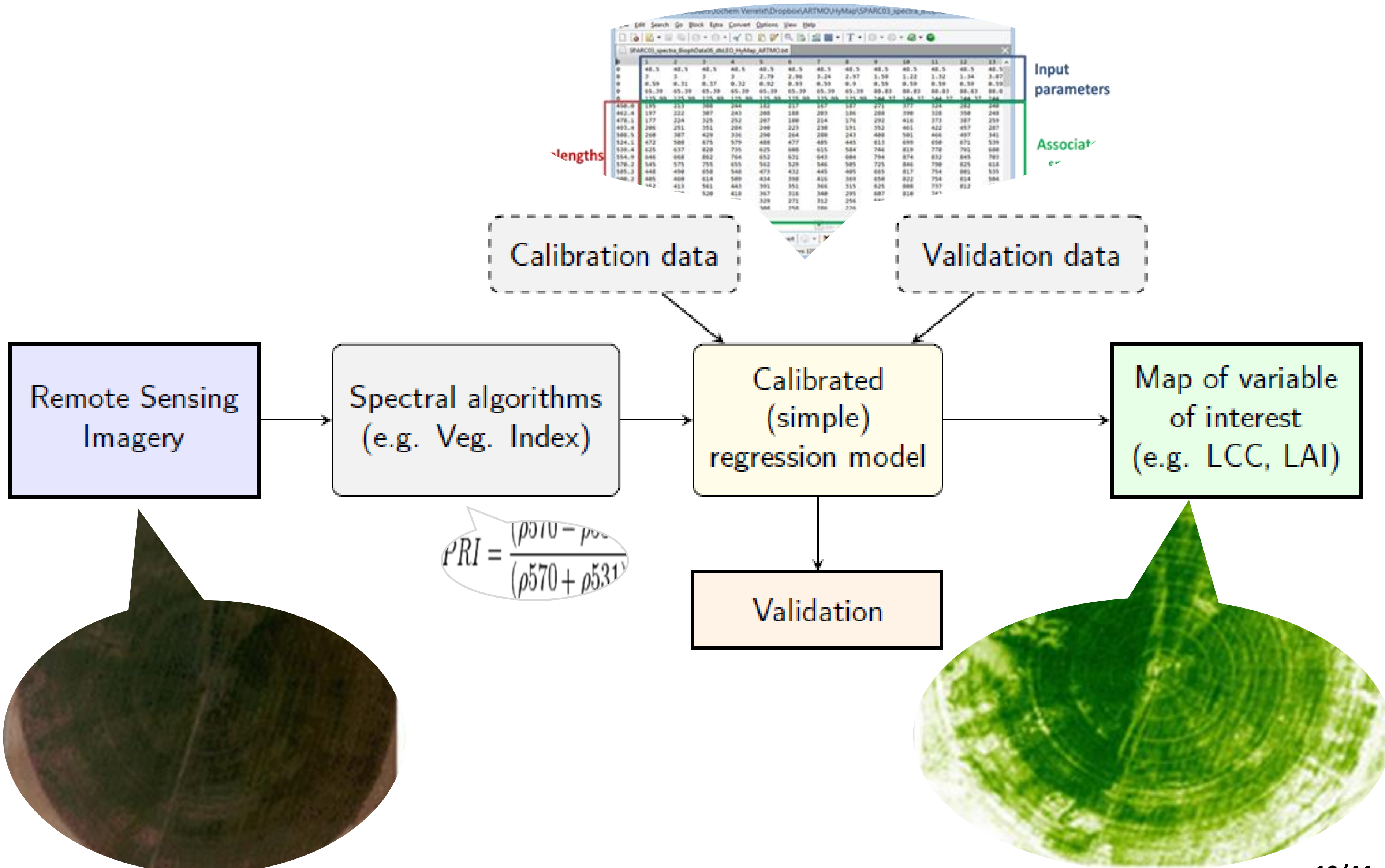
Parametric regression:



(Fonte: http://odis.ca/ndvi.html)

Strengths	Weaknesses
• Simple and comprehensive regression models; little knowledge of user required. • Fast in processing • Computationally inexpensive	• Makes only poorly use of the available information within the spectral observation; at most a spectral subset is used. Therefore, they tend to be more noise-sensitive as compared to full-spectrum methods • Parametric regression puts boundary conditions at the level of chosen bands, formulations and regression function. • Statistical function accounts for one variable at a time. • A limited portability to different measurement conditions or sensor characteristics • No uncertainty estimates are provided. Hence the quality of the output maps remains unknown.

Parametric regression



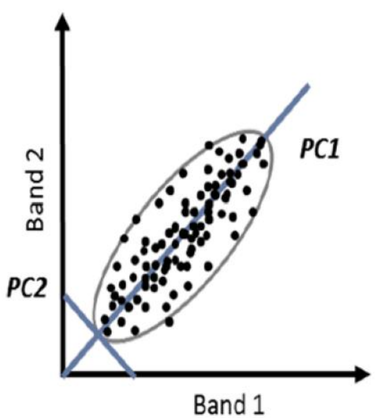
Non-parametric models (1/2):

Data-driven methods: Do not assume explicit feature relations

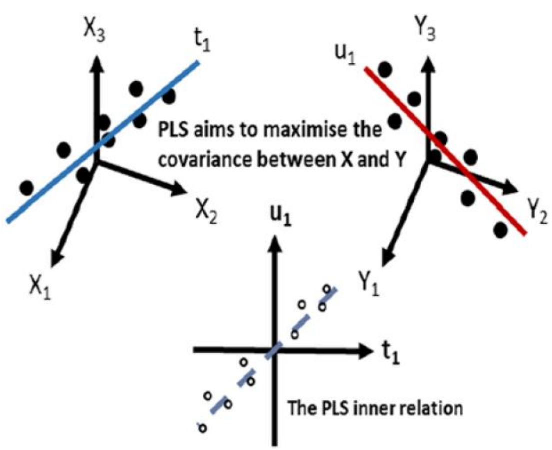
Linear nonparametric models:

- Stepwise multiple linear regression (SMLR)
- Principal component regression (PCR)
- Partial least squares regression (PLSR)
- Ridge regression (RR)
- Least Absolute Shrinkage and Selection Operator (LASSO)

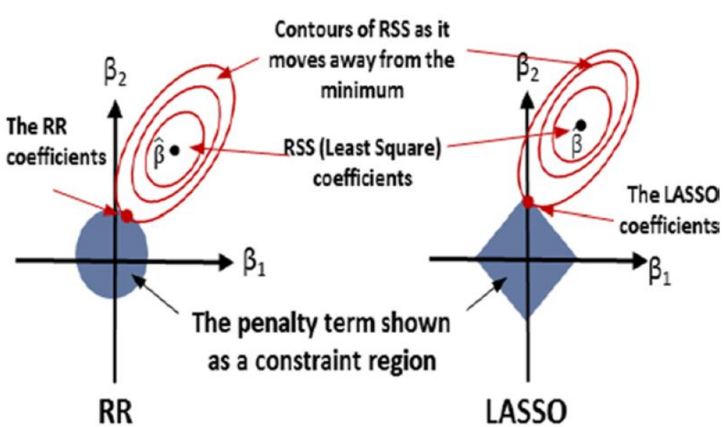
PCR



PLSR



RR & LASSO

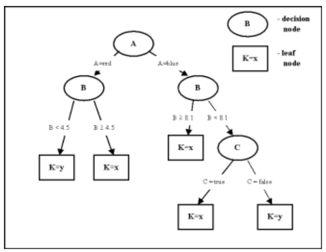


Non-parametric models (2/2):

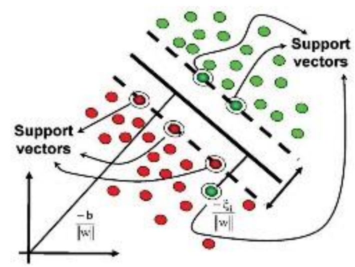
Data-driven methods: Do not assume explicit feature relations

Non-linear nonparametric models:

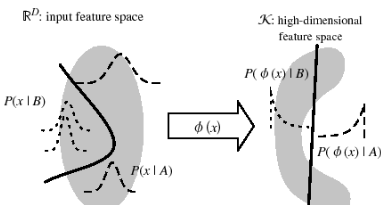
Decision Trees (DT)



Support vector regression (SVR)

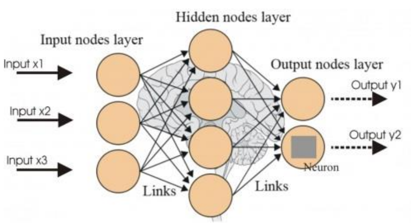


Gaussian processes regression (GPR)

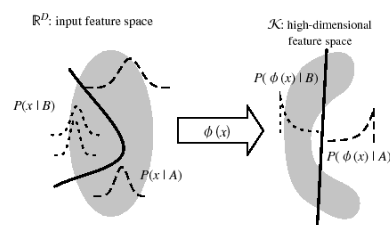


$$K(\mathbf{x}_i, \mathbf{x}_j) = \nu \exp\left(- \sum_{b=1}^{B} \frac{\left(\mathbf{x}_i^{(b)} - \mathbf{x}_j^{(b)}\right)^2}{2\sigma_b^2} \right)$$

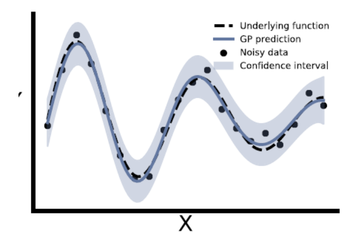
Neural networks (NN)



Kernel ridge regression (KRR)



$$K(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\|\mathbf{x}_i - \mathbf{x}_j\|^2 / (2\sigma^2)).$$



<u>Also:</u>

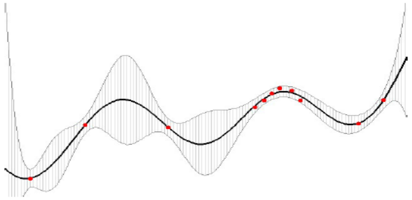
- Elastic Net (ELASTICNET)
- Bagging trees (BAGTREE)
- Boosting trees (BOOST)
- Neural networks (NN)
- Extreme Learning Machines (ELM)
- Relevance Vector Machine (RVM)
- Gaussian process Regression (GPR)
- Variational Heteroscedastic Gaussian Process Regression (VHGPR)

Gaussian Processes Regression (GPR)

- A **GPR** model is a **probabilistic** (**Bayesian**) model directly in function space, with no intermediate model or model parameters.

- **GPR** are **equivalent** to **kernel ridge regression**, least square **suport vector machines** (SVM), **Kriging**, large **neural networks** (NN) and very closely related to **SVM regulazation networks**.

