Physiography improvements in numerical weather prediction digital twin engines

Thomas Rieutord, Geoffrey Bessardon, Emily Gleeson

thomas.rieutord@met.ie
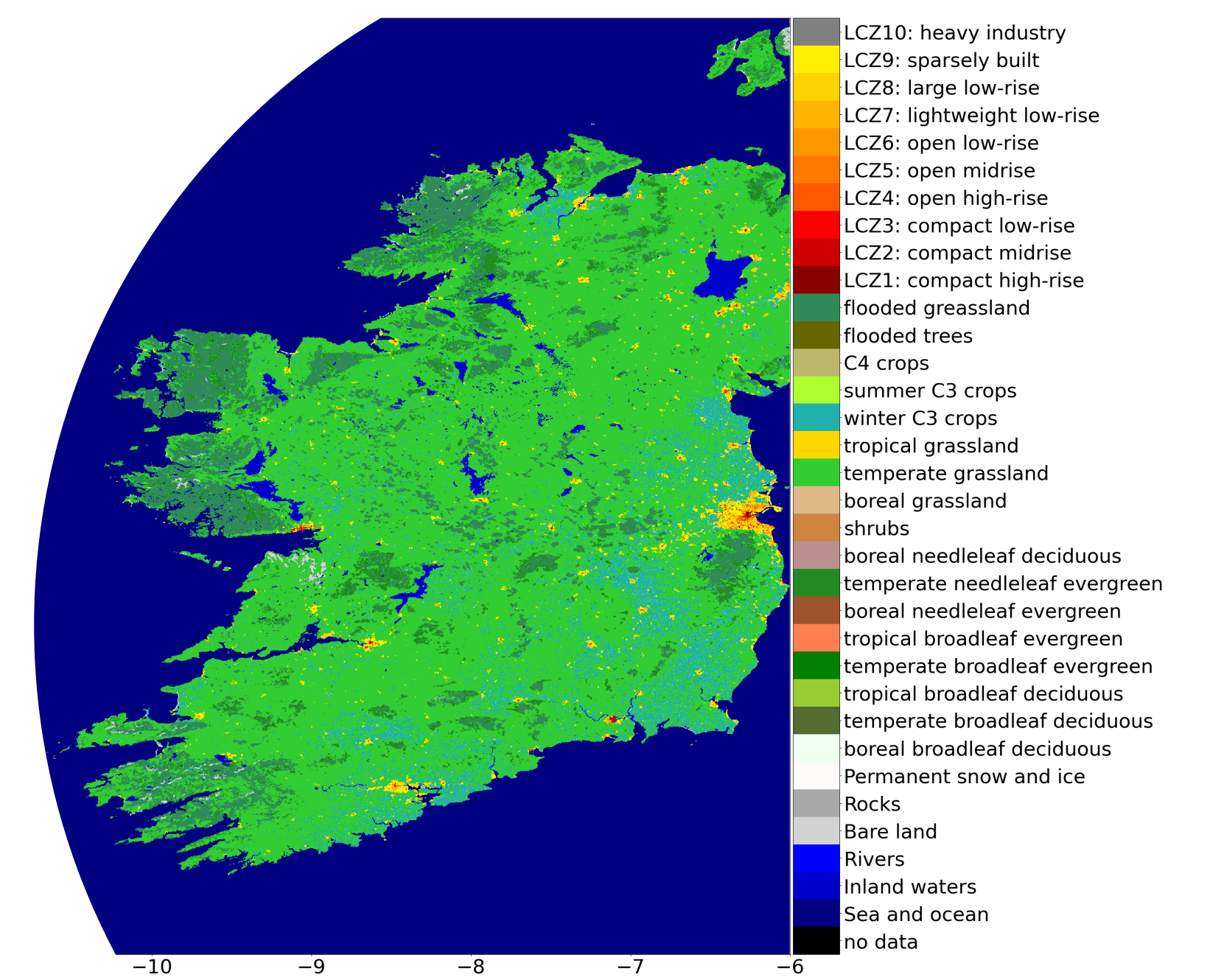
Poster displayed at the EGU General Assembly Vienna, 24 April 2023

Sharing is encouraged

Met Éireann

Funded by the European Union — Destination Earth — implemented by ECMWF · eesa · EUMETSAT

Introduction

Land cover has a large impact on atmosphere (fluxes of heat, moisture, momentum...).

Weather prediction models for local scale are now at >1km resolution and use land cover at 300m resolution.

Next generation of weather models will be too close to 300m.

