A DEEP LEARNING APPROACH FOR CROP TYPE MAPPING BASED ON COMBINED TIME SERIES OF SATELLITE AND WEATHER DATA

Nicoletta Addimando1*, Michael Engel2, Frederic Schwarz1, Matej Batić3

1meteoblue AG, Basel, Switzerland - nicoletta.addimando, frederic.schwarz@meteoblue.com
2Chair of Remote Sensing Technology, Department of Aerospace and Geodesy, Technical University of Munich, Germany - m.engel@tum.de
3Sinergise LTD, Ljubljana, Slovenia – matej.batic@sinergise.com
* Corresponding author. Email: nicoletta.addimando@meteoblue.com

KEY WORDS: Earth Monitoring, Satellite Image Time Series, Weather Time Series, Crop Dataset, Deep Learning, Crop Type Mapping, Sentinel-2

ABSTRACT:

Global Earth Monitor (GEM, Horizon 2020) takes advantage of the large volumes of available Earth Observation (EO), weather, climate and other non-EO data to establish economically viable continuous monitoring of the Earth. Within the GEM framework, the development of scalable and cost-effective solutions is being tested on several use-cases, with crop identification being one of them. Crop identification uses a combination of EO and weather data to enable automatic identification of crops. The use case supports operational decisions when managing crops and the monitoring of actual or planned agricultural land use (e.g., Common Agricultural Policy monitoring). Satellite data and weather data come at very different temporal and spatial resolutions: Sentinel-2 constellation nominally provides an observation of a field every 5 days at 10 m spatial resolution, while weather data has continuous hourly time series at multi-km spatial resolution. We have designed ad-hoc routines to spatially aggregate satellite data at field level and to systematically compose layers of different time discretization series, so that each EO is associated with a complete time series (of opportune length) of weather variables at daily resolution. For each field, we extract the time series of the median over field pixels of Sentinel-2 L1C bands, cloud mask and cloud probability. For this, we use a flexible encoding-decoding framework (FlexMod, designed by TUM as part of GEM project): multiple encoders are designed for features of different time length (namely EO data and weather variables) and are then passed to the decoder via a mediator. Thanks to the flexible design of FlexMod framework, different models and architectures can be easily tested by simply defining new encoders and/or decoders. We then train the deep model, exploiting the potential of FlexMod to test different feature extractors, temporal encoding frameworks and decoders and present a comparison between the results obtained training a long-short term memory (LSTM) implementation (BreizhCrops, Rußwurm et al. 2020) and a Self-attention transformer model (Vaswani et al. 2017), the latter showing the best performances with accuracy 0.904 and Cohen’s kappa 0.824. We moreover investigate the role of weather data by benchmarking results against those obtained with just satellite imagery. To better appraise the influence of weather data, we analyse how perturbing weather data in the testing dataset affects the final results. So far, we obtain in both cases very similar accuracies and Cohen’s kappa. A deeper analysis of crop-specific scores (precision, recall, F1) suggests that the training and testing datasets are too limited in terms of size and crop variability to draw any general conclusion over the role of weather. As future developments, once the EuroCrops datasets are ready, we plan to expand the training and testing dataset to cover a higher variability of climatological areas and increase the numerosity of the so far under-represented crops, in the attempt to draw more general conclusions around the influence of weather and the predictability of specific crop classes. Moreover, given the encouraging scores, we aim to perform crop type mapping at least at European scale, thanks to the availability of the EuroCrops data and the cost-effective big data solutions developed during GEM project.

1. INTRODUCTION

1.1 Global Earth Monitor project

Global Earth Monitor (GEM, Horizon 2020) takes advantage of the large volumes of available Earth Observation (EO), weather, climate and other non-EO data to enable economically viable continuous monitoring of the Earth, driven by the transition from traditional “strip mode” monitoring to “spot mode” monitoring. This GEM approach is based on the drill down mechanism: fast (and cheap) global monitoring at low resolution, finding the areas of interest (AOI) to perform spot monitoring with (appropriately) high resolution data and more elaborate machine learning (ML) models. Such processes can be run continuously on a monthly, weekly, or even daily basis provided they work in a sustainable way - adding more value than their cost - at least on a continental if not global scale, able to automatically improve accuracy and detect changes as they occur.
A proprietary concept of Adjustable Data Cubes (a combination of static and dynamic data cubes using Sentinel Hub batch processing) has been integrated with the EO-oriented open-source Machine Learning (ML) framework eo-learn (Singerise development team, 2019). Modern ML technologies and approaches are used to construct global, scale-independent interpretation models with the special focus on causality and change detection.

