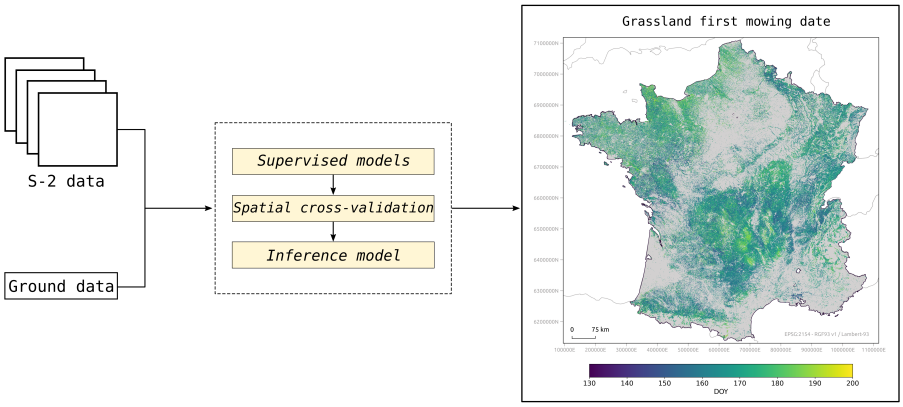
Nationwide operational mapping of grassland mowing events combining machine learning and Sentinel-2 time series

Henry Rivas, Mathieu Fauvel, Vincent Thiérion, Millet Jerôme, Laurence Curtet



Graphical Abstract

Nationwide operational mapping of grassland mowing events combining machine learning and Sentinel-2 time series

Henry Rivas, Hélène Touchais, Vincent Thierion, Jerome Millet, Laurence Curtet, Mathieu Fauvel

Highlights

Nationwide operational mapping of grassland mowing events combining machine learning and Sentinel-2 time series

Henry Rivas, Hélène Touchais, Vincent Thierion, Jerome Millet, Laurence Curtet, Mathieu Fauvel

- Estimation of grassland first mowing date using regression algorithm
- Time aware deep-learning architectures were the most accurate models
- Oversampling techniques did not improve predictions of extreme mowing dates
- Threshold-based methods underperformed compared to all supervised models
- LTAE performed reliably across all unknown sites, demonstrating transferability

Nationwide operational mapping of grassland mowing events combining machine learning and Sentinel-2 time series

Henry Rivas[a,*], Hélène Touchais[a], Vincent Thierion[a], Jerome Millet[b], Laurence Curtet[c], Mathieu Fauvel[a]

[a] *Centre d'Etudes Spatiales de la Biosphère (CESBIO) Université de Toulouse, CNES/CNRS/INRAE/IRD/UT3-Paul Sabatier 31401 Toulouse France*
[b] *Office Français de la Biodiversité (OFB) Direction de la Recherche et de l'Appui Scientifique, 79360 Villiers-en-Bois France*
[c] *Office Français de la Biodiversité (OFB) Direction de la Recherche et de l'Appui Scientifique, 01330 Montfort, Birieux France*

Abstract

Grassland dynamics are modulated by management intensity and impact overall ecosystem functioning. In mowed grasslands, the first mowing date is a key indicator of management intensity. The aim of this work was to assess several supervised regression models for mapping grassland first mowing date at national-level using Sentinel-2 time series. Three deep-learning architectures, two conventional machine learning models and two threshold-based methods (fixed and relative) were compared. Algorithms were trained/calibrated and tested from field observations, using a spatial cross-validation approach. Our findings showed that time aware deep-learning models -Lightweight Temporal Attention Encoder (LTAE) and 1D Convolutional Neural Network (1D-CNN)- yielded higher performances compared to Multilayer Perceptron, Random Forest and Ridge Regression models. Threshold-based methods under-performed compared to all other models. Best model (LTAE) mean absolute error was within six days with a coefficient of determination of 0.52. Additionally, uncertainties were accentuated at extreme mowing dates, which were underrepresented in the data set. Oversampling techniques did not improve predicting extreme mowing

*Corresponding author

dates. Finally, the best prediction accuracy was obtained when the number of clear dates surrounding the mowing event was greater than 2. Our outputs evidenced time aware deep-learning models' potential for large-scale grassland mowing monitoring. A national-level map was produced to support birdlife monitoring or public policies for biodiversity and agro-ecological transition in France.

Keywords: Regression, Deep-learning models, Mowing dates mapping, Grassland management intensity, Satellite image time series

1. Introduction

Grasslands cover approximately 40% of the Earth's land area, encompassing nearly 70% of the global agricultural land area, and they are distributed on all continents and all latitudes [1, 2]. Grassland dynamics are modulated by management intensity and influence overall ecosystem functioning [3, 4], biodiversity [5], carbon sequestration [6], water quality [7], and more [3]. Grasslands are subject to management practices such as mowing or grazing or a combination of both. These practices are primarily driven by grassland landscape maintenance as well as by ecosystem service of provisioning supplied from grasslands. The choice of a specific practice therefore impact biodiversity [8, 9]. For instance, in mowed grasslands, the first mowing event date is a key indicator of plot management intensity [10]. Consequently, this timing information could be critical for assessing above-mentioned topics [11, 5, 10].

In France, the *Observatoire National de l'Ecosystème Prairie de Fauche*[1] (ONEPF) conducts birdlife monitoring in mowed grasslands and has related breeding failures to early mowing date. Indeed, Broyer et al. [10] demonstrated that early mowing intercepts birds' reproductive period and interrupts their breeding process. Usually, this monitoring is supported by local field observation campaigns. However, they are time-consuming and difficult to repeat regularly.

[1] National Observatory of Mowed Grassland Ecosystems : `https://www.ofb.gouv.fr`

Thus they are not spatially/temporally exhaustive. Hence, field campaigns need to be complemented with other data acquisition process.

Remote sensing data enable regular and global-scale monitoring, facilitating tracking of vegetation dynamics at high spatial resolution and frequent revisit. For instance, Sentinel-2 mission provides cost-free high resolution data (10m as spatial resolution and 5 days revisit), allowing intra-plot level observations. Such satellite data have proven their interest to monitor vegetation at large scale [12]. In this paper, the underlying idea was that satellite time series could be used to infer grassland first mowing date at national-level using supervised machine learning algorithm, supporting birdlife monitoring in France.

Grassland mowing events and intensity estimation have been investigated using satellite image time series (SITS), mainly through features sensitive to vegetation status, such as Normalized Difference Vegetation Index (NDVI), Enhanced Vegetation Index (EVI), Leaf Area Index (LAI) and more [13]. These methods usually exploit the temporal information contained in SITS to detect moving events: a significant variation is usually associated to an event. Methods differ in how such drop is computed. For instance, Estel et al. [14] assessed annual mowing frequency using temporal change analysis based on spline-adjusted MODIS NDVI time series. Their approach involved identifying mowing events as instances where a local minimum exhibited a change relative to its preceding peak. The results showed an overall accuracy of 80%, which decreases as the frequency of events increases. In northern Switzerland, Kolecka et al. [15] also estimated mowing frequency employing similar temporal change analysis, but based on raw Sentinel-2 NDVI time series. A drop greater than 0.2 in NDVI value between two consecutive cloud-free acquisition dates was counted as a mowing event. Their method accurately identified 77% of observed events and highlighted that false detection can occur due to residual cloud presence, while sparse time series led to the omission of mowing events. In Griffiths et al. [16], mowing events frequency and timing were mapped in Germany using 10-day composite Harmonized Landsat and Sentinel-2 NDVI time series. Deviation from a hypothetical bell-shaped curve and the current polynomial-fitted curve

were evaluated. An event was counted when the difference exceeded 0.2× NDVI. Findings revealed consistent spatial patterns in mowing frequency (indicating extensive and intensive management). However, estimated dates exhibited significant discrepancies compared to observed dates (mean absolute error -MAE- greater than fifty days), which could be due to lower temporal resolution of Sentinel-2 before 2017 and the absence of reliable ground data for calibration and validation. Stumpf et al. [17] mapped grassland management (grazing or mowing) and its intensity based on biomass productivity and management frequency, respectively. The latter were extracted from Landsat ETM + and Landsat OLI NDVI composite time series. As in previous cases, a management event was counted when NDVI loss is higher than a threshold, which was based on the probability density function of all NDVI changes across the time series and was specified for p = 0.01. Their approach yielded management patterns consistent with several management-related indicators (species richness, nutrient supply, slope, etc). Recently, Watzig et al. [18] estimated mowing events in Austria, using Sentinel-2 NDVI time series and implementing difference analysis between an idealized unmowed trajectory and actual NDVI values. An event was recorded if the difference became too significant. Commission errors due to residual clouds were reduced via a subsequent binary classification of each estimated event using a gradient boosting algorithm trained over cloudy plots. Findings indicated an overall accuracy of 80% in correct event detection, with estimated dates closely aligning with observed dates (MAE $<$ 5 days).

Previous methods exploit only optical modality and can be limited by clouds cover. A strategy to reduce cloud-related gap in optical time series is to combine optical data from different sensors. For instance, Schwieder et al. [19] combined Sentinel-2 and Landsat-8 EVI time series for mowing events detection in Germany. They analyzed the deviation between actual observations and an idealized temporal profile (unmowed regime). An event was recorded when the difference exceeded the mean value of all absolute residuals. Overall, detected mowing dates exhibited an average absolute difference lower than 12 days compared to observed dates. Mowing events were detected with an aver-

age F-score of 0.60, while the estimation of their frequency showed a mean error lower than 40% of the actual number of mowing events. The authors highlighted that performance was lower in areas with less clean observations due to clouds covers.

To cope with optical sensors limitation, Vroey et al. [20] developed an algorithm for detecting mowing events across Europe using jointly raw Sentinel-2 NDVI and Sentinel-1 VH-coherence time series. A mowing event was deemed when temporal change in NDVI and VH-coherence standard deviation values exceeded given relative thresholds. VH-coherence standard deviation was calculated from residuals of the six preceding observations. These residuals capture disparities between linear-fitted values and actual values. In the final estimation, Sentinel-1 outputs were considered when Sentinel-2 omitted events due to cloud cover. Results demonstrated synergy between optical and radar data in detecting mowing events (F1-score of 79%). Using only Sentinel-2 data achieved maximum precision, but combining both sensors boosted recall significantly. Also relying on optical and radar data synergy, Reinermann et al. [21] mapped mowing frequency across Germany, from Sentinel-2 EVI and Sentinel-1 PolSAR entropy time series separately. A mowing event was counted when temporal change in EVI computed between consecutive local maxima and minima exceeded a given relative threshold. S1-based detection was used to find potentially missed mowing events in cloudy gaps (> 25 days) in optical observations. Findings showed that S2-based method correctly detected 60.3% of mowing events with an F1-Score of 0.64. However, combining S1 and S2 increased recall but also caused more false positives, lowering precision. Alternatively, Garioud et al. [22] jointly used Sentinel-1 and -2 SITS as well as climatic and topographic data to reconstruct continuous Sentinel-2 NDVI time series for mowing date estimation, based on NDVI drop analysis.

While threshold-based methods have been widely investigated, they are limited in their used because of manual tuning of the threshold parameter, which can be difficult when working on large areas. This issues is exacerbated when multi-modal time series are used. Supervised learning approaches have been

investigated in the last decade to circumvent this issue. Komisarenko et al. [23] estimated mowing events timing at plot level in Estonia, using a 1-D Convolutional Neural Networks (CNN) on Sentinel-2- and Sentinel-1-based features time series. Although fourteen features were used, NDVI and the harmonic mean of VV and VH coherence were found to be the most relevant. Their approach yielded an accuracy of 73%. Most of the incorrectly estimated events were observed when optical time series were sparse or the size of the plot was small. Lobert et al. [24] also used a similar deep-learning model (1-D CNN) on Sentinel-2/Landsat-8- and Sentinel-1-based features time series for mowing event frequency and timing detection. Among all tested feature combinations, the highest overall accuracy was reached when combined NDVI, backscatter cross-ratio and coherence with an F1-Score of 0.84. Estimated mowing dates showed a MAE of 3.79 days compared with the observed dates. In terms of management intensity, low-intensity grasslands were overestimated, while high-intensity grasslands were underestimated. Following a similar approach, Holtgrave et al. [25] tested four machine learning algorithms for mowing event detection in Germany. Sentinel-2/Landsat-8, Sentinel-1- and weather-based features time series were analyzed. Mowing events were detected by a binary classification approach (mown or unmown) for each observation in the time series, using the adjacent observations as predictors. 1D-CNN and Long Short-Term Memory algorithms provided the best results in terms of classification accuracy. Mowing events could be detected with F1-Score of up to 89% and first cut with up to 94%.