Thus, we leverage AI and open data to produce land cover for weather forecast
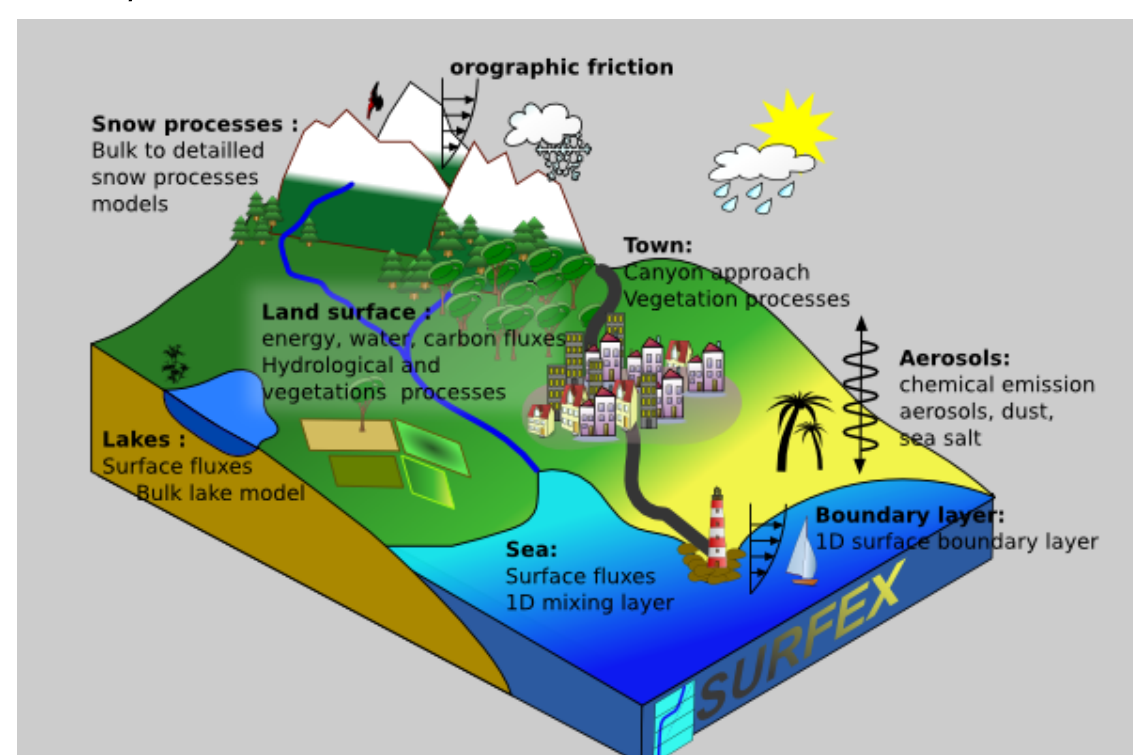
Machine learning learning brings large improvement when enough data is available.

Many data are freely accessible to study the land cover (high resolution maps, satellite images...)

But, nothing cover the specifics needs of weather prediction models

Abstract

The next generation of numerical weather prediction models (so-called digital twin engines) will reach hectometric scale, for which the existing physiography databases are insufficient. Our work leverages machine learning and open-access data to produce a more accurate and higher resolution physiography database. One component to improve is the land cover map. The reference data gathers multiple high-resolution thematic maps thanks to an agreement-based decision tree. The input data are taken from the Sentinel-2 satellite. Then, the land cover map generation is made with image segmentation. This work implements several algorithms of different families to study their suitability to the land cover classification problem. To date, the best performing model is a U-net, with an accuracy of 67%. Compared to existing work, this work is innovative in the reuse of open data in nearly all components: the reference map, the source data and the model starting weights.

Material

Existing satellite data

Direct access
⊕ Geographical coverage, bands available, up to date
⊖ High pre-processing, clouds

Benchmarck dataset
⊖ Low pre-processing, cloudless, leaderboard
⊕ Geographical coverage, bands available, task design

