Contents lists available at ScienceDirect

Remote Sensing of Environment

journal homepage: www.elsevier.com/locate/rse

Adaptive fusion of multi-modal remote sensing data for optimal sub-field crop yield prediction

Francisco Mena [a,b,*,1], Deepak Pathak [a,b,1], Hiba Najjar [a,b], Cristhian Sanchez [a,b], Patrick Helber [c], Benjamin Bischke [c], Peter Habelitz [c], Miro Miranda [a,b], Jayanth Siddamsetty [b], Marlon Nuske [b], Marcela Charfuelan [b], Diego Arenas [b], Michaela Vollmer [b], Andreas Dengel [a,b]

[a] Department of Computer Science, University of Kaiserslautern-Landau (RPTU), Kaiserslautern, Germany
[b] SDS, German Research Center for Artificial Intelligence (DFKI), Kaiserslautern, Germany
[c] Vision Impulse GmbH, Kaiserslautern, Germany

ARTICLE INFO

Keywords:
Multi-modal learning
Multi-sensor data fusion
Remote sensing
Deep neural networks
Crop yield prediction

ABSTRACT

Accurate crop yield prediction is of utmost importance for informed decision-making in agriculture, aiding farmers, industry stakeholders, and policymakers in optimizing agricultural practices. However, this task is complex and depends on multiple factors, such as environmental conditions, soil properties, and management practices. Leveraging Remote Sensing (RS) technologies, multi-modal data from diverse global data sources can be collected to enhance predictive model accuracy. However, combining heterogeneous RS data poses a fusion challenge, like identifying the specific contribution of each modality in the predictive task. In this paper, we present a novel multi-modal learning approach to predict crop yield for different crops (soybean, wheat, rapeseed) and regions (Argentina, Uruguay, and Germany). Our multi-modal input data includes multi-spectral optical images from Sentinel-2 satellites and weather data as dynamic features during the crop growing season, complemented by static features like soil properties and topographic information. To effectively fuse the multi-modal data, we introduce a Multi-modal Gated Fusion (MMGF) model, comprising dedicated modality-encoders and a Gated Unit (GU) module. The modality-encoders handle the heterogeneity of data sources with varying temporal resolutions by learning a modality-specific representation. These representations are adaptively fused via a weighted sum. The *fusion* weights are computed for each sample by the GU using a concatenation of the multi-modal representations. The MMGF model is trained at sub-field level with 10 m resolution pixels. Our evaluations show that the MMGF outperforms conventional models on the same task, achieving the best results by incorporating all the data sources, unlike the usual fusion results in the literature. For Argentina, the MMGF model achieves an R^2 value of 0.68 at sub-field yield prediction, while at the field level evaluation (comparing field averages), it reaches around 0.80 across different countries. The GU module learned different weights based on the country and crop-type, aligning with the variable significance of each data source to the prediction task. This novel method has proven its effectiveness in enhancing the accuracy of the challenging sub-field crop yield prediction. Our investigation indicates that the gated fusion approach promises a significant advancement in the field of agriculture and precision farming.

1. Introduction

Phenomena observed on the Earth have complex interactions, and hence their observation requires a multi-faceted measurement approach. For instance, the development of a farm crop is affected by human practices, weather conditions, soil structure, and other aspects. To cover some of these interacting factors, the availability of diverse and rapidly increasing Remote Sensing (RS) sources (Camps-Valls et al., 2021) has enabled multiple observations for an object of study. In the machine learning (ML) domain, this scenario is called Multi-Modal Learning (MML; Li et al. (2022)). However, in RS-based applications, the modalities can be rather heterogeneous. The differences in spatial and temporal resolution could be significant, and establishing the complementary and supplementary information between sensors for a predictive task is non-trivial (Mena et al., 2024).

In this paper, we focus on the challenging machine learning task of multi-modal crop yield prediction. The crop productivity can be

[*] Corresponding author at: Department of Computer Science, University of Kaiserslautern-Landau (RPTU), Kaiserslautern, Germany.
E-mail addresses: f.menat@rptu.de (F. Mena), deepak_kumar.pathak@dfki.de (D. Pathak).

[1] Both authors contributed equally to this work.

https://doi.org/10.1016/j.rse.2024.114547
Received 9 October 2023; Received in revised form 11 October 2024; Accepted 28 November 2024

influenced by, but not limited to, environmental conditions, soil properties, and management practices. Therefore, the effectiveness of the predictive models relies on how well it combines the task-related information from the multi-modal data. The standard approach in the related research involves extracting a set of domain-specialized features from each modality, concatenating and feeding them to a single model, such as classical ML models (Bocca and Rodrigues, 2016; Cai et al., 2019; Maimaitijiang et al., 2020). Recent research has used deep neural networks (Gavahi et al., 2021), such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs). Additionally, some works (Maimaitijiang et al., 2020; Chu and Yu, 2020; Shahhosseini et al., 2021) use a feature-level fusion strategy, where encoders are used to learn a new set of features for each modality, before merging them. In this way, the fusion is done at the intermediate layers of neural network models instead of at the input layer. However, these techniques use a static fusion function (such as concatenation or average operators), which ignores the variable impact that each modality has on the yield value.

Nowadays, variants of the original attention mechanism (Bahdanau et al., 2015) has been applied to crop yield prediction with the purpose of highlighting input features. For instance, Lin et al. (2020) apply attention weights across time in time-series data and then perform a weighted sum by using gated recurrent units in county-level corn yield prediction. On the other hand, Ma et al. (2023) apply attention weights across different sensors in field-level winter wheat yield prediction. In addition, Feng et al. (2021) apply a guided weights across input features in county-level winter wheat yield prediction, where the learned weights are computed from the geographical coordinates and year of the data. The work of Feng et al. (2021) claim the usage of a weighted neural network, which might be related to our work. However, the weights we use in our work are applied across modalities at feature-level for explicit fusion, i.e. weigh the contribution of each individual modality in the fusion, in contrast to the weights of Feng et al. (2021) that are applied across all features at input-level, i.e. the input features are just scaled.

Our case study consists of the crop yield prediction utilizing multiple RS data sources, each characterized by distinct spatial and temporal resolution. The target data is the crop yield from fields rasterized at a spatial resolution of 10 m/px, which we refer to as **sub-field** level data. For the sub-field level predictions, we focus on a pixel-wise prediction approach. We use multiple crop-types (soybean, wheat, and rapeseed) and regions (Argentina, Uruguay, and Germany) grouped into four dataset combinations. We use multi-spectral optical images from the Sentinel-2 (S2) mission to provide the main information about the Earth surface. In addition, we collected different RS data sources (modalities) to enhance the modeling of the yield prediction task, and provide further information that the optical image might not capture directly. We include weather features during the growing season, and static information from soil properties, and elevation maps. For this, we use temporal data from seeding to harvesting, encompassing the entire crop growing season. Notably, the beginning and end of the growing season depend on several factors, including the region, crop-type and farmer's practices.

We propose a MML model that performs data fusion at the feature-level using Gated Units (GUs; Arevalo et al. (2020)). The features are learned by dedicated modality-encoders, allowing to handle the heterogeneous nature of modalities with different temporal resolution and data distributions. Since the RS data used is fairly diverse, the GU module is included to fuse the learned features based on data-driven fusion weights. This allows an adaptive fusion of the multi-modal high level features based on each sample (pixel). For the evaluation, we use a cross-validation splits of the fields in each dataset. The metrics R^2, MAE, and MAPE are computed at the field and sub-field level (pixels). Our results show an overall improvement compared to previous approaches based on single-modal learning, input-level fusion (Pathak et al., 2023), and other conventional feature-level fusion approaches. Thus, the best results are obtained by feeding all modalities, unlike common results in the literature. The R^2 values are around 0.80 across all datasets at field-level evaluation. While at sub-field level, the score is 0.68 for Argentina and around 0.44 for Uruguay and Germany. The key contributions of this paper are as follows:

1. We propose a two-component model that (i) learns a high-level representation for each modality via dedicated encoders, and (ii) learns to adaptively fuse this data with a weighted sum computed with a gating mechanism (Gated Unit). To the best of our knowledge, the proposed Multi-Modal Gated Fusion (**MMGF**) is the first model applying an adaptive fusion approach (via gating mechanism) to multi-modal crop yield prediction.
2. We evaluated in three countries and three crop-types, showing that our best results are consistently obtained using all modalities over regions and crops, contrary to previous results in the literature (Bocca and Rodrigues, 2016; Kang et al., 2020; Pathak et al., 2023; Ma et al., 2023). In addition, we obtained overall improvements compared to an input-level fusion baseline, single-modal models and other fusion approaches.
3. The proposed model allows a simple interpretation through the analysis of the gated fusion weights. These data-dependent weights act as a proxy to the modality contribution in the yield prediction (Section 6.1). The GU module in the proposed MMGF learns different fusion weights distribution depending on the country and crop-type.

The paper is organized as follows: In Section 2, related works in crop yield prediction and adaptive fusion are presented. While Section 3 describes the data and study, Section 4 explains the proposed approach. Experimental settings and main results are described in Section 5. Additional analysis is presented in Section 6. Finally, the conclusion about the work are in Section 7.

2. Related work

Recently, there has been an increase of MML models applied to different tasks with RS data. Since the approaches usually incorporate domain knowledge, they vary from task to task (Mena et al., 2024). In the following, we briefly discuss some related works in MML for RS data.

MML for crop yield prediction. The standard approach in crop yield prediction is to build a specialized domain-specific set of features across the growing season and then apply standard ML models (Bocca and Rodrigues, 2016; Cai et al., 2019), e.g. random forest and Multi Layer Perceptron (MLP), or deep neural networks (Gavahi et al., 2021; Pathak et al., 2023), e.g. with convolutional (CNN) or recurrent (RNN) operations. In these cases, the input-level fusion (Feng et al., 2021) is used to merge the information coming from multiple RS sources, i.e. all the input features are concatenated and then fed to a single model. However, certain approaches involve the fusion of hidden features in intermediate layers of neural network models, a concept known as feature-level fusion (Maimaitijiang et al., 2020; Mena et al., 2024). This strategy needs a sub-model for each modality (referred to as a modality-encoder), which learns new features. For instance, Yang et al. (2019) use two-modal data, RGB and multi-spectral images, to predict the crop yield at county level, where a 2D CNN is used on each modality-encoder. Some works apply feature-level fusion with other types of modality-encoder, depending on the data used. Maimaitijiang et al. (2020) use MLPs for multi-modal vector data, while (Chu and Yu, 2020) incorporate an independent RNN for meteorological dynamic features and an MLP for county information. Shahhosseini et al. (2021) use a 1D CNN (across time) for weather data, 1D CNN (across depths) for soil properties, and an MLP for vector data. Additionally, some works group modalities based on their information content. For instance, Wang et al.

(2020) group static (soil properties) and dynamic (optical and meteorological) features into a two-modal model that performs feature-level fusion with an MLP and Long-Short Term Memory (LSTM) as modality-encoders. Subsequent works (Cao et al., 2021; Srivastava et al., 2022) have applied a similar group of static and dynamic modalities for the yield prediction with the feature-level fusion. However, instead of grouping modalities, in this work we process the features of each modality with a dedicated model, and then employ an adaptive fusion approach.

Adaptive fusion and attention in RS. As different variations of neural networks models have been used in literature, different forms of the attention mechanism (Bahdanau et al., 2015) have shown state-of-the-art results in RS applications. The temporal attention pooling is used by some works (Lin et al., 2020; Garnot and Landrieu, 2020; Ofori-Ampofo et al., 2021; Garnot et al., 2022) to aggregate dynamic (time-series) RS data. Attention mechanisms have also been used to highlight input features (Feng et al., 2021), spatio-temporal RS data (Wang et al., 2022), or different sensors (Ma et al., 2023). Motivated by attention mechanisms and mixture-of-expert models (Jacobs et al., 1991), some studies have explored gated fusion approaches in pursuit of adaptively fusing multi-modal features (Zhang et al., 2020; Zheng et al., 2021; Hosseinpour et al., 2022). The main idea is to highlight (apply an adaptive weight to) the most relevant information of each modality and aggregate them, e.g. with a linear sum (Arevalo et al., 2020). Furthermore, recent literature is exploring the connection between attention and explainability. For instance, explanation through attention is used for land-use and land-cover classification (Méger et al., 2022) and crop-type mapping (Rußwurm et al., 2020; Obadic et al., 2022). This is mainly because the data-driven weights operate as a proxy for feature importance by identifying which features are most used for model prediction. Our work considers a tailored of the adaptive fusion based on a gating mechanism. To the best of our knowledge, this is the first work applying the gated fusion approach to the multi-modal crop yield prediction task.

MML for land-use and land-cover task. In the widely studied land-use and land-cover mapping, the target data focuses on a specific and limited time frame. Typically, the multi-modal data consist of static images captured by a variety of sensors, including multi-spectral optical or radar images, or elevation maps. Chen et al. (2017) propose one of the first models that use deep neural networks on the modality-encoders (concretely 2D CNN) of feature-level fusion (Mena et al., 2024) with two-sensor data. Later, to fuse sensors with different resolution, Benedetti et al. (2018) include an auxiliary classifier for each modality that is feed with the learned features and acts as a regularization. (Wu et al., 2021) use a variation of this approach with auxiliary reconstructions of the learned features. On the other side, Audebert et al. (2018) propose to fuse across all layers in 2D CNN modality-encoders with a central model. They use skip-connections between individual modalities and post-fusion layers, where (Hosseinpour et al., 2022) later extend the merge across all skip-connections layers. To account for different levels of information that modalities might have, Wang et al. (2022) merge the learned features in a hierarchical way. In addition, the decision-level fusion (merge class predictions) has been explored without significant advantages compared to feature fusion (Audebert et al., 2018; Ofori-Ampofo et al., 2021). These works usually consider a pixel-wise prediction as our application. However, we are considering a time-dependent target with dynamic features.

Recent works in crop yield prediction focuses on feature fusion at field level by grouping dynamic and static features (Wang et al., 2020; Cao et al., 2021; Srivastava et al., 2022). However, the standard merger are static functions, such as simple concatenation, while in other RS applications more sophisticated approaches have been used. Since the gated fusion (through the adaptive weighted sum) suits the dynamic significance that each modality could have for prediction, we use it for a sub-field crop yield prediction case study.

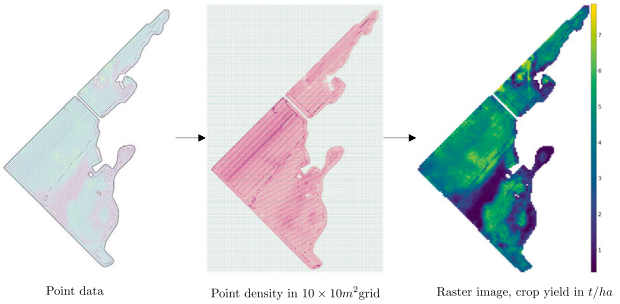

Point data
Point density in 10 × 10m^2grid
Raster image, crop yield in t/ha
1
2
3
4
5
6
7

Fig. 1. The figure depicts the rasterization process for the ground truth yield data, where a cleaned point vector data (left) for a field is aligned with the satellite image, and the mean of all yield points within one pixel is assigned to that pixel. The resulting raster image is referred to as the yield map in this study.