In the literature, there is no consensus on the optimal satellite data for mowing event detection in grasslands: for instance, [26] concluded that "*due to differences in grazing intensity and moisture conditions [...] it seems infeasible to determine a single wavelength or polarization that is best suited to detect cutting events*". Yet, some studies demonstrated Sentinel-1 data potential, due to their sensitivity to changes in vegetation cover structure, and their insensitivity to clouds [27]; but performance varies locally due to factors like soil moisture, vegetation water content, roughness, etc. Furthermore, some au-

thors combined Sentinel-2 and Sentinel-1 data to reduce cloud effects in time series, and found enhanced performance in some cases [20, 13, 23, 24, 25] or degraded performances [28]. Others studies combined data from different optical sensors [16, 17, 19]. Overall, most authors agree that optical data alone can effectively detect mowing events, provided that enough cloud-free observations are available [15, 16, 18, 17, 19].

Recent research have clearly showed the superiority of supervised machine learning-based approaches in mowing detection from remote sensing data. Such approaches were mainly formulated as a classification problem: *was there a mowing event during a given temporal sequence or not?* [24, 25]. Yet, the number of temporal sequences to be processed, their length and their possible overlap impact significantly the final accuracy, posing challenges in optimization at large-scale.

This paper targets the estimation of the first mowing event date using optical time series and machine learning algorithm formulated as a regression problem, to support birdlife monitoring in France. The choice to use only optical SITS was motivated by the significant additional computing cost to perform jointly radar and optical data at large-scale in regards to the limited reported improvement in the literature. Furthermore, regression algorithms were selected in contrast to classification ones in order to reduce the number of non-learnable parameters (such as those related to the temporal sequences) and to ease the learning step at large-scale.

In this study, a large field survey (more than 2 000 plots) was conducted on different and distant landscapes, to construct one of the largest data set of mowing events for mainland France. We compared several algorithms for predicting first mowing date: two threshold-based, two conventional machine learning-based and three deep-learning-based from which two process the temporal information specifically. The accuracy of each method was assessed using spatial cross-validation for several metrics. We provided an analysis of the influence of the clouds cover on the prediction accuracy for the best regression model. A national-level map was produced using the learned model and qualitative and

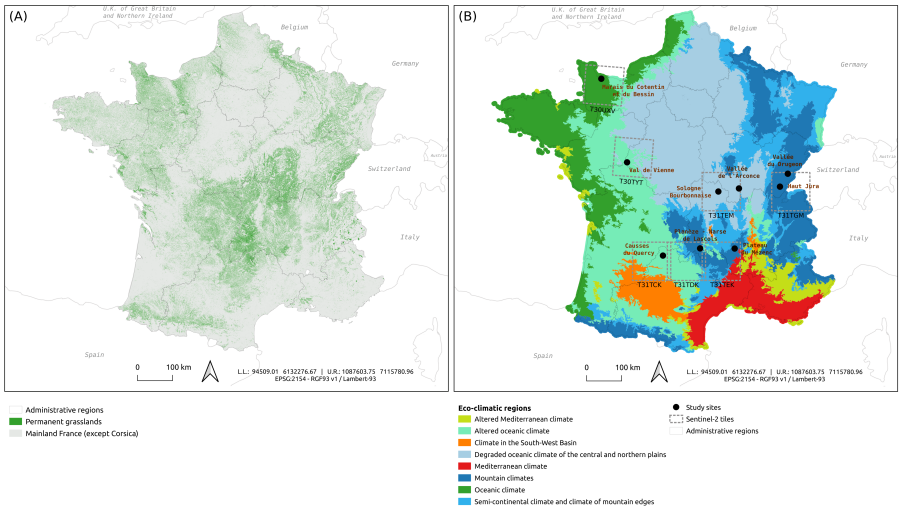


Figure 1: (A) Study area location. The gray color represents the delimitation of mainland France (except Corsica), while the green color represents the permanent grassland plots declared in the LPIS 2022. (B) Study sites location. The black dots represent the observation sites in 2022. The boxes in dashed gray lines represent the Sentinel-2 tiles that intercept each observation site. The color palette represents the eco-climatic regions in mainland France, as defined in [29].

quantitative accuracy were discussed.

The remainder of the paper is organized as follows. Section 2 presents study area as well as reference and remote sensing data. Then competitive algorithms for predicting first mowing date are presented in section 3. Results and discussion are reported in section 4 and section 5. Finally, conclusions and perspectives are drawn in section 6.

2. Materials

2.1. Study area

Our study area covers permanent grasslands across the mainland France (except Corsica), which represent 68.5% of the total grassland area -including permanent, temporary and other grasslands-, declared in the Land Parcel Identification System - LPIS [30] in 2022 (Figure 1). According to climate normals, annual rainfall is around 800-1 000 (mm), with a contrast between the western ($>$ 1 000 mm) and the southeastern (600-800 mm) regions. The average annual

temperature is about 11-13 °C, with 20-25 °C in summer and 5-10 °C in winter (`https://meteofrance.com/climat/normales/france`).

In LPIS, permanent grasslands are defined as surfaces with uninterrupted herbaceous cover for at least 6 years and are identified at the plot level with class code 18. These permanent grasslands alone account for approximately 27.5% (76 835 km^2) of the entire agricultural area reported in 2022. Grasslands cover regions that are less suitable for agricultural activities due to unfavorable climatic or site conditions (high altitudes, steep slopes, poor or wet soils). In mainland France, permanent grasslands are found in mountain chains in the center (Massif Central), western (Massif Armoricain), eastern (Jura and Vosges), Alps and Pyrenees, as well as in plains and wet regions (Figure 1). According to the LPIS, at least 75% of permanent grassland plots cover 2.80 hectares or less, and the largest plots -exceeding 20.0 hectares- are concentrated mainly in the center and eastern regions. Grasslands undergo various management practices such as mowing, grazing or a combination of both; and the intensity of these practices varies across plots, influenced by climate, site conditions and farmer decisions (fertilization, irrigation, …). Lower altitudes tend to offer more favorable conditions for mowing and more intensive management, except for damp or wet areas. In mainland France, grassland growing season spans from spring to autumn (March to October) and mowed grasslands are mainly managed extensively, with one or two mowing events per year (up to six mowing events in intensive management).

From birdlife diversity view point, the first mowing event date could be more important than mowing events frequency along growing season [10]. In our study area, intensive management is characterized by the first mowing event occurring before end of June, whereas extensive management is characterized by the first mowing event occurring after that date. Extensive management practices are beneficial for biodiversity [31, 11] and birdlife [10], and are actively promoted by the European Common Agricultural Policy (CAP) through incentive payment mechanisms.

Table 1: Satellite data preprocessing and derived features according to implemented approach.

Approach	Spectral bands	Features	Preprocessing
Machine learning	B2, B3, B4, B5, B6, B7, B8, B8A, B11, B12	1st derivative	linear interpolation
Threshold-based	-	NDVI	linear interpolation
Threshold-based	-	NDVI	raw data

2.2. Satellite data

All available Sentinel-2 (L2A) surface reflectance images, captured throughout the growing season (from the beginning of January to the end of September 2022) and intercepting mainland France, were used. This dataset comprised ninety tiles and seven of them, intercepting field observation sites, were used for training and testing models (Figure 1). An average of sixty images were available for each tile for the considered temporal period. All spectral bands (except B1, B9 and B10) were used, after resampling 20m resolution bands to 10m resolution[2] to uniform pixel sizes on a common geographical grid.

These images had been preprocessed using MAJA algorithm [33] for atmospheric correction and cloud detection, and were downloaded from THEIA platform (`https://www.theia-land.fr`). All images were provided with a mask layer for clouds and shadows. For each tile, cloud- and shadow-free time series with a regular 10-day time interval were generated using a linear interpolator, as done in [34] or [35]. In addition to spectral bands, we also computed their temporal derivative -using finite differences-, as well as the Normalized Difference Vegetation Index - NDVI [36] (Table 1).

2.3. Reference data

In 2022, the French Biodiversity Agency (`https://www.ofb.gouv.fr`) coordinated an intensive campaign of field observations throughout the mainland French territory, involving local government agencies participating in the

[2]using a bicubic interpolation, as implemented in the Orfeo ToolBox and its SuperImpose application [32]

National Observatory of Mowed Grasslands Ecosystem network. A total of eight sites (from north to south: Marais du Cotentin et du Bessin, Val de Vienne, Sologne Bourbonnaise, Vallée de l'Arconce, Vallée du Drugeon, Haut Jura, Plateau du Mézenc and Planèze - Narse de Lascols) were monitored. They come from four different eco-climatic regions (Figure 1) and they have a significant altitudinal gradient (Table 2). Observations were conducted once a week from May to August, for a total of 2 227 plots, with 1 605 mowed plots and a balanced distribution among sites (Table 2). For each specific site, observed plots were chosen based on accessibility and the local observer's prior knowledge of the area.

Plot boundaries were obtained from the 2020 LPIS. This database provides spatialized information on agricultural plot boundaries and crop types, but does not provide information about management practices. For permanent grasslands, a declared plot can be managed with two practices simultaneously (i.e., spatially separated mowing and grazing within the same plot). Therefore, prior to the field observation campaign, we visually assessed each chosen plot using a national database of aerial imagery (BD ORTHO, `https://geoservices.ign.fr/bdortho`) and Google Earth, to identify, as far as possible, sub-plots with the most homogeneous spatial structure. For each actual plot, a total of eleven observations were conducted throughout the growing season. At each weekly visit, current management practice (mowing or grazing) and the corresponding date were recorded.

Based on these records, each plot was labeled as *mowed*, *grazed* or *mixed* (mowing + grazing). Then, these labeled plots were grouped into two management practice categories: *mowed* -including *mowed* and *mixed* plots (70.5% of plots)- and *unmowed* -including *grazed* plots-. A management practice could be ongoing during the visit or have occurred between the current visit and the previous visit. Consequently, in these *mowed* plots, the observed date for a mowing event may have an uncertainty of few days. Here, 87% of *mowed* plots had one mowing event. The remaining plots had two mowing events and only the first event was used in the experiments.

Table 2: Statistical description of the observed sites. The values represent the number of mowed plots (# plots), average plot area (Av. area), and approximate altitude. Tile column represents Sentinel-2 tile intercepting an observed site.

Site	**Tile**	**# mowed plots**	**Av. area** (Ha)	**Altitude** (m)
Marais du Cotentin et du Bessin	T30UXV	136	1.39	2-50
Val de Vienne	T30TYT	239	1.06	30
Sologne Bourbonnaise	T31TEM	119	2.47	230-280
Vallée de l'Arconce	T31TEM	174	2.23	280-390
Vallée du Drugeon	T31TGM	219	3.87	800-850
Haut Jura	T31TGM	213	2.55	800-950
Plateau du Mézenc	T31TEK	255	1.66	1100-1300
Planèze - Narse de Lascols	T31TDK	217	1.40	1000-1050
Causses du Quercy	T31TCK	33	0.50	309-775
Total		1 605		

Additionally, Causses du Quercy site was included (in the south, Figure 1), where 38 plots were observed with a lower temporal resolution. Here, observations were provided by the local observatory of the *Parc Naturel Régional des Causses du Quercy*, independently of the main observation campaign at the above-mentioned sites. In this site, 87% of observed plots were *mowed*, and all had one mowing event.

A unique site identifier was assigned to plots located within the same Sentinel-2 tile in order to separate training and testing samples based on the tile membership. In the following, the term *site* is used to denote plots belonging to the same Sentinel-2 tile.

The prediction of first mowing event date is done at pixel-level. Thus, all pixels in observed *mowed* plots were selected, and their spectro-temporal profiles were extracted along with their corresponding first mowing event dates, serving as predictor and target values, respectively. Mowing event dates span from May 10th, day of the year (DOY) 130, to August 2nd, DOY 214, comprising a total of 328 451 pixels derived from the 1 605 observed *mowed* plots (Figure 2). 80% of the mowing dates fell between DOY 150 and 188. The remaining occurrences (20%) fell at the extremes of the distribution. The average observed date was June 16th (DOY 167), while the median was June 15th (DOY 166).