BigEarthNet
A Large-Scale Sentinel Benchmark Archive
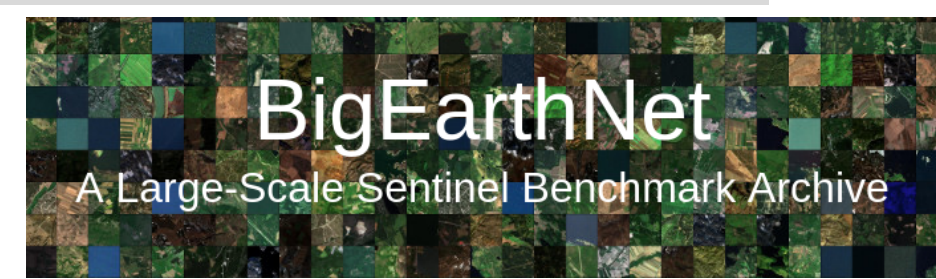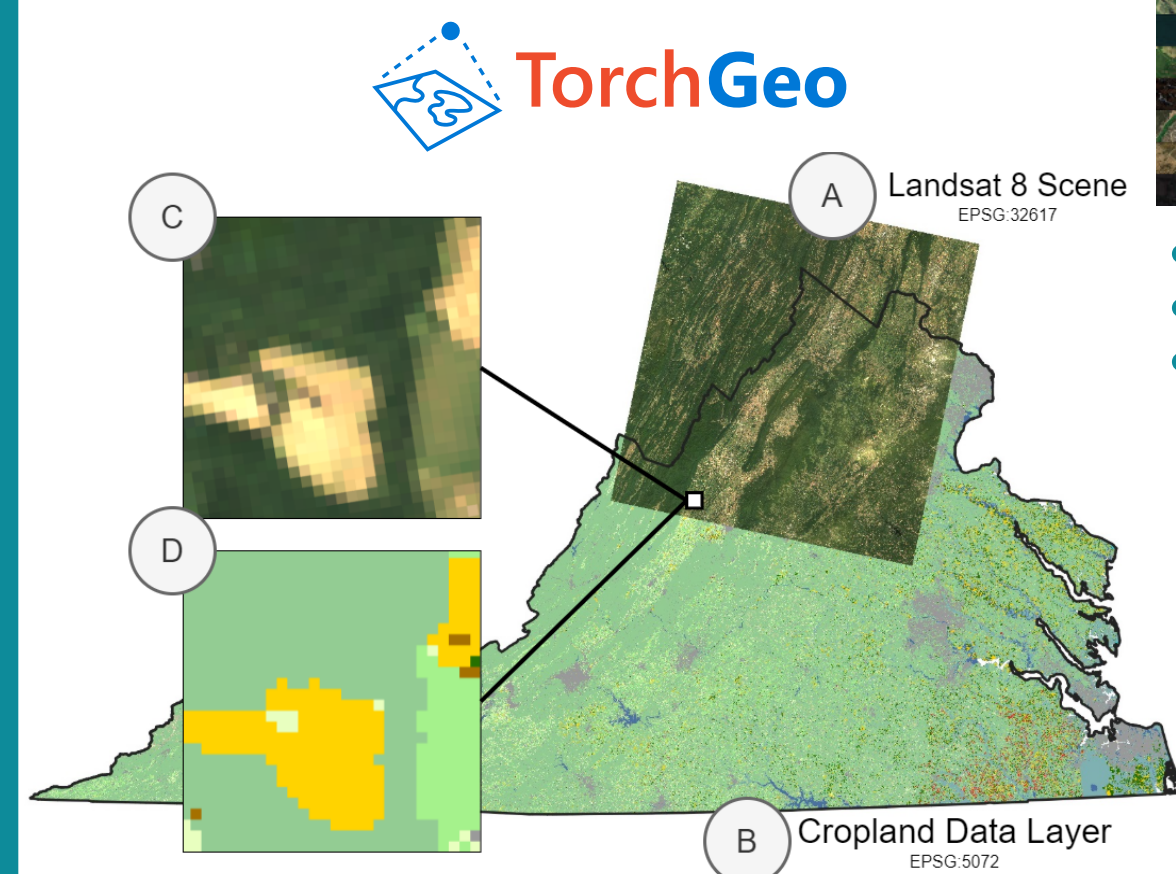
- 590,326 patches
- 1.2km x 1.2 km
- Multiple CORINE labels
- Sentinel 1-2 all bands
- No train/test split
- 10 European countries

TorchGeo

A — Landsat 8 Scene EPSG 32617
B — Cropland Data Layer EPSG 5072

Maps are often on different grids, with different coordinate reference systems. TorchGeo and Rasterio help making patches of pixel-aligned maps

Existing land cover maps

ESA WorldCover — LUCAS
Crops maps — Google Dynamic World — S2GLC
ESRI — CORINE — Forest maps
Urban maps — National maps
ELC10

None of the existing maps has the appropriate labels for SURFEX*. However, some labels are comparable when aggregated