The development of scalable and cost-effective solutions is tested on five use-cases: the built-up area use-case exploits appropriate ML models to detect urban elements; the land cover classification use-case aims to perform a baseline multi-user/multi-purpose land cover classification; the map-making use-case enables systematic lead detection of changes from mid- to very-high-resolution imagery; the conflict pre-warning use-case combines data mining techniques, EO, weather and other geospatial data with open sources of information (e.g., distribution of ethnicities or religion) to support the detection of hot-spot areas and support political decision-making. The crop identification use-case, object of this paper, is presented in the next paragraph.

### 1.2 Crop identification use-case

Crop identification uses a combination of EO and weather data to enable automatic identification of crops. The service can support crop and land management, operational decisions when managing crops, the monitoring of actual vs. planned or reported agricultural land use (e.g., Common Agricultural Policy monitoring), environmental monitoring and governance, commodity supply and price prediction.

From the EO perspective, agriculture is a complex phenomenon which poses unique challenges. For example, the same crop type can have different temporal and spectral appearance due to local land management, genotype features, site conditions or environmental factors such as weather. Temporal information is usually the key to differentiating individual crop types, making use of unique differences in seasonal growing characteristics. The goal is to develop a model to be run at least at European scale (for those states that share crop declarations data publicly) that combines satellite data and weather variables to discriminate between crop species, with uncertainties considered acceptable for operational purposes.

### 2. DATA AND METHOD

#### 2.1 Data

Satellite data and weather data come at very different temporal and spatial resolutions: Sentinel-2 constellation nominally provides an observation of a field every 5 days at 10 m spatial resolution, while weather data has continuous hourly time series at multi-km spatial resolution.

We have designed ad-hoc routines to spatially aggregate satellite data at field level and to systematically compose layers of different time discretization series, so that each satellite observation is associated with a complete time series (of opportune length) of weather features at daily resolution.

#### 2.1.1 Satellite data:

For each field, we extract the time series of the median of field pixels of Copernicus Sentinel-2 L1C top-of-atmosphere reflectance, cloud mask and cloud probability, for a total of 15 variables (Copernicus Sentinel data, 2019). For doing this we take advantage of Sentinel Hub's Statistical API (Singerise, 2020), that enables the retrieval of statistics of band values and derived indices over a specified geographic area and time range.

#### 2.1.2 Weather data:

Using meteoblue dataset API (meteoblue, 2017), complete time series of daily weather data at 4 km resolution (NEMS4 model, meteoblue, 2008) are associated to each field observation, following a systematic layer composition approach, illustrated in Figure 1. The challenge when dealing with incomplete data is not to lose information (e.g., with accumulation techniques) as well as not to introduce artificial information (e.g., with interpolation techniques). To maximize and preserve the information content of weather and satellite data, we therefore extend dataset A (green squares in Figure 1, representing EO data) with a consistent number of instances of dataset B (blue squares in Figure 1, representing weather data), so that the time instances of B before the time instance of A could be looked at as additional channels of A. Note that the operation denoted by the red arrows is not necessarily equal to identity. That is, encoding, compression or accumulation techniques may come into play whereas their effect on the information contained by the data must be considered as described above. Using the encoding-decoding framework, presented in Section 2.2, many different techniques can be tried.