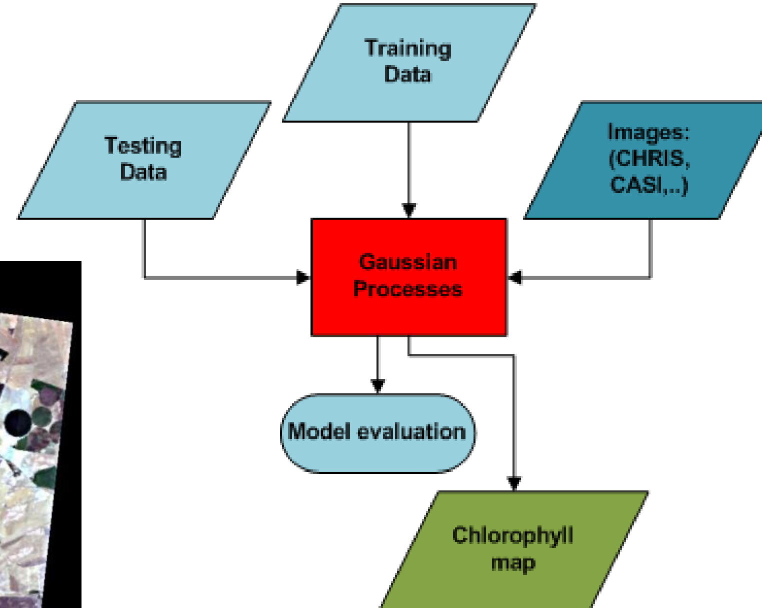
- **GPR alleviates** some **shortcomings** of the previous methods, while maintaining very good numerical performance and stability:
  - GPR is far more **simple than NN**, and needs **less sample points** ☺
  - **Not only a mean prediction** for each sample (**pixel**), but also a full distribution over the output values **including an uncertainty of the prediction** (**confidence interval**). ☺
  - GPR **provide a ranking of features (bands) and samples (spectra)**, thus partly **overcoming the blackbox problem**. ☺
  - http://www.rainsoft.de/projects/gausspro.html



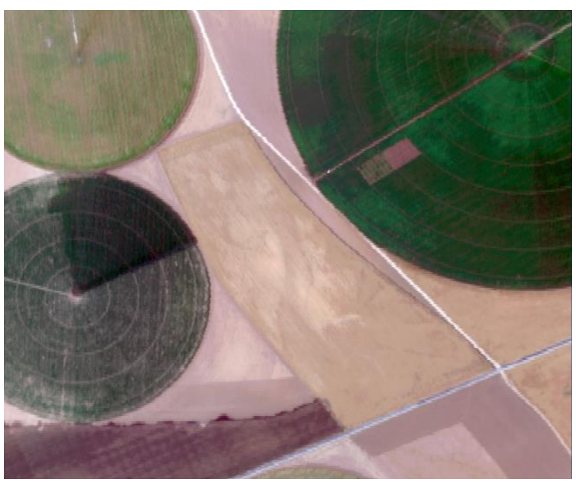
A regression curve plot by the "Gaussian Process Regression Applet" using 11 data points. One can observe that uncertainty goes down when multiple data points are aggregated together.

More info on: http://www.gaussianprocess.org/
Rasmussen and C Williams, **Gaussian Processes for Machine Learning**, 2006

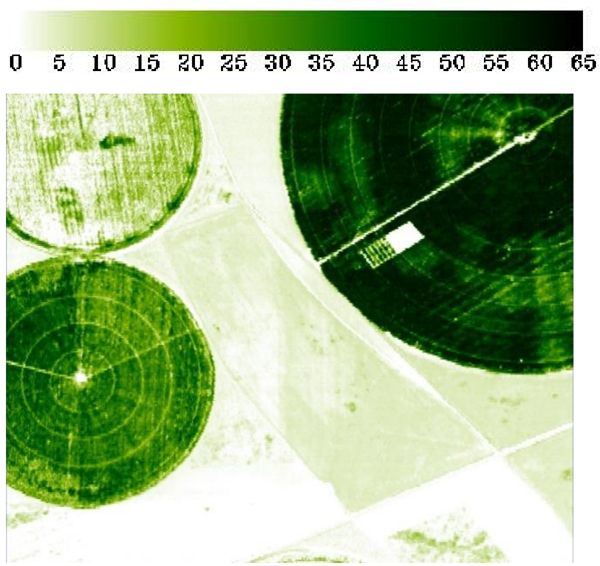
Example GPR



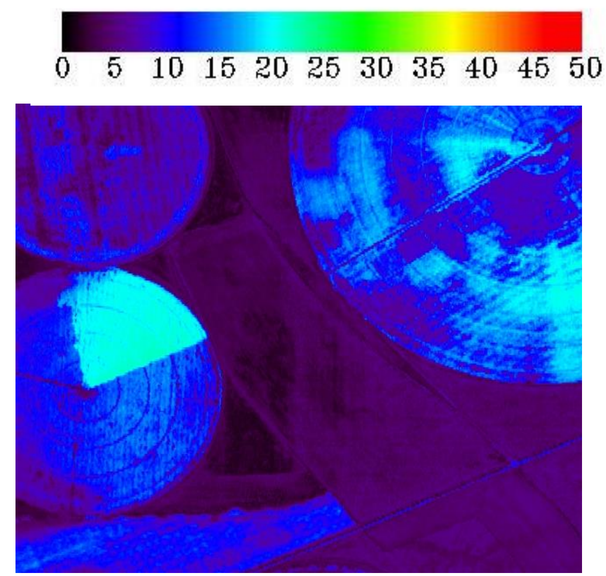
RGB CASI



Chl [µg/cm^2]



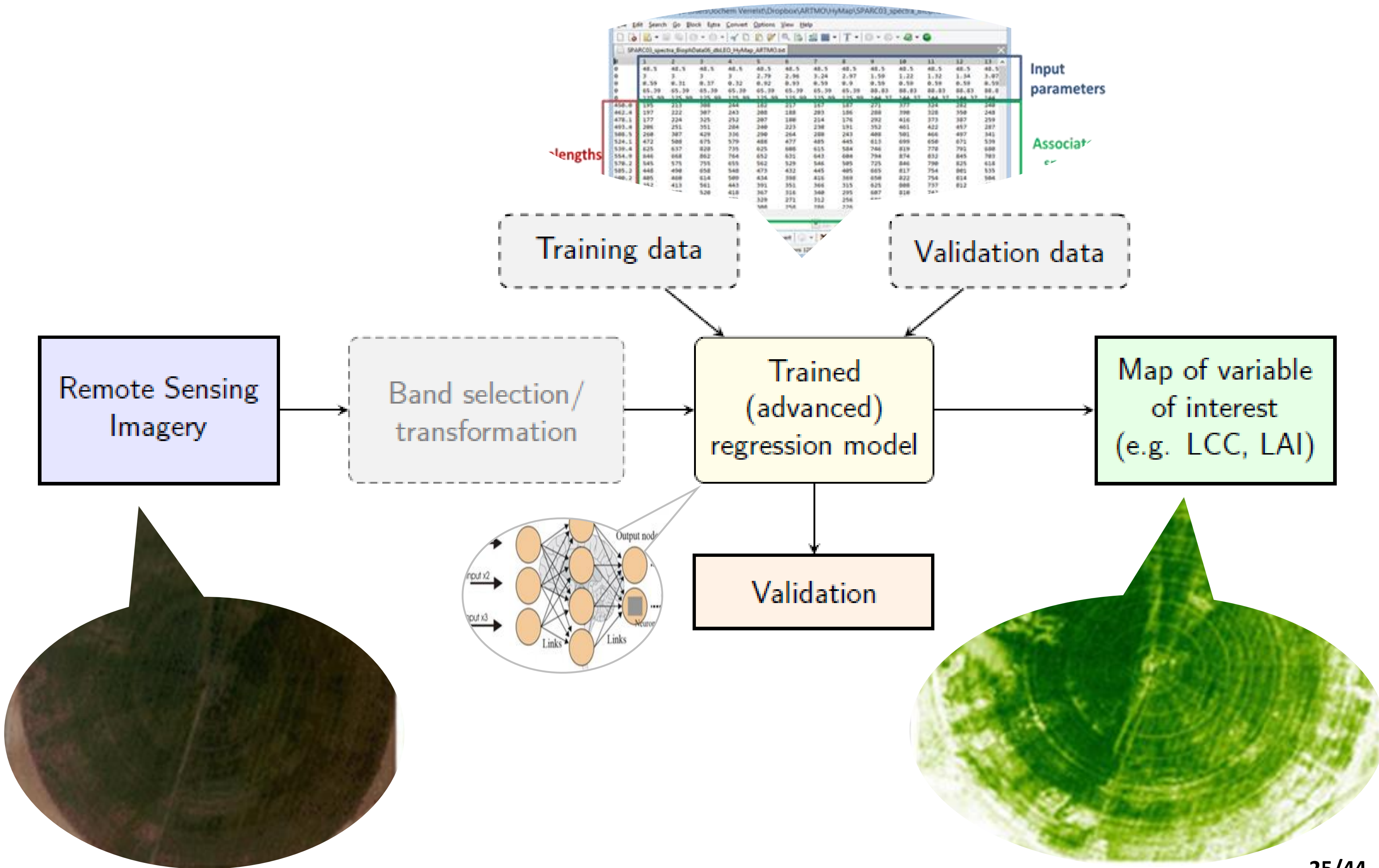
St Dev
Chl [µg/cm2]

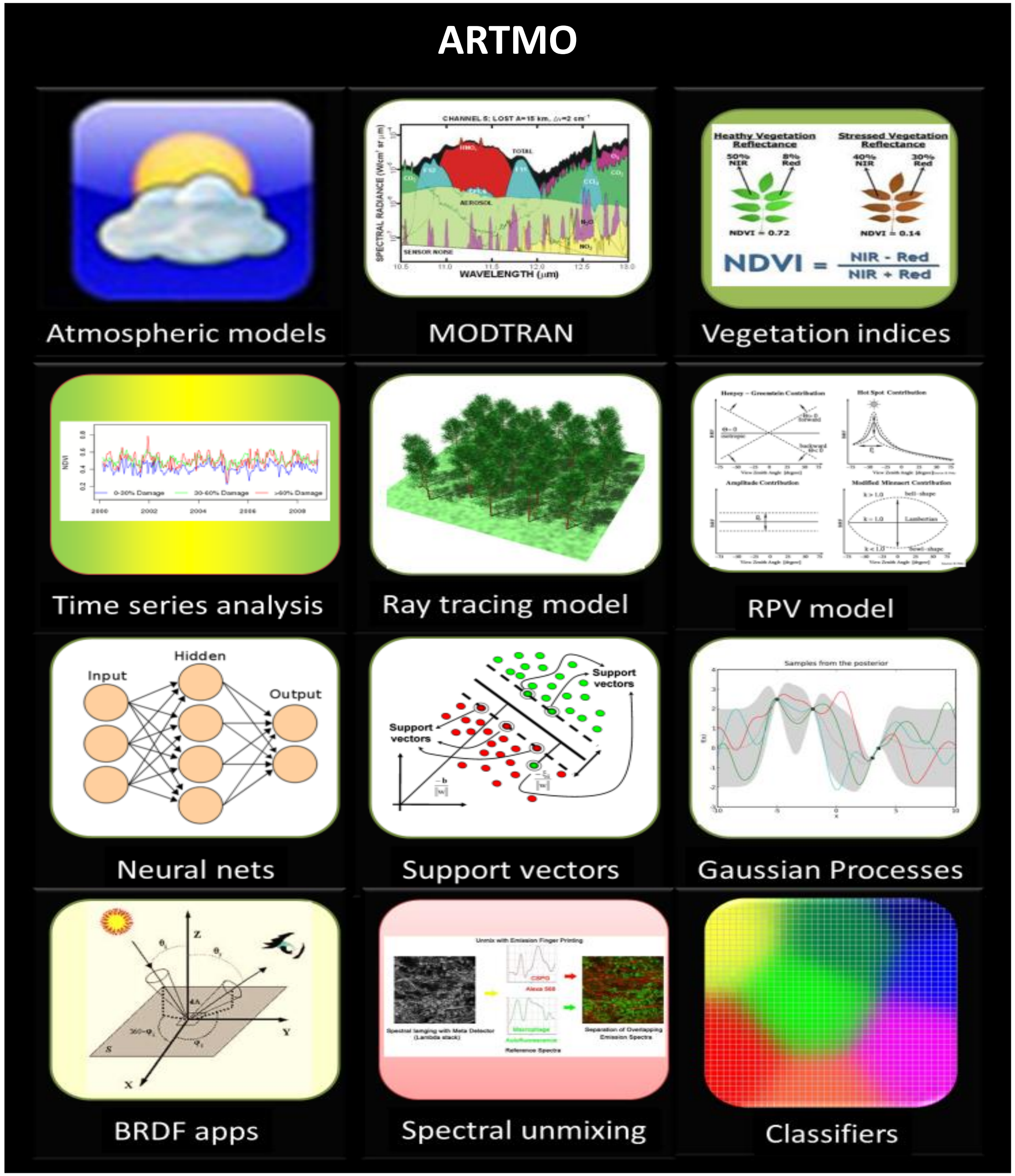


Non-parametric regression:

Strengths	Weaknesses
• Full-spectrum methods. They make use of the complete spectral information. • Advanced, adaptive (non-linear) models are built. • Methodologically, accurate and robust performance is enabled. • Some MLRAs cope well with datasets showing redundancy and high noise levels. • Once trained, imagery can be processed time efficient. • Some of the non-parametric methods (e.g. ANNs, decision trees) can be trained with a high number of samples (typically >1,000,000). • Some MLRAs provide insight in model development (e.g. GPR: relevant bands; decision trees: model structure). • Some MLRAs can provide multiple-outputs (e.g. PLRS, ANN, SVR, GPR and KRR) • Some MLRAs provide uncertainty intervals (e.g. GPR).	• Training can be computational expensive. • Hypercomplex models can be generated. Their generic potential is limited and hence they do not generalize well, based on the training data (problem of over-fitting). • Some regression algorithms are difficult (or even impossible) to train with a high number of samples. • Expert knowledge is required, e.g. for tuning. However, toolboxes exist automating some of the steps in this sub-process. • Some of the methods can be considered as black boxes. • Some regression algorithms elicit instability when applied with datasets statistically deviating from the datasets used for training.

