



UNIVERSITY  
OF TRENTO

---

DEPARTMENT OF INFORMATION ENGINEERING AND COMPUTER SCIENCE  
DEPARTMENT OF INDUSTRIAL ENGINEERING  
FONDAZIONE BRUNO KESSLER

**Doctorate Program in Industrial Innovation**

ENOGISAI: AN ARTIFICIAL  
INTELLIGENCE FRAMEWORK FOR  
PREDICTIVE AGRONOMICS

Luca Coviello

Academic Advisor  
Prof. Mauro Brunato  
Università di Trento

Industrial Advisor  
Dr. Cesare Furlanello  
HK3 Lab & LIGHT Center

---

January 2025





# Abstract

Ensuring the quality and health standards necessary to feed the growing population is a critical challenge. Smart farming technologies combined with artificial intelligence (AI) represent an opportunity to improve crop quality and yield while efficiently and sustainably using the available planetary resources.

This thesis describes EnogisAI: a comprehensive framework for developing and adopting agronomic AI solutions in the field. EnogisAI encompasses all the critical aspects of an AI solution development, starting from the data integration processes through model definition, training, and validation to the development of Decision Support Systems (DSSs) for farmers and agronomists. EnogisAI outputs are integrated within the Enogis agronomic management web platform, used to manage one-fourth of the total Italian viticulture, along with other crops.

Relevant external environmental and weather data sources (e.g., Sentinel-2, ECMWF forecast, ERA5 reanalysis, TINITALY/01 digital terrain model) are integrated and linked with Enogis agronomic in-field collected data. Data is ingested into the data platform and is served through API interfaces to the production ecosystem.

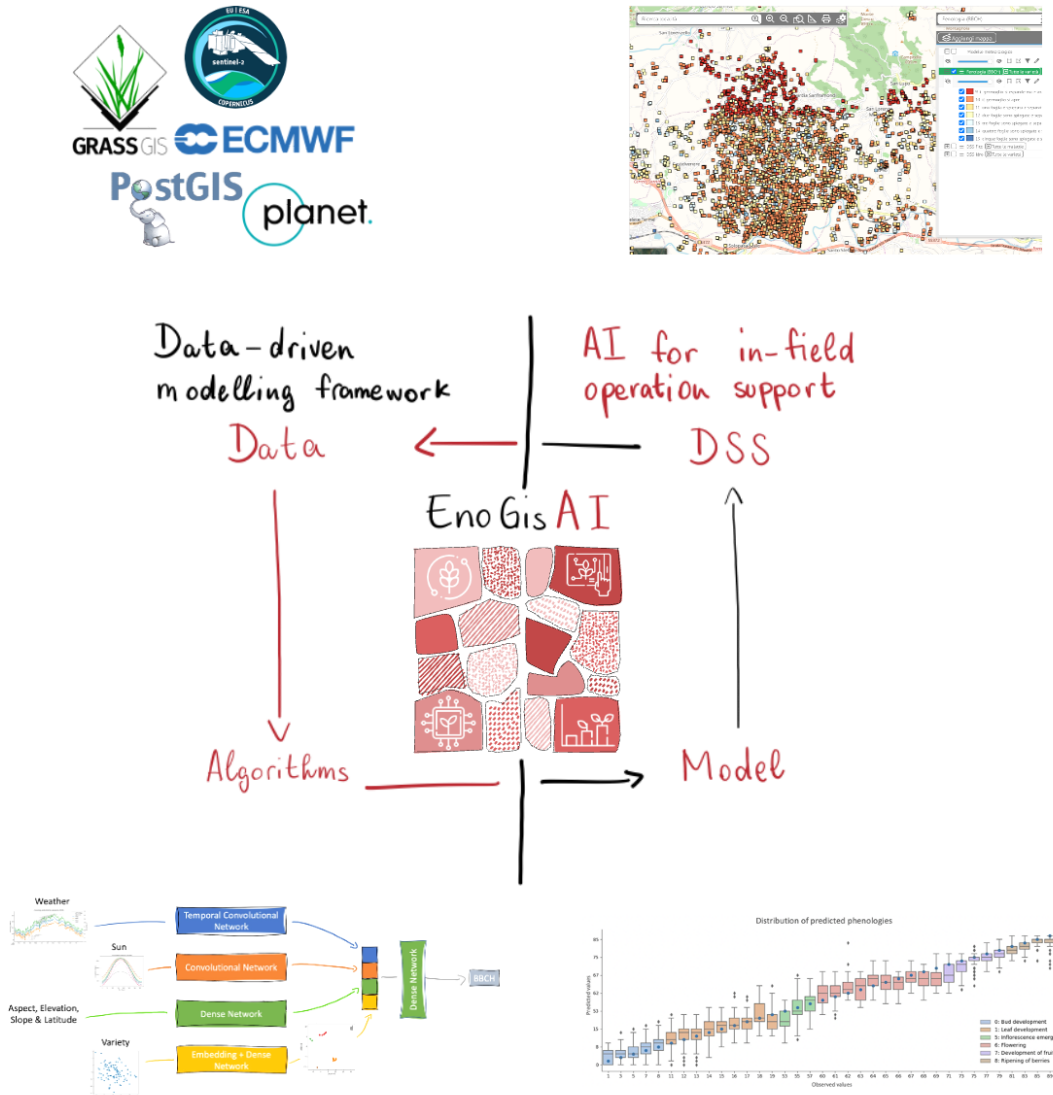
Machine and deep learning models are developed to address relevant agronomic challenges for Enogis users. Specifically, in this work, we describe (i) machine learning models for field zone identification for homogeneous management and yield prediction ( $R^2 > 0.76$ ), (ii) deep learning models for grapevine development monitoring using remote sensing ( $R^2 = 0.96$ ) and in-field images, (iii) a soil water balance assessment tools for crop stress detection and irrigation planning, and (iv) machine and deep learning models for reconstructing and forecasting pollen concentration time series ( $R^2 = 0.82$  and  $R^2 = .72$ , respectively).

Finally, based on the performance and technological maturity levels, a selection of the resulting models is further developed into agronomic DSSs for crop development monitoring, disease protection, and soil water assessment. The DSSs are computed and updated daily for 121296 agricultural fields. This way, model results are accessible by end users through various tools, such as the Enogis WebGIS platform and field management tools, increasing the impact on sustainable agricultural practices.

EnogisAI is the result of industrial research and innovation efforts. As such, all methodologies are designed and built based on the needs resulting from interacting with end users and stakeholders and in collaboration with technological patterns leaders in the agrometeorology ecosystem and international research experts. As an additional result, adopting an open innovation approach, a consistent part of the presented thesis is available either as conference proceedings or research articles.

# Keywords

agriculture, smart farming, artificial intelligence, decision support systems



**Figure 1:** Outline of EnogisAI: a framework for developing and adopting agronomic AI solutions in the field. EnogisAI describes all the aspects of an AI pipeline: data integration, model definition, training, and validation, and the development of Decision Support Systems (DSSs) for farmers and agronomists.

# Contents

|          |                                                                       |           |
|----------|-----------------------------------------------------------------------|-----------|
| <b>1</b> | <b>Introduction</b>                                                   | <b>1</b>  |
| 1.1      | Context . . . . .                                                     | 1         |
| 1.2      | Agricultural challenges and the role of technology . . . . .          | 2         |
| 1.3      | The European context . . . . .                                        | 3         |
| 1.4      | Innovation at Enogis . . . . .                                        | 3         |
| 1.5      | The EnogisAI framework . . . . .                                      | 5         |
| 1.6      | EnogisAI innovative aspects . . . . .                                 | 8         |
| 1.7      | Structure of the thesis . . . . .                                     | 9         |
| <b>2</b> | <b>State of the art</b>                                               | <b>11</b> |
| 2.1      | Crop development monitoring . . . . .                                 | 12        |
| 2.2      | Field management and yield estimation with satellite images . . . . . | 14        |
| 2.3      | Soil water stress estimation . . . . .                                | 15        |
| 2.4      | Pollen concentration time series . . . . .                            | 16        |
| <b>3</b> | <b>The EnogisAI data platform</b>                                     | <b>19</b> |
| 3.1      | Internal data . . . . .                                               | 21        |
| 3.1.1    | Field records . . . . .                                               | 21        |
| 3.1.2    | Field observations . . . . .                                          | 22        |
| 3.1.3    | Agricultural activities . . . . .                                     | 24        |
| 3.2      | External data . . . . .                                               | 26        |
| 3.2.1    | Weather . . . . .                                                     | 26        |
| 3.2.2    | Satellite . . . . .                                                   | 29        |
| 3.2.3    | Environmental descriptors . . . . .                                   | 30        |
| 3.3      | Derived products . . . . .                                            | 32        |
| 3.3.1    | Solar irradiation maps . . . . .                                      | 32        |
| 3.3.2    | Climatic and bioclimatic indices . . . . .                            | 34        |
| 3.4      | Data platform technical aspects . . . . .                             | 37        |
| 3.4.1    | Timescale database for weather time series . . . . .                  | 38        |
| <b>4</b> | <b>Agricultural task, data, methodologies, and algorithms</b>         | <b>45</b> |
| 4.1      | Field management zone identification . . . . .                        | 47        |
| 4.2      | Grapevine development monitoring . . . . .                            | 50        |
| 4.2.1    | Remote sensing neural network for phenology prediction . . . . .      | 51        |
| 4.2.2    | In-field image classification . . . . .                               | 54        |

|          |                                                                          |            |
|----------|--------------------------------------------------------------------------|------------|
| 4.3      | Satellite-based durum wheat yield prediction . . . . .                   | 55         |
| 4.4      | Soil water balance assessment . . . . .                                  | 58         |
| 4.5      | Pollen seasonal time series reconstruction . . . . .                     | 61         |
| 4.6      | Time series forecasting of pollen concentrations . . . . .               | 63         |
| <b>5</b> | <b>EnogisAI models and results</b>                                       | <b>65</b>  |
| 5.1      | Field management zone identification . . . . .                           | 67         |
| 5.2      | Grapevine development monitoring . . . . .                               | 70         |
| 5.2.1    | Remote sensing neural network for phenology prediction . . . .           | 70         |
| 5.2.2    | In-field image classification . . . . .                                  | 76         |
| 5.3      | Satellite-based durum wheat yield prediction . . . . .                   | 79         |
| 5.4      | Soil water balance assessment . . . . .                                  | 84         |
| 5.5      | Pollen seasonal time series reconstruction . . . . .                     | 89         |
| 5.6      | Time series forecasting of pollen concentrations . . . . .               | 94         |
| <b>6</b> | <b>Decision Support Systems</b>                                          | <b>101</b> |
| 6.1      | Phenology DSS . . . . .                                                  | 103        |
| 6.2      | Disease coverage DSS . . . . .                                           | 106        |
| 6.3      | Soil water balance DSS . . . . .                                         | 112        |
| 6.4      | Interactive data exploration tool for enhanced decision-making . . . . . | 114        |
| 6.5      | Technical implementation details . . . . .                               | 118        |
| <b>7</b> | <b>Conclusions</b>                                                       | <b>121</b> |
|          | <b>Bibliography</b>                                                      | <b>125</b> |
| <b>A</b> | <b>Additional figures</b>                                                | <b>143</b> |

# List of Tables

|     |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                   |    |
|-----|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----|
| 3.1 | Summary of agronomic data samples (and corresponding year) collected with the Enogis software. Entries are sorted by count (descending). . .                                                                                                                                                                                                                                                                                                                                                                                                                                      | 22 |
| 3.2 | External data products integrated into the EnogisAI data platform. Static datasets are manually organized, while, for the others, we developed specific data ingestion workflows. All data products were used in the later model development phases. . . . .                                                                                                                                                                                                                                                                                                                      | 26 |
| 4.1 | Sentinel-2 characteristics of the spectral bands used for the NDVI and NDRE index calculation. The two indices were used to construct the pipeline’s input time series. The Red Edge band was resampled with linear interpolation to 10m. . . . .                                                                                                                                                                                                                                                                                                                                 | 48 |
| 5.1 | External validation results computed on 2019 data. Input sources are ordered by cardinality. MAE and RMSE are measured in BBCH phases. Bold values indicate the best value for each metric. The full model reached the best MAE and $R^2$ performance (1.72 and 0.96). The model using only weather and morphology reached the second-best performance ( $RMSE = 2.33$ and $R^2 = 0.96$ ). . . . .                                                                                                                                                                                | 71 |
| 5.2 | Performance for the best resulting model on the external validation (cfr. Table 5.1), and on the two real-world scenario datasets (RWS-incl. and RWS-excl). Reported results include model experimentation using the latest observed phenology value (w/ past) as input. <b>MAE and RMSE are expressed in BBCH.</b> The effect of using past phenologies observed in the season (w/past column) significantly improves performance on the two RWS scenarios but negatively impacts the external validation dataset. . . . .                                                       | 74 |
| 5.3 | Performance for the best resulting model on the external validation and the two real-world scenario datasets (RWS-incl. and RWS-excl). Reported results include model experimentation using the latest observed phenology value (w/ past) as input. <b>MAE and RMSE are expressed in days.</b> MAE values are associated with $\pm 7$ days on the real-world scenarios. The effect of using past phenologies observed in the season (w/past column) significantly improves performance on the two RWS scenarios but impact negatively on the external validation dataset. . . . . | 75 |

|      |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                        |    |
|------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----|
| 5.4  | Test performance metrics of the best performing VGG-16 and ResNet-50 models. ResNet-50 converged with 17 epochs (compared to 65 of VGG-16) and resulted in lower MAE and RMSE (3.63 and 5.26). . . .                                                                                                                                                                                                                                                                                                                                                                   | 77 |
| 5.5  | Results for yield prediction applying PLS to time series derived from remotely sensed Sentinel-2 L2A and Planet data. The best resulting model is Sentinel-2's NDRE for both considered windows (T1: weeks 10-36, T2: weeks 10-23). . . . .                                                                                                                                                                                                                                                                                                                            | 80 |
| 5.6  | Regression results for yield with different temporal windows using the image-based neural networks. VGG-16 is the best-performing model on T1, while MobileNetV2 is the best one on T2. VGG-19 is the second-best on both temporal windows. . . . .                                                                                                                                                                                                                                                                                                                    | 81 |
| 5.7  | Missing data imputation performance with weather information (ConvAE + Weather) and without weather information (ConvAE) when removing <b>small</b> windows of consecutive days from original data. Metrics are reported for the 1995–2010 data set (5-fold validation ( <i>val</i> ) and test sets ( <i>test</i> )) and for the 2010 test data set (103 stations not present in the training distribution ( <i>test103</i> )). In the case of small windows removal, the weather component does not significantly improve the reconstruction processes. . . . .       | 90 |
| 5.8  | Missing data imputation performance with weather information (ConvAE + Weather) and without weather information (ConvAE) when removing <b>large</b> windows of consecutive days from original data. Metrics are reported for the 1995–2010 data set (5-fold validation ( <i>val</i> ) and test sets ( <i>test</i> )) and for the 2010 test data set (103 stations not present in the training distribution ( <i>test103</i> )). In the case of large windows removal, the weather component contributes significantly to improving the reconstruction process. . . . . | 90 |
| 5.9  | Annual total pollen count estimation results reported in terms of $R^2$ , RMSE, and MAE. The DL model takes the additional weather information as input. The DL model results in better performance compared to the GM model, up to 70% of missing data for $R^2$ and RMSE, and up to 50% for MAE. . . . .                                                                                                                                                                                                                                                             | 91 |
| 5.10 | MAPE for map generation validation results. The DL model shows an improvement for all the data removal percentages, with the exception of 60% (DL MAPE is bigger than Input MAPE by 2.90%). On average, the GM model decreases the accuracy of generated maps. . . . .                                                                                                                                                                                                                                                                                                 | 92 |
| 5.11 | Pearson R and $R^2$ metrics for the CatBoost (CB) and deep learning (DL) models. Numbers in the model versions refer to the number of past days used as input, while <i>np</i> indicates that the daily total pollen concentration was not used as an input variable. The Pearson's R p-values are all significant at the .001 level. All models beat the persistence model, and using pollen improves over not using it (CB_7 vs CB_7np and DL_7 vs DL_7np). Increasing the input context length over 7 days does not significantly improve model results. . . . .    | 95 |

# List of Figures

|     |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                             |    |
|-----|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----|
| 1.1 | The EnogisAI framework consists of four components. First, data from internal and external sources are integrated and organized to power further application development. Second, AI and mechanistic algorithms use and are built on the data component. Third, the training and validation of the algorithms yield models that address specific agricultural tasks. Finally, models are integrated into Decision Support Systems (DSSs) solutions provided to end users (agronomists and farmers). The framework can be seen as a combination of a data-driven modeling framework (left side) aimed at providing AI solutions for in-field operation support (right side). | 6  |
| 3.1 | Example of filled disease form in the Enogis field notebook tool. Data collected in this and similar forms are then used for successive data modeling and analysis phases.                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                  | 25 |
| 3.2 | Example of real irrigation form in the Enogis field notebook tool. Data collected in this and similar forms are then used for successive data modeling and analysis phases.                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                 | 25 |
| 3.3 | Resulting sun irradiation maps around the Adige valley, Trentino-South Tyrol region, North Italy, computed for Mar 21 (a), Jun 21 (b), Sep 21 (c), and Dec 21 (d). Maps were computed for all day of year and the entire Italian territory using the <i>r.sun</i> GRASS module.                                                                                                                                                                                                                                                                                                                                                                                             | 33 |
| 3.4 | Winkler index comparison between ERA5-Land (horizontal resolution: 9 km) (a) and VHR-REA_IT (horizontal resolution: 2.2 km) (b) on the Abruzzo region. The Winkler index is upscaled to 500 m, applying a dry adiabatic gradient to the input temperature data. VHR-REA_IT expresses better the extreme values of the index, especially along the coastline. The red box in the corner map indicates the region where the Winkler index was computed with VHR-REA_IT data.                                                                                                                                                                                                  | 36 |

|     |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                             |    |
|-----|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----|
| 3.5 | Timescale performance test results, expressed in milliseconds. Faster benchmarks are <b>blue</b> , while slower ones are <b>red</b> . Timescale (tsdb) improves performance for 10 out of 16 benchmarks (10.24% faster on average in the cold environment), while Timescale with compression (tsdb-c) is systematically better compared to PostgreSQL (postgres) (45% faster on average in the cold environment and 35% faster on the warm one). Each row represents a benchmark experiment at a different level (database or application access interface), a different dataset (wide or long), run on all three setups (PostgreSQL, TimescaleDB, and TimescaleDB with compression) two consecutive times (cold and warm). The color scale is normalized per benchmark (row). Boxplots (top) summarize the results of each setup for both cold and warm scenarios. . . . . | 42 |
| 4.1 | Data processing pipeline for the field management zone (MZ) identification task. Sentinel-2 satellite data are organized into one time series for each corresponding ground sampling pixel. Time series of NDVI and NDRE vegetation indices are decomposed with Functional PCA. The resulting FPCA scores are then clustered with K-means to identify homogeneous MZ. Finally, these are then displayed on a map. . . . .                                                                                                                                                                                                                                                                                                                                                                                                                                                   | 48 |
| 4.2 | The MZ identification pipeline was developed considering 12 organic durum wheat fields in the Marche region, Central Italy. The fields were included in the BioCereals 4.0 innovation project. Total considered area: 25.7 ha (median: 1.31 ha, min: 0.27 ha, max: 7.26 ha). . . . .                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                        | 49 |
| 4.3 | Phenology data statistics for 44941 phenology samples collected from 2012 to 2020. Map of the observed phenologies with heatmaps showing the density of the three most sampled areas (Trentino-South Tyrol, Veneto and Friuli-Venezia Giulia, and Sicily) (a). Boxplot distributions of each collected phenology phase (b). . . . .                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                         | 51 |
| 4.4 | Multi-modal deep learning architecture of the remote sensing phenology model composed of four branches for representation learning and a final sub-network for latent data fusion and phenology prediction. Temperature and precipitation data are processed by a Temporal Convolutional Network ( <b>blue</b> ). Annual solar irradiation profiles are processed with a convolutional neural network ( <b>orange</b> ). Tabular characteristics such as aspect, elevation, slope, and latitude are modeled by a dense network ( <b>green</b> ). Finally, the cultivar categorical variable is associated with learnable embedding vectors ( <b>yellow</b> ). . . . .                                                                                                                                                                                                       | 52 |
| 4.5 | Solar profiles expressed as $Wh * m^{-2} * day^{-1}$ for 10 random fields, spanning from the minimum to the maximum cumulative yearly potential computed on fields present in the dataset. Solar profiles are computed with the <i>r.sun</i> GRASS module considering the surrounding mountain ranges for a range of 25 km around each point. . . . .                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                       | 53 |



|      |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                   |    |
|------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----|
| 4.6  | Distribution of BBCH values of the 7713 images with phenological information collected with the Enogis software. Images are collected for agronomic purposes in different time periods from different clients, geographical areas, and agronomists operating with different agronomical objectives. This heterogeneity in data collection processes introduces a high degree of variability in the collected phenology phases and visual characteristics of dataset images. . . . .                                               | 55 |
| 4.7  | Data processing pipeline for the durum wheat yield prediction task. Sentinel-2 and PlanetScope satellite data are organized into one time series for each ground pixel. Two approaches are tested. First ( <b>pink</b> ), the NDVI and NDRE time series are decomposed with Functional PCA. The result is used as input to the PLS regression algorithm to predict yield per pixel. Second ( <b>green</b> ), spectral data are arranged into images and processed with image-based neural networks to estimate the yield. . . . . | 56 |
| 4.8  | Reference yield map of the considered area after remapping yields to the Sentinel-2 reference grid. Yields are measured in kg/ha (a). Example of an image generated for a single grid cell containing data of the 12 Sentinel-2 bands (columns) for 35 weeks (rows) (b). . . . .                                                                                                                                                                                                                                                  | 57 |
| 4.9  | Station and data density for the 95 stations of 1995-2010 (a) and the 162 stations of 2010 (b) datasets. The five classes in the legend refer to 0% (smaller dots) to 100% (larger dots) of data density (i.e., 100 - the percentage of missing data) with a step of 20% (i.e., class 1: is 0-20% of present data; class 5: 80-100% of present data). . . . .                                                                                                                                                                     | 61 |
| 4.10 | ConvAE model architecture. Input ( <b>green</b> block, left) contains the pollen concentration observations with missing values to reconstruct for a single city and year and, optionally, precipitation and temperature data. The encoder ( <b>blue</b> ) learns and generates the intermediate representation used to estimate the missing pollen concentration values by the decoder ( <b>orange</b> ). The network outputs reconstructed seasonal pollen concentration values ( <b>green</b> block, right). . . . .           | 62 |
| 4.11 | Geographical distribution/location of the aerobiological stations that provided data for the analysis. The colors refer to the number of years considered with the available daily pollen concentrations . . . . .                                                                                                                                                                                                                                                                                                                | 63 |
| 5.1  | FPCA results on a selected test site using NDRE as vegetation index. The mean function indicates the mean of the vegetation index time series. The scree-plot shows the amount of information captured by each principal component. 3 components are enough to capture 95% of the variance. Finally, the eigenfunctions plot shows the dynamics of curves over time captured by each component. The x-axis weeks are computed starting from the first collected satellite image. . . . .                                          | 68 |

|     |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                  |    |
|-----|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----|
| 5.2 | Functions projected with FPCA and peak distribution in a critical period: middle March to middle of June. It is possible to appreciate the different trends that characterize each grid cell (e.g., the peak is reached at different week numbers) (a). Distributions of week numbers associated with peaks of the reconstructed functions (b). . . . .                                                                                                                                                                                                                                                                                          | 69 |
| 5.3 | Results of clustering FPCA scores with K-means into two clusters for a sample field. The (a) side shows the NDRE index of each grid cell during the monitored period, curves are colored based on the clustering results, and dashed lines represent cluster averages . The (b) left side shows the difference between pixel FPCA scores to the mean profile, while the right side shows the clustering of FPCA scores over the two classes for the selected sample field. . . . .                                                                                                                                                               | 69 |
| 5.4 | Examples of comparison between observed (blue) and predicted (orange) phenologies for four testing fields. BBCH values on the y-axis are represented interpreting the code as an integer value to highlight the difference between vegetative and reproductive phases. Predicted phenologies in the second and fourth images tend to anticipate observed values. Observed phenologies in the third image exhibit a less smooth trend; this may be associated with data collection errors that also contribute to noise in the dataset. . . . .                                                                                                   | 72 |
| 5.5 | Distribution of predicted values for each observed phenological phase. The blue dots represent the correct prediction for each phenology. The highest model uncertainty is associated with BBCH 18, 55, 60, 63, and 71. . . . .                                                                                                                                                                                                                                                                                                                                                                                                                  | 73 |
| 5.6 | Confusion matrices of the VGG-16 (left) and ResNet-50 (right) models calculated on the Moncaro test dataset. VGG-16 overfitted the first vegetative phases and is skewed toward predicting phases in the 13-18 range. On the other hand, ResNet-50 worked better with later phenologies. In particular, it predicts BBCH 55 (inflorescences swelling, flowers closely pressed together) due to the presence of emerged inflorescence in the images, while farmers and agronomists still focused on vegetative phenologies). Phenology values below 8 and above 72 were never observed nor predicted and are excluded from visualization. . . . . | 78 |
| 5.7 | Sample image where the ResNet-50 model wrongly predicts a reproductive phase (BBCH 55: inflorescences swelling, flowers closely pressed together) with clusters in evidence. The image was annotated with BBCH 26: sixteen leaves unfolded. The full image (a) was used as model input. The red area represents a part of the image that shows clusters (b). . .                                                                                                                                                                                                                                                                                 | 79 |
| 5.8 | Scatter plot of predicted and measured scaled yields using Sentinel-2 NDRE vegetation index (T2 time window). The model obtained an $R^2$ of 0.755. The blue line is the best-fit line, while the green one represents the 1:1 line. . . . .                                                                                                                                                                                                                                                                                                                                                                                                     | 81 |
| 5.9 | Scatter plot for VGG-16 model with linear interpolation over T1 temporal window ( $R^2 = 0.843$ ). The dashed line represents the best-fit line. .                                                                                                                                                                                                                                                                                                                                                                                                                                                                                               | 82 |

|      |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                              |    |
|------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----|
| 5.10 | Soil water content (blue) simulation on 2022 using the ASCE method to compute $ET_0$ using ERA5-Land (top) and ERA5-Land substituting the total precipitation variable with station recorded precipitation (bottom). The large overestimation of rainy days (orange) in ERA5-Land causes an extreme difference starting in the May-September period, leading to an overestimation of the available water content. Vertical lines (gray) indicate daily precipitation larger than 10 mm. The central dashed line (black) represents the stress threshold, i.e. when the root zone depletion grows over the readily available water (RAW). The top and bottom horizontal dotted lines (gray) represent field capacity and wilting point, respectively. . . . . | 85 |
| 5.11 | Soil water content (blue) simulation on 2022 using the ASCE method to compute $ET_0$ with ERA5-Land and replacing precipitation (orange) with observed values (top). The median of the observed soil available water was computed on the nine selected stations at 10 (green), 20 (violet), and 30 (red) cm depths (bottom). The black line is computed as the mean of the three depths. The top simulation output (blue) is presented in transparent blue for easier comparison. . . . .                                                                                                                                                                                                                                                                    | 86 |
| 5.12 | Comparison between simulations run with ERA5-Land and station precipitation data with ASCE methodology (top), Hypermeteo data with MHS3 methodology (middle), and ECMWF data with MHS3 methodology (bottom). The three simulations present comparable patterns, with ERA5-Land (with station precipitation) slightly underestimating the available water. . . . .                                                                                                                                                                                                                                                                                                                                                                                            | 87 |
| 5.13 | 16-year mean annual total ragweed pollen concentrations for Europe using a base-10 logarithmic color scale, with the geographical location of the 67 aerobiological stations with raw (a), GM restored (b), and DL restored (c) data sets, July 15 - October 15, 1995-2010, (pollen grains $m^{-3}$ ). Black dots indicate the location of the aerobiological stations. Maps computed with GM and DL restored present a higher pollen concentration, . . . . .                                                                                                                                                                                                                                                                                               | 93 |
| 5.14 | Modeled $R^2$ for each day, averaged over the 23 cities. $R^2$ values are computed using only the daily total pollen concentrations greater than 0. CB models (circle symbol and connected lines) work better for the 1-3 forecast days. DL models (square symbol and dashed lines) work better for the 5-14 forecast days (starting with DL_7). . . . .                                                                                                                                                                                                                                                                                                                                                                                                     | 96 |
| 5.15 | Differences in the model performance ( $R^2$ for each city averaged over days) for the best CB and DL models. DL_7 (orange) outperforms CB_7 (blue) for 14 cities. Brisbane (Australia) and Bursa (Turkey) present the largest differences in the performance. $R^2$ values differences are computed using only daily total pollen concentrations greater than 0. . . . .                                                                                                                                                                                                                                                                                                                                                                                    | 96 |

- 5.16 Predictions of CB\_7 (**blue** dashed line) and DL\_7 (**orange** dashed line) models for Amiens in the year 2013. The background (**gray**) represents the observed pollen count. The forecasts are presented for 1 (a), 5 (b), 10 (c), and 14 (d) days in advance. . . . . 97
- 5.17 Clustering results for all combinations of city-day. The numbers in the squares are the serial numbers of the 19 clusters from 0 to 18, comprising groups of various environmental variables. A value of -1 indicates points classified as noise from the DBSCAN algorithm (points that are not part of a group of at least 4 points). For example, cluster 0 indicates that the same features (i.e., the same environmental variables) are important for 17 cities for predicting the daily total pollen concentration for Day 1, i.e., for the next day. At the same time, for Day 2, the same environmental variables in cluster 0 are important only for six cities. Larger squares indicate a higher predictive performance for the city-day combination (smallest square: Seoul (South Korea), day 12,  $R^2 = 0$ ; biggest square: Mexico City (Mexico), Day 1,  $R^2 = 1$ ; median size square: Busan (South Korea), Day 7,  $R^2 = .506$ ). The dendrogram at the top indicates cities that are similar in terms of the clusters they are assigned to. . . . . 98
- 5.18 The top 10 most important environmental variables in decreasing order of the median of their feature importance. Values from left to right are used to predict the daily total pollen concentrations, cluster 0, and city-day combinations. Feature importance values were transformed using the  $\log_{10}(x + 1)$  transformation for visualization purposes. Feature importance values of the daily total pollen concentration (past) both for cluster 0 and all other clusters are presented with a strip plot to highlight the number of points belonging to each group. The diagram on the right indicates what city and forecast day pairs are included in the cluster. . . . . 99
- 6.1 Detail view of observed (**blue** and **green** lines) and predicted phenologies (dotted **gray** line) for a sample field during the 2021 agricultural season. The predicted values closely follow the observed ones. Around May 31<sup>st</sup>, the phenological model is late compared to the observed value. . . . . 103
- 6.2 Testing area (spanning 15 km in width and 10 km in height) located in southern Italy, with BBCH values displayed for April 21, 2022. On the right side, the layer list showcases all available DSS, with the phenology layer labeled "Fenologia BBCH". This layer includes a dropdown menu to filter by cultivar (by default, all cultivars are selected). . . . . 104
- 6.3 Feature info box displaying the BBCH trends for two selected fields (**yellow** contours, left side of the picture) for the 2022 agricultural season. The area spans 1.9 km in width and 1.2 km in height. Fields belong to the Lambrusco (**blue**) and Malvasia Bianca (black) varieties. . . . . 105

|      |                                                                                                                                                                                                                                                                                                                                                                                                                                                   |     |
|------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----|
| 6.4  | Testing area located in southern Italy, with coverage values displayed for May 25, 2022. The three color classes encode the number of days fields are protected against diseases: less than 3 ( <b>red</b> ), between 3 and 4 ( <b>yellow</b> ), or more than 4 days ( <b>green</b> ).                                                                                                                                                            | 107 |
| 6.5  | Feature info box displaying the disease coverage DSS as of May 25, 2022. The selected field (left side of the figure) is yellow, thus associated with 3-4 days of coverage. This is an analysis of a past season; for this reason, the timeline chart displays coverage values for the rest of the season.                                                                                                                                        | 108 |
| 6.6  | Example of dashboard summarizing DSS output stats in terms of fields ( <i>appezzamenti</i> ) and surface ( <i>superficie</i> ). The three color classes encode the number of days fields are protected against diseases: less than 3 ( <b>red</b> ), between 3 and 4 ( <b>yellow</b> ), or more than 4 days ( <b>green</b> ).                                                                                                                     | 109 |
| 6.7  | Enogis field notebook tool showing a selection of agricultural companies present in Enogis for the 2022 agricultural season. The last column ( <i>DSS Fito</i> ) shows an indicator summarizing the status of all fields belonging to each company. The three color classes encode the number of days fields are protected against diseases: less than 3 ( <b>red</b> ), between 3 and 4 ( <b>yellow</b> ), or more than 4 days ( <b>green</b> ). | 110 |
| 6.8  | Downy mildew simulation for an example field. The solid line ( <b>violet</b> ) represents the coverage associated with actual treatments, while the dashed line ( <b>fuchsia</b> ) is associated with simulated treatments. Treatments are associated with days where the coverage is maximized to 100% (e.g., on May 28 <sup>th</sup> ). In this case, the number of actual treatments is 7, while the number of simulated treatments is 3.      | 111 |
| 6.9  | Testing area located in southern Italy, with soil water status values displayed for May 25, 2022. The three color classes encode the number of days to reach the crop stress threshold: less than 3 ( <b>red</b> ), between 3 and 4 ( <b>yellow</b> ), or more than 4 days ( <b>green</b> ).                                                                                                                                                      | 112 |
| 6.10 | Feature info box displaying the disease coverage DSS as of May 25, 2022. The selected field (left side of the figure) is red, thus associated with less than 3 days to reach the stress threshold. This is an analysis of a past season; for this reason, the timeline chart displays the soil water balance values for the rest of the season.                                                                                                   | 113 |
| 6.11 | Classification results using the Winkler index as input. Categories and rules are defined by the users as follows: Winkler<1400 ( <b>blue</b> ), 1400<Winkler<1600 ( <b>green</b> ), 1600<Winkler<1800 ( <b>yellow</b> ), Winkler>1800 ( <b>red</b> ). The info box shows the values associated with the selected <b>yellow</b> area. The top part displays the input values of the DSS, and the three bottom boxes show the output of the DSS.   | 116 |

|      |                                                                                                                                                                                                                                                                                                                                                                                                                                                                     |     |
|------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----|
| 6.12 | DSS example output for July 20, 2024. The map displays six classes resulting from the combination of two biomass adoption levels and three soil management suggestions. The area refers to the Offida DOC region. The application displays a feature info box for the "remove - high biomass" class ( <i>rompere - alta biomassa</i> in the figure, dark orange), displaying the values of the input data (top box) and the DSS outputs (two bottom boxes). . . . . | 117 |
| A.1  | Comparison between ASCE and MHS3 methods for estimating $ET_0$ computed on the reference plot data for year 2022. Both methods used as input the same ERA5-Land variables. . . . .                                                                                                                                                                                                                                                                                  | 143 |
| A.2  | Soil available water simulation on 2022 using ASCE to compute $ET_0$ with ERA5-Land (top), ERA5-Land substituting the total precipitation variable with station recorded precipitation (center), and ERA5-Land settings to 0 the precipitation for days that precipitation values were not recorded by the station (bottom). . . . .                                                                                                                                | 144 |
| A.3  | Simulation on 2022 using Soil available water simulation on 2022 using ASCE to compute $ET_0$ with ERA5-Land (top), ERA5-Land substituting the total precipitation variable with station recorded precipitation (center), and using MHS3 on station temperature and precipitation data (bottom). . . . .                                                                                                                                                            | 145 |
| A.4  | Total annual ragweed pollen concentrations for Europe on base-10 logarithmic color scale, with the geographical locations of the 162 aerobiological stations, using the (a) raw (a), GM restored (b), and (c) DL restored data sets, July 15 - October 15, 2010, (pollen grains $m^{-3}$ ). . . . .                                                                                                                                                                 | 146 |

# Chapter 1

## Introduction

### 1.1 Context

This PhD thesis discusses the development of data-intensive and Artificial Intelligence (AI) solutions for agriculture. The work was developed in the context of the "Doctorate Program in Industrial Innovation" at the University of Trento. The Industrial Innovation course is a multidisciplinary program that involves multiple scientific, engineering, and economics departments within the university. Topics covered in the PhD program are developed in partnership with companies eager to invest in research and development initiatives. These collaborations aim to generate practical, transferable outcomes that integrate seamlessly into the company processes, promoting innovation and enhancing operational efficiency.