![Figure 1. Systematic layer composition approach for dealing with incomplete data. In our use case, the green and the blue squares symbolize satellite observations and weather variables, respectively.](image)

An opportune time series length is defined for each of the 17 weather variables considered, the choice of the length based on agronomical considerations over the longer- and shorter-term effect that each weather variable may reasonably have on crop development and its appearance on satellite imagery (Table 1).

| Weather variables | Time series length |
|-------------------|--------------------|
| Precipitation, radiation | 30 days |
| Temperature (min, max, mean), growing degree days (T base 5 °C), temperature range | 14 days |
| Wind speed (mean, maximum), wind gust (mean, maximum), number of frost days, number of icing days, number of heat days, number of tropical nights | 7 days |
| Soil moisture, relative humidity | 3 days |

**Table 1.** Weather variables and associated time series length.
2.1.3 Ground-truth data: to train the model and validate the results, we used a crop database of Slovenia (INSPIRE metadata Slovenia, 2019). The dataset is composed by field polygons associated to the recorded crop that was grown during 2019. To build a homogeneous crop database that overcomes the country-specific classification system, crop fields were re-labelled according to a Hierarchical Crop and Agriculture Taxonomy (HCAT). This taxonomy, based on the EAGLE-Matrix and EU regulations, is the one adopted in the EuroCrops project (Schneider et al. 2021) and is propaedeutic to the future expansion of crop identification to Europe. Its 81 classes are represented in Figure 2, where the circles represent further distinction levels from 0 (inner circle) to 4 (outer circle).

2.2 Encoding-decoding framework

To handle this kind of multi-dimensional layered data, we use a flexible encoding-decoding framework (FlexMod, ongoing design by TUM as part of GEM project): multiple encoders are designed for features of different shape, so time length, discretization, or dimension, (namely EO data and weather variables) and are then passed to the decoder thanks to a mediator, as shown in Figure 3.

Thanks to the flexible object-oriented design of the FlexMod framework, different models and architectures can be easily tested by simply defining new encoders, mediators and/or decoders. This flexibility is based on the ability to import any package, file, or even specific parts (method, class) of a file stored at any place on a machine. Although this is a technical detail, we emphasize this as the foundation of our flexible and agile development for many different configurations, of which we highlight some.

The FlexMod itself only uses other models and combines them to a larger one which intrinsically enables to standardize things down the processing pipeline like the training or inference loops. Other standardized procedures include rasterization techniques to make maps out of our classification results for customers. Especially if these monitoring tasks shall be done on demand by a distributed system, standardized model frameworks foster easy implementation of coordinator nodes forwarding jobs for calculation.

From the ML perspective, the FlexMod is a rather simple encoder-decoder structure where multiple encoders are used for multimodal data. For the sake of completeness, we note that the encoders could work as some scalers, standard methods or identity transforms as well. The mediator, however, could work as another encoder. In any case, the mediator concatenates the outputs of the foregoing encoders or feature extractors to one tensor. That is forwarded to the decoder (or classifier) then. In our experiment, the encoders are thought of as feature extractors from multimodal data whose results are concatenated by the mediator and analysed by the decoder then. The available configurations are illustrated in Figure 3.

Figure 2. A representation of the EuroCrops taxonomy classes and levels (Schneider et al. 2021).

Figure 3. Basic concept of the FlexMod Framework fostering flexibility and standardization. Blue bars denote data or tensors not true to scale.

2.3 Performance evaluation

Experiments are evaluated by computing and comparing the accuracy, Cohen’s kappa, precision, recall and F1 scores, presented below.

2.3.1 Accuracy: the overall accuracy of the prediction, defined as the fraction of correct predictions (Pedregosa et al., 2011):

$$\text{Accuracy}(y, \hat{y}) = \frac{1}{n_{\text{samples}}} \sum_{i=0}^{n_{\text{samples}}-1} 1(\hat{y}_i = y_i)$$ (1)

where $\hat{y}$ = predicted value of the -th sample  
$y$ = true value of the -th sample  
n_{\text{samples}} = number of samples  
$1(x)$ = indicator function

2.3.2 Cohen’s kappa: expressing the level of agreement between two annotators versus the possibility of the agreement occurring by chance (Cohen, 1960; Artstein and Poesio, 2008; Pedregosa et al., 2011):

$$\text{Cohen’s kappa} = \frac{p_o - p_e}{1 - p_e}$$ (2)

where $p_o$ = empirical probability of agreement on the label assigned to any sample  
$p_e$ = expected agreement when both annotators assign labels randomly

This score is deemed important in evaluating the results minimizing the bias due to strongly unbalanced classes in terms of samples’ numerosity (Powers, 2015). The model prediction
may in fact score high accuracies just for the fact that the annotator is always assigning the most numerous label.