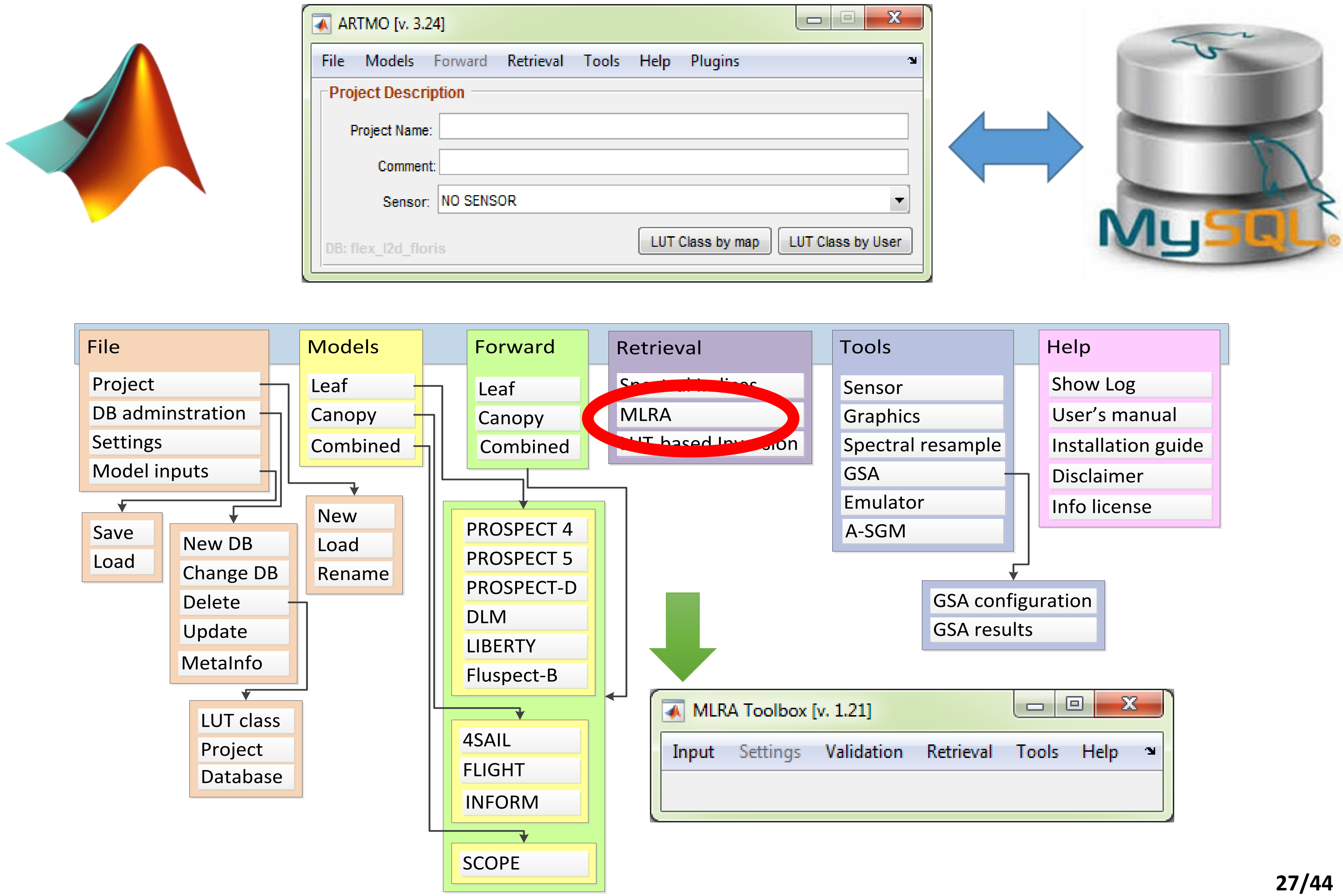
Non-parametric regression



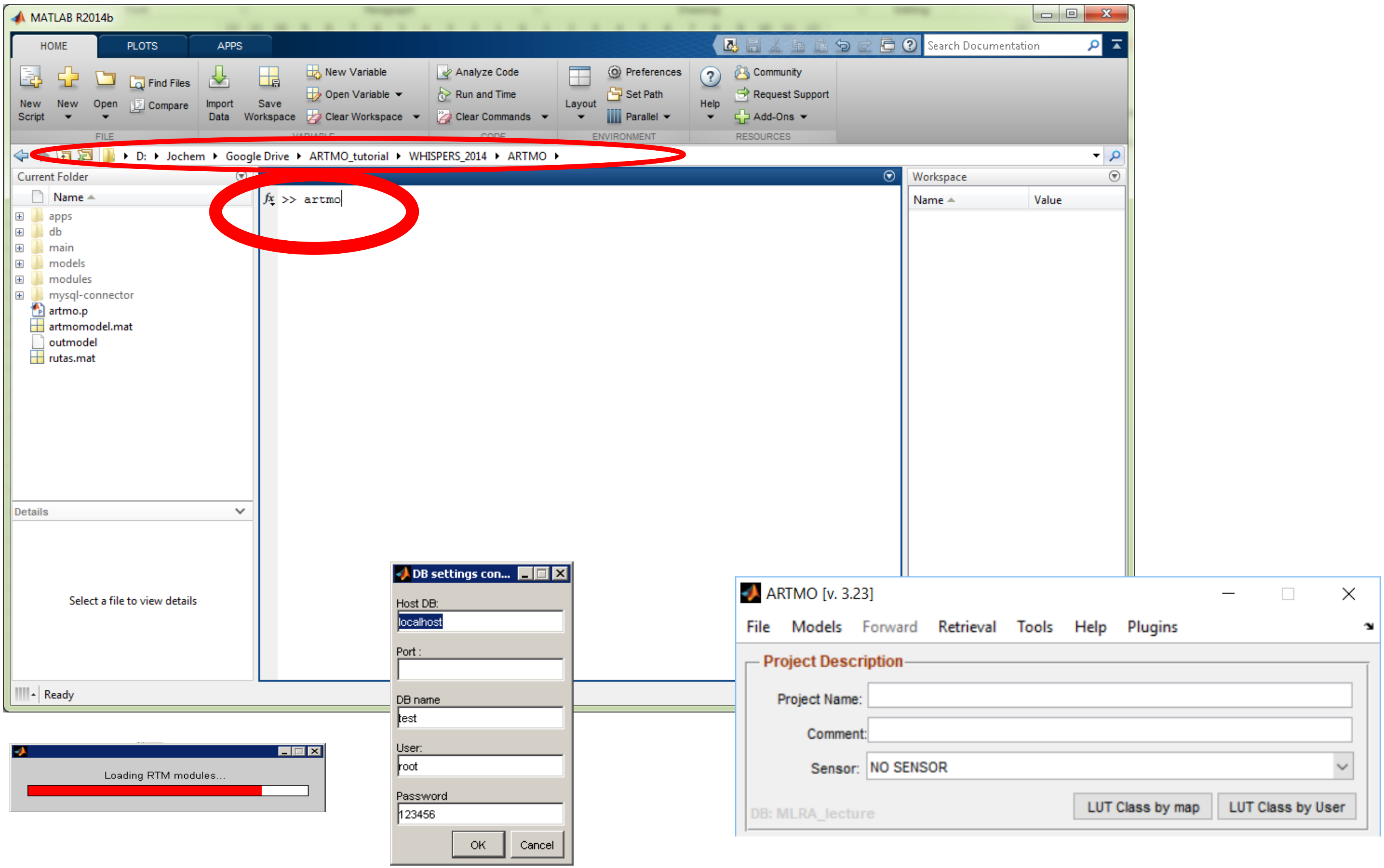

ARTMO
Atmospheric models
MODTRAN
Vegetation indices
Time series analysis
Ray tracing model
RPV model
Input
Hidden
Output
Neural nets
Support vectors
Support vectors
Support vectors
Gaussian Processes
BRDF apps
Spectral unmixing
Classifiers

ARTMO v. 3.24

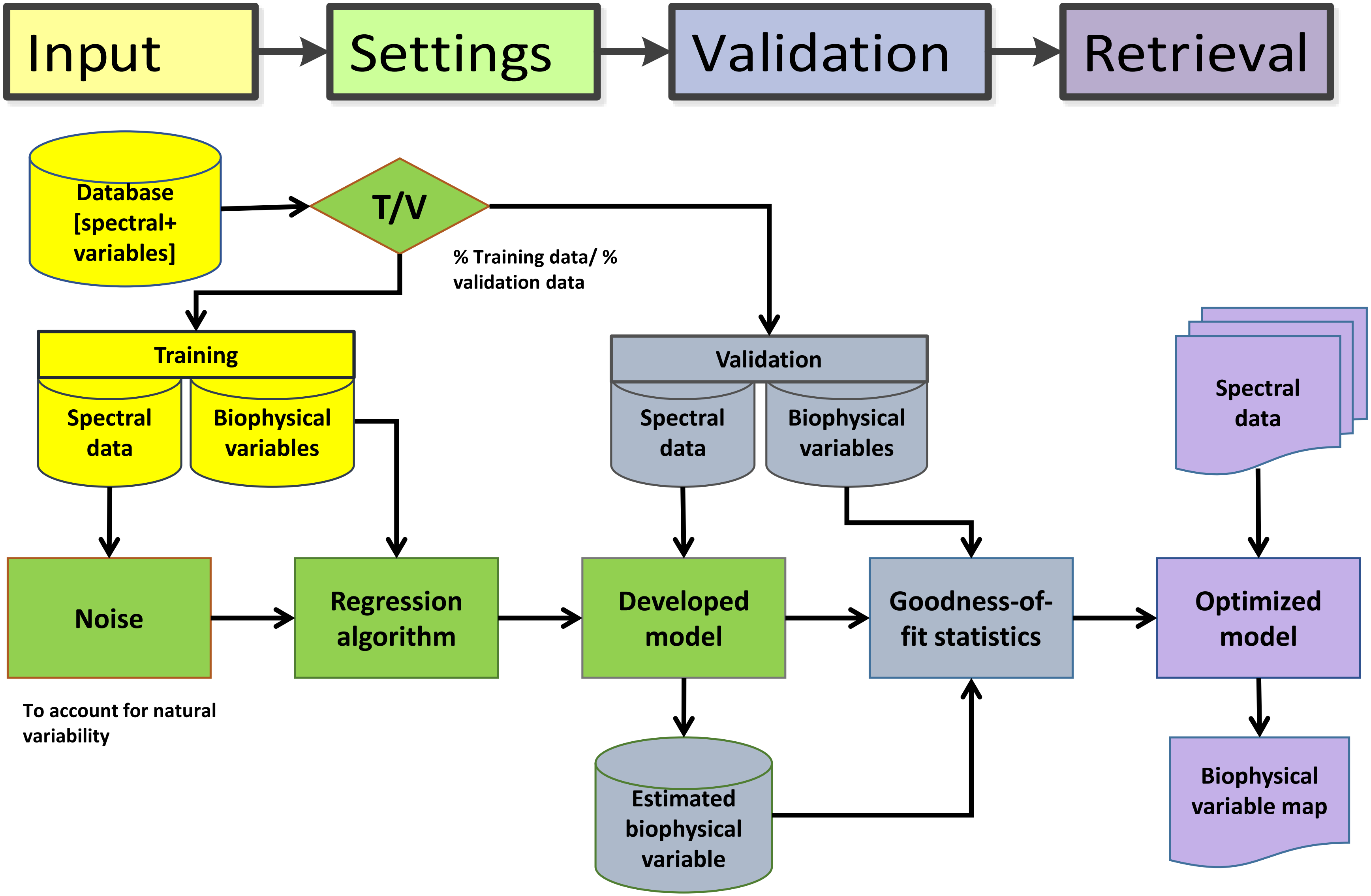
http://ipl.uv.es/artmo/



Open Matlab and provide the ARTMO path
In Matlab Command Window: artmo



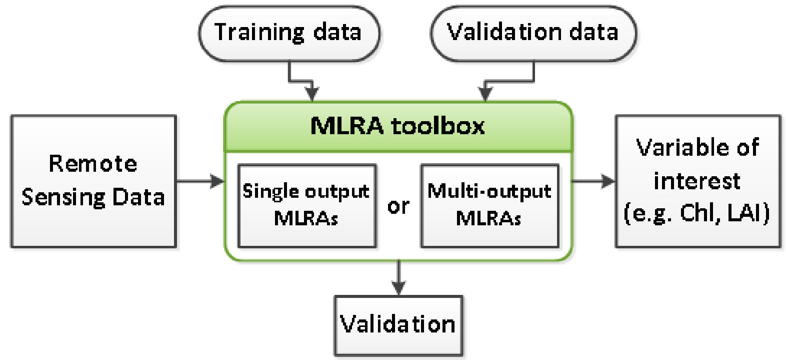
Schematic overview for systematic evaluation of nonparametric regression models to estimate biophysical variables