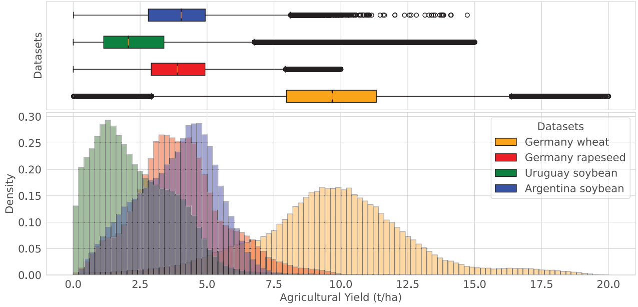


Fig. 2. Crop yield distribution per pixel (at 10 m resolution) in the four datasets considered in this study.

3. Data

Our case study consists of the crop yield prediction task based on multiple RS data sources, with different spatial and temporal resolutions. The target data is the crop yield at sub-field level (pixel-wise yield values at 10 m spatial resolution) over different countries and crop-types.

3.1. Crop yield data

The crop yield data corresponds to the target (ground-truth) variable to be predicted by machine learning models. This yield data originally comes from combine harvesters at a sub-field level as point vector data. The combine harvester equipped with yield monitors records point data at consistent intervals with a high spatial resolution as it moves across the field during the harvesting process. All points collected are characterized by different features such as the geolocation, yield moisture, and amount of yield. We pre-processed the points by (i) re-projecting the reference coordinate system to respective UTM zones, (ii) removing wrong values based on position, timestamp, moisture, and yield (like biological infeasible crop yield), and (iii) filtering based on a statistical threshold following (Sanchez et al., 2023). The statistical threshold used correspond to remove all points for which the associated yield value is outside the range of three standard deviations around the mean within a field. Furthermore, we rasterized the point vector data into pixels, aligning it with the available satellite images, illustrated in Fig. 1. For each raster pixel, the mean yield of all points within the $10 \times 10\,\text{m}^2$ area is assigned as its value. The resulting rasterized image, with a resolution of 10 meters per pixel (m/px), is referred to as the yield map in this study. The unit of these yield maps is tons per hectare (t/ha).

Table 1
Descriptive factors of the four datasets in this study, with different combinations of country and crop-type.

Name	Country	Crop-type	Years	Total area	Fields	Pixels	Yield value	Total area	Growing season
							mean ± std.	average across field	
ARG-S	Argentina	soybean	2017–2022	15351 ha	190	∼1.4 M	3.86 ± 1.49	79.5 ha	156 days
URU-S	Uruguay	soybean	2018–2021	28358 ha	486	∼1.8 M	2.35 ± 1.59	58.3 ha	169 days
GER-R	Germany	rapeseed	2016–2022	3221 ha	111	∼0.3 M	4.01 ± 1.67	29.0 ha	335 days
GER-W		wheat	2016–2022	3240 ha	188	∼0.3 M	9.64 ± 2.95	17.2 ha	306 days

Crop-type and region. Unlike common crop-region specific use-cases in the literature, we use field data across different countries (Argentina, Uruguay, and Germany), crop-types (soybean, rapeseed, and wheat), and years (from 2016 until 2022). We gathered this data from three data providers, with each provider supplying data for one country. Each provider collects data from various farmers, potentially utilizing combine harvesters from different manufacturers. This data collection process directly influences the quality of the sourced data. Therefore, we categorize this dataset into four distinct sets, each pertaining to different regions and crops. Fig. 2 shows the variability of the yield data from the different datasets used in this study. The distribution of the yield, e.g. mean and standard deviation, changes between country and crop-type. Table 1 presents the total number of fields and yield pixels in each dataset considered in this study. While there are a larger number of pixels of soybean crops (in Argentina and Uruguay), the crops in Germany have fewer number of pixels. The field data of this study is fairly diverse, as the area covered per field is different in each combination, and, over the years, the fields are distributed in different geographical locations within each country. Furthermore, the seeding and harvesting dates are different across the fields and countries. We present this in more detail in Fig. B.13.

3.2. Multi-modal input data

We use dynamic (time-dependent) and static input data collected from different RS sources. Since there are numerous RS sources available nowadays, this work is limited to the selected sources. Our selection criteria hinge on the global availability of the RS data, coupled by its usefulness in estimating crop yield. The aim is to have a better representation and modeling of the crop yield drivers through the growing season, and therefore improve the crop yield prediction. Table 2 displays an overview of the multi-modal RS data used in this study.

3.2.1. S2-based optical image

We use multi-spectral optical information coming from the Sentinel-2 (S2) mission. Specifically, we use the surface reflectance imaging product (level-2 A) from the S2 data, which is available from 2016. For this, we collected a Satellite Image Time Series (SITS) from seeding to harvesting date of each field, with approximately 5-days revisit time. We use all 12 spectral bands of the L2 A product being agnostic to their contribution to crop yield prediction, see Table 2, where the bands with lower spatial resolutions (B1, B5-B7, B8 A, B9-B12) are up-sampled to the ones with higher (at 10 m/px). Finally, the number of images in the SITS per field ranges from 11 to 78 for Argentina, from 21 to 82 for Uruguay, and from 17 to 140 for Germany, depending on the crop growing season.

Thanks to the Scene Classification Layer (SCL) contained in the L2 A product, we can calculate the cloud coverage of the S2-based SITS. The SCL is an additional layer on the S2 data that assigns a label between 12 options (see Table 2) for each pixel in the image. By considering labels related to occlusion factors (cloud, shadow, snow, and errors) a percentage of pixels within the fields related to cloud occlusion is computed for each SITS. We refer to this as **field cloud coverage**. In Fig. 3, we show the monthly cloud coverage for the fields across the growing season. In Germany, there is a longer growing season since it contains winter crops,[2] where we see a high cloud coverage for the winter months (from December to next-year February). These charts show the diversity of the S2 optical images across the growing season for different countries.

3.2.2. Additional modalities

Weather. We utilize meteorological factors obtained from the climate source, ECMWF ERA5 (Hersbach et al., 2020). This source is based on the assimilation of various observations from satellites, ground-based weather stations, and other sources into a consistent numerical weather model, which has been available from 1979. The raw data collected for each field is at hourly temporal resolution, from which we extract four daily features based on temperature and precipitation, see Table 2. The temperature values correspond to the air temperature at 2 m above the land surface in a raw grid with a spatial resolution of 30 km/px.

Digital elevation model (DEM). We use topographic information collected from the NASA source, Space Shuttle Radar Topography Mission (SRTM) (Farr and Kobrick, 2000). Here, elevation information was collected by bouncing radar signals to the Earth's surface. We use five features extracted with the RichDEM tool[3] at a spatial resolution of 30 m/px, see Table 2. This information is static across time (uni-temporal).

Soil map. We utilize chemical and physical soil properties obtained from the global source, SoilGrids (Poggio et al., 2021) available at spatial resolution of 250 meters per pixel. This data source integrates sampled ground-based data with satellite measurements using ML-based predictions to derive global soil properties. SoilGrids offers these properties as static over time. However, in reality, certain chemical properties, such as pH and soil organic matter carbon, exhibit dynamic behavior. According to Poggio et al. (2021), spatial variation outweighs temporal variation, indicating that these properties remain relatively stable over time. Hence, in our study, we treat all soil properties as static (uni-temporal) throughout the crop growing season. We use eight soil properties across three depth intervals, as presented in Table 2. Considering domain expertise, we use the top layer features up to a depth of 30 cm: 0–5 cm, 5–15 cm, and 15–30 cm.

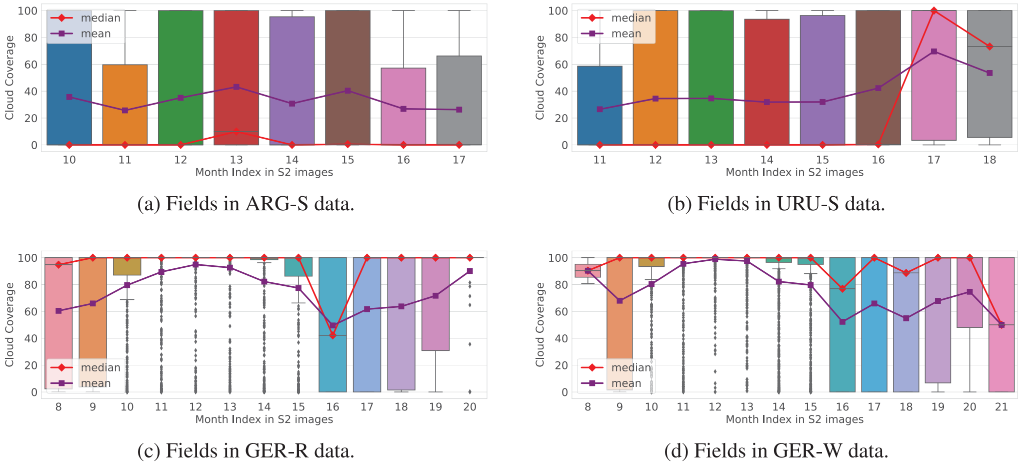
4. Method description

We use a pixel-wise approach for the prediction task. The approach uses a feature-level fusion with an adaptive fusion approach, which shows good results in RS applications (Section 2).

Spatial alignment. To harmonize the different spatial resolutions for the pixel-wise approach, we spatially align the multi-modal data before feeding them to ML models. For DEM and soil modalities, a cubic spline method is used to interpolate to the spatial resolution of the S2 images (10 m/px). While for weather, the value from the centroid of the field is repeated across all the field pixels to match the same spatial resolution. See Fig. 4 for an illustration.

[2] Winter crops are a type of crop planted in early winter and harvested in early summer that could handle cooler temperatures, frost, and shorter days.

[3] http://github.com/r-barnes/richdem (Accessed 18 of April 2024).



(a) Fields in ARG-S data.

(b) Fields in URU-S data.

(c) Fields in GER-R data.

(d) Fields in GER-W data.

Fig. 3. Field cloud coverage across the growing season. Each point in a boxplot represents the cloud coverage of a field in the corresponding month. The month index goes from (1) January in the seeding year to (24) December of the following year.

Table 2
Summary of the collected modalities in our study, with the respective resolutions and used features.

Modality	Data source	Resolution		Features
		Spatial	Temporal	
Optical	S2 S2 SCL	10 m 10 m	5 days 5 days	B1, B2, B3, B4, B5, B6, B7, B8, B8A, B9, B11, B12. Categorical label among: no data, saturated/defective, dark area pixel, cloud shadows, vegetation, not vegetated, water, unclassified, cloud medium probability, cloud high probability, thin cirrus, snow.
Weather	ERA5	30 km	daily	Mean, maximum and minimum temperature, cumulative precipitation.
DEM	SRTM	30 m	–	Aspect, curvature, digital surface model, slope, topographic wetness index.
Soil	SoilGrids	250 m	–	Cation exchange capacity, volumetric fraction of course fragments, clay, nitrogen, soil pH, sand, silt, soil organic carbon. (each at three depths)

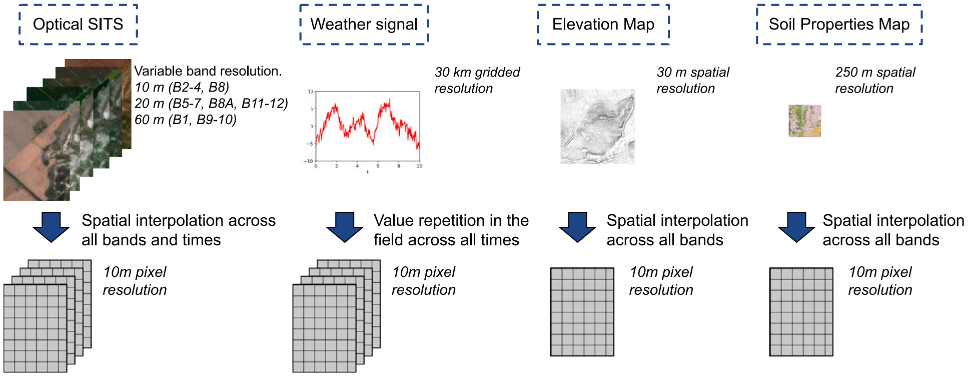


Fig. 4. Illustration of spatial alignment applied to the four input modalities for a specific field. After this process, all modalities have a spatial resolution of 10 m/px in each field.

4.1. Data formulation and notation

Consider the following MML scenario with N labeled pixels, $\mathcal{D} = \{\mathcal{X}^{(i)}, y^{(i)}\}_{i=1}^{N}$, $y^{(i)} \in \mathbb{R}+$ the ground-truth crop yield for the ith pixel, and $\mathcal{X}^{(i)} = \left\{\mathbf{X}_{\mathrm{S2}}^{(i)}, \mathbf{X}_{\mathrm{W}}^{(i)}, \mathbf{x}_{\mathrm{D}}^{(i)}, \mathbf{x}_{\mathrm{S}}^{(i)}\right\}$ its corresponding multi-modal input data. Let B_m be the number of bands or features in each modality $m \in \{\mathrm{S2}, \mathrm{W}, \mathrm{D}, \mathrm{S}\}$ (see Table 2). The S2-based optical modality for the ith pixel is a multivariate time-series of length $T_{\mathrm{S2}}^{(i)}$: $\mathbf{X}_{\mathrm{S2}}^{(i)} \in \mathbb{R}^{T_{\mathrm{S2}}^{(i)} \times B_{\mathrm{S2}}}$, and the weather modality is a multivariate time-series of length $T_{\mathrm{W}}^{(i)}$: $\mathbf{X}_{\mathrm{W}}^{(i)} \in \mathbb{R}^{T_{\mathrm{W}}^{(i)} \times B_{\mathrm{W}}}$. Note that the temporal resolution between modalities and pixels have not been aligned. On the other hand, DEM and soil modalities are constant variables over time for the ith pixel, $\mathbf{x}_{\mathrm{D}}^{(i)} \in \mathbb{R}^{B_{\mathrm{D}}}$ and $\mathbf{x}_{\mathrm{S}}^{(i)} \in \mathbb{R}^{B_{\mathrm{S}}}$. Additionally, $\mathsf{C} : \mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}^{4d}$ is the concatenation function, and $\mathsf{S} : \mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}^{4 \times d}$ the stacking function.

4.2. Feature-level learning

In order to fuse modalities with different temporal resolution and number of features, we learn a single high-level representation for each modality on a d-dimensional vector space, named modality-representation. We use one dedicated encoder model for each modality built with neural networks, named modality-encoder, see Fig. 6 for an illustration. This modality-encoder is a function $E_{\theta_m} : X_m \rightarrow \mathbb{R}^d$, with the corresponding learnable parameters θ_m for the m modality.

$$\mathbf{z}_{\mathrm{S2}}^{(i)} = E_{\theta_{\mathrm{S2}}}\left(\mathbf{X}_{\mathrm{S2}}^{(i)}\right) \in \mathbb{R}^d \tag{1}$$

$$\mathbf{z}_{\mathrm{W}}^{(i)} = E_{\theta_{\mathrm{W}}}\left(\mathbf{X}_{\mathrm{W}}^{(i)}\right) \in \mathbb{R}^d \tag{2}$$

$$\mathbf{z}_{\mathrm{D}}^{(i)} = E_{\theta_{\mathrm{D}}}\left(\mathbf{x}_{\mathrm{D}}^{(i)}\right) \in \mathbb{R}^d \tag{3}$$

$$\mathbf{z}_{\mathrm{S}}^{(i)} = E_{\theta_{\mathrm{S}}}\left(\mathbf{x}_{\mathrm{S}}^{(i)}\right) \in \mathbb{R}^d \tag{4}$$

These learned representations ($\mathbf{z}_m$) allow the model to handle the heterogeneous nature of the modalities (different temporal resolutions, magnitudes, and data distributions). The modality-encoder gives the model the chance to extract information with a specific and dedicated sub-model. Furthermore, the modality-encoder can use different complexities and types of architecture (*asymmetric* network).

