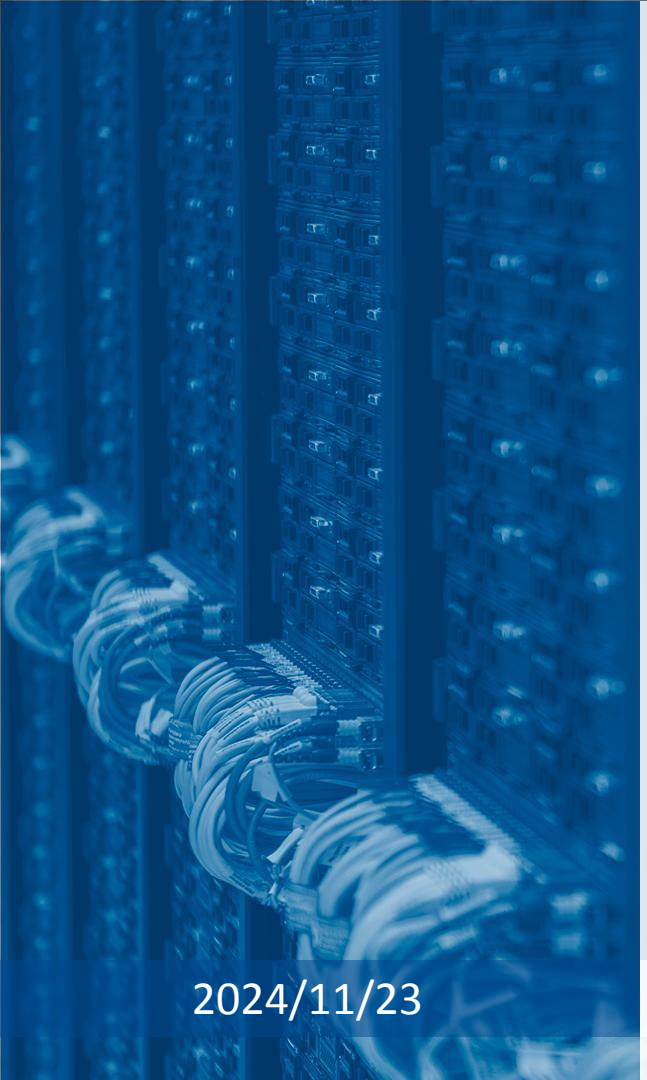


Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Downscaling historical Land Cover and Leaf Area Index data with machine learning

Amirpasha Mozaffari, Vinayak Huggannavar, Stefano Materia, Lina Teckentrup, Iria Ayan, Etienne Tourigny

2024/11/23

ACAM 2024

CERISE
(2023 - 2026)

High- resolution Land cover and Leaf Area Index
1925 - 2020

The CERISE project (grant agreement No101082139) is funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the Commission. Neither the European Union nor the granting authority can be held responsible for them.

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

https://visibleearth.nasa.gov/images/153312/marvelous-lenticu

About CERISE

The scope of CERISE (2023 - 2026) is to enhance the quality of the Copernicus Climate Change Service (C3S) reanalysis and seasonal forecast portfolio, with a focus on land-atmosphere coupling.

- Next generation of ECMWF reanalysis (ERA6 - Land)
- Seasonal Forecast (SEAS6)
- …

ECMWF · Barcelona Supercomputing Center · cmcc · Deutscher Wetterdienst · METEO FRANCE · Met Office · The Danish Meteorological Institute · Estellus · IPMA · Norwegian Meteorological Institute · SMHI · NILU

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

About CERISE

CERISE has the following key objectives:

- Develop unified multivariate land data assimilation systems for global and regional reanalysis.
- Improve coupled surface-atmosphere data assimilation and use of Earth system interface observations.
- Explore balanced land-atmosphere initialization for seasonal prediction, focusing on reforecasts and real-time conditions.
- **Extend slow-varying surface variables (LC, LAI, lake cover) data back to 1925 for reanalysis.**
- Deliver prototype coupled Earth system and land surface reanalysis datasets at global and regional scales.
- Provide multi-system seasonal forecast da[illegible]s with innovative land initialization.
- Develop diagnostic methods to assess reana[illegible]sis prototypes and seasonal forecasts.
- Offer recommendations for Copernicus Climate [illegible]hange [illegible]implem[illegible]tation.

Some terminology

Land Use (LU):

Human activities or management practices that alter or maintain land, such as agriculture, forestry, or urban development.

Land Cover (LC):

The physical surface characteristics of Earth's land, including vegetation, water bodies, and urban areas.

Leaf Area Index (LAI):

A dimensionless measure of leaf canopy coverage, representing the total leaf area per unit ground surface area.

Datasets

ESA CCI Land Cover yearly maps 300m resolution from 1992 - 2020 that includes 38 classes of LC in discrete values (0 - 220)

CONFESS LAI 5-days maps at 1km resolution from 1980 - 2020 that has continuous values between (0 - 6)

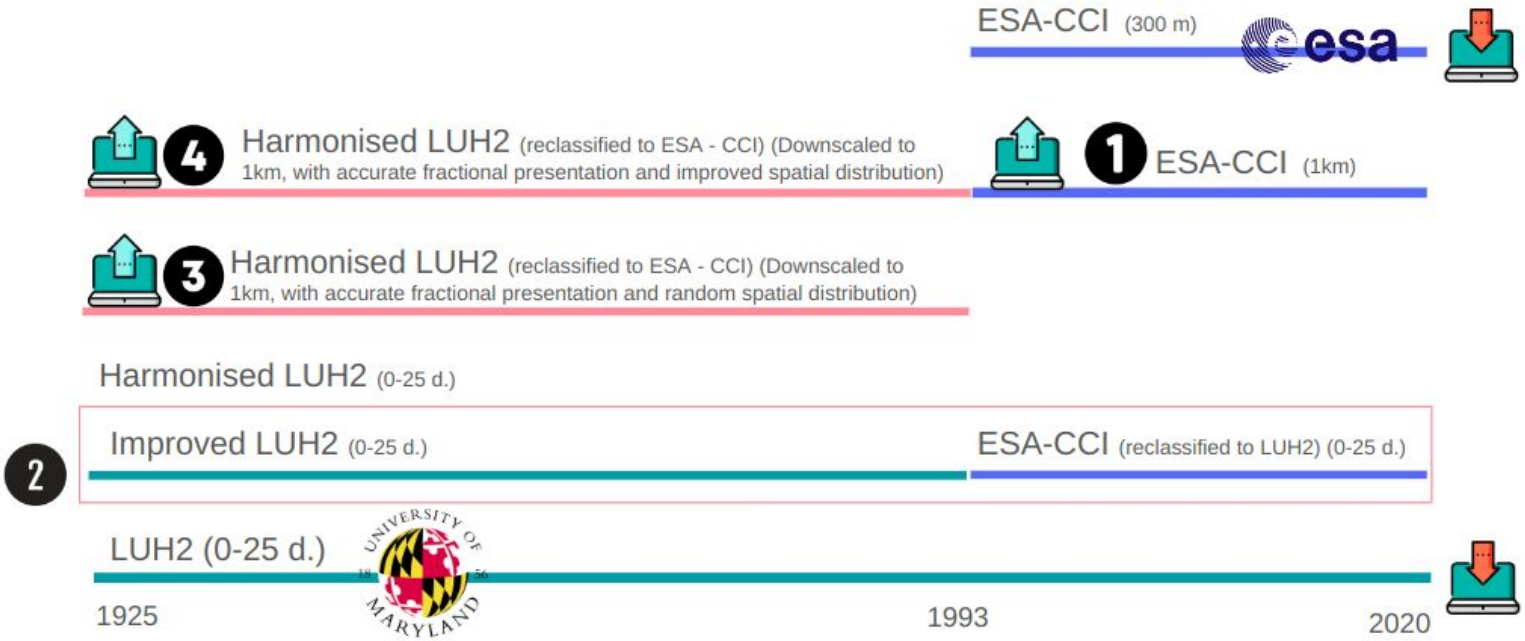
LUH2f Land Use yearly maps 0.25 x 0.25 degrees resolution from 950 - 2100 that has fractional values for 8 classes in continuous values (0 -1)