As a PhD Executive candidate, I conducted my efforts entirely at Enogis Srl, the project's partner company, where I led the artificial intelligence research and development efforts. Enogis operates in the agricultural field management sector, with a strong focus on grapevines, providing its customers with a software solution for precision viticulture to record and analyze field observations and activities and simplified field management. Around one-fourth of the Italian viticulture surface is managed by Enogis software.

The objective of the work discussed in this thesis is twofold: first is to advance the research field of computer science, particularly artificial intelligence, applied to agriculture; and second, to bridge the gap between research and innovation, effectively delivering new tools and solutions to end users, enhancing agricultural practices and productivity. The thesis describes EnogisAI: a framework to streamline the development and simplify the adoption of artificial intelligence solutions for precision agriculture.

EnogisAI encompasses all aspects of AI development, starting from data collection and curation, through model development and validation, to solution deployment for end users: farmers and agronomists.

## 1.2 Agricultural challenges and the role of technology

The global agriculture system is undergoing significant transformations, influenced by various trends that shape its future. According to the United Nations, the population was estimated to reach 9.7 billion by 2050 [1], peaking around 10.3 in the mid-2080s and gradually decreasing to 10.2 billion by the end of the century [2]. As the world population continues to rise the demand for food is increasing, necessitating increased agricultural yields and the maintaining of health and quality standards. The challenge becomes even more evident when considering that the number of undernourished people in 2023 is estimated to range between 713 and 757 million (around 1 out 11 people) [3].

Climate change is another factor posing pressure on global agricultural production. Rising levels of temperatures lead to additional pest infestations, increased evapotranspiration, and reduced crop durations [4]. Additionally, the derived processes of coastal erosion and desertification limit the expansion of cropland that can be devoted to agriculture.

Food systems consistently contribute to pollution, accounting for one-third of the global greenhouse gas (GHG) emissions, with 32% resulting from land use and land use change related to agricultural production [5]. Furthermore, the contributions of agriculture to GHG emission, water pollution, and biodiversity loss are estimated to become unsustainable by 2050 if no mitigation strategies are adopted [6]. Although the study showed that no single measure among those proposed can keep food systems within environmental limits, it highlighted the adoption of technological solutions can contribute significantly to the aim. Technological measures include improvement in agricultural yields, more efficient fertilizer application, and changes in management practices (irrigation, fertilization, and pesticide application) [7, 8, 9, 10].

An important aspect of technological measures in agriculture is the application of digital technologies, commonly referred to as digital agriculture. Digital agriculture refers to information systems that analyze, monitor, and forecast changes at field scales by integrating sensed and modeled data [11]. Agricultural processes are closely linked with environmental, climatic, and genetic variables, encompassing a level of variability that was challenging to manage with traditional information and communication technologies. With the release of public satellite imagery and the wider adoption of in-



the-field sensor devices, the research community focused on applying advanced artificial intelligence techniques to agriculture [12] and investigating their limitations [13].

### 1.3 The European context

The European Green Deal is a comprehensive policy that aims at transforming Europe into the first climate-neutral continent by 2050 and to cut by at least 55% GHG emissions by 2030, compared to 1990 levels. The Green Deal strongly focuses on the need to transition to a more sustainable food production system, underlining the need to avoid food waste and decrease water, air, and soil pollution. This transition is also supported by increasing the area dedicated to organic farming. To effectively reach the objectives, the Green Deal identifies data-driven decision-making as a key technological tool that must rely on accessible and interoperable data combined with artificial intelligence.

The effects of climate change on European agriculture vary regionally. Considering viticulture, northern regions may be able to expand the area suitable for viticulture, while southern regions could face a decrease in area suitability and crop quantity and quality [14]. Further variability is expected within the same region due to differences in the environment and microclimate conditions. An example is the changes in grapevine growing dynamics deriving from altitude that will have a strong impact on viticulture in orographically complex areas [15].

Following this initiative, the European Union funded the Italian Center for Agritech, which selects and funds agritech projects that are aligned with the Green Deal objectives. The National Center for Agritech is an ecosystem of universities, research centers, and hub companies with the common aim of transferring innovation and technologies to private and public agricultural actors to strengthen Italian agriculture.

### 1.4 Innovation at Enogis

European and national initiatives, such as those previously described, significantly influence the development and adoption of smart agriculture solutions, also known as Agriculture 4.0. This is evident in the Italian market, where in 2023, the market value of agriculture 4.0 rose by 19% (reaching 2.5 billion euros). In the same year, the surface cultivated with 4.0 solutions increased to 9% and the value generated by Decision Support Systems and management software grew while that associated with hardware solutions decreased [16].

In such a striving ecosystem, Enogis decided to fund the PhD topic "Deep Learning for Predictive Agronomics" for the Industrial Innovation PhD program at the University of Trento. Enogis is a small company (around 10 people) that offers a suite for agricultural field management, strongly focusing on viticulture. Enogis originated as a project in 2011 and subsequently evolved into a formal company in 2018. As of 2024, around one-fourth of the total Italian viticulture area is managed with the Enogis software, with clients in almost all Italian regions.

The Enogis software is a comprehensive management tool designed for precision agriculture, enabling users - such as farmers and agronomists - to record both tabular and georeferenced information. This includes critical data on the appearance of diseases, activities performed, and production costs. Enogis features multiple modules, such as a cartographic platform, a cost management system, and a soil analysis module. These modules can be customized and tailored specifically to meet the unique needs of clients, enhancing the software's adaptability and effectiveness in various agricultural contexts. This flexibility allows users to optimize their farming practices and make informed decisions based on accurate data.

The Enogis development life cycle employs the Scrum methodology, which facilitates collaborative and incremental work. The development team is structured with multiple feature and product owners alongside a minimal number of hierarchical levels. This streamlined structure contributes to reaching a high level of technical innovation, enabling the team to pivot solutions and enhance features tailored to specific agricultural and management use cases. This agility is fundamental for Enogis, as Italian agricultural companies, especially wineries, often implement diverse management workflows based on factors such as territory, agronomic objectives, and climate.

With the funding of the PhD opportunity, Enogis sought to enhance its technical team by adding a specialist focused on artificial intelligence (AI) solutions who could develop and integrate such solutions with the existing ecosystem. For these reasons, the selection process investigated candidate competencies regarding research and innovation experience in developing artificial intelligence applied to agriculture. Therefore, this thesis describes research efforts regarding new AI techniques and models to be applied to contexts relevant to Enogis, as well as the processes needed to integrate these models into operational solutions.

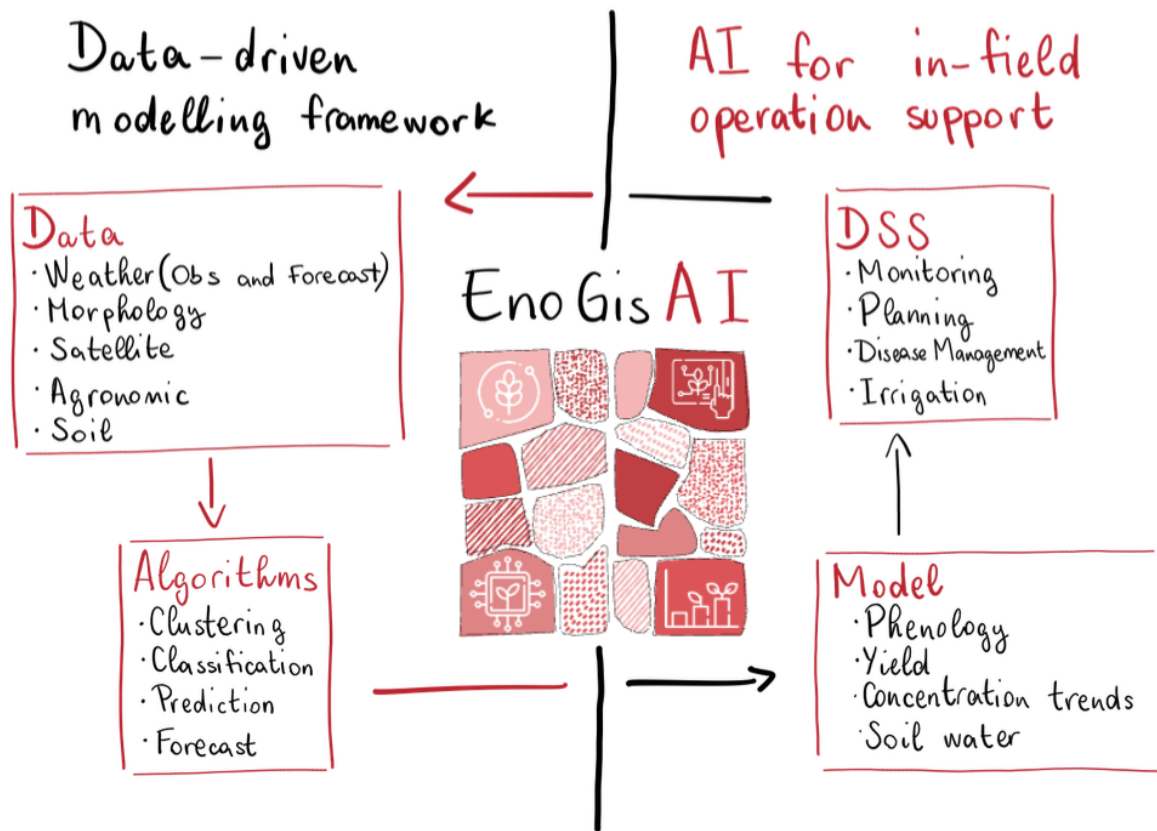
## 1.5 EnogisAI: a framework for developing and adopting agronomic AI solutions in the field

The Enogis platform started collecting agricultural field data in 2015. Initially, data was primarily archived, serving as a knowledge base for end users to enable the visualization and analysis of spatiotemporal observations. This, in some cases, proved enough to gather an idea of more general trends and insights.

Interest and an opportunity to further develop the platform to provide data-driven tools and Decision Support Systems (DSSs) derived from the combined push of public funding opportunities and market demands. The integration of AI into industrial sectors raised the interest of agricultural companies in adopting and experimenting with AI solutions to address increasing climate and management challenges. Additionally, advancements in machine and deep learning methodologies applied to agriculture, along with the release of new satellite-based and environmental data products, increased the transferability of research findings into market-ready solutions. Moreover, policies aimed at advancing agriculture's transition towards a 4.0 integration funded the development of new technologies and platforms and the acquisition of these tools by agricultural entities.

Building on the current global and local agricultural challenges and innovation trends, Enogis dedicated the last four years to creating EnogisAI: a framework for developing and adopting agronomic AI solutions in the field (Figure 1.1). EnogisAI is divided into four components, each representing a moment in the development of AI pipelines: data, algorithms, models, and DSSs. We envision this framework as a pathway for transforming data into value for end users. We begin with multiple data archives with the potential to enhance the reusability and interoperability of data, progressing through the definition of algorithms and the generation of models that utilize this data to fuel DSS applications for end users to answer concrete questions. Our goal is then to restart the entire process using DSS outputs, user feedback, and lessons learned to enhance existing solutions and unlock new opportunities for further agricultural applications.

The first described component is data, a key ingredient of any AI solution. In agricultural applications, data can represent environmental factors (e.g., soils, orography, weather, and climate), crop-related elements (e.g., cultivar, genetics, training systems), and agricultural practices (e.g., irrigation methods, agrochemical treatments, and fertilization techniques). Data can be collected using proximity or remote sensing



**Figure 1.1:** The EnogisAI framework consists of four components. First, data from internal and external sources are integrated and organized to power further application development. Second, AI and mechanistic algorithms use and are built on the data component. Third, the training and validation of the algorithms yield models that address specific agricultural tasks. Finally, models are integrated into Decision Support Systems (DSSs) solutions provided to end users (agronomists and farmers). The framework can be seen as a combination of a data-driven modeling framework (left side) aimed at providing AI solutions for in-field operation support (right side).

instruments (e.g., weather stations and satellites) or generated through modeling processes (e.g., weather forecasts and soil modeling). Combining input sources enables the generation of bioclimate and environmental indices that describe further relevant agronomic factors. Finally, aspects such as historical availability, geographical extent, and spatiotemporal resolution significantly influence their application to develop solutions for specific agricultural challenges. The EnogisAI provides a data platform to integrate, document, and serve all relevant data sources for the development of relevant DSS.

Building on the significance of data, AI modeling processes rely on the development of task-specific algorithms. This includes the creation of both supervised and unsupervised pipelines aimed at responding to relevant agronomic questions. Pipelines need to

address the complexities of data in terms of spatiotemporal resolution, availability, and volume. The EnogisAI platform serves as a collection of best practices, development tools, and reusable algorithm libraries that standardize and streamline the creation and integration of new and existing algorithms. Within Enogis, there is a strong encouragement to reuse validated model components, which accelerates the development of new applications and facilitates the adaptation of existing solutions to new use cases. Special emphasis is placed on the use of neural networks, a class of models renowned for their flexibility in handling multimodal data.

By interacting with farmers and technicians in various agricultural working conditions and with different aims, a clear demand for diverse solutions becomes evident. Among all the identified needs, during the PhD program, we focused on:

- Field management zone identification,
- Grapevine development monitoring,
- Satellite-based durum wheat yield prediction,
- Soil water balance assessment,
- Pollen seasonal time series reconstruction,
- Time series forecasting of pollen concentrations.

These tasks were selected based on the business and application needs identified while interviewing and engaging with Enogis clients and partners. Each task involved a combination of research and innovation activities, resulting in varying levels of integration within the DSS tools provided to users. Specifically, the grapevine development monitoring and soil water balance assessment solution were successfully integrated into the Enogis platform and are utilized daily by end users. Conversely, the pollen time series reconstruction and forecasting tasks represent two research studies that served as essential modeling playgrounds, allowing for comparison between neural networks with other machine and statistical learning tools.

Lastly, the models that were considered more mature and relevant were further developed into DSS instruments available on the Enogis platform. Enogis provides various tabular-based and map-based tools for managing the agricultural company. An example is the cartographic WebGIS tool that visualizes user fields and a set of other environmental and agronomic information, including notes recorded in the field by users or the output of data processing pipelines (e.g., satellite-based vegetation indices). Among the generated models (third step of EnogisAI), we selected the grapevine development monitoring and the soil water balance assessment to be integrated into three DSSs.

The first two are the integration of the models into user-friendly tools for monitoring, inspecting, and comparing grapevine development and soil water balance of fields. The third one consists of a disease coverage solution that models the lasting effects of agro-chemical treatments against the most relevant diseases and suggests new treatments considering the crop development stage.

## 1.6 EnogisAI innovative aspects

Transferring research results into market-ready products is a complex process. In the context of AI models for agriculture, these models are often tested and validated in controlled environments that differ significantly from real-world use case conditions. Additionally, research output models may not always meet the precision, accuracy, or efficiency levels required by industrial standards. These challenges can slow the adoption of research solutions, adding to the difficulties of replicating and integrating external results into established company workflows, which are often characterized by limited flexibility.

For these reasons, EnogisAI follows an innovative approach that consists of introducing AI and AI-related best practices into each aspect of the framework.

Regarding data, we introduced the EnogisAI data platform, which aims at ingesting data of different natures from various sources and with different update frequencies. Moreover, the platform seamlessly serves data to the whole Enogis ecosystem by adopting the same authentication and authorization schemes and providing an API interface. Further, the platform adopts state-of-the-art data storage solutions to efficiently archive and serve time series data to both research and production systems.

EnogisAI acted as a development environment for new algorithms and techniques operating on data ingested by the data platform. The innovations involved the definition of new data processing pipelines and new model architectures.

Regarding the former case, we can cite the field management zone (FMZ) identification and durum wheat yield prediction. These two studies developed unsupervised (FMZ) and supervised (yield prediction) approaches based on satellite-derived vegetation indices, namely the Normalized Difference Vegetation Index (NDVI) and the Normalized Difference Red Edge Index (NDRE). Indices time series were analyzed first by applying Functional Principal Component Analysis (FPCA) for data normalization and data projection step. The output was fed to machine learning models for clustering (K-means) and regression (Partial Least Square) tasks.

Regarding the latter case, i.e. proposing new model architectures for agricultural

tasks, we developed a multi-modal deep learning architecture for crop development monitoring. In addition, two other neural networks were developed for pollen concentration time series reconstruction and forecasting. These were compared to simple machine and statistical learning models. Efforts were devoted to the explainability of model prediction dynamics to gather additional insights on the training process and environmental and climate impacts on model outputs.

The machine and deep learning model experimentation and training processes followed state-of-the-art designs to ensure proper validation and avoid data leaks between training and validation data sets. Models were designed and trained with a focus on real use-case scenarios to maximize usefulness and adoption by end users. All models were integrated into EnogisAI that acts as a collector of the resulting solutions, ensuring a smooth transition from modeling to production environments.

Models are gradually transformed into DSS tools consulted by end users directly in the Enogis app. Proper integration into the already existing Enogis ecosystem, especially within the tools that users are already accustomed to, significantly increases the likelihood of DSS utilization by end users. This, in turn, enhances the impact of DSSs on the efficiency and sustainability of agricultural management practices. For these reasons, efforts were devoted to proposing and introducing new technological standards to the Enogis ecosystem. Examples are the adoption of JSON Web Tokens for authorization purposes, model deployment solutions, and centralized data lakes for accessing DSS results.

Finally, the DSS and, in general, the whole framework was developed for use in the whole Italian context, eventually scaling to other European countries in the future. This is made possible by prioritizing access and integration of data sources that cover extensive areas and that are not focused on specific territories. Nonetheless, the framework is designed to allow the usage of data sets characterized by limited geographical applicability but improved descriptive accuracy.

## 1.7 Structure of the thesis

All chapters of the thesis are centered around the concept of EnogisAI. For this reason, instead of presenting each of the works developed during the PhD program as standalone pieces, the thesis adopts a thematic approach while describing each of the EnogisAI components.

Chapter 2 outlines existing literature on AI's application in agriculture, with a focus on the research tasks addressed in the thesis.

Chapter 3 describes the data sources and records used in the development and production environments of EnogisAI, both internal and external. Internal data are available within the Enogis context (e.g., field activities data collected through the agronomic seasons), while external data belong to public or commercial sources. The chapter also describes the preprocessing and harmonization steps undertaken to integrate the data into EnogisAI and preliminary statistical insights useful to describe the data.

The research methods to develop AI solutions for agriculture are presented in Chapter 4. These include the problem and objective formalization of each analyzed scenario, the description and selection of AI algorithms, and the analytical techniques used to evaluate the performance of AI models.

Chapter 5 outlines the resulting models learned and constructed from the available data. It presents specific model applications relevant to the agricultural field and their results and limitations.

Chapter 6 extends the previous concept by incorporating AI models into applications tailored for agronomists and farmers. This thesis treats model predictions as ingredients for powering Decision Support Systems (DSS) to assist field monitoring, activity planning, and activity efficiency. The aim is to demonstrate the effectiveness of AI in enhancing vineyard operations and their implication within the viticulture industry. The chapter also explores the potential benefits and challenges of integrating AI technologies in vineyard management.

The key findings and overall contributions of this thesis to AI in viticulture are summarized in Chapter 7. The chapter also identifies the limitations of the current research and proposes directions for future studies, emphasizing the need for research in AI applied to agriculture to address future challenges.

Each chapter (except for Chapter 2) presents two boxes at the beginning, one summarizing the highlights and key points of the chapter and one underlining the personal contributions of the candidate.



# Chapter 2

## State of the art

Agriculture plays a fundamental role in feeding the world’s population while aiming for the sustainable use of available resources. The contributions of agriculture to GHG emission, water pollution, and biodiversity loss are estimated to become unsustainable by 2050 if no mitigation strategies are adopted [6]. The growing population and climate change are among the most urgent challenges that the sector must confront [1, 4]. The advancement and adoption of technology can enable smart farming and enhance agricultural productivity and sustainability [6, 11, 17].

Other aspects to carefully consider during the development of smart farming tools are the adoption barriers related to the agricultural industry and the socio-economic aspects of farmers and agricultural companies. Factors such as age, instruction levels, attitude toward information collection, the perceived complexity of new tools, and intensity of field labor highly impact the adoption of smart farming tools in existing companies [18].

Machine learning and artificial intelligence were applied for crop disease, water management, and yield management [19]. However, the study also showed the transition from research results to market-ready products is associated with a wide number of challenges. Examples are farmer’s data privacy, size of available historical records, quality of data, data acquisition difficulties, and the scalability of AI algorithms to other spatial and climatic contexts.

The EnogisAI platform design aims to tackle these challenges by providing a solution that considers all aspects of the artificial intelligence development process and that integrates smoothly with solutions already adopted by end users.

Considering that the thesis explores new AI solutions for agriculture and aims to transfer these innovations into industrial applications within the context of Enogis, we

primarily focused on relevant challenges related to the Italian agriculture and viticulture sectors. However, this focus does not preclude the potential reuse of model components or research results for other sectors and applications. Specifically, we mainly focused on (i) the use of remote sensing and field images to monitor crop growth monitoring, (ii) field management and yield estimation with satellite images, and (iii) soil water stress estimation.

Additionally, we addressed two issues related to pollen concentration time series, i.e. reconstruction and forecast. The algorithms developed in this context contribute to the thesis framework in two ways. First, they demonstrate the feasibility of using AI to reconstruct corrupted time series and generate autoregressive models. Second, they deal with pollen data, which can be considered a proxy and enabler of agricultural dynamics, making it relevant for integration into more agricultural-specific workflows.

This chapter reviews the literature related to the presented tasks.

## 2.1 Crop development monitoring

Phenology is the study of crop development resulting from climatic inputs [20]. Phenology is extremely important to growers for two main reasons. First, the efficiency of many in-field activities, such as fertilization, irrigation, and pesticide application, heavily depends on crop growth processes. Second, the ripening phase impacts the final yield quality and quantity. Grapevine phenology is highly affected by climatic variables, and their role is expected to become even more crucial due to the projected effects of climate change [14].

Monitoring and forecasting grapevine phenology can help growers plan all in-field activities more effectively. Consequently, farmers would fulfill seasonal production objectives more easily, substantially improving both economic efficiency and environmental sustainability.

Recently, the research community focused on the development of data-driven analyses for modeling plant development. The two main approaches consist of using in-situ and remote sensing time series data or images collected in the field.

Regarding the first approach, studies have applied machine and statistical models to predict the dates of specific phenological stages such as bud break [21, 22], flowering [23, 22], veraison [23, 22], and ripening [23, 24].

Other studies focused on estimating more phenological stages throughout the season from temperature data. Blanco et al. [25] used remotely-sensed temperature time series to predict 6 phenological phases for grapevine. The deep learning model reached a mean

absolute error of 5.96 days on the validation set, outperforming a Growing Degree Days (GDDs) approach (mean absolute error of 35.48 days). Campillay-Llanos et al. [26] calibrated a monomolecular equation to predict 9 phenology classes from GDD. The work performed validation considering 8 varieties of sweet cherry, reaching a mean absolute error of 0.79-1.10 phenological classes.

More recently, Yang et al. [27] and Han et al. [28] modeled the complete crop phenological cycle by applying deep learning models to image data. Deep learning architectures, or neural networks, represent a family of complex models directly learned from data. These models are characterized by high generalization capabilities and have reached impressive results also in analyzing time-series data in other fields [29, 30]. Despite the proven benefits of applying deep networks to time-series modeling, little attention was given to the development of a multi-modal model for forecasting all phenological stages.

The other common approach consists of estimating phenology from images taken with a camera or smartphone. The main advantage of this approach is the opportunity to visually inspect the input data of the model's estimation.

Field images may be acquired through fixed cameras [31], manual collection [28], or mounting devices on agricultural vehicles [32] or Unmanned Aerial Vehicles (UAV) [27].

Collected images are then used to train deep learning models for estimating crop phenological aspects. The most used deep learning models are usually inherited from the broader image classification literature. Examples are VGG and ResNet, used for phenology of the apple flower [33], classification of jujube fruits according to their level of maturity [34], classification of phenology in horticultural crops [35], and phenotyping of vine in the field [32].

In the context of grapevine image analysis, Milella et al. applied AlexNet, VGG-16, VGG-19, and GoogLeNet for cluster segmentation, with VGG-19 reaching 91.52% of accuracy and 88.52% of recall [32]. Further, Shen et al. successfully applied the Mask R-CNN instance segmentation model for detecting grape bunches, obtaining a mean average precision value of 81.53% [36]. In addition, they estimated the ripening stage of clusters based on the hue value of the HSV color space. Schieck et al. used the ResNet, DenseNet, and InceptionV3 deep networks to predict phenologies of single bunches [37]. The work compares the three methods with human evaluation with a special focus on fruit development stages (i.e., from "fruit set" to "majority of berries touching"). ResNet results as the best phenology classification model reaching an accuracy of 88.1%.

Although deep learning models are expected to yield better results compared to

other approaches, they also require larger datasets for training compared to machine learning models. To address this data-limitation issue there exists publicly available datasets containing grapevine images and phenological information [38, 39, 40, 37]. These datasets represent useful resources for extending the training and validation datasets for deep networks.

## 2.2 Field management and yield estimation with satellite images

Remote sensing images present an important opportunity for continuously monitoring crops with adequate spatial, spectral, and temporal resolutions. While UAVs collect very high-resolution data, sometimes at the submeter level, they are also associated with higher costs and smaller geographical extent compared to satellite missions.

The Sentinel missions are part of the European Copernicus program, and provide satellite images at a resolution from 10 to 60 m, with a revisit frequency of 5 days at the equator (and 2-3 days at mid-latitudes). Satellite images and, in particular, satellite-derived vegetation indices were used to detect the phenological development of winter wheat and winter barley using Sentinel-1 and Sentinel-2 data [41]. Sentinel-2 data was also used for quantifying hail [42] and frost [43] crop damages. In [44], the authors applied random forests, support vector machines, and boosted regressions for estimating wheat grain yield at the sub-field level. The work used Sentinel-2 bands together with other vegetation indices as input variables and considered temporal sequences of images for the prediction. Random forest outperformed the other models reaching  $R^2 = 0.89$ , and  $RSME = 0.74$  t/ha.

While Sentinel-2 is commonly adopted for its medium-high resolution and open access, other works experimented with higher-resolution missions. For example, Amankulova et al. compared Sentinel-2 with Landsat-8 and PlanetScope data [45]. Landsat-8 is characterized by a 30 m horizontal resolution and a total of 9 spectral bands. PlanetScope Super Dove satellites provide eight spectral bands with a pixel size of 3 m and a daily revisit time. The study adopted the random forests algorithm for estimating soybean yield and yield variability within fields using the different remote sensing sources. Sentinel-2 resulted in the best-performing one, with RMSE ranging from 0.069 to 0.188 t/ha for the four considered fields.

Further, Farmonov et al. proved that combining in the same network Sentinel-2 and PlanetScope data could improve the performance of the model, compared to using only

one of the sources [46].

Remote sensing data and vegetation indices play a key role in discriminating among field areas that require different management practices (e.g., fertilization or irrigation). For example, Nutini et al. combined Sentinel-2 images and in-field smartphone measurements to generate prescription maps with the suggested amount of nitrogen to apply for each rice field sub-area [47]. The approach resulted in increased yield (0.2-0.5 t/ha) and nitrogen use efficiency (-7.8%), leading to improved economic and environmental sustainability. Sentinel-2 image data was proved to be an adequate tool also for grapevine management zone identification, although the use of UAVs yields better management maps [48].

Integrating satellite images over time enables a comprehensive analysis of seasonal crop variations by utilizing pixel-based time series data [49, 50]. This approach allows for building more robust models that can be updated during the season while considering the whole crop evolution.

Time-series dimensionality reduction algorithm can be used to reduce data complexity and simplify modeling tasks, especially in data-limited contexts. To this end, functional-PCA (FPCA), the extension of PCA to functional data, was successfully applied for vegetation classification using satellite data [51, 52].

## 2.3 Soil water stress estimation

Water management is critical in agriculture, as it directly influences crop health, yield, and overall farm productivity. Effective water management practices ensure that crops receive the right amount of water at the right time, which is essential for sustaining plant health and maximizing yields.

Agricultural outputs and sustainability practices can widely benefit from precise soil moisture monitoring. This can give hints on the water requirements for crops and the amount of available water in the root zone area [53]. The adoption of in-field sensors revolutionized agricultural practices by providing real-time data on soil moisture levels, crop water requirements, and environmental conditions. These sensors enable farmers to monitor their fields more accurately and adjust irrigation schedules accordingly. By integrating sensor data with decision support systems, farmers can optimize water use, reduce waste, and improve crop health [54].

Recently, remote sensing approaches were developed for estimating soil moisture on larger geographical extents. Kisekka et al. applied random forests to Landsat satellite data for estimating soil moisture in vineyards [55]. The random forests model was built

considering additional covariates such as weather time series and soil properties. Yin et al. [56] estimated soil moisture in the root zone of alfalfa by applying a deep learning model to UAV visible, multispectral, and thermal data. The main drawbacks of remote sensing reside in the limited revisit time of satellites and the high costs of operating UAV flights with high frequency.

Another option for modeling soil moisture is to adopt mechanistic agronomic models. These describe soil water balance processes by physically describing the role of soil properties and simulating soil moisture and other crop-related variables, such as evapotranspiration or water requirements. Further, modeling techniques allow for simulating and evaluating various climate and irrigation scenarios. To this aim, the SIMDualKc model was used to discriminate between the transpiration and evaporation contribution to total evapotranspiration for grapevine [57] and other fruits [58, 59]. The SIMDualKc method is based on the FAO-56 notebook for estimating evapotranspiration and soil water balance [60]. The SIMDualKc model works with hourly or daily time steps and relies on precise field and crop parametrization. An open question regards the usability and performance of the model in large extent contexts by adopting open geographical data that describe the territory eliminating the necessity of field-specific analysis on the soil and other agronomic relevant variables.

## 2.4 Pollen concentration time series

Pollen plays a key role in multiple aspects of society, including health, environment, and agriculture. Considering agriculture, pollinator diversity is linked to increased yield quality and quantity [61, 62], while changes in pollen seasonality due to climate change may impact agricultural seasonality [63, 64, 65]. Given the links between aero-allergenic pollen and weather, accurate and time-sensitive pollen forecasts would simplify the monitoring of flowering dynamics, especially for fruit trees [66, 67].

Pollen concentration forecast is commonly addressed with process- or observation-based methods [68]. The former aims at modeling the physical processes that connect climate, environment, and physiological traits of pollen production to estimate development dynamics. The latter is based on data collection processes and consists of training statistical and machine learning methods to predict future concentration.

Recently, the research community focused mainly on developing machine and deep learning models for estimating pollen concentration forecasts under different conditions. Compared to mechanistic models, AI methods (1) can deal with the complexity of the release and dispersion of airborne pollen [69], (2) are computationally efficient [70],

(3) can be easily integrated into operational use of the models [71], and (4) provide substantially higher forecasting power compared to traditional regression models [72]. The main disadvantage of AI methods is their lower interpretability [71, 73], their limited effectiveness where smaller data is present (e.g., at the beginning and end of the pollen seasons) [74], and their sensitivity to the training data [71, 74].

Zewdie et al. developed random forests and deep learning models for predicting *Ambrosia* pollen concentration for Tulsa, Oklahoma (USA), reaching a Pearson correlation coefficient of 0.82 and 0.81, respectively [75]. A neural network developed for predicting *Parietaria* pollen concentration for the next three days obtained an average  $R^2$  of 0.50, in the Atlantic coastal regions of Spain [76]. Navares and Aznarte in [77] and Muzalyova et al. [78] focused on next-day pollen concentration estimation obtaining an  $R^2$  of 0.67 (in the best-performing location) and 0.62. The first employed random forest and focused on the *Plantago* pollen in the Madrid region, Spain, while the second developed a neural network and considered the *Betula* and *Poaceae* pollen concentrations on a 3-hourly resolution for Augsburg, Germany. The lower  $R^2$  value obtained by Muzalyova et al. [78] is likely associated with using sub-daily values and the deriving modeling complexity. Finally, Lops et al. in [79] have developed a convolutional neural network that obtained Pearson correlation coefficients of 0.83 and 0.69 for predictions for the 1st and 7th day.

A potential limitation in the usability of historical pollen databases for analyzing climatic and environmental contexts or for developing forecasting solutions is the presence of missing data. Gaps in these datasets can reduce their reliability and accuracy, particularly when attempting to identify long-term trends or correlations with climate variables. Missing data may arise due to inconsistent sampling originating from mechanical or social challenges associated with data collection processes. Pollen records are often compiled by manually counting the number of pollen particles present on adhesive rotational tapes using a microscope. Such manual processes are error-prone and can result in missing pollen observation even for multiple consecutive days.

Missing data imputation strategies can be developed with statistical interpolation methods or machine learning. Picornell et al. tested various interpolation methods for restoring pollen missing data in specific time windows (linear interpolation, moving mean interpolation, and spline interpolation), by interpolating observed data and decomposed seasonal trends, and interpolating data with nearby locations data [80].

Moving mean interpolation resulted in a lower reconstruction error compared to the other methods. Navares et al. compared a convolutional neural network (CNN) approach against more established geostatistical interpolation methods (i.e., kriging

and inverse distance weighted interpolation) [81]. The study did not consider past and future values when reconstructing missing data but focused solely on the surrounding context of specific dates. The CNN resulted in being the best-performing model, however, the study suggested that incorporating weather observation or forecasts could further improve performance. The use of multi-variate or multi-modal data for data imputation represents a complex data integration and modeling task. In other fields, deep autoencoders were successfully employed to address this further complexity and represent an interesting opportunity for multi-modal pollen concentration reconstruction [82, 83, 84, 85, 86].



# Chapter 3

## The EnogisAI data platform

### Highlights

- The data platform catalogs all the available data integrated into EnogisAI for analysis, model development, and operational usage.
- Data characteristics include spatial resolution (DTM: 20 m, ERA5: 9 km), update frequency (e.g., static, annually, daily), and the distinction between raw or derived products (e.g., bioclimatic indices and solar irradiation maps) and observed or forecasted ones (e.g., pluviometer data, ERA5 reanalysis, HRES forecast product).
- Weather and sensor time series data were migrated to TimescaleDB (total of more than 2 billion data points; current rate: 3.3 million points per day). Compared to PostgreSQL, TimescaleDB offers efficient data archiving (6.19x compression factor) and querying (40% faster).
- Data integration modules were developed for satellite, weather station, and field sensor device data (total of 11 providers), scheduled and executed with Celery, and monitored through Sentry to ensure quality of service.

**Personal contribution:** I integrated the 11 described data sources into the EnogisAI data platform. I developed the integration modules for weather and sensor data into the platform and managed the migration from PostgreSQL to TimescaleDB. I am the operative manager of the system.

---

Agriculture refers to a series of biological processes mainly linked to non-controlled environments (i.e., open fields). Several factors can influence these processes: cultivated crop species and cultivars, micro and macro climate and weather patterns, surrounding environment, and human activity. From a temporal point of view, inter- and intra-year variability of such factors impact agricultural processes. One of this thesis aims is to provide a set of instruments to end-users (farmers and agronomists) to monitor crops and predict their evolution accurately. Consequently, a key aspect of EnogisAI is the integration of various and heterogeneous data sources to allow easy integration with data analysis and model development pipelines. For this reason, the first component of EnogisAI described in this thesis is the EnogisAI data platform. The data platform can be considered an abstraction layer that integrates all data sources that can eventually be accessed in specific project pipelines.