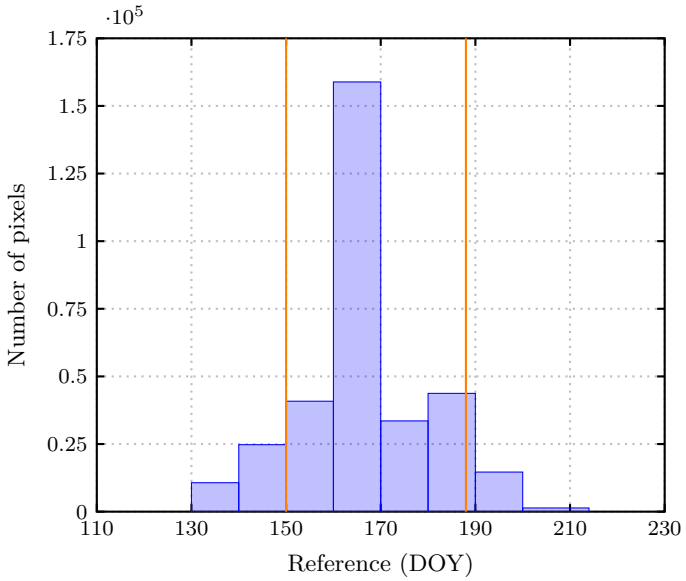


Figure 2: Distribution of pixel-level first mowing event dates observed in *mowed* plots across all sites in 2022. The orange vertical lines indicate the 10^{th} and 90^{th} percentiles of the data, respectively.

3. Methods

3.1. Mowing events prediction

Several supervised regression models were investigated, from conventional machine learning to recent deep-learning ones, and unsupervised threshold-based methods. Prediction performances against ground observations were then compared. Following Fauvel et al. [28], we set up a spatial cross-validation to estimate the prediction accuracy (Figure 3). All observations from a site were excluded from reference data before training/calibrating models. Then, models were tested and assessed using the excluded site-specific observations. In other words, all observations from excluded site were used as testing data, and all observations from non-excluded sites were used as training data. This split was repeated seven times, so that each site was excluded once and considered as testing data (i.e., 7-fold spatial cross-validation). We computed average prediction accuracy using all sites scores (spatial folds), with each site's score being estimated as the average of individual evaluations on fifty bootstrap test set.

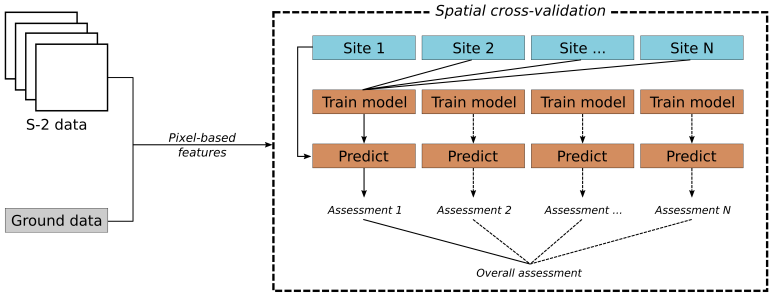


Figure 3: Workflow for predicting first mowing event date. The box in dashed lines represents the spatial cross-validation approach implemented in this study. Here, the different steps (model training, prediction and assessment) implemented for a given site (spatial fold) are illustrated. The solid lines represent implementation for site 1, while dashed lines represent implementation for all remaining sites. In this example, all observations from site 1 were used as testing data, while all observations from the remaining sites were used as training data.

A comprehensive overview of algorithms used to predict the first mowing event date is given in the following.

3.1.1. Machine learning approach

We implemented five supervised regression models from the literature: two conventional machine learning models -Random Forest (RF) and Ridge Regression- and three cutting-edge deep-learning architectures -Multilayer Perceptron (MLP), 1D Convolutional Neural Network (1D-CNN, [37]) and Lightweight Temporal Attention Encoder (LTAE, [38])-.

Ridge Regression and MLP were used as a baseline, while RF was chosen due to its good behavior in large-scale prediction [34, 28]. 1D-CNN and LTAE were selected for their capacity to model temporal information, leveraging convolutional techniques and attention mechanisms, respectively. A brief review of these models is given in section Appendix A, while their hyperparameters are presented in Table 3.

Deep-learning models were trained on 200 epochs with a batch size of 4 096, using the Adam optimization algorithm [39]. In learning process, 10% of training data were used to form a validation set, which was used to perform early stopping and to reduce the learning rate by a factor of 10 when learning

stagnated[3].

For all supervised models, oversampling techniques for minority ranges of mowing dates were also investigated. More specifically, two oversampling techniques for classification problems were tested: SMOTE [40] and ADASYN [41] algorithms, relying on a convex combination of existing samples. Since these oversampling techniques were defined for classification problem, we created 10 *fake* classes by dividing the interval of mowing date values into 10 sub-intervals of equal width and assigned each pixel to a class corresponding to the number of the interval in which its label fell[4]. For each oversampling technique, we used the implementation provided by the `imbalanced-learn` library [42]. All classes were oversampled, except for the majority class, to obtain an equal number of samples in each class.

Lastly, in order to quantify the learning capacity of the above machine learning algorithms, we also reported results for naive regressor[5], that makes prediction using a simple rule: predict the mean value of the training data. It can be seen as a linear model with only an intercept parameter. Theoretically, such model should have a R^2 equal to zero. It is called SimpleMean in the following.

3.1.2. Threshold-based approach

Threshold-based methods are well-known in vegetation dynamics studies, and were widely used in mowing event frequency and timing detection [43]. We implemented a recent specific mowing event detection algorithm introduced by Vroey et al. [20] as an integral monitoring tool within Sen4CAP program (`http://esa-sen4cap.org`). The main idea developed in [20] was to quantify temporal loss of NDVI, and to consider a mowing event when this loss is higher than a threshold value. In our study, the threshold value was set automatically using grid-search on training data (Table 3). Two types of thresholds methods

[3] *Reduce on plateau* strategy implemented here `https://pytorch.org/docs/stable/generated/torch.optim.lr_scheduler.ReduceLROnPlateau.html`

[4] Other number of bins were investigated, as well as different binning strategy, providing similar or worst results. For clarity we only reported results for 10 equal bins.

[5] https://scikit-learn.org/stable/modules/generated/sklearn.dummy.DummyRegressor.html

Table 3: Algorithm-specific hyperparameters. The **Value** column reports the selected value or the search range for the algorithm, with the following notation `start:step:end`. For Ridge Regression and threshold methods, cross-validation was used to select the best value.

Algorithm	**Hyperparameters**	**Value**	**Package**
Random Forest	Number of trees	100	Scikit-Learn
Rigde Regression	Regularization	1000:500:15500	Scikit-Learn
1D-CNN	Learning rate	1e-3	Pytorch
LTAE	Learning rate	1e-3	Pytorch
MLP	Learning rate	1e-4	Pytorch
Fixed threshold	Minimum loss of NDVI	0.10:0.01:0.40	
Relative threshold	Minimum loss of NDVI	10:5:50 %	

were used:

1. A fixed threshold corresponding to fix loss of NDVI. Threshold value is expressed in NDVI.

2. A relative threshold corresponding to relative loss of NDVI, taking into account the pixel amplitude (minimum and maximum value). Threshold value is expressed as a percentage of the amplitude.

It was adapted to detect only the first mowing event date, since it was primarily designed to detect mowing event time interval. The main differences compared to original algorithm are detailed in section Appendix B.

Threshold-based algorithms were calibrated and tested using the same training and testing data used for machine learning-based algorithms, respectively.

3.2. Assessment of mowing events

The deviation between predicted and observed first mowing dates was assessed using four standard metrics: Mean Absolute Error (MAE), Root Mean Square Error (RMSE), Max error and the coefficient of determination (R^2), defined as:

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}(|y_i - \hat{y}_i|),$$
$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2},$$
$$\text{Max error} = \max_{i\in\{1,\ldots,n\}}(|y_i - \hat{y}_i|),$$
$$\text{R}^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2},$$

where $\hat{y}_i$ and y_i are predicted and observed first mowing dates at pixel i, respectively, n is the number of pixels in testing data. In R^2 formula, $\bar{y}$ is the average of observed dates in testing data. The MAE, RMSE and Max error should be minimized while the R^2 should be minimized.

4. Results

In this section, a comprehensive overview of the quantitative results derived from all implemented algorithms is presented. We further provide a more detailed analysis of the results obtained for the best model. Finally, a qualitative assessment of the prediction map is conducted.

4.1. Evaluation of algorithms for mowing events prediction

Figure 4 shows best results obtained for each model with and without sample augmentation. Results for all models and configurations can be found in Appendix D, Table D.4. Results were averaged over the 7-fold spatial cross-validation runs, as detailed in 3.1.

From Figure 4, non-linear machine learning models obtained the best performances, by a clear margin, with a R^2 above 0.4. Linear models reached lower accuracy, Ridge Regression model being slightly better than SimpleMean, as expected. Worst results, for any quality index, were obtained for threshold-based

methods. Among non-linear models, deep-learning ones (LTAE, 1D-CNN and MLP) yielded higher performances compared to conventional RF, and architectures that take into account the temporal dimension (LTAE and 1D-CNN) were the most accurate models, with a slight advantage for LTAE. Results in terms of MAE followed the same trend for R^2.

Oversampling techniques improved accuracy only for MLP; other models did not show notable accuracy gains with augmented data. Additionally, there was no significant difference in accuracy improvement between SMOTE and ADASYN. Oversampling did not improve predictions for extreme mowing dates (early/late), as shown in the next section.

In terms of maximum error, among the best algorithms, 1D-CNN demonstrated the lowest values, followed by MLP and LTAE. While LTAE offered the best predictions on average, it occasionally resulted in higher errors in certain cases.

Figure D.11 in Appendix D presents the prediction accuracy at the site level (spatial fold). Within-site variability can be observed, and LTAE did not offer the highest accuracy for each site. Yet, on average, LTAE provided the best results for MAE, RMSE or R^2. In the next section, we provide a more detailed analysis of LTAE predictions.

4.2. Mowing events prediction across sites for LTAE

Figure 5 shows the joint density between predicted and observed dates for each site, along with the marginal density for both predicted and observed dates, as well as for the corresponding training data. On average, predictions were accurate since the modes of the joint density were on the identity line. Yet, for smaller R^2 (T30UXV, T30TYT and T31TEM), we observed a clear overestimation of early dates (predicted dates were later than observed ones), as well as an underestimation of late dates (predicted dates were earlier than observed ones).

The possible low number of clear dates in optical SITS is usually a factor explaining the performances of prediction methods. Figure 6 shows MAE

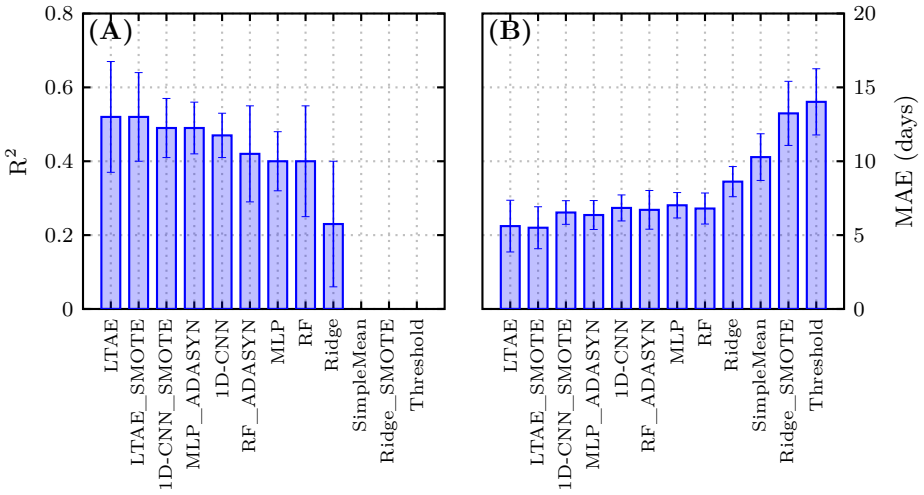


Figure 4: Algorithm-specific statistical summary in terms of (A) R^2 and (B) MAE. The values represent weighted means of all sites. A site-specific score was weighted using the number of pixels used for the evaluation. A site-specific score represents the mean of fifty individual evaluations (from 50 folds by bootstrapping 70% of observations). Here, values less than zero are not shown.

distribution as a function of the number of clear dates in a temporal window surrounding the event[6]. The worst average MAE was obtained for very low number of clear dates (0 to 2). Once the number of clear dates reached 3, MAE remained low, with a median value below 7 days. Interestingly, a slight degradation of MAE was observed when the number of clear dates is greater than 8.