Temporal modality-encoder. For multivariate time-series $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_T\} \in \mathbb{R}^{T \times B}$, with the tth observation $\mathbf{x}_t \in \mathbb{R}^B$, recurrent layers (or RNN) could be used to extract temporal high-level representations $\mathbf{h}_t \in \mathbb{R}^d$. With $H^{(l)}$ a recurrent unit (e.g. LSTM) at layer l, and $\mathbf{h}_0^{(l)} = \mathbf{0}\ \forall l$, the hidden state at a time t and layer l can be expressed by

$$\mathbf{h}_t^{(l)} = \begin{cases} H^{(l)}\left(\mathbf{x}_t, \mathbf{h}_{t-1}^{(l)}\right), & l = 1 \\ H^{(l)}\left(\mathbf{h}_t^{(l-1)}, \mathbf{h}_{t-1}^{(l)}\right), & l \in \{2, 3, \ldots, L\} \end{cases} . \tag{5}$$

Then, the last hidden state could be used to extract a single vector representation $\mathbf{a}^{(L)} = \mathbf{h}_T^{(L)} \in \mathbb{R}^d$. Additionally, attention-based approaches, such as temporal attention pooling ($\mathbf{a} = \sum_t \alpha_t \mathbf{h}_t$) could be used (in Section 6.2 we show an empirical comparison). Given the dynamic features of optical and weather modalities, we use these types of architectures in $E_{\theta_{\mathrm{S2}}}$, $E_{\theta_{\mathrm{W}}}$.

Static modality-encoder. For vector data $\mathbf{x} \in \mathbb{R}^B$, fully connected layers (or MLP) are used to extract a high-level representation at the output layer $\mathbf{a}^{(L)} \in \mathbb{R}^d$. With $H^{(l)}$ a linear projection followed by a nonlinear activation function on layer l, the output of a layer l could be written as

$$\mathbf{a}^{(l)} = \begin{cases} H^{(l)}(\mathbf{x}), & l = 1 \\ H^{(l)}\left(\mathbf{a}^{(l)}\right), & l \in \{2, 3, \ldots, L\} \end{cases} . \tag{6}$$

Given the static features of DEM and soil modalities, we use these architectures in $E_{\theta_{\mathrm{D}}}$, $E_{\theta_{\mathrm{S}}}$.

4.3. Gated fusion: Adaptive fusion with gating mechanism

The learned modality-representations can be fused with the concatenation merge function, as usual in crop yield prediction (Wang et al., 2020; Shahhosseini et al., 2021; Srivastava et al., 2022):

$$\mathbf{z}_{\mathrm{F}}^{(i)} = \mathsf{C}\left(\mathbf{z}_{\mathrm{S2}}^{(i)}, \mathbf{z}_{\mathrm{W}}^{(i)}, \mathbf{z}_{\mathrm{D}}^{(i)}, \mathbf{z}_{\mathrm{S}}^{(i)}\right) . \tag{7}$$

However, this static merge function does not align with the variable contributions that each modality has in predicting crop yield (Kang et al., 2020; Pathak et al., 2023). Additionally, it ignores the real-time environment of EO, where different phenomena (e.g. clouds in optical images or noise in measurement) can affect the data quality (Ofori-Ampofo et al., 2021; Ferrari et al., 2023). Therefore, inspired by the Gated Unit (GU) modules (Arevalo et al., 2020), we propose an adaptive fusion approach via gating mechanisms, the so-called gated fusion.

We use a gating function $G_{\theta_{\mathrm{G}}} : \mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}^4$ that takes the four modality-representations, and generates four pixel-specific values $\alpha_m^{(i)} \in [0, 1]$ with $m \in \{\mathrm{S2}, \mathrm{W}, \mathrm{D}, \mathrm{S}\}$, which we refer to as **gated fusion weights.** The following computation is used

$$\alpha^{(i)} = G_{\theta_{\mathrm{G}}}\left(\mathbf{z}_{\mathrm{S2}}^{(i)}, \mathbf{z}_{\mathrm{W}}^{(i)}, \mathbf{z}_{\mathrm{D}}^{(i)}, \mathbf{z}_{\mathrm{S}}^{(i)}\right) = \operatorname{softmax}\left(\mathsf{C}\left(\mathbf{z}_{\mathrm{S2}}^{(i)}, \mathbf{z}_{\mathrm{W}}^{(i)}, \mathbf{z}_{\mathrm{D}}^{(i)}, \mathbf{z}_{\mathrm{S}}^{(i)}\right)^{\top} \theta_{\mathrm{G}}\right) , \tag{8}$$

with learnable parameters $\theta_{\mathrm{G}} \in \mathbb{R}^{4d \times 4}$. This GU module learns a weight distribution over the modalities for each sample: $\sum_{m \in \{\mathrm{S2,W,S,D}\}} \alpha_m^{(i)} = 1$. As proposed, this function can be easily extended to use any kind of weight calculation. See Fig. 5 for an illustration. Then, these fusion weights are applied to the stacked vectors to obtain the fused representation $\mathbf{z}_{\mathrm{F}}^{(i)} \in \mathbb{R}^d$:

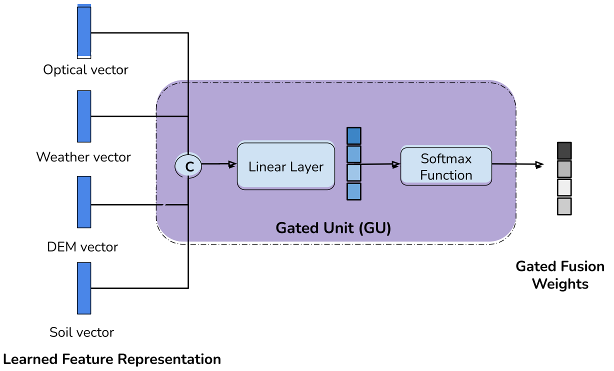


Fig. 5. Illustration of the proposed gating mechanism. The four modality-representations are merged (C) and linearly projected to a four-dimensional vector. Then, the *softmax* function is applied to obtain the normalized fusion weights.

$$\mathbf{z}_{\mathrm{F}}^{(i)} = \mathsf{S}\left(\mathbf{z}_{\mathrm{S2}}^{(i)}, \mathbf{z}_{\mathrm{W}}^{(i)}, \mathbf{z}_{\mathrm{D}}^{(i)}, \mathbf{z}_{\mathrm{S}}^{(i)}\right)^{\top} \alpha^{(i)} = \sum_{m \in \{\mathrm{S2,W,D,S}\}} \alpha_m^{(i)} \cdot \mathbf{z}_m^{(i)} . \tag{9}$$

Thus, the data-driven fusion weights are applied to the modality-representations as an adaptive weighted sum (9). The gating mechanism plays a pivotal role in both weight learning and enabling adaptive fusion for each sample (data-driven). In our approach, a pixel is a sample. Therefore, the model highlights the modalities depending on the information contained in the multi-modal pixel-level data. For example, for a cloudy pixel, the model could learn to assign a lower weight to the optical modality α_{S2} and higher to the other modalities ($\alpha_{\mathrm{W}}, \alpha_{\mathrm{S}}, \alpha_{\mathrm{D}}$). While, for a cloudless pixel, learn to assign a higher weight to the optical modality and distribute the rest to the complementary modalities.

The gated fusion approach introduced in Arevalo et al. (2020) shares similarities with ours, as both methods utilize the concatenation of features, C, inside the GU module that calculates the fusion weights, α_m. However, their proposed fusion weights are individually normalized via a *sigmoid* activation function, i.e. $\alpha_m^{(i)} \in [0, 1]$ and $0 \leq \sum_m \alpha_m^{(i)} \leq 4$, in contrast to our work where the weights are complementary normalized to sum up one. In addition, the weights are applied (as in (9)) after another non-linear mapping in the representations, $\tilde{\mathbf{z}}_m^{(i)} = \tanh(\mathbf{W}_g \mathbf{z}_m^{(i)})$ (Arevalo et al., 2020). In Section 6.2 we show that our design of fusion weights calculation and modality-representation improves the performance of this strategy.

4.4. Prediction and optimization

After obtaining the fused representation, it is fed to fully connected neural network layers. These layers, represented by a function $F_{\theta_{\mathrm{F}}} : \mathbb{R}^d \rightarrow \mathbb{R}^1$ parameterized by θ_{F}, serve as a prediction head to estimate the crop yield for each pixel i: $\hat{y}^{(i)} = F_{\theta_{\mathrm{F}}}\left(\mathbf{z}_{\mathrm{F}}^{(i)}\right)$. Since we are designing a predictive model that is fed with the multi-modal data and has all the previous components, $\hat{y}^{(i)} = P_{\Theta}(\mathcal{X}^{(i)})$ with $\Theta = \left\{\theta_{\mathrm{S2}}, \theta_{\mathrm{W}}, \theta_{\mathrm{D}}, \theta_{\mathrm{S}}, \theta_{\mathrm{G}}, \theta_{\mathrm{F}}\right\}$, we could minimize a loss function $\mathcal{L}$ over training pixels to learn it end-to-end. We use the following

$$\mathcal{L}(\Theta; \mathcal{D}) = \frac{1}{N} \sum_{i=1}^{N} \operatorname{MSE}\left(y^{(i)}, P_{\Theta}(\mathcal{X}^{(i)})\right) , \tag{10}$$

with the Mean Squared Error (MSE) as a loss function, $\operatorname{MSE}(y, \hat{y}) = (y - \hat{y})^2$. We named our model that unifies these different components

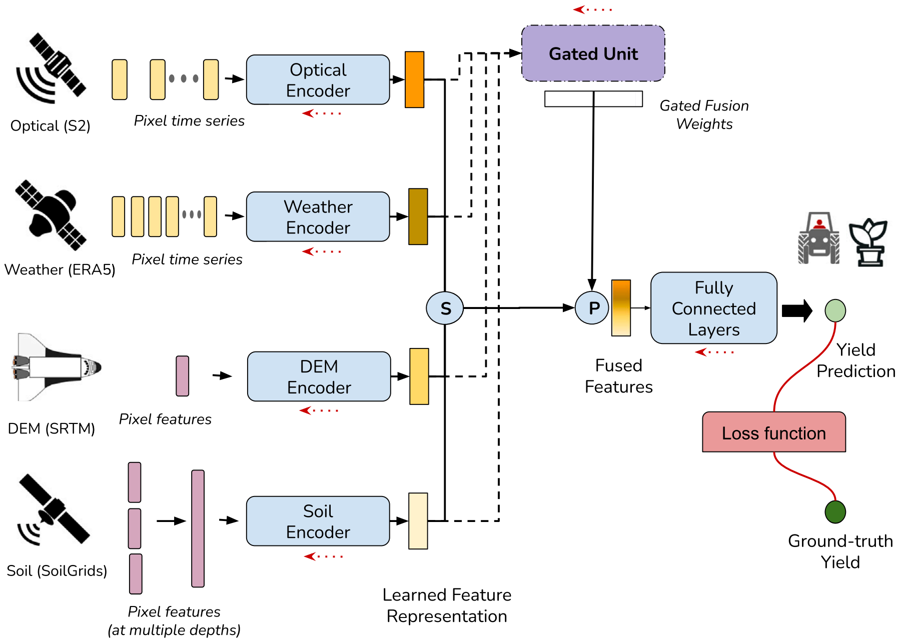


Fig. 6. Illustration of the proposed Multi-Modal Gated Fusion (MMGF) model with the four modalities used. “S” represents the vector stacking operation, and “P” the dot product. The forward pass is shown with a black arrow, while the dotted arrow shows the additional connections for the GU. The model is learned end-to-end by comparing the prediction with the ground truth. The red dotted arrows illustrate the backward pass of the loss function through the model components.

(modality-encoders, the GU, and a prediction head) as Multi-Modal Gated Fusion (MMGF), see Fig. 6 for an illustration.

5. Experiments

5.1. Data preparation and evaluation

For the S2 optical data, we use two different versions. **S2-R** as the raw SITS (including the SCL) up to 5-days temporal resolution. **S2-M** as a monthly-based sampling SITS to use with the input-level fusion as described in Pathak et al. (2023). In the second setting, a sample for each month is selected based on the lowest cloud coverage, across a two calendar year period (24 time-steps). The two years are defined by seeding and harvesting, so that the harvesting month always falls in the second year (Helber et al., 2023).

To avoid overfitting to the different magnitudes and scale of the multi-modal data, we re-scale each numerical band or feature in the input data into a [0,1] range (Zhang et al., 2021): $x_{\text{norm}} = (x - \min(x))/(\max(x) - \min(x))$. For the S2 optical images, we calculate the maximum and minimum value across time and samples in the S2-M representation, which contains cloud-free values. In addition, for dynamic modalities (optical and weather), we pad the sequences with a masked value (−1). For the categorical information contained in the S2-based SCL, we codify the 12 labels as a one-hot encoding vector, with one additional category used in the padded times. In the S2 data, the sequence length is 150 including padding, while for the weather is 500.

For the evaluation, we use a stratified-group K-fold cross validation ($K = 10$). The grouping is based on the field identifier and stratified at farm identifier.[4] This means that all pixels within a field are used either for training or validation. We quantitatively evaluate model performance by using standard regression error metrics. The coefficient of determination (R^2), Mean Absolute Percentage Error (MAPE) in %, and Mean Absolute Error (MAE), in t/ha, are measured between ground-truth yield values, y, and model predictions, $P_\Theta(\mathcal{X})$.

$$R^2 = 1 - \frac{\sum_{i=1}^{N_{\text{val}}} (y^{(i)} - P_\Theta(\mathcal{X}^{(i)}))^2}{\sum_{i=1}^{N_{\text{val}}} (y^{(i)} - \bar{y})^2} \tag{11}$$

$$MAPE = \frac{1}{N_{\text{val}}} \sum_{i=1}^{N_{\text{val}}} \frac{|y^{(i)} - P_\Theta(\mathcal{X}^{(i)})|}{y^{(i)}} \tag{12}$$

$$MAE = \frac{1}{N_{\text{val}}} \sum_{i=1}^{N_{\text{val}}} |y^{(i)} - P_\Theta(\mathcal{X}^{(i)})| \tag{13}$$

N_{val} are the number of samples and $\bar{y}$ is the field average in the validation split. These metrics are evaluated at sub-field level (pixel-wise comparison, N_{val}= number of pixels) and field level (comparison of field-averaged values, N_{val}= number of fields). The aggregated results with the standard deviation across folds are presented.