HILDA+ Land use yearly maps at 1km resolution from 1899 - 2020 that has 7 classes in discrete values (0 -77)

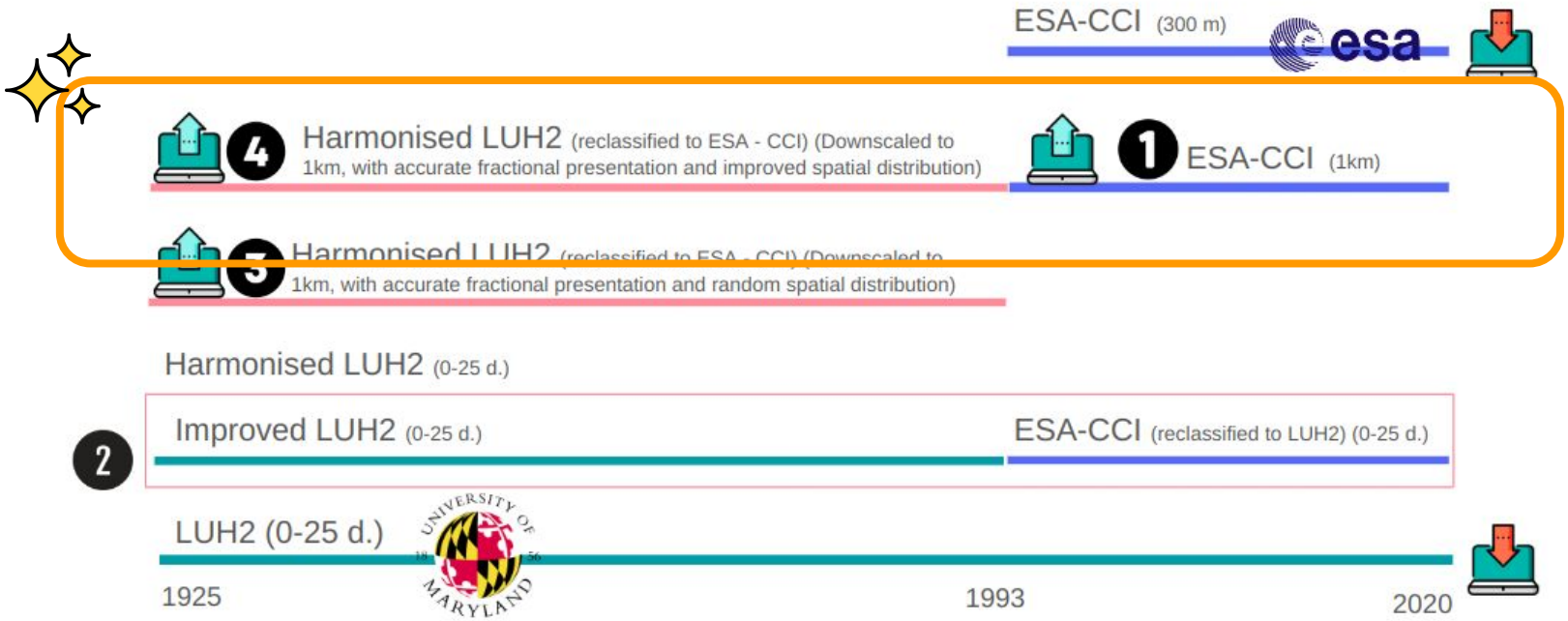
NEW

BSC
Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Original work plan



~~Original~~ ML work plan



Preprocessing

Upscaling ESA CCI Land Cover to 1km resolution aligned with CONFESS LAI grid

Downscale LUH2f to 1km resolution using nearest neighbor, aligned with CONFESS LAI

Remapping HILDA+ to same 1km CONFESS LAI grid.

All the three dataset are in the same grid and resolution of the target data.

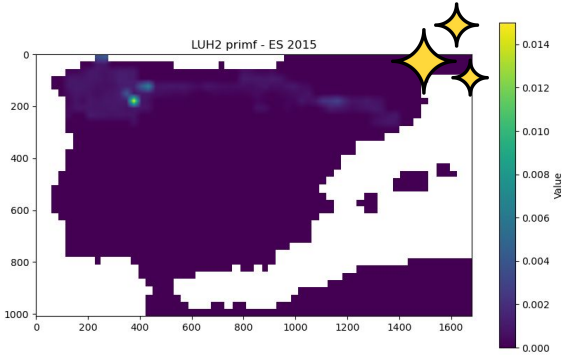
Early results 1

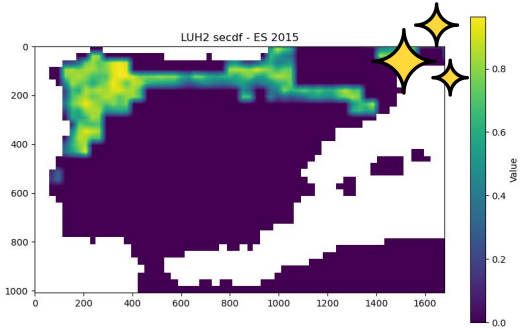
Trained on 5 years (2010 - 2014) , tested on 1 year (2015)

On a moving window of 1200 km x 1000 km

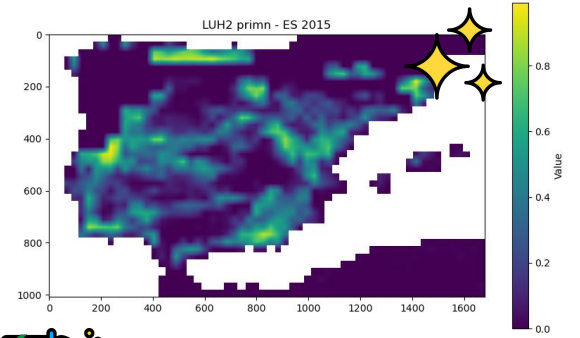
- LUH2f, 14 fractional features
- HILDA+ , 1 LU state features
- Target: LC state (38 classes)