Data sources can be organized into internal and external. Internal sources represent all the data collected within the company boundaries through Enogis company processes or software products used by the end users. Examples are all field activities recorded in the electronic field records (EFR) software provided by Enogis or manual data assimilation procedure of specific clients' data (e.g., soil maps derived from analysis in the particular territory). External sources include all data that can provide additional information and context to improve agricultural modeling processes. Examples are weather data, remote sensing data, and static environmental descriptors.

Another aspect to consider of the data is the geographical domain. Since Enogis operates mainly in Italian territory, the ingested data usually covers various portions of such territory. The percentage of covered territory differs due to business or technical reasons. First, there are the Italian regions where Enogis is not operating yet and was not interested in expanding during the PhD program. These are Liguria, Umbria, Lazio, Basilicata, and Calabria. Second, some data necessitate significant storage volume or processing computational resources due to their inherent characteristics (e.g., satellite data at medium/high resolution). For this reason, it is preferable to only retrieve and store this kind of data for part of the considered territory, leading to discontinuous monitoring areas but improving costs and benefits evaluations. This decision proves strategic when considering that Enogis aims to improve its economic sustainability through all its innovation processes in the medium to long term. All data ingestion processes are extensible, and the areas that are initially left out can be integrated later.

The data platform aims to describe how to retrieve, organize, and access all relevant data sources. As noted earlier, Enogis has a distributed software architecture consisting of multiple services deployed over different virtual machines (VMs). The idea behind

developing the EnogisAI data platform is that it should integrate as smoothly as possible with the existing software and hardware infrastructure. This need, combined with the high heterogeneity of data in terms of type (e.g., time series, geographical data, tabular records), led to the design of a centralized platform that describes how to retrieve, access, and use the data. By realizing this concept, it is possible to integrate data in multiple modeling workflows while eventually allowing the existing software to integrate the data platform approach.

This chapter presents the internal and external data sources we considered, underlines their role in the context of this thesis and EnogisAI, and concludes with an outline of the technical solutions we used to integrate them.

## 3.1 Internal data

The Enogis suite aims to help farmers and technicians manage their vineyards (and other agriculture fields in a more limited fashion). Consequently, it models a large number of entities and their relationships. The main entity is the field, where all other information is centered. The AI models developed in EnogisAI are mostly designed to work at the field level, as opposed to work for gridded areas (e.g., weather models). The rationale is that while environmental models have limited dependency on human input, agriculture is strongly influenced by proactive activities. Moreover, vineyards of different cultivars react in various ways to human inputs, introducing the need for modeling approaches that consider this aspect. The reliance on field-specific and activity-related data in the developed models restricts their applicability to areas where vineyards already exist. This aspect poses a challenge in creating effective solutions for vineyard expansion planning and similar contexts.

The field and associated relevant entities are presented next.

### 3.1.1 Field records

Enogis’s field records contain all relevant agronomical and geographical information. Users can operate on fields in graphical (WebGIS software) or tabular environments, offering advantages and disadvantages. Examples of field metadata are administrative area information (municipality, province, etc.), species (there are other possibilities in addition to vine), cultivar, training system (guyot, pergola, etc.), number of plants, cultivated surface, planting year, soil information, and information on the irrigation systems. Fields are associated with a single farm, which is associated with a single win-

ery. Fields may change between agricultural seasons depending on farmers’ needs and decisions; thus, year-to-year matching is usually performed with a fuzzy logic matching on geographical and farming attributes.

### 3.1.2 Field observations

The Enogis software enables farmers and agronomists to collect relevant events and observations directly from the field using a mobile app. The primary interest when collecting such data is to trace field conditions during the current seasons and compare them to previous seasons. For this reason, the software provides the capability to visualize the collected samples on a map, giving the opportunity to identify spatial patterns that may explain the onset of disease or other behaviors. The logical progression is to leverage the intrinsic value of collected data to generate predictive insights.

Users are interested in monitoring various aspects. To this end, Enogis offers structured forms to compile agronomic management practices checklists, organize general observations, store pre-harvest and yield estimation analyses, and collect phenology and pathology samples. Further, clients may require additional customization in the common scenario when they are used to retrieve and analyze additional information, especially for specific use cases (e.g., specific micro-climates, quality projects, or pathological issues). Consequently, users are granted access to custom additional sheets or edited versions of the described ones. A brief description of the standard data sheets is given below, with an indication of how many samples are collected for each category (Table 3.1).

The **agronomic management** practices checklist enables monitoring of actual vigor and irrigation stress levels, reporting on the completion of winter activities, defining the overall management of the vineyard (e.g., row and sub-row management, pruning methods), and indicating any operations that are needed to enhance field condition.

**Table 3.1:** Summary of agronomic data samples (and corresponding year) collected with the Enogis software. Entries are sorted by count (descending).

| Data collection form    | Count | Since | To   |
|-------------------------|-------|-------|------|
| Pre-harvest             | 47529 | 2000  | 2024 |
| Agronomic management    | 18161 | 2015  | 2024 |
| Phenology and pathology | 12695 | 2016  | 2024 |
| Yield estimation        | 10063 | 2015  | 2024 |
| General observation     | 2769  | 2015  | 2024 |

**Phenology and pathology** forms are usually compiled during the whole season. They permit the description of the growing phases of the vineyard (phenology) and the monitoring of the emergence of eventual pathologies. The phenology is expressed with the BBCH scale [87, 88], a categorical scale that encodes all plant development stages into a two-digit system. The first digit (from 0 to 9) describes the principal growth stage (i.e., 0: Sprouting/Bud development, 1: Leaf development, 5: Inflorescence emerge, 6: Flowering, 7: Development of fruits, 8: Ripening, 9: Senescence), while the secondary digit indicates the secondary growth stage.

These act as a percentage indicator or ordinal numbers of the principal stage. The BBCH scale is categorical, thus distances between the numbers do not reflect real distances. Each principal growth stage can have different durations, and successive phenological phases can be relatively close or distant in time. The same form allows the recording of pathologies, which are usually closely associated with phenology, along with a disease severity grade.

**Yield estimation** sheets are used to record plant productivity features, including graft failure percentage, average cluster weight, number of clusters per plant, and field homogeneity status. These features are especially monitored after extreme weather events, such as droughts or hail, in order to monitor the impact on grape productivity.

The **pre-harvest analysis** consists of laboratory analysis conducted close to harvest to improve the understanding of the ripening profiles of vineyards. These analyses are usually performed with spectrometers (e.g., WineScan SO2, Foss, Denmark<sup>1</sup>) and allow to record various quality parameters of sample harvested grapes. Although portable spectrometers exist, laboratory ones are still identified as more reliable from wineries [89, 90]. However, such instruments are usually expensive and thus available in a limited number. This leads to critical operation planning to avoid sampling deterioration caused by heavy analysis loads and long queues. Pre-harvest analysis results include (but are not limited to) pH, malic acid, anthocyanins, polyphenols, and soluble solids content. The georeferencing of these parameters is possible as long as the geographic and field information of the grape samples are tracked through the lab process. The Enogis software can connect to the most adopted lab instruments, so results are automatically uploaded to the users' systems.

The **general observation** sheet is the less structured form, and it allows for free note-taking. Such a format is particularly useful for documenting all those situations that do not follow under a specific category but are nonetheless worth capturing. This

---

<sup>1</sup><https://www.fossanalytics.com/en/products/winescan-so2>

form could eventually be a feature development trigger for more specific sheets. Given its generic and unstructured nature, this form is of limited use in further analysis.

As a general note, all in-field data collection forms offer the capability to store images as additional evidence for further analysis or archiving purposes. For example, phenology or pathology images could be later used for remote or (even automatic) classification of field conditions.

In this thesis, phenological, pre-harvesting, and yield estimation data have been used for training and validating machine learning models, as explained in the next chapters.

#### 3.1.3 Agricultural activities

Italian farmers are legally obliged to maintain a field notebook in which they record various agricultural activities performed during the season. The Enogis suite offers an electronic field records tool (Italian: *Registro di Campagna*) to plan, note, and archive such activities efficiently. In addition, the software runs several tests to confirm activity compliance. These tests originate in the production specifications, which contain all the indications and restrictions that farmers must comply with. Examples include limits for pesticide use, time period allowed for specific activities, names of allowed products, and irrigation limits. Further, the specifications can be issued by political actors (e.g., the European Union and national or local governments), vineries, or production consortia. These last two have the only option to issue more stricter rules.

Among all data recorded in field notebooks, only disease treatments and irrigations are discussed here, being involved in the rest of the thesis.

Treatment forms are designed to record essential information, including the date, field details, weather conditions, operator identity, products used (together with the amounts), and diseases addressed (Figure 3.1). Following practical farmers' approaches, more products can be used to address multiple diseases in a single treatment record. The number of treatments performed during a season can vary significantly due to climatic and agronomic factors. For instance, it is more common to treat fungal infections in more humid regions. Overall, the number can range from 3-4 to as many as 15-20 and over per season.

Irrigation forms have a simpler structure compared to treatment ones. They document field details, single or multiple dates (or time windows for recurrent irrigation activities), water volume, and eventual system loss. (Figure 3.2). Similarly to treatments, different climates and agronomic factors (e.g., cultivars and agronomic objectives) impact the number of irrigations. Numbers range between 0-1 and 10-12 irrigations per

Torna alla lista Trattamento: 2022-05-18 Stampa PDF

Anticipazione culturale (attività post-raccolta) ☐

Appezamenti Scegli 6 Appezamenti, Sup. 0.3982 ha (3982 m<sup>2</sup>)

Tipo trattamento [A] Trattamenti e fitoregolatori Data 18/05/2022 Concentrazione 1 Modalità di distribuzione Scrivi e seleziona

Ettoltri d'acqua 1,991 Ettoltri / ha 5

Meteo Clicca e seleziona

Note

Solo agrofarmaci ammessi e non revocati ☒ Solo avversità valide ☒

| Operatore | Ore operatore | Attrezzatura      | Ore lavoro attrezzatura | Attrezzatura pulita | Trattrice | DPI utilizzati |
|-----------|---------------|-------------------|-------------------------|---------------------|-----------|----------------|
|           | 0,00          | GNAGNARELLA 300PP | 0,00                    |                     | Search... |                |

Aggiungi

Agrofarmaci usati nei trattamenti 2 Prodotti concimazioni 0 Biostimolanti usati nei trattamenti 0

| Revisione agrofarmaco                               | Avversità              | Dose totale                                    | Dose ettolitro                     | Dose ettaro                        | Etichette e pericoli                                  | Proprietà                |
|-----------------------------------------------------|------------------------|------------------------------------------------|------------------------------------|------------------------------------|-------------------------------------------------------|--------------------------|
| Attenzione: sono visualizzati solo prodotti ammessi |                        | <input checked="" type="checkbox"/> tot.[kg-l] | <input type="checkbox"/> [kg-l/ha] | <input type="checkbox"/> [kg-l/ha] |                                                       | terzista                 |
| MICROTHIOL DISPERS (fino al 05/07/2024)             | OIDIO DELLA VITE       | 1,0                                            | 0,502                              | 2,511                              | <span>Info</span> <span>Info</span> <span>Info</span> | <input type="checkbox"/> |
| POLYRAM DF (fino al 28/11/2024)                     | PERONOSPORA DELLA VITE | 0,8                                            | 0,402                              | 2,009                              | <span>Info</span> <span>Info</span> <span>Info</span> | <input type="checkbox"/> |
| Scrivi e seleziona...                               | Scrivi e seleziona     |                                                |                                    |                                    |                                                       | <input type="checkbox"/> |
| Scrivi e seleziona...                               | Scrivi e seleziona     |                                                |                                    |                                    |                                                       | <input type="checkbox"/> |
| Scrivi e seleziona...                               | Scrivi e seleziona     |                                                |                                    |                                    |                                                       | <input type="checkbox"/> |
| Scrivi e seleziona...                               | Scrivi e seleziona     |                                                |                                    |                                    |                                                       | <input type="checkbox"/> |
| Scrivi e seleziona...                               | Scrivi e seleziona     |                                                |                                    |                                    |                                                       | <input type="checkbox"/> |
| Scrivi e seleziona...                               | Scrivi e seleziona     |                                                |                                    |                                    |                                                       | <input type="checkbox"/> |
| Scrivi e seleziona...                               | Scrivi e seleziona     |                                                |                                    |                                    |                                                       | <input type="checkbox"/> |

Cancellazione Salva e continua Salva e aggiungi un altro Salva

**Figure 3.1:** Example of filled disease form in the Enogis field notebook tool. Data collected in this and similar forms are then used for successive data modeling and analysis phases.

Torna alla lista Apporto idrico: 2024-06-14

Anticipazione culturale (attività post-raccolta) ☐

Appezamenti Scegli 2 Appezamenti, Sup. 0.9859 ha (9859 m<sup>2</sup>)

Note

Date o periodo

Date o periodo Data singola Data singola 14/06/2024

Singolo apporto

Tipo di apporto Volume totale del singolo apporto Perdita % 0,0

Unità di misura Metri cubi [m<sup>3</sup>] Superficie totale Valore 0.2

**Figure 3.2:** Example of real irrigation form in the Enogis field notebook tool. Data collected in this and similar forms are then used for successive data modeling and analysis phases.

season.

As a concluding note, both treatments and irrigations are used as input information for the DSS, which is presented in the following chapters.

## 3.2 External data

Agricultural development is closely linked to weather and environmental variables and processes. To effectively develop AI solutions that are able to accurately model crop and pest dynamics, it is paramount to include such variables. This section presents the external data sources used in EnogisAI. These sources may include static variables, such as environmental descriptors (e.g., altitude), as well as dynamic ones, such as geographical time series (e.g., weather observations). Table 3.2 summarizes the external data products integrated into the EnogisAI data platform and later used for further modeling.

### 3.2.1 Weather

Weather data can be classified based on various characteristics. Here, we adopt station data in comparison to model data, as well as observations in contrast to forecasts.

In the following, we describe the weather data sources considered and integrated into the EnogisAI data platform during the program.

**Table 3.2:** External data products integrated into the EnogisAI data platform. Static datasets are manually organized, while, for the others, we developed specific data ingestion workflows. All data products were used in the later model development phases.

| Category    | Name              | Typology                  | Horizontal res. | Temporal res.       |
|-------------|-------------------|---------------------------|-----------------|---------------------|
| Weather     | ERA5              | Reanalysis                | 9 km            | hourly              |
|             | HRES              | Forecast                  | 9 km            | 1-6 hours           |
|             | Weather stations  | Real-time                 | -               | 1- 60 min.          |
| Satellite   | Sentinel-2        | Passive sensor            | 10 - 60 m       | 2-3 days at mid-lat |
|             | Planetscope       | Passive sensor            | 3 - 5 m         | daily               |
| Environment | TINITALY/01       | DTM                       | 10 m            | static              |
|             | GLO-90            | DTM                       | 90 m            | static              |
|             | EU-SoilHydroGrids | Soil hydraulic properties | 250 m           | static              |
|             | GCN250            | Curve numbers             | 250 m           | static              |



## Weather stations

In-field weather stations are useful instruments frequently adopted by agronomists and farmers. On the one hand, they collect detailed data regarding precision and time resolution (stations can generally collect data starting from a 1-minute frequency and up to 1 hour). These characteristics prove especially useful in those territorial contexts where weather monitoring is crucial for crop survival or ensuring yield quality. On the other hand, the associated hardware and maintenance costs make them hardly scalable in the territory. For this reason, they are usually placed in limited numbers and critical areas, which may not deliver generally applicable insights, especially for areas farther away from the station.

Weather stations can monitor a wide number of parameters relevant to agriculture, the most common being air temperature, air humidity, and precipitation. Depending on the needs for specific applications or environmental conditions, stations can have additional sensors to monitor variables such as solar radiation, wind intensity and direction, leaf wetness, and soil temperature and water content. These measures can also be aggregated and combined to compute additional weather indices, such as growing degree days or potential evapotranspiration ( $ET_0$ ).

Modern weather stations are usually distributed with connectivity hardware that allows easy integration on commercial web platforms for data visualization and exploration. These platforms usually offer an API interface that is used to integrate observations within the EnogisAI ecosystem. The general approach is that upon the request of clients to integrate their stations in Enogis, we develop a data collection workflow that continuously downloads data from the source into the data platform.

## Weather models

The major limitation of weather stations is the very limited access to forecast values. While some initiatives exist for adding forecast capabilities to weather stations [91, 92, 93], the major issue is the general lack of context of surrounding conditions. A single weather network cannot usually offer enough power to predict future weather stations accurately.

Global or regional weather forecast models are often used to overcome this limitation. These are based on mathematical and physical equations, model a large number of atmospheric parameters, and are defined over a grid surface.

The main weather forecast model considered in this thesis is the ECMWF High Resolution (HRES) model. HRES provides forecasts for the next 10 days, at 1-hour

intervals for the first 90 hours, 3-hour intervals from hours 93 to 144, and 6-hour intervals from 144 to 240. The model is run every 6 hours starting at 00 UTC (Coordinated Universal Time).

While weather stations suffer from scalability problems in terms of spatial coverage, weather models solve this issue at the cost of offering low spatial resolution. The main reason is that they are based on physical simulations that are computationally costly. The horizontal resolution of HRES is  $0.1^\circ$  on the lat/lon grid (original resolution is 9 km), which may be impractical for specific agriculture modeling tasks. Orographically complex areas (such as the Alps or the Apennines) are associated with high elevation, aspect, and slope variability, leading to the emergence of micro-climates that can hardly be observed with a 9 km grid. Nevertheless, while other models exist for the Italian context (e.g., ICON [94]), ECMWF HRES was chosen for its limited costs, service robustness, and lead time prediction (10 days).

The HRES parameters integrated into the data platform are relative humidity, air temperature, and total precipitation.

## Reanalyses

Another advantage of weather models over stations is the availability of past observations, known as reanalyses. The modeling techniques adopted in this thesis (ML and DL) require training on historical observations, a process facilitated by reanalysis. Reanalyses are generated by integrating various data sources, such as ground observations, and utilizing weather models constrained by actual observed data during their initialization and evolution. This process results in model-derived observations that cannot be as precise as weather stations, but that cover the whole interest surface (in this case, the Italian territory).

Given the choice of the operational forecasting system (HRES), we decided to adopt ERA5-Land, a global-land reanalysis dataset providing hourly data with the same horizontal resolution of HRES ( $0.1^\circ$ , or approximately 9 km) [95]. ERA5-Land, and ERA5 (global reanalysis also considering seas and oceans) [96] are generated by ECMWF (European Centre for Medium-Range Weather Forecasts) and distributed by the Climate Data Store, a system belonging to the Copernicus earth observation component (part of the European Union's Space program). In the context of the thesis, ERA5-Land is mainly used for reconstructing past weather conditions when training AI models or generating climatic indices aggregating 20 or 30 years.

Another data product used to generate higher resolution climatic indices is the

VHR-REA\_IT reanalysis [97]. This dataset was obtained by dynamically rescaling ERA5 reanalysis ( $0.25^\circ$ ) to a horizontal resolution of  $0.02^\circ$  (approximately 2.2 km). The rescaling was obtained using COSMO-CLM, the regional COSMO model in CLimate Mode [98]. Given the higher resolution of VHR-REA\_IT compared to ERA5-Land, it is possible to generate valuable insights on past weather or climate conditions. However, ERA5-Land is still preferred for modeling tasks since it is more similar to HRES, considering they operate both at the same resolution and utilize similar underlying modeling processes.

## Hypermeteo

During the PhD program, some specific use cases raised the need and will to use higher resolution data. For this reason, Enogis started a partnership with HyperMeteo<sup>2</sup>, an Italian start-up (born from Radarmeteo<sup>3</sup>) that provides reanalysis and forecast data at  $0.01^\circ$  and  $0.04^\circ$  horizontal resolutions, and hourly or daily time resolutions. They assert their models are particularly effective because they integrate diverse external sources for reanalysis and forecast improvements. Given the superior load and the associated service cost, Hypermeteo data was not integrated for the whole Italian territory but only for commercially relevant areas, mainly consisting of Veneto, Friuli-Venezia Giulia, and Sardinia regions. Reanalysis (previous 5 days) and forecast (future 7 days) data is integrated daily via API. Differently from HRES, due to the relevance of the commercial project where this data is used, the integrated parameters are total precipitation and potential evapotranspiration ( $ET_0$ ).  $ET_0$  is computed following the modified Hargreaves–Samani 3 formula [99, 100, 101].

### 3.2.2 Satellite

In the context of agriculture, aerial images offer the opportunity to monitor crops and soils. Commonly available photography collection vehicles are unmanned aerial vehicles (UAVs), aircraft, and satellites [102]. Among the available methods, satellites offer low-cost images, continuous monitoring, and a historical archive. Drawbacks are the lower resolution of sensed data and the potential impact of cloud coverage.

---

<sup>2</sup><https://www.hypermeteo.com/>

<sup>3</sup><https://www.radarmeteo.com/>

#### **Sentinel-2**

The Copernicus program includes the Sentinel-2 space mission, consisting of two identical satellites (Sentinel-2 A and Sentinel-2 B) operated by the European Space Agency (ESA). These satellites are equipped with multi-spectral optical sensors for monitoring land and agriculture changes.

Sentinel-2 senses data at 13 spectral bands at different resolutions: RGB and Near Infra-Red (total of 4 bands) are sensed at 10 m, Visible and Near Infra-Red (VNIR, 4 bands) and Short Wave Infra-Red (SWIR, 2 bands) are sensed at 20 m. These bands are usually employed for agriculture and land applications, while the remaining 3 bands, sensed at 60m, are mainly focused on atmospheric correction and cloud detection. The revisit frequency of Sentinel-2 images is 5 days at the equator (10 days for each satellite) and 2-3 days at mid-latitudes, such as Italy. Sentinel-2 A and B satellites were launched in June 2015 and January 2017, respectively, and since December 2018, the mission systematically computes the atmospherically corrected Level-2A products (L2A).

#### **Planet**

Another satellite data provider is Planet<sup>4</sup>, a private company that operates a large constellation of *Dove* satellites. Planet data has a resolution of 3-4 m (depending on the specific instruments), senses a maximum of 8 visible and NIR bands, and has a daily revisiting time. The direct drawbacks compared to Sentinel-2 are the greater volume of data to handle (and the derived computational load) and acquisition costs.

### **3.2.3 Environmental descriptors**

An important ingredient for agricultural modeling is the description of the surrounding environment. Such description can be achieved from multiple perspectives, including elevation, exposure, soil content characterization, and potential solar irradiance.

#### **Digital Terrain Model**

To try to overcome the limitations posed by the weather data resolution, we integrated TINITALY/01, the Italian Digital Terrain Model (DTM) with horizontal resolution at 10 m (rescaled at 20 m for performance reasons) [103, 104].

The DTM is the main source for downscaling weather information, which can happen directly or indirectly. An example of direct use is the application of the dry adiabatic

---

<sup>4</sup><https://www.planet.com/>

gradient correction to temperature data, by which it is possible to obtain a more faithful representation of real temperatures. Contrarily, an example of indirect use is the combination of weather data with original resolution and the DTM as input sources to modeling algorithms. In this scenario, during the training process, the model has to figure out how to combine the two sources to maximize its performance. Intuitively, the model is left with approximating the dry adiabatic gradient correction, estimating the best parameters for each sub-area.

### **European Soil Hydraulic Database**

An additional source integrated into the data platform is the European Soil Hydraulic Database (EU-SoilHydroGrids ver1.0) [105]. It describes the hydraulic parameters of the soil for continental Europe at 250 m resolutions and for seven depth layers (0, 5, 15, 30, 60, 100, and 200 cm). Specifically, the parameters contained are water content (at saturation (THS), field capacity (FC), and wilting point (WP)), saturated hydraulic conductivity (KS), Mualem-van Genuchten parameters for the description of the moisture retention, and unsaturated hydraulic conductivity curves.

EU-SoilHydroGrids data covers continental Europe and is distributed in a tiled format; thus, all tiles belonging to the Italian territory were integrated into the data platform. An interesting aspect of EU-SoilHydroGrids is that it was derived with pedotransfer functions calibrated on soil information of 134 to 6,074 soil samples and validated on 57 to 2,357 samples, depending on the type of soil hydraulic property [106, 105].

The parameters that were ingested into the EnogisAI data platform are THS, FC, WP, and KS, important components for soil water content applications presented in the following chapters. The THS, FC, and WP conditions refer to the amount of water that can be stored in soil and used by crops. THS is associated with the soil water content following rainfall or irrigation. Since the soil is saturated, the water is free to move and percolates downward due to gravity (matric potential is 0). Matric potential indicates the pressure potential between water and soil particles. As water is drained, the matric potential increases up to the FC condition when the soil holds all the water contained. Water is then lost with increasing difficulty through soil evaporation and crop transpiration processes up to the WP, where the crop can no longer extract the available water and permanently wilts. Finally, KS describes how easily water can flow through the soil when saturated by water content.

#### Global Curve Number

Irrigation applications rely on accurate soil water balance models. To improve the rain processes simulation, we implemented a rainfall-runoff module based on the curve number (CN) method to enhance the rain processes simulation. CNs are derived from watersheds' hydrologic response potential and were initially developed by the US Department of Agriculture [107, 108, 109].

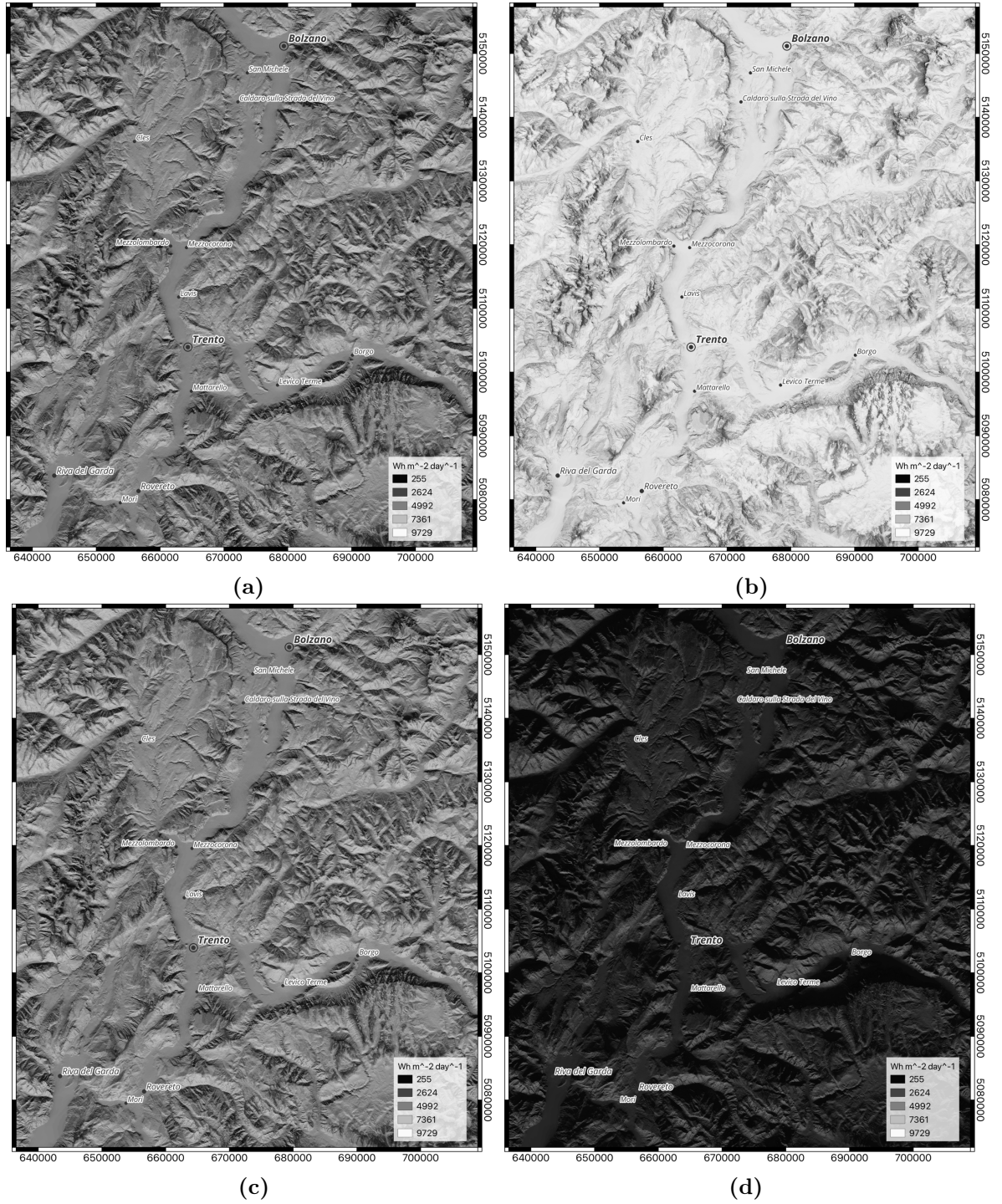
The data product integrated into the data platform is GCN250, a 250 m global gridded curve numbers dataset [110]. GCN250 is generated using (i) a satellite-derived land cover dataset (ESA CCI-LC 2015)[111], (ii) the Hydrologic Soil Groups at 250 m dataset (HYSOGs250m) representing soil runoff potential, and (iii) plant functional types [112], a land-cover dataset employed for Earth System Model that groups similar plants based on phenology or other characteristics. Since CNs depend on the antecedent runoff conditions (ARC), GCN250 consists of three different maps (dry, average, and wet) to be used in combination with observed or modeled precipitation data following the classification of [113].

## 3.3 Derived products

In addition to internal and external data, derived products were calculated and integrated into the Enogis workflows. Derived products have the potential to describe the territory more precisely, starting from one or simpler products. Here, we consider two derived product types later used in specific Enogis projects (described in the following chapters): first, solar potential irradiation maps, and second, climatic indices.

### 3.3.1 Solar irradiation maps

Solar potential irradiation maps are daily products used to indicate direct irradiation for specific latitudes and atmospheric conditions, considering the shadowing effect of the surrounding terrain. Solar radiation maps were computed using the GRASS GIS software (version 7.8.4) [114]. The computation is based on elevation, aspect, slope, and horizon maps, with aspect and slope maps being computed starting from the DTM (TINITALY/01 [103, 104]) on the whole Italian territory. Horizon and solar maps are instead computed for 665 grid chunks of side 25km to limit the memory requirements of the processes, with a considered context of an additional radius of 25km (to improve the resulting quality of points on the border of the 25 km grid cell). The Copernicus Global Digital Terrain Model at 90m resolution (GLO-90 [115]) was used to extend the



**Figure 3.3:** Resulting sun irradiation maps around the Adige valley, Trentino-South Tyrol region, North Italy, computed for Mar 21 (a), Jun 21 (b), Sep 21 (c), and Dec 21 (d). Maps were computed for all day of year and the entire Italian territory using the *r.sun* GRASS module.

TINITALY/01 DTM out of the Italian political border and improve the accuracy of the horizon and solar maps such areas.

Horizon maps describe the horizon height (in radians) of specific directions, which is needed to consider the shadowing effect. A total of 24 maps were computed with the *r.horizon* GRASS module, one each 15 degrees.

Sun maps instead describe the direct solar irradiation, expressed as  $\text{Wh} \cdot \text{m}^{-2} \cdot \text{day}^{-1}$ , and are generated with the *r.sun* GRASS module [116]. The *r.sun* module accepts the Linke atmosphere turbidity factor to account for the differences in terms of atmospheric conditions among mountainous, rural, urban, and industrial areas. The value is set to 3, an average between rural and urban areas. The final output is 365 maps for each of the 665 chunks at 20m horizontal resolution that can be used to reconstruct sun irradiation profiles of georeferenced points or whole fields. The four maps of the day of the years that mark the beginning of the four seasons are presented (Figure 3.3). The focus is given on the Trentino-South Tyrol region, a geographically complex area in the North Italy Alps. The maps highlight not only the expected difference between different periods of the year but also the magnitude of the shadowing effect associated with different sun incidence angles (note the changing amount of darker areas in the bottom right corner, close to Levico Terme and Borgo). The *r.sun* maps are less variable in more flat areas (e.g., Pianura Padana) but are useful to capture sun irradiation profiles for hilly and mountainous areas. As a technical note, starting from a DTM of 20m (compressed size of 1.6 GB), the total disk space required by all (compressed) chunk maps is 873 GB, with the total processing taking around 3 weeks (using 64 GB of RAM and 20 CPUs).

A crucial aspect is that these maps are static in the sense that they do not include real weather conditions. As such, they indicate potential direct irradiation and not real irradiation. This concept is similar to sun exposure, but it takes into account the obstructing and shading effects of the surrounding terrain.

The generated product is used to provide geostatistical insights in field description dashboards and as input data to improve environmental modeling in later application stages.

#### 3.3.2 Climatic and bioclimatic indices

Climatic and bioclimatic indices are the other derived products relevant to this thesis work. They are used to describe the suitability of the environment for grapevine cultivation, both in general and for specific cultivars.



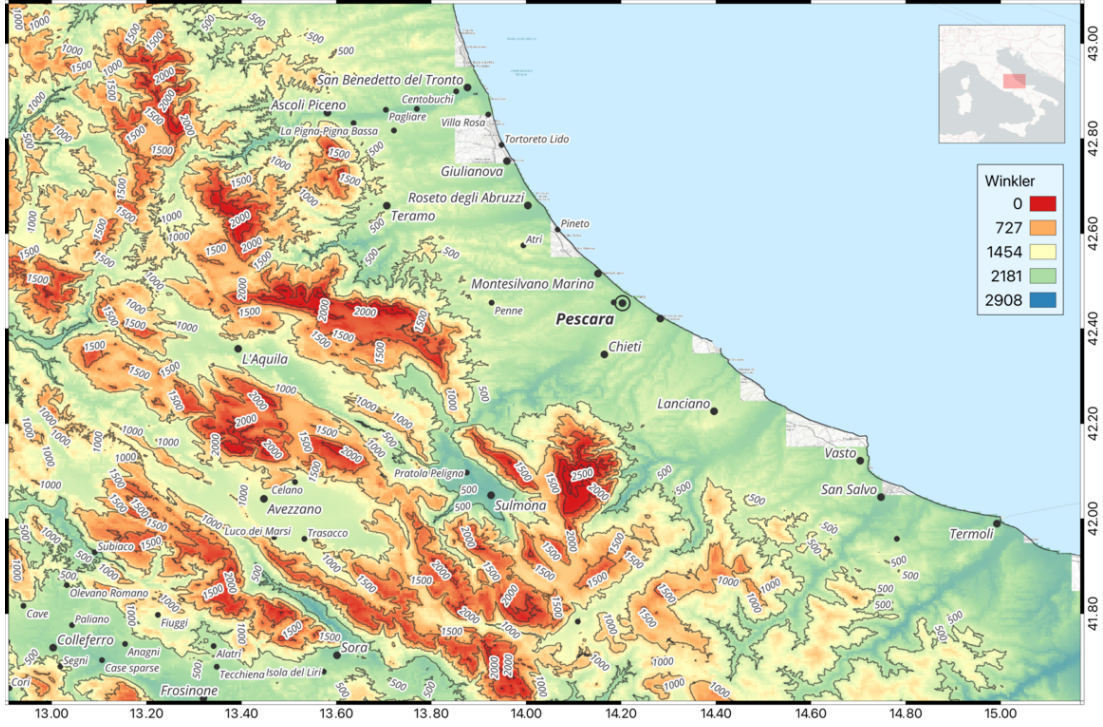
The role of such indices in viticulture is central: they can provide insights on (i) what areas are suitable for cultivation as of now, (ii) how cultivated regions are developing, and (iii) what new areas are becoming suitable for viticulture [117, 118].