5.2. Compared methods

As a baseline, we select the most similar approach in sub-field level yield prediction, the Input-level Fusion (IF) proposed in Pathak et al. (2023). For this, the S2-M is used, where the static modalities (DEM and soil features) are vectorized (*flattened*) and repeated for each month along the 24 time-step representation. The weather data is aggregated by summing the daily features between the dates of the selected optical images in each month. This generates a multivariate time-series data, where each time-step and pixel has the features from weather, DEM, and soil concatenated with the S2-M features. Since (Pathak et al., 2023) showed that a subset of the modalities have to be used for a better prediction performance, we follow the best combination that they identified for each dataset. It is important to note that we trained IF with S2-M instead of S2-R as in the MMGF. We do not see a direct way to perform IF with S2-R modality for a fair comparison. In this case, two models are selected: LSTM for IF (LSTM-IF) and Gradient

[4] Four our study, a farm represents either a set of fields operated by a farmer or geographically nearby fields.

Table 3
Sub-field level performance. R^2 of the crop yield prediction at sub-field level for different models and combination of modalities. *The multi-modal data used in IF models are: S2-M and DEM in ARG-S, S2-M and soil in GER-R, and all modalities in URU-S and GER-W. The highest mean in each dataset is highlighted in bold.

Model	Modalities	ARG-S	URU-S	GER-R	GER-W	Overall
LSTM	S2-M	0.61 ±0.11	0.38 ±0.08	0.35 ±0.13	0.32 ±0.09	0.41
LSTM	S2-R	0.67 ±0.05	0.41 ±0.06	**0.46** ±0.11	0.41 ±0.08	0.49
GBDT-IF	S2-M+varies*	0.58 ±0.11	**0.42** ±0.07	0.42 ±0.08	0.37 ±0.10	0.45
LSTM-IF	S2-M+varies*	0.65 ±0.08	0.41 ±0.07	0.45 ±0.10	0.35 ±0.12	0.47
Concat-FF	S2-R+All	0.67 ±0.05	**0.42** ±0.07	0.44 ±0.15	0.43 ±0.10	0.49
Avg-FF	S2-R+All	0.66 ±0.06	**0.42** ±0.06	**0.46** ±0.15	0.42 ±0.13	0.49
MMGF	S2-R+All	**0.68** ±0.05	**0.42** ±0.06	**0.46** ±0.15	**0.44** ±0.10	**0.50**

Table 4
Field-level performance. R^2 of the crop yield prediction at field level for different models and combination of modalities. *The multi-modal data used in IF models are: S2-M and DEM in ARG-S, S2-M and soil in GER-R, and all modalities in URU-S and GER-W. The highest mean in each dataset is highlighted in bold.

Model	Modalities	ARG-S	URU-S	GER-R	GER-W	Overall
LSTM	S2-M	0.74 ±0.12	0.69 ±0.14	0.65 ±0.18	0.60 ±0.25	0.67
LSTM	S2-R	**0.84** ±0.08	0.77 ±0.09	0.77 ±0.13	0.72 ±0.11	0.78
GBDT-IF	S2-M+varies*	0.72 ±0.14	0.77 ±0.08	0.69 **±0.09**	0.68 ±0.12	0.71
LSTM-IF	S2-M+varies*	0.82 ±0.12	0.74 ±0.12	0.78 **±0.09**	0.66 ±0.28	0.75
Concat-FF	S2-R+All	0.83 ±0.10	0.77 ±0.10	0.77 ±0.11	0.77 ±0.10	0.78
Avg-FF	S2-R+All	0.82 ±0.14	**0.78** ±0.08	0.78 ±0.13	0.74 ±0.22	0.78
MMGF	S2-R+All	**0.84** ±0.11	**0.78** ±0.07	**0.80** ±0.13	**0.80** ±0.09	**0.80**

Boosting Decision Tree for IF (GBDT-IF). The GBDT-IF model is fed with flattened (across time) vectors (Feng et al., 2021). In addition, to evaluate the benefit of the MML scenario, two single-modal models are used. This corresponds to a model with a LSTM encoder that is feed with either S2-M or S2-R data.

We consider alternative feature-level fusion approaches for sub-field level yield prediction. We use the feature-level fusion with the commonly employed concatenation as merge function (Concat-FF). In addition, we consider the sum of features with a constant uniform weight (i.e. feature average) as a merge function in the feature-level fusion (Avg-FF).

Implementation. The MMGF model uses different modality-encoders. For S2 and weather modalities, we use RNN-based modality-encoders consisting of two LSTM layers with 128 units each. We show a few experiments with the Transformer model in Section 6.2 without achieving significant improvements. For DEM and soil we use MLP as encoders, with one hidden layer of 128 units. On each modality-encoder, there is an output linear layer projecting the data to $d = 128$ dimensions. For the prediction head, we use an MLP with one hidden layer of 128 units and an output layer with a single unit. As suggested by previous works (Chen et al., 2017; Maimaitijiang et al., 2020), we include 30% of dropout on the modality-encoders and Batch-Normalization (BN) on all MLPs. In Table A.12 we show the relevance of this regularization in our model. With this implementation, the LSTM in the single-modal model with S2-R has 228K parameters. The LSTM of the LSTM-IF increased the number of parameters to 238K, while for the MMGF, the parameters of the four LSTM reach to 483K, and 2.1K in the GU. These models are trained a maximum of 50 epochs, with an early stopping criterion in the loss function (MSE). This function is optimized with ADAM (Kingma and Ba, 2015), a batch-size of 1024, a learning rate of 10^{-3}, and a weight decay of 10^{-4}. The hyper-parameters were tuned in the ARG-S data. The GBDT-IF is implemented in the LightGBM Python library and all other neural networks with the PyTorch library.

5.3. Quantitative results

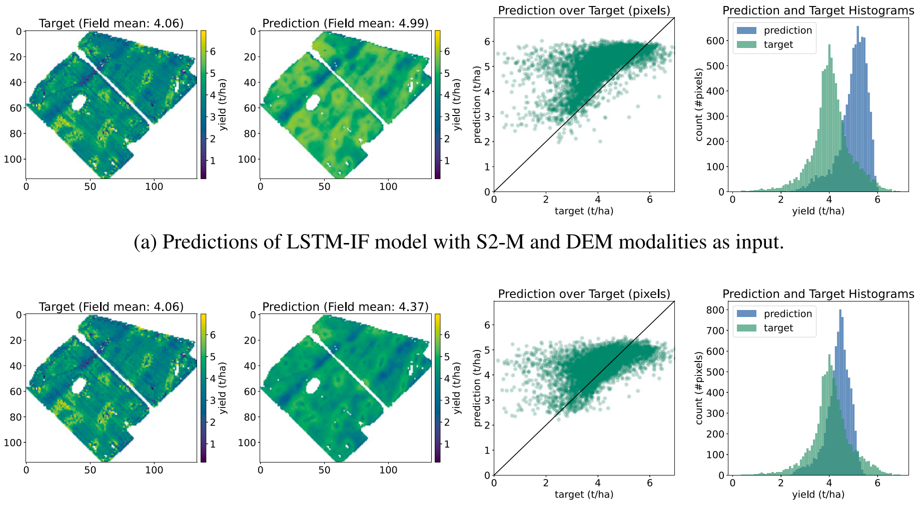
The aggregated R^2 results for all datasets are displayed in Table 3 for sub-field level and in Table 4 for field level. We observe that the proposed MMGF obtains the best performance across all datasets and metrics regarding the compared methods, as indicated in the Overall column, that correspond to the average metrics across datasets. The same evidence is observed in the other regression metrics (See Tables A.8, A.9, A.10, A.11 in). The R^2 is around 0.80 across all datasets at the field level, while at sub-field level it is 0.68 for ARG-S and around 0.44 for the rest. The lower values of sub-field level R^2 compared to field level reflects the complexity of predicting the crop yield to the high spatial resolution of 10 m/px.

When comparing the MMGF to the single-modal models, we observe both high and small prediction improvements, reflecting that the benefit obtained from the MML scenario relies on each setting. The best-performing combination of modalities for the MMGF is obtained by using all modalities (Table 5), in contrast to IF models, where the best-performing combination is obtained with a sub-set of them, depending on the country and crop-type (Pathak et al., 2023). For instance, for ARG-S the best combination is S2-M and DEM modalities, and for GER-R it is S2-M and soil, we compare this in more detail in Section 6.2. Compared to other FF models, the improvement of the MMGF is minor, in some cases the results are the same. However, the best baseline model depends on the dataset and evaluation metric. For instance, Concat-FF is among the bests at sub-field level R^2 in URU-S, and Avg-FF is among the bests at field-level R^2 in URU-S. In contrast, the MMGF consistently obtains the best results overall cases. Furthermore, we notice that the proposed MMGF performs better in cloudy fields (Fig. A.11), as well as over unseen years (Fig. A.12). These results exhibit the effectiveness of the MMGF to adaptively fuse the features based on the information of each sample. In this way, the practitioner can avoid expending time on trying different fusion strategies and modalities combination for each particularly setting.

Different results are obtained in each dataset. In GER-W fields is where the MMGF obtains the greatest improvements as compared to other fusion methods. Regarding the LSTM-IF, there are 0.07 and 0.12 points of improvement for sub-field and field R^2. On the other hand, in ARG-S and GER-R fields minor improvements are obtained, around 0.02 and 0.01 points in R^2. A similar pattern is observed when comparing the MMGF with single-modal models. In addition, we show the statistical significance of this improvements in . This difference is statistically significant only in some cases, mainly between the MMGF with the IF

Table 5
Multi-modal combinations. Crop yield prediction performance in ARG-S fields with different modalities as input data. The LSTM indicates the single-modal model with the optical data (S2). The best results are in bold.

Model	Modalities	*Sub-Field*			*Field*		
		MAPE (%)	MAE (t/ha)	R^2 –	MAPE (%)	MAE (t/ha)	R^2 –
LSTM	S2-M without SCL	25	0.69	0.61	11	0.40	0.74
	S2-M with SCL	25	0.68	0.61	11	0.39	0.75
	S2-R cloudless	24	0.66	0.63	9	0.33	0.81
	S2-R without SCL	23	0.62	0.67	9	0.31	0.82
	S2-R with SCL	23	0.62	0.67	9	0.31	**0.84**
LSTM-IF	S2-R+weather	25	0.67	0.63	11	0.38	0.78
	S2-R+DEM	24	0.65	0.65	9	0.33	0.82
	S2-R+soil	25	0.68	0.61	10	0.37	0.76
	S2-R+weather, DEM, soil	24	0.66	0.63	11	0.40	0.76
Concat-FF	S2-R+weather	**22**	0.62	0.66	10	0.34	**0.84**
	S2-R+DEM	23	0.63	0.66	10	0.34	0.81
	S2-R+soil	23	0.65	0.65	10	0.34	0.80
	S2-R+weather, DEM, soil	**22**	0.62	0.67	9	0.32	0.83
MMGF	S2-R+weather	**22**	0.62	0.67	9	0.32	0.83
	S2-R+DEM	**22**	0.62	0.67	9	**0.30**	**0.84**
	S2-R+soil	23	0.63	0.66	9	0.32	0.83
	S2-R+weather, DEM, soil	**22**	**0.61**	**0.68**	**8**	**0.30**	**0.84**



(a) Predictions of LSTM-IF model with S2-M and DEM modalities as input.

(b) Predictions of MMGF model with S2-R, weather, DEM, and soil modalities as input.

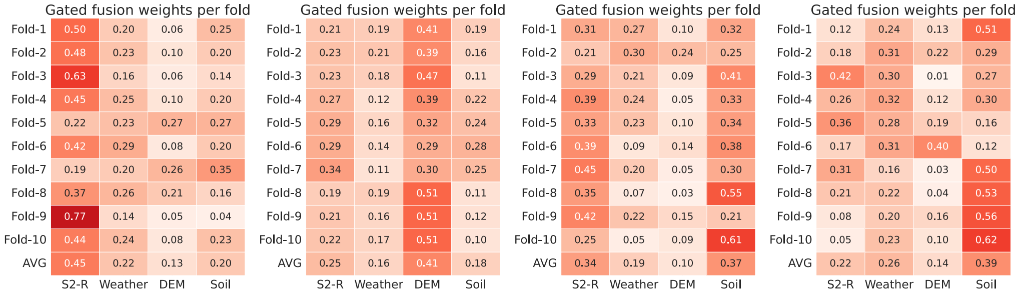
Fig. 7. Sub-field crop yield prediction for a random ARG-S field. The columns from left to right are the ground truth yield map, the predicted yield map, prediction and target scatter, plot of prediction (blue) and target (green) distribution.

and single-modal models. In addition, we notice that despite ARG-S and URU-S have the same crop-type, the results are quite different. The models in URU-S, with more fields and data pixels, perform worse than in ARG-S. This might be caused by the different geographic patterns in each region. However, the error can also be associated with the data quality and internal noise factors, as the label providers from each country are different (Section 3.1). The datasets also influence the model convergence. Whereas LSTM-IF converges in average (across folds) in less than 23 epochs, MMGF need 29, 25, 39, and 36 for ARG-S, URU-S, GER-R and GER-W respectively.

5.4. Qualitative results

In Fig. 7 we compare the yield map predictions of two fusion models (LSTM-IF and MMGF) for a single field in ARG-S data. The purpose is to qualitative inspect if the model learns the in-field variability. The proposed MMGF (Fig. 7(b)) predicts a better yield map (first and second column in the plot) than the LSTM-IF model (Fig. 7(a)). Indeed, the LSTM-IF tends to predict higher yield values for the selected field (third column in the plot), while the predictions of the MMGF are closer to the target data. This is also observed with the better yield distribution alignment (fourth column in the plot) by the MMGF model (Fig. 7(b)). The same evidence is shown in other datasets and fields (see Fig. B.14, B.15, B.16, and B.17 for some examples).

We visualize the fusion weights α_m computed by the GU (from (8) in Section 4.3) to analyze what the MMGF model learned. To get a general overview, we compare the fusion weights distribution across the 10 validation folds in Fig. 8. We observe that the weights are not perfectly aligned across folds, mainly because each validation fold contains a unique set of fields, and the fusion weights are distributed differently from one field to another (Fig. B.18). This visualization suggests that the MMGF model assigns higher fusion weights to different modalities, within and across the datasets: optical modality is particularly high in the Argentina fields, DEM modality is predominant in Uruguay, while soil and weather modalities share high weights in the crops of German fields. This crop-dependent behavior has also been observed in the



(a) Fields in ARG-S data. (b) Fields in URU-S data. (c) Fields in GER-R data. (d) Fields in GER-W data.

Fig. 8. Summary of fusion weights averaged across all pixels in each fold. The 10 folds used for evaluation in the cross-validation are displayed for the fields in each dataset. The intensity of the color increases from 0 to 1.

attention weights computed along the temporal dimension for crop classification (Garnot and Landrieu, 2020).

6. Analysis

In the following, we present an analysis of the data-driven gated fusion weights in Section 6.1, and an ablation study in Section 6.2.