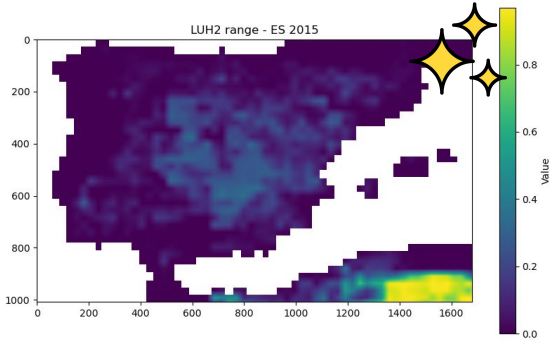
Random Forest Classifier (200 estimator)

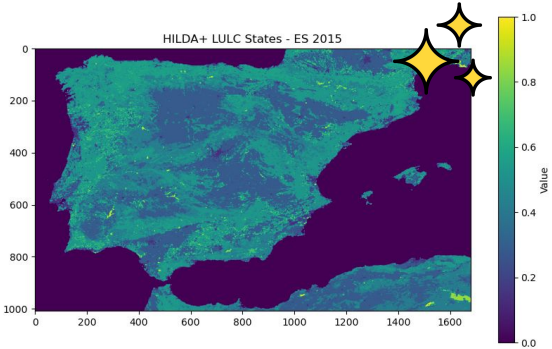


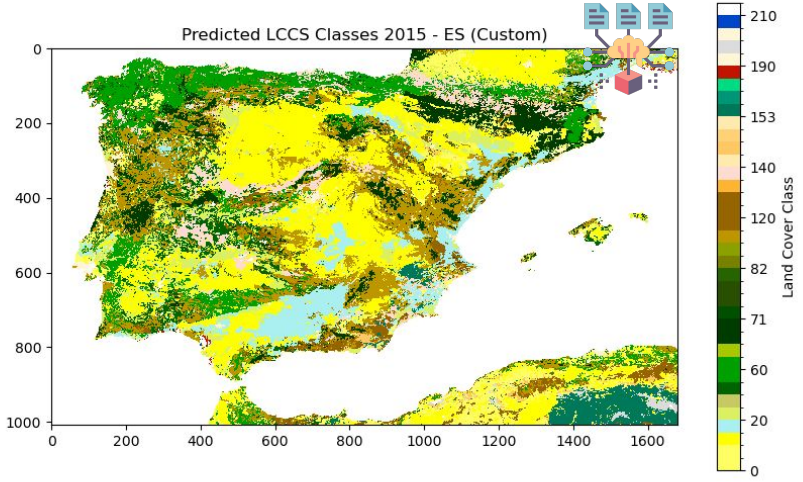


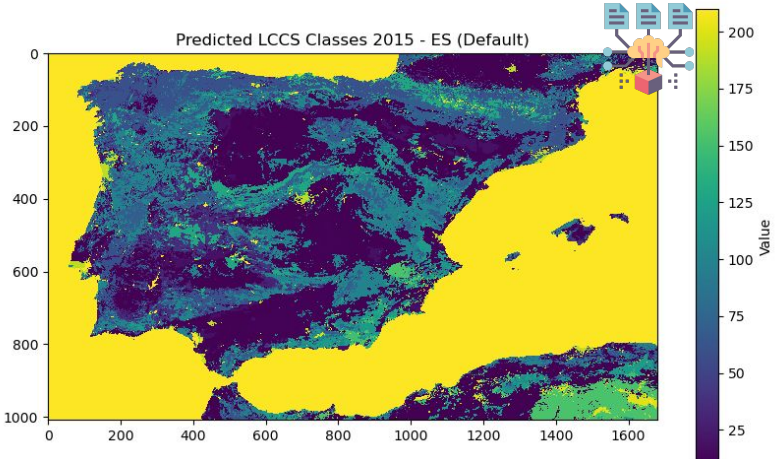
These are examples of the model features that are fed to the LC model.

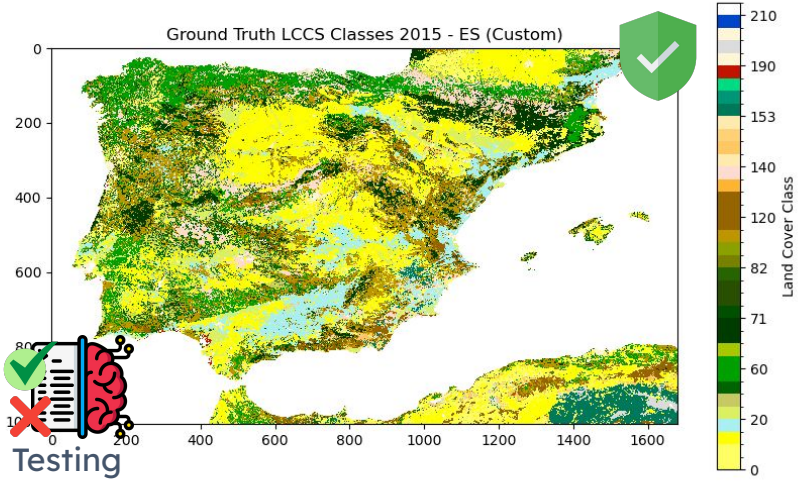


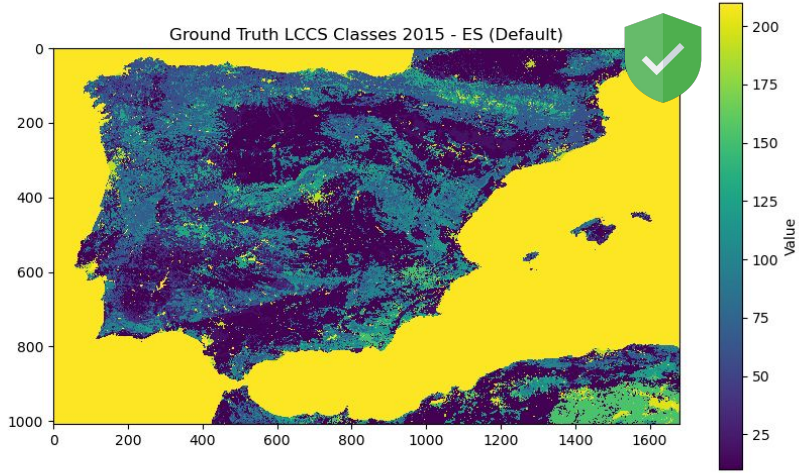












Testing

These are predicted and ground truth LC for test year 2015 with average 93% accuracy across classes.