Figure 7 shows pixel-wise average of LTAE attention score. The attention score reflects which part of the temporal signal was used for the prediction, and it is computed at pixel-level. Implemented LTAE used 4 heads, i.e., 4 ranges of the temporal signal can be selected. For the first attention head, selected observations -from interpolated SITS- were located in the month of July, whatever the predicted date. We observed more variation for the three other heads. For the second attention head, selected observations were distributed between April

[6]0 clear dates means that the 10-day linearly interpolated SITS used for prediction was constructed using Sentinel-2 acquisitions outside the temporal window.

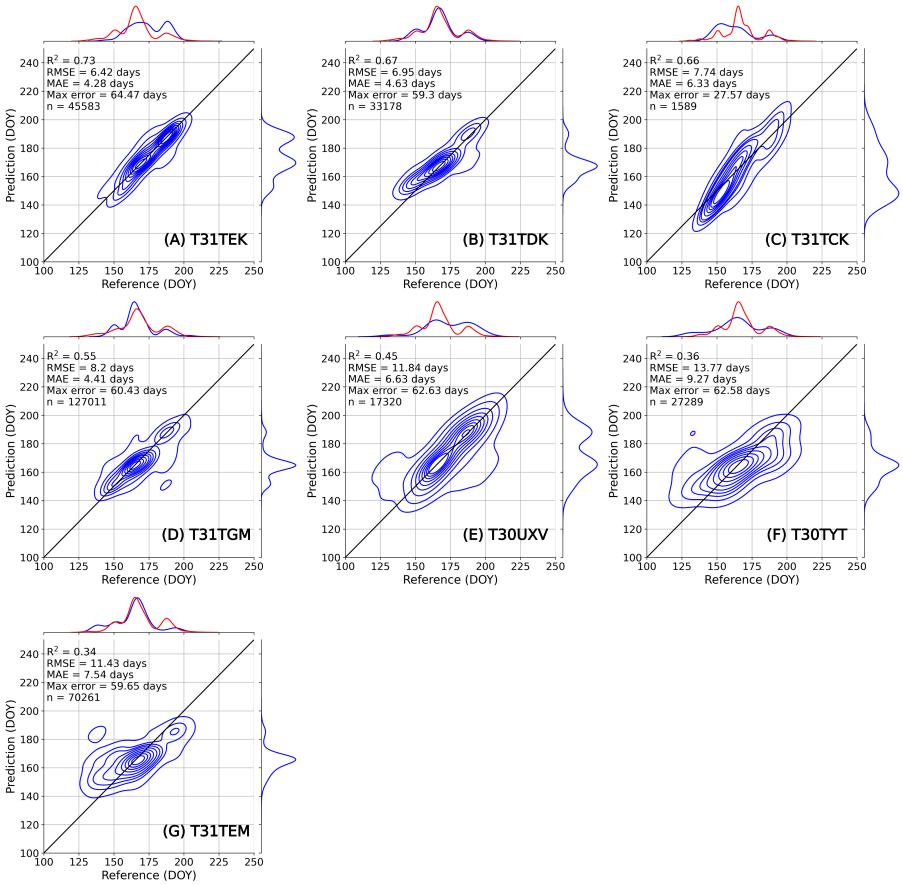


Figure 5: Site-specific LTAE pixel-level predictions. For each graph, horizontal axis represents the reference datasets (testing in blue and training in red), while the vertical axis represents the prediction. Each graph corresponds to a site (spatial fold). For instance, in (A), T31TEK testing data included all samples from that specific site, while training data included all samples from all remaining sites, as discussed in section 3.1. Predicted and observed dates are expressed in days of year (DOY). Joint and marginal densities were computed using Scipy `gaussian_kde` function [44].

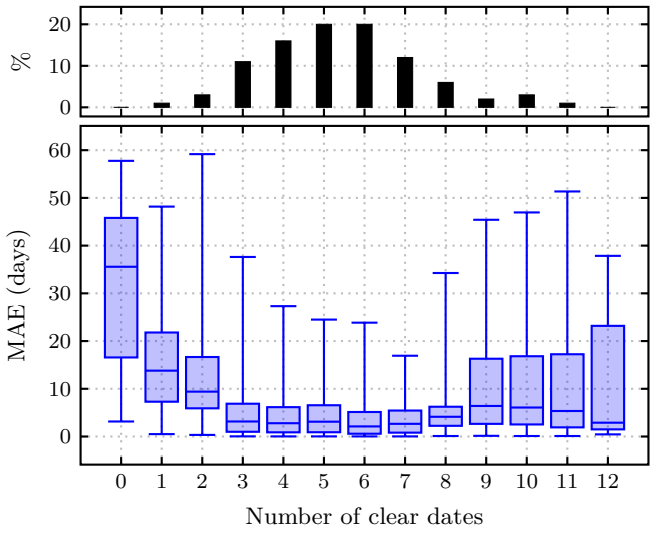


Figure 6: MAE distribution as a function of the number of clear dates in a temporal window surrounding the event (in blue). The temporal window was defined as $[y-20$ days, $y+20$ days] with y the observed mowing date, and MAE is defined in 3.2. The number of clear dates can reach up to 12 within a 40-day period due to Sentinel-2 orbit overlap. Frequency of each category (number of clear dates), expressed as a percentage of the total number of samples, is shown in black.

and June (included). For the third attention head, selected observations were located between February and June, for predicted early dates; while between June and September for predicted late dates. For the last attention head, selected observations were distributed between April and June. Overall, no clear pattern was observed linking attention scores and predicted dates.

4.3. Spatialized prediction analysis for LTAE

In this section, the prediction map was analyzed qualitatively with a particular focus on the spatial distribution of intra-plot predictions. For each spatial cross-validation fold (testing data), an individual prediction map was generated for each observed plot. All plot-level prediction maps are fully available for visualization in the supplementary material called `site_predictions.pdf`, which contains prediction for LTAE, 1D-CNN and RF. After a careful visual analysis of LTAE predictions, four main cases were identified and illustrated in Figure 8.

1. In most of the plots, intra-plot predictions were homogeneous showing a

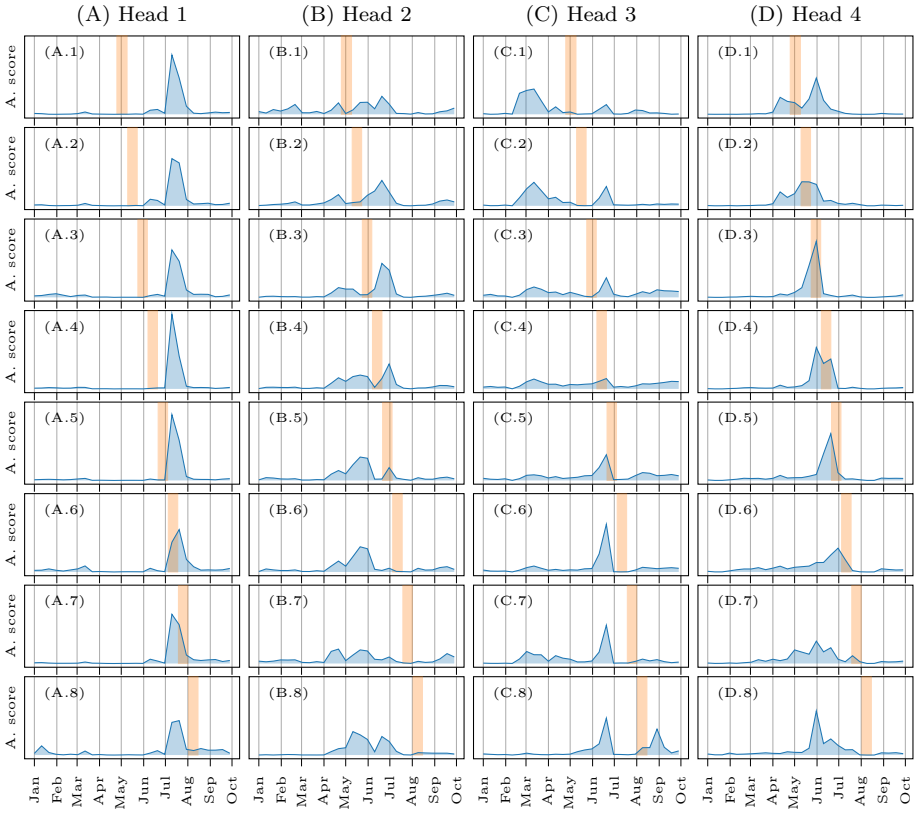


Figure 7: LTAE attention mechanism across 4 heads. For each observation in the interpolated SITS (x-axis), attention score (y-axis) was computed at pixel-level; and the average attention score from all pixels -in testing data- is depicted in blue. SITS observations with high attention score were the most significant for the prediction process. Each column (A-D) represents a specific attention head, while each row (1-8) represents a specific range of predicted dates (in orange).

coherence in terms of spatial configuration and were in agreement with the reference, as illustrated in the first row (A.1 to A.3) of Figure 8.

2. For some plots, predictions exhibited two distinct spatial patterns within a given plot. Usually, one spatial pattern matched with the reference and the other one did not. This case is illustrated in the second row (B.1 to B.3) of Figure 8.

3. Occasionally, heterogeneous intra-plot predictions were obtained with a global disagreement with the reference. This situation is shown in the third row (C.1 to C.3) of Figure 8.

4. Finally, homogeneous intra-plot predictions with large deviation from the reference can be found. For most of such plots, it corresponded to early/late mowing dates, as it is displayed in the last row (D.1 to D.3) of Figure 8.

The two last cases occurred most often at site T31TEM, where prediction accuracy was the lowest (R^2=0.34 and MAE=7.54 days).

4.4. Mowing events prediction across mainland France

The LTAE model was selected to generate a prediction map across mainland France, because of its performances compared to others models previously evaluated in this study. All referenced samples (328 451 pixels across seven sites) were used to train model. Then the learned model was used to predict the first grassland mowing date for all *mowed grassland* pixels. Such pixels corresponds to pixels classified as *mowed* in a grassland management map, as described in Appendix C. The prediction map is freely accessible at `https://zenodo.org/records/11034387`. An overview of the map is given in Figure 9.

It is difficult to assess the quality of the map at large-scale. Accuracy metrics and spatial homogeneity have been reported in the previous section for the 7 tiles. As a sanity check, mowing dates for 45 plots (corresponding to 27 478

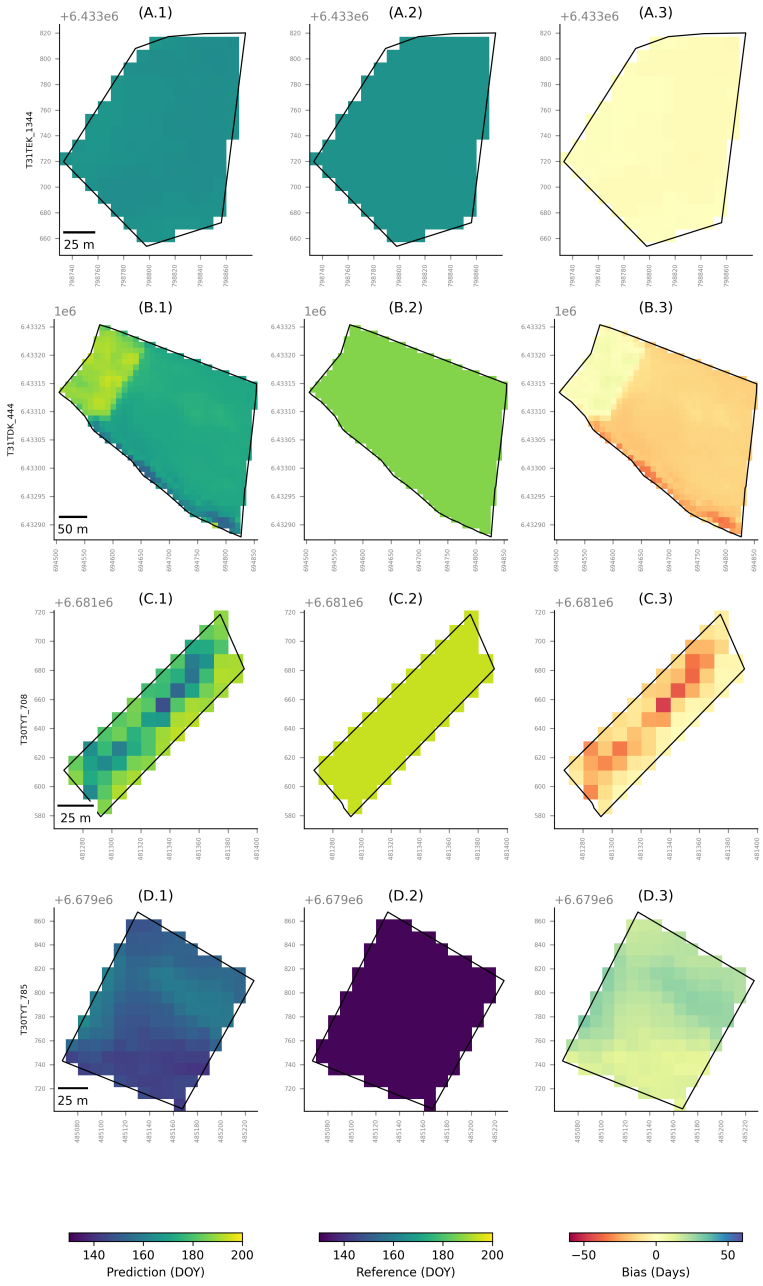


Figure 8: Intra-plot LTAE outputs. The columns (1-3) represent prediction, reference and bias (prediction − reference) values, respectively. The predicted and observed dates are expressed in Days Of Year (DOY), and bias is expressed in days. A negative bias indicates the predicted date was earlier than the observed date, while a positive bias means the predicted date was later than the observed date. The rows (A-D) represent four selected plots (T31TEK_1344, T31TDK_444, T30TYT_708 and T30TYT_785).