Climatic indices are computed on multiple years of data (at least 5) to account for the shifting conditions and avoid focusing on specific years' behavior. The considered indices in EnogisAI are:

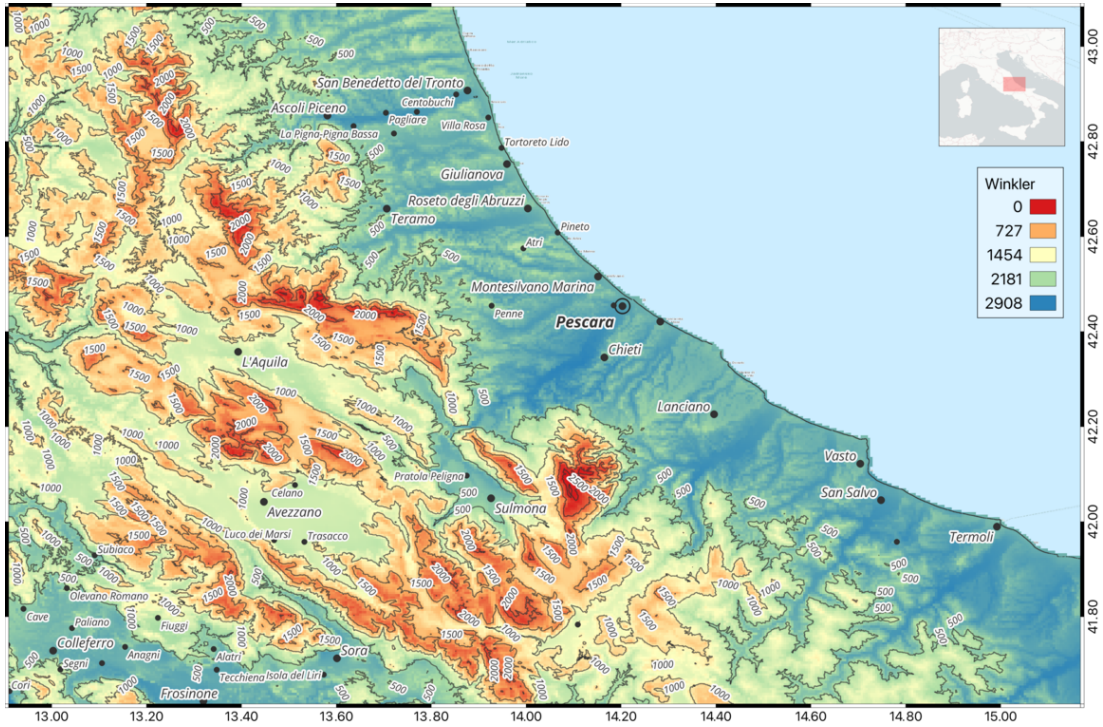
1. Cumulative precipitation: calculates the average total amount of yearly precipitation, which is crucial for understanding water availability for grapevine cultivation.
2. Winkler index [119]: calculates growing degree days (GDDs) above a base temperature of 10°C from April to October and provides insights into what cultivars are preferred based on the results and their ripening period (early vs. late). The Winkler index is calculated using the formula  $\sum_{i=Apr1}^{Oct31} (TG_i - 10 | TG_i > 10)$ , with  $TG_i$  being the mean daily temperature (in °C).
3. Huglin heliothermal index [120]: similar to the Winkler index, but uses the average between maximum and average temperatures to estimate the daytime temperature (when plant photosynthesis occurs), together with an estimation of the day length. It is defined only for latitudes between 40°N and 50°N. The Huglin index is calculated using the formula  $\sum_{i=Apr1}^{Sep30} ((TX_i + TG_i/2 - 10) * k)$ , where  $TG_i$  and  $TX_i$  represent the mean and maximum daily temperatures (in °C), and  $k$  is the day length estimation as a function of latitude, computed as  $1 + (|lat| - 40)/10 * 0.06$ .
4. Gladstone index (biologically effective degree-day) [121]: another GDD variant that incorporates temperature range adjustment (restricted to 10°C to 19°C), a maximum of daily cumulated growing degree days, and latitude correction. The Gladstone is computed as  $\sum_{i=Apr1}^{Oct31} \min((\max((TX_i + TN_i)/2 - 10, 0) * k) + TR_{adj}, \text{degdays}_{max})$ , where  $TX_i$  and  $TN_i$  are the maximum and minimum daily temperatures,  $k$  is the day length estimation computed as  $1 + |lat|/50 * 0.06$ , and  $TR_{adj}$  is the range adjustments that boost small and large temperature swings (for brevity, readers are referred to the original paper).
5. De Martonne aridity index [122]: it helps in evaluating the aridity of a region. It is defined as the ratio of total annual precipitation to the sum of the annual mean temperature and a base temperature of 10°C. The index is calculated as  $\frac{P}{T+10}$ , where  $P$  represents the annual precipitation (in mm) and  $T$  represents the mean annual temperature (in °C).

Climatic indices are computed on the ERA5-Land and VHR-REA\_IT datasets,

### 3.3. Derived products



(a)



(b)

**Figure 3.4:** Winkler index comparison between ERA5-Land (horizontal resolution: 9 km) (a) and VHR-REA\_IT (horizontal resolution: 2.2 km) (b) on the Abruzzo region. The Winkler index is upscaled to 500 m, applying a dry adiabatic gradient to the input temperature data. VHR-REA\_IT expresses better the extreme values of the index, especially along the coastline. The red box in the corner map indicates the region where the Winkler index was computed with VHR-REA\_IT data.

considering 10 (2012 to 2021) and 23 (2001-2022) years of data, respectively. Maps are generated using the *xclim* scientific Python package that provides algorithmic solutions for computing climate indices [123]. Maps are generated with 500 m resolution by applying a dry adiabatic gradient to the input temperature data. We computed a bilinear interpolation of the ERA5-Land and VHR-REA\_IT DTMs and applied the gradient based on the altitude difference between the original and calculated DTMs with a lapse rate of  $-9.76\text{ }^{\circ}\text{C}/\text{km}$ . The interpolation to 500 m resolution was not applied to precipitation records.

The results obtained with the two datasets were compared, focusing on the Abruzzo region (Figure 3.4). The use of VHR-REA\_IT data improves the results in two aspects. First, there is a higher variability and degree of details in the latter map, and second, there is better coverage of the coastline compared to ERA5-Land (geometrical flaws can be appreciated on the top and bottom sections of the coastline). The ERA5-Land (compared to ERA5) does not provide complete geographical coverage of global land but only for those grid cells that comprise at least half a percentage of land.

The increased resolution data offers higher environmental modeling capabilities at a higher computational resource cost. The original resolution of ERA5-Land data is 9 km, while for VHR-REA\_IT, it is 2.2 km, resulting in a 16x scaling factor (in both time and space required for the computation).

### 3.4 Data platform technical aspects

The data sources discussed exhibit varying characteristics regarding update rate, temporal frequency, and geographical scope (point vs. area), among others. Intrinsic differences among the data types are not ideal for a unique standardized way of storing and retrieving data in a distributed manner. In addition, the technological context where data is used adds to the needs and requirements for such a hypothetical system. In fact, depending on the developed application type and the programming and abstraction levels considered, different technologies could be available and preferable. Examples are direct database connections, API/web clients, and shared disks/locations access. As a result, the EnogisAI data platform consists of a centralized catalog of (i) the integrated and readily available data, (ii) the description of the sources they belong to, (iii) a code repository with documentation and examples on how to access the data and how to download additional records, and (iv) a description of the Enogis internal services that are using the data. The advantages of this data collection and description catalog are twofold: it provides both technical and business benefits. The data platform

became the central knowledge base of all internal and external data used in the Enogis AI-related efforts. As such, it allows the extension of client-specific databases with additional context and the introduction of new analyses for clients in a seamless way. Nonetheless, it helps in writing industrial and research project proposals as it offers an initial presentation of Enogis’s capabilities and how they could be extended.

In addition to the catalog, more specific software architectures and standards were adopted for the particular data sources and application use cases. These include the use of VMs with dedicated hardware (e.g., GPUs), software requirements (e.g., specific database management systems (DBMS)), and integration with monitoring services (e.g., Sentry).

#### 3.4.1 Timescale database for weather time series

A core component of the Enogis usually adopts PostgreSQL as a relational DBMS (RDBMS), with the PostGIS extension to handle geographic relational data records. Focusing on time series data, RDBMSs tend to provide limited access performance considering their common architectural design patterns and the nature of time series compared to relational records. RDBMSs are optimized to organize, archive, and retrieve relational data, impacting how data is saved to storage disks. Contrarily, time series have different characteristics: there are frequent insert operations (compared to update ones in RDBMS), data relations are limited to a few tables (compared to wide database structures), and they are usually queried (and eventually aggregated) by time period (and not by ID).

To leverage the use of the existing PostgreSQL clusters used to store Enogis data, we adopted TimescaleDB [124], a PostgreSQL extension for time series data. Timescale impacts the underlying storage system (i) partitioning data across time and other meta-data and compressing data for efficient storage. Timescale introduces the concept of *hypertable*, a virtual view of regular PostgreSQL tables called chunks, where data is stored. Chunks can be compressed with different algorithms (also depending on their data format) [125] to save space and improve retrieving performance.

To deal with the weather data sources, we adopted TimescaleDB for the time series data. In particular, TimescaleDB is used for two weather and sensor data projects. First, the *Measurements* project acts as the main central hub for Italian weather. It collects data for 557 weather stations distributed over the Italian territory, plus weather reanalysis and forecast of Italy from ECMWF, as well as Veneto, Friuli-Venezia Giulia, and Sardinia regions from Hypermeteo. In total, its archive contains 973 million data

points (considering the period from 2020 to 2024 included), while the current ingestion rate is about 1.3 million data points per day. Second, the *DataCollector* project collects ECMWF data for 4 Mediterranean countries (Spain, France, Italy, and Greece), plus weather and other monitored variables collected from 1636 intelligent devices (weather stations or sensors). In total, its archive contains 1.1 billion data points (considering the period from 2020 to 2024 included), while the current ingestion rate is about 2 million data points per day.

Compression is a crucial feature of TimescaleDB, especially for weather data. One of the available algorithms is the XOR compression [125], which aims to lower the information needed to represent an object by storing only the XOR’ed difference from the previous one. This approach proves extremely effective for those data that vary slightly over time (e.g., temperatures) or are mostly sparse variables (e.g., precipitation). In fact, PostgreSQL uses the double precision data type (8 bytes) to store floating point numbers (time series values) in our setting. Taking into account precipitation, ca. 84% of data is 0. For each window of continuous 0 values, ideally, we would need 0 bits to represent the value all but the first values.

Another interesting aspect of TimescaleDB is the performance boost it can provide. As data is divided into chunks, TimescaleDB can use multi-threading to read and process data faster. One key factor to consider is the magnitude of performance gain, considering the contrary effects of compression and multi-threading.

The *Measurements* and *DataCollector* projects were initially developed and deployed using standard PostgreSQL instances. A migration to TimescaleDB of the two projects was planned after 3 years of operative data collection and serving due to the amount of required storage and the slow query execution speed. Before migrating the two projects to the new database, experiments were conducted to assess the effects of adopting the latest technology regarding querying performance and storage use. In the following, we depict the methodology used in these tests and the obtained results.

### **TimescaleDB performance experimental design**

The tests were conducted on two datasets, each sampled from two distinct projects. The first dataset, referred to as the *wide* dataset, contains 4 months of hourly data (from September 2022 to January 2023) across Italy, Spain, France, and Greece, corresponding to 18,076 stations, resulting in a total of ca. 52 million observations. The second dataset, referred to as the *long* dataset, contains 1 year (from March 2022 to March 2023) of hourly data across Italy, corresponding to 3,112 stations, resulting in

27 million observations. The two datasets reflect the two different characteristics of the deployed systems. *Measurements* archived 4 years of observations for the Italian territory. *DataCollector* instead ingests data relative to four Mediterranean countries for a more limited period.

The tests were run on a VM dedicated to AI development and experimentation, providing 20 CPU and 64 GB of RAM. The hardware specifications ensured a responsive testing environment, although they did not match the production environments. Consequently, we run the experiments on three setups, each resembling (i) the existing software architecture, (ii) a TimescaleDB instance, and (iii) a TimescaleDB instance with data compression enabled.

The tests are run for required storage space and querying time at two different levels. First, at the database level, to observe the direct effect of the technological shift, and second, at the application level, to compare the performance changes when migrating the data access with the software framework that automatically translates higher-level API queries into SQL queries. In this specific case, the application layer is developed in Django, a web framework written in Python that follows the Model View Controller (MVC) paradigm. The most relevant advantage of Django (and more in general of MVCs) in this context is that it allows access to database data directly in Python without needing to deal with SQL. Each benchmark is executed two consecutive times. The first time, the experiment is run after database initialization. Thus, the execution is expected to be slower as the database has no cached data compared to the second run, where the database may have cached data. These two setups are referenced as cold and warm scenarios.

Finally, all testing environments were constructed using Docker container technology, which ensured software isolation and enhanced replicability across different setups.

The benchmarks consist of 3 queries run at both the database and application layer and two additional application layer queries corresponding to application logic employed when serving data via API to external systems. Benchmarks are designed to work either vertically against a single series of data (or all the series of a station) or horizontally aggregating multiple series data, aiming at reproducing real system usage scenarios. A series is defined as all a station's data for a single monitored variable (e.g., air temperature). In particular, the benchmarks measure the time needed to:

1. Count all the values for each year;
2. Count all data of a series;
3. Aggregate all data of a single series into count, min, max, and average values;

4. Perform daily aggregation of all series of a station (similar to API usage);
5. Retrieve all station data without performing aggregation (similar to API usage).

Benchmarks 1 and 2 assess the differences between the setups when executing a full scan. With TimescaleDB, this operation is expected to be executed in parallel. Benchmark 3 instead operates on a single series of data. According to TimescaleDB design principles, this operation should be restricted to only a limited number of chunks (partitioned tables) and then be faster than the PostgreSQL version. Benchmarks 4-5 involve the Django framework and are intended to preventively identify any potential issue in querying data after the migration. Finally, benchmarks 1-3 are run at both levels, while benchmarks 4-5 are run only at the application level.

### **TimescaleDB performance experimental results**

On average, TimescaleDB (tsdb) without compression improves performance on the cold scenario by 10% but decreases it by 25% compared to PostgreSQL (postgres, Figure 3.5). It tends to perform better on the *wide* dataset, although marginally (or not at all) when querying data of a single series. On the *long* dataset, TimescaleDB is faster for benchmarks involving the whole database, while it is slower when dealing with single station data. The reason may be related to the multiprocessing capability and the division of the database table into chunks. First, TimescaleDB can read disk data and process them in parallel, an advantage that tends to disappear when the number of data, and thus of chunks, is low. In that case, PostgreSQL still performs better.

The best performance time was systematically obtained by TimescaleDB with compression enabled (tsdb-c), faster by 40% considering the warm and cold scenarios. The rationale behind this result has to be researched in terms of the amount of data to be read from the disk. In the uncompressed case, the system directly reads all (needed) compressed data. In the compressed case, instead, the same data has to be read but with a fraction of the actual bytes while still having to decompress at the database level the resulting payload. Therefore, the benchmark results suggest that reading more bytes is slower (due to disk speed) than reading fewer bytes and decompressing them (thanks to CPUs and algorithm efficiency). The observed results remain valid under production conditions, as the VMs utilize the same disk and CPU architectures.

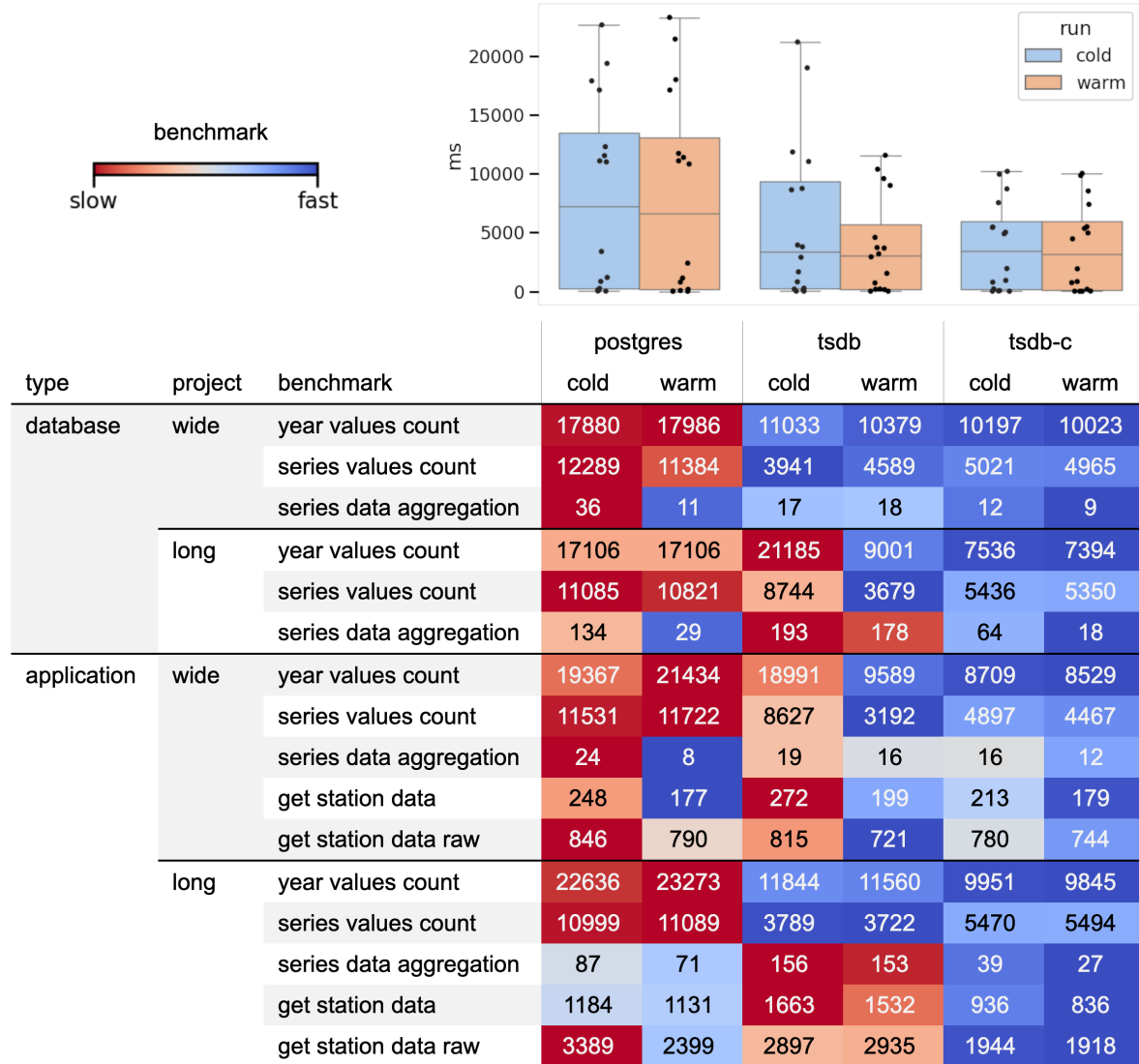
Benchmarks were executed against cold and warm environments: all three setups show a caching behavior with different impacts depending on the executed query (Figure 3.5).

The other dimension the systems were compared to is the required storage space.



### 3.4. Data platform technical aspects

The migration from PostgreSQL to TimescaleDB did not highlight any benefit in terms of storage space. In particular, it showed a 1.09x compression factor for the *wide* dataset and a 1.06x one for the *long* one. When enabling compression, however, TimescaleDB



**Figure 3.5:** Timescale performance test results, expressed in milliseconds. Faster benchmarks are blue, while slower ones are red. Timescale (tsdb) improves performance for 10 out of 16 benchmarks (10.24% faster on average in the cold environment), while Timescale with compression (tsdb-c) is systematically better compared to PostgreSQL (postgres) (45% faster on average in the cold environment and 35% faster on the warm one). Each row represents a benchmark experiment at a different level (database or application access interface), a different dataset (wide or long), run on all three setups (PostgreSQL, TimescaleDB, and TimescaleDB with compression) two consecutive times (cold and warm). The color scale is normalized per benchmark (row). Boxplots (top) summarize the results of each setup for both cold and warm scenarios.



obtained a compression factor of 5.19x on the *wide* dataset and 7.18 on the *long* one. The difference is probably to be researched in the number of resulting chunks (and thus overhead) due to the larger number of stations.

### **Weather and sensors time series data integration and monitoring**

The TimescaleDB database instances are populated with continuously monitored weather and sensor data. The different sources have been integrated by developing source-specific data connectors run at fixed intervals. Most modules implement a pull logic, where we can download data calling external APIs, with a few exceptions that use a push policy via File Transfer Protocol (FTP). Excluding custom data sources that were collaboratively designed and built with other companies, at Enogis, I developed the integration for the following public or commercial services: Pessl instruments, Sencrop, Agricultural Support, Altoadige, Cesens, Farmsense, Hypermeteo, Netsens, 3acloud, Polaris, and The Things Network. The ingestion data flows are scheduled to be executed at different intervals depending on the associated data sampling frequency, API limits, and other characteristics. To develop and manage the flows, I created an internal Django module to ingest data and serve the collected data to other internal or external services. The application uses Celery and Celery Beat [126] to schedule and run the flows and Sentry to monitor their execution status.



# Chapter 4

## Agricultural task, data, methodologies, and algorithms

### Highlights

- We developed AI models that can contribute to agricultural management at various stages, e.g., crop development, crop management, and yield prediction.
- The EnogisAI data platform supports the development of new computational pipelines by simplifying the integration of data products for methodology testing.
- Machine and deep learning techniques were selected for their flexibility in combining diverse data sources. We addressed the complexity of time series employing unsupervised approaches, such as Functional PCA and K-means, and supervised ones, such as Gradient Boosting, Partial Least Squares, and Deep Neural Networks.
- EnogisAI serves as a modeling environment supporting the creation of reusable and extendable model components. The proposed methodologies are built on top of the data engineer aspects of the framework, and operational models benefit from this standardized building environment.
- Four of the presented models resulted in conference proceedings or research articles. These are (i) the durum wheat field management zone and (ii) crop yield pipelines, and (iii) the daily pollen concentration reconstruction and (iv) forecast.

---

**Personal contribution:** I actively contributed to the development activities of all models presented in the chapter. In particular, I contributed to the development of the field management zone pipeline. I developed the remote sensing crop development model and supervised the work of a junior scientific programmer for the in-field image classification. I developed the machine learning pipeline for crop yield prediction. I adapted the soil water balance model to work with the single crop coefficient approach. I contributed to the design, training, and evaluation of machine and deep learning models for pollen research activities. During the research, I reviewed the literature, prepared data, developed the code, and performed the experiments. I contributed to the writing and reviewing of the four published articles.

This chapter outlines the specific tasks and objectives, focusing on the chosen algorithmic methodologies, experimental designs, and task-specific details of the data products used to develop and validate the algorithms. Our ultimate goal is to provide users with agronomic DSSs that continuously monitor, predict, and recommend the most appropriate actions. To reach this objective, we integrated the data described in the EnogisAI data platform (Chapter 3), developing a number of machine and deep learning techniques and mechanistic models. Specifically, during the PhD program, we addressed several key challenges, including field management zones identification, crop development monitoring, soil water balance assessment, and yield prediction. The range of solutions developed during the program highlights the potential of AI to deliver significant benefits to multiple aspects of agricultural management. AI can be applied at the different phases of the field operations, as well as during subsequent food processing and delivery activities.

The EnogisAI framework is designed with two key objectives: (i) to enable the reuse of developed modules across the national territory and (ii) to adapt the modules to different input data, crops, or contexts. Currently, the work in this phase follows a research and development approach that includes literature review, experimentation, and validation. In addition to work focused on the agricultural aspects relevant to Enogis, the approach was expanded to include two international pollen-related research initiatives. While these projects share minimal overlap with the current Enogis products and data, they provided an ideal testing ground for building machine learning algorithms focused on time series reconstruction and trend forecasting. Moreover, the efforts devoted to these two works are well connected to the intelligent agriculture field. First, time se-

ries reconstruction and forecasts are topics of high interest in the field, considering the frequency at which data corruption occurs. Scenarios include malfunctioning weather stations and the impossibility of sensing data with satellite instruments (e.g., cloud coverage). Second, while studying pollen dynamics may appear relevant only for public health, the same dynamics are also connected to agricultural processes. Thus, it would be worth experimenting with using pollen-derived signals and insights as input variables for other models and assessing the contribution to the final performance.

The goal is to transform any research findings into products or services that can be integrated within the broader Enogis ecosystem, adding significant value. For this reason, we consider EnogisAI as a modeling environment where all components follow consistent best practices and standard procedures concerning low-level implementation and design decisions. As a result, each workflow or analysis pipeline can be easily customized to incorporate additional data sources or adjust modeling choices, such as hyperparameters, training algorithms, and model selections.

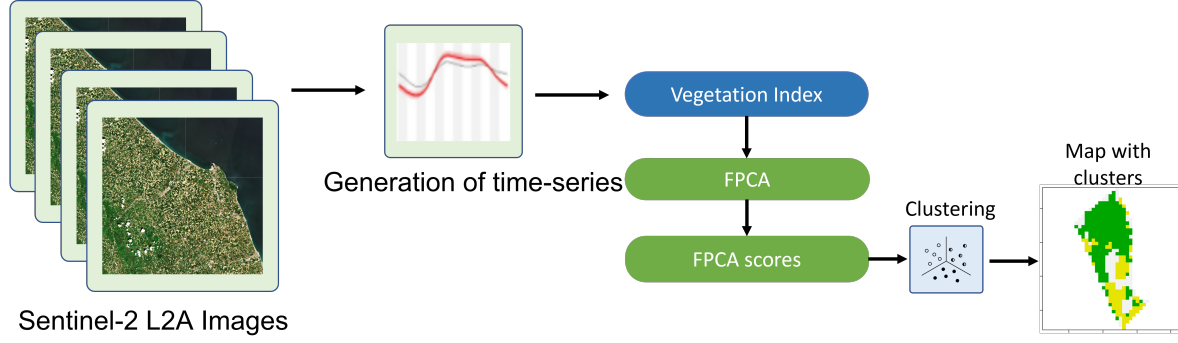
Each section describes one of these challenges and depicts the data and algorithms used.

## 4.1 Field management zone identification

Management zone (MZ) identification refers to the ability to cluster field areas that behave homogeneously during the agricultural season, whether the focus is on crop development, quality, soil nutrients, or other. The pipeline designed for this task operates entirely unsupervised, requiring minimal user inputs at the operational time (Figure 4.1). This approach eliminates the need to collect data (whether historical or live) from the selected fields. At the same time, users can interact with the system by changing simple hyper-parameters to extract valuable insights. The work presented next was published as a conference proceeding in the 2024 IEEE International Workshop on Metrology for Agriculture and Forestry (MetroAgriFor2022) [127]. I performed data cleaning and pre-processing and developed the FPCA+K-means pipeline. The methodology was built in collaboration with APRA and *Università Politecnica delle Marche* within the BioCereals 4.0 project (project financed by Regione Marche in the context of *Programmi di sviluppo rurale regionali approvati 2014-2020*; duration 4 years).

The study area for MZ identification comprises 12 organic durum wheat fields (Fig. 4.2), resulting in a total area of 25.7 ha (median: 1.31 ha, min: 0.27 ha, max: 7.26 ha). MZs for each field were identified by applying Functional Principal Components Analysis (FPCA) to satellite-derived indices.

#### 4.1. Field management zone identification



**Figure 4.1:** Data processing pipeline for the field management zone (MZ) identification task. Sentinel-2 satellite data are organized into one time series for each corresponding ground sampling pixel. Time series of NDVI and NDRE vegetation indices are decomposed with Functional PCA. The resulting FPCA scores are then clustered with K-means to identify homogeneous MZ. Finally, these are then displayed on a map.

The indices considered were the Normalized Difference Vegetation Index (NDVI) and the Normalized Difference Red Edge (NDRE). These are computed as:

$$\text{NDVI} = \frac{\text{NIR} - R}{\text{NIR} + R} \quad (4.1)$$

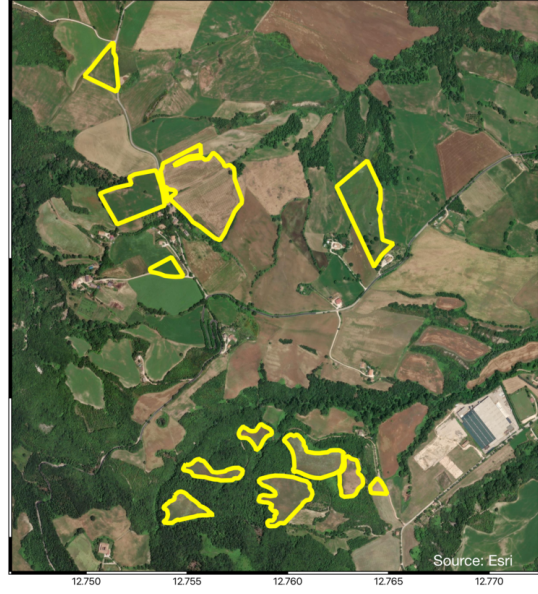
$$\text{NDRE} = \frac{\text{NIR} - \text{RE}}{\text{NIR} + \text{RE}} \quad (4.2)$$

where *NIR* is the near infra-red band, *R* is the red band, and *RE* is red edge band. The indices were computed on Sentinel-2 data (Table 4.1) and had 10 m resolution except for the red edge, which had 20 m spatial resolution. The red edge (B5) band was resampled at 10 m to obtain NDVI and NDRE outputs at 10 m resolution.

NDVI is an index used to estimate and monitor biomass levels and overall plant health, especially in early growth stages [128, 129]. NDRE is a similar index that uses the red edge band instead of the red one, which can better penetrate through leaves

**Table 4.1:** Sentinel-2 characteristics of the spectral bands used for the NDVI and NDRE index calculation. The two indices were used to construct the pipeline’s input time series. The Red Edge band was resampled with linear interpolation to 10m.

| Band Name      | Band Number | Wavelength | Spatial Resolution |
|----------------|-------------|------------|--------------------|
| Red            | B4          | 665 nm     | 10 m               |
| Red Edge       | B5          | 705 nm     | 20 m               |
| Near Infra-red | B8          | 842 nm     | 10 m               |



**Figure 4.2:** The MZ identification pipeline was developed considering 12 organic durum wheat fields in the Marche region, Central Italy. The fields were included in the BioCereals 4.0 innovation project. Total considered area: 25.7 ha (median: 1.31 ha, min: 0.27 ha, max: 7.26 ha).

and becomes more useful in later development stages [130].

Field surfaces are split into grid cells of side 10 m following the Sentinel-2 spatial reference system, and a time series was extracted for each point, obtaining 2566 grid cells. Satellite images were collected from October 2019 to August 2020, resulting in 52 images with a 2-3 days revisit time. Vegetation indices time series were constructed for each grid cell of each field. For each field, data of specific visiting dates were retained if the cloud coverage was below 30%, ensuring the accuracy and reliability of the vegetation assessments.

NDVI and NDRE time series are input to FPCA, an extension of Principal Component Analysis to functional data (e.g., time series) [131]. Similarly to PCA, it decomposes the functional data into principal components (continuous functions), allowing for a better data variance study. FPCA aims to find a limited number of principal components that can capture a large portion of data variability. The principal components are called eigenfunctions and indicate the directions with the most variation in the function space. Eigenfunctions are associated with eigenvalues, a measure of explained variability. Additionally, FPCA projects each grid cell time series to a low-dimensional (corresponding to the number of components) representation of the data. The coordinates are also called scores and depict the structure of each time series given the extracted eigenfunction and eigenvalues. Continuing the similarity with PCA, one

is left with deciding the amount of principal components to consider based on the explained variance. In this study, we selected the number of components that explained at least 90% variability.

The last step in the MZ identification pipeline is clustering the resulting FPCA-reconstructed time series. In this case, FPCA can be seen as a preprocessing step that yields a set of scores for each time series. The FPCA scores were clustered using K-means, a top-down clustering algorithm that requires the prior selection of the desired number of resulting classes. This parameter typically depends on the specific application and input dataset. It allows agronomists and farmers to experiment with the pipeline to identify the number that better fits the scenario and the field activities. The underlying idea is that the identified management zone can exhibit early or late growth patterns and varying potentials for quantity or quality. By employing and experimenting with K-means clustering, users can discover the inherent groupings within the data, facilitating a better understanding of the agricultural dynamics.

The pipeline was developed using the R programming language, version 4.0.3.

## 4.2 Grapevine development monitoring

Monitoring crop development is a key agricultural activity of extreme importance to growers for two main reasons. First, the efficiency of many in-field activities, such as fertilization, irrigation, and pesticide application, heavily depends on crop growth processes. Second, the ripening phase impacts the final yield quality and quantity.

Crop development resulting from climatic inputs is expressed through phenology, [20]. The monitoring and forecasting of grapevine phenology can help growers plan all in-field activities more effectively. This section describes two approaches based on deep networks to monitor and forecast phenology, especially applied to grapevines. The first is a remote sensing-based approach, utilizing in-field data solely as model targets for the training and validation phases. The second approach relies on in-field images collected manually by agronomists or farmers or automatically through installed cameras. Manual photo collection requires visiting the field more often, while camera installation can provide a more systematic and continuous data collection at higher hardware and maintenance costs.

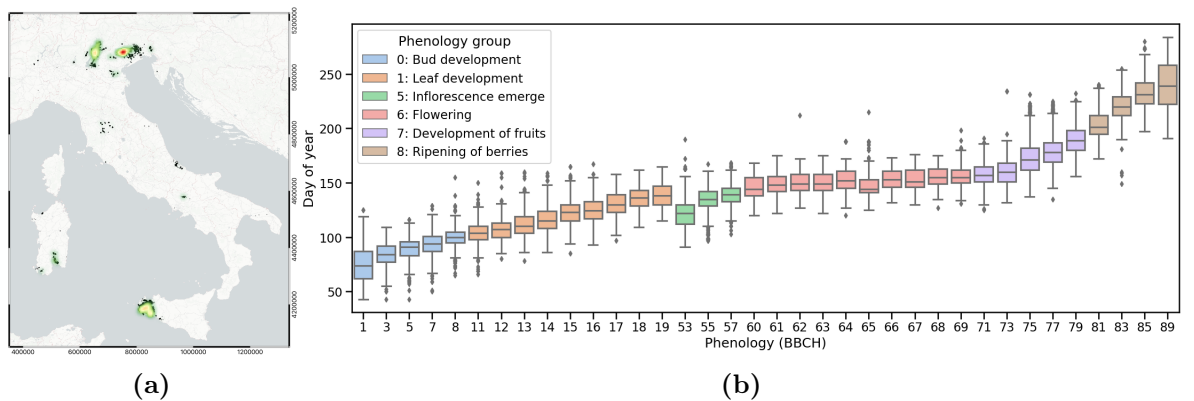


### 4.2.1 Remote sensing neural network for phenology prediction

The remote sensing phenology prediction task was developed considering 44941 phenology samples collected in the agronomic seasons from 2012 to 2020 (9 years of data, with 99.5% of data collected starting from 2015).