*SURFEX: Surface model used in numerical weather prediction

ECOSG label	ECOSG thematic label	S2GLC label	Esri 2020	Corine primary/secondary/tertiary mix	ESA WorldCover
Sea and oceans					
Lakes	Water bodies	Water bodies	Water	Inland Waters	Open Water
Rivers				Marine Waters	
Bare land				Beaches dunes sands	
	Bare land	Natural material surfaces	Bare ground	Sparsely vegetated areas	Barren / sparse vegetation
Rocks				Burnt areas	Moss and lichen?
				Bare rocks	
Permanent snow and ice	Snow	Permanent snow-covered surfaces	Snow/Ice	Glaciers and perpetual snow	Snow and ice
boreal broadleaf deciduous					
temperate broadleaf deciduous		Broadleaf tree cover			
tropical broadleaf deciduous					
temperate broadleaf evergreen	Forest		Trees	Forests	Trees
tropical broadleaf evergreen		Coniferous tree cover			
boreal needleleaf evergreen					
temperate needleleaf evergreen					
boreal needleleaf deciduous					
shrubs	Shrubs	Moors and heathland	Shrub/Scrub	Moors and heathland	Shrubland
				Transitional woodland-shrub	
boreal grassland		Herbaceous vegetation		Natural grassland	
temperate grassland	Grassland	Sclerophyllous vegetation	Grass	Pastures	Grassland
tropical grassland				Sclerophyllous vegetation	
winter C3 crops		Cultivated areas		Arable land	
summer C3 crops	Crops	Vineyards	Crops	Permanent crops	Cropland
C4 crops				Heterogeneous agricultural areas	
Flooded trees	Flooded vegetation	Marshes	Flooded Vegetation	Inland Wetlands	Mangroves
Flooded grassland		Peatbogs		Maritime wetlands	Herbaceous wetland
LCZ1: compact high-rise					
LCZ2: compact midrise					
LCZ3: compact low-rise					
LCZ4: open high-rise					
LCZ5: open midrise					
LCZ6: open low-rise	Urban	Artificial surfaces	Built area	Artificial surfaces	Built - up
LCZ7: lightweight low-rise					
LCZ8: large low-rise					
LCZ9: sparsely built					
LCZ10: heavy industry					

Method

Classification: a collection of methods

Image segmentation

Source → Target

Pixelwise classification
- Random forest with 50 trees, 2.5Mpx in training set
- MLP with 2 hidden layers of 128 neurons each, trained on 65Mpx for 100 epochs

Transfer learning
- Pre-trained classifier used as backbone in segmentation network
- Backbone: ResNet50 trained on BigEarthNet (RGB bands, only first label) for 30 epochs (10 with last 2 layers unfrozen, 20 with all unfrozen)
- Segmentation: Fully Convolutional Network, trained for 30 epochs (backbone frozen) on BigEarthNet with ECO-SG labels

"Vanilla" U-net trained from scratch
- Trained for 100 epochs on BigEarthNet (RGB bands) with ECO-SG labels
- 90/10 % train/test split, CosineAnnealing scheduler

Reference map: built with an agreement-based decision tree

Datasets
Reprojected + downscaled/upscaled to 60-m resolution

Start here:
- Thematic data
- WorldCover10
- Other 10m land cover
- ECOCLIMAP-SG

Processing
1. Do high resolution maps match?
2. Is there any more reliable thematic map?
3. Does the WC10 map agree with thematic map?
4. Does the WC10 map agree with ECO-SG?

Outcome
- Quality flag: 1 — Highest quality — Apply thematic label
- Quality flag: 2
- Quality flag: 3 — Apply ECO-SG label
- Quality flag: 10
- Quality flag: 11
- Quality flag: 12 — Lowest quality

First results

Reference map

Sentinel 2 (Dublin city center)

ECO-SG (current reference map)

ECO-SG+ (future reference map)

U-net trained from scratch

Training UNet with SGD on BEN-ECOSG-120x120-full-0

◀ Learning curves
▼ Prediction examples

Sentinel image

ECOCLIMAP-SG

UNet trained from scratch

- 29: LCZ9: sparsely built
- 17. temper.
- 15. shrubs
- 13. temper.
- 1. sea and
- 32: LCZ9: s
- 28. LCZ5: 0
- 21. C4 crop
- 19. winter C

Conclusions, prospects

Conclusions

- Improvement of the land cover component of the physiography benefits from **loads of open access data** and many machine learning works.

- Multiple open access land cover maps have been used to **create a reference map (ECO-SG+)** with the appropriate labels for weather prediction thanks to an **agreement-based decision tree.**

- While ECO-SG+ was under construction, machine learning algorithms have been trained on the BigEarthNet patches with the ECO-SG labels. To date, **the best performing model is a U-net**, with an overall accuracy of 67%.

- Despite the poor resolution of ECO-SG, the U-net **correctly represents some features** of the land.

Prospects

- Improve the **reference**: train on ECO-SG+ instead of ECO-SG.

- Improve the **input**: use more bands than RGB, try more appropriate benchmark datasets than BigEarthNet (e.g. SEN12MS), ensure the same input data will be available for inference

- Improve the **training**: perform hyperparameter tuning for all the models implemented, perform ablation studies, try other approaches (contrastive learning), compare with similar methods

- Improve **collaboration**: ensure readability and reusability of the code, keep regular contact with other teams in the topic.