Early results 1

Trained on 5 years (2010 - 2014) , tested on 1 year (2015)

On a moving window of 1200 km x 1000 km

- LUH2f, 14 fractional features
- HILDA+ , 1 LU state features -> ESA CCI 300m is used to train it!
- Target: LC state (38 classes)

Random Forest Classifier (200 estimator)

Early results 2

Trained on 20 years (1992 - 2015) , tested on 3 year (1993, 2004, 2013)

On a moving window of 400 km x 400 km

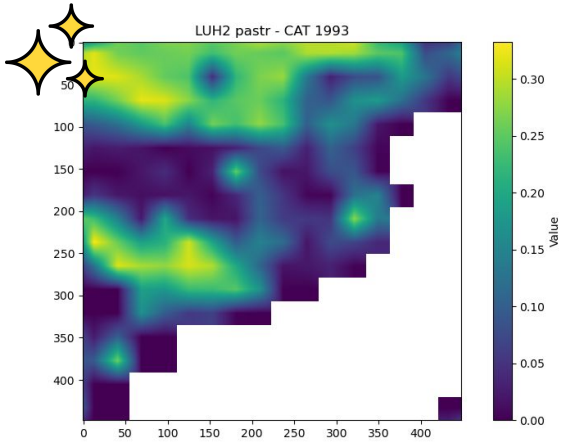
- LUH2f, 14 fractional features
- HILDA+ in-land water mask (1 feature)
- Target: LC state (38 classes)

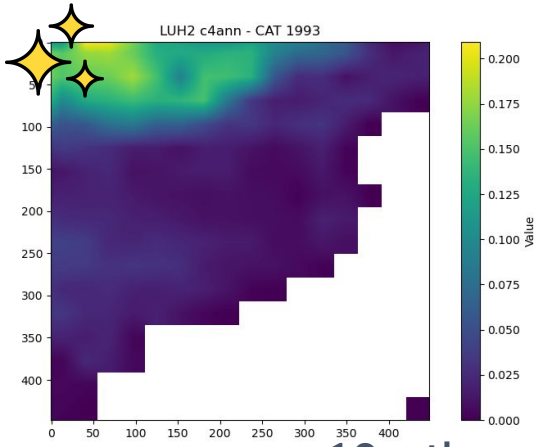
XGboost Classifier (2000 estimator)

No autoregressive feature as we are exploring the viability of the data that we rely on.

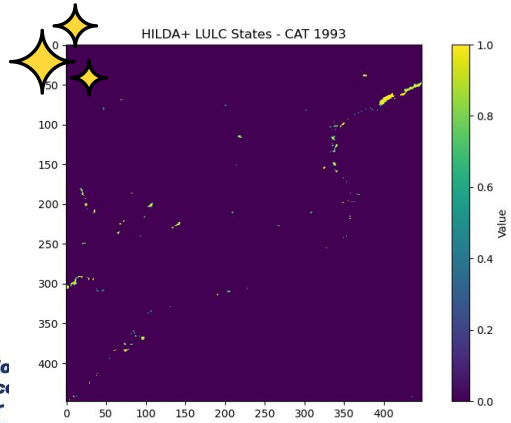
XGBoost

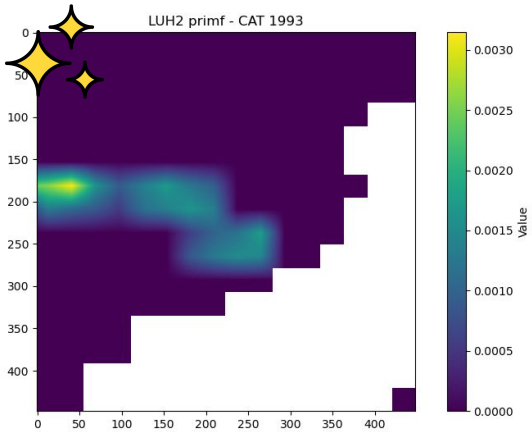
These are examples of the model features that are fed to the LC model.