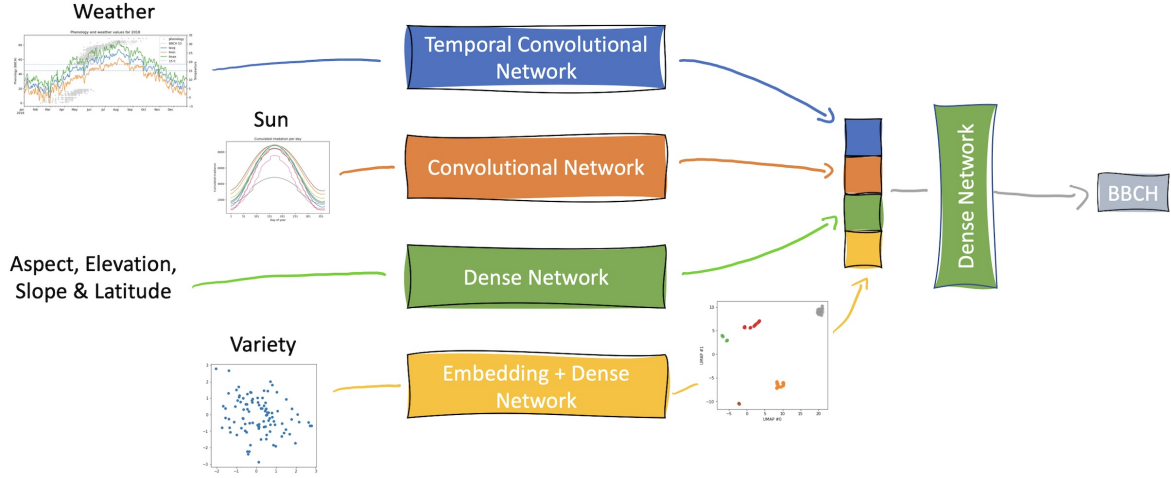
Samples were collected over different areas of the Italian territory, with prevalence in three areas: Trentino-South Tyrol, Veneto and Friuli-Venezia Giulia, and Sicily, each corresponding to different environmental and climatic conditions (Fig. 4.3a). Each sample is geo-referenced and is associated with a specific vineyard. For each vineyard, multiple measurements may be collected during the agronomic season. Each sample consists of a BBCH value belonging to all BBCH grapevine phenological stages, with the exception of the final stage, senescence, where no data are collected (due to low agronomic interest). Considering that a vineyard may exhibit a range of phenological conditions (although similar), the BBCH is expected to reflect the average state of the vineyard. Critical growth stages to monitor for grapevines are the emergence of inflorescence (principal growth stage 5) and the development of fruits (principal growth stage 7). Each field present in the phenological dataset was augmented with additional data coming from other sources in the EnogisAI data platform, in particular weather, sun irradiation profiles, and DTM-related variables (aspect, elevation, and slope).

We pose the task of phenology prediction as a regression problem. Consequently, we developed a multi-modal deep learning architecture capable of processing multiple input sources and output the estimated phenology. The model is composed of specific deep network branches for weather and sun data, environmental variables (elevation,



**Figure 4.3:** Phenology data statistics for 44941 phenology samples collected from 2012 to 2020. Map of the observed phenologies with heatmaps showing the density of the three most sampled areas (Trentino-South Tyrol, Veneto and Friuli-Venezia Giulia, and Sicily) (a). Box-plot distributions of each collected phenology phase (b).

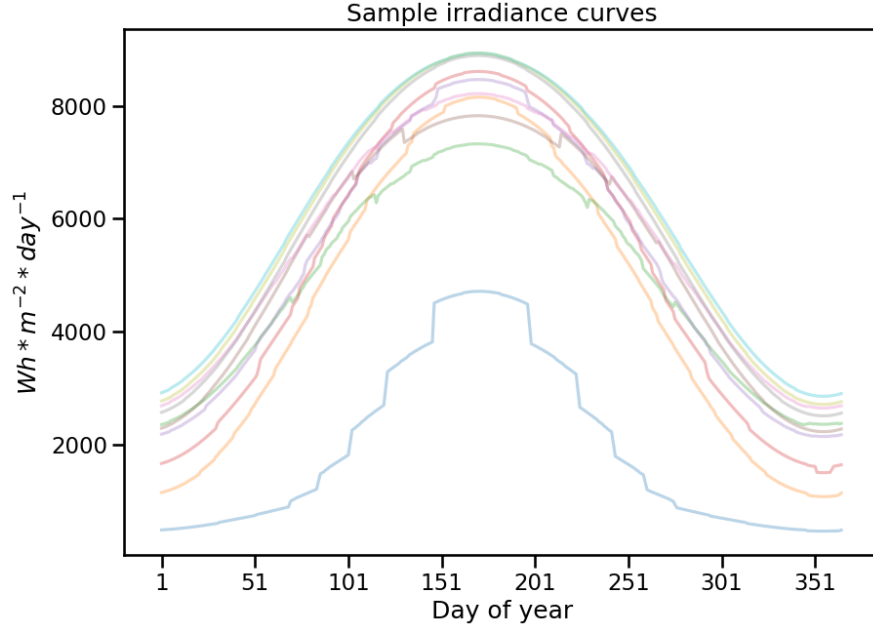
## 4.2. Grapevine development monitoring



**Figure 4.4:** Multi-modal deep learning architecture of the remote sensing phenology model composed of four branches for representation learning and a final sub-network for latent data fusion and phenology prediction. Temperature and precipitation data are processed by a Temporal Convolutional Network (blue). Annual solar irradiance profiles are processed with a convolutional neural network (orange). Tabular characteristics such as aspect, elevation, slope, and latitude are modeled by a dense network (green). Finally, the cultivar categorical variable is associated with learnable embedding vectors (yellow).

aspect, slope, and latitude), and cultivars. The output of the model is a phenology value in the BBCH scale (Fig. 4.4). Since the grapevine-specific BBCH scale presents missing values in its numerical encoding, we mapped the values to an ordinal and contiguous scale used for training and evaluating the model.

Weather data is processed with a Temporal Convolutional Network (TCN). TCNs have been proven to converge in fewer epochs compared to recurrent networks (e.g., LSTMs) while requiring a smaller number of parameters to train [132]. In addition, they tend to work better with longer input time series, as is the case of daily measures since the beginning of the year (compared to sentences usually made of 15-20 words). TCNs are based on dilated and causal convolutions, where the convolution operation considers only past observations, and its receptive field grows exponentially with the number of layers. This allows the model to generate a prediction covering all input time steps with a number of layers proportional to the square root of the input sequence length. The input weather data consists of four daily features extracted from ERA5-Land: daily average, minimum, and maximum temperatures, daily total precipitation, and optionally, solar irradiance data. Solar irradiance data was either used as an additional input to the weather or as a standalone network input. The input time series has a dimensionality of  $\mathbb{R}^{C \times D}$ , where  $C$  is 4 or 5 (depending on whether sun data is used)



**Figure 4.5:** Solar profiles expressed as  $Wh * m^{-2} * day^{-1}$  for 10 random fields, spanning from the minimum to the maximum cumulative yearly potential computed on fields present in the dataset. Solar profiles are computed with the *r.sun* GRASS module considering the surrounding mountain ranges for a range of 25 km around each point.

and  $D$  is the number of days from the beginning of the year to the date of the given phenological observation in train, or the desired prediction date in operational mode. Note that the weather time series has a resolution of 0.1 degrees (approximately 11.1 km), while solar irradiance has a resolution of 20 m. Consequently, the model can use solar irradiance to model specific micro-climates associated with the same weather data. Examples are areas close to mountain ranges, with a high morphological variability not captured by the relatively high weather resolution.

Yearly solar profiles are processed with a convolutional neural network. Compared to the weather network, we give as input the radiance time series consisting of 365 values independent of the day of the year of the phenology observation (Fig. 4.5). Since solar profiles represent the potential radiance, we treat it as a signature of a specific location. The reasoning is that this signature holds information about the whole development cycle of the crop, and it can be useful to discriminate between different micro-climates resulting from different exposure or environmental shading effects (e.g., mountain ranges).

The cultivars of vineyards, being a categorical variable, are converted to trainable embeddings and processed by a dense network. Embeddings are 4-dimensional vectors initialized with a normal distribution and updated during training. Cultivars character-

ized by similar behaviors are expected to be clustered together, as it commonly results from text [133].

A dense sub-network with two hidden layers processes the aspect, elevation, slope, and latitude variables.

Finally, the outputs of each network branch are concatenated together, and a dense network generates the phenology prediction. The model is trained in an end-to-end fashion.

The contribution of different variables to the final prediction was evaluated through an ablation study, in which various combinations of variables were tested.

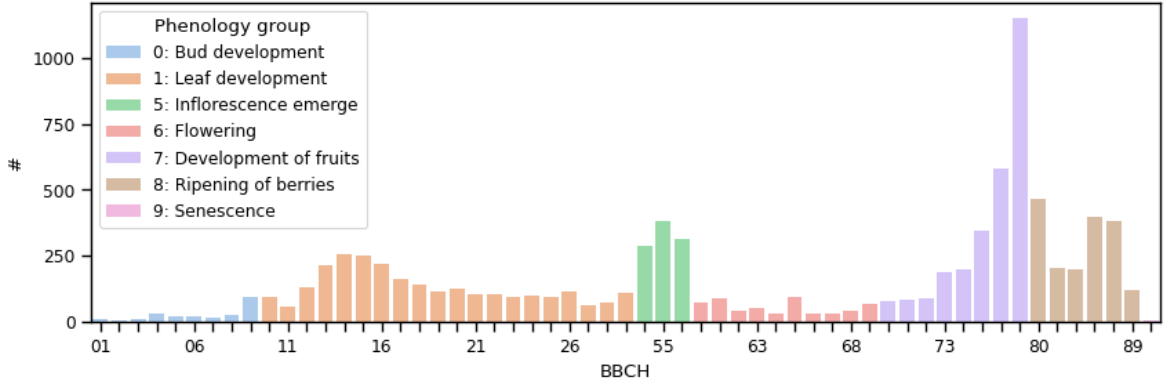
The models were developed using the PyTorch Lightning deep learning framework (Pytorch version 2.0.1, Lightning version 2.0.5) [134] and the Hydra configuration framework to run the ablation study [135]. The models were trained using two NVIDIA GPUs (Quadro RTX 4000).

### 4.2.2 In-field image classification

The second approach developed for phenology estimation consists of in-field image classification. The methodology was built in collaboration with APRA and *Università Politecnica delle Marche* within the Agrifood project (financed by *Ministero delle Imprese e del Made in Italy* and *Regione Marche*; duration 3 years).

The considered dataset comprises 7713 images with associated phenological observations collected with the Enogis software. No field contextual information was considered in this case. Images were either taken manually with smartphones or with installed webcams. Phenologies are recorded using a customized BBCH scale. The principal growth stage 1 (i.e., from 11 to 19) indicates the number of leaves unfolded and spread away from the shoot (i.e., 11 means one leaf is unfolded, 19 means nine leaves unfolded). In this case, BBCH 20 to 29 are also recorded, indicating up to 19 leaves unfolded (BBCH 29) (Fig. 4.6). BBCH up to 29 belong to the *vegetative* phase, while successive BBCH refers to the *reproductive* phase. The vegetative and reproductive phases often significantly overlap, as the latter begins while the plant continues to develop new leaves. To focus mainly on the reproductive phases, we excluded any vegetative observations recorded after the first reproductive event for each field during that agricultural season.

To estimate the phenology, we considered two computer vision models: Very Deep Convolutional Networks (VGG-16) [136] and Residual Networks (ResNet-50) [137]. Both are based on convolutional neural networks and are widely used in the agricultural literature for image and phenology classification (phenology of the apple flower



**Figure 4.6:** Distribution of BBCH values of the 7713 images with phenological information collected with the Enogis software. Images are collected for agronomic purposes in different time periods from different clients, geographical areas, and agronomists operating with different agronomical objectives. This heterogeneity in data collection processes introduces a high degree of variability in the collected phenology phases and visual characteristics of dataset images.

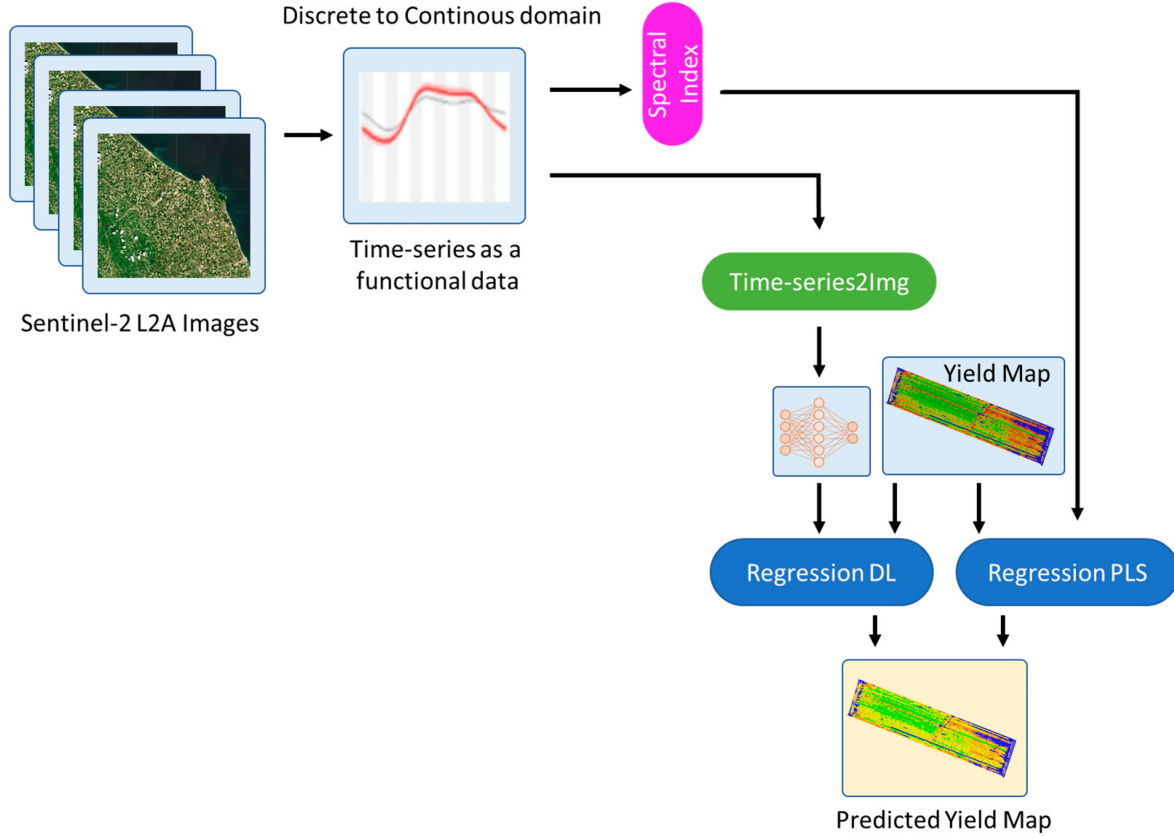
[33], classification of jujube fruits according to their level of maturity [34], classification of phenology in horticultural crops [35], phenotyping of the vine in the field [32]). VGG architectures follow a more primitive design approach consisting of a number of convolutional layers used to extract features regarding the content of the image and dense (or fully connected) layers for final classification. ResNet introduced the concept of skip-connections, which allowed for the increase in the number of convolutional layers used in deep networks to improve the result.

The models were developed with the PyTorch deep learning framework. The dataset was divided into 7494 images for train and 219 for test. To increase the robustness of the model to changes in environmental conditions (e.g., light, colors, image size) and to prevent overfitting on the training set, various data augmentation techniques (e.g., Color Jitter, Gaussian Blur, Random Autocontrast, and Random Horizontal Flip) were also applied.

### 4.3 Satellite-based durum wheat yield prediction

Crop yield prediction is an activity with the potential to impact the whole agricultural value chain. For this reason, the focus is on developing technological approaches capable of forecasting yield in advance (Figure 4.7). We developed a satellite-based approach for organic durum wheat in central Italy as an extension of the previous work with FPCA and Sentinel-2 data described in Section 4.1. A second approach consisted of arranging

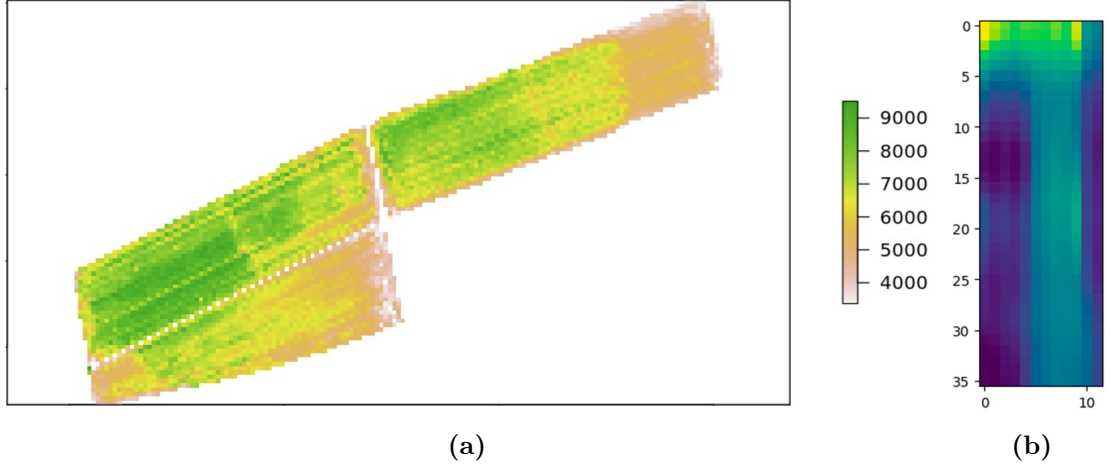
### 4.3. Satellite-based durum wheat yield prediction



**Figure 4.7:** Data processing pipeline for the durum wheat yield prediction task. Sentinel-2 and PlanetScope satellite data are organized into one time series for each ground pixel. Two approaches are tested. First (pink), the NDVI and NDRE time series are decomposed with Functional PCA. The result is used as input to the PLS regression algorithm to predict yield per pixel. Second (green), spectral data are arranged into images and processed with image-based neural networks to estimate the yield.

satellite time series as images and processing them with image-based neural networks to estimate the yield. Part of the work described next is published as a research article in *Agronomy* (MDPI) [138], which compares Partial Least Squares (PLS) regression and a deep learning network for predicting wheat yield. I developed the FPCA and PLS yield estimation pipeline, training and validating the models. The methodology was built in collaboration with APRA and *Università Politecnica delle Marche* within the BioCereals 4.0 project (project financed by Regione Marche in the context of *Programmi di sviluppo rurale regionali approvati 2014-2020*; duration 4 years).

The study area of this work consists of three geographically close testing fields located in central Italy (a total of 50 hectares). The data is sourced from a company specializing in agriculture 4.0 data collection, which has requested that specific details regarding the location remain confidential. The area is characterized by a Mediter-



**Figure 4.8:** Reference yield map of the considered area after remapping yields to the Sentinel-2 reference grid. Yields are measured in kg/ha (a). Example of an image generated for a single grid cell containing data of the 12 Sentinel-2 bands (columns) for 35 weeks (rows) (b).

anean climate with a mean annual temperature of about 15 °C., dry summers, and rainfall (about 770 mm) concentrated in autumn and winter. Yield data were collected with an intelligent combine harvester that maps dry and wet weight with a 7.6 m header. A total of 47462 points were collected over the three fields, with a frequency of 1 second and an average speed of 1.43 m/s. Data represents an estimate of real yields as harvester speed, GPS signal quality, and sensor calibration may have influenced the collection process. To use yield data as model targets, we rasterized the point geographical yields into grid cells of the same spatial reference systems of the selected satellite sources (Fig. 4.8a).

Sentinel-2 and Planet were selected as satellite data sources. The former provides data at 10-20 m horizontal resolutions, while the latter works at 3.125 m. For both missions, we collected data from November 2016 to July 2017 and generated weekly time series of NDVI, while NDRE was computed only for Sentinel-2.

The first tested yield prediction methodology consists of the combination of FPCA and PLS. FPCA was applied to both vegetation indices, and the number of functional principal components was selected using the elbow method. Yield predictions are performed pixel-wise using the original satellite-derived vegetation index or its FPCA-derived representation as input to PLS. PLS is a statistical regression algorithm with a decomposition step to reduce multicollinearity in the input features.

The second methodology instead consists of employing computer vision models, in particular VGG and MobileNet [139], to images depicting the spectral and temporal information (Fig. 4.8b). Instead of computing vegetation indices, for each pixel, we

generate an image of size  $m \times n$  where  $m$  is the overall number of weeks starting from the reference image and  $n$  is the number of considered satellite bands. Differently from the previous approach, this one was tested using only Satellite-2 data (using bands B01, B02, B03, B04, B05, B06, B07, B08, B8A, B09, B11, and B12). Bands at 20 m resolution were resampled with linear interpolation to match the 10 m resolution of the other bands. These images are given as input to architectures that use VGG and MobileNetV2 as feature extractors and perform a regression on top of their results. The idea is that instead of computing vegetation indices as hand-crafted features, we let the deep models learn the relevant features directly from the raw input (i.e., satellite-sensed bands).

## 4.4 Soil water balance assessment

The quantification of soil available water is a key result for agronomists. In viticulture, a common approach is to base the analysis on the FAO56 report [60] that describes the single and dual crop coefficient ( $K_c$ ) methods to estimate actual crop evapotranspiration ( $ET_c$ ). Evapotranspiration depicts the amount of water that is lost to the atmosphere through soil evaporation and crop transpiration.

The single crop coefficient approach combines crop transpiration and soil evaporation effects into a single  $K_c$  coefficient. On the other hand, in the dual crop coefficient approach, the two effects are modeled with the basal crop coefficient ( $K_{cb}$  and the soil water evaporation coefficient ( $K_e$ ). Thus, the single  $K_c$  coefficient follows  $K_c = K_{cb} + K_e$ . This thesis aims to introduce intelligent solutions into agriculture (with a special focus on the Enogis market) by prioritizing the geographical coverage of the solution over its complexity. Considering that the dual  $K_c$  requires more precise agronomic modeling of the considered crop, we started by focusing especially on the single  $K_c$  approach. For this reason, we put special emphasis on the single coefficient equations.

In the single crop coefficient approach actual crop evapotranspiration is computed as  $ET_c = K_c ET_0$ .  $ET_0$  is the reference crop evapotranspiration [60, 140] and is computed with the Penman-Monteith method [140]. The  $K_c$  coefficient is usually selected based on the crop, its growth stages, the crop density, fraction of land cover, height, age, and training system [141]. As a best practice, one should start by selecting from literature a suitable  $K_c$  to use as a starting point, to later calibrate and find a condition-specific adjusted  $K_c$  to use. Often, however, this is difficult or impossible due to a lack of collected data or even collection tools.

FAO56 also addresses the computation of crop evapotranspiration under soil water



stress conditions  $K_{c\ adj} = K_s K_c ET_0$ , where  $K_s$  is the water stress coefficient.  $K_s < 1$  for soil water limiting conditions, and equals 1 when there is no stress. The water stress coefficient can be estimated as

$$K_s = \frac{TAW - D_{r,i}}{TAW - RAW} \quad (4.3)$$

where  $TAW$  is the total available water,  $RAW$  is the readily available water, and  $D_{r,i}$  is the root zone depletion (i.e., water shortage relative to field capacity).

Total available water is defined as the amount of water roots are able to extract from the soil, defined by the field capacity and wilting point. After substantial rainfall or irrigation, the soil would reach the saturation condition and begin to drain until field capacity is reached. Field capacity denotes the amount of water that soil manages to hold against gravitational forces. At this point crop is able to extract water, and the water content in the root zone decreases to the wilting point. At this point, the available water is held to the soil particles with such a force that the crop is not able to retrieve it.  $TAW$  ( $mm$ ) is then computed as

$$TAW = 1000(\theta_{FC} - \theta_{WP})Z_r \quad (4.4)$$

where  $\theta_{FC}$  and  $\theta_{WP}$  are the water content at field capacity and wilting point ( $m^3/m^{-3}$ ), and  $Z_r$  is the rooting depth ( $m$ ).

Readily available water ( $RAW$ ) is the fraction of  $TAW$  that a crop can extract without suffering water stress. Crop experiences stress when the soil water content reaches below a threshold, and it cannot be efficiently used to respond to the transpiration demand.  $RAW$  is computed as

$$RAW = p\ TAW \quad (4.5)$$

where  $p$  depends on the crop type (for grapevine is 0.45, refer to table 22 of [60]). Finally,  $D_{r,i}$  refers to the root zone depletion of the  $i$ -th day and is one of the most important indicators for farmers.

At this point, it is clear that the estimation of  $D_{r,i}$  and  $K_s$  require a daily water balance computation. FAO56 defines the water balance model as

$$D_{r,i} = D_{r,i-1} - (P - RO)_i - I_i - CR_i + ET_{c,i} + DP_i \quad (4.6)$$

where  $D_r$  is the root zone depletion at the end of the day,  $P$  is the precipitation,  $RO$  is the runoff,  $I$  is the irrigation depth,  $CR$  is the capillary rise from the groundwater

table,  $ET_c$  is the crop evapotranspiration, and  $DP$  is water loss in the root zone by deep percolation. All quantities are measured in mm.  $D_r$  ranges from zero (at field capacity) to  $TAW$  (at the wilting point). Consequently, processes that increase the water amount in soil (e.g.,  $P$  and  $I$ ) are subtracted from  $D_{r,i-1}$  (i.e., the previous day root depletion), while those decreasing the water amount are added (e.g.,  $RO$  and  $ET_c$ ).

Runoff occurs when the total precipitation exceeds the soil’s capacity to store water and the rate at which water can infiltrate the soil.  $RO$  is estimated using the curve number (CN) method [109], which considers soil type, vegetation type, and antecedent runoff condition (i.e., dry, average, and wet) [142, 109].

Capillary rise from the groundwater table is usually considered 0 when the water table is found more than 1 m below the root zone’s lowest point. In this work, capillary rise contribution is ignored.

Deep percolation is the downward movement of water lost by the soil, occurring after heavy precipitation or irrigation events. Following FAO56, we simply consider that if the soil becomes saturated ( $\theta_S$ ), the next day, excess water is percolated, and the soil is at field capacity ( $\theta_{FC}$ ). If  $D_r > 0$  (i.e., the root zone is below field capacity) then  $DP = 0$ .

The single crop coefficient model was implemented starting from `pyfao56` (version 1.2), a Python library originally written for FAO-56 dual crop coefficient approach [143]. The single crop coefficient computing capability was added to the library, together with the runoff module (based on the curve number methodology), performance improvements due to an optimization of the employed data structures, and caching logic.

To run the model on the whole Italian territory, we extracted the environmental parameters from descriptors integrated into the EnogisAI data platform (Chapter 3), and the agronomic and crop parameters from the literature. In particular: soil matric potentials  $\theta_S$ ,  $\theta_{FC}$ , and  $\theta_{WP}$  are extracted from `EU_SoilHydroGrids` [105]; curve numbers for all three antecedent runoff conditions (CN ARCI, ARCII, ARCIII) are extracted from `GCN250` [110]; the single  $K_c$  crop coefficient is extracted from [141]; tree height ( $h$ ) and root height ( $Z_r$ ) are extracted from [57]; and  $p$ , length of the initial ( $L_{ini}$ ), development ( $L_{dev}$ ), mid ( $L_{mid}$ ), and final ( $L_{end}$ ) crop stages are extracted from FAO56 [60].

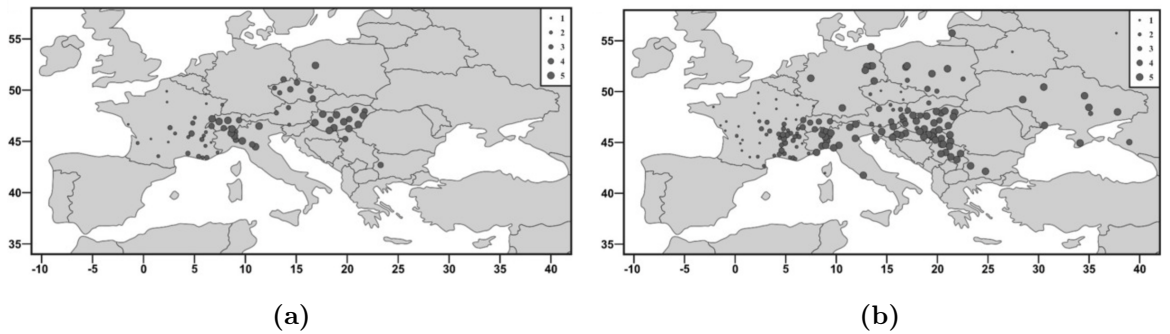
FAO56 recommends the use of the Penman-Monteith method to compute  $ET_0$ , and the `pyfao56` library utilizes the ASCE Standardized Reference ET Equation [144] to compute  $ET_0$ . While the ERA5-Land reanalysis provides all the weather variables needed as input to the Penman-Monteith model, the forecasted variables are not avail-

able in the EnogisAI context. To compute  $ET_0$ , we used a calibrated version of the modified Hargreaves–Samani 3 formula [99], which considers only temperature and precipitation data.

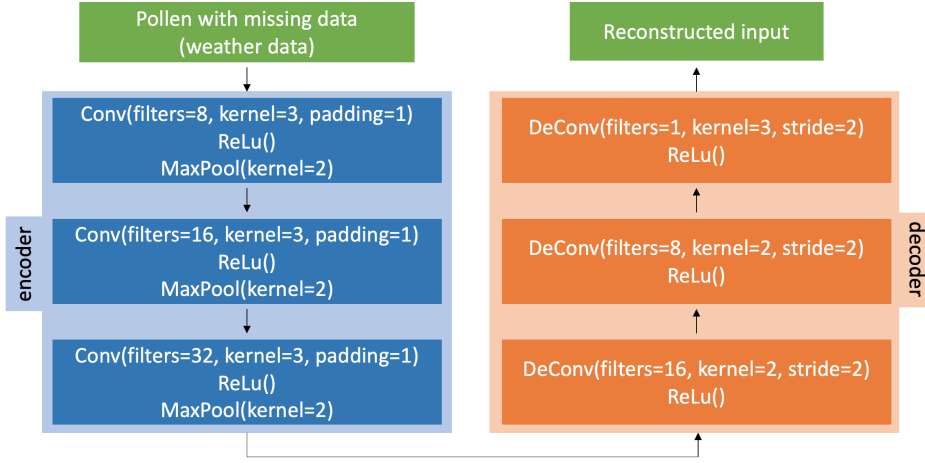
## 4.5 Pollen seasonal time series reconstruction

Data quality is a common issue at all phases of a data analysis workflow. The problem is even more pronounced when relevant observations are missing. A use case tackled during the PhD is the analysis of annual total pollen counts computed on aerobiological station data. The pollen season is characterized by narrow peaks that often account for the majority of total annual counts. Additionally, data collection is conducted manually, making it susceptible to errors. To estimate missing data in historical time series, we proposed a methodology based on denoising autoencoders. The methodologies explained here were published as a research article in *Science of The Total Environment* (Elsevier, Impact Factor: 8.2) [145], where I developed and validated the pollen reconstruction methods. The research is an effort of an international community of researchers active in the pollen research field.

We considered two datasets consisting of European aerobiological stations. The first one contains 67 stations with 16 years of data (1995 to 2010), while the second one contains 162 stations for the year 2010 (Fig. 4.9). Both contain daily ragweed pollen counts, used to generate annual total pollen statistics and maps (measured in pollen grains  $\cdot m^{-3}$ ). Each location is subject to different climatic conditions impacting the pollen season’s start, end, and duration. The period from July 15 to October 15 was selected as a uniform study period across all stations. The data exhibits substantial



**Figure 4.9:** Station and data density for the 95 stations of 1995-2010 (a) and the 162 stations of 2010 (b) datasets. The five classes in the legend refer to 0% (smaller dots) to 100% (larger dots) of data density (i.e., 100 - the percentage of missing data) with a step of 20% (i.e., class 1: is 0-20% of present data; class 5: 80-100% of present data).



**Figure 4.10:** ConvAE model architecture. Input (green block, left) contains the pollen concentration observations with missing values to reconstruct for a single city and year and, optionally, precipitation and temperature data. The encoder (blue) learns and generates the intermediate representation used to estimate the missing pollen concentration values by the decoder (orange). The network outputs reconstructed seasonal pollen concentration values (green block, right).

missing information, with approximately 50% of the dataset containing less than 60% of daily measurements.

Pollen quantity maps are generated by spatially interpolating the annual total pollen counts of every station. High-resolution raster maps were generated through Bayesian Gaussian Process Modeling using the scattered measured data points as input. For this purpose, we used the spTimer R package [146]. The study’s objective is to improve the map generation process using reconstructed data.

We proposed a Denoising Convolutional Auto-Encoder (ConvAE) to reconstruct the missing data. ConvAE is a fully convolutional 1D network that aims to reconstruct the missing data modeling the temporal dependency in the time series (Figure 4.10). The network is trained by randomly removing daily input concentrations and trying to reconstruct the original time series. The idea is that the model should be able to reconstruct the actual missing values at test time. Missing data may present various patterns. Thus, we developed two main perturbation schemes: the first selects random days to remove concentrations during training, and the second removes consecutive days’ windows. The network was designed to accept temperature and precipitation data as additional inputs to consider a wider context for reconstructing data, leading to a performance improvement. The Blended ECA dataset [147] was selected as a weather source.

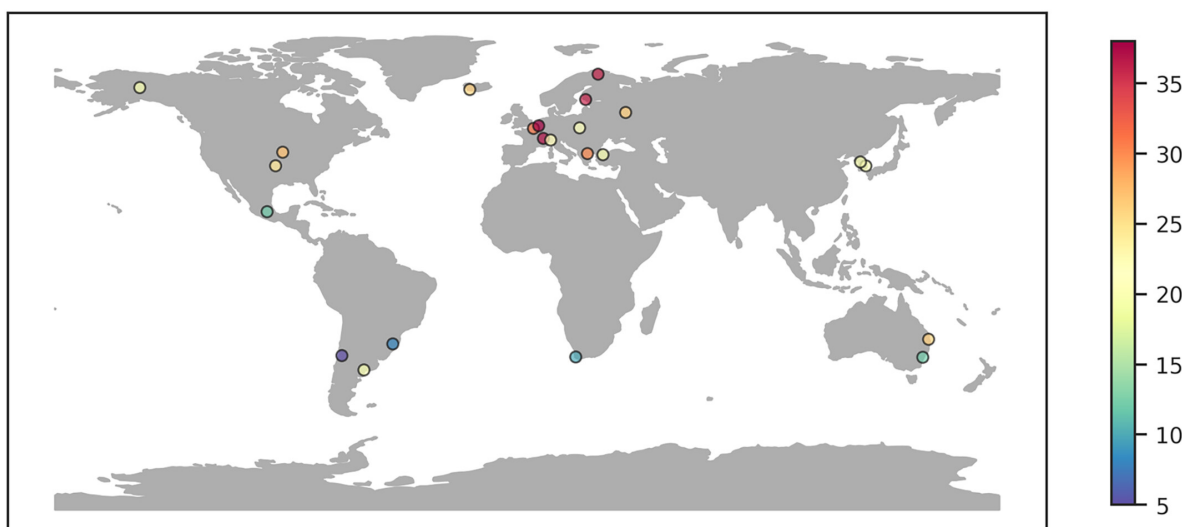
The deep learning reconstruction approach was compared against a statistical one. While the former aims to impute missing daily data, the latter models a Gaussian distribution over the data that may provide inaccurate daily estimates but reliable annual total pollen counts.

## 4.6 Time series forecasting of pollen concentrations

Forecasting pollen concentration is a crucial activity for public health. Accurate predictions enable a better management of allergy seasons, reducing health risks associated with rhinitis, asthma, and other physiological consequences.

Here, we describe two machine learning approaches for forecasting pollen concentrations for the future 7 to 28 days. The work was led in collaboration with an international research team and was published in *Allergy* (Wiley, Impact Factor: 12.6) [148]. I contributed to the design, development, and validation of the models and to researching explainability methodologies. The research is an effort of an international community of researchers active in the pollen research field.

Pollen data were collected from 23 monitoring stations across the Americas, Europe, Asia, Africa, and Oceania continents. Pollen counts are expressed as daily sums of all the taxa recorded at each station. A total of 200 taxa are monitored across the 23 stations, with specific taxa depending on the local climatic conditions. Multiple years



**Figure 4.11:** Geographical distribution/location of the aerobiological stations that provided data for the analysis. The colors refer to the number of years considered with the available daily pollen concentrations

of data are available, spanning from a minimum of 5 years for Santiago (Chile) to a maximum of 38 years for Brussels (Belgium) (Fig. 4.11).

In addition to pollen counts, we extracted 22 environmental variables from the ERA5-Land reanalysis. The objective is to predict future pollen concentration given the latest observations and future environmental forecasts. For this reason, we used ERA5-Land data both as past observations and future forecasts in the model training and validation phases.