2.3.3 Precision and recall: they represent respectively the ability of the classifier not to label as positive a sample that is negative and to find all the positive samples (Olson and Delen, 2008; Pedregosa et al., 2011):

\[
\text{Precision} = \frac{tp}{tp + fp} \\
\text{Recall} = \frac{tp}{tp + fn}
\]

where \( tp \) = number of true positives, \( fp \) = number of false positives, \( fn \) = number of false negatives.

2.3.4 F1-score: the weighted harmonic mean of precision and recall, where precision and recall are evenly weighted (Shantanu and Sunita, 2004; Wikipedia, 2022; Pedregosa et al., 2011):

\[
F1 = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}
\]

3. RESULTS AND CONCLUSIONS

We present a selection of experiments (Table 2) carried out on the study area of Slovenia, using the crop database of year 2019, constituted by a total of 590720 fields and 45 different crops. We test the usage of Long-Short Term Memory (LSTM) implementation (Breizhcrops, Rußwurm et al. 2020) and a Self-attention Transformer model (Vaswani et al. 2017) as decoders. As feature encoders, we try Multi-layer Perceptron (MLP) and Temporal Attention Encoder (TAE).

| Features          | Model          | Test | Satellite | Weather | Encoder | Decoder   |
|-------------------|----------------|------|-----------|---------|---------|-----------|
| 1                  |                | yes  | yes       | MLP     | LSTM    |
| 2                  |                | yes  | yes       | MLP     | Transformer classifier |
| 3                  |                | yes  | yes       | TAE     | Transformer classifier |
| 4                  |                | yes  | -         | TAE     | Transformer classifier |
| 5                  |                | yes  | perturbed in test subset | TAE | Transformer classifier |

Table 2. Summary of the experiments presented.

3.1 Training and testing subset

We split the ground-truth dataset in training, validation and testing subset. For doing this, we use Torch’s random split utility (PyTorch development team, 2019), setting the seed to grant reproducibility. This choice ensures the presence of all crop classes in each subset (see Table 6 in Appendix for more details). Training is performed on 60% of the data; the remaining 40% is further split in a validation subset (20% of the data), for monitoring model convergence, and a testing subset (20% of the data), where scores are evaluated.

3.2 Model

Crop identification is performed on the testing subset using LSTM and Self-attention Transformer model as decoders (test 1 and 2 in Table 2). The resulting scores are compared in Table 3, where we report overall accuracy and Cohen’s kappa computed considering all crop classes and excluding the class “pasture / meadow”. We report also results without the “pasture / meadow” class because it is by far the most numerous (more than 8 times the numerosity of other classes, see Figure 4) and is also one of the few representing a level 1 distinction, together with “arable crops”, “permanent crops” and “mushrooms, energy crops and genetically modified crops” (Figure 2).


![Figure 4. Number of samples per crop class in the testing dataset. For visualization purposes the y axis is limited to 8000. Two classes are more numerous: grain maize (9047 samples) and pasture / meadow (77684 samples).](image)

| Including “pasture/meadow” | Excluding “pasture/meadow” |
|-----------------------------|----------------------------|
| Model | Accuracy | Cohen’s kappa | Accuracy | Cohen’s kappa |
| LSTM    | 0.888    | 0.796       | 0.728    | 0.687       |
| Transformer | 0.902  | 0.819       | 0.750    | 0.715       |

Table 3. Overall accuracy and Cohen’s kappa obtained on the testing subset using an LSTM and a Transformer model, both including and excluding the crop class “pasture/meadow” from scores computation.

The transformer classifier performs better, with scores overall 2-3% higher than LSTM. This is generally true also for class-specific scores (Figure 5).
These tests are based on MLP feature encoders. We also test the usage of a different encoder, in particular a Temporal Attention encoder, always in association with the better performing Transformer decoder (test 2 and 3 in Table 2). Prediction results are compared in Table 4 and Figure 6. The Transformer encoder performs only minimally better than MLPs, particularly on the most numerous classes. Apart from that, scores are very similar. Confusion matrices obtained with the transformer encoder-decoder model are reported in Appendix, Figure 8 and Figure 9.