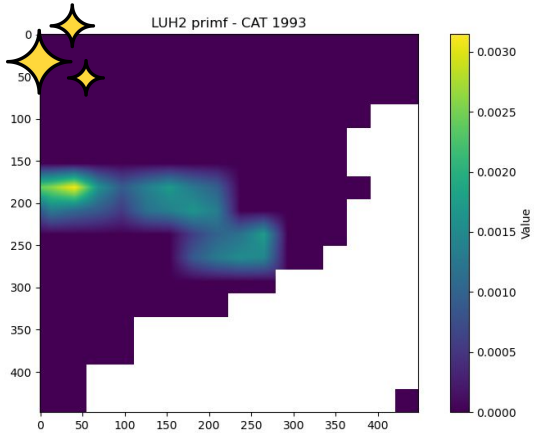




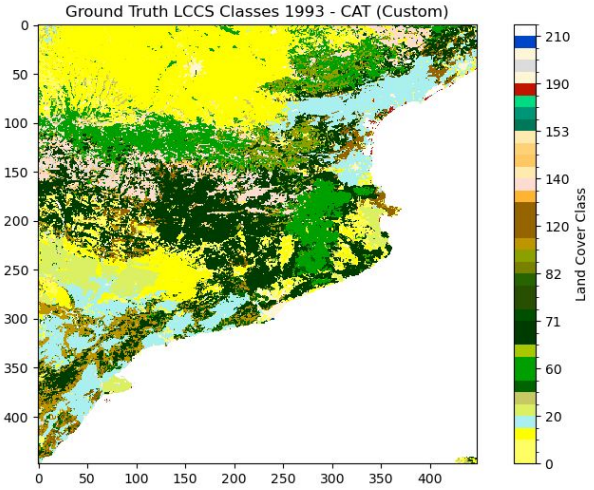
+ 10 others

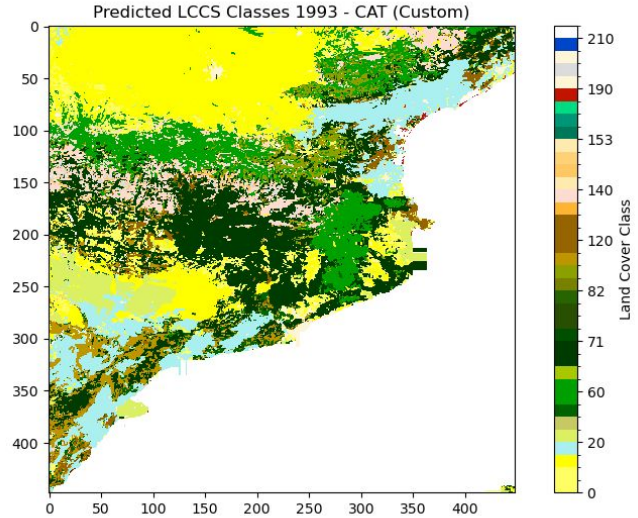




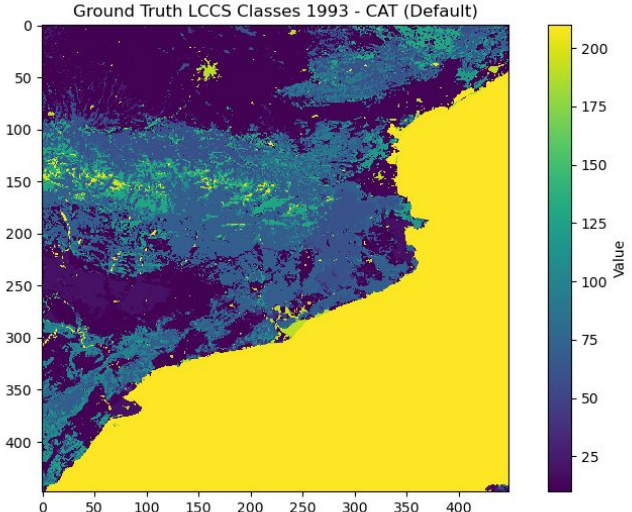


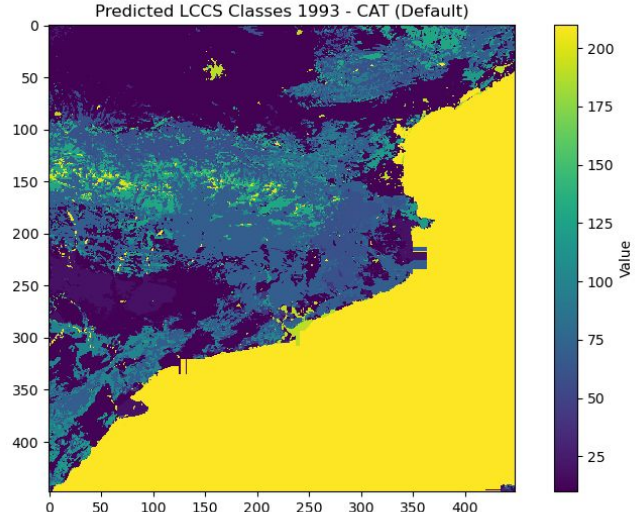
Testing

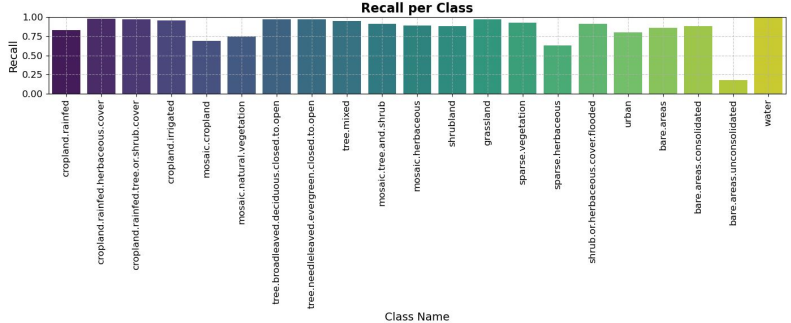


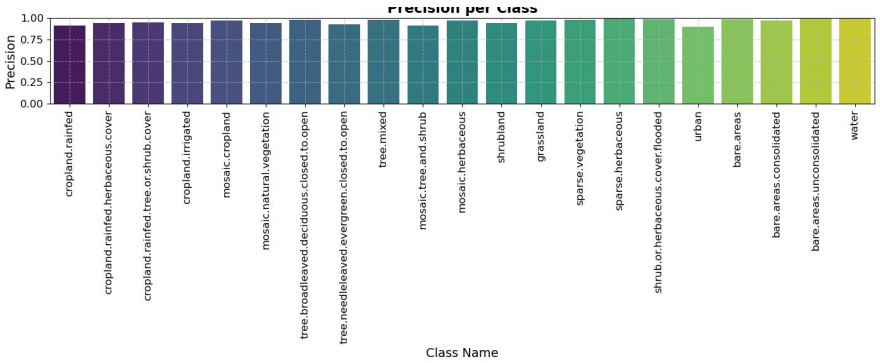


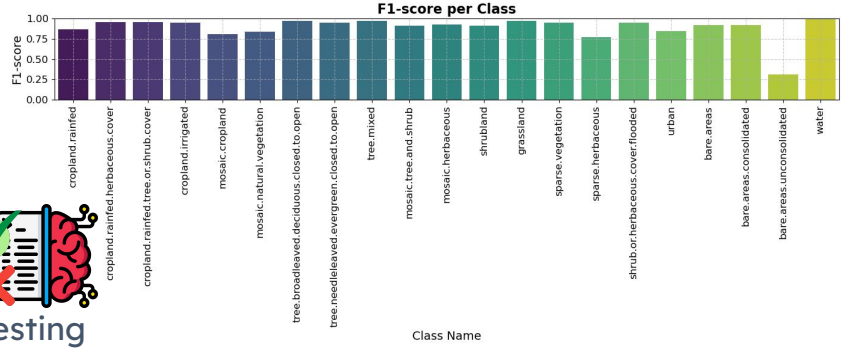
Predicted and ground truth LC for test year 1993.

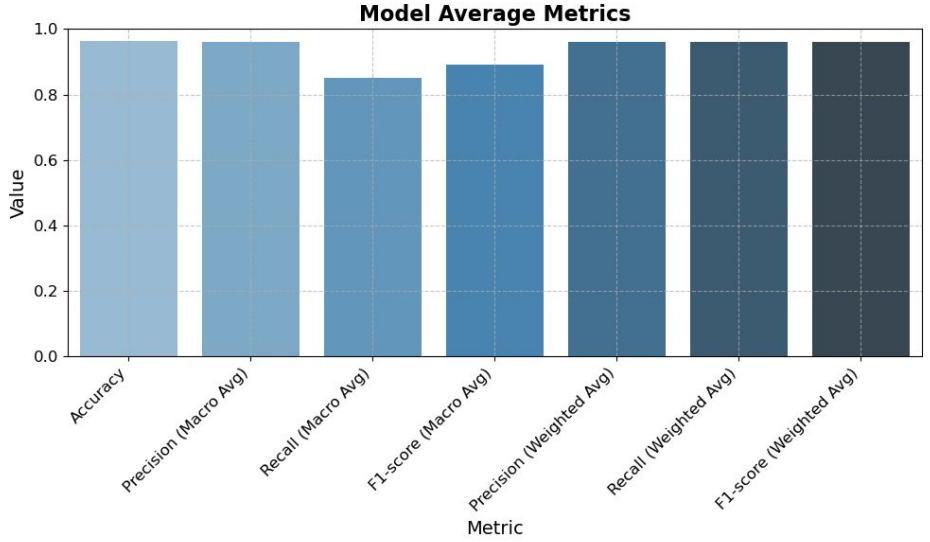


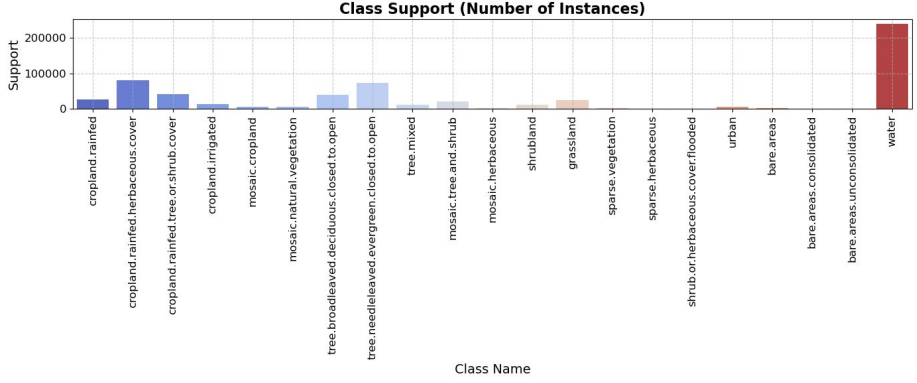



Recall per Class
Recall
1.00
0.75
0.50
0.25
0.00
cropland.rainfed
cropland.rainfed.herbaceous.cover
cropland.rainfed.tree.or.shrub.cover
cropland.irrigated
mosaic.cropland
mosaic.natural.vegetation
tree.broadleaved.deciduous.closed.to.open
tree.needleleaved.evergreen.closed.to.open
tree.mixed
mosaic.tree.and.shrub
mosaic.herbaceous
shrubland
grassland
sparse.vegetation
sparse.herbaceous
shrub.or.herbaceous.cover.flooded
urban
bare.areas
bare.areas.consolidated
bare.areas.unconsolidated
water
Class Name