6.1. Gated fusion weights analysis with linear regression

We further investigate the influence of the gated fusion weights in the GU, as well as the contribution of each modality in the final prediction, by retraining a variation of the MMGF model. For this, we replaced the MLP in the prediction head (see Section 4.4) with a linear layer with weight parameters $\mathbf{w} \in \mathbb{R}^d$ and bias parameter $b \in \mathbb{R}^1$. We refer to this model as MMGF with Linear Regression (MMGF-LR). The prediction head (now a linear regression) allows us to interpret the contribution of each modality. Given the fused representation $\mathbf{z}_\mathrm{F} \in \mathbb{R}^d$, the modality-representations $\{\mathbf{z}_{\mathrm{S2}}, \mathbf{z}_\mathrm{W}, \mathbf{z}_\mathrm{D}, \mathbf{z}_\mathrm{S}\}$ and the corresponding gated fusion weights $\{\alpha_{\mathrm{S2}}, \alpha_\mathrm{W}, \alpha_\mathrm{D}, \alpha_\mathrm{S}\}$, the expression for the crop yield prediction can be written as follows:

$$\hat{y} = \mathbf{w}^\top \cdot \mathbf{z}_\mathrm{F} + b = \sum_{i=1}^{d} w_i \times z_{\mathrm{F},i} + b, \tag{14}$$

where, if we expand $\mathbf{z}_\mathrm{F}$ as the weighted sum of the modality-representations (9), we obtain:

$$\hat{y} = \sum_m \mathbf{w}^\top \cdot \alpha_m \mathbf{z}_m + b \tag{15}$$

$$= \alpha_{\mathrm{S2}} \mathbf{w}^\top \cdot \mathbf{z}_{\mathrm{S2}} + \alpha_\mathrm{W} \mathbf{w}^\top \cdot \mathbf{z}_\mathrm{W} + \alpha_\mathrm{D} \mathbf{w}^\top \cdot \mathbf{z}_\mathrm{D} + \alpha_\mathrm{S} \mathbf{w}^\top \cdot \mathbf{z}_\mathrm{S} + b. \tag{16}$$

The prediction in Eq. (16) is like a weighted combination of the prediction of individual modalities by using a shared linear regressor (with $\mathbf{w}$ and b). As this weighted combination is based on a gated layer, this expression is similar to mixture of expert models (Yuksel et al., 2012). Furthermore, we can consider $C_m = \mathbf{w}^\top \cdot \mathbf{z}_m$ and $\alpha_m C_m$ as the scalar contribution ($\in \mathbb{R}^1$) to the prediction by a particular modality m before and after applying the fusion weights respectively. Using this simplified version of the model (MMGF-LR) allows to mathematically express the crop yield predictions in terms of individual modality-representations and data-driven fusion weights, making the fusion process simpler to track back and interpret.

To mitigate the variability in the results stemming from the multiple repetitions, we analyze a single fold in each data — the one yielding the highest sub-field R^2 score. We illustrate the α_m, C_m, and $\alpha_m C_m$ aggregated at field level[5] for ARG-S fields in Fig. 9. This analysis suggests that the learned fusion weights (α_m) linearly scales the contribution of each modality (C_m) and yet allows each modality-encoders to learn representations ($\mathbf{z}_m$) with distinct scales. This experiment demonstrates that the GU module has a tendency to assigns higher weights to the optical modality compared to the others. However, we can see how the learned data-driven fusion weights complement each other in some fields. For instance, when the GU assigns a small value to the optical modality (Fig. 9(a)), the soil modality has a larger fusion weights (with the DEM modality in some cases). This is also reflected in the total modality contribution ($\alpha_m C_m$) despite a significant modality-representation contribution (C_m) of DEM and weather modalities (Fig. 9(a) middle and last row). We observe a similar pattern in the other crop types (see Fig. B.19 for GER-W and GER-R). We reorder the ARG-S fields according to the yield prediction value of the MMGF-LR model in Fig. 9(b). We notice a strong relationship between the yield predicted value and the fusion weights. When predicting a low yield value, the MMGF-LR model tends to use more information from static modalities, with high total representation contribution ($\alpha_m C_m$) to DEM and soil. Thus, when predicting a high yield value, the MMGF-LR model tends to use more information the temporal modalities (S2-R and weather).

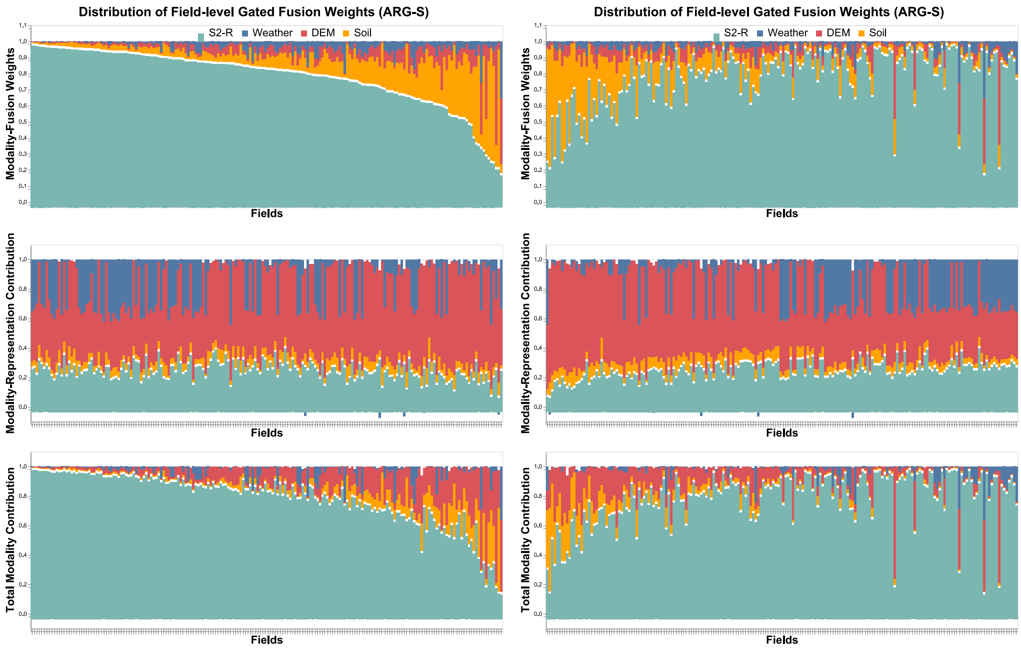
Similarly, we display the α_m, C_m, and $\alpha_m C_m$ for the different datasets in Fig. 10. We notice that for soybean crops (in Argentina and Uruguay), the model tends to assign a higher fusion weights and total contribution to the S2-R modality compared to wheat and rapeseed (in Germany). Similarly, additional modalities have a higher fusion weights and total contribution in the German crops compared to soybean crops. For instance, while weather modality is contributing more in wheat crops, soil (and DEM) does it in rapeseed crops. These results reflect that the MMGF-LR model tends to use S2-R modality for soybean prediction, while in rapeseed and wheat prediction it uses auxiliary modalities for a better modeling. This can be caused by the high cloud coverage in German fields compared to soybean fields (Fig. 3).

6.2. Ablation study

In the following, we present different experiments with the proposed MMGF. The main purpose is to assess the impact of the individual components that contribute to the overall model performance. We focus on the cross-validation setting with the ARG-S fields.

In Table 5 we include the results obtained with different multimodal combinations. These results illustrate that the MMGF model is stable when feeding all the heterogeneous modalities. There are other multi-modal combinations and models that also obtain the best performance in some metrics. However, the best result across all scenarios is obtained when feeding all data to our MMGF model. In contrast to some models compared and previous results in the RS literature (Bocca and Rodrigues, 2016; Kang et al., 2020; Pathak et al., 2023), the MMGF model does not saturate through giving more input data. This effect could be attributed to the GU in the gated fusion approach, since it could select which modalities to merge depending on each sample, in contrast to the static merge function of Concat-FF that obtains lower

[5] We assume that within each field, fusion weights for different modalities are roughly aligned, as shown in Fig. B.18.



(a) Order based on descending weight given to S2-R. (b) Order based on ascending yield prediction value.

Fig. 9. Fusion weights and contributions of the MMGF-LR model over different ARG-S fields. Gated fusion weights from the GU (α_m) at the top, the contribution before applying the fusion weights (C_m) at the middle, and the final contribution in the predicted yield ($\alpha_m C_m$) at the bottom. The sub-field values are averaged per field. The C_m and $\alpha_m C_m$ are scaled by $1/\sum_m |C_m|$ and $1/\sum_m |\alpha_m C_m|$ respectively into a $[-1, 1]$ range. The field bars are sorted (from left to right) by (a) descending weight given to the predominant modality, and (b) ascending yield prediction value.

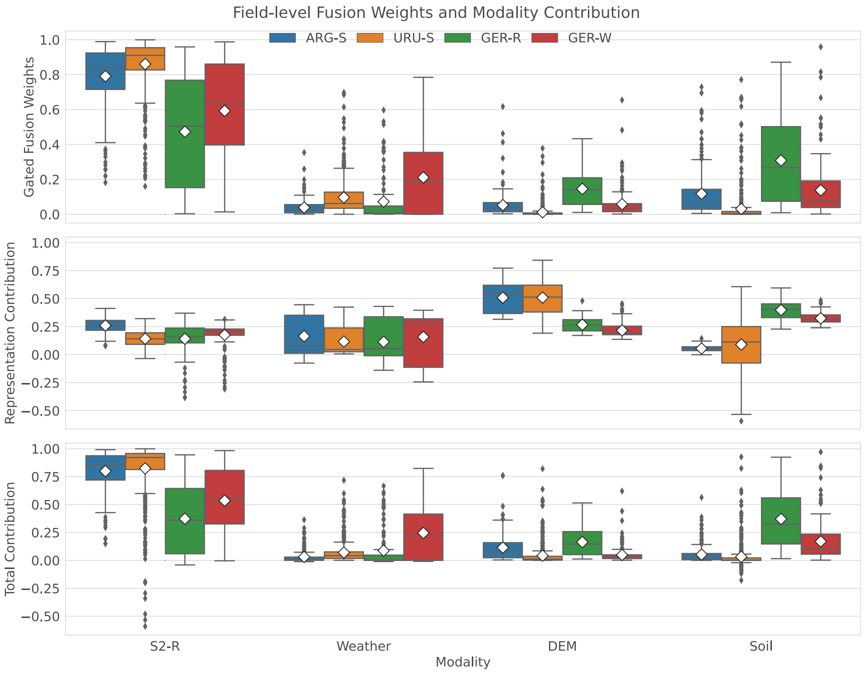


Fig. 10. Fusion weights and contributions of the MMGF-LR model over different country-crop combinations. Gated fusion weights from the GU (α_m) at the top, the contribution before applying the fusion weights (C_m) at the middle, and the final contribution in the predicted yield ($\alpha_m C_m$) at the bottom. The sub-field values are averaged per field using the best performance fold. The C_m and $\alpha_m C_m$ are scaled by $1/\sum_m |C_m|$ and $1/\sum_m |\alpha_m C_m|$ respectively into a range $[-1, 1]$.

Table 6
Merge function alternatives. Crop yield prediction performance in ARG-S fields with different types of merge functions in the feature-level fusion. The adaptive fusion correspond to a weighted sum with weights computed by different criterion. The MMGF configuration correspond to the last row. The best results are in bold.

Approach	Merge	*Sub-Field*			*Field*		
		MAPE (%)	MAE (t/ha)	R^2 –	MAPE (%)	MAE (t/ha)	R^2 –
Feature Fusion	Product	23	0.64	0.65	10	0.34	0.81
	Maximum	23	0.63	0.67	10	0.34	0.81
	Weighted Concatenation	24	0.66	0.63	11	0.38	0.78
Adaptive Fusion	Scaled-dot Product Attention	23	0.63	0.67	10	0.33	0.82
	Arevalo et al. (2020) version	23	0.62	0.67	9	**0.30**	0.83
	Gated Fusion per feature	23	0.63	0.66	9	0.33	0.81
	Gated Fusion	**22**	**0.61**	**0.68**	**8**	**0.30**	**0.84**

Table 7
Encoder alternatives. Crop yield prediction performance in ARG-S fields by varying the temporal modality-encoder between LSTM and Transformer in the MMGF model. The selected MMGF configuration is the one with LSTM. The best results are in bold.

Model	Modalities	*Sub-Field*			*Field*		
		MAPE (%)	MAE (t/ha)	R^2 –	MAPE (%)	MAE (t/ha)	R^2 –
Transformer	S2-R	23	0.63	0.66	10	0.35	0.78
→ MMGF	S2-R+All	**22**	**0.61**	0.67	10	0.35	0.78
LSTM	S2-R	23	0.62	0.67	9	0.31	**0.84**
→ MMGF	S2-R+All	**22**	**0.61**	**0.68**	**8**	**0.30**	**0.84**

results. The motivation of the GU comes from whether it allows passing the information through a channel (Arevalo et al., 2020), so it can reduce non-relevant information and focus on the most important one. This effect is observed in the other datasets as well.

Motivated by the work of Mena et al. (2023) in crop classification, different merge functions are compared in Table 6. The product and maximum replaced the weighted sum (C in (7)) as static merge functions. In addition, we consider applying learnable weights followed by a concatenation (instead of summing), inspired by Ma et al. (2023), Feng et al. (2021). As adaptive fusion alternatives, we include the vanilla GU proposal with the sigmoid activation function (Arevalo et al., 2020), and a weighted sum with the weights calculated via the scaled-dot product attention (Vaswani et al., 2017). In the scaled-dot product attention, a learnable query is used to calculate attention weights and pull over the multi-modal representations. In addition, we show a per-feature gated fusion which calculates and applies fusion weights on each feature for each modality (feature-specific) instead of the global fusion weights for each modality. The feature-specific fusion weights may suffer from overfitting since there is an increase in the number of parameters in the GU module, from 2.1K (in the global) to 262K (in the feature-specific). Overall, the proposed gated fusion obtains the best predictions compared to these alternative merge functions, illustrating the effectiveness of the MMGF for crop yield prediction.

Table 7 shows the results of the LSTM being replaced with the Transformer model (Vaswani et al., 2017) as the encoders of S2 and weather modalities. Similar to Vision Transformer (Dosovitskiy et al., 2020), we use a class token to extract a single representation from the complete time series. In sub-field level metrics, the MMGF with Transformer has a similar performance to LSTM, contrary to field-level, where the LSTM has the best performance. The similarity in performance between Transformer-based model and neural networks without attention is something already observed in MML with RS data (Zhao and Ji, 2022).

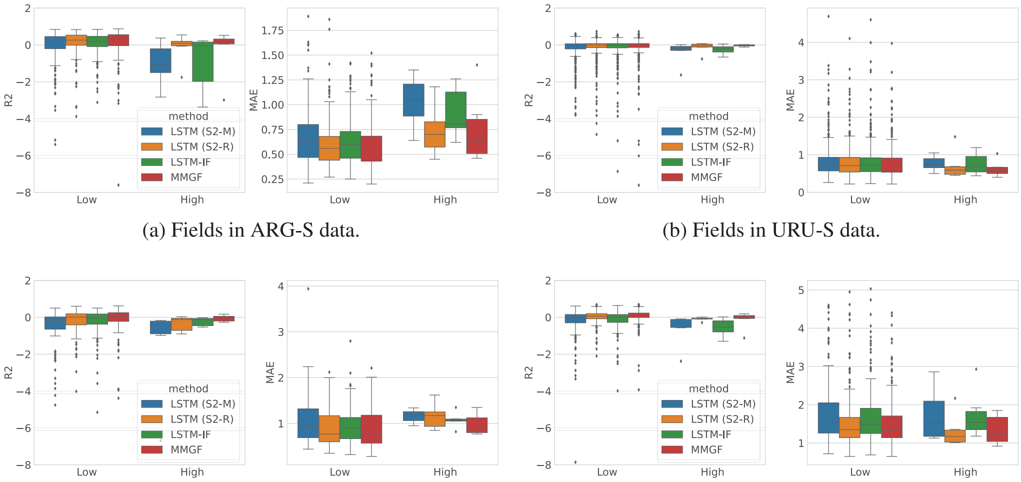
7. Final remarks

We present a study of crop yield prediction at a sub-field level (10 m) by leveraging a variety of RS sources: optical, weather, soil, and DEM. Our model, named Multi-Modal Gated Fusion (MMGF), comprises two integral components. The first one involves feature-level learning, where dedicated encoders learn high-level representations for each modality, effectively handling varying temporal resolutions and data distributions. The second component involves gated fusion, where a gating mechanism (the GU) learns data-driven weights to fuse multi-modal features adaptively, allowing customized aggregations depending on input data. In addition, the GU module allows a simple interpretation of what was learned by the model. Our evaluation on four country-crop combinations shows that our MMGF achieves optimal crop yield predictions when all modalities are used. Such a consistent outcome is uncommon in yield prediction (Bocca and Rodrigues, 2016; Kang et al., 2020; Pathak et al., 2023; Ma et al., 2023), where factors such as region and modalities often shape the effectiveness of solutions (Mena et al., 2024).