We developed and compared two machine learning approaches. First, we built a gradient-boosted model (GB) based on the CatBoost [149, 150], which creates an ensemble of regression trees that reduces overfitting risk and improves accuracy. Second, we designed a deep learning (DL) model that follows an encoder-decoder architecture that uses 1-d convolutions and long short-term memory (LSTM) layers. The GB and DL models consider as input two windows: one of 7 past days and one of either 7, 14, or 28 future days, depending on the experiment. Pollen counts are used only in the past window, while environmental variables and the day of the year indicator are used in both windows. The GB model consists of one predictor for each future day (i.e., 7, 14, or 28). In contrast, the DL model unrolls the prediction day by day using the same latent representation computed from the input series. This behavior is expected to improve performance for longer outputs.

In addition to forecasting pollen concentration, efforts were devoted to explaining the modeling prediction dynamics. With this objective, we selected the best-performing CB model to assess variable importance for the 22 environmental parameters in predicting the daily total pollen concentration for all 23 cities. We computed the variable importance for each city and forecast day combination. Variable importance was computed with the *PredictionValuesChange* method of the CatBoost library. Variable importance vectors of each city (resp. each forecast day) were compared against those of all other cities (resp. forecast days) in a one-vs-all fashion. Significant differences were detected by applying a two-sample Kolmogorov–Smirnov test. This first approach can highlight the key role of specific variables for specific cities or forecast days.

Further, we tried to identify groups where similar dynamics occur. First, we applied the Uniform Manifold Approximation and Projection (UMAP) dimensionality reduction algorithm [151] to the feature importance vectors to condense the information into a latent space. Subsequently, we applied the DBSCAN clustering algorithm [152] to categorize similar projected vectors, aiming at identifying analogous prediction patterns. DBSCAN is an algorithm for identifying clusters of arbitrary shape.

## Chapter 5

# Models performance and their applicability in the Enogis context

### Highlights

- **Field MZ identification:** 3 FPCA components explained the 95% of time series variance. K-means clustered FPCA scores to generate MZ maps. The clusters resembled characteristics of the original time series, such as the index peak value week.
- **Crop development monitoring - remote sensing:** the model using all input sources performed best in the ablation study for feature importance ( $R^2=0.96$ ). The most critical macro stages are 5 (inflorescence emergence) and 6 (flowering). The model predicts phenologies in real-world validation scenarios with MAE=7 days ( $R^2>0.9$ ).
- **Crop development monitoring - in-field images:** ResNet-50 (MAE=3.63, RMSE=5.26) performed better than VGG-16 (MAE=3.87, RMSE=5.58). While the Agrifood project aimed at reaching MAE<3, the result fell within the satisfactory range. Prediction errors are associated with the overlap of vegetative and reproductive phases found in images. Further development and learning strategies have the potential to improve results.

■ **Crop yield prediction:** The FPCA projection method was applied to predict yield by applying PLS. Sentinel-2 NDRE obtained the best performance ( $R^2=0.755$  on the restricted temporal window T2) compared to Sentinel-2 NDVI and Planet NDVI. Image-based deep learning model using all Sentinel-2 available bands improved the result, reaching  $R^2=0.838$  on T2.

■ **Soil water balance assessment:** the single  $K_c$  method added to py-fao56 was tested with ERA5-Land, pluviometer data, ECMWF forecasts, and Hypermeteo reanalysis and forecast. ERA5-Land data does not represent a good baseline since it diffuse precipitation on multiple days, increasing the daily water intake by precipitation. On the other hand, ECMWF and Hypermeteo data present precipitation and model output trends sim-

ilar to those derived from the pluviometer.

■ **Pollen time series reconstruction:** The weather input information improved the results of the DL architecture when large windows of data are missing ( $R^2=0.82$ ). The DL model outperformed the GM model when evaluating the map reconstruction processes. This effect is associated with the more robust and accurate DL model outputs.

■ **Pollen concentration forecast:** The DL model and CatBoost (CB) models obtained similar results in predicting pollen concentration for the next 14 days ( $R^2$  of 0.716 and 0.712, respectively). CB performed better in the 1-3 days range, while DL performed better in the 5-14 range. By applying UMAP and DBSCAN, we were able to identify patterns of variable importance for cities and future forecast days.

**Personal contribution:** I designed the machine learning experiments and the evaluation processes of all the models described here. I implemented the data preparation and analysis pipelines. I developed the neural network architectures and evaluated model performance. For the crop yield prediction task, I developed 100% of the FPCA and PLS pipeline while I made minor contributions to the image-based deep learning approach.



This chapter presents the results of the methodologies presented in Chapter 4. We report experimental design, results, and discussion for each approach, focusing on the potential or actual integration within the Enogis ecosystem.

During the PhD program, the developed methodology reached different technology readiness levels (TRLs) depending on the results, the market interest in specific tools, and development priorities. Given its centrality in agronomic practices, the remote sensing phenological model reached the TRL9 (actual system proven in operational environment). Consequently, it is integrated and operated with daily steps, and final users can access and use the results. Field management zone identification in the context of grapevines reached a TRL5 (technology validated in relevant environment). While the solution was developed and deployed for durum wheat, it has not yet been fully adapted to grapevines. As of now, it was developed and integrated to validate the results with client-specific use cases. The two pollen research results are at the other end of the spectrum (TRL4, technology validated in the lab). These two pieces of research describe technological approaches relevant to environmental or agricultural applications, but they have not yet been further developed and integrated into Enogis.

Additional details on how the described methodologies are integrated into the Enogis ecosystem are depicted in Chapter 6.

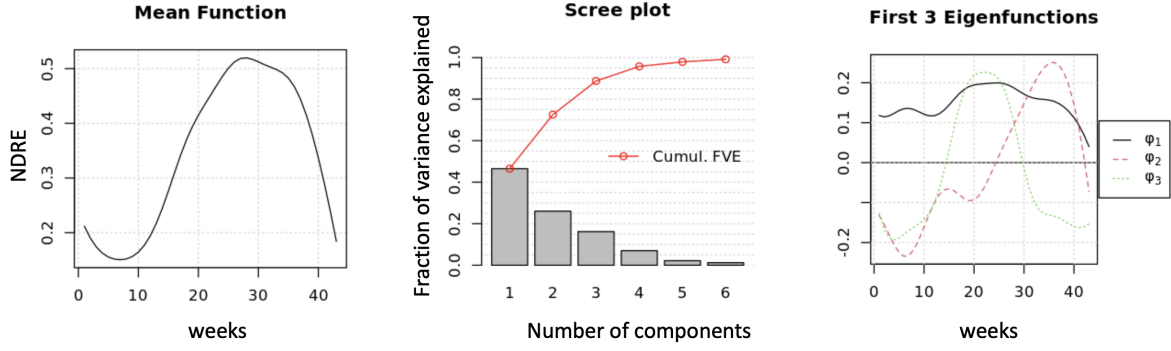
## 5.1 Field management zone identification

### Results

This section presents the results of using FPCA to derive a management zones map over selected fields starting from remote sensing time series (Section 4.1). The work presented next was published as a conference proceeding in the 2024 IEEE International Workshop on Metrology for Agriculture and Forestry (MetroAgriFor) [127]. I performed data cleaning and pre-processing and developed the FPCA+K-means pipeline. The methodology was built in collaboration with APRA and *Università Politecnica delle Marche* within the BioCereals 4.0 project (project financed by Regione Marche in the context of *Programmi di sviluppo rurale regionali approvati 2014-2020*; duration 4 years).

The proposed pipeline is applied field-wise to operate by considering solely field sensed data. We expect users to run the pipeline considering single fields to avoid interference in the data analysis phase of differently performing fields. In fact, when considering multiple fields - especially in organic agriculture - different fields may exhibit

### 5.1. Field management zone identification



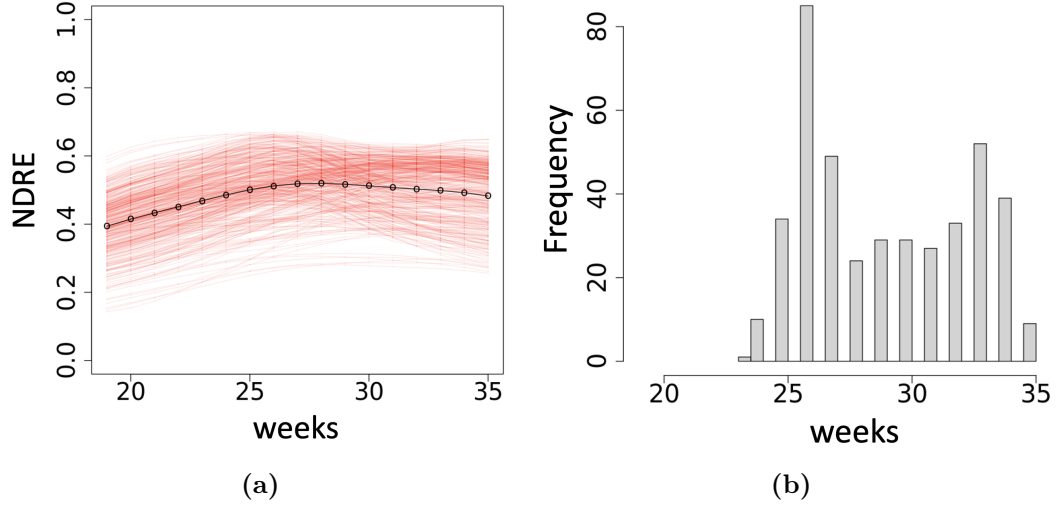
**Figure 5.1:** FPCA results on a selected test site using NDRE as vegetation index. The mean function indicates the mean of the vegetation index time series. The scree-plot shows the amount of information captured by each principal component. 3 components are enough to capture 95% of the variance. Finally, the eigenfunctions plot shows the dynamics of curves over time captured by each component. The x-axis weeks are computed starting from the first collected satellite image.

very different conditions and performance regarding crop development and yield. We present results for a sample field using the NDRE index. In general, NDRE and NDVI obtained similar results.

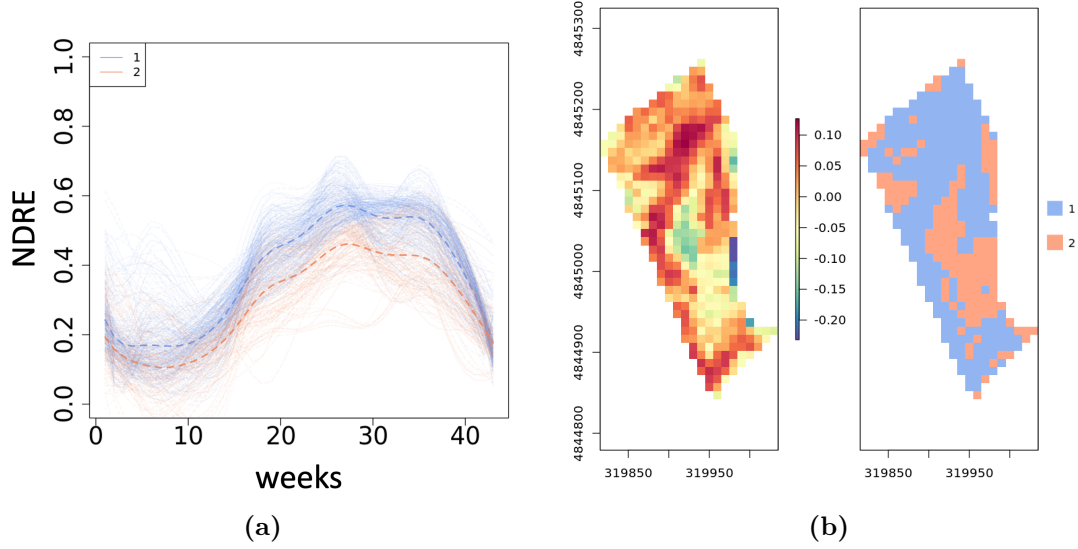
Figure 5.1 shows the output of FPCA applied to a specific field over the agronomic season using the NDRE vegetation index. The mean function highlights the field’s overall NDRE profile. Moreover, the peak phase is relatively wide, suggesting high peak-week variability. The second and third eigenfunctions are used to identify the peak week more precisely. In this case, three components are enough to explain 90% of the variance.

Figure 5.2a shows the behaviors of the simplified NDRE index during a critical period for wheat growing: the mean (black) and pixel functions (red) belonging to the testing area are the base to extract the FPCA scores. The index simplification procedures consist of using only the selected components to reproject the data back to the original space. Notably, it is possible to identify trends, maximum values, and peak weeks. Further, figure 5.2b also shows the temporal distribution of peaks derived from the reconstructed function using three principal components. This kind of information is already interesting for an agronomist to evaluate the process through which different areas reach the maximum value for the considered vegetation index.

The obtained FPCA scores are then fed to the K-means clustering algorithm. We randomly initialized the K-means centroids and selected  $N=2$  clusters. The results of the whole pipeline can be analyzed both temporally and spatially. First, it is important to check what vegetation index functions are included in each cluster (Figure 5.3a).



**Figure 5.2:** Functions projected with FPCA and peak distribution in a critical period: middle March to middle of June. It is possible to appreciate the different trends that characterize each grid cell (e.g., the peak is reached at different week numbers) (a). Distributions of week numbers associated with peaks of the reconstructed functions (b).



**Figure 5.3:** Results of clustering FPCA scores with K-means into two clusters for a sample field. The (a) side shows the NDRE index of each grid cell during the monitored period, curves are colored based on the clustering results, and dashed lines represent cluster averages. The (b) left side shows the difference between pixel FPCA scores to the mean profile, while the right side shows the clustering of FPCA scores over the two classes for the selected sample field.

For the considered case, one of the two clusters (i.e., *cluster 1*) contains all the NDRE functions that reached a higher NDRE value. Symmetrically, *cluster 2* contains the functions that reached a lower NDRE value.

From the spatial point of view, we compare the clustering results with the differences between each reconstructed function and the mean function (Figure 5.3b). This deviation map is strongly correlated with the final cluster map and can be used by agronomists to evaluate the overall field performance.

### Discussion

In this work, we presented preliminary results regarding the use of FPCA to evaluate the performance of a field starting from remotely sensed time series. One interesting aspect of FPCA is its capability to consider the data as a continuous function, helping to fill the gap in case of short periods of missing data. The FPCA scores could be used to generate cluster maps that reflect different behaviors. This kind of analysis could be performed during the season, and it could, in particular, be used to design management zones and create prescription maps.

Considering the limited computational time and resources needed to generate the clustering maps, we envision a solution that can work together with the final users. For example, a farmer or agronomist could interact with the tool by changing the hyper-parameters (e.g., the number of clusters) to analyze the output of the clustering algorithms and what features it uses to discriminate the clusters. This would allow for obtaining a final result that reflects the employed agricultural strategies and would ease technology adoption in the field. In general, deeper investigations and validations are required before adopting this approach in real agricultural practices.

## 5.2 Grapevine development monitoring

### 5.2.1 Remote sensing neural network for phenology prediction

#### Results

This subsection presents the result of the remote sensing deep learning model for phenology prediction (Section 4.2.1).

The models were trained using four distinct data splits: (i) data from 2012 to 2017 was used for the model training phase (8277 samples), (ii) data from 2018 was used for model validation during the training phase (also referred to as internal validation,

6482 samples), (iii) data from 2019 was used for best architecture selection (external validation, 6134 samples), and (iv) data from 2020 was used to implement a real-world scenario validation, 5094 samples). In addition, fields used in the training dataset were excluded from the validation datasets, with the exception of the real-world scenario dataset (year 2020). This strategy effectively minimized the risk of data leakage between the training and evaluation phases. Given that phenological observations share similarities in temporal and spatial dimensions, splitting observations from geographically close fields (or close dates) into separate training and validation sets could result in performance evaluations that do not accurately reflect real-world conditions.

Internal validation is used to select the best training hyper-parameters for each set of input sources, starting with only weather and adding the other sources incrementally. The input sources are weather, sun, sun-year, morphology, and cultivar. Sun refers to using the solar irradiation profiles up to the phenology observation day of the year, while sun-year refers to using the solar profile for the whole year. On the other hand, external validation is used to compare the best models obtained for each input combination.

The year 2020 was exclusively used to test the best-performing model selected for production use. Model performance was assessed on two subsets of 2020 data: fields included in previous years' datasets (i.e., those already used during the training and

**Table 5.1:** External validation results computed on 2019 data. Input sources are ordered by cardinality. MAE and RMSE are measured in BBCH phases. Bold values indicate the best value for each metric. The full model reached the best MAE and  $R^2$  performance (1.72 and 0.96). The model using only weather and morphology reached the second-best performance (RMSE = 2.33 and  $R^2 = 0.96$ ).

| Input sources                              | MAE         | RMSE        | $R^2$       |
|--------------------------------------------|-------------|-------------|-------------|
| weather                                    | 2.07        | 2.67        | 0.94        |
| weather + sun                              | 2.09        | 2.77        | 0.94        |
| weather + sun-year                         | 1.83        | 2.42        | 0.95        |
| weather + morphology                       | 1.77        | <b>2.33</b> | <b>0.96</b> |
| weather + cultivar                         | 2.64        | 3.34        | 0.91        |
| weather + sun + morphology                 | 2.10        | 2.79        | 0.94        |
| weather + sun + cultivar                   | 2.01        | 2.63        | 0.95        |
| weather + sun-year + morphology            | 3.71        | 4.49        | 0.84        |
| weather + sun-year + cultivar              | 1.96        | 2.55        | 0.95        |
| weather + morphology + cultivar            | 3.37        | 4.59        | 0.83        |
| weather + sun + morphology + cultivar      | <b>1.72</b> | 2.36        | <b>0.96</b> |
| weather + sun-year + morphology + cultivar | 2.07        | 2.65        | 0.95        |

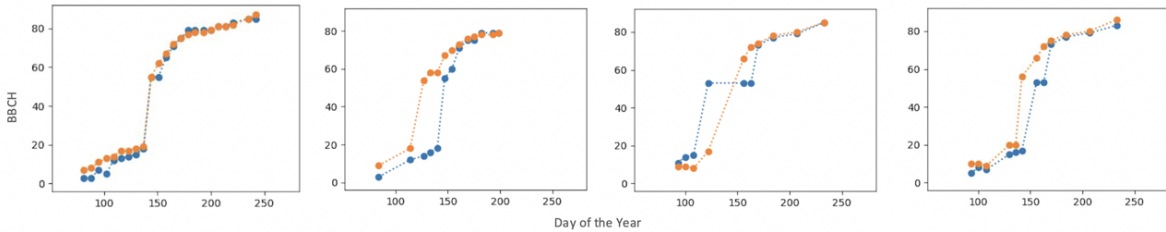
## 5.2. Grapevine development monitoring

validation phases, 3616 samples) and fields not included in past years' data (i.e., comparable to newly acquired customers or fields, 1478 samples). This evaluation aims to assess the model while minimizing selection bias.

Results are reported using Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and  $R^2$  (Table 5.1). Since the BBCH scale is expressed using codes, each scale value is transformed and associated with a consecutive integer number such that successive codes have a distance of 1. This means that the difference between BBCH 19 (9 or more leaves unfolded) and BBCH 53 (inflorescences clearly visible), being two consecutive phenological phases, is 1. Metrics are computed by comparing the observed phenological value of a specific date with the model prediction for that date. Notably, the combination of weather and morphology data reaches the best RMSE (2.33) and  $R^2$  (0.96) values, along with the second-best MAE (1.77) value. In contrast, incorporating weather, sun, morphology, and cultivar improves MAE to 1.72 (a reduction of 0.05) but slightly increases RMSE to 2.36 (0.03 higher). The  $R^2$  remains unchanged at 0.96. This behavior is likely associated with including the sun and cultivar variables, which allow for more precise field modeling at the cost of slightly losing robustness.

After evaluating both models, we selected the model incorporating weather, sun, morphology, and cultivars as the best model architecture. This model uses a broader range of input sources, significantly improving its generalization capabilities and overall predictive power.

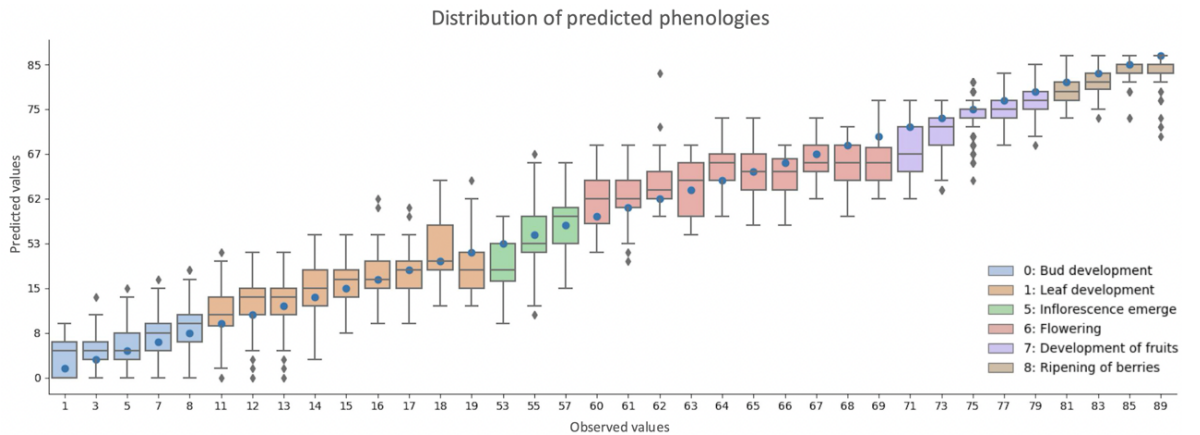
By comparing observed and predicted phenologies for specific fields, it is possible to gather insights into how the model works and qualitatively assess the performance (Figure 5.4). The first example shows model predictions that closely follow the seasonal trend of the considered field. The second one presents anticipated crop development



**Figure 5.4:** Examples of comparison between observed (blue) and predicted (orange) phenologies for four testing fields. BBCH values on the y-axis are represented interpreting the code as an integer value to highlight the difference between vegetative and reproductive phases. Predicted phenologies in the second and fourth images tend to anticipate observed values. Observed phenologies in the third image exhibit a less smooth trend; this may be associated with data collection errors that also contribute to noise in the dataset.

predictions (orange points are above blue points) up to the 150<sup>th</sup> day of the year. The third one exhibits a less smooth observational trend compared to the others. Such observations appear less natural than others; however, it is difficult to assess whether they represent an error or reflect the different interests of the technicians who performed the measurements. Similar samples may impact the model's training and evaluation processes. The last example shows an anticipation trend in the central part of the seasons, around the 130<sup>th</sup> to 160<sup>th</sup> days of the year.

The output distributions of the models for each phenological phase better highlight errors, anticipations, and delays in modeled crop development (Figure 5.5). The analysis was performed on the external validation dataset (year 2019). Larger boxplots (indicating more spread in output distributions) are associated with the shift of focus from leaf development (macro phase 1) to inflorescence emergence (macro phase 5) and flowering (macro phase 6). Flowering phases represent the most critical and variable and are thus the most difficult to predict. On the other hand, ripening (macro phase 8) is associated with the smaller boxplots, which indicates higher model accuracy. Most observed phenological phases are contained in the central 50% of the prediction distributions (the blue dot is inside the box). Boxplots skewed towards higher BBCH values indicate that the model tends to anticipate crop development (predicted values are higher than observed ones). Examples are BBCH 11 (the first leaf unfolded and spread away from the shoot), 18 (the eighth leaf unfolded and spread away from the shoot), and 64 (40% of flower-hoods fallen), among the others. The opposite (the model delays prediction) happens for BBCH 19 (the ninth leaf unfolded and spread away from the shoot), 53 (inflorescences clearly visible), and 71 (Fruit set: young fruits begin to



**Figure 5.5:** Distribution of predicted values for each observed phenological phase. The blue dots represent the correct prediction for each phenology. The highest model uncertainty is associated with BBCH 18, 55, 60, 63, and 71.

swell, remains of flowers lost).

The model was then tested on the real-world scenario data (year 2020) on the two described scenarios: fields included in the training dataset and fields not included in the training dataset (Table 5.2, *no past* column). On the former, the model reached  $MAE = 2.21$ ,  $RMSE = 3.13$ , and  $R^2 = 0.92$ . On the latter, the model reached  $MAE = 2.18$ ,  $RMSE = 3.08$ , and  $R^2 = 0.92$ . While the performance metrics are lower, they still fall within a satisfactory range, indicating that the model remains usable and practical for its intended applications.

Finally, we explored the effect of using the latest available phenological observation for a specific field and its corresponding date as input. The test was performed with the best-performing architecture found in the previous validation. The underlying rationale is that incorporating the most recent observation should calibrate the model and, therefore, improve its predictions. Notably, this effect is visible on the real-world scenario dataset (year 2020), while performance decreases on the external validation dataset (year 2019), suggesting that this approach may require further investigations (Table 5.2, *w/ past* column).

In addition to comparing the observed and predicted values of specific dates, the model can also be evaluated from a day-of-the-year perspective, which translates to measuring MAE and RMSE in days instead of BBCH values. We generated a model prediction for each day of the year (from February to November) and for each field (we recall that a field holds one or more phenological observations). This process resulted in a dataset of  $N \times 303$  samples, where  $N$  represents the number of fields in the considered dataset (external validation or real-world scenario) and 303 corresponds to the number of days between February, 1<sup>st</sup> to November, 30<sup>th</sup>. Clearly, for each field and agricultural season considered, the same phenological value is repeated for multiple

**Table 5.2:** Performance for the best resulting model on the external validation (cfr. Table 5.1), and on the two real-world scenario datasets (RWS-incl. and RWS-excl.). Reported results include model experimentation using the latest observed phenology value (*w/ past*) as input. **MAE and RMSE are expressed in BBCH.** The effect of using past phenologies observed in the season (*w/past* column) significantly improves performance on the two RWS scenarios but negatively impacts the external validation dataset.

| Dataset     | MAE     |         | RMSE    |         | $R^2$   |         |
|-------------|---------|---------|---------|---------|---------|---------|
|             | no past | w/ past | no past | w/ past | no past | w/ past |
| Ext. val.   | 1.72    | 1.77    | 2.36    | 2.48    | 0.96    | 0.95    |
| RWS - incl. | 2.21    | 1.94    | 3.13    | 2.77    | 0.92    | 0.94    |
| RWS - excl. | 2.18    | 2.05    | 3.08    | 2.99    | 0.92    | 0.92    |



days, as phenological phases do not increase with a constant daily frequency. Our next objective was to align each observed phenological phase with only one reference date to compute the error. On the one hand, if the model predicts a specific phase later than when it was observed we can consider the first prediction date as the reference date. On the other hand, supposing the model predicts a specific phase earlier than when it was observed, we can select the prediction date that minimizes the difference with the observation date. This choice is based on the fact that there is a misalignment between the development of vines up to a specific phase and the later observation of that phase in the field. Consequently, we implicitly model the observation date, not the actual development date.

In the real-world scenario with fields included in the training distribution, the model reached MAE and RMSE values of 6.80 and 9.78, respectively (Table 5.3, *no past* column). Values are comparable to those obtained on the external validation dataset (6.21 and 9.64). The MAE values indicate that, on average, the model’s predictions are off by less than 7 days, whether in advance or delayed. The model performs worse in the real-world scenario with previously unconsidered fields: MAE and RMSE increase to 7.93 and 11.40. This effect was expected, but the performance difference suggests that the model can be used in production systems, also considering that all  $R^2$  values are higher than 0.9.

The impact of using the latest observed phenology as input to the model mirrors the trends observed when evaluating the model on BBCH values: metrics show improvement for the two real-world scenarios but decrease for the external validation dataset (Table 5.3, *w/ past* column). Again, this indicates that this approach may require further investigation.

**Table 5.3:** Performance for the best resulting model on the external validation and the two real-world scenario datasets (RWS-incl. and RWS-excl). Reported results include model experimentation using the latest observed phenology value (*w/ past*) as input. **MAE and RMSE are expressed in days.** MAE values are associated with  $\pm 7$  days on the real-world scenarios. The effect of using past phenologies observed in the season (*w/past* column) significantly improves performance on the two RWS scenarios but impact negatively on the external validation dataset.

| Dataset     | MAE     |         | RMSE    |         | $R^2$   |         |
|-------------|---------|---------|---------|---------|---------|---------|
|             | no past | w/ past | no past | w/ past | no past | w/ past |
| Ext. val.   | 6.21    | 6.78    | 9.64    | 10.66   | 0.96    | 0.94    |
| RWS - incl. | 6.80    | 6.47    | 9.78    | 10.04   | 0.95    | 0.94    |
| RWS - excl. | 7.93    | 7.03    | 11.40   | 10.29   | 0.91    | 0.93    |

### Discussion

This study described a phenological multi-modal model based on weather, environmental, and cultivar data. By integrating multiple data sources, the model aims to provide a robust framework for understanding and predicting phenological events. As a result, it represents a valuable tool for researchers and practitioners in the field of phenology.

The performance assessment of the model indicated its potential for application in live systems for predicting phenological phases. The results demonstrated that the model can effectively capture the dynamics of phenological changes, suggesting that it can be reliably used in agricultural decision-making processes. The model was trained with values sampled from the Italian territory, making it suitable for use in the same geographical domain. Regarding the weather domain, the model was designed to output the phenology value corresponding to the reference date of the input weather time series. Phenology estimation can be extended by using weather forecasts as input, effectively increasing its applicability and impact on vineyard management.

Further research will explore the use of past phenological observational values as inputs to enhance model performance in more detail. By incorporating historical data, the model may better account for temporal variations and improve its predictive accuracy. A deeper performance analysis across different cultivars is essential for understanding the variability in phenological responses. After proper validation, examining how specific varieties react to environmental factors could help researchers identify which cultivars are best suited for particular conditions.

### 5.2.2 In-field image classification

#### Results

This subsection compares VGG-16 and ResNet-50 when predicting specific phenological phases (Section 4.2.2). As described in the previous chapter, we selected the VGG-16 and ResNet-50 architectures and fine-tuned these architectures starting from weights learned on ImageNet. Models were trained using the Stochastic Gradient Descent (SGD) algorithm and experimenting with different learning rate values. We posed the model optimization problem as a classification problem (each phenology representing a class) and maximized the negative log-likelihood loss during training. This project was developed in collaboration with the Moncaro vinery (Marche region, central Italy) in the context of the Agrifood project. Therefore, experiments were conducted using the Enogis dataset (7422 photos) as training and all of Moncaro's images (291 photos) as

**Table 5.4:** Test performance metrics of the best performing VGG-16 and ResNet-50 models. ResNet-50 converged with 17 epochs (compared to 65 of VGG-16) and resulted in lower MAE and RMSE (3.63 and 5.26).

| Model     | Epochs | Minutes per epoch | MAE  | RMSE |
|-----------|--------|-------------------|------|------|
| VGG-16    | 65     | 11                | 3.87 | 5.58 |
| ResNet-50 | 17     | 6                 | 3.63 | 5.26 |

a test. VGG-16 took 11 minutes per epoch (10 minutes for training and 1 minute for validation), while ResNet-50 took 6 minutes per epoch (5 minutes for training and 1 minute for validation). Models were trained for 150 epochs, employing an early stopping mechanism, using a single Quadro RTX 4000 GPU with 8 GB of RAM. Models implementation, as well as training and validation experiments, were developed in Python with the PyTorch deep learning framework.

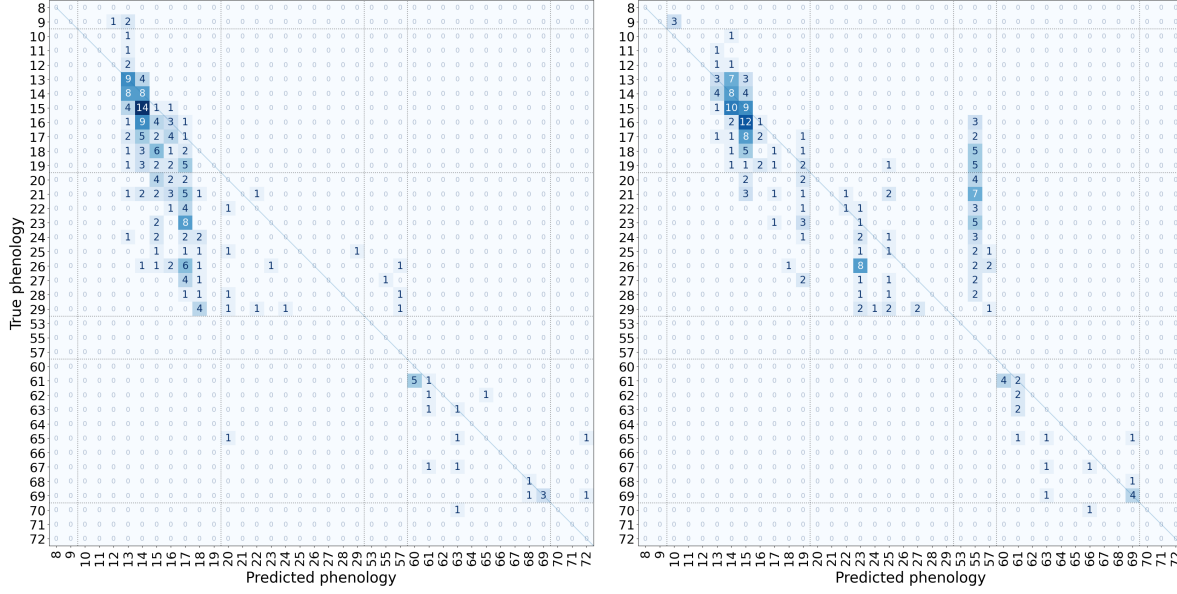
For evaluating the models' performance using Mean Absolute Error (MAE) and Root Mean Square Error (RMSE). The first metric indicates the model's accuracy, while the second indicates the robustness of the predictions. The results reported are expressed in phenological phases.

In the case of VGG-16, the best model (validated on Moncaro's data) was obtained after 65 epochs. The values of MAE and RMSE obtained are respectively 3.87 and 5.58 (Table 5.4). In the ResNet-50 case, the best model obtained was after 17 epochs. The values of MAE and RMSE are 3.63 and 5.26, respectively. ResNet-50 performed better than VGG-16 on Moncaro's data, with a difference between MAE and RMSE values of 0.24 and 0.32, respectively.

The confusion matrices show the different behaviors of the models concerning the phenological phases (Figure 5.6). Regarding the vegetative phases, the VGG-based model tends to correctly identify the early stages of leaf development (BBCH 10-18), while it is late in predicting the subsequent vegetative phenological phases (BBCH 19-26).

In the case of ResNet-50, there are some examples where the model predicts phenological reproductive stages (BBCH 53-57) for images associated with vegetative stages (BBCH 16-29). This behavior is probably associated with some images containing both vegetative and reproductive annotations. During the dataset's cleaning phase, the annotations were unified in favor of the vegetative stages. However, the model seems skewed towards predicting phenological reproductive phases when clusters are present in the images (Figure 5.7).