Precision per Class
Precision
1.00
0.75
0.50
0.25
0.00
cropland.rainfed
cropland.rainfed.herbaceous.cover
cropland.rainfed.tree.or.shrub.cover
cropland.irrigated
mosaic.cropland
mosaic.natural.vegetation
tree.broadleaved.deciduous.closed.to.open
tree.needleleaved.evergreen.closed.to.open
tree.mixed
mosaic.tree.and.shrub
mosaic.herbaceous
shrubland
grassland
sparse.vegetation
sparse.herbaceous
shrub.or.herbaceous.cover.flooded
urban
bare.areas
bare.areas.consolidated
bare.areas.unconsolidated
water
Class Name


F1-score per Class
F1-score
1.00
0.75
0.50
0.25
0.00
cropland.rainfed
cropland.rainfed.herbaceous.cover
cropland.rainfed.tree.or.shrub.cover
cropland.irrigated
mosaic.cropland
mosaic.natural.vegetation
tree.broadleaved.deciduous.closed.to.open
tree.needleleaved.evergreen.closed.to.open
tree.mixed
mosaic.tree.and.shrub
mosaic.herbaceous
shrubland
grassland
sparse.vegetation
sparse.herbaceous
shrub.or.herbaceous.cover.flooded
urban
bare.areas
bare.areas.consolidated
bare.areas.unconsolidated
water
Class Name


Model Average Metrics
Value
1.0
0.8
0.6
0.4
0.2
0.0
Accuracy
Precision (Macro Avg)
Recall (Macro Avg)
F1-score (Macro Avg)
Precision (Weighted Avg)
Recall (Weighted Avg)
F1-score (Weighted Avg)
Metric


Class Support (Number of Instances)
Support
200000
100000
0
cropland.rainfed
cropland.rainfed.herbaceous.cover
cropland.rainfed.tree.or.shrub.cover
cropland.irrigated
mosaic.cropland
mosaic.natural.vegetation
tree.broadleaved.deciduous.closed.to.open
tree.needleleaved.evergreen.closed.to.open
tree.mixed
mosaic.tree.and.shrub
mosaic.herbaceous
shrubland
grassland
sparse.vegetation
sparse.herbaceous
shrub.or.herbaceous.cover.flooded
urban
bare.areas
bare.areas.consolidated
bare.areas.unconsolidated
water
Class Name

Testing

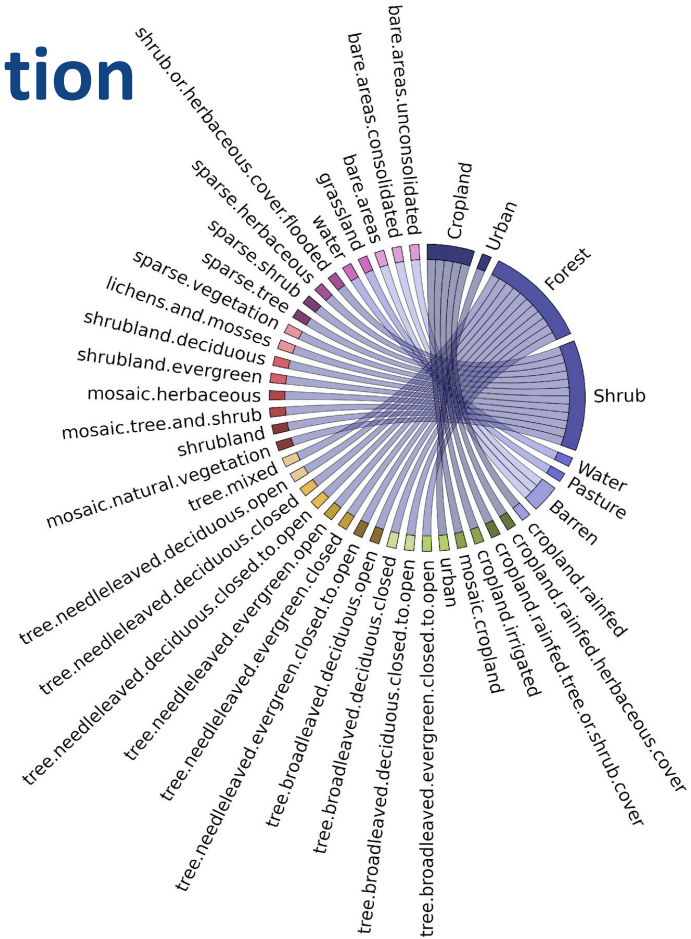
Inference validation

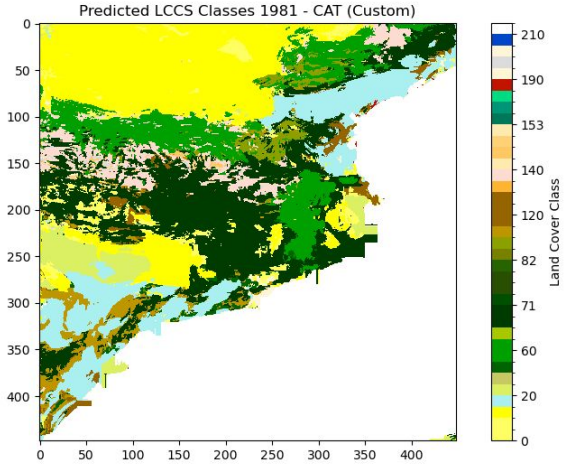
Basic classes (global comparison)

A cross-walk translation of LC to basic LU that enable us to verify the inference results 1925 - 1991 with HILDA+ LU dataset.

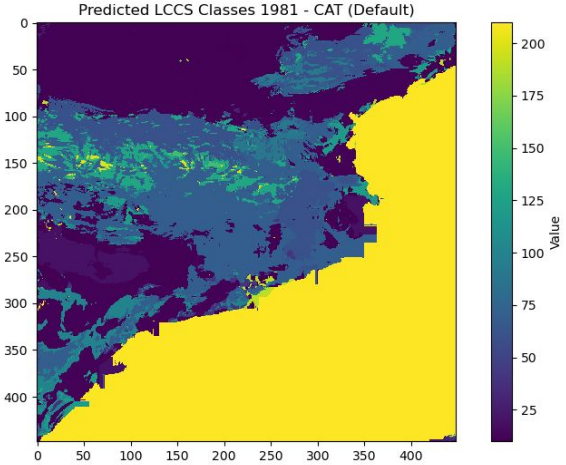
Kernel density (point comparison)

Autoregressive computed basic classes in single points using the LUH2f transition data starting from 1992 and going backward.