Notably, results at the sub-field level underscore the increased complexity of the crop yield prediction compared to the field level. This insight prompts future researchers to prioritize this scenario, such as with image-based or neighborhood-based mapping. Furthermore, given the valuable interpretability of the data-driven gated fusion weights, future research will explore this for enhancing model explainability. Thus, some researchers (Jain and Wallace, 2019; Wiegreffe and Pinter, 2019) have already challenged the conventional notion of using attention weights as feature importance scores. Their framework is a promising future step towards explainability in multi-modal crop yield prediction.

Abbreviations

BN	Batch-Normalization
CNN	Convolutional Neural Network
DEM	Digital Elevation Model
GU	Gated Unit
LSTM	Long-Short Term Memory
MLP	Multi-Layer Perceptron
ML	Machine Learning
MML	Multi-Modal Learning
RS	Remote Sensing
RNN	Recurrent Neural Network
SCL	Scene Classification Layer
SITS	Satellite Image Time Series
S2	Sentinel-2



(a) Fields in ARG-S data.

(b) Fields in URU-S data.

(c) Fields in GER-R data.

(d) Fields in GER-W data.

Fig. A.11. Results at different cloud coverage. Field-level R^2 and MAE are separated into two cloud coverage categorization. "High" considers fields with cloud coverage above the 5th highest cloud coverage value, while "Low" considers fields with a cloud coverage same or below that threshold.

CRediT authorship contribution statement

Francisco Mena: Writing – original draft, Visualization, Software, Methodology, Investigation, Conceptualization. **Deepak Pathak:** Writing – review & editing, Visualization, Validation, Software, Methodology, Conceptualization. **Hiba Najjar:** Writing – review & editing, Visualization, Validation, Software. **Cristhian Sanchez:** Software. **Patrick Helber:** Visualization, Validation, Software, Funding acquisition, Data curation. **Benjamin Bischke:** Visualization, Validation, Software, Funding acquisition, Data curation. **Peter Habelitz:** Visualization, Validation, Software. **Miro Miranda:** Data curation. **Jayanth Siddamsetty:** Writing – review & editing. **Marlon Nuske:** Writing – review & editing, Supervision, Project administration, Funding acquisition, Conceptualization. **Marcela Charfuelan:** Writing – review & editing, Project administration, Data curation. **Diego Arenas:** Writing – review & editing, Supervision. **Michaela Vollmer:** Project administration. **Andreas Dengel:** Supervision, Resources, Funding acquisition.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The research results presented are part of a large collaborative project on agricultural yield predictions, which was partly funded through the ESA InCubed Programme (https://incubed.esa.int/) as part of the project AI4EO Solution Factory (https://www.ai4eo-solution-factory.de/). F. Mena and H. Najjar acknowledge support through a scholarship of the University of Kaiserslautern-Landau. The research carried out in this work was partly funded by companies, BASF Digital Farming, John Deere, and MunichRe.

Appendix A. Extended results

In this section, we present additional quantitative results to complement the analysis.

Results per dataset. The crop yield prediction performance for different metrics (MAE, MAPE, R^2) at field and sub-field level is presented in Table A.8 for ARG-S fields, Table A.9 for URU-S fields, Table A.10 for GER-R fields, and Table A.11 for GER-W fields. Depending on the metric, different models get better or worse performance. In addition, we assess the statistical significance of the MMGF model by comparing it against other models with the Wilcoxon signed-rank test (Wilcoxon, 1992) in the paired 10-fold cross validation experiments.

Results per cloud coverage. We split the validation fields into their field cloud coverage (see Section 3.2.1). The field with the 5th highest cloud coverage is selected as a threshold in each dataset, therefore the high/low categorization is relative to the country and crop-type. Fig. A.11 displays R^2 and MAE metrics for these grouped fields, comparing the LSTM-based models to our model. For all the datasets, it can be seen that the prediction performance in high cloud coverage fields is similar between the models, and that the main difference (and greater error) is coming from the fields with more occluded optical images. In these cloudy fields, the proposed MMGF takes the most advantage of the additional information given in the multi-modal data, obtaining errors among the lowest for the compared methods. This suggests that a better way of combining the multi-modal data particularly benefits the case where one of the modalities has a higher chance of missing information (due to occlusion).

Results per year. We perform a leave-one-year-out (LOYO) cross-validation experiment to compare the prediction for each year in the datasets. In this experiment, all fields with a harvesting date[6] in a specific year were chosen for validation. In addition, it is important to note that not every crop field is available for a farm throughout the years (see Fig. B.13). Here, we compare LSTM-IF and MMGF in the ARG-S fields with MAPE as a score that normalizes the magnitude of the target variable, illustrated in Fig. A.12. We observe that the MMGF model performs better than the LSTM-IF, as the previous evaluation reflects. This holds for the results across all years except for 2017 and 2021 at field-level metrics. Whereas the MMGF model obtains the best yield prediction performance in the most recent year (2022), the most difficult years for prediction are 2018 at the field level and 2021 at the sub-field level. As a reference, the 10-fold cross-validation mean performance is shown with dashed lines. Since the models perform

[6] Please note that the growing season of a crop could extend from one year to the following.

Table A.8
Crop yield prediction performance in the **ARG-S** fields for different models and combinations of modalities. The best result is in bold. The p-value shows the statistical significance of the MMGF with respect to other methods.

		Sub-Field				*Field*			
Model	Modalities	MAPE	MAE	R^2	R^2	MAPE	MAE	R^2	R^2
		(%)	(t/ha)	-	p-value	(%)	(t/ha)	-	p-value
LSTM	S2-M	25 ±4	0.69 ±0.07	0.61 ±0.11	0.002	11 ±3	0.40 ±0.09	0.74 ±0.12	0.010
LSTM	S2-R	23 ±3	0.62 ±0.05	0.67 ±0.06	0.092	9 ±3	0.31 ±0.07	**0.84** ±0.08	0.341
GBDT-IF	S2-M+DEM	27 ±5	0.72 ±0.07	0.58 ±0.11	0.001	12 ±3	0.71 ±0.07	0.72 ±0.14	0.012
LSTM-IF	S2-M+DEM	24 ±3	0.65 ±0.05	0.65 ±0.08	0.010	9 ±3	0.33 ±0.08	0.82 ±0.12	0.312
Concat-FF	S2-R+All	**22** ±3	0.62 ±0.06	0.67 ±0.05	0.098	9 ±2	0.32 ±0.08	0.83 ±0.10	0.138
Avg-FF	S2-R+All	23 ±4	0.62 ±0.08	0.66 ±0.06	0.082	9 ±3	0.32 ±0.11	0.82 ±0.14	0.116
MMGF	S2-R+All	**22** ±3	**0.61** ±0.07	**0.68** ±0.05		**8** ±2	**0.30** ±0.11	**0.84** ±0.08	

Table A.9
Crop yield prediction performance in the **URU-S** fields for different models and combinations of modalities. The best result is in bold. The p-value shows the statistical significance of the MMGF with respect to other methods.

		Sub-Field				*Field*			
Model	Modalities	MAPE	MAE	R^2	R^2	MAPE	MAE	R^2	R^2
		(%)	(t/ha)	-	p-value	(%)	(t/ha)	-	p-value
LSTM	S2-M	100 ±18	0.80 ±0.07	0.38 ±0.08	0.001	21 ±3	0.40 ±0.06	0.69 ±0.14	0.030
LSTM	S2-R	95 ±14	**0.77** ±0.06	0.69 ±0.14	0.032	**18** ±3	**0.35** ±0.06	0.77 ±0.09	0.419
GBDT-IF	S2-M+All	102 ±17	0.78 ±0.06	**0.42** ±0.07	0.081	20 ±4	**0.35** ±0.06	0.77 ±0.08	0.216
LSTM-IF	S2-M+All	99 ±15	0.78 ±0.06	0.41 ±0.06	0.010	20 ±2	0.37 ±0.06	0.74 ±0.12	0.076
Concat-FF	S2-R+All	95 ±15	**0.77** ±0.06	**0.42** ±0.07	0.267	19 ±3	**0.35** ±0.06	0.77 ±0.10	0.179
Avg-FF	S2-R+All	96 ±16	**0.77** ±0.07	**0.42** ±0.06	0.267	19 ±5	**0.35** ±0.06	**0.78** ±0.08	0.439
MMGF	S2-R+All	**94** ±14	**0.77** ±0.06	**0.42** ±0.06		19 ±3	**0.35** ±0.05	**0.78** ±0.07	

Table A.10
Crop yield prediction performance in the **GER-R** fields for different models and combinations of modalities. The best result is in bold. The p-value shows the statistical significance of the MMGF with respect to other methods.

		Sub-Field				*Field*			
Model	Modalities	MAPE	MAE	R^2	R^2	MAPE	MAE	R^2	R^2
		(%)	(t/ha)	-	p-value	(%)	(t/ha)	-	p-value
LSTM	S2-M	42 ±16	1.01 ±0.21	0.35 ±0.13	0.014	18 ±11	0.60 ±0.16	0.65 ±0.18	0.006
LSTM	S2-R	**36** ±12	**0.90** ±0.18	**0.46** ±0.11	0.312	15 ±8	0.51 ±0.17	0.77 ±0.13	0.080
GBDT-IF	S2-M+soil	40 ±14	0.94 ±0.19	0.42 ±0.08	0.188	18 ±11	0.58 ±0.18	0.69 ±0.09	0.042
LSTM-IF	S2-M+soil	39 ±12	0.93 ±0.19	0.45 ±0.10	0.348	15 ±8	0.47 ±0.14	0.78 ±0.09	0.461
Concat-FF	S2-R+All	37 ±11	0.93 ±0.20	0.44 ±0.15	0.348	14 ±8	0.49 ±0.18	0.77 ±0.11	0.278
Avg-FF	S2-R+All	**36** ±11	0.91 ±0.18	**0.46** ±0.15	0.539	14 ±8	0.48 ±0.19	0.78 ±0.13	0.577
MMGF	S2-R+All	**36** ±9	**0.90** ±0.15	**0.46** ±0.17		**13** ±7	**0.43** ±0.15	**0.80** ±0.13	

Table A.11
Crop yield prediction performance in the **GER-W** fields for different models and combinations of modalities. The best result is in bold. The p-value shows the statistical significance of the MMGF with respect to other methods.

		Sub-Field				*Field*			
Model	Modalities	MAPE	MAE	R^2	R^2	MAPE	MAE	R^2	R^2
		(%)	(t/ha)	-	p-value	(%)	(t/ha)	-	p-value
LSTM	S2-M	30 ±5	1.79 ±0.22	0.32 ±0.09	0.001	10 ±4	0.91 ±0.23	0.60 ±0.25	0.005
LSTM	S2-R	**27** ±4	1.59 ±0.26	0.41 ±0.08	0.084	8 ±3	0.72 ±0.24	0.72 ±0.11	0.057
GBDT-IF	S2-M+All	29 ±4	1.76 ±0.24	0.37 ±0.10	0.005	10 ±2	0.86 ±0.22	0.68 ±0.12	0.002
LSTM-IF	S2-M+All	29 ±6	1.71 ±0.28	0.35 ±0.12	0.001	9 ±3	0.84 ±0.19	0.66 ±0.28	0.019
Concat-FF	S2-R+All	**27** ±5	**1.58** ±0.26	0.43 ±0.10	0.236	8 ±2	0.70 ±0.16	0.77 ±0.10	0.116
Avg-FF	S2-R+All	**27** ±5	1.60 ±0.25	0.42 ±0.13	0.193	8 ±3	0.71 ±0.22	0.74 ±0.22	0.116
MMGF	S2-R+All	**27** ±2	**1.58** ±0.16	**0.44** ±0.09		**7** ±5	**0.64** ±0.26	**0.80** ±0.10	

better in the standard cross-validation, it illustrates the difficulty of the LOYO evaluation. This means that learning to predict an unseen year is harder than focusing on a random prediction.

Regularization effect. In Table A.12 different regularization techniques are compared. The results prove that different techniques are essential to avoid model overfitting, especially when having several parameters in the MMGF model. Concretely, BN contributes more to improving prediction performance than Dropout. However, when both techniques are combined, the regularization techniques boost their individual improvements and obtain the best result.

Appendix B. Further visualization

Some additional charts are included in this section to highlight some patterns in the data.

B.1. Data visualization

Input data. Fig. B.13 illustrates the crop growing season in each field (from seeding to harvesting), to illustrate how diverse the collected data is regarding the temporal axis (years) and over regions.