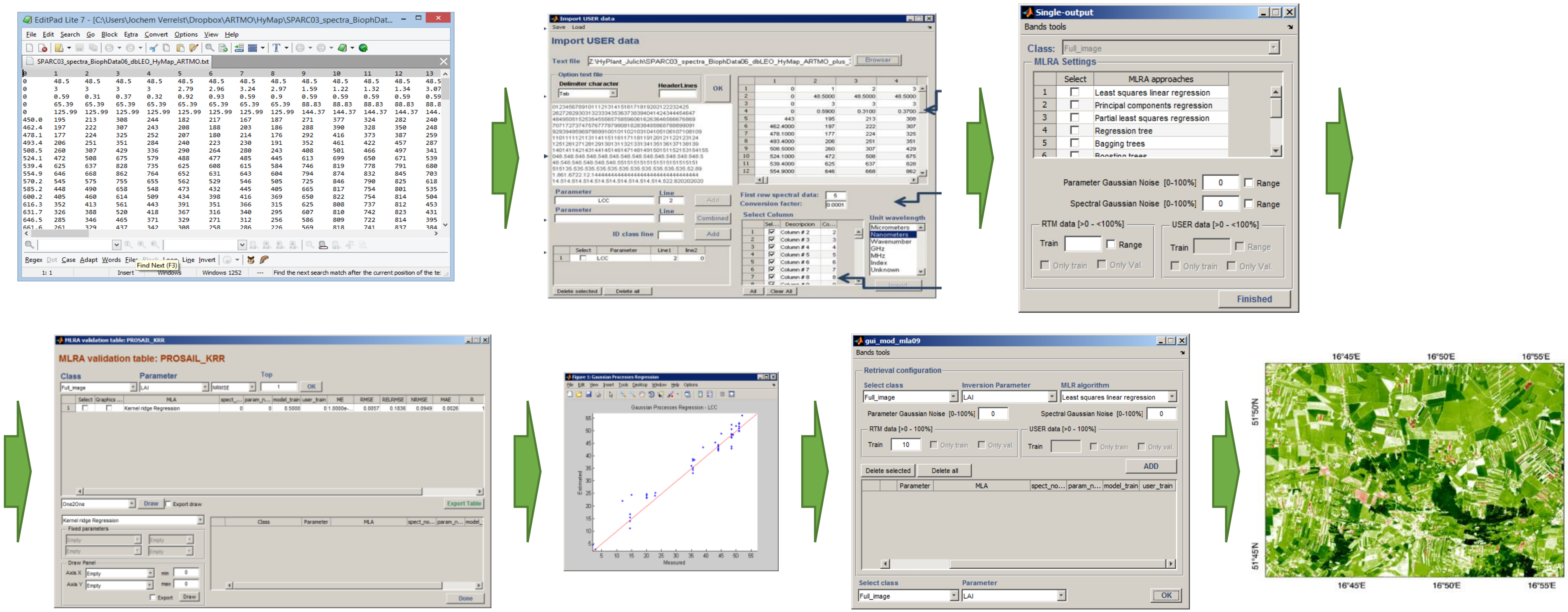


Outlook:

MLRA mapping based on User data

- User Input data
- MLRA setting
- Validation
- Mapping



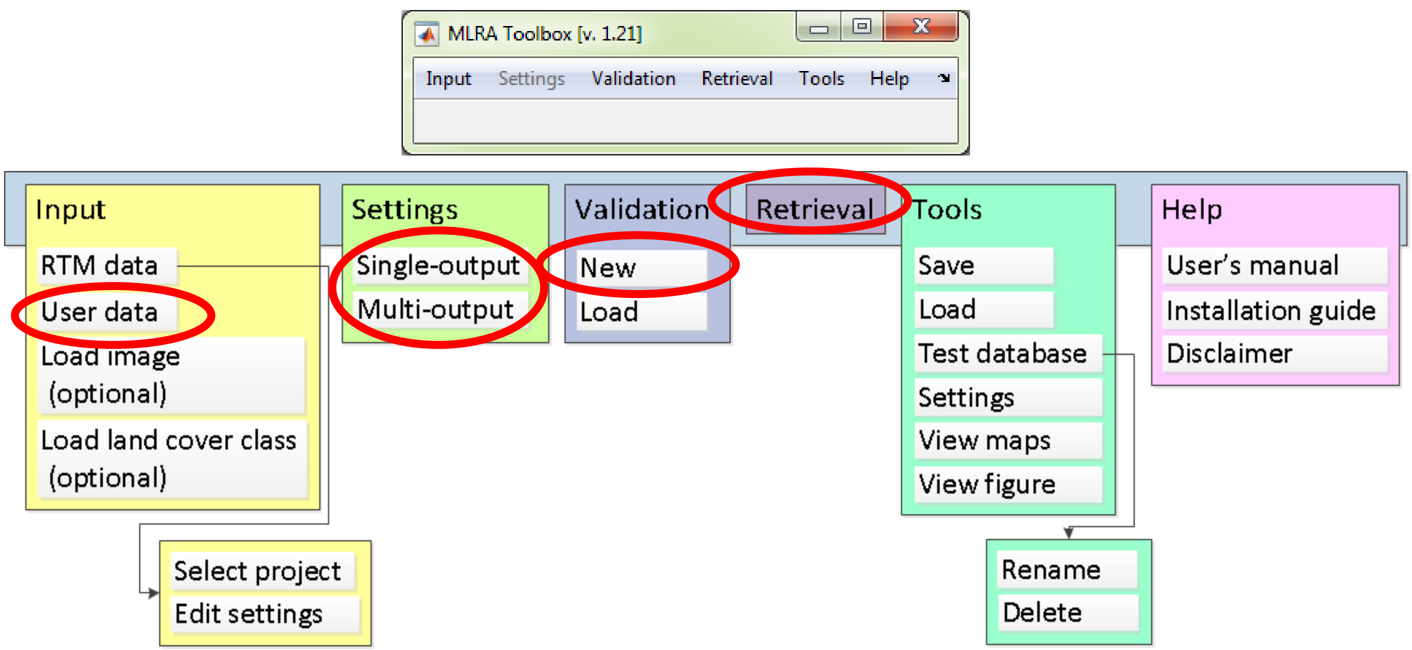


MLRA mapping based on User data

To train and validate single-output and multi-output MLRA models and apply it to an image using a field dataset.

The procedure will be as follows:

1. ***User data:*** *Insert field data for training and validation*
2. ***Single-output:*** *Choose single-output MLRA models and define training/testing partitioning*
3. ***Validation:*** *Validate the defined MLRA strategies*
4. ***Retrieval:*** *Apply the best one to a remote sensing image.*

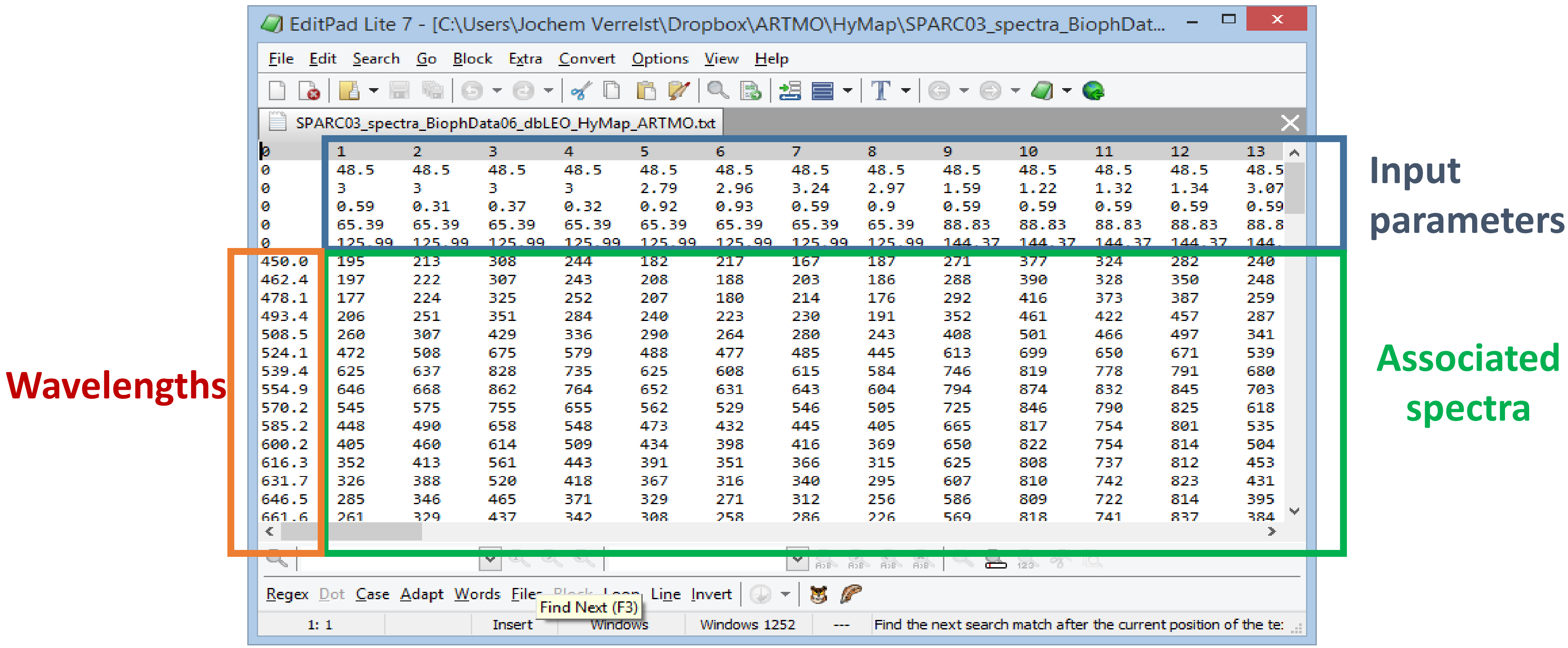


Input: User data (e.g. field data)

User data for training and validation requires one input file, including:

1. **Biophysical parameters (e.g., LAI, chlorophyll content,…)**
2. **Associated spectra (e.g., obtained from a remote sensing image)**

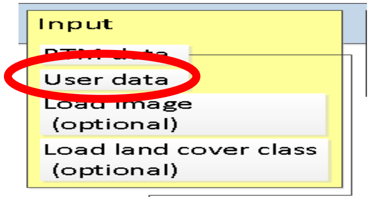
User data need to be organized in a matrix format in plain text file, according to example below:

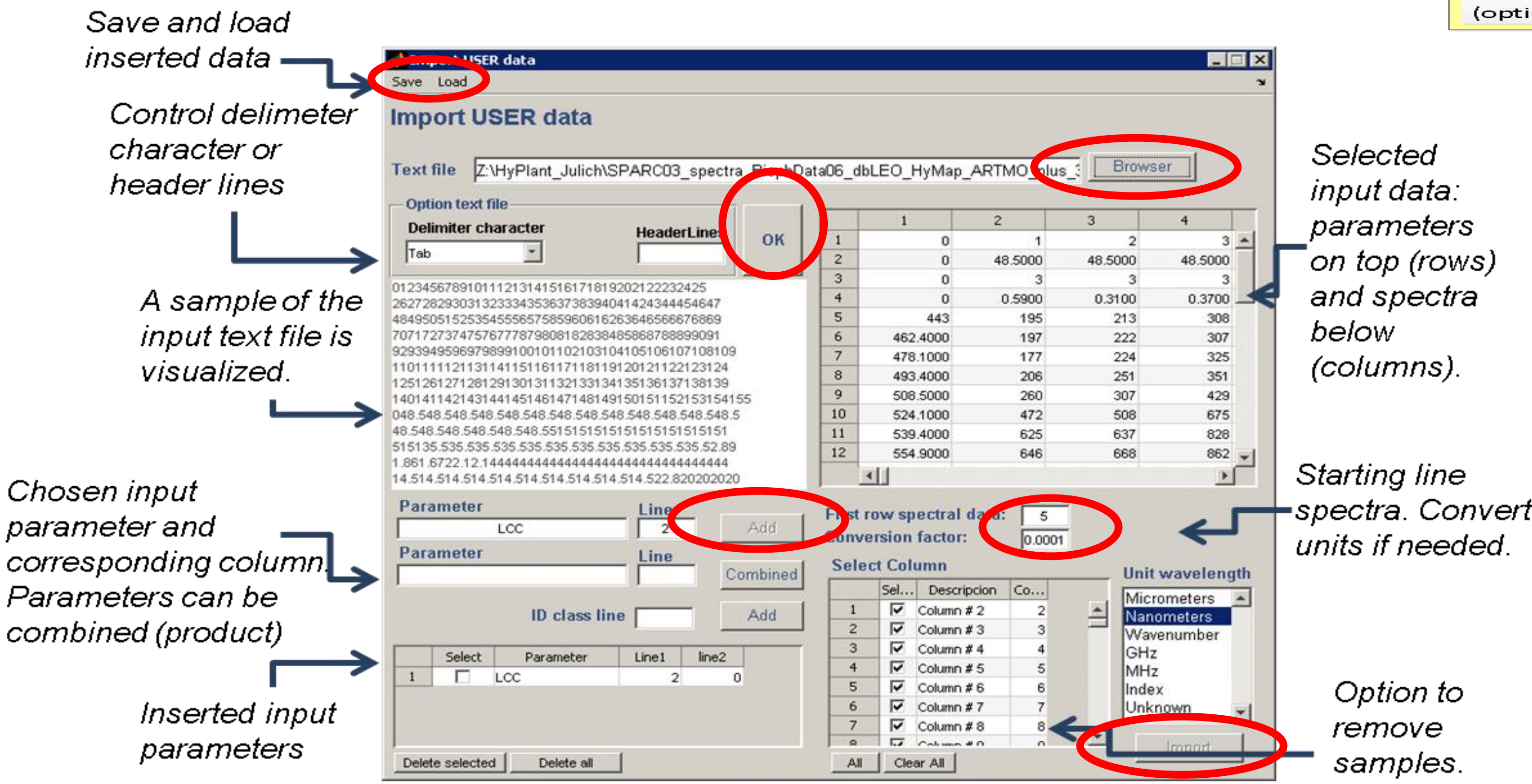


0	1	2	3	4	5	6	7	8	9	10	11	12	13
0	48.5	48.5	48.5	48.5	48.5	48.5	48.5	48.5	48.5	48.5	48.5	48.5	48.5
0	3	3	3	3	2.79	2.96	3.24	2.97	1.59	1.22	1.32	1.34	3.07
0	0.59	0.31	0.37	0.32	0.92	0.93	0.59	0.9	0.59	0.59	0.59	0.59	0.59
0	65.39	65.39	65.39	65.39	65.39	65.39	65.39	65.39	88.83	88.83	88.83	88.83	88.8
0	125.99	125.99	125.99	125.99	125.99	125.99	125.99	125.99	144.37	144.37	144.37	144.37	144.
450.0	195	213	308	244	182	217	167	187	271	377	324	282	240
462.4	197	222	307	243	208	188	203	186	288	390	328	350	248
478.1	177	224	325	252	207	180	214	176	292	416	373	387	259
493.4	206	251	351	284	240	223	230	191	352	461	422	457	287
508.5	260	307	429	336	290	264	280	243	408	501	466	497	341
524.1	472	508	675	579	488	477	485	445	613	699	650	671	539
539.4	625	637	828	735	625	608	615	584	746	819	778	791	680
554.9	646	668	862	764	652	631	643	604	794	874	832	845	703
570.2	545	575	755	655	562	529	546	505	725	846	790	825	618
585.2	448	490	658	548	473	432	445	405	665	817	754	801	535
600.2	405	460	614	509	434	398	416	369	650	822	754	814	504
616.3	352	413	561	443	391	351	366	315	625	808	737	812	453
631.7	326	388	520	418	367	316	340	295	607	810	742	823	431
646.5	285	346	465	371	329	271	312	256	586	809	722	814	395
661.6	261	329	437	342	308	258	286	226	569	818	741	837	384

- Make sure to fill up the whole Matrix! In case of empty cells, use **NaN** and remove those samples in the following step.
- Make sure that wavelengths are the same as the remote sensing image! They need to match. A band selection or band transformation can be later done in **Settings**.

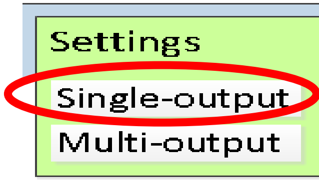
Import User data window





1. **Browser**: Import User data file.
2. Inspect if right data in left panel. By clicking on **OK** data will appear in right panel.
3. Define a row with a parameter to its line. Click on **Add**. Multiple parameters can be define by repeating this step. Parameters can be combined.
4. **Define the row where spectra starts.**
5. If needed, **convert** spectral data.
6. Option to **remove** samples.
7. Configured input data can be **save**d and **load**ed as .m file.
8. Finally, click on **Import**.

Single-output Settings

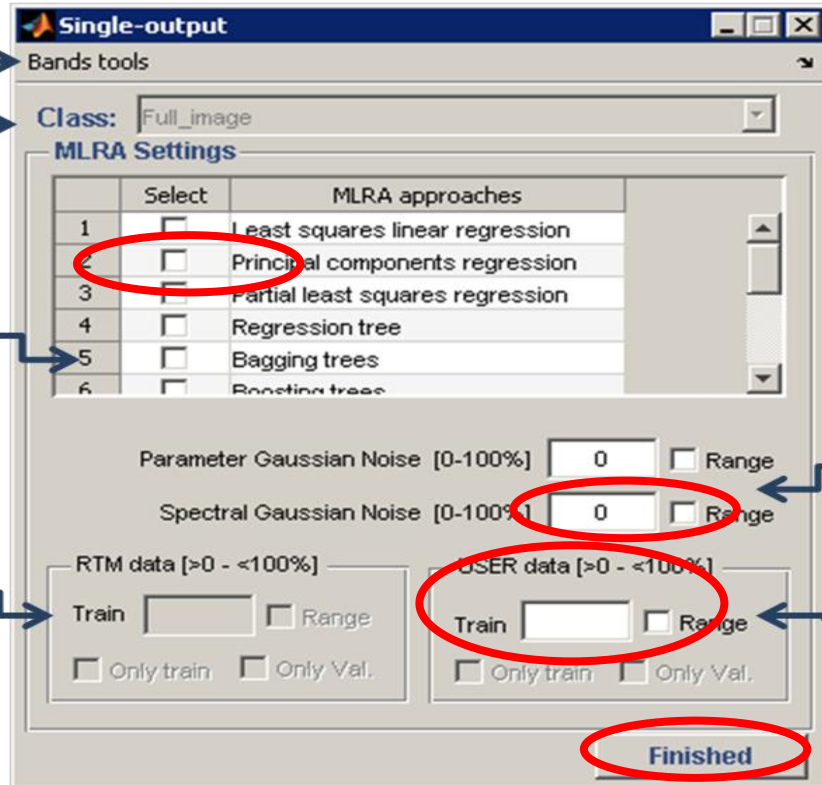


Option to apply a band subselection or transformation

If a land cover map has been provided, per clas scan be configured.

Multiple regressors can be selected

If RTM data is inserted it can serve for training or validation

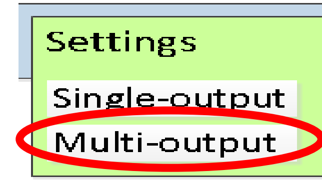


Options to add noise to parameters and spectral data

Options to control the training/validation partitioning for user data

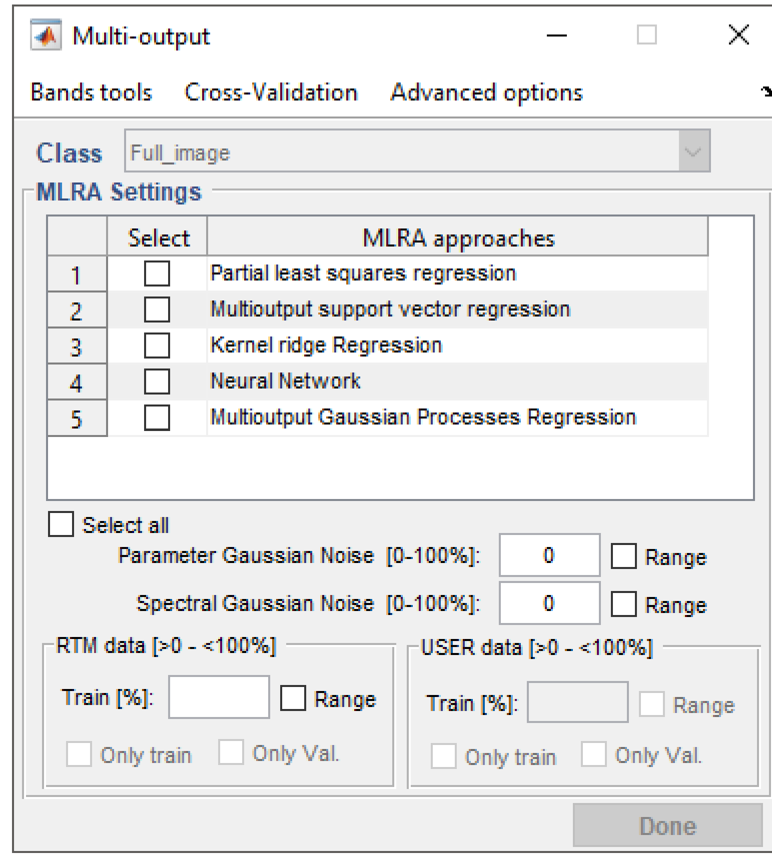
1. Select the **MLRA** to be trained and validated.
2. The option to add Gaussian **noise** is provided. A range of noise scenarios can be applied.
3. Select the User data **training/validation partitioning**. This will randomly partition the input data in a training and validation dataset. **Make sure to keep some data for validation (thus < 100% training).** Also a range of training/validation partitioning scenarios can be applied. **If no validation is required, go directly to Retrieval.**
4. Click on **Finished.**