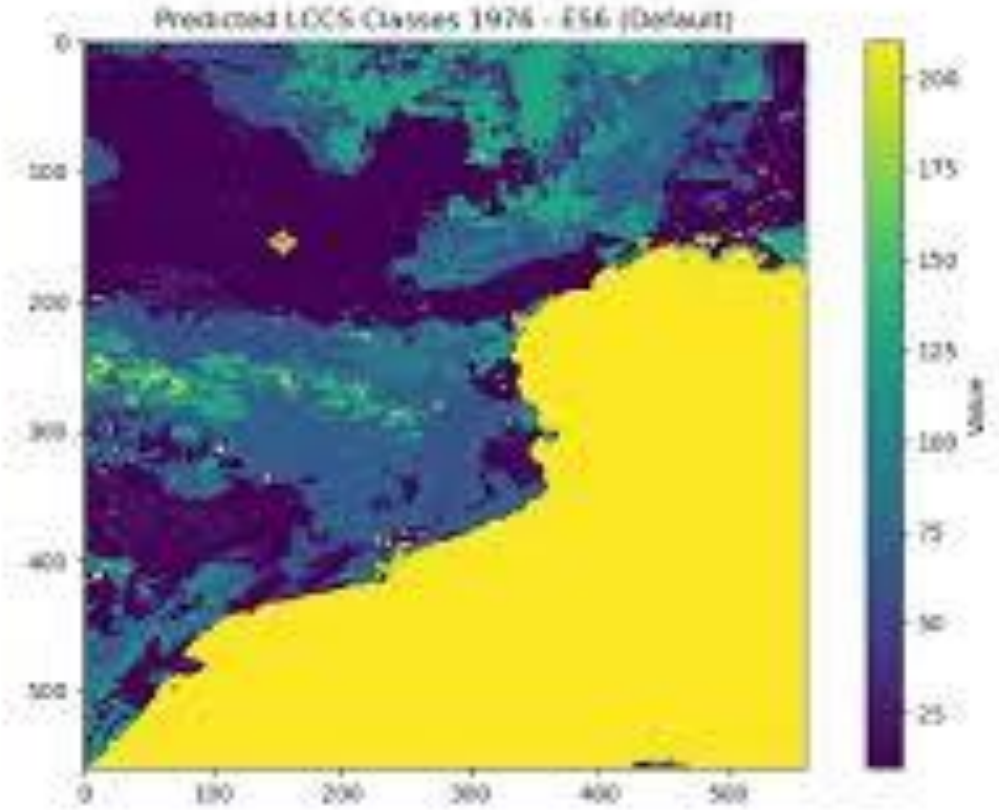
Reserve for comparison with HILDA+ basic classes



Reserve for comparison with HILDA+ basic classes

These are inference for year 1981, but we are missing metrics on performance for now.

Inference



Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Inference 1925 - 1991 Land cover based on LUH2f with water-mask

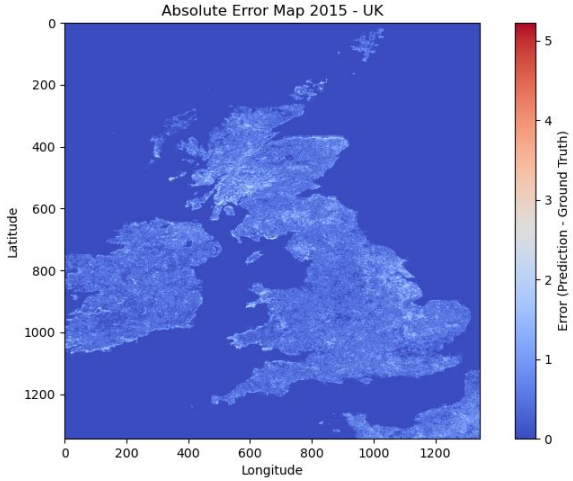
Early results 3

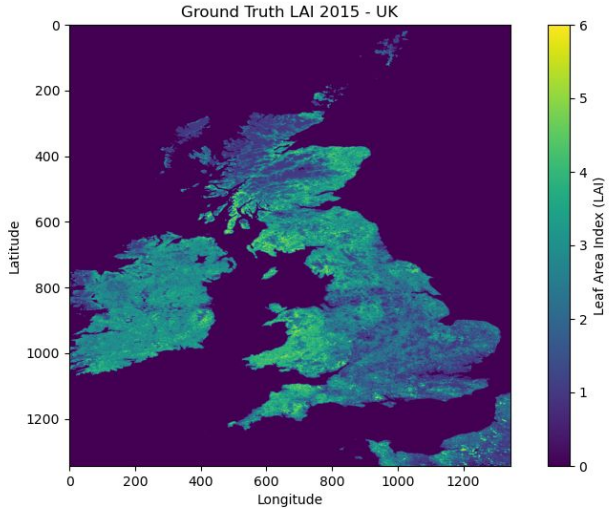
Trained on 5 years (2010 - 2014) , tested on 1 year (2015)

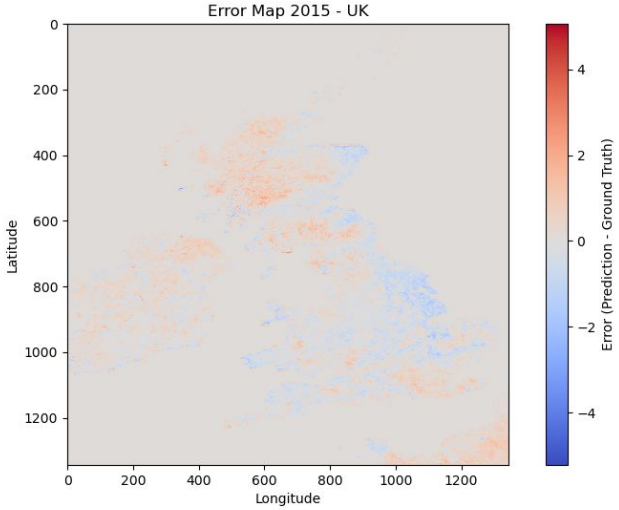
On a moving window of 1200 km x 1000 km

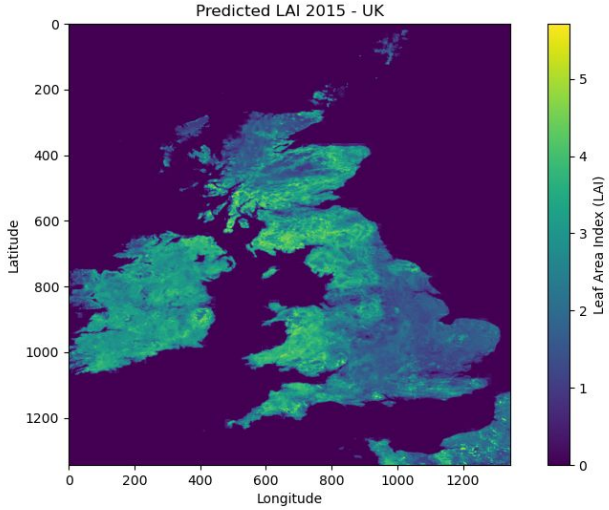
- LUH2f, 14 fractional features
- HILDA+ land use class (1 feature)
- Target: CONFESS LAI (continuous value)

Random Forest Regressor (200 estimator)









These are ground truth and predicted LAI for The British Isles for test year 2015 for LAI, and the error maps, with R^2 Score of 0.894 and 89% accuracy.