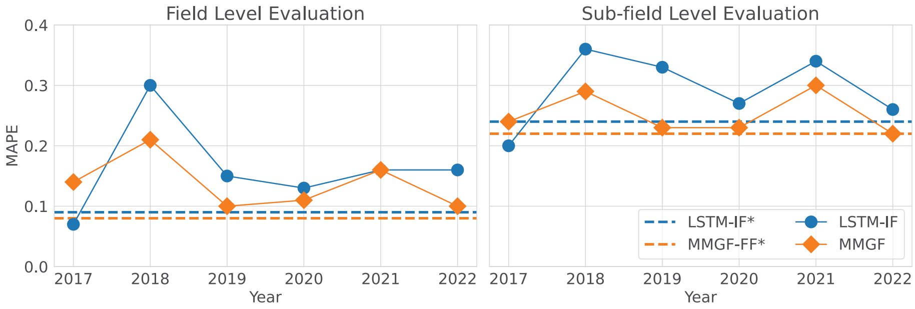
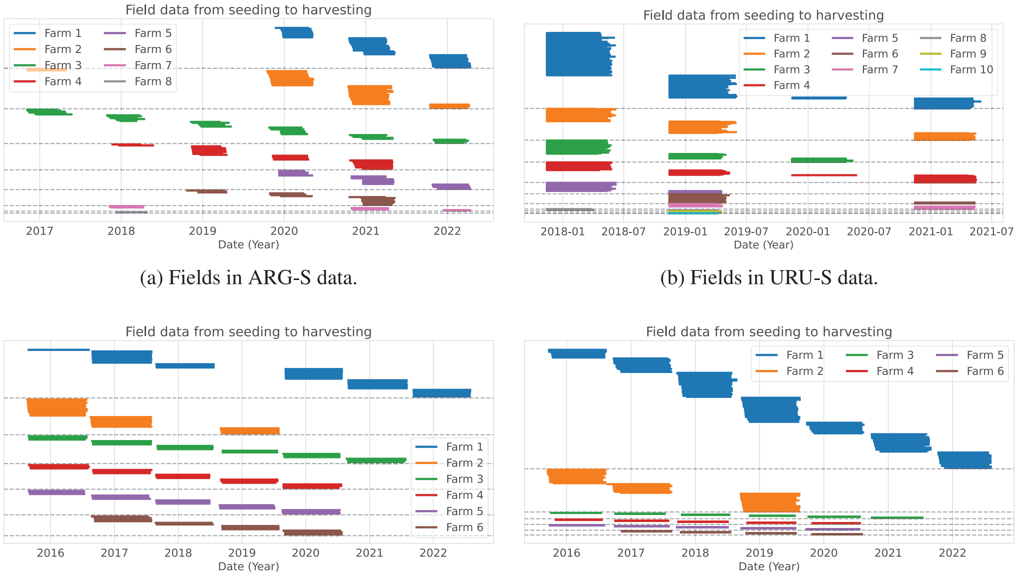


Fig. A.12. Results at different years. MAPE of the crop yield prediction across years in the leave-one-year-out evaluation for ARG-S fields. *As a comparison, the dashed lines show the model performance in the stratified 10-fold cross-validation. The LSTM-IF uses S2-M with DEM modalities, while the MMGF use all modalities (S2-R, weather, DEM, soil). There are 8, 12, 19, 50, 73, and 27 (validation) fields respectively for each year from 2017 to 2022.

Table A.12
Regularization contributions. Crop yield prediction performance in ARG-S fields by varying ML techniques in the MMGF model. The best results are in bold.

Technique	*Sub-Field*			*Field*		
	MAPE (%)	MAE (t/ha)	R^2 –	MAPE (%)	MAE (t/ha)	R^2 –
Without Regularization	23	0.64	0.64	10	0.77	0.36
With Dropout	24	0.65	0.64	10	0.35	0.79
With BN	23	0.63	0.66	10	0.35	0.80
With BN and Dropout	**22**	**0.61**	**0.68**	**8**	**0.30**	**0.84**



(a) Fields in ARG-S data.

(b) Fields in URU-S data.

(c) Fields in GER-R data.

(d) Fields in GER-W data.

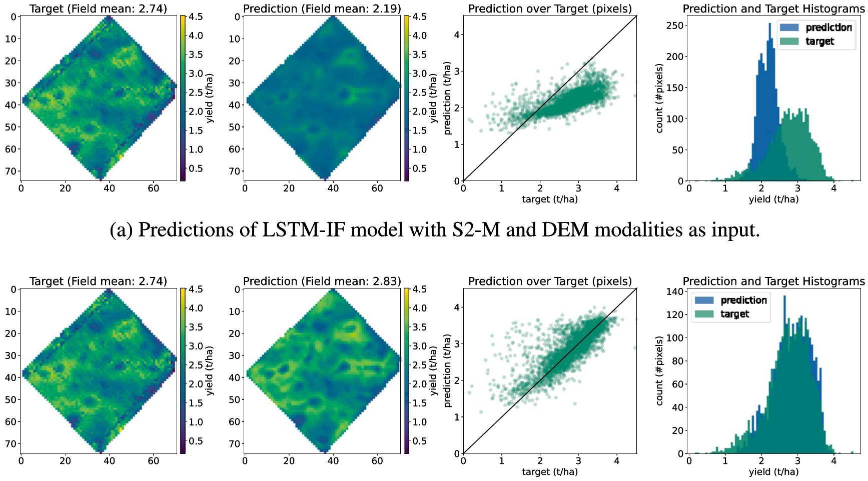
Fig. B.13. Fields from seeding to harvesting grouped by farms. Here, a farm represents either a set of fields operated by a farmer or geographically nearby field. There are a different number of fields per year with different growing seasons.

B.2. Model-based visualization

Crop yield predictions. Since we use a pixel-wise prediction approach for the field images, we visualize the field prediction to qualitatively inspect if the model learns the in-field variability. We created a field yield map based on the sub-field predictions of two models (LSTM-IF and MMGF). In Fig. B.14, B.15, B.16, and B.17 we show examples of these yield map predictions in random fields of the ARG-S, URU-S, GER-R, and GER-W data respectively.

Gated fusion weights at sub-field level. Fig. B.18 displays a stacked bar plot on four randomly selected fields per dataset. Spatially close pixels (pixels within the same field) have a similar distribution of weights. These results were consistent across a larger set of fields, which we inspected separately.

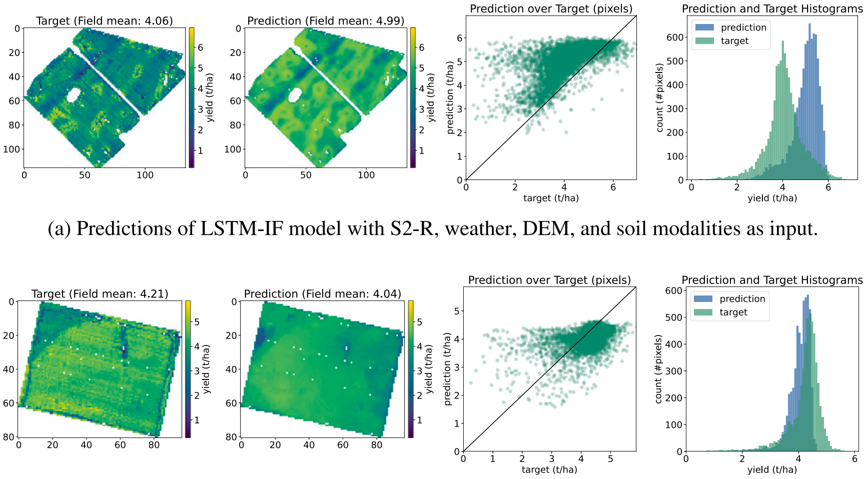
Gated fusion weights analysis. In Fig. B.19 we display the modality contributions based on the MMGF-LR model (α_m, C_m, and $\alpha_m C_m$ explained in Section 6.1) for the GER-R and GER-W fields.



(a) Predictions of LSTM-IF model with S2-M and DEM modalities as input.

(b) Predictions of MMGF model with S2-R, weather, DEM, and soil modalities as input.

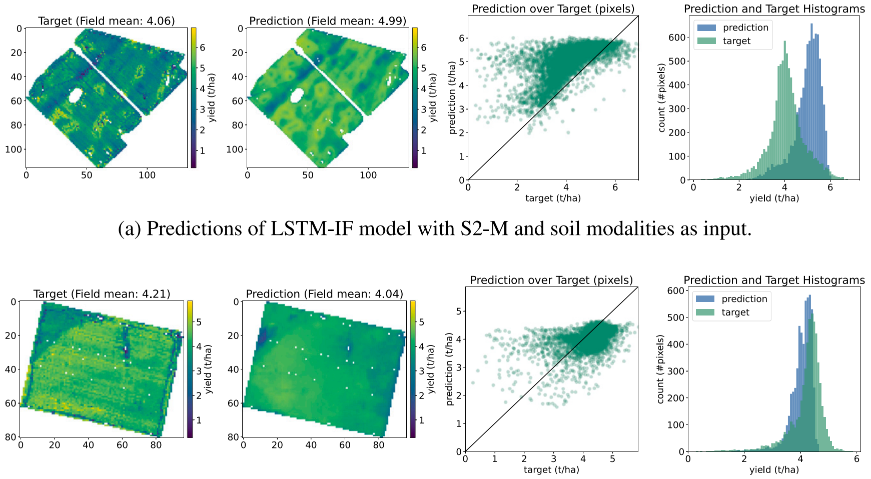
Fig. B.14. Field-level crop yield prediction for a field in **ARG-S** data. The columns from left to right are the ground truth yield map, the predicted yield map, prediction and target scatter, and prediction-target (blue–green) distributions.



(a) Predictions of LSTM-IF model with S2-R, weather, DEM, and soil modalities as input.

(b) Predictions of MMGF model with S2-R, weather, DEM, and soil modalities as input.

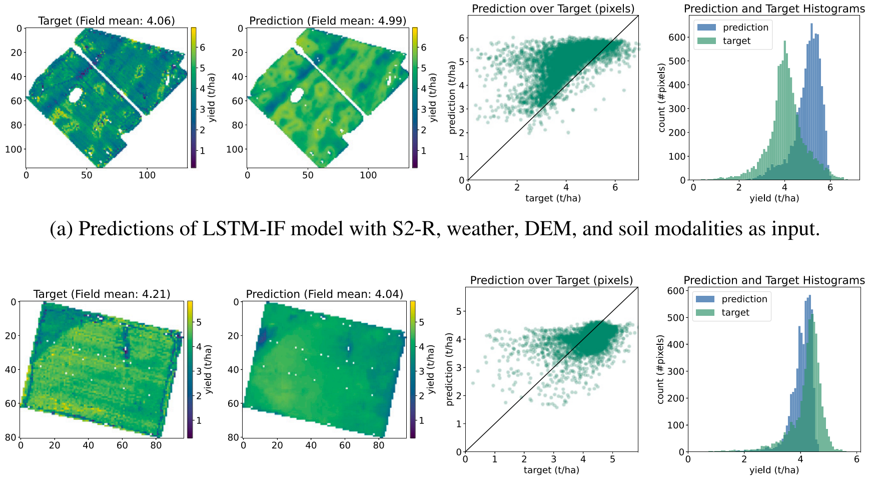
Fig. B.15. Field-level crop yield prediction for a field in **URU-S** data. The columns from left to right are the ground truth yield map, the predicted yield map, prediction and target scatter, and prediction-target (blue–green) distributions.



(a) Predictions of LSTM-IF model with S2-M and soil modalities as input.

(b) Predictions of MMGF model with S2-R, weather, DEM, and soil modalities as input.

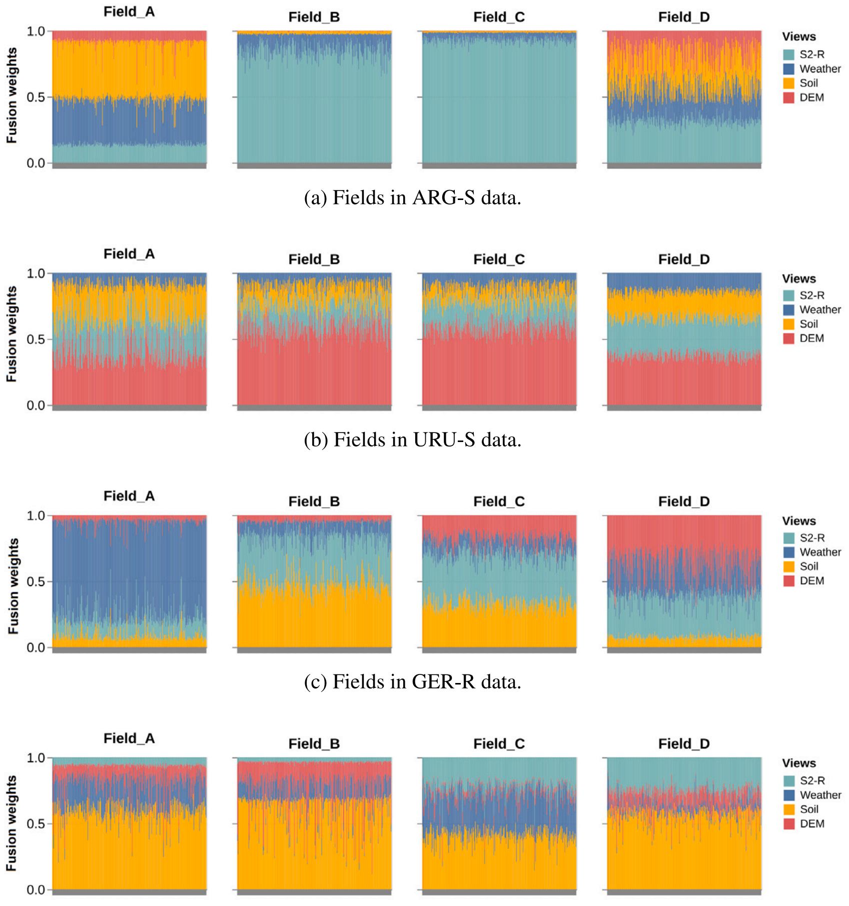
Fig. B.16. Field-level crop yield prediction for a field in **GER-R** data. The columns from left to right are the ground truth yield map, the predicted yield map, prediction and target scatter, and prediction-target (blue–green) distributions.



(a) Predictions of LSTM-IF model with S2-R, weather, DEM, and soil modalities as input.

(b) Predictions of MMGF model with S2-R, weather, DEM, and soil modalities as input.

Fig. B.17. Field-level crop yield prediction for a field in **GER-W** data. The columns from left to right are the ground truth yield map, the predicted yield map, prediction and target scatter, and prediction-target (blue–green) distributions.



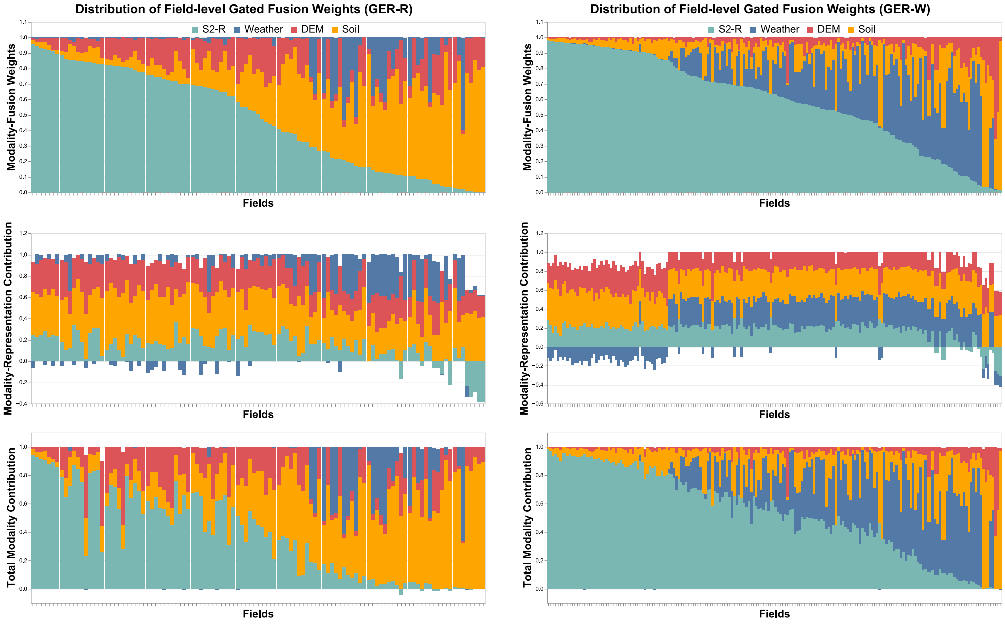
(a) Fields in ARG-S data.

(b) Fields in URU-S data.

(c) Fields in GER-R data.

(d) Fields in GER-W data.