## 5.2. Grapevine development monitoring

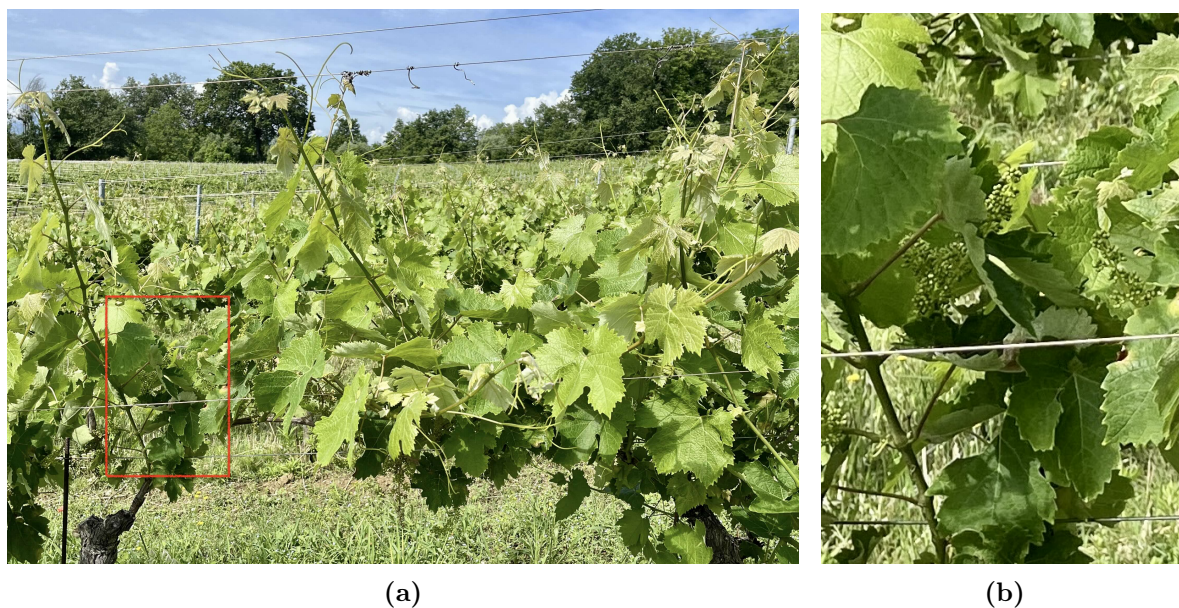


**Figure 5.6:** Confusion matrices of the VGG-16 (left) and ResNet-50 (right) models calculated on the Moncaro test dataset. VGG-16 overfitted the first vegetative phases and is skewed toward predicting phases in the 13-18 range. On the other hand, ResNet-50 worked better with later phenologies. In particular, it predicts BBCH 55 (inflorescences swelling, flowers closely pressed together) due to the presence of emerged inflorescence in the images, while farmers and agronomists still focused on vegetative phenologies). Phenology values below 8 and above 72 were never observed nor predicted and are excluded from visualization.

The target key performance indicator (KPI) for the models was established as an error margin of  $\pm 3$  BBCH values in model predictions. Both models, particularly ResNet, achieved satisfactory performance but did not meet the specified threshold. However, deeper architectures within the same model families (e.g., VGG-19 and ResNet-101) could be employed, along with newer architectures that may yield better performance outcomes. In addition, the dataset's dual annotation (vegetative and reproductive) complicates the data modeling and training processes, contributing to decreased performance. One potential solution to this challenge is to develop model architectures with multiple outputs that simultaneously classify vegetative and reproductive stages.

## Discussion

Models were trained and evaluated using fixed-installation cameras and smartphone photos. Further developments can evolve into the development of continuous monitoring applications to provide users with up-to-date daily insights. Users would be able to monitor the fields and early identify critical conditions remotely. Moreover, cameras may be equipped with additional spectral length sensors to develop models for com-



**Figure 5.7:** Sample image where the ResNet-50 model wrongly predicts a reproductive phase (BBCH 55: inflorescences swelling, flowers closely pressed together) with clusters in evidence. The image was annotated with BBCH 26: sixteen leaves unfolded. The full image (a) was used as model input. The red area represents a part of the image that shows clusters (b).

puting other crop indicators (e.g., stress indices). Following appropriate model training and validation, using the same software and hardware infrastructure

On the other hand, the obtained performance shows promising potential for integration into mobile apps or chatbots. Such a service could provide users with assisting tools for compiling phenological observations. Furthermore, it could simplify the adoption of a multi-sample strategy for phenological estimation, leading to more efficient and accurate results.

### 5.3 Satellite-based durum wheat yield prediction

#### Results

In this section, we present the results for the two different approaches used for estimating durum wheat yield: first, FPCA combined with PLS, and second, deep learning applied to satellite images (Section 4.3).

The first part presents the results of PLS regression on NDVI and NDRE functional time series, while the second reports on image-based regression using deep learning techniques. Part of the work described next is published as a research article in *Agronomy* (MDPI) [138], which compares Partial Least Squares (PLS) regression and a deep

### 5.3. Satellite-based durum wheat yield prediction

learning network for predicting wheat yield. I developed the FPCA and PLS yield estimation pipeline, training and validating the models. The methodology was built in collaboration with APRA and *Università Politecnica delle Marche* within the BioCereals 4.0 project (project financed by Regione Marche in the context of *Programmi di sviluppo rurale regionali approvati 2014-2020*; duration 4 years).

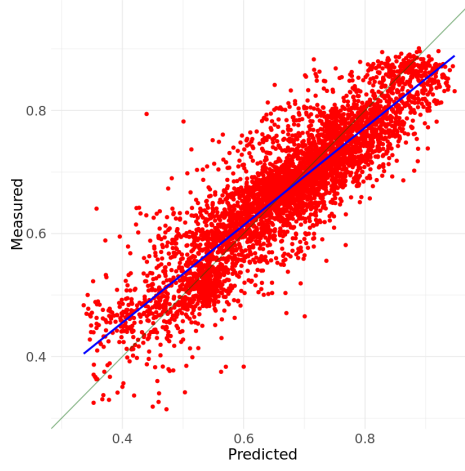
The metrics used to evaluate the models are Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and correlation ( $R^2$ ). Only areas with valid yields ranging from 2000 kg/ha to 9000 kg/ha, considered reasonable values, were processed. Other values may be considered outliers due to potential issues arising from the acquisition process, such as the tractor's speed variations while collecting data.

We consider different temporal windows to understand the relationship between seasonal data and the final yield. The period "weeks 10-36" (T1) represents the entire season, excluding the preliminary stages (seeding and emergence). Of course, during the season, the relevant window is "weeks 10-23" (T2); during this period, correction actions (agronomic operations) could be performed.

We performed the PLS analysis starting from the functional representation of the NDVI and NDRE, considering that they proved to have high correlations with the final yield. This kind of approach could also be extended to other indices. Results are reported for 10-fold cross-validation from the time series derived using the pipeline described in the previous section. Predicted values are compared against yield values collected at the end of the agricultural season. For all the reported cases, the correlation with the final yield varies from 0.74 to 0.79 when using Sentinel-2 data, while it drops to 0.59-0.61 when using Planet data (Table 5.5). The best-performing index considering both time windows is Sentinel-2's NDRE (Figure 5.8). It is interesting to note that NDRE outperformed NDVI for T2, reflecting the significance of the NDRE index in

**Table 5.5:** Results for yield prediction applying PLS to time series derived from remotely sensed Sentinel-2 L2A and Planet data. The best resulting model is Sentinel-2's NDRE for both considered windows (T1: weeks 10-36, T2: weeks 10-23).

| Satellite  | Index | Temporal window | MAE   | RMSE  | $R^2$ |
|------------|-------|-----------------|-------|-------|-------|
| Sentinel-2 | NDVI  | T1              | 0.042 | 0.056 | 0.792 |
| Sentinel-2 | NDRE  | T1              | 0.043 | 0.056 | 0.792 |
| Planet     | NDVI  | T1              | 0.060 | 0.081 | 0.614 |
| Sentinel-2 | NDVI  | T2              | 0.049 | 0.063 | 0.737 |
| Sentinel-2 | NDRE  | T2              | 0.047 | 0.061 | 0.755 |
| Planet     | NDVI  | T2              | 0.064 | 0.083 | 0.586 |



**Figure 5.8:** Scatter plot of predicted and measured scaled yields using Sentinel-2 NDRE vegetation index (T2 time window). The model obtained an  $R^2$  of 0.755. The blue line is the best-fit line, while the green one represents the 1:1 line.

supporting the estimation of the final yield throughout the season.

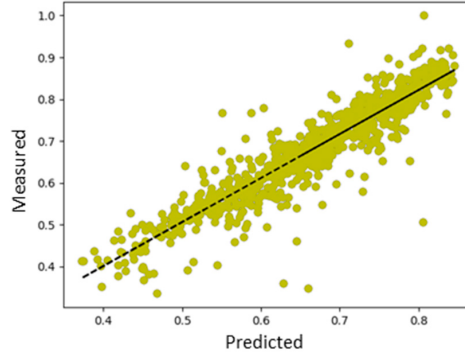
We now focus on the regression based on the image approach, which considers all the Sentinel-2 bands with different configurations in terms of weeks. The main difference with the previous methodology is that we consider all the bands over the selected time frame, trying to capture the complex interaction of time and spectral axes. VGG-16, VGG-19, and MobileNetV2 were adapted to perform a regression task by replacing the final fully connected layers. These models were trained utilizing the Gradient Descent and the Adam optimization algorithm. The input images were resized using linear interpolation to match the square input size of the selected networks.

Results are summarized highlighting the effects of different temporal windows and

**Table 5.6:** Regression results for yield with different temporal windows using the image-based neural networks. VGG-16 is the best-performing model on T1, while MobileNetV2 is the best one on T2. VGG-19 is the second-best on both temporal windows.

| Model       | Temporal window | MAE   | RMSE  | $R^2$ |
|-------------|-----------------|-------|-------|-------|
| VGG-16      | T1              | 0.033 | 0.047 | 0.843 |
| VGG-19      | T1              | 0.038 | 0.051 | 0.812 |
| MobileNetV2 | T1              | 0.044 | 0.056 | 0.758 |
| VGG-16      | T2              | 0.042 | 0.053 | 0.783 |
| VGG-19      | T2              | 0.034 | 0.048 | 0.821 |
| MobileNetV2 | T2              | 0.033 | 0.045 | 0.838 |





**Figure 5.9:** Scatter plot for VGG-16 model with linear interpolation over T1 temporal window ( $R^2 = 0.843$ ). The dashed line represents the best-fit line.

model contributions (Table 5.6). VGG-16 and MobileNetV2 were the best-performing models on periods T1 (Figure 5.9) and T2, respectively. It is interesting to consider that the  $R^2$  values in both the T1 and T2 windows in the best cases are over 0.83.

## Discussion

One of the main limitations is that the models were trained using two different temporal windows. Of course, the most relevant window is mid-season, when the farmer still has time to evaluate potential operations. Estimating the values of each band over future intervals may help exploit the benefits of a full-season model that has demonstrated higher performance in terms of  $R^2$ . The developed models can estimate yields in surrounding areas with the same crop variety and similar soil conditions. The approach mentioned above could also be used in cases involving multiple seasons. It is possible to have a multi-channel image as input (one channel per season), or seasonal images can be stacked on the time axis.

The obtained results confirm that by using remotely sensed time series based on Sentinel-2 L2A data, it is possible to estimate the crop yield with acceptable accuracy (values greater than 0.8 for  $R^2$  suggest that the estimated yield could be used to make decisions). The obtained values in terms of  $R^2$ , MAPE, and RMSE in both temporal windows suggest that predicted maps could support farmers and consortia. Knowing the potential yield, farmers can allocate resources such as water, fertilizers, and pesticides (where allowed), reducing costs and minimizing environmental impact. Yield prediction is fundamental for consortia in planning the logistics of harvesting, storage, and transportation. It could also be used to optimize contracts with farmers. Of course, in cases involving consortia where several farmers from different areas/districts confer



their harvest, we should consider that more complex trials should be executed, considering multiple crop varieties and other agronomic practices (e.g., traditional, organic, zero tillage). As already mentioned, calibrating collected data (not only yield) plays a significant role, and preliminary checks on the data quality must be performed.

Planet data was evaluated as a higher-resolution data source using the first methodological framework. The main limitation is the number of sensed bands, primarily focused on the visible and infrared ranges [46]. Future works may further investigate the combination of Planet and Sentinel-2 with topographic and climatic variables to estimate the yield correctly [153].

Deep learning models may eliminate the need to identify and select the most effective vegetation index, given that they can consider all the spectral bands over time. To enable the use of such models, it is paramount to adopt consistent and broad sampling strategies to ensure adequate data collection for modeling objectives.

Finally, we estimated the yield considering different temporal windows. Of course, considering the overall season, we derive the best results according to the selected metrics. However, this approach cannot be used during the season; problems must be detected as soon as possible so that corrective actions (agronomic operations) can be planned and executed to mitigate the effect of the final yield. For this reason, the estimation must consider shorter temporal windows. Deep learning algorithms may be used to predict future values; consequently, the predicted curves could be used to forecast the final yield using the presented models. On the contrary, the models could be trained on shorter windows, but in this case, we are not considering the historical data as in the case of the forecast. Crop rotation limits the capability to refine models considering historical data (multiple seasons) and could impose constraints regarding temporal spacing (e.g., a three-year crop rotation significantly reduces the amount of historical yield data that could be used to train models). In this context, we consider that remote and proximal sensing will play a significant role in predicting the yield and using digital solutions will support the transition to smart farming, also in areas where traditional approaches dominate today [154].

As a final note on crop yield modeling, we want to stress that at Enogis, we are currently devoting efforts to developing a grapevine yield forecast model based on multi-modal deep neural networks. The idea is to combine the lessons learned within the context of this project with the more general temporal and environmental deep architecture used for phenology forecasting. This ongoing approach addresses multiple limitations highlighted in the discussions, such as the opportunity to provide forecasts throughout the seasons continuously and the integration of other environmental descriptors.

## 5.4 Soil water balance assessment

### Results

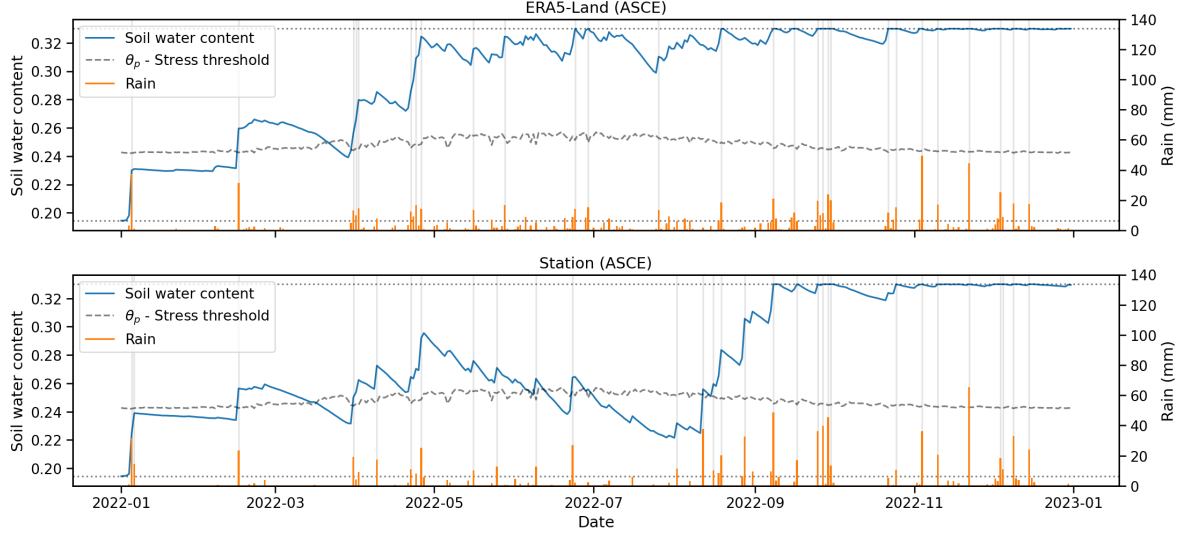
This section presents the results obtained on the soil water balance assessment model (Section 4.4). We implemented the single  $K_c$  model described in the FAO56 report [60].

Water balance model performance highly depends on accurate estimation of reference and crop evapotranspiration (respectively  $ET_0$  and  $ET_c$ ). We take as reference the Penman-Monteith method to calculate  $ET_0$  (computed following [144] methodology, referenced as ASCE in the following) and compare it to a modified Hargreaves-Shamani method (MHS3) [99]. According to the literature, we observed that MHS3 tends to overestimate ASCE when both are computed using the ERA5-Land weather variables in the specific area (Figure A.1) [99, 101].

The model's validation requires monitoring soil water availability directly in the field, which requires significant effort from farmers and agronomists. Another option is to take the available water content measurements of in-field weather stations as the ground truth value. These may be equipped with capacitive probes that continuously measure the percentage of available water content.

Here, we show the model simulations on a field in the Friuli-Venezia Giulia region in northeastern Italy, an area devoted to the Glera cultivar for producing the Prosecco sparkling wine. The field was selected based on the following criteria: it belongs to the Prosecco DOC area (the initial focus of model development) and is close to precipitation and soil moisture stations. We selected the closest precipitation station, which is 1.6 km distant from the field; it recorded data from the beginning of January 2022 to the end of September 2024. Instead, for soil moisture, we selected the nine closest stations to the field (minimum distance: 1.4 km, maximum distance: 3.6 km) and computed the median of daily measurements to account for missing data and measurement outliers. The stations recorded data from mid-June 2022 to the end of September 2024 at three different depth levels: 10 cm, 20 cm, and 30 cm.

Simulations are run with reanalysis, measured data, and forecasted weather data. ERA5-Land provided data for implementing the single  $K_c$  model in the `pyfao56` python library using the Penman-Monteith for estimating  $ET_0$  (following ASCE methodology). From ERA5-Land, we extracted daily surface pressure, 2-meter air temperature, 2-meter dewpoint temperature, total precipitation, 10-meter horizontal wind component, 10-meter vertical wind component, surface net short-wave (solar) radiation, and surface net long-wave (thermal) radiation. The precipitation station provided localized precipitation records, which are expected to be more precise than those of ERA5-Land



**Figure 5.10:** Soil water content (blue) simulation on 2022 using the ASCE method to compute  $ET_0$  using ERA5-Land (top) and ERA5-Land substituting the total precipitation variable with station recorded precipitation (bottom). The large overestimation of rainy days (orange) in ERA5-Land causes an extreme difference starting in the May-September period, leading to an overestimation of the available water content. Vertical lines (gray) indicate daily precipitation larger than 10 mm. The central dashed line (black) represents the stress threshold, i.e. when the root zone depletion grows over the readily available water (RAW). The top and bottom horizontal dotted lines (gray) represent field capacity and wilting point, respectively.

since its working horizontal resolution is about 9 km. Finally, weather forecasts for the same area from ECMWF and Hypermeteo were used to assess the model behavior in operational settings. The simulations are run settings with the initial soil moisture set to the wilting point, which effectively emphasizes water infiltration and retention dynamics.

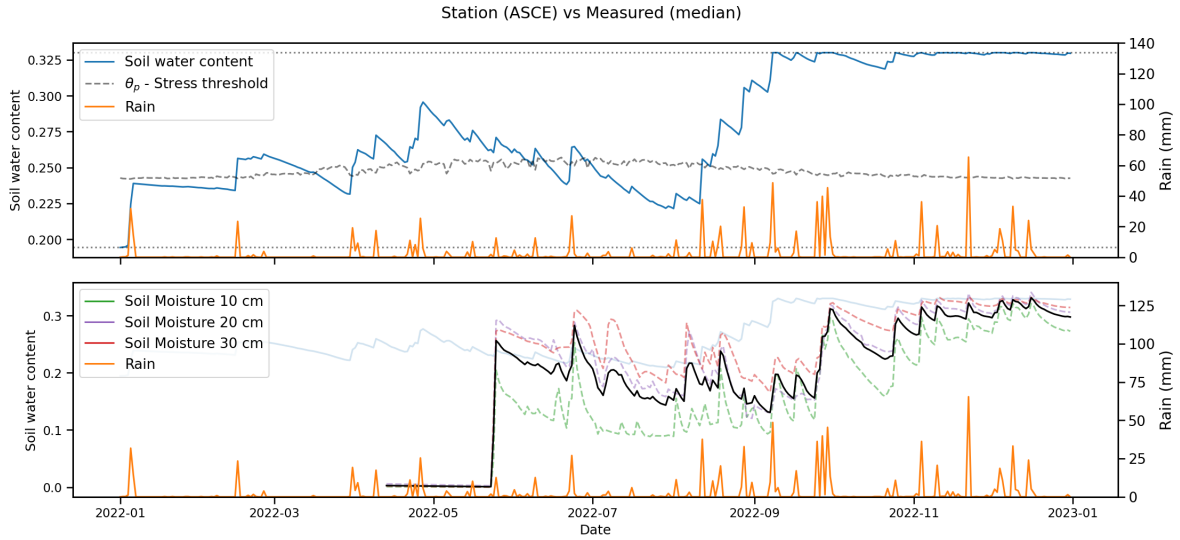
For the first comparison, we run model simulations for the year 2022, characterized by lower daily and annual total precipitation compared to 2023 and 2024. Figure 5.10 shows the simulation results obtained using ERA5-Land to compute  $ET_0$  with the Penman-Montieth method (top) and substituting ERA5-Land precipitation with station precipitation (bottom). The two simulations present a significant difference in the season's central and most critical part. This discrepancy probably derives from ERA5 generally overestimating the frequency of precipitation and underestimating the intensity of heavy precipitation events [155, 156]. A decrease in the discrepancy was observed by setting to 0 precipitation values of ERA5-land for dates where no precipitation was recorded by the station (Figure A.2).

The overestimation effect of MHS3 (compared to ASCE) impacts the model simulation output (Figure A.3, middle vs bottom). However, the effect is less prominent

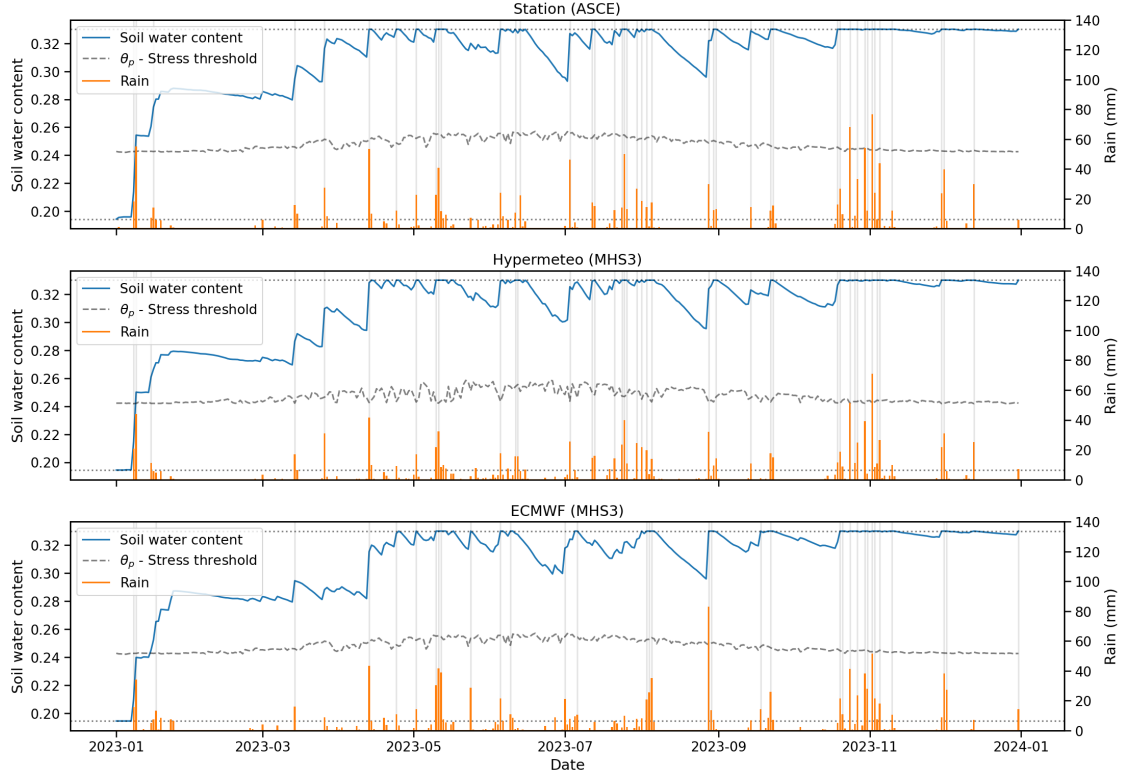
compared to the precipitation data quality effect previously observed (Figure A.3, top vs middle).

The overestimation effect of  $ET_0$  from the MHS3 method, when compared to ASCE, has a noticeable impact on the simulation outputs (Figure A.3, middle vs bottom). However, this effect is less pronounced than the influence of precipitation data quality, which was previously observed (Figure A.3, top vs middle). This suggests that while the  $ET_0$  overestimation is a factor to consider, the accuracy of precipitation data plays a more critical role in determining the overall reliability of the model simulations.

Next, we compared the simulation output (computed using the ASCE method on ERA5-Land data substituting precipitation with station observations) with the measured available water content (Figure 5.11). While the simulation and observation trends are similar, the more significant difference is in the range of values that characterize each source. The model output ranges from a minimum of 20% (initial condition set to wilting point) to a maximum of 33% (field capacity). The observed values range from a minimum of 9% to a maximum of 33% at the season's end. It is important to consider that while the simulation models up to 1 m, field probes only observe water available up to a depth of 30 cm, which may result in differences in the range of measured and simulated waters. However, in this case, the difference is related to using the



**Figure 5.11:** Soil water content (blue) simulation on 2022 using the ASCE method to compute  $ET_0$  with ERA5-Land and replacing precipitation (orange) with observed values (top). The median of the observed soil available water was computed on the nine selected stations at 10 (green), 20 (violet), and 30 (red) cm depths (bottom). The black line is computed as the mean of the three depths. The top simulation output (blue) is presented in transparent blue for easier comparison.



**Figure 5.12:** Comparison between simulations run with ERA5-Land and station precipitation data with ASCE methodology (top), Hypermeteo data with MHS3 methodology (middle), and ECMWF data with MHS3 methodology (bottom). The three simulations present comparable patterns, with ERA5-Land (with station precipitation) slightly underestimating the available water.

EU-SoilHydroGrids dataset to estimate soil hydraulic properties. Another difference can be observed after mid-August: the model tends to quickly increase the available water compared to the probes after heavy precipitation events. The single  $K_c$  model does not accurately describe field surface conditions (e.g., active ground cover) which may impact runoff and absorption dynamics.

Finally, we focus the evaluation on the operational point of view, comparing the available water content simulations computed on reanalysis and forecast data. The simulation based on reanalysis used the ERA5-Land data, substituting precipitation data with station observations and computed  $ET_0$  with the ASCE methodology. Thus, it can be considered the most accurate simulation scenario we can obtain.

Then, we run two simulations based on forecast data from Hypermeteo and ECMWF. The Hypermeteo forecast data considered in this contest consists of hourly precipitation and  $ET_0$  computed with the MHS3 methodology. Hypermeteo uses a grid of approximately 4 km. ECMWF data consists of hourly temperature and precipitation data on

a grid of 9 km.  $ET_0$  was computed with the MHS3 methodology. Forecast simulations are run for the year 2023 due to data availability.

While data are generally available for the next 7 to 14 days, depending on the source, the forecast horizon considered in the analysis is 1 day, i.e., the best-case scenario regarding data quality and weather prediction confidence. The aim is to compare the single  $K_c$  model results obtained on reanalysis and forecast data.

The Hypermeteo simulation follows the reference simulation more closely than the ECMWF one (Figure 5.12). This is probably due to the higher resolution of the grid. The analysis suggests that for the considered area, the difference in the methodology adopted for computing evapotranspiration does not significantly impact simulation dynamics compared to input data quality.

## Discussion

Soil water balance assessment is crucial for crop stress and irrigation management. The described single crop coefficient model represents a starting point for providing monitoring capabilities on the whole Italian territory. Model validation focused on comparing reference evapotranspiration computed using the Penman-Montieth method (following the ASCE standard) with a modified version of the Hargreaves-Samani method. The validation was conducted on a single field in the Prosecco DOC area, which was selected considering the availability of precipitation and soil moisture monitoring stations.

We found that the quality of input data has more impact on modeling dynamics compared to the evapotranspiration method. However, some authors suggested that when using methods different than Penman-Montieth, it is advisable to perform a method calibration considering local data [99, 101]. The calibration should be performed using historically validated weather station data that usually provide more information, although highly geographically localized.

The conducted analysis implies that ERA5-Land precipitation is overestimated in frequency, an aspect that reduces data suitability for this kind of modeling. Mitigation strategies, such as filtering low-precipitation values or using data from other sources, may be implemented.

The simulation shows an acceptable behavior compared to the observed soil moisture for the year 2022. Absolute values are comparable, although the model tends to overestimate the available water, which may depend on the estimation of soil hydraulic properties. However, the stress threshold is computed starting from the same input parameters. This may imply that the model accounts for inaccuracies in estimating

the field capacity and wilting point thresholds. Consequently, the soil water assessment model could effectively provide valuable insights to end users.

A more extensive validation is required to better assess model capabilities and investigate its limits. Validation will be performed considering other fields close to precipitation and soil moisture monitoring stations, preferably in various geographical areas, depending on data availability.

The model considers various parameters, such as the training system, active ground cover, and crop root depth. These parameters can be used where available to initialize the model and obtain a better balance estimate.

## 5.5 Pollen seasonal time series reconstruction

### Results

This section presents the results of time series reconstruction and concentration map generation achieved through deep learning models (Section 4.5). These results are compared with those obtained from the initially developed Gaussian model within the context of the study. The methodologies explained here were published as a research article in *Science of The Total Environment* (Elsevier, Impact Factor: 8.2) [145], where I developed and validated the pollen reconstruction methods. The research is an effort of an international community of researchers active in the pollen research field.

Two testing datasets were generated, namely *test* and *test103*. The first one comprises station years data drawn from the 1995-2010 dataset, while the second contains 2010 data of the 103 stations that belong to the 162 stations dataset but are not included in the 67 stations dataset.

Model performance is assessed through four metrics: Root Mean Squared Error (aRMSE), the average of Normalized Root Mean Squared Error (aNRMSE), the coefficient of determination ( $R^2$ ), and the Pearson Correlation Coefficient (Pearson R). While Pearson R and  $R^2$  are standard measures computed on the whole set of the reconstructed values, aRMSE and aNRMSE are calculated by averaging RMSE and NRMSE values obtained on each reconstructed series. Results are presented for two scenarios: removing the small window of consecutive days (Table 5.7) and removing large windows of successive days (Table 5.8). The weather input feature positively impacted the large window removal scenario, while models performed similarly in the small one. Intuitively, the improvement in aNRMSE is mainly associated with high pollen concentration values, i.e.,  $> 500$  pollen grains  $m^{-3}$ . The best-performing archi-

### 5.5. Pollen seasonal time series reconstruction

**Table 5.7:** Missing data imputation performance with weather information (ConvAE + Weather) and without weather information (ConvAE) when removing **small** windows of consecutive days from original data. Metrics are reported for the 1995–2010 data set (5-fold validation (*val*) and test sets (*test*)) and for the 2010 test data set (103 stations not present in the training distribution (*test103*)). In the case of small windows removal, the weather component does not significantly improve the reconstruction processes.

| Models          | Data sets | aRMSE | aNRMSE | R <sup>2</sup> | Pearson R |
|-----------------|-----------|-------|--------|----------------|-----------|
| ConvAE          | val       | 18.13 | 0.23   | 0.67           | 0.82      |
|                 | test      | 16.90 | 0.23   | 0.62           | 0.80      |
|                 | test103   | 28.61 | 0.20   | 0.64           | 0.82      |
| ConvAE+ Weather | val       | 18.09 | 0.23   | 0.67           | 0.82      |
|                 | test      | 16.84 | 0.23   | 0.64           | 0.81      |
|                 | test103   | 28.20 | 0.21   | 0.65           | 0.82      |

texture to reconstruct missing data in the considered datasets was obtained by training the model to reconstruct small windows of missing data using weather data.

The final aim of the GM and DL models was to estimate a more accurate annual total pollen count (ATPC) value to improve the accuracy of generated maps. We compared the reconstruction abilities of the two models, considering a set of 25 stations from the *test103* dataset. These stations had no missing data for the 2010 season; therefore, their ATPC is considered correct. We removed different percentages of data (from 10% to 80%) from the time series of the 25 stations and asked the models to reconstruct the real ATPC, evaluating two scenarios. First, we assessed model accuracy in reconstructing ATPC values with R<sup>2</sup>, RMSE, and MAE. Second, we compared the

**Table 5.8:** Missing data imputation performance with weather information (ConvAE + Weather) and without weather information (ConvAE) when removing **large** windows of consecutive days from original data. Metrics are reported for the 1995–2010 data set (5-fold validation (*val*) and test sets (*test*)) and for the 2010 test data set (103 stations not present in the training distribution (*test103*)). In the case of large windows removal, the weather component contributes significantly to improving the reconstruction process.

| Models           | Data sets | aRMSE | aNRMSE | R <sup>2</sup> | Pearson R |
|------------------|-----------|-------|--------|----------------|-----------|
| ConvAE           | val       | 23.99 | 0.34   | 0.39           | 0.63      |
|                  | test      | 19.92 | 0.32   | 0.40           | 0.65      |
|                  | test103   | 39.83 | 0.29   | 0.19           | 0.46      |
| ConvAE + Weather | val       | 24.58 | 0.33   | 0.31           | 0.60      |
|                  | test      | 16.93 | 0.23   | 0.65           | 0.82      |
|                  | test103   | 28.45 | 0.20   | 0.64           | 0.81      |



**Table 5.9:** Annual total pollen count estimation results reported in terms of  $R^2$ , RMSE, and MAE. The DL model takes the additional weather information as input. The DL model results in better performance compared to the GM model, up to 70% of missing data for  $R^2$  and RMSE, and up to 50% for MAE.

| Removed Data, % | Method | $R^2$  | RMSE    | MAE     |
|-----------------|--------|--------|---------|---------|
| 10              | DL     | 0.9971 | 203.47  | 123.03  |
|                 | GM     | 0.9654 | 705.25  | 460.60  |
| 20              | DL     | 0.9912 | 343.39  | 204.52  |
|                 | GM     | 0.9503 | 789.72  | 489.66  |
| 30              | DL     | 0.9802 | 521.27  | 323.33  |
|                 | GM     | 0.9130 | 1002.28 | 545.31  |
| 40              | DL     | 0.9678 | 680.54  | 449.96  |
|                 | GM     | 0.9299 | 976.64  | 595.34  |
| 50              | DL     | 0.9302 | 1008.30 | 644.69  |
|                 | GM     | 0.8430 | 1460.19 | 861.59  |
| 60              | DL     | 0.8745 | 1348.70 | 917.29  |
|                 | GM     | 0.8574 | 1410.63 | 886.81  |
| 70              | DL     | 0.8149 | 1635.62 | 1196.53 |
|                 | GM     | 0.7901 | 1732.98 | 1042.87 |
| 80              | DL     | 0.5262 | 2626.26 | 1971.28 |
|                 | GM     | 0.6071 | 2292.06 | 1325.30 |

Mean Average Percentage Error (MAPE) between the maps generated with real and perturbed data and the MAPE between the maps generated with real and reconstructed data. A lower value of the MAPE computed with reconstructed values indicates that the reconstruction process improves the map generation process. Both experiments were repeated 10 times, and the results were averaged.

We found that DL model performance was consistently better than GM for estimating ATPC with 10% to 70% data removal considering  $R^2$  and RMSE. With 80% missing data, the GM approach works better since DL underestimates the real value (Table 5.9). The simulation shows that with large percentages of missing data, performance becomes increasingly similar and tends to overlap in terms of standard deviation, with DL exhibiting a more robust behavior.

When evaluating the map reconstruction process, we found that the data reconstruction with GM resulted in a larger error than the data reconstruction compared to DL (Table 5.10). The negative values of the GM+ and DL+ columns (i.e., improvement of the restoration techniques compared to perturbed data) can be interpreted as the accuracy of the map restoration with the help of the given model has weakened with the given percentage data removal.

For every station in the restored databases, the missing data for given days were restored by entering the values generated by both the Gaussian method (GM) and deep learning (DL). We prepared mean annual total ragweed pollen concentration maps for Europe for the 16 years 1995-2010 (67 stations) and 2010 (162 stations). Maps were produced using both the raw data sets and the data sets restored with either the Gaussian method (GM) or deep learning (DL). The question arises as to whether restoration of the missing data changed the mapped pollen patterns.

If we take the 16-year (1995-2010) mean annual total ragweed pollen data sets including all stations, we find that the maps of the raw data set (Figure 5.13a) were very similar to both the restored (GM) (Figure 5.13b) and the restored (DL) (Figure 5.13c) data sets. However, the maps for the restored (GM) and restored (DL) data sets show higher pollen concentrations in Western Europe and the Pannonian Basin. This suggests that data restoration increases the mean seasonal pollen integrals per station.