Early results 4

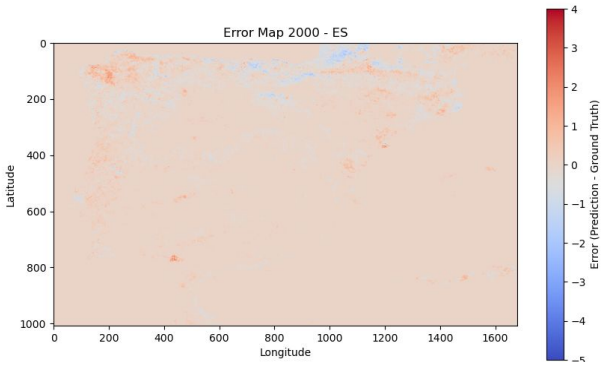
Trained on 13 years (2001 - 2014) , tested on 1 year (2000)

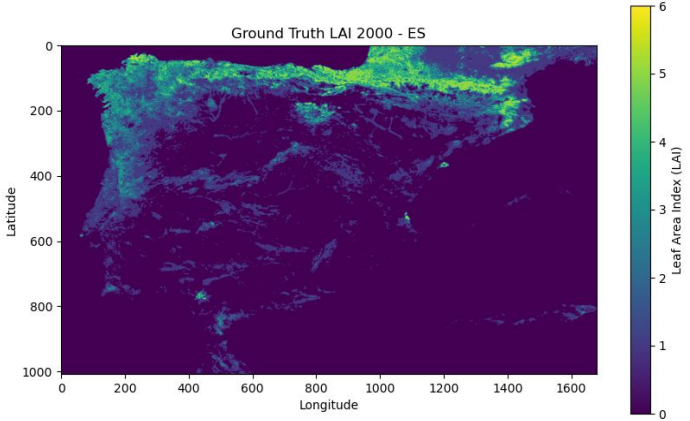
On a moving window of 1200 km x 1000 km

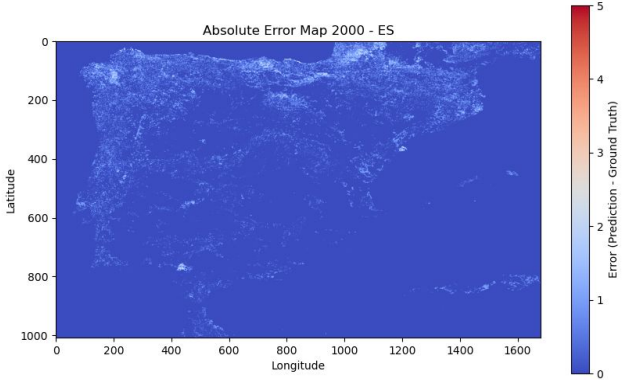
- LUH2f, 14 fractional features
- HILDA+ land use class (1 feature)
- Target: CONFESS LAI (continuous value)
- Last known state of LAI (this case year 2001)

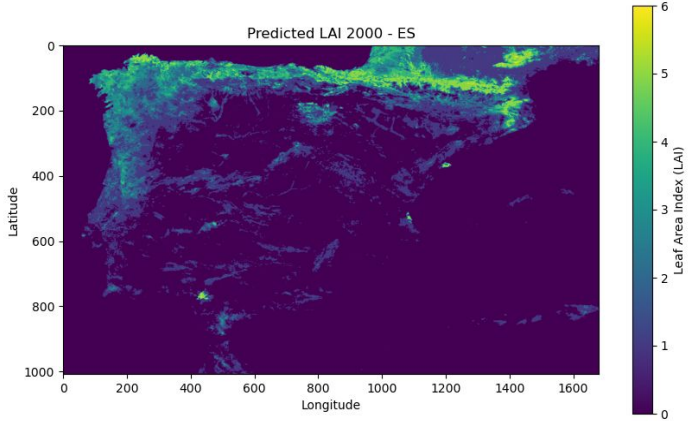
XGboost running on GPU (2000 estimator)

XGBoost









These are ground truth and predicted LAI for Iberian Peninsula for test year 2000 for LAI, and the error maps, with R^2 Score of 0.90 and 94% accuracy.

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Conclusion so far

- Low resolution data LU seems to have enough information to enable a sufficient reconstruction, but fails to provide enough to reduce the search domain that can consistently resolve without knowing the spatio-temporal relationship. -> It is necessary to pass the temporal relationship & last known state to model for autoregressive reconstruction.
- Simple ML methods (RF, XGB) seems to have a good shoot on regional reconstruction of LC & LAI
- To exploit the data that we have, we have to move to GPU and out of core memory loading as these data are big.

To-do(s):

- Developing secondary network to imitate the yearly cycle of LAI variation
- Developing the out of the core data loading that allows us to handle big dataset and train on model .
- Developing the verification pipeline that we can estimate the accuracy of predictions for inference years versus published LU and numerical results.
- Validate and analysis of using the inference data as inputs for IFS model.
- Integrating the time-varying lake cover into LC datasets.

https://www.visibleearth.nasa.gov/images/7421/barcelona-spain/7421t

CONCERTO

(2025 - 2028)

NEXT

TerraDT

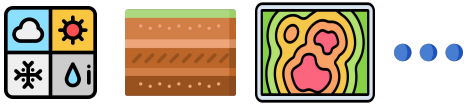
Digital Twin of Earth system for Cryosphere, Land surface and related interactions

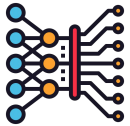
(2025 - 2028)

Accurate High- resolution Land Use Land cover and Leaf Area Index and their emulators 1850 - 2100

What will be new?

More relevant data! Climate and weather, topography, soil type, anthropogenic footprint,…





More advance and deep ML architecture to exploit all these data,

And building emulators for LC, LU & LAI for IFS & ICON

Thank you all!

amirpasha.mozaffari@bsc.es

https://www.visibleearth.nasa.gov/images/151257/shrinking-reservoirs-in-catalonia/151260l

Earth Sciences Department

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

COORDINATED BY ECMWF

Co-funded by the European Union

The CERISE project (grant agreement No101082139) is funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the Commission. Neither the European Union nor the granting authority can be held responsible for them.