Fig. B.18. The gated fusion weights distribution of 300 randomly sampled pixels from 4 random fields of the different datasets used. The x-axis displays the different pixels, while the y-axis the fusion weights.



(a) Fields from GER-R data.

(b) Fields from GER-W data.

Fig. B.19. Field level gated fusion weights at the top (α_m), the contribution before applying the fusion weights at the middle (C_m), and the total contribution at the bottom ($\alpha_m C_m$). The values are from the GU in the MMGF-LR model. The bars are ordered (from left to right) in descending order of the weight given to the predominant modality (S2-R). The C_m and $\alpha_m C_m$ are scaled by $1/\sum_m |C_m|$ and $1/\sum_m \alpha_m C_m|$ respectively into a range $[-1, 1]$ for better visualization.

Data availability

The authors do not have permission to share data.

References

Arevalo, J., Solorio, T., Montes-y Gomez, M., González, F.A., 2020. Gated multimodal networks. Neural Comput. Appl. 32, 10209–10228. http://dx.doi.org/10.1007/s00521-019-04559-1.

Audebert, N., Le Saux, B., Lefèvre, S., 2018. Beyond RGB: Very high resolution urban remote sensing with multimodal deep networks. ISPRS J. Photogramm. Remote Sens. 140, 20–32. http://dx.doi.org/10.1016/j.isprsjprs.2017.11.011.

Bahdanau, D., Cho, K.H., Bengio, Y., 2015. Neural machine translation by jointly learning to align and translate. In: International Conference on Learning Representations. ICLR.

Benedetti, P., Ienco, D., Gaetano, R., Ose, K., Pensa, R.G., Dupuy, S., 2018. M3Fusion: A deep learning architecture for multiscale multimodal multitemporal satellite data fusion. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 11, 4939–4949. http://dx.doi.org/10.1109/JSTARS.2018.2876357.

Bocca, F.F., Rodrigues, L.H.A., 2016. The effect of tuning, feature engineering, and feature selection in data mining applied to rainfed sugarcane yield modelling. Comput. Electron. Agric. 128, 67–76. http://dx.doi.org/10.1016/j.compag.2016.08.015.

Cai, Y., Guan, K., Lobell, D., Potgieter, A.B., Wang, S., Peng, J., Xu, T., Asseng, S., Zhang, Y., You, L., et al., 2019. Integrating satellite and climate data to predict wheat yield in Australia using machine learning approaches. Agricult. Forest. Meterol. 274, 144–159. http://dx.doi.org/10.1016/j.agrformet.2019.03.010.

Camps-Valls, G., Tuia, D., Zhu, X.X., Reichstein, M., 2021. Deep Learning for the Earth Sciences: A Comprehensive Approach To Remote Sensing, Climate Science and Geosciences. John Wiley & Sons, http://dx.doi.org/10.1002/9781119646181.ch1.

Cao, J., Zhang, Z., Tao, F., Zhang, L., Luo, Y., Zhang, J., Han, J., Xie, J., 2021. Integrating multi-source data for rice yield prediction across China using machine learning and deep learning approaches. Agricult. Forest. Meterol. 297, 108275. http://dx.doi.org/10.1016/j.agrformet.2020.108275.

Chen, Y., Li, C., Ghamisi, P., Jia, X., Gu, Y., 2017. Deep fusion of remote sensing data for accurate classification. IEEE Geosci. Remote Sens. Lett. 14, 1253–1257. http://dx.doi.org/10.1109/LGRS.2017.2704625.

Chu, Z., Yu, J., 2020. An end-to-end model for rice yield prediction using deep learning fusion. Comput. Electron. Agric. 174, 105471. http://dx.doi.org/10.1016/j.compag.2020.105471.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al., 2020. An image is worth 16x16 words: Transformers for image recognition at scale. In: International Conference on Learning Representations. ICLR.

Farr, T.G., Kobrick, M., 2000. Shuttle Radar Topography Mission produces a wealth of data. EOS Trans. Am. Geophys. Union 81, 583–585. http://dx.doi.org/10.1029/EO081i048p00583.

Feng, L., Wang, Y., Zhang, Z., Du, Q., 2021. Geographically and temporally weighted neural network for winter wheat yield prediction. Remote Sens. Environ. 262, 112514. http://dx.doi.org/10.1016/j.rse.2021.112514.

Ferrari, F., Ferreira, M.P., Almeida, C.A., Feitosa, R.Q., 2023. Fusing Sentinel-1 and Sentinel-2 images for deforestation detection in the Brazilian amazon under diverse cloud conditions. IEEE Geosci. Remote Sens. Lett. 20, 1–5. http://dx.doi.org/10.1109/LGRS.2023.3242430.

Garnot, V.S.F., Landrieu, L., 2020. Lightweight temporal self-attention for classifying satellite images time series. In: Advanced Analytics and Learning on Temporal Data: ECML-PKDD Workshop. Springer, pp. 171–181. http://dx.doi.org/10.1007/978-3-030-65742-0_12.

Garnot, V.S.F., Landrieu, L., Chehata, N., 2022. Multi-modal temporal attention models for crop mapping from satellite time series. ISPRS J. Photogramm. Remote Sens. 187, 294–305. http://dx.doi.org/10.1016/j.isprsjprs.2022.03.012.

Gavahi, K., Abbaszadeh, P., Moradkhani, H., 2021. DeepYield: A combined convolutional neural network with long short-term memory for crop yield forecasting. Expert Syst. Appl. 184, 115511. http://dx.doi.org/10.1016/j.eswa.2021.115511.

Helber, P., Bischke, B., Habelitz, P., Sanchez, C., Pathak, D., Miranda, M., Najjar, H., Mena, F., Siddamsetty, J., Arenas, D., Vollmer, M., Charfuelan, M., Nuske, M., Dengel, A., 2023. Crop yield prediction: An operational approach to crop yield modeling on field and subfield level with machine learning models. In: IEEE International Geoscience and Remote Sensing Symposium. IGARSS, pp. 2763–2766. http://dx.doi.org/10.1109/IGARSS52108.2023.10283302.

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., Peubey, C., Radu, R., Schepers, D., et al., 2020. The ERA5 global reanalysis. Q. J. R. Meteorol. Soc. 146, 1999–2049. http://dx.doi.org/10.1002/qj.3803.

Hosseinpour, H., Samadzadegan, F., Javan, F.D., 2022. Cmgfnet: A deep cross-modal gated fusion network for building extraction from very high-resolution remote sensing images. ISPRS J. Photogramm. Remote Sens. 184, 96–115. http://dx.doi.org/10.1016/j.isprsjprs.2021.12.007.

Jacobs, R.A., Jordan, M.I., Nowlan, S.J., Hinton, G.E., 1991. Adaptive mixtures of local experts. Neural Comput. 3, 79–87. http://dx.doi.org/10.1162/neco.1991.3.1.79.

Jain, S., Wallace, B.C., 2019. Attention is not Explanation. In: Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. NAACL-HLT, pp. 3543–3556. http://dx.doi.org/10.18653/v1/N19-1357.

Kang, Y., Ozdogan, M., Zhu, X., Ye, Z., Hain, C., Anderson, M., 2020. Comparative assessment of environmental variables and machine learning algorithms for maize yield prediction in the US midwest. Environ. Res. Lett. 15, 064005. http://dx.doi.org/10.1088/1748-9326/ab7df9.

Kingma, D.P., Ba, J., 2015. Adam: A method for stochastic optimization. In: International Conference on Learning Representations. ICLR, https://dblp.org/rec/journals/corr/KingmaB14.html.

Li, J., Hong, D., Gao, L., Yao, J., Zheng, K., Zhang, B., Chanussot, J., 2022. Deep learning in multimodal remote sensing data fusion: A comprehensive review. Int. J. Appl. Earth Obs. Geoinf. 112, 102926.

Lin, T., Zhong, R., Wang, Y., Xu, J., Jiang, H., Xu, J., Ying, Y., Rodriguez, L., Ting, K., Li, H., 2020. DeepCropNet: A deep spatial–temporal learning framework for county-level corn yield estimation. Environ. Res. Lett. 15, 034016. http://dx.doi.org/10.1088/1748-9326/ab66cb.

Ma, J., Liu, B., Ji, L., Zhu, Z., Wu, Y., Jiao, W., 2023. Field-scale yield prediction of winter wheat under different irrigation regimes based on dynamic fusion of multimodal UAV imagery. Int. J. Appl. Earth Obs. Geoinf. 118, 103292. http://dx.doi.org/10.1016/j.jag.2023.103292.

Maimaitijiang, M., Sagan, V., Sidike, P., Hartling, S., Esposito, F., Fritschi, F.B., 2020. Soybean yield prediction from UAV using multimodal data fusion and deep learning. Remote Sens. Environ. 237, 111599. http://dx.doi.org/10.1016/j.rse.2019.111599.

Méger, N., Courteille, H., Benoit, A., Atto, A., Ienco, D., 2022. Explaining a deep spatiotemporal land cover classifier with attention and redescription mining. In: The XXIV International Society for Photogrammetry and Remote Sensing (ISPRS) Congress. pp. 673–680. http://dx.doi.org/10.5194/isprs-archives-XLIII-B3-2022-673-2022.

Mena, F., Arenas, D., Nuske, M., Dengel, A., 2023. A comparative assessment of multi-view fusion learning for crop classification. In: IEEE International Geoscience and Remote Sensing Symposium. IGARSS, pp. 5631–5634. http://dx.doi.org/10.1109/IGARSS52108.2023.10282138.

Mena, F., Arenas, D., Nuske, M., Dengel, A., 2024. Common practices and taxonomy in deep multi-view fusion for remote sensing applications. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 479, 7–4818. http://dx.doi.org/10.1109/JSTARS.2024.3361556.

Obadic, I., Roscher, R., Oliveira, D.A.B., Zhu, X.X., 2022. Exploring self-attention for crop-type classification explainability. ArXiv preprint arXiv:2210.13167.

Ofori-Ampofo, S., Pelletier, C., Lang, S., 2021. Crop type mapping from optical and radar time series using attention-based deep learning. Remote Sens. 13 (4668), http://dx.doi.org/10.3390/rs13224668.

Pathak, D., Miranda, M., Mena, F., Sanchez, C., Helber, P., Bischke, B., Habelitz, P., Najjar, H., Siddamsetty, J., Arenas, D., Vollmer, M., Charfuelan, M., Nuske, M., Dengel, A., 2023. Predicting crop yield with machine learning: an extensive analysis of input modalities and models on a field and subfield leve. IEEE International Geoscience and Remote Sensing Symposium. IGARSS, pp. 2767–2770. http://dx.doi.org/10.1109/IGARSS52108.2023.10282318.

Poggio, L., De Sousa, L.M., Batjes, N.H., Heuvelink, G., Kempen, B., Ribeiro, E., Rossiter, D., 2021. SoilGrids 2.0: producing soil information for the globe with quantified spatial uncertainty. Soil 7, 217–240. http://dx.doi.org/10.5194/soil-7-217-2021.

Rußwurm, M., Körner, M., 2020. Self-attention for raw optical satellite time series classification. ISPRS J. Photogramm. Remote Sens. 169, 421–435. http://dx.doi.org/10.1016/j.isprsjprs.2020.06.006.

Sanchez, C., Pathak, D., Miranda, M., Charfuelan, M., Helber, P., Nuske, M., Bischke, B., Habelitz, P., Rahman, N., Mena, F., Najjar, H., Siddamsetty, J., Arenas, D., Vollmer, M., Dengel, A., 2023. Influence of data cleaning techniques on sub-field yield predictions. In: IEEE International Geoscience and Remote Sensing Symposium. IGARSS, pp. 4852–4855. http://dx.doi.org/10.1109/IGARSS52108.2023.10282955.

Shahhosseini, M., Hu, G., Khaki, S., Archontoulis, S.V., 2021. Corn yield prediction with ensemble cnn-dnn. Front. Plant Sci. 12, 709008. http://dx.doi.org/10.3389/fpls.2021.709008.

Srivastava, A.K., Safaei, N., Khaki, S., Lopez, G., Zeng, W., Ewert, F., Gaiser, T., Rahimi, J., 2022. Winter wheat yield prediction using convolutional neural networks from environmental and phenological data. Sci. Rep. 12 (3215), http://dx.doi.org/10.1038/s41598-022-06249-w.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I., 2017. Attention is all you need. Adv. Neural Inf. Process. Syst. (NeurIPS) 30.

Wang, X., Feng, Y., Song, R., Mu, Z., Song, C., 2022. Multi-attentive hierarchical dense fusion net for fusion classification of hyperspectral and LiDAR data. Inf. Fusion 82, 1–18. http://dx.doi.org/10.1016/j.inffus.2021.12.008.

Wang, X., Huang, J., Feng, Q., Yin, D., 2020. Winter wheat yield prediction at county level and uncertainty analysis in main wheat-producing regions of China with deep learning approaches. Remote Sens. 12, 1744. http://dx.doi.org/10.3390/rs12111744.

Wiegreffe, S., Pinter, Y., 2019. Attention is not not explanation. In: Conference on Empirical Methods in Natural Language Processing and International Joint Conference on Natural Language Processing. ECML-IJCNLP, pp. 11–20. http://dx.doi.org/10.18653/v1/D19-1002.

Wilcoxon, F., 1992. Individual comparisons by ranking methods. In: Breakthroughs in Statistics: Methodology and Distribution. Springer, http://dx.doi.org/10.2307/3001968.

Wu, X., Hong, D., Chanussot, J., 2021. Convolutional neural networks for multimodal remote sensing data classification. IEEE Trans. Geosci. Remote Sens. 60, 1–10. http://dx.doi.org/10.1109/TGRS.2021.3124913.

Yang, Q., Shi, L., Han, J., Zha, Y., Zhu, P., 2019. Deep convolutional neural networks for rice grain yield estimation at the ripening stage using UAV-based remotely sensed images. Field Crops Res. 235, 142–153. http://dx.doi.org/10.1016/j.fcr.2019.02.022.

Yuksel, S.E., Wilson, J.N., Gader, P.D., 2012. Twenty years of mixture of experts. IEEE Trans. Neural Netw. Learn. Syst. 23, 1177–1193. http://dx.doi.org/10.1109/TNNLS.2012.2200299.

Zhang, P., Du, P., Lin, C., Wang, X., Li, E., Xue, Z., Bai, X., 2020. A hybrid attention-aware fusion network (HAFNet) for building extraction from high-resolution imagery and LiDAR data. Remote Sens. 12 (3764), http://dx.doi.org/10.3390/rs12223764.

Zhang, M., Li, W., Tao, R., Li, H., Du, Q., 2021. Information fusion for classification of hyperspectral and LiDAR data using IP-CNN. IEEE Trans. Geosci. Remote Sens. 60, 1–12. http://dx.doi.org/10.1109/TGRS.2021.3093334.

Zhao, L., Ji, S., 2022. CNN, RNN, or ViT? An evaluation of different deep learning architectures for spatio-temporal representation of Sentinel time series. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 16, 44–56. http://dx.doi.org/10.1109/JSTARS.2022.3219816.

Zheng, X., Wu, X., Huan, L., He, W., Zhang, H., 2021. A gather-to-guide network for remote sensing semantic segmentation of RGB and auxiliary image. IEEE Trans. Geosci. Remote Sens. 60, 1–15. http://dx.doi.org/10.1109/TGRS.2021.3103517.