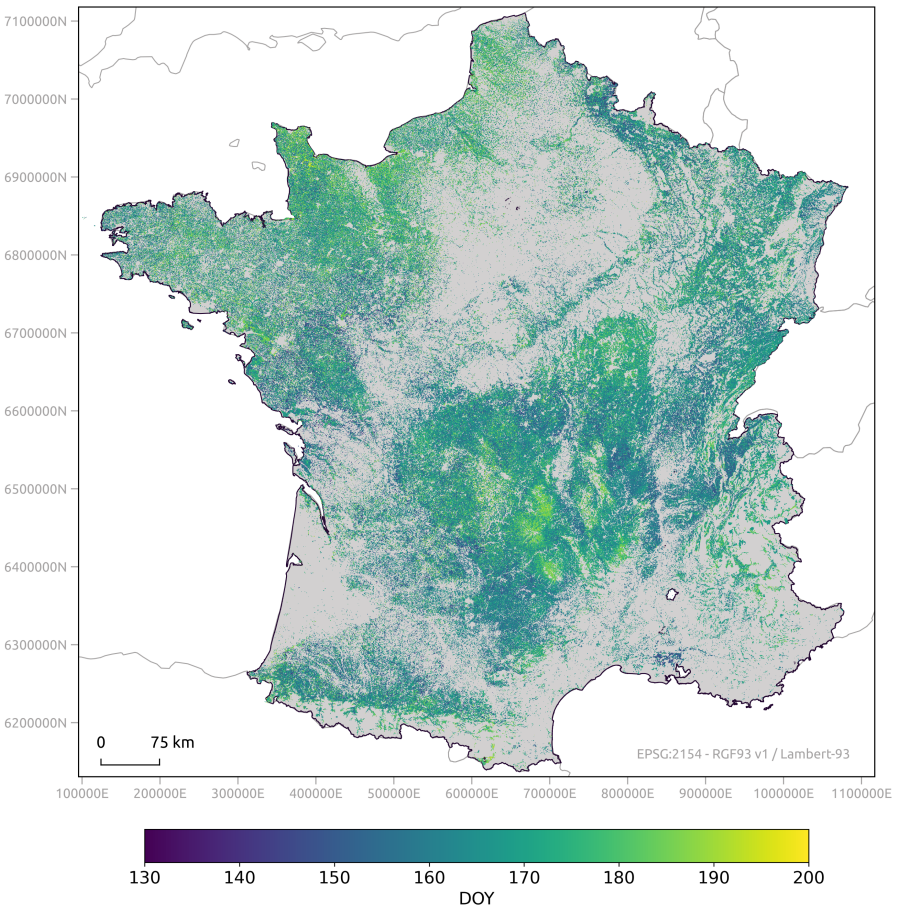


Figure 9: Map of grassland first mowing date in mainland France in 2022. This map was generated by combining LTAE model and Sentinel-2 time series.

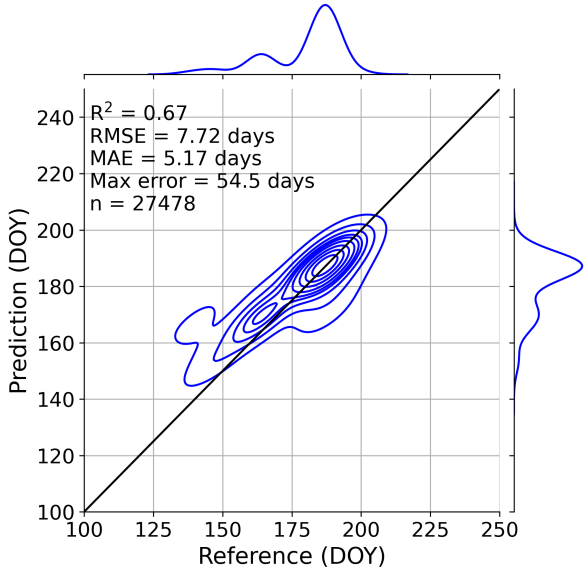


Figure 10: Comparison between predicted (y-axis) and observed (x-axis) dates from the two independent sites. Observations from both sites were concatenated and are not shown separately in this graph. Predicted and observed dates are expressed in Days Of Year (DOY). The number of pixels used for the evaluation is represented by n. Joint and marginal densities were computed using Scipy `gaussian_kde` function [44].

pixels) have been extracted from an experimental research platform[7] located in the tile T31TDL not seen during the training process, for the same year. These are two experimental sites (Laqueuille and Marcenat) and are actively monitored by local research units. Observed dates were the true mowing date (without uncertainty).

Prediction accuracy is given in the Figure 10. Numbers were in agreement with those of the previous section. In particular, the tile T31TDL is spatially closed to the tile T31TDK and the accuracy metrics were similar.

[7]INRAE-Herbipole experimental farm `https://doi.org/10.15454/1.5572318050509348E12`

5. Discussion

5.1. Machine learning algorithms

Deep-learning models shown superior results in terms of prediction accuracy. Contrary to RF, they have the ability to automatically learn data representations at multiple levels of abstraction through the use of multiple processing layers [45], which could enable them to capture patterns within the data more effectively compared to conventional machine learning models. In mowing event detection, Holtgrave et al. [25] have also reported this superior capability of deep-learning models. In our study, architectures that exploited temporal dimension (LTAE and 1D-CNN) demonstrated higher performance compared to a baseline deep-learning model (MLP). This temporal information's contribution to improving mowing events prediction was also observed in Komisarenko et al. [23], Lobert et al. [24]. This could be due to their ability to capture temporal patterns and dependencies (e.g., biomass variation) within the time series, as already noted in other applications [46, 37, 47, 48]. However in our study, temporal derivatives of each spectral band were provided as additional feature to each algorithm, thus LTAE and 1D-CNN may capture other temporal dependency than first derivative.

Attention scores of LTAE, show in Figure 7, tend to indicate that acquisition between April and September (included) are the most important for the prediction, whatever the mowing date to be predicted. From the first head, it seems that the month of July contains very predictive information, since it is used for every range of prediction. To a lesser extent, May and June data are also very important for the prediction. If early dates are considered, one attention head use acquisition from mid-February to end of March. However, this observation should be mitigated with the limited accuracy associated to this range of dates.

Threshold-based method exhibited the worst performance in detecting first mowing event date compared to all supervised models. Methods based on expert rules are sensitive to local conditions (pedoclimatic and management) and are difficult to generalize. Additionally, these methods are particularly affected

by the lack of cloud-free observations during the event, leading to significant detection uncertainties or non-detection, as mentioned in many studies related to detection of agricultural practices or phenological events [49, 50, 51, 52]. The lack of cloud-free observations during the event is an inherent limitation of optical remote sensing data (it can be clearly visualized in the supplementary material called `site_ndvi_sits.pdf`). Although this affects both thresholding and machine learning methods, the latter have the ability to learn from the available data (with or without cloud-free observations during the event) which could overcome this limitation to some extent.

5.2. Accuracy of the mowing events prediction

Uncertainties were more important when predicting early and late mowing dates. A clear trend in overestimating early mowing dates as well as underestimating late mowing dates was observed for each model. Owing to the reference data distribution (see Figure 2) early/late mowing dates were underrepresented and therefore more difficult to learn for the supervised models. Yet rare, maximum errors were obtained for such extreme dates.

Oversampling techniques were investigated in this study to alleviate such imbalanced issues in the training data [53]. Such approaches have found to be useful in classification problems [54, 55] and also for remote sensing data [56] to construct well balanced training data. However, these techniques did not help to reduce this issue in our large-scale regression scenario. Only MLP model exhibited an improvement when oversampling data, but its accuracy remained lower than those obtained with time aware deep-learning models.

5.3. Influence of the number of clear dates on prediction accuracy

The number and distribution of cloud-free satellite observations during the growing season are critical for reliably detecting patterns of change, mainly during the event of interest. Cloud cover modulates optical satellite data availability, resulting in dense or sparse time series according to location. Temporal gap in optical time series affects both the threshold method and the supervised

models performances. In this paper, models' underperformance could also be attributed to persistent cloud cover during the event.

Overall, the highest prediction errors were observed when the number of cloud-free satellite observations was limited around the true mowing date (i.e., 0-2 clear dates), but they decreased as the number of clear dates increased (i.e., 3-8 clear dates), as illustrated in Figure 6 and in agreement with Schwieder et al. [19], Kolecka et al. [15], Komisarenko et al. [23]. Surprisingly, prediction errors increased slightly on average when the number of clear dates increased (i.e., 9-12 clear dates). We explained this phenomenon by factors related to the temporal gap filling. When the number of clear dates was higher than 8, the data were coming from two overlapping orbits. Images from each orbit have different acquisition angles which can induce variations in the signal. Furthermore, even with more temporal acquisition, the linear gap filling algorithm only uses the two closest post/past dates to perform the temporal interpolation. With very closed dates, the smoothing effect of the linear filter is limited and more noise can be included in the interpolated time series[8].

Usually, main strategies implemented to cope with sparse time series are (i) combining optical data from different sensors and (ii) combining optical and radar data. Regarding combining optical data from different sensors, Harmonized Landsat and Sentinel-2 (HLS) data (30 m, 3–4-day revisit) [57] have already yielded promising results in vegetation monitoring [52, 16], where Sentinel-2 data alone omitted events due to sparse cloud-free observations [52]. Schwieder et al. [19] combined Sentinel-2 and Landsat-8 EVI time series to map successfully mowing events in Germany. As an alternative to sensor combination, PlanetScope data (3m, daily revisit) have also demonstrated great potential in vegetation monitoring [58, 59], which could be useful in mowing detection. Yet, these data are commercial and could be a limitation for large-scale implementa-

[8]Site-specific prediction maps can be accessed in supplementary material called `site_predictions.pdf`, while NDVI image time series in `site_ndvi_sits.pdf`. For clarity, illustrated NDVI image time series were generated with ten satellite observations, five before and five after the event. A *readme.txt* file is also available and provides more details on these figures.

tion in support of public policy. Regarding combining optical and radar data, implemented strategies are diverse. Usually, when using threshold-based methods, optical and radar data are separately used, but their results are merged [20, 21]. This leads to a decrease in omission of events but may increase false detections due to the noisy character of radar data. Technically, when using supervised models, optical and radar time series are conventionally concatenated, and could improve models' performances [24, 23, 25]. Garioud et al. [22] presented a different approach, where Sentinel-1 and -2 time series as well as climatic and topographic data were used to reconstruct continuous Sentinel-2 NDVI time series for mowing date estimation. Although the above-mentioned studies agreed that combining optical and radar data globally improved mowing event detection, factors such as topography, soil moisture and vegetation water content influence the results locally [27].

From an operational view point, including Sentinel-1 SITS in the processing comes with a high computational burden, since Sentinel-1 and Sentinel-2 images need to be projected on the same pixels grid and Sentinel-1 SITS required additional noise filtering. In mainland France, very cloudy pixels are rare (less than 5% of the pixels) and it is questionable that the effort is worth it. Uncertainties in predicting extreme mowing dates (early/late) are more critical for a birdlife monitoring perspective.