Table 4. Comparison of overall accuracy and Cohen’s kappa obtained on the testing subset with MLP and Transformer feature encoders. Both tests use a Transformer model as decoder. We present results including and excluding the crop class “pasture/meadow” from scores computation.

| Feature encoder | Including “pasture/meadow” | Excluding “pasture/meadow” |
|-----------------|-----------------------------|-----------------------------|
|                 | Accuracy        | Cohen’s kappa | Accuracy        | Cohen’s kappa       |
| MLP             | 0.902           | 0.819         | 0.750           | 0.715                |
| Transformer     | 0.904           | 0.824         | 0.757           | 0.722                |

3.3 The role of weather

To assess and quantify the value added of weather, prediction results are benchmarked against those obtained training the model using only satellite data. We present this comparison using the most promising model presented so far (self-attention transformer decoder, with self-attention transformer feature encoder when using weather data). In addition, we also try perturbing the weather features of the testing subset with Gaussian white noise (Peebles, 2001; NumPy Developers, 2022) with zero mean and standard deviation equal to half the variability range of each weather variable. The same random perturbation is applied to all the weather observations of a given field. The resulting scores (test 3, 4 and 5 in Table 2) are reported in Table 5 and Figure 7.

Table 5. Overall accuracy and Cohen’s kappa obtained on the testing subset using a Transformer model trained with both satellite and weather data and only with satellite data. In the last row, the scores obtained perturbing the weather data of the testing subset with Gaussian white noise.

|                          | Including “pasture/meadow” | Excluding “pasture/meadow” |
|--------------------------|-----------------------------|-----------------------------|
|                          | Accuracy        | Cohen’s kappa | Accuracy        | Cohen’s kappa |
| Satellite + weather      | 0.904           | 0.824         | 0.757           | 0.722         |
| Satellite only           | 0.901           | 0.818         | 0.747           | 0.711         |
| Satellite + perturbed    | 0.901           | 0.817         | 0.744           | 0.708         |

Satellite and weather data.
The overall accuracy and Cohen’s kappa (Table 5) seem to suggest that weather is not playing a role: scores obtained using only satellite data are very similar to those obtained also using weather data. In addition, even perturbing the weather data in the testing subset, accuracy and Cohen’s kappa remain almost unchanged. From a deeper analysis of the class-specific scores of Figure 7, though, we can derive some initial considerations, to be confirmed by future, more extensive testing. Weather seems to be less influential for winter species: this is particularly evident when looking at scores for the three most numerous classes of winter crops (winter common wheat and spelt, winter barley, winter triticale and winter rapeseed, Figure 4). The fact that weather is less relevant in distinguishing winter crops, is expected: winter crops in fact go through a long period of dormancy and their appearance is unchanged for the first 40-70 days of the year (Liu et al., 2021).

On the other hand, also the very numerous maize classes show little or no sensitivity to the presence of weather data, pointing more strongly towards the thesis that weather is not adding value to the analysis.

The model seems to benefit from the presence of weather data in those cases where the performances of the model with satellite data only are relatively low (e.g., winter rye, summer barley, summer oats, summer triticale, peas, beans and lupins, sugar beet, sunflowers), suggesting that weather could help detect crops that are most hard to distinguish from Earth Observation only. Despite this, it remains difficult to explain why predictions perturbing weather data present in most cases scores very similar to those obtained on the unperturbed dataset. The most sensible reasoning is that the dataset in Slovenia is too small to pick apart differences in crop phenomenologies that arise from local weather phenomena.

### 3.4 Conclusions and future development

As part of the GEM Horizon 2020 project, we developed a novel, flexible encoding-decoding framework, that allows to easily setup experiments to train, test and compare different models. Building on API services of Sinergise and meteoblue, we can efficiently associate EO and weather data to ground-truth datasets. The available crop databases, progressively collected and harmonized within the EuroCrops project, are pre-processed and prepared for training machine learning models. The models predict, based on satellite data and weather model data, what crop is grown during a particular year, in a certain field. We validate the results on subset of the crop database, not used for training. Performances of different encoder-decoder models are evaluated on a subset of Slovenia dataset of year 2019, the best results obtained for classification of 45 crop types so far with a Self-attention (transformers) encoder-decoder showing accuracy of 0.904 and Cohen’s kappa 0.824.