Concerning the mean annual total ragweed pollen data sets for the year 2010, including all stations, the high similarity is obvious between the maps based on the raw data set (Figure A.4a) vs. both the restored(GM) (Figure A.4b) and the restored(DL) (Figure A.4c) data sets. The maps based on both restoration procedures show slightly higher values over France, producing a smaller gradient in the pollen concentrations from east to west after the restoration. We suppose that the pollen concentrations in

**Table 5.10:** MAPE for map generation validation results. The DL model shows an improvement for all the data removal percentages, with the exception of 60% (DL MAPE is bigger than Input MAPE by 2.90%). On average, the GM model decreases the accuracy of generated maps.

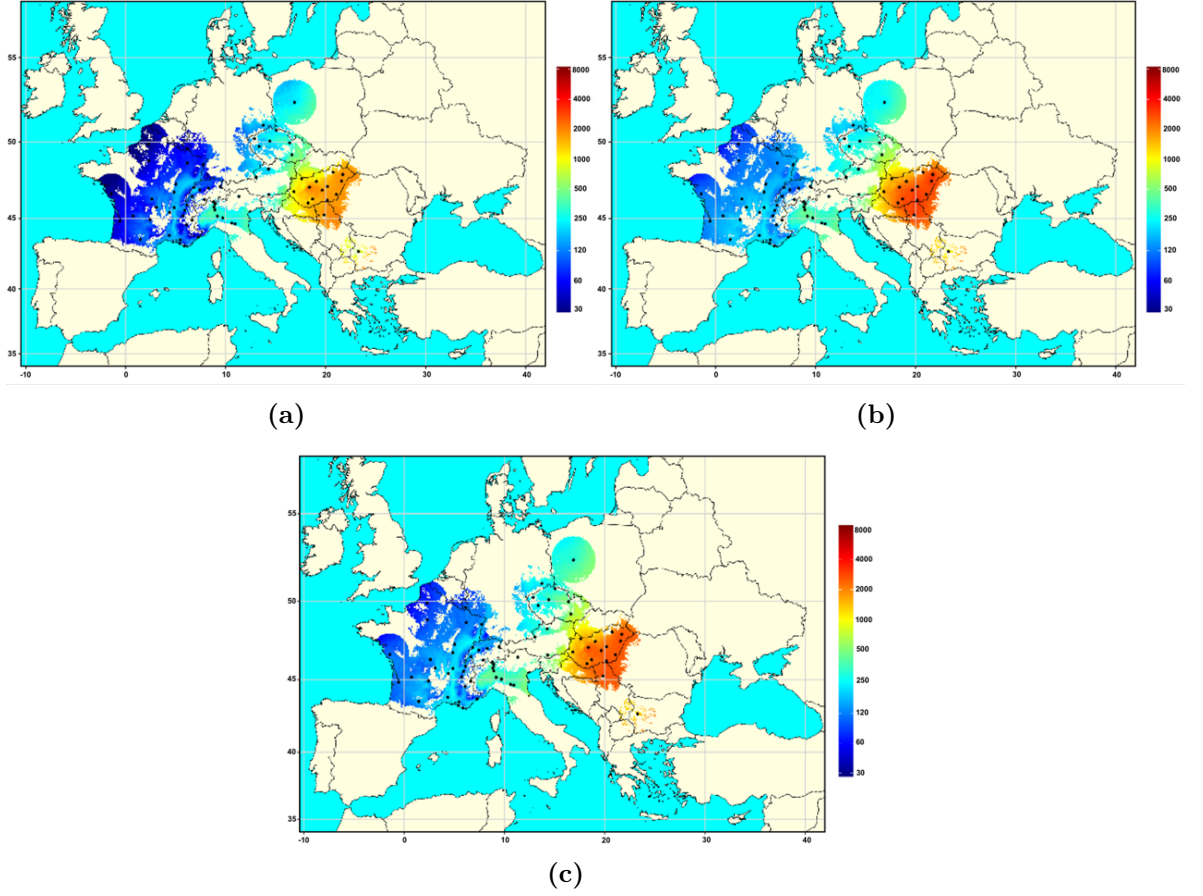
| Removed data % | Input | GM    | DL    | GM+    | DL+   |
|----------------|-------|-------|-------|--------|-------|
| 10             | 4.18  | 40.14 | 2.93  | -35.96 | 1.25  |
| 20             | 12.31 | 31.06 | 2.25  | -18.75 | 10.06 |
| 30             | 16.87 | 28.34 | 12.91 | -11.47 | 3.97  |
| 40             | 22.18 | 20.88 | 4.02  | 1.30   | 18.16 |
| 50             | 25.95 | 23.38 | 12.76 | 2.58   | 13.20 |
| 60             | 23.56 | 28.15 | 26.54 | -4.59  | -2.98 |
| 70             | 25.03 | 21.49 | 15.24 | 3.53   | 9.78  |
| 80             | 31.50 | 26.02 | 22.29 | 5.48   | 9.20  |
| Median         | 22.87 | 27.08 | 12.83 | -1.64  | 9.49  |

Input: MAPE between the real maps and the maps generated with perturbed data

GM: MAPE between the real maps and the data reconstructed with GM

DL: MAPE between the real maps and the data reconstructed with DL

GM+ and DL+: improvement in map accuracy (difference with the Input column)



**Figure 5.13:** 16-year mean annual total ragweed pollen concentrations for Europe using a base-10 logarithmic color scale, with the geographical location of the 67 aerobiological stations with raw (a), GM restored (b), and DL restored (c) data sets, July 15 - October 15, 1995-2010, (pollen grains  $\text{m}^{-3}$ ). Black dots indicate the location of the aerobiological stations. Maps computed with GM and DL restored present a higher pollen concentration,

Western Europe have been underestimated due to more missing values here. Similarly, pollen concentrations at Minsk (Belarus) and Moscow (Russia) may have also been underestimated due to more missing values in their raw data sets.

## Discussion

In this study, we developed a comprehensive pipeline that integrates temporal, climatic, and spatial data to predict pollen levels across various regions in Europe. This data-driven approach enhances our understanding of how environmental changes affect pollen distribution across Europe. Additionally, we envision applying this technique in agricultural contexts within EnogisAI. First, a multivariate time-series reconstruction model could be applied to restore missing data sourced from external inputs (e.g., mal-

functioning weather stations). Second, this approach could serve as a foundation for super-resolution models that aim to increase the temporal frequency of low-frequency data, such as vegetation indices computed from satellite observations. For instance, Sentinel-2 satellites have a revisit time of 2-3 days at mid-latitudes; however, due to cloud cover, many acquired products may be partly or completely unusable. Assuming that vegetation typically does not change abruptly, it would be feasible to correlate multiple weather and environmental variables to model the missing NDVI values, potentially achieving a daily frequency. This information could then be integrated into subsequent modeling phases.

## 5.6 Time series forecasting of pollen concentrations

### Results

This section presents the experiments and results obtained on the daily pollen concentration task using gradient boosting and deep learning approaches (Section 4.6). The work was led in collaboration with an international research team and was published in *Allergy* (Wiley, Impact Factor: 12.6) [148]. I contributed to the design, development, and validation of the models and to researching explainability methodologies. The research is an effort of an international community of researchers active in the pollen research field.

Data was split into train, validation, and test datasets based on the years available for each monitored station. The first 50% of data was selected for training, the next 25% of data for validation, and the last 25% of data for testing. For example, Tulsa (USA) provided data for the years 1996–2019: 1996–2007 was used as a training set, 2008–2013 as a validation set, and 2014–2019 as a test set.

Metrics are reported using Pearson R and  $R^2$ . To establish a general model performance comparison, we first show the experimental results as an average of the results obtained over the forecast days and cities to show the global performance of the models (Table 5.11). All models outperform the persistence model (PERS), with or without the past daily total pollen concentration. We report performance metrics computed on the whole test dataset, as is usually done in the literature. In the following, we emphasize model results on days with a pollen concentration greater than 0 (57.76% of the total dataset).

Using past pollen information as an input feature improved the forecasting performance of CatBoost (CB\_7) and deep learning (DL\_7) models compared to models

**Table 5.11:** Pearson R and  $R^2$  metrics for the CatBoost (CB) and deep learning (DL) models. Numbers in the model versions refer to the number of past days used as input, while *np* indicates that the daily total pollen concentration was not used as an input variable. The Pearson’s R p-values are all significant at the .001 level. All models beat the persistence model, and using pollen improves over not using it (CB\_7 vs CB\_7np and DL\_7 vs DL\_7np). Increasing the input context length over 7 days does not significantly improve model results.

| Model       | Pearson R <sup>a</sup> | $R^2$ <sup>a</sup> | Pearson R <sup>b</sup> | $R^2$ <sup>b</sup> |
|-------------|------------------------|--------------------|------------------------|--------------------|
| CB_7        | 0.712                  | 0.486              | 0.687                  | 0.449              |
| CB_7np      | 0.639                  | 0.372              | 0.607                  | 0.326              |
| CB_14       | 0.712                  | 0.485              | 0.687                  | 0.448              |
| CB_28       | 0.714                  | 0.490              | 0.690                  | 0.452              |
| DL_7        | 0.716                  | 0.495              | 0.694                  | 0.462              |
| DL_7np      | 0.650                  | 0.403              | 0.619                  | 0.361              |
| DL_14       | 0.705                  | 0.478              | 0.680                  | 0.439              |
| DL_28       | 0.702                  | 0.463              | 0.678                  | 0.424              |
| Persistence | 0.469                  | -0.063             | 0.430                  | -0.137             |

<sup>a</sup> The metrics are computed considering all (including 0) daily total pollen concentrations.

<sup>b</sup> The metrics are computed considering only the daily total pollen concentrations greater than 0.

without past pollen information (CB\_7np and DL\_7np) (Table 5.11).

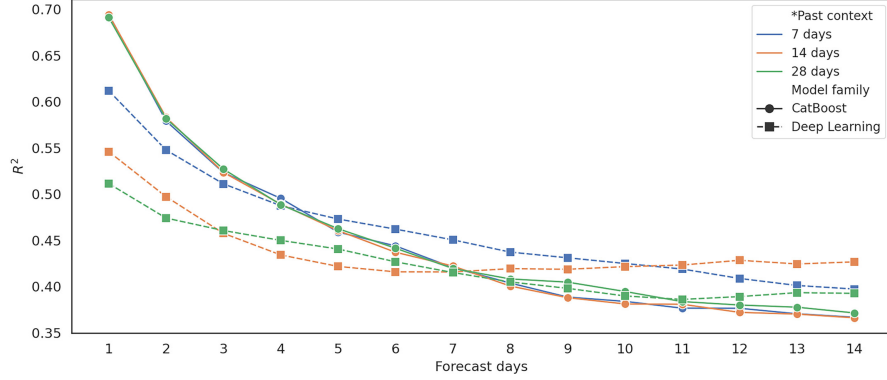
Both the CB and the DL models showed that increasing the past temporal context over 7 days improved the forecasting results only marginally (Table 5.11). The CB and DL model classes are comparable, although the DL model class is associated with slightly higher  $R^2$  (best  $R^2$ : .462 for DL\_7) and Pearson R (best R: .694 for DL\_7).

Since increasing the past input context improves predictive performance only marginally, we focused on models using  $p = 7$  past input days (i.e., CB\_7 and DL\_7) and, therefore, a lower number of model parameters.

Analyzing the daily averaged performance, the CB\_7 models performed better when forecasting in the range of 1-3 days ( $R^2(\text{CB}_7) \approx .69$ ,  $R^2(\text{DL}_7) \approx .61$  for prediction on the 1st day), and show similar performance for Day 4 ( $R^2(\text{CB}_7) \approx .50$ ,  $R^2(\text{DL}_7) \approx .49$  for prediction on the 4th day), while the DL\_7 model obtains better performance for Days 5-14 ( $R^2(\text{CB}_7) \approx .37$ ,  $R^2(\text{DL}_7) \approx .40$  for prediction on the 14th day) (Figure 5.14).

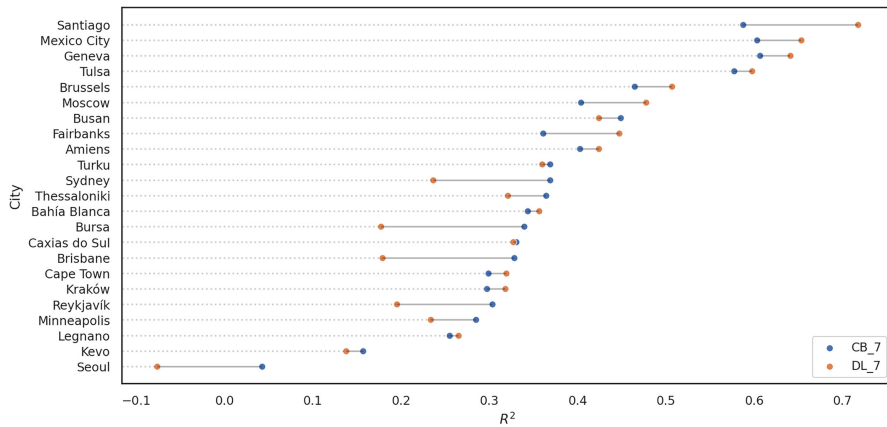
The CB and DL forecasting performance shows different behavior depending on the city (Figure 5.15). Interestingly, the DL\_7 model outperforms the CB\_7 model in 14 out of 23 cities in terms of  $R^2$ . The most significant differences in the model performance data sets can be observed in decreasing order for Bursa (Turkey), Brisbane (Australia),

## 5.6. Time series forecasting of pollen concentrations

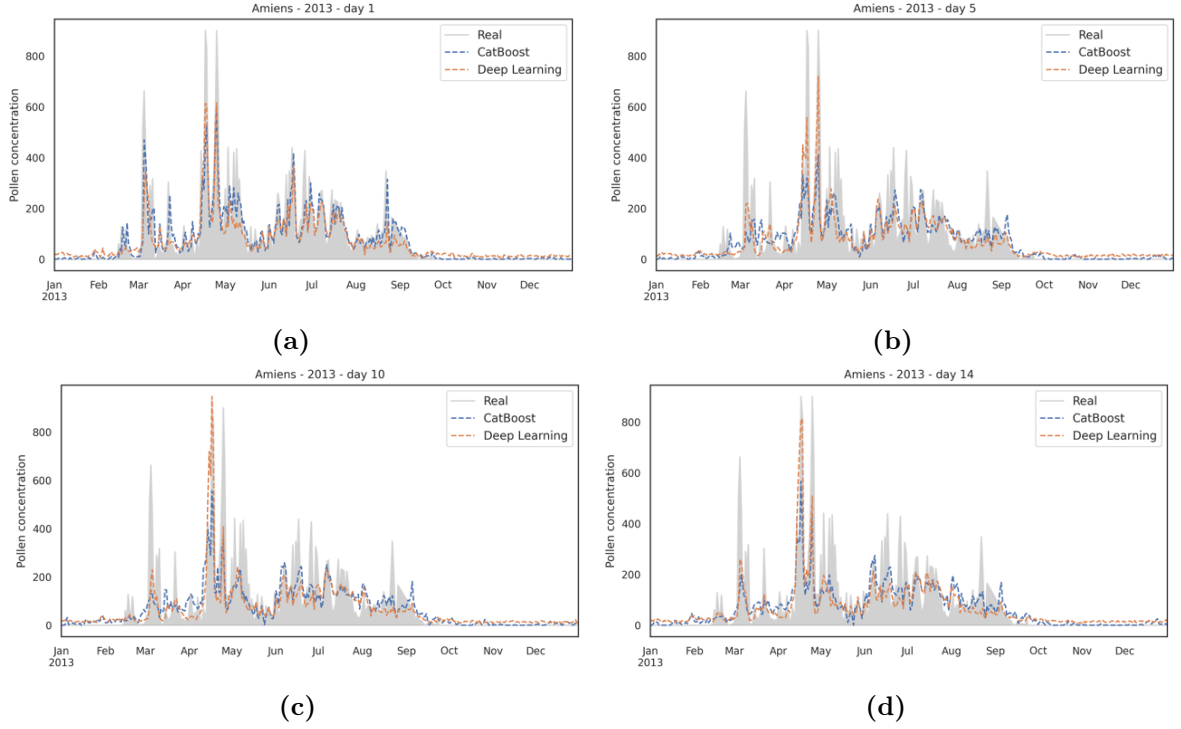


**Figure 5.14:** Modeled  $R^2$  for each day, averaged over the 23 cities.  $R^2$  values are computed using only the daily total pollen concentrations greater than 0. CB models (circle symbol and connected lines) work better for the 1-3 forecast days. DL models (square symbol and dashed lines) work better for the 5-14 forecast days (starting with DL\_7).

Sydney (Australia), Santiago (Chile), Reykjavík (Iceland), and Seoul (South Korea) (Figure 5.15). At the same time, where the performance difference was very large, in five out of these six cities, the DL\_7 model performed very poorly compared to the CB\_7 model. This may reflect that the DL model suffers more in cities with fewer data available for training or sudden temporal changes in pollen concentration, leading to substantial differences in the performance of the two models in inter-city comparisons (Figure 5.15). However, when the predictive performance of the two models is similar for a specific city, the output time series closely follows the same trend and differs mostly for daily values (Figure 5.16). Notably, we did not find any correlation between



**Figure 5.15:** Differences in the model performance ( $R^2$  for each city averaged over days) for the best CB and DL models. DL\_7 (orange) outperforms CB\_7 (blue) for 14 cities. Brisbane (Australia) and Bursa (Turkey) present the largest differences in the performance.  $R^2$  values differences are computed using only daily total pollen concentrations greater than 0.



**Figure 5.16:** Predictions of CB\_7 (blue dashed line) and DL\_7 (orange dashed line) models for Amiens in the year 2013. The background (gray) represents the observed pollen count. The forecasts are presented for 1 (a), 5 (b), 10 (c), and 14 (d) days in advance.

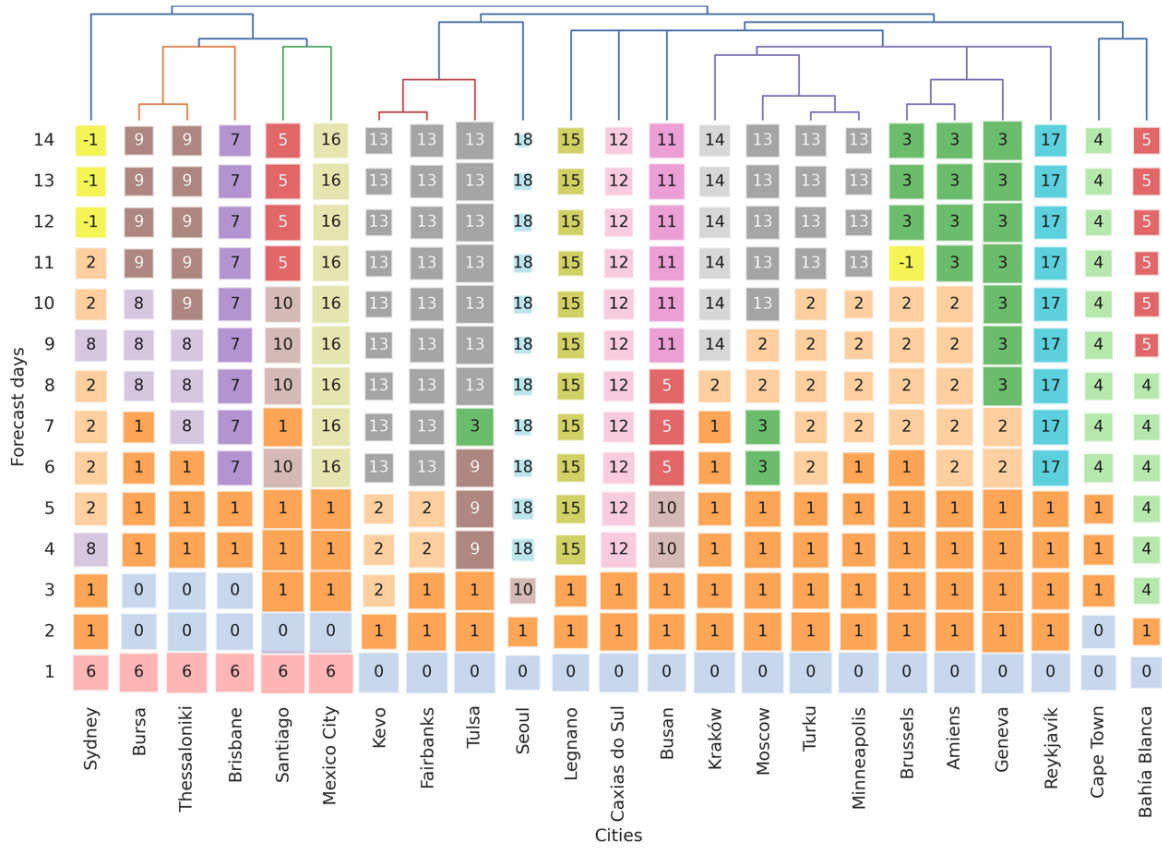
the number of years available for each station and the model performance of each city.

We used the CB\_7 model to rank the environmental factors that affect the pollen concentration forecast for cities and days. The five most important variables in determining the future daily total pollen concentrations are the past daily total pollen concentration, the future 2 m temperature (i.e., the future temperature at 2 m height above the surface), the past 2 m temperature, the past soil temperature level 3 (namely, the soil temperature 28-100 cm below the surface), and the past soil temperature level 1 (0-7 cm below the surface)

The next step consisted of detecting similarities between the cities and model dynamics over the forecast days. We obtained 20 clusters from the application of the UMAP (parameters:  $n\_neighbors = 7$ ,  $min\_dist = 0.05$ ,  $n\_dim = 2$ ) and DBSCAN algorithms (parameters:  $eps = 0.4$ ,  $min\_samples = 4$ ) to the projected samples, . Each cluster grouped similar variable importance vectors associated with cities and forecast days. For example, cluster 0 indicates that the same environmental variables are important for 17 cities out of 23 for predicting the daily total pollen concentration for the 1st forecast day and, in some cases, also for the following days (Figure 5.17).

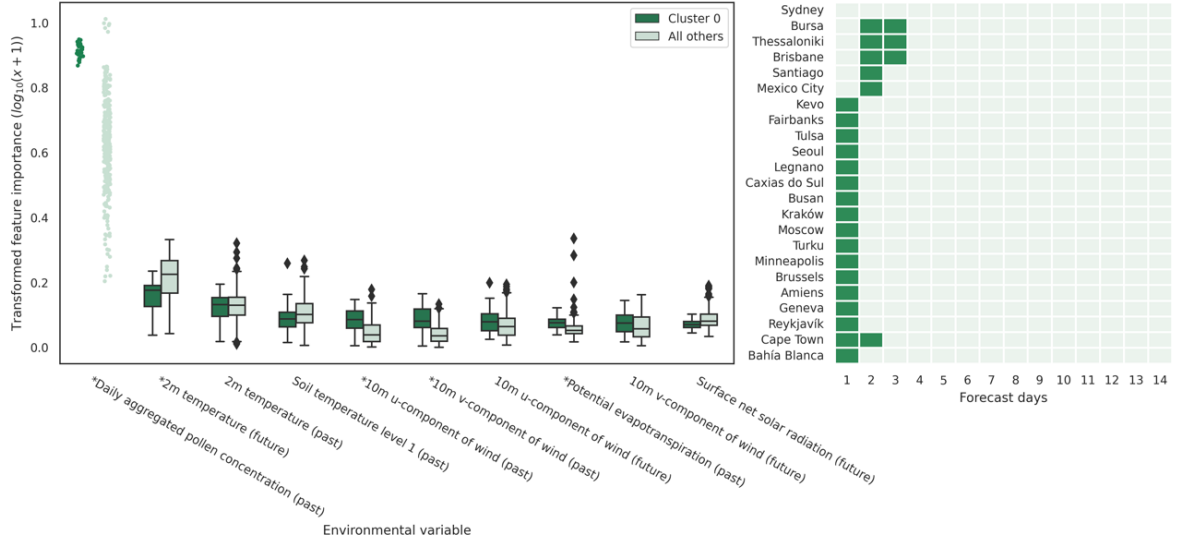
## 5.6. Time series forecasting of pollen concentrations

Additionally, the same features that are useful for predicting the daily total pollen concentration for the 1st forecast day are also relevant for predicting pollen concentration for the second forecast day at Cape Town (South Africa), Brisbane (Australia), Mexico City (Mexico), Bursa (Turkey), Thessaloniki (Greece), and Santiago (Chile) (Figure 5.17). Cluster composition was characterized by comparing the distribution of feature importance values of the environmental variables of all cities and forecast days in a given cluster with the distribution of feature importance values of all other cities and forecast days. For example, we compare the distribution of the environmental



**Figure 5.17:** Clustering results for all combinations of city-day. The numbers in the squares are the serial numbers of the 19 clusters from 0 to 18, comprising groups of various environmental variables. A value of -1 indicates points classified as noise from the DBSCAN algorithm (points that are not part of a group of at least 4 points). For example, cluster 0 indicates that the same features (i.e., the same environmental variables) are important for 17 cities for predicting the daily total pollen concentration for Day 1, i.e., for the next day. At the same time, for Day 2, the same environmental variables in cluster 0 are important only for six cities. Larger squares indicate a higher predictive performance for the city-day combination (smallest square: Seoul (South Korea), day 12,  $R^2 = 0$ ; biggest square: Mexico City (Mexico), Day 1,  $R^2 = 1$ ; median size square: Busan (South Korea), Day 7,  $R^2 = .506$ ). The dendrogram at the top indicates cities that are similar in terms of the clusters they are assigned to.





**Figure 5.18:** The top 10 most important environmental variables in decreasing order of the median of their feature importance. Values from left to right are used to predict the daily total pollen concentrations, cluster 0, and city-day combinations. Feature importance values were transformed using the  $\log_{10}(x + 1)$  transformation for visualization purposes. Feature importance values of the daily total pollen concentration (past) both for cluster 0 and all other clusters are presented with a strip plot to highlight the number of points belonging to each group. The diagram on the right indicates what city and forecast day pairs are included in the cluster.

variable importance of the elements of cluster 8 (there are a total of 8 elements from cluster 8; see Figure 5.17) with the distribution of the environmental variable importance of all other cities and forecast days (23 cities x 14 forecast days = 322 elements - 8 elements = 314 elements; see Figure 5.17). A two-sample Kolmogorov-Smirnov test was conducted to identify only the significant variables' importance for each cluster, selecting the environmental variables associated with a  $p < .005$ , or 0.5%. For example, for city-day pairs included in cluster 0, model predictions are heavily based on the past pollen concentration values and are less sensitive to, for example, the forecasted temperature (i.e., 2 m temperature (future)) (Figure 5.18, dark boxplots), compared to the distribution of its feature importance in other clusters (Figure 5.18, light boxplots).

## Discussion

Building on previous pollen-related research, this study investigates AI methodologies for forecasting pollen concentrations over the next 7 to 28 days, with a focus on explaining the influence of environmental variables across different climatic conditions. A natural progression of this study is to integrate these explanatory techniques into EnogisAI

to assess the contribution of various environmental and climatic factors in monitoring crop growth and forecasting yield. Understanding modeling dynamics makes two key objectives achievable: (i) defining strategies to improve model performance through the inclusion of additional data or the adoption of alternative design approaches and (ii) gaining deeper insights into agricultural processes, which will facilitate the creation of more intuitive and explainable tools for farmers. Both of these initiatives have the potential to increase digital technology adoption in agriculture by addressing the barriers related to trust and usability.

## Chapter 6

# Decision Support Systems for optimizing agricultural activities

### Highlights

- DSSs were developed for phenology estimation, disease coverage monitoring, and soil moisture analysis. All DSS were integrated into the Enogis software and provided to users.
- By providing actionable insights, the DSS influences field management strategies, optimizing crop health and effectively maximizing yields.
- Users can perform broad comparisons across fields or conduct in-depth analyses, enabling informed decisions based on comprehensive data.
- The DSS is computed with daily frequency on 121296 fields (relative to 27999 agricultural companies). The phenology DSS was activated for 118206 fields, the disease coverage one for 36398 fields (total of 93651 treatments), and the soil moisture one for 61430 fields.

**Personal contribution:** I implemented the EnogisAI DSS system and the distributed and parallel computing infrastructure to daily update all DSSs. I developed the DSS APIs for integration with the Enogis software and contributed to the design and implementation of the JWT authorization system.

---

This chapter provides an overview of the AI tools designed for in-field support that are deployed within EnogisAI for Enogis users, including farmers, agronomists, and agricultural consultants. The goal of these tools is to provide users with decision support systems (DSS) to enhance their decision-making processes, optimize agricultural practices, and improve overall productivity. The chapter will detail the functionalities and benefits of these AI tools, illustrating how they can effectively address the unique challenges faced in the agricultural sector.

DSS outputs can be presented in a variety of formats, including geographic visualizations on the Enogis WebGIS tool, statistical charts, and concise indicators displayed in tables and summary lists. The effectiveness of these visualizations can vary depending on the specific tasks users are performing within Enogis. For instance, users may be interested in understanding the status of all vineyards, investigating development dynamics, or monitoring potential future changes. The DSSs described in this chapter aim at providing efficient tools for answering these questions.

From the perspective of this thesis, and building on the concept of Technology Readiness Levels (TRL) discussed in Chapter 5, we have transitioned the models that achieved a satisfactory maturity level into production. This includes the phenology model and a pesticide coverage system (initially developed in collaboration with Fondazione Bruno Kessler and Università di Trento as part of the Enophit project). Both systems used the remote sensing BBCH model discussed in Section 4.2.1 as a fundamental component. The integration of the model allows for a better assessment of plant growth stages and leads pesticide usage optimization, ultimately supporting more sustainable farming methods.

As part of the innovation initiatives within EnogisAI, we decided to expand the selection to include the soil water balance module described in Section 4.4, aimed at providing an effective irrigation management tool. Although the results of this module have not yet been extensively validated, we chose to further develop and refine the system by making it available to end users and gathering their feedback. In 2024, we launched the system in the Prosecco DOC area, utilizing Hypermeteo weather data, while for the rest of Italy, we employed ECMWF data. This marks the first experimental year of implementation, allowing us to assess the system’s performance in real-world conditions.

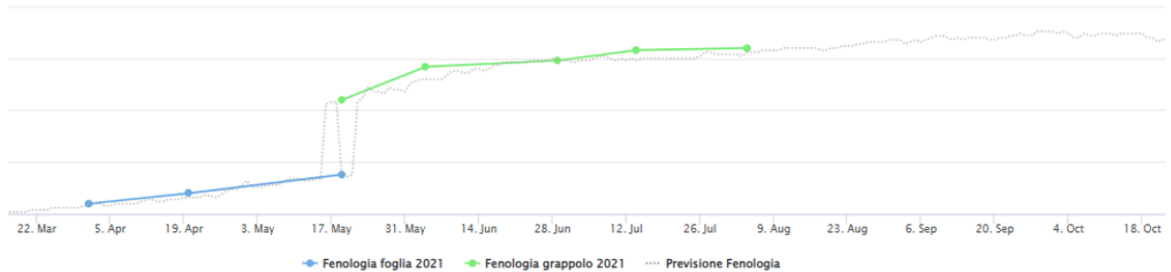
Finally, we introduce a tool that enables end users to explore various data and model products directly through the Enogis WebGIS platform. This tool allows users to define custom rules for classifying and analyzing specific geographical areas, such as a DOC area. By facilitating interactive data exploration, users can gain deeper insights

into spatial information and make informed decisions based on tailored analyses. In addition, by using this tool agronomists can unlock their knowledge and experience to integrate and empower it with data analysis outputs interactively. As a result, we are effectively developing a user-centered intelligent data analysis framework where end users are not merely passive participants but can actively engage with the underlying data and knowledge base in a continuous query refinement loop. This approach empowers users to explore and manipulate data dynamically, fostering a more interactive and insightful analytical experience.

## 6.1 Phenology DSS

The phenology DSS aims to provide users with a clear picture of crop development stages across all vineyards. The DSS was integrated into two main sections of the Enogis application. First, it was integrated into the phenology and pathology sheet (Section 3.1.2), and second, in the WebGIS module. Each phenology and pathology sheet is associated with a vineyard and usually contains multiple field observations during the agronomical season, ranging from a minimum of 1-3 to a maximum of 20-30, depending on the agronomical interest in the area or field.

The integration within the sheet is designed to add contextual information regarding phenology, showing estimates of past and future crop development dynamics. Users can use the sheet to record phenological data for each sampling date, which is visually represented on an interactive line chart alongside the corresponding dates (see Figure 6.1). The y-axis displays BBCH categorical codes in their integer form, effectively underlining the distinctions between vegetative and reproductive phases. Specific BBCH codes and their description are interactively shown when hovering over the chart. To enhance

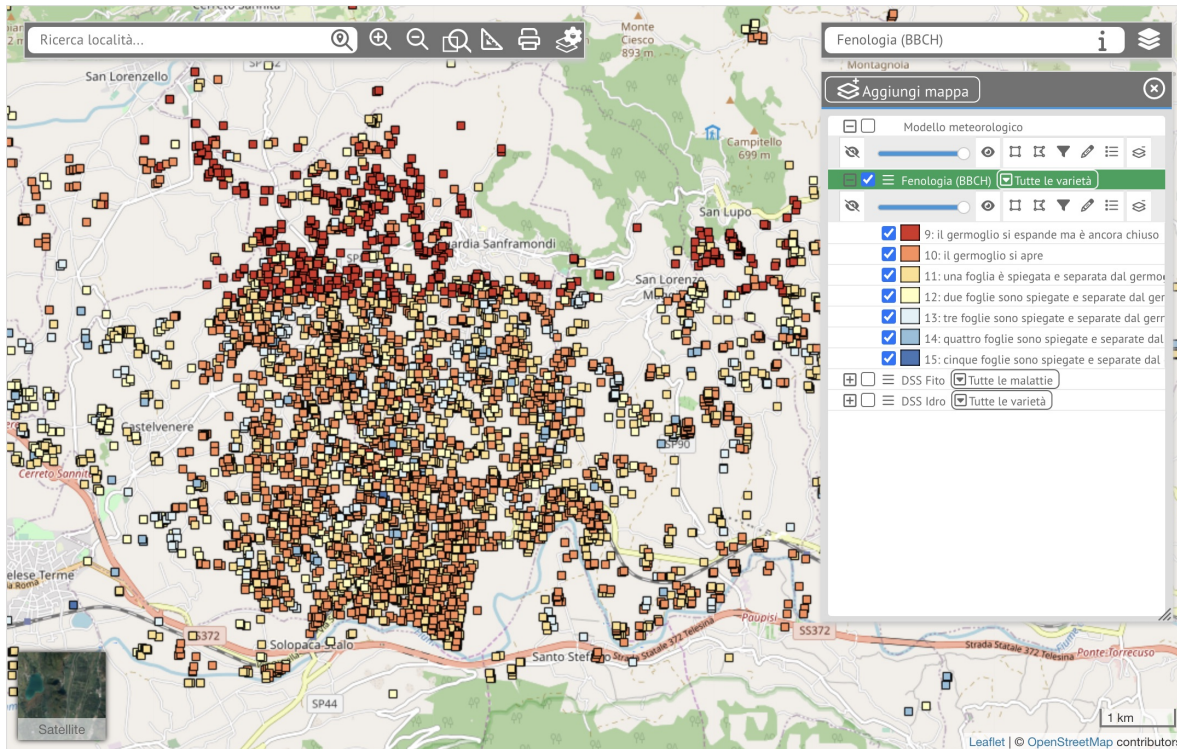


**Figure 6.1:** Detail view of observed (blue and green lines) and predicted phenologies (dotted gray line) for a sample field during the 2021 agricultural season. The predicted values closely follow the observed ones. Around May 31<sup>st</sup>, the phenological model is late compared to the observed value.

## 6.1. Phenology DSS