5.4. Importance of the field survey

In our study, field observation campaign conducted on 2 265 plots (section 2.3) aimed to represent the greatest possible diversity of mowing dates as well as site conditions (i.e., altitudinal gradient, flooded and dry grassland, early and late mowing dates, etc.), which required a great human and time effort. However, despite this effort, the representativeness of the reference data needs to be improved, mainly for minority ranges of mowing dates. In this context, efforts to build a more diversified reference dataset could be oriented towards a citizen/collaborative science, involving citizens and established observation networks, such as the emerging initiatives of the French Biodiver-

sity Agency[9], or those already consolidated in similar topics for phenological observations (e.g., French Phenology Network-TEMPO[10], National Phenology Network-USA/NPN[11], UK Phenology Network[12], Pan-European Phenological Network-PEP725[13], etc.).

As shown in Figure 8-B.(1-3), some grasslands have bimodal mowing management. Such situation was not considered neither in the learning nor in the validation processes. This inaccurate/incomplete reference map is mainly due to inaccessibility to the plot or non-visibility of the intra-plot bimodal mowing management during the observer's visit. The LTAE model was able to reliably capture the intra-plot mowing management, despite it was learned with few label noise (i.e. wrong mowing date). Yet rare, this situation was not taken into account in the validation process.

Regarding accuracy of observed date, revisit frequency depends on the observation campaign protocol and the observer's availability. Here, observed dates included an average uncertainty of seven days, due to weekly revisits. Consequently, this uncertainty in the reference data is involved in the learning process, which could partially degrades the performance of the model. A more reliable approach would be to use the true mowing date rather than observed mowing date. Such field data is inaccessible by field survey unless farmers voluntarily declare and share it, which could be raised under public policies such as the CAP (`https://agriculture.ec.europa.eu`) in Europe for example. Hence, in our study, supervised models were trained to predict the observed date which is not necessarily the actual date of the event, but a close approximation of it. We believe this a cost to pay for covering significant heterogeneous grassland practices. It should be note that the quality of the prediction was good, and of the same magnitude, for the data set with the true mowing date, see section 4.4.

[9] `https://www.ofb.gouv.fr`
[10] `https://tempo.pheno.fr`
[11] `https://www.usanpn.org`
[12] `https://naturescalendar.woodlandtrust.org.uk`
[13] `http://www.pep725.eu`

It indicates that the field survey uncertainty is not critical.

6. Conclusion

This paper focused on nationwide mapping of grassland first mowing event date, combining machine learning and Sentinel-2 time series. Among implemented algorithms, non-linear machine learning models obtained higher performances compared to linear ones. Deep-learning models yielded higher performances compared to conventional Random Forest model, and time aware architectures (LTAE and 1D-CNN) were the most accurate models. Overall, uncertainties were accentuated at extreme mowing dates (early/late), which were underrepresented in the reference data. Oversampling techniques demonstrated no significant improvement in predicting these extreme mowing dates, except for MLP. Regarding transferability, LTAE model exhibited reliable performance across all spatial folds, but accuracy may degrade for area underrepresented in training process. In our study, LTAE model's best prediction occurred when the number of clear dates was greater than 2 within a 40-day temporal window surrounding the mowing event (which occurred in more than 95% of tested samples).

End-to-end learning such as in [60] has shown significant improvement for classification purpose. Rather than interpolating the data before and independently of the learning process, the reconstruction is learned jointly with the classification task. Such approach should be considered in the future, either for a mono-sensor or multi-sensor, as it has outperformed common "reconstruct" then "learn" strategy discussed in previous studies. We expect it will leverage the slight performance loss observed when the number of clear dates was high.

Our findings evidenced time aware deep-learning models' potential to nationwide grassland mowing monitoring. Although our approach should be adapted for predicting all mowing events during growing season, predicted first mowing event date is a key indicator of plot management intensity; and could support birdlife monitoring or public policies for biodiversity and agro-ecological transi-

tion in France. For long-term monitoring, LTAE transferability into an unknown year needs to be investigated.

7. Acknowledgement

The authors would like to thank OFB's fieldworker (departmental services 03, 14, 15, 25, 37, 39, 43, 50, 71) and Permanent Center for Environmental Initiatives of Val de Vienne for the field campaign, the INRAe-Herbipôle Team for sharing their data set, the Theia Center for making available the Sentinel-2 SITS, the CNES for funding H. Touchais and H. Rivas work and for providing the HPC facility. M. Fauvel wants to thank the CESBIO AI Team for fruitful discussions and the IOTA^2 core team for their helps, and the French Biodiversity Agency (OFB) for co-funding this project.

8. Supplementary materials

The supplementary files `site_predictions.pdf` and `site_ndvi_sits.pdf` are available at the following link `https://zenodo.org/records/11034843`. The python scripts implementing the different algorithms are available at `https://src.koda.cnrs.fr/pzzkfwbr/mowing-event-detection`. Finally, the prediction raster map, as well as per French county vector files (plots summary) are provided here: `https://zenodo.org/records/11034387`.

References

[1] J. M. Suttie, S. G. Reynolds, C. Batello, Grasslands of the World, volume 34, Food & Agriculture Org., 2005.

[2] R. P. White, S. Murray, M. Rohweder, S. Prince, K. Thompson, et al., Grassland ecosystems, World Resources Institute Washington, DC, USA, 2000.

[3] Y. Zhao, Z. Liu, J. Wu, Grassland ecosystem services: a systematic review of research advances and future directions, Landscape Ecology 35 (2020) 793–814.

[4] J. Bengtsson, J. M. Bullock, B. N. Egoh, C. S. Everson, T. Everson, T. O'Connor, P. J. O'Farrell, H. G. Smith, R. Lindborg, Grasslands—more important for ecosystem services than you might think, Ecosphere (2019). doi:10.1002/ecs2.2582.

[5] N. Klein, C. Theux, R. Arlettaz, A. Jacot, J. Pradervand, Modeling the effects of grassland management intensity on biodiversity, Ecology and Evolution (2020). doi:10.1002/ece3.6957.

[6] Y. Yang, D. Tilman, G. N. Furey, C. Lehman, Soil carbon sequestration accelerated by restoration of grassland biodiversity, Nature Communications (2019). doi:10.1038/s41467-019-08636-w.

[7] J. Huang, J. Zhan, H. Yan, F. Wu, X. Deng, Evaluation of the impacts of land use on water quality: A case study in the chaohu lake basin, The Scientific World Journal (2013). doi:10.1155/2013/329187.

[8] X. Wang, F. Y. Li, K. Tang, Y. Wang, S. Guga, Z. Bai, T. Baoyin, Land use alters relationships of grassland productivity with plant and arthropod diversity in inner mongolian grassland, Ecological Applications (2019). doi:10.1002/eap.2052.

[9] E. Metera, T. Sakowski, K. Sloniewski, B. Romanowicz, Grazing as a tool to maintain biodiversity of grassland - a review, Animal Science Papers and Reports 28 (2010) 315–334.

[10] J. Broyer, L. Curtet, M. Boissenin, Does breeding success lead meadow passerines to select late mown fields?, Journal of Ornithology 153 (2012) 817–823.

[11] P. Buri, J. Humbert, M. Stańska, I. Hajdamowicz, E. Tran, M. H. Entling, R. Arlettaz, Delayed mowing promotes planthoppers, leafhoppers and spi-

ders in extensively managed meadows, Insect Conservation and Diversity (2016). doi:10.1111/icad.12186.

[12] L. Kooistra, K. Berger, B. Brede, L. V. Graf, H. Aasen, J.-L. Roujean, M. Machwitz, M. Schlerf, C. Atzberger, E. Prikaziuk, D. Ganeva, E. Tomelleri, H. Croft, P. Reyes Muñoz, V. Garcia Millan, R. Darvishzadeh, G. Koren, I. Herrmann, O. Rozenstein, S. Belda, M. Rautiainen, S. Rune Karlsen, C. Figueira Silva, S. Cerasoli, J. Pierre, E. Tanır Kayıkçı, A. Halabuk, E. Tunc Gormus, F. Fluit, Z. Cai, M. Kycko, T. Udelhoven, J. Verrelst, Reviews and syntheses: Remotely sensed optical time series for monitoring vegetation productivity, Biogeosciences 21 (2024) 473–511. URL: https://bg.copernicus.org/articles/21/473/2024/. doi:10.5194/bg-21-473-2024.

[13] S. Reinermann, S. Asam, C. Kuenzer, Remote sensing of grassland production and management—a review, Remote Sensing (2020). doi:10.3390/rs12121949.

[14] S. Estel, S. Mader, C. Levers, P. H. Verburg, M. Baumann, T. Kuemmerle, Combining satellite data and agricultural statistics to map grassland management intensity in europe, Environmental Research Letters (2018). doi:10.1088/1748-9326/aacc7a.

[15] N. Kolecka, C. Ginzler, R. Pazúr, B. Price, P. H. Verburg, Regional scale mapping of grassland mowing frequency with sentinel-2 time series, Remote Sensing (2018). doi:10.3390/rs10081221.

[16] P. Griffiths, C. Nendel, J. Pickert, P. Hostert, Towards national-scale characterization of grassland use intensity from integrated sentinel-2 and landsat time series, Remote Sensing of Environment (2020). doi:10.1016/j.rse.2019.03.017.

[17] F. Stumpf, M. K. Schneider, A. Keller, A. Mayr, T. Rentschler, R. Meuli, M. E. Schaepman, F. Liebisch, Spatial monitoring of grassland manage-

ment using multi-temporal satellite imagery, Ecological Indicators (2020). doi:10.1016/j.ecolind.2020.106201.

[18] C. Watzig, A. Schaumberger, A. Klingler, A. Dujakovic, C. Atzberger, F. Vuolo, Grassland cut detection based on sentinel-2 time series to respond to the environmental and technical challenges of the austrian fodder production for livestock feeding, Remote Sensing of Environment (2023). doi:10.1016/j.rse.2023.113577.

[19] M. Schwieder, M. Wesemeyer, D. Frantz, K. Pfoch, S. Erasmi, J. Pickert, C. Nendel, P. Hostert, Mapping grassland mowing events across germany based on combined sentinel-2 and landsat 8 time series, Remote Sensing of Environment (2022). doi:10.1016/j.rse.2021.112795.

[20] M. D. Vroey, L. D. Vendictis, M. Zavagli, S. Bontemps, D. Heymans, J. Radoux, B. Koetz, P. Defourny, Mowing detection using sentinel-1 and sentinel-2 time series for large scale grassland monitoring, Remote Sensing of Environment (2022). doi:10.1016/j.rse.2022.113145.

[21] S. Reinermann, U. Gessner, S. Asam, T. Ullmann, A. Schucknecht, C. Kuenzer, Detection of grassland mowing events for germany by combining sentinel-1 and sentinel-2 time series, Remote Sensing 14 (2022). doi:10.3390/rs14071647.

[22] A. Garioud, S. Giordano, S. Valero, C. Mallet, Challenges in Grassland Mowing Event Detection with Multimodal Sentinel Images, in: 2019 10th International Workshop on the Analysis of Multitemporal Remote Sensing Images (MultiTemp), IEEE, Shanghai, France, 2019, pp. 1–4. URL: https://hal.science/hal-02387167. doi:10.1109/Multi-Temp.2019.8866914.

[23] V. Komisarenko, K. Voormansik, R. Elshawi, S. Sakr, Exploiting time series of sentinel-1 and sentinel-2 to detect grassland mowing events using deep learning with reject region, Scientific Reports (2022). doi:10.1038/s41598-022-04932-6.

[24] F. Lobert, A.-K. Holtgrave, M. Schwieder, M. Pause, J. Vogt, A. Gocht, S. Erasmi, Mowing event detection in permanent grasslands: Systematic evaluation of input features from sentinel-1, sentinel-2, and landsat 8 time series, Remote Sensing of Environment (2021). doi:10.1016/j.rse.2021.112751.

[25] A.-K. Holtgrave, F. Lobert, S. Erasmi, N. Röder, B. Kleinschmit, Grassland mowing event detection using combined optical, sar, and weather time series, Remote Sensing of Environment (2023). doi:10.1016/j.rse.2023.113680.

[26] A. Taravat, M. P. Wagner, N. Oppelt, Automatic grassland cutting status detection in the context of spatiotemporal sentinel-1 imagery analysis and artificial neural networks, Remote Sensing 11 (2019). URL: https://www.mdpi.com/2072-4292/11/6/711. doi:10.3390/rs11060711.

[27] M. D. Vroey, J. Radoux, P. Defourny, Grassland mowing detection using sentinel-1 time series: Potential and limitations, Remote Sensing (2021). doi:10.3390/rs13030348.

[28] M. Fauvel, M. Lopes, T. Dubo, J. Rivers-Moore, P.-L. Frison, N. Gross, A. Ouin, Prediction of plant diversity in grasslands using sentinel-1 and -2 satellite image time series, Remote Sensing of Environment 237 (2020) 111536. URL: https://www.sciencedirect.com/science/article/pii/S0034425719305553. doi:https://doi.org/10.1016/j.rse.2019.111536.