We moreover investigate the role of weather data by benchmarking results against those obtained with just satellite imagery. To better appraise the influence of the weather data we analyse how perturbing weather data in the testing dataset affects the final results. So far, we obtain in both cases very similar accuracies and Cohen’s kappa, suggesting that the training and testing datasets are too limited in terms of size and crop variability to draw any general conclusion over the role of weather.

As future developments, we aim to expand the training and testing dataset to cover a higher variability of climatological areas and increase the numerosity of the so far under-represented crops. We would like, for example, to train different models on different climatological areas and benchmark these results against the usage of a single model. This will allow us to draw more general conclusions around the influence of weather and the predictability of specific crop classes.

Independently from considerations around weather, given the encouraging scores, we plan to perform crop type mapping at European scale, thanks to the cost-effective big data solutions developed during GEM project.

### ACKNOWLEDGEMENTS

This work was carried out in the framework of the European Union’s Horizon 2020 Global Earth Monitor project, grant agreements No. 101004112.

### REFERENCES

Artstein, R., Poesio, M., 2008. “Inter-coder agreement for computational linguistics”. Computational Linguistics 34(4):555-596.

Cohen, J., 1960. “A coefficient of agreement for nominal scales”. Educational and Psychological Measurement 20(1):37-46, doi:10.1177/001316446002000104.

Copernicus Sentinel data, 2019, processed by ESA, retrieved from Sentinel-Hub.

GEM Global Earth Monitor, European Union project, https://www.globalearthmonitor.eu/
**APPENDIX**

Complete list of crop classes (Slovenia 2019 database) and their numerosity in the training and testing subset, as a result of a random split.

| Crop class                                      | Training | Testing |
|------------------------------------------------|----------|---------|
| summer common wheat and spelt                   | 359      | 119     |
| winter common wheat and spelt                    | 15299    | 5075    |
| summer durum wheat                              | 9        | 2       |
| winter durum wheat                              | 30       | 8       |
| summer rye                                      | 7        | 2       |
| winter rye                                      | 713      | 227     |
| summer barley                                   | 569      | 176     |
| winter barley                                   | 15449    | 5343    |
| summer oats                                     | 827      | 264     |
| winter oats                                     | 381      | 119     |
| grain maize                                     | 27207    | 9047    |
| summer triticale                                 | 147      | 43      |
| winter triticale                                 | 3957     | 1340    |
| millet                                          | 84       | 28      |
| other cereals for the production of grain        |          |         |
| peas, field beans and sweet lupins              | 303      | 98      |
| other dry pulses                                | 17       | 5       |
| potatoes                                        | 3190     | 1067    |
| sugar beet                                      | 71       | 25      |
| fodder roots and brassicas                      | 415      | 138     |
| cotton                                          | 392      | 131     |
| rapseseed and turnip rapseseed                  | 17       | 5       |
| summer rapseseed                                | 31       | 7       |
| winter rapseseed                                | 1277     | 437     |
| sunflower and yellow bloomer                    | 242      | 84      |
| soya                                            | 634      | 213     |
| other oil seed crops                            | 71       | 24      |
| flax                                            | 22       | 6       |
| hemp                                            | 240      | 83      |
| aromatic, medicinal and culinary plants         |          |         |
| other industrial crops                          | 6764     | 2220    |
| fresh vegetables, melons and strawberries        | 2917     | 941     |
| radish/turnips                                  | 6        | 1       |
| cucurbits                                       | 3011     | 1007    |
| flowers and ornamental plants                   | 135      | 51      |
| green maize                                     | 19290    | 6403    |
| other plants harvested green                    | 76       | 19      |
| fallow land (not crop)                          | 978      | 303     |
| pasture / meadow                                | 233128   | 77684   |
| permanent crops                                 | 2        | 1       |
| olive plantations                               | 1110     | 383     |
| vineyards                                       | 14160    | 4735    |
| nurseries                                       | 164      | 61      |
| other permanent crops                           | 20       | 9       |

**Table 6.** Numerosity of each crop class in the training and testing subsets.
Confusion matrices obtained with the best performing model (transformer encoder-decoder), using satellite and weather data.

**Figure 8.** Confusion matrix normalized by target label.

**Figure 9.** Confusion matrix normalized by predicted label.