Multi-output Settings



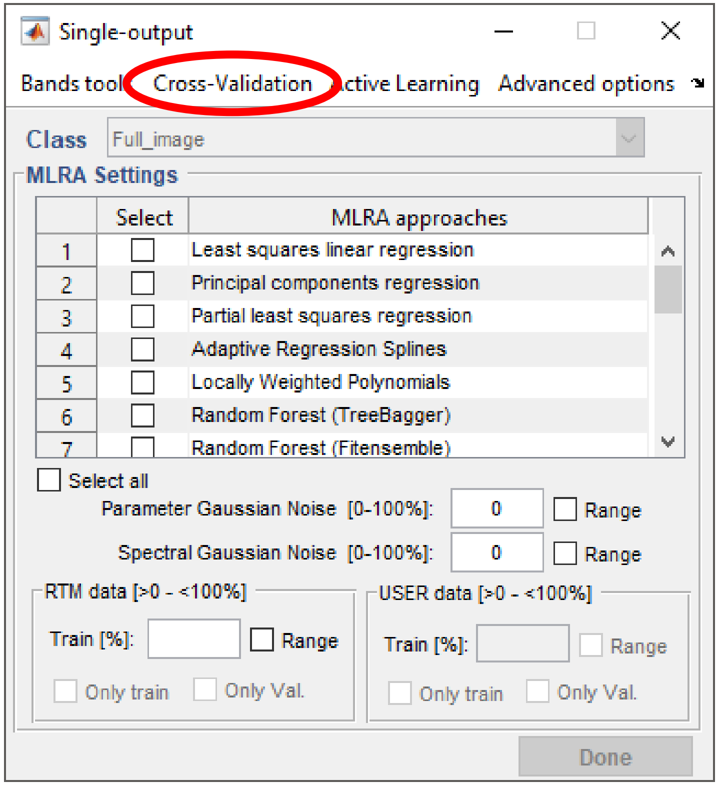
Prepare models for multiple variables. When multiple input variables have been selected, the following regression algorithms provide multiple-outputs with the same single model:

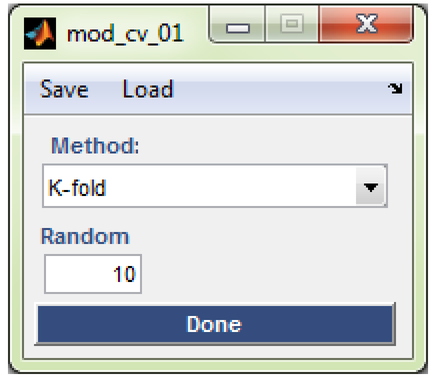
- partial least square regression (PLSR)
- Multioutput support vector regression (MSVR)
- neural networks (NN)
- kernel ridge regression (KRR)
- Multi-output Gaussian process regression (MGPR)

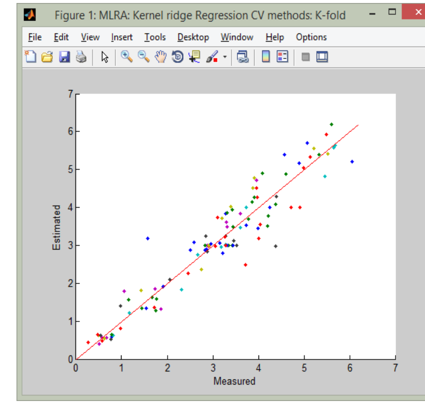


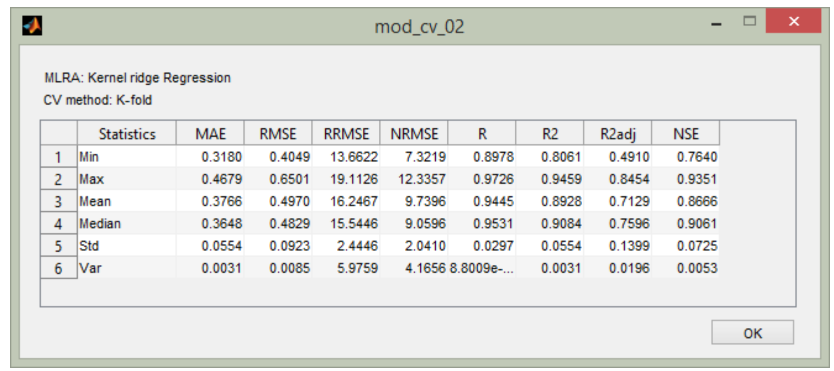
The same options as Single-output are provided.

Cross-validation sub-sampling options







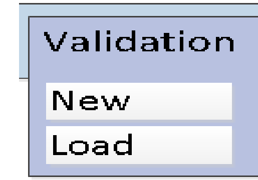


	Statistics	MAE	RMSE	RRMSE	NRMSE	R	R2	R2adj	NSE
1	Min	0.3180	0.4049	13.6622	7.3219	0.8978	0.8061	0.4910	0.7640
2	Max	0.4679	0.6501	19.1126	12.3357	0.9726	0.9459	0.8454	0.9351
3	Mean	0.3766	0.4970	16.2467	9.7396	0.9445	0.8928	0.7129	0.8666
4	Median	0.3648	0.4829	15.5446	9.0596	0.9531	0.9084	0.7596	0.9061
5	Std	0.0554	0.0923	2.4446	2.0410	0.0297	0.0554	0.1399	0.0725
6	Var	0.0031	0.0085	5.9759	4.1656	8.8009e-...	0.0031	0.0196	0.0053

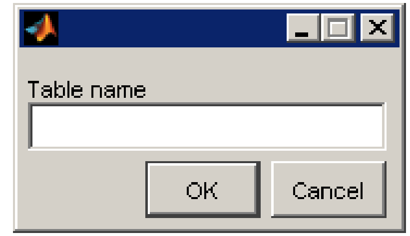
- **K-fold:** The partition divides the observations into *k* disjoint subsamples (or *folds*), chosen randomly but with roughly equal size.

- **Leave-one-out:** Leave-one-out is a special case of 'K-fold', in which the number of folds equals the number of observations.
- **Hold-out:** This partition divides the observations into a training set and a test (or *holdout*) set.

Validation



✓Start a **New** validation: provide a name

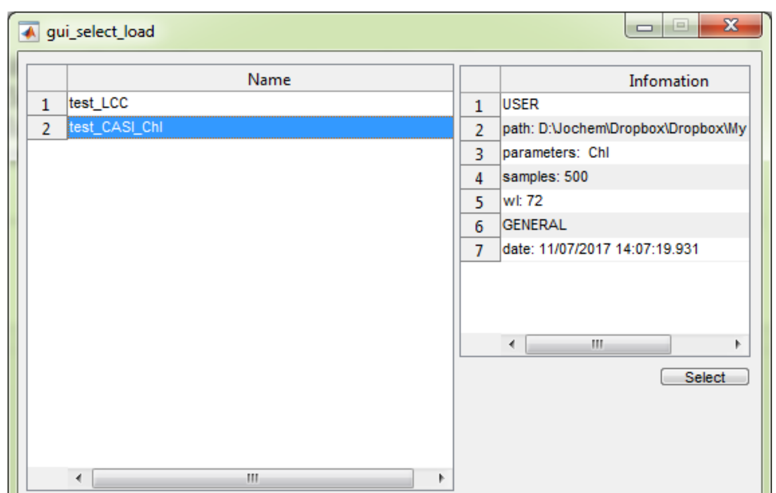


✓All MLRA scenarios will be trained using training data and validated against validation data according to goodness-of-fit indicators:

- **R, R^2, RMSE, RRMSE, NRMSE, ME, MAE**

✓Results will be automatically stored in a MySQL table.

✓When finished, an overview table will appear (*see further*). Such overview table can also be consulted when selecting: **Load.** A window with generated validation results and metadata will appear: (see next slide)



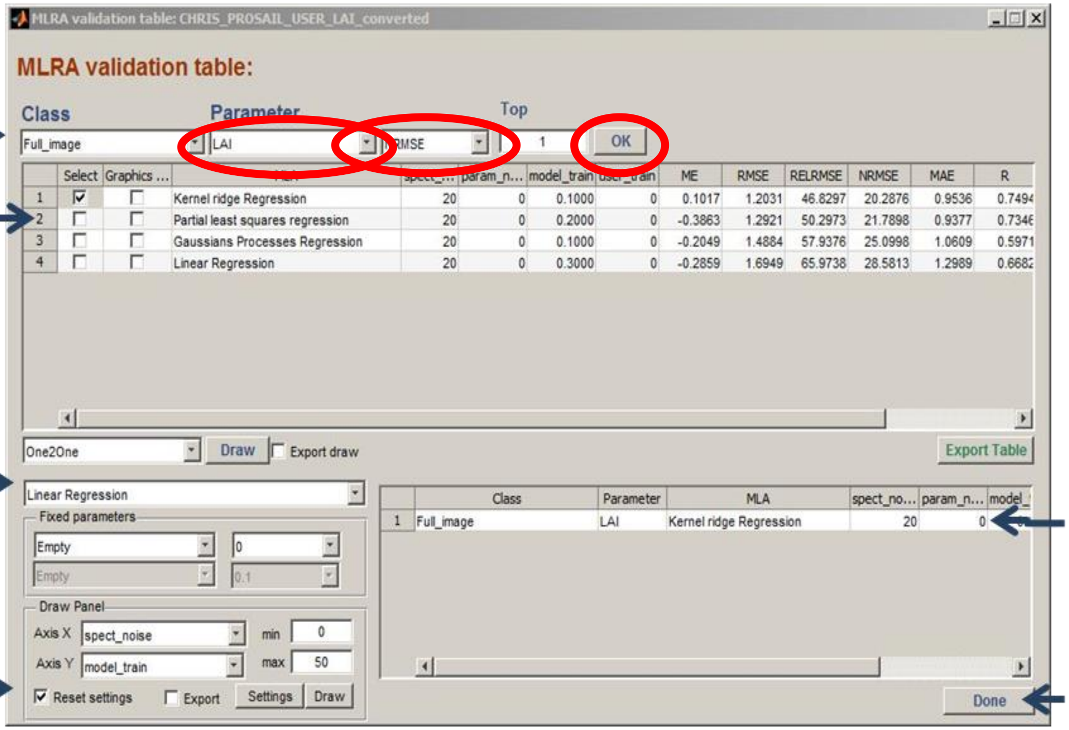
MLRA validation

Options to organize statistics per Class, Parameter, stat. and # of best results

Overview of best performing MLRAs. The check boxes allow to select a MLRA model for retrieval, the graphics for graphics outputs.

Graphic options: 1:1-line, sigmas (GPR), and 2D correlation matrices.

Options to control graphics properties.



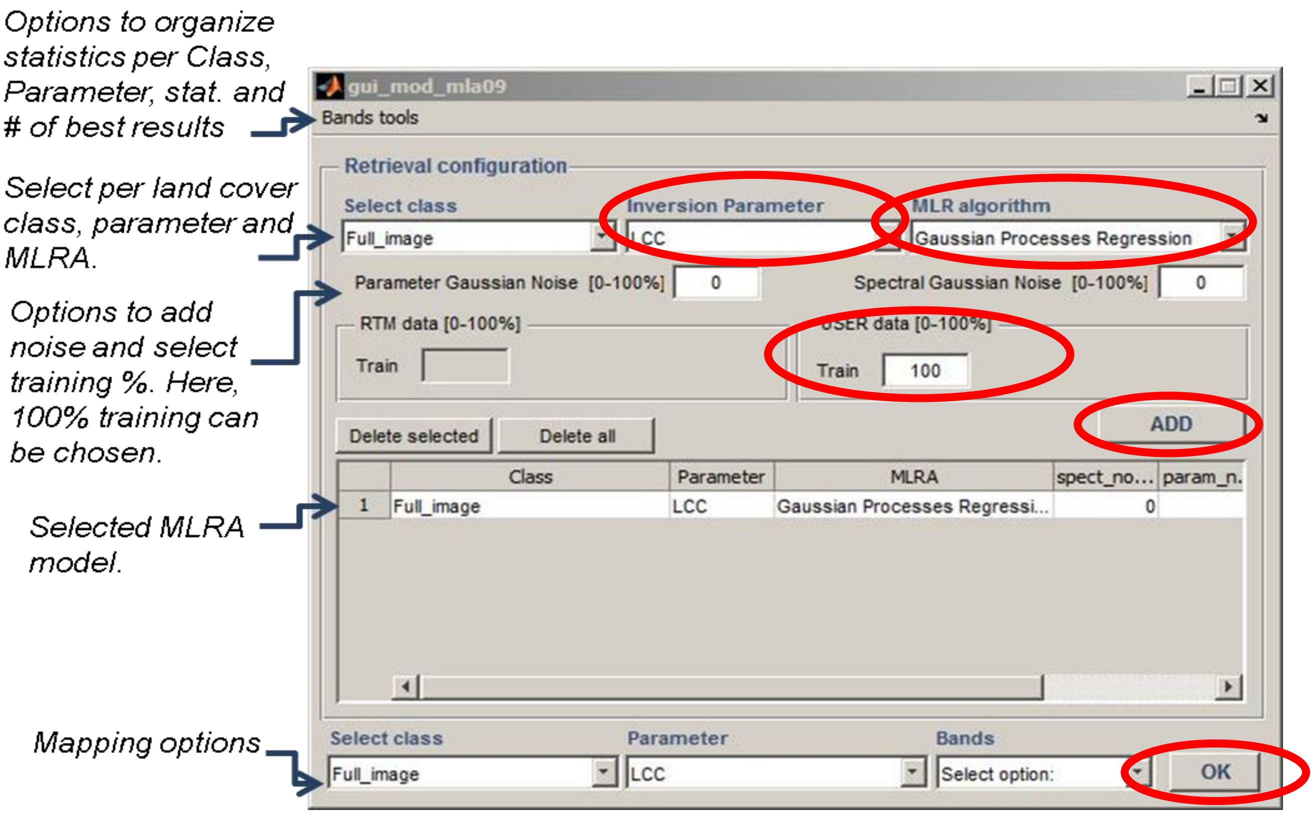
Option to export a table to a .txt file.