[29] D. Joly, T. Brossard, H. Cardot, J. Cavailhes, M. Hilal, P. Wavresky, Les types de climats en france, une construction spatiale, Cybergeo: European Journal of Geography (2010). URL: https://journals.openedition.org/cybergeo/23155#annexe. doi:10.4000/cybergeo.23155, publisher: CNRS-UMR Géographie-cités 8504.

[30] P. Cantelaube, M. Carles, Le registre parcellaire graphique: des données

géographiques pour décrire la couverture du sol agricole, Le Cahier des Techniques de l'INRA (2014) 58–64.

[31] J. S. Petermann, O. Y. Buzhdygan, Grassland biodiversity, Current Biology 31 (2021) R1195–R1201.

[32] O. D. Team, Orfeo toolbox 8.1.2, 2023. URL: https://doi.org/10.5281/zenodo.8178641. doi:10.5281/zenodo.8178641.

[33] V. Lonjou, C. Desjardins, O. Hagolle, B. Petrucci, T. Tremas, M. Dejus, A. Makarau, S. Auer, MACCS-ATCOR joint algorithm (MAJA), in: A. Comerón, E. I. Kassianov, K. Schäfer (Eds.), Remote Sensing of Clouds and the Atmosphere XXI, volume 10001, International Society for Optics and Photonics, SPIE, 2016, p. 1000107. URL: https://doi.org/10.1117/12.2240935. doi:10.1117/12.2240935.

[34] J. Inglada, A. Vincent, M. Arias, B. Tardy, D. Morin, I. Rodes, Operational high resolution land cover map production at the country scale using satellite image time series, Remote Sensing 9 (2017). URL: https://www.mdpi.com/2072-4292/9/1/95. doi:10.3390/rs9010095.

[35] V. Bellet, M. Fauvel, J. Inglada, Land Cover Classification with Gaussian Processes using spatio-spectro-temporal features, IEEE Transactions on Geoscience and Remote Sensing (2023). URL: https://hal.science/hal-03781332. doi:10.1109/TGRS.2023.3234527.

[36] J. W. Rouse, R. H. Haas, J. A. Schell, D. W. Deering, et al., Monitoring vegetation systems in the great plains with erts, NASA Spec. Publ 351 (1974) 309.

[37] C. Pelletier, G. I. Webb, F. Petitjean, Temporal convolutional neural network for the classification of satellite image time series, Remote Sensing 11 (2019) 523.

[38] V. S. F. Garnot, L. Landrieu, Lightweight temporal self-attention for classifying satellite images time series, in: V. Lemaire, S. Malinowski, A. Bagnall,

T. Guyet, R. Tavenard, G. Ifrim (Eds.), Advanced Analytics and Learning on Temporal Data, Springer International Publishing, Cham, 2020, pp. 171–181.

[39] D. Kingma, J. Ba, Adam: A method for stochastic optimization, in: International Conference on Learning Representations (ICLR), San Diega, CA, USA, 2015.

[40] L. O. H. N. V. Chawla, K. W. Bowyer, W. P. Kegelmeyer, Smote: synthetic minority over-sampling technique, Journal of artificial intelligence research (2002). doi:10.1613/jair.953.

[41] E. A. G. Haibo He, Yang Bai, S. Li, Adasyn: Adaptive synthetic sampling approach for imbalanced learning, IEEE International Joint Conference on Neural Networks (2008). doi:10.1109/IJCNN.2008.4633969.

[42] G. Lemaître, F. Nogueira, C. K. Aridas, Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning, Journal of Machine Learning Research 18 (2017) 1–5. URL: http://jmlr.org/papers/v18/16-365.html.

[43] Z. Wang, Y. Ma, Y. Zhang, J. Shang, Review of remote sensing applications in grassland monitoring, Remote Sensing (2022). doi:10.3390/rs14122903.

[44] P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S. J. van der Walt, M. Brett, J. Wilson, K. J. Millman, N. Mayorov, A. R. J. Nelson, E. Jones, R. Kern, E. Larson, C. J. Carey, İ. Polat, Y. Feng, E. W. Moore, J. VanderPlas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, E. A. Quintero, C. R. Harris, A. M. Archibald, A. H. Ribeiro, F. Pedregosa, P. van Mulbregt, SciPy 1.0 Contributors, SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python, Nature Methods 17 (2020) 261–272. doi:10.1038/s41592-019-0686-2.

[45] Y. LeCun, Y. Bengio, G. E. Hinton, Deep learning, Nature (2015). doi:10.1038/nature14539.

[46] T. Lin, R. Zhong, Y. Wang, J. Xu, J. Xu, Y. Ying, L. F. Rodríguez, K. C. Ting, H. Li, Deepcropnet: A deep spatial-temporal learning framework for county-level corn yield estimation, Environmental Research Letters (2020). doi:10.1088/1748-9326/ab66cb.

[47] C. Liao, J. Wang, Q. Xie, A. A. Baz, X. Huang, J. Shang, Y. He, Synergistic use of multi-temporal radarsat-2 and venμs data for crop classification based on 1d convolutional neural network, Remote Sensing 12 (2020) 832.

[48] L. Zhong, L. Hu, H. Zhou, Deep learning based multi-temporal crop classification, Remote sensing of environment 221 (2019) 430–443.

[49] N. Younes, K. E. Joyce, S. W. Maier, All models of satellite-derived phenology are wrong, but some are useful: A case study from northern australia, International Journal of Applied Earth Observation and Geoinformation (2021). doi:10.1016/j.jag.2020.102285.

[50] K. Kowalski, C. Senf, P. Hostert, D. Pflugmacher, Characterizing spring phenology of temperate broadleaf forests using landsat and sentinel-2 time series, International Journal of Applied Earth Observation and Geoinformation (2020). doi:10.1016/j.jag.2020.102172.

[51] Q. Zhou, J. Rover, J. Brown, B. Worstell, D. Howard, Z. Wu, A. L. Gallant, B. Rundquist, M. Burke, Monitoring landscape dynamics in central us grasslands with harmonized landsat-8 and sentinel-2 time series data, Remote Sensing 11 (2019) 328.

[52] F. Gao, M. Anderson, C. Daughtry, A. Karnieli, D. Hively, W. Kustas, A within-season approach for detecting early growth stages in corn and soybean using high temporal and spatial resolution imagery, Remote Sensing of Environment 242 (2020) 111752.

[53] E. Jafarigol, T. Trafalis, A review of machine learning techniques in imbalanced data and future trends, ArXiv abs/2310.07917 (2023). URL: https://api.semanticscholar.org/CorpusID:263909438.

[54] R. Mohammed, J. Rawashdeh, M. Abdullah, Machine learning with oversampling and undersampling techniques: Overview study and experimental results, in: 2020 11th International Conference on Information and Communication Systems (ICICS), 2020, pp. 243–248. doi:10.1109/ICICS49469.2020.239556.

[55] A. A. Rahman, S. S. Prasetiyowati, Y. Sibaroni, Performance analysis of the imbalanced data method on increasing the classification accuracy of the machine learning hybrid method, Jipi (Jurnal Ilmiah Penelitian Dan Pembelajaran Informatika) (2023). doi:10.29100/jipi.v8i1.3286.

[56] X. Hao, L. Liu, R. Yang, L. Yin, L. Zhang, X. Li, A review of data augmentation methods of remote sensing image target recognition, Remote Sensing 15 (2023). URL: https://www.mdpi.com/2072-4292/15/3/827. doi:10.3390/rs15030827.

[57] M. Claverie, J. Ju, J. G. Masek, J. L. Dungan, E. F. Vermote, J.-C. Roger, S. V. Skakun, C. Justice, The harmonized landsat and sentinel-2 surface reflectance data set, Remote Sensing of Environment 219 (2018) 145–161. URL: https://www.sciencedirect.com/science/article/pii/S0034425718304139. doi:https://doi.org/10.1016/j.rse.2018.09.002.

[58] F. Gao, X. Zhang, Mapping crop phenology in near real-time using satellite remote sensing: Challenges and opportunities, Journal of Remote Sensing (2021). doi:10.34133/2021/8379391.

[59] Y. Cheng, A. Vrieling, F. Fava, M. Meroni, M. Marshall, S. Gachoki, Phenology of short vegetation cycles in a kenyan rangeland from planetscope and sentinel-2, Remote Sensing of Environment 248 (2020) 112004. URL: https://www.sciencedirect.

com/science/article/pii/S0034425720303746. doi:https://doi.org/10.1016/j.rse.2020.112004.

[60] V. Bellet, M. Fauvel, J. Inglada, J. Michel, End-to-end Learning for Land Cover Classification using Irregular and Unaligned SITS by Combining Attention-Based Interpolation with Sparse Variational Gaussian Processes, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing (2023). URL: https://hal.science/hal-04112115.

[61] T. Hastie, R. Tibshirani, J. Friedman, The Elements of Statistical Learning, Springer Series in Statistics, Springer New York Inc., New York, NY, USA, 2001.

[62] A. Ivanda, L. Šerić, M. Bugarić, M. Braović, Mapping chlorophyll-a concentrations in the kaštela bay and brač channel using ridge regression and sentinel-2 satellite images, Electronics 10 (2021) 3004.

[63] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, Scikit-learn: Machine learning in Python, Journal of Machine Learning Research 12 (2011) 2825–2830.

[64] L. Breiman, Random forests, Machine Learning (2001). doi:10.1023/a:1010933404324.

[65] M. Belgiu, L. Drăguţ, Random forest in remote sensing: A review of applications and future directions, ISPRS journal of photogrammetry and remote sensing 114 (2016) 24–31.

[66] A. Zhang, Z. C. Lipton, M. Li, A. J. Smola, Dive into Deep Learning, Cambridge University Press, 2023. https://D2L.ai.

[67] N. Kussul, M. Lavreniuk, S. Skakun, A. Shelestov, Deep learning classification of land cover and crop types using remote sensing data, IEEE Geoscience and Remote Sensing Letters 14 (2017) 778–782.

[68] C. Zhang, X. Pan, H. Li, A. Gardiner, I. Sargent, J. Hare, P. M. Atkinson, A hybrid mlp-cnn classifier for very fine resolution remotely sensed image classification, ISPRS Journal of Photogrammetry and Remote Sensing 140 (2018) 133–144.

[69] C. Zhang, I. Sargent, X. Pan, H. Li, A. Gardiner, J. Hare, P. M. Atkinson, Joint deep learning for land cover and land use classification, Remote sensing of environment 221 (2019) 173–187.

[70] B. Vinayak, H. S. Lee, S. Gedem, Prediction of land use and land cover changes in mumbai city, india, using remote sensing data and a multilayer perceptron neural network-based markov chain model, Sustainability 13 (2021) 471.

[71] T. Kattenborn, J. Leitloff, F. Schiefer, S. Hinz, Review on convolutional neural networks (cnn) in vegetation remote sensing, ISPRS journal of photogrammetry and remote sensing 173 (2021) 24–49.

[72] D. Guidici, M. L. Clark, One-dimensional convolutional neural network land-cover classification of multi-seasonal hyperspectral imagery in the san francisco bay area, california, Remote Sensing 9 (2017) 629.

[73] S. Ofori-Ampofo, C. Pelletier, S. Lang, Crop type mapping from optical and radar time series using attention-based deep learning, Remote Sensing 13 (2021) 4668.

[74] Z. Li, G. Chen, T. Zhang, Temporal attention networks for multitemporal multisensor crop classification, Ieee Access 7 (2019) 134677–134690.

[75] S. K. McFeeters, The use of the normalized difference water index (ndwi) in the delineation of open water features, International Journal of Remote Sensing (1996). doi:10.1080/01431169608948714.

[76] R. Escadafal, Remote sensing of arid soil surface color with landsat thematic mapper, Advances in space research 9 (1989) 159–163.

[77] J. Inglada, A. Vincent, M. Arias, B. Tardy, iota2-a25386, 2016. URL: https://doi.org/10.5281/zenodo.58150. doi:10.5281/zenodo.58150.

Appendix

Appendix A. Review of models

Appendix A.1. Conventional machine learning models

Ridge Regression is a regularized linear model that seeks a linear relationship between the predictors (here the Sentinel-2 spectro-temporal features) and the output (here the observed mowing date) [61]. A regularized version was used to cope with the high number of spectro-temporal features [61, Chapter 3]. This method serves as a baseline for supervised model: its learning capacity is limited w.r.t. other non-parametric regression methods but has provided accurate results for some case, such as *chlorophyll-a* concentration mapping [62]. The regularization parameter value was selected using 10-folds cross-validation on the training data, as implemented in Scikit-learn [63].

Random Forest is a non-parametric and non-linear regression model introduced by Breiman [64]. It is an ensemble-based model learning multiple independent decision trees, using bootstraps of training samples and features. It has been widely used in remote sensing time series applications, mainly for land cover/use mapping [34] and estimation of continuous variables [65]. Several hyperparameters can be selected for training. The most important one is the number of decision trees in the forest. As shown in Inglada et al. [34], Fauvel et al. [28], setting it to a large value is enough to provide accurate results. In this experiment we found that 100 trees was a good compromise: increasing the values did not lead to an improvement of the precision while the processing complexity (time and memory footprint) was much higher. Random Forest was implemented in [63].

Appendix A.2. Deep-learning models

One conventional and two advanced DL models were implemented: a Multilayer Perceptron (MLP), a 1D Convolutional Neural Network (1D-CNN) and the Lightweight Temporal Attention Encoder (LTAE), respectively. MLP was composed of three "linear layer + batchnormalization layer + rectified linear activation layer" modules and last linear output layer [66]. Such architecture has been widely used in remote sensing for land cover/use mapping [67, 68, 69] or land cover/use changes analysis [70].

1D-CNN was defined to perform along the temporal dimension, as in [71, 67, 72, 48, 47, 37], to take into account the temporal dependence between the acquisition dates. From the MLP configuration, we replace the linear layer by a 1D convolutional layer and add max-pooling operation, as usually done with CNN models [66].

LTAE used temporal attention mechanism to make use of the acquisition dates [38]. Attention mechanism has showed to perform really well for land-cover mapping [73, 38, 74, 35]. The same architecture proposed by Garnot and Landrieu [38] was used in this work, the last layer and loss function was modified to perform regression rather than classification.

Appendix B. Threshold-based method

We implemented a specific mowing event detection algorithm introduced by Vroey et al. [20] and integrated into the Sen4CAP toolbox (`http://esa-sen4cap.org`) to facilitate the monitoring of grassland management activities across Europe, aligning with the European Common Agricultural Policy. In our study, this method was adapted to detect first mowing event date, since it was primarily designed to detect mowing event time interval.

Vroey et al. [20] proposed two independent change detection algorithms, whereby raw Sentinel-2 NDVI and Sentinel-1 VH-coherence time series were evaluated separately. In the final product, Sentinel-1 outputs were considered only when Sentinel-2 omitted events due to cloud cover. Here, we reproduced

and adapted their Sentinel-2-based algorithm for evaluating pixel-based time series, as opposed to the original method that used object-based approaches.

To account for a mowing event, the original algorithm performed the following steps:

1. Each observation NDVI(t) is compared to the last available cloud-free observation NDVI($t-1$).

2. If the loss of NDVI, between NDVI(t) and NDVI($t-1$), is greater than 0.15 NDVI (NDVI(t) $<$ NDVI($t-1$) $-$ 0.15), a mowing event is considered. As an additional condition, two consecutive mowing events must be separated by a minimum temporal distance of 28 days, and if a mowing event is detected within the time interval $[t-1, t]$, it is assumed that the actual event took place within 60 days before t. If $[t-1, t]$ spans more than 60 days, the detection interval is adjusted to $[t-60, t]$. For each detected mowing event, the confidence level was estimated through a normalization function as follows:

$$f(x; \min, \max) = \max - (\max - \min) \times \exp(-x), \tag{B.1}$$

where x is the difference NDVI($t-1$) $-$ 0.15 $-$ NDVI(t), $[\min, \max]$ were set to fit the confidence limits from 0.5 to 1.

The first mowing event among the four most confident detections was retained, as opposed to the original method that retained all four most confident detections. In contrast to the original method, where the time interval $[t-1, t]$ was kept for each detected mowing event, we retained the specific date t. Therefore, in our study, additional checks in step 2 were ignored.

Appendix C. Grassland management map

A map of grassland management practices -*mowed* or *unmowed*- was generated to constrain mowing date prediction to areas of mowed grassland. We

performed a pixel-based classification task within a nationwide grassland mask (Figure 1), derived from permanent grassland plots declared in the 2022 LPIS (section 2.1). This database provides spatialized information on agricultural plot boundaries and crop types, but does not provide information about management practices.

Here, we trained a Random Forest classifier using a grassland management practices dataset, derived from ground observations in 2022 (section 2.3). In this reference dataset, *mowed* class included 1 605 plots and *unmowed* class 660 plots (Table 2). Reference data were split into a 70% training dataset and a 30% test dataset, ensuring classes and sites representation through stratified sampling. Sentinel-2-based time series were used as predictor. In this dataset, in addition to spectral bands, we also computed three spectral indices: Normalized Difference Vegetation Index - NDVI [36], Normalized Difference Water Index - NDWI [75] and Brightness Index - BI [76].

The classification was done using IOTA^2 software [77]. Grassland management map achieved an overall precision of 90%, with *mowed* class showing an F-score of 0.93 and *unmowed* class exhibiting an F-score of 0.81. Findings showed that *mowed* class was slightly overestimated. In addition, in each plot of the initial grassland mask, resulting classes exhibited a coherent spatial structure, showing unimodal or bimodal intra-plot management patterns.

It is important to note that all quantitative evaluation results presented in this paper were not based on the grasslands management map, as they were computed on the reference data (observed *mowed* plots). This map was used for visual evaluation only.

Appendix D. Tables and figures

Table D.4: Algorithm-specific statistical summary assessed at pixel-level. Each score value represents weighted mean of all sites. A site-specific score was weighted using the number of pixels used for the evaluation. A site-specific score represents mean of fifty individual evaluations (from 50 folds by bootstrapping 70% of observations). MAE, RMSE and Max_error are represented in days. The lines are sorted based on R^2 values (descending order).

Algorithm	MAE	RMSE	R^2	Max_error
LTAE	5.63	9.13	0.52	59.58
LTAE_SMOTE	5.54	9.24	0.51	62.09
LTAE_ADASYN	5.51	9.32	0.50	64.31
MLP_ADASYN	6.36	9.39	0.49	54.00
1D-CNN_SMOTE	6.53	9.40	0.49	48.59
1D-CNN_ADASYN	6.67	9.46	0.48	48.57
MLP_SMOTE	6.50	9.50	0.47	54.23
1D-CNN	6.84	9.60	0.47	49.30
RF_ADASYN	6.71	10.04	0.42	61.30
RF	6.80	10.09	0.40	56.69
MLP	7.02	10.11	0.40	54.28
RF_SMOTE	6.97	10.64	0.34	61.53
Ridge	8.62	11.45	0.23	72.57
SimpleMean	10.28	13.81	-0.10	42.06
Ridge_SMOTE	13.24	16.71	-0.76	87.89
Threshold	14.02	19.66	-1.36	73.24
Ridge_ADASYN	17.11	21.17	-2.03	110.89

Table D.5: Algorithm-specific statistical summary assessed at plot-level. Each score value represents weighted mean of all sites. A site-specific score was weighted using the number of plots used for the evaluation. A site-specific score represents mean of fifty individual evaluations (from 50 folds by bootstrapping 70% of observations). MAE, RMSE and Max_error are represented in days. The lines are sorted based on R^2 values (descending order).

Algorithm	MAE	RMSE	R^2	Max_error
LTAE_ADASYN	5.42	8.56	0.61	41.32
LTAE_SMOTE	5.28	8.66	0.61	47.93
LTAE	5.52	8.95	0.58	48.48
RF_ADASYN	6.19	9.19	0.56	46.06
RF_SMOTE	6.51	9.59	0.51	46.73
1D-CNN_ADASYN	6.71	9.62	0.51	42.70
MLP_SMOTE	6.46	9.66	0.51	45.73
1D-CNN_SMOTE	6.60	9.68	0.51	43.09
RF	6.61	9.73	0.50	45.76
MLP_ADASYN	6.45	9.75	0.50	48.05
1D-CNN	6.95	9.94	0.48	44.05
MLP	6.86	10.04	0.47	44.05
Ridge	8.04	11.01	0.36	43.25
SimpleMean	10.78	14.47	-0.10	41.14
Ridge_SMOTE	11.57	14.63	-0.21	45.53
Threshold	12.98	17.24	-0.64	56.38
Ridge_ADASYN	14.42	17.80	-0.93	50.07

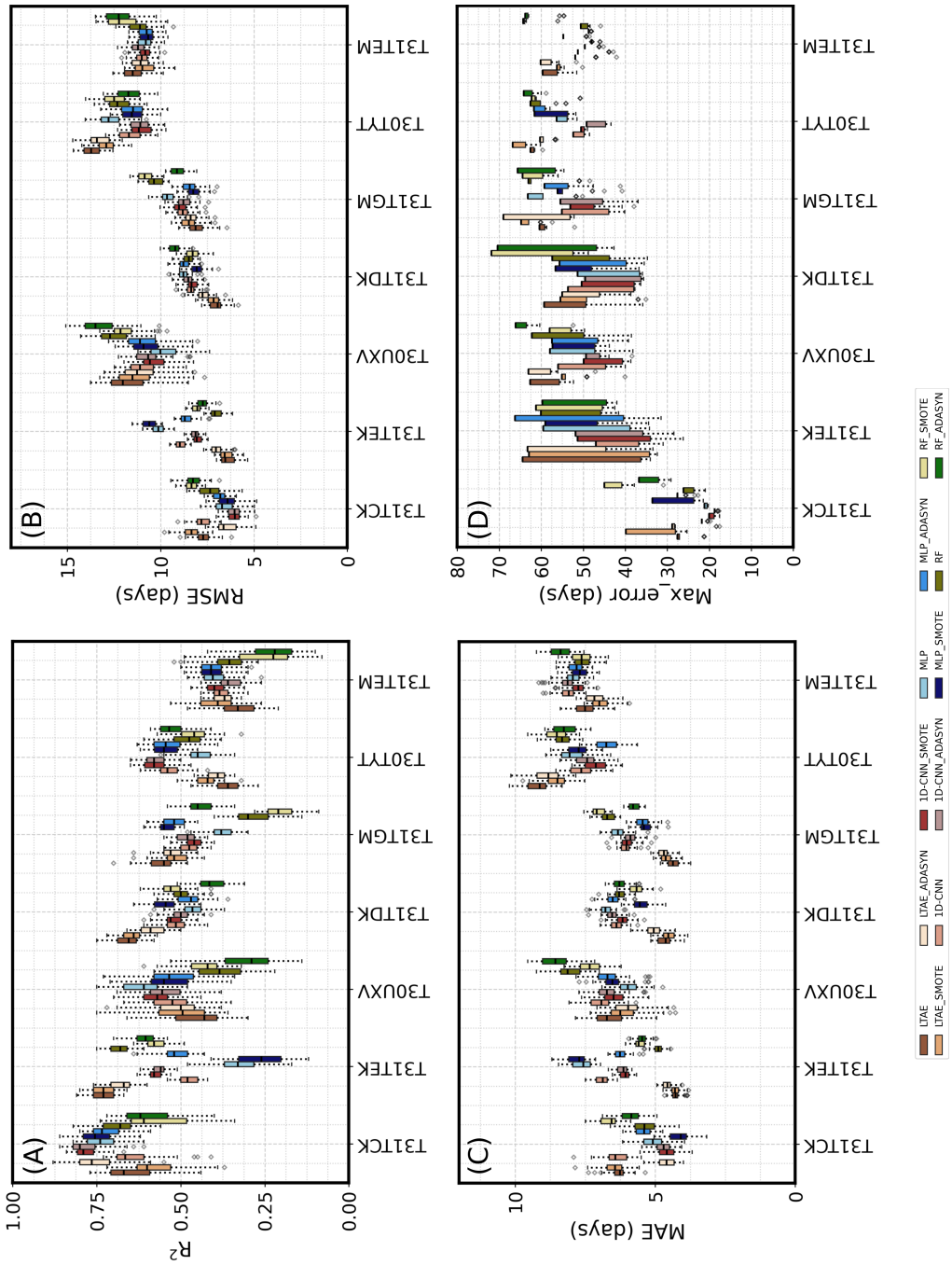


Figure D.11: Site-specific outputs in terms of (A) R^2, (B) RMSE, (C) MAE and (D) Max_error. The sites are represented on the x-axis. The color palette represents the algorithms. For each site, fifty individual evaluations were conducted (from 50 folds by bootstrapping 70% of observations). Ridge Regression, SimpleMean and Threshold outputs are not shown in this figure.