Selected MLRA model. With **Done** it will be moved to ***Retrieval.***

1. Choose how to **sort outputs,** according to parameter, statistic and number of top results per regression algorithm. Click on **OK**.
2. **Select** a **MLRA scenario for retrieval** (e.g. the top performing one). It will move to lower panel. When clicking on **Done** it will move to the Retrieval window (slide 11).
3. Select a MLRA scenario for **Graphics plottings: 1:1-line measured vs. predicted.** For GPR additional band relevance information will be provided. **Make sure to have User data loaded, because the selected model will be regenerated.**
4. In case ranges were introduced (noise, training/validation partitioning), validation results can be plotted in a **2D-matrix**. Results are plotted according to selected parameter and statistic.

Retrieval

Retrieval



Instead of going through the validation procedure, **one can also choose to immediately train a regression model and apply it to a remote sensing image**. User data has for training to be first inserted.

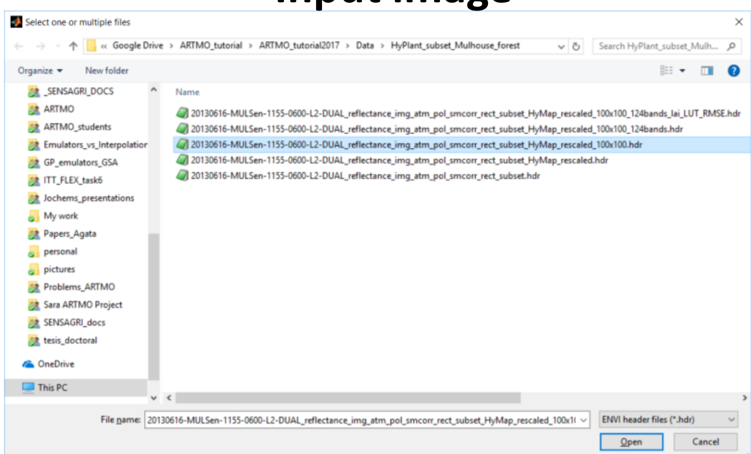
1. Select the **parameter** and **the regression model**.
2. Optionally **noise** can be added.
3. Select the **training partitioning**. Here 100% training data can be applied. The configuration need to be **ADD**ed and chosen model will appear in the down panel. **In case a model has been selected during the validation step, it will directly appear in that panel.**
4. Band tools options are provided (e.g. spectral subset, PCA).
5. When clicking on **OK**, the mapping procedure will start (see next slide).

Retrieval

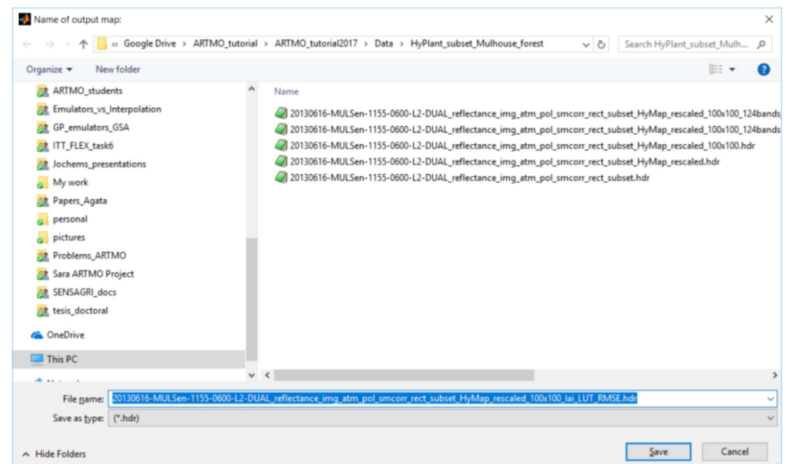
The mapping of selected biophysical parameter requires the following steps:

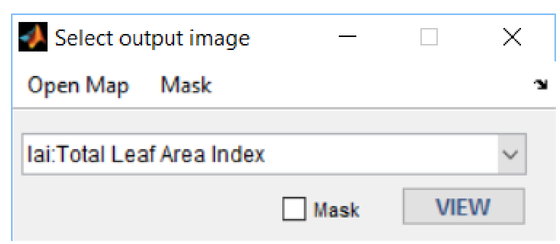
1. Select the directory with **Input images**
2. Images according to **TIFF or ENVI file format** (including .hdr file) will be identified and listed. **Multiple images can be selected**. They will be processed one-after-another.
3. Select the **Output directory**
4. When the processing is done, the output maps can be viewed. Select one through **Open Map** and click on **PREVIEW**.

Input image



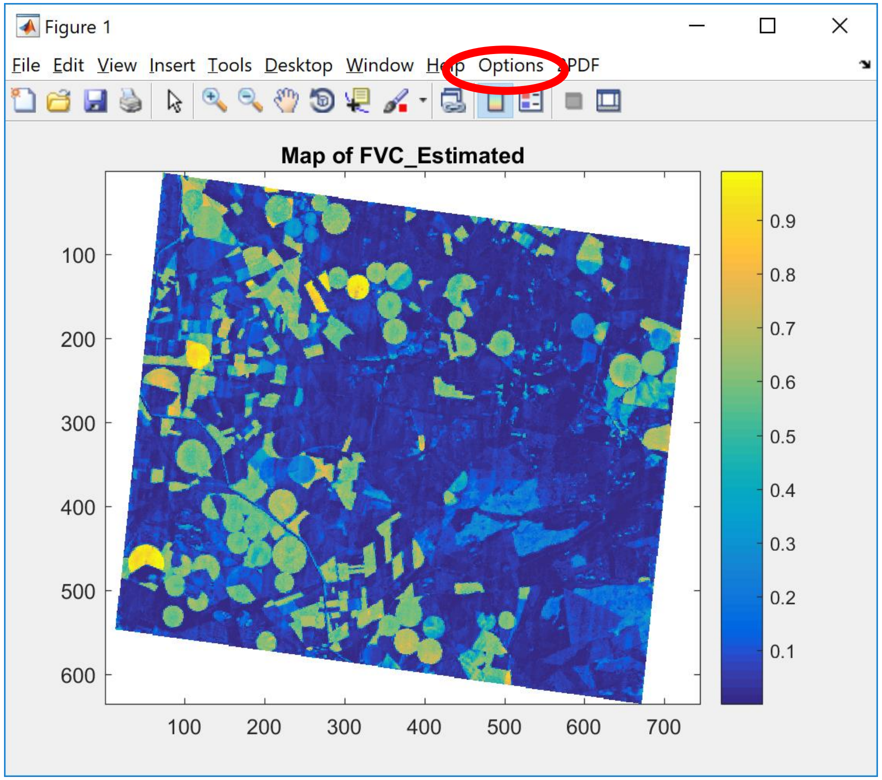
Output map

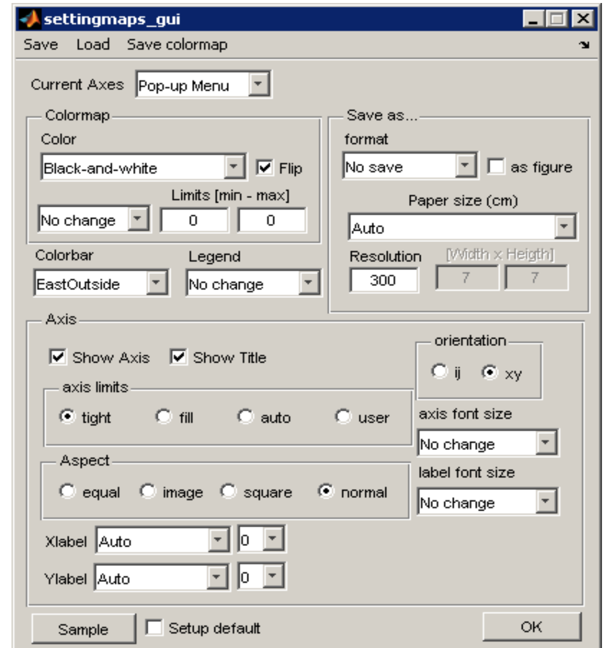




A drop-down list will show the provided output layers. One output map can then be peviewed.

Final maps

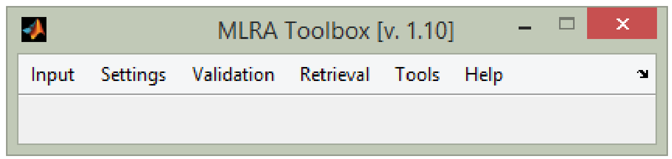


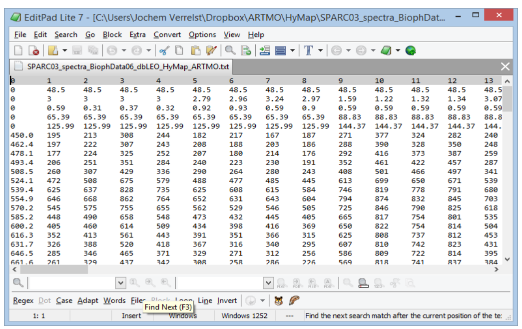


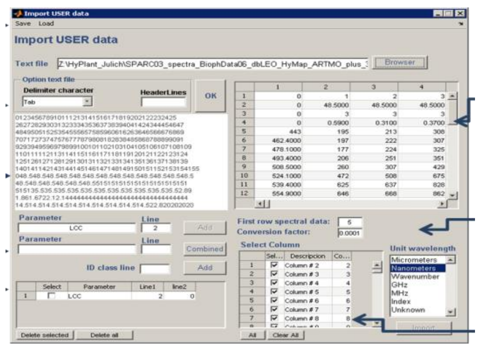
- Visualization of an output layer. In **Options**, map properties can be controlled (e.g. color scale, color table).
- Make sure to orient the map according to **ij** for **correct orientation**.
- The map can be **saved** according to various vector or bitmap formats. Redundant white space around the figure will be automatically removed.
- Settings can be set as **default** – will be automatically applied to subsequent maps.
- Click on **Sample** to visualize the map. Click on **OK** to save it away.

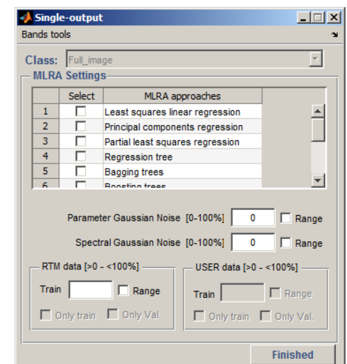
Exercise

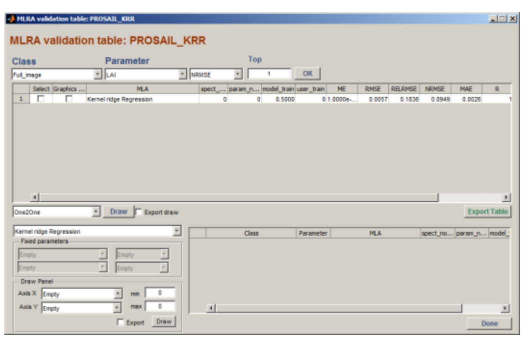
- ✓ Evaluate the performance of MLRAs using a field dataset and S2 spectral data.
- ✓ Apply the best performing regression algorithm to S2 images.

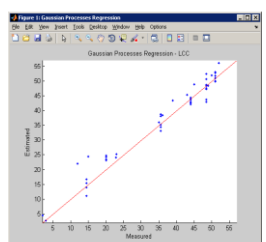


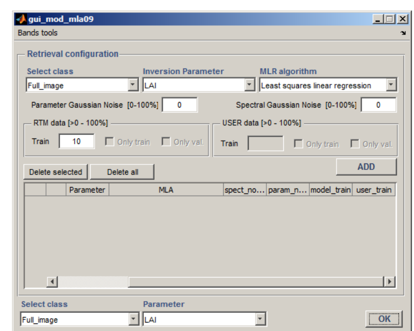


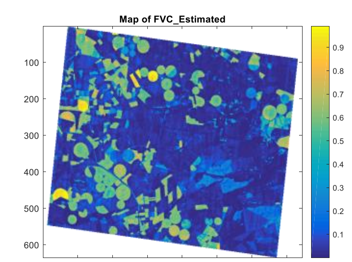




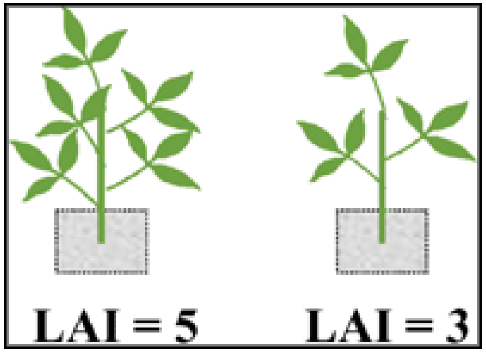






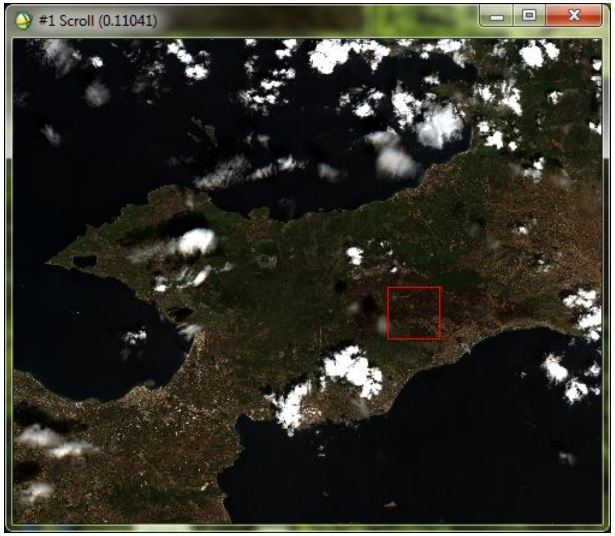


exercise

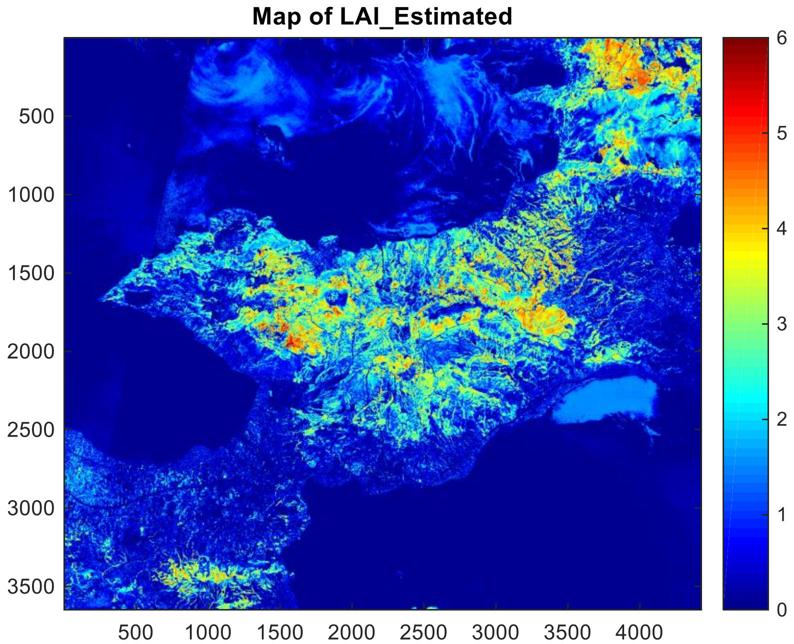


Subsets S2 before after Kineta fire

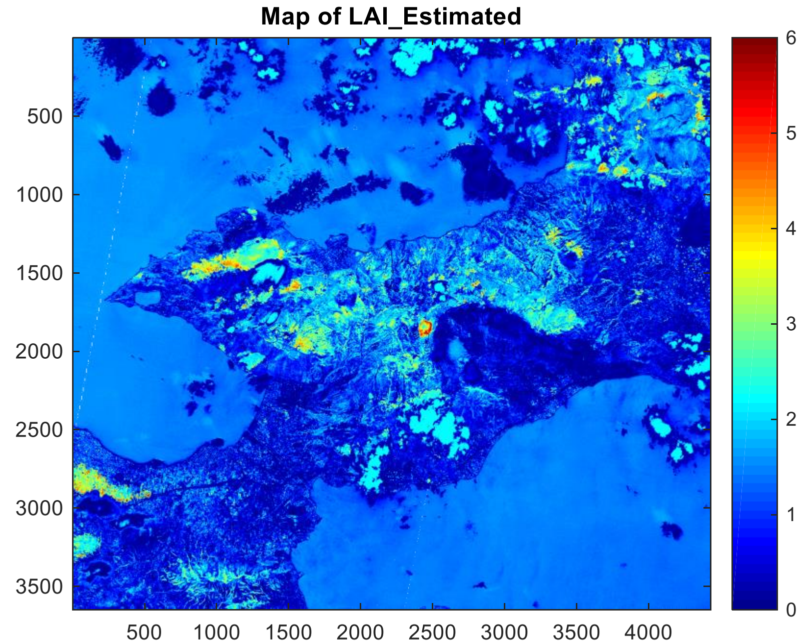
Map LAI before and after using a machine learning regression algorithm (MLRA)



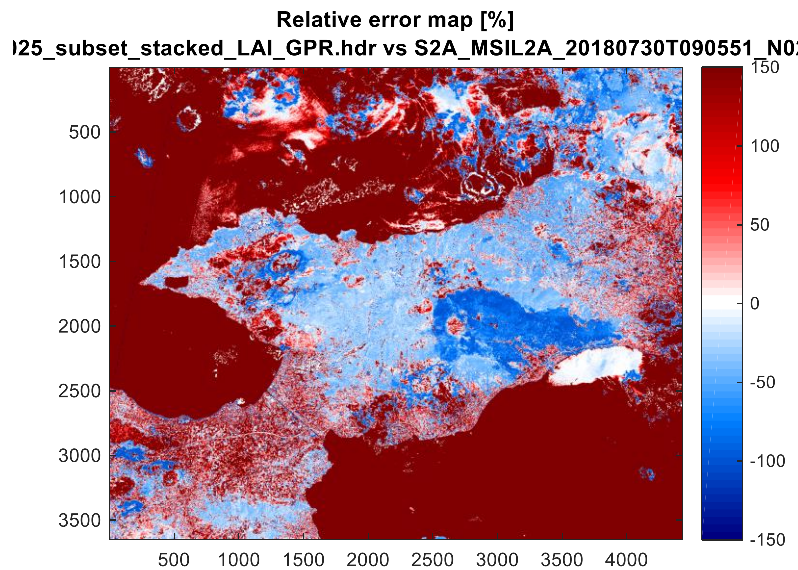
LAI before

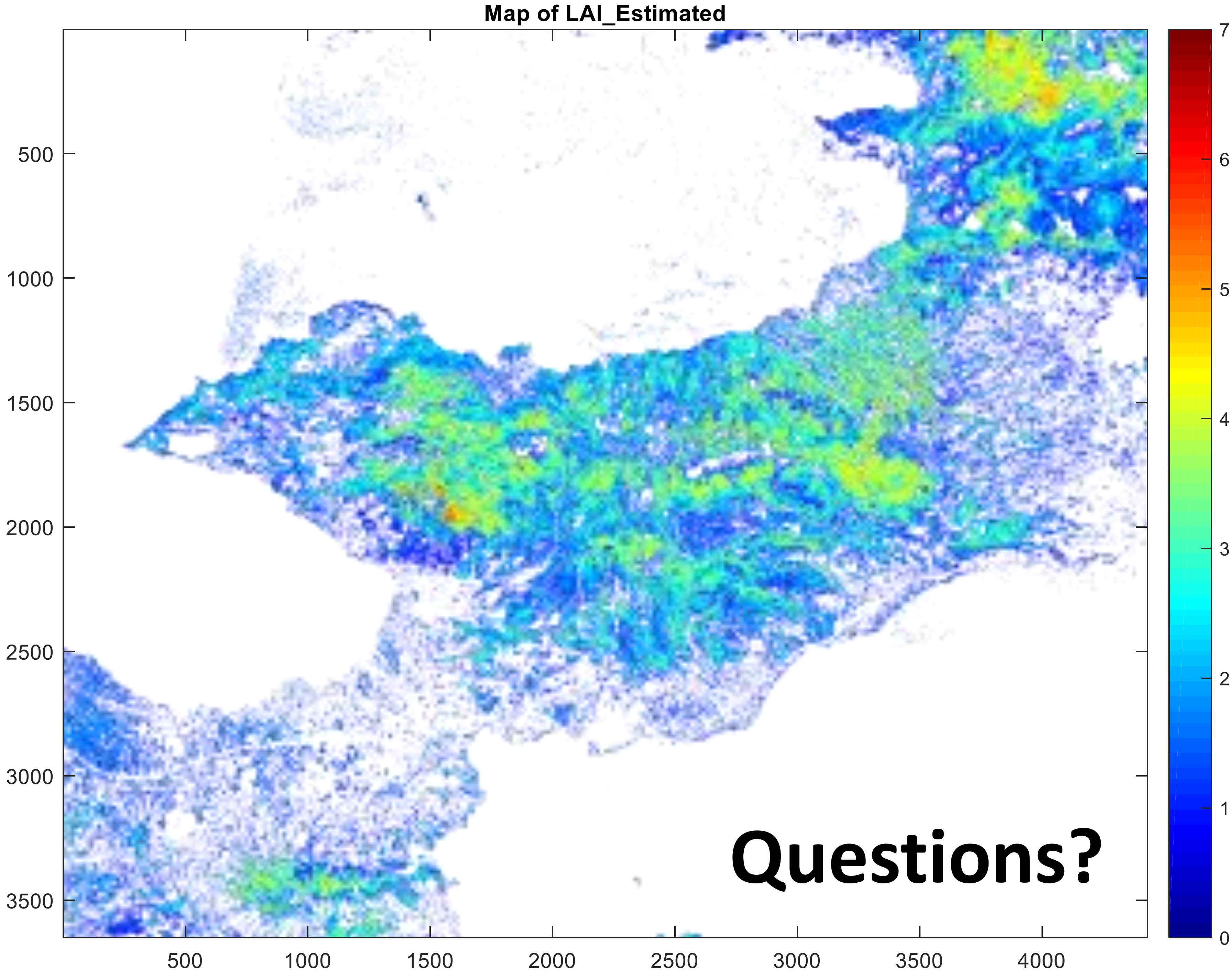


LAI after



Relative differences




Map of LAI_Estimated
500
1000
1500
2000
2500
3000
3500
4000
7
6
5
4
3
2
1
0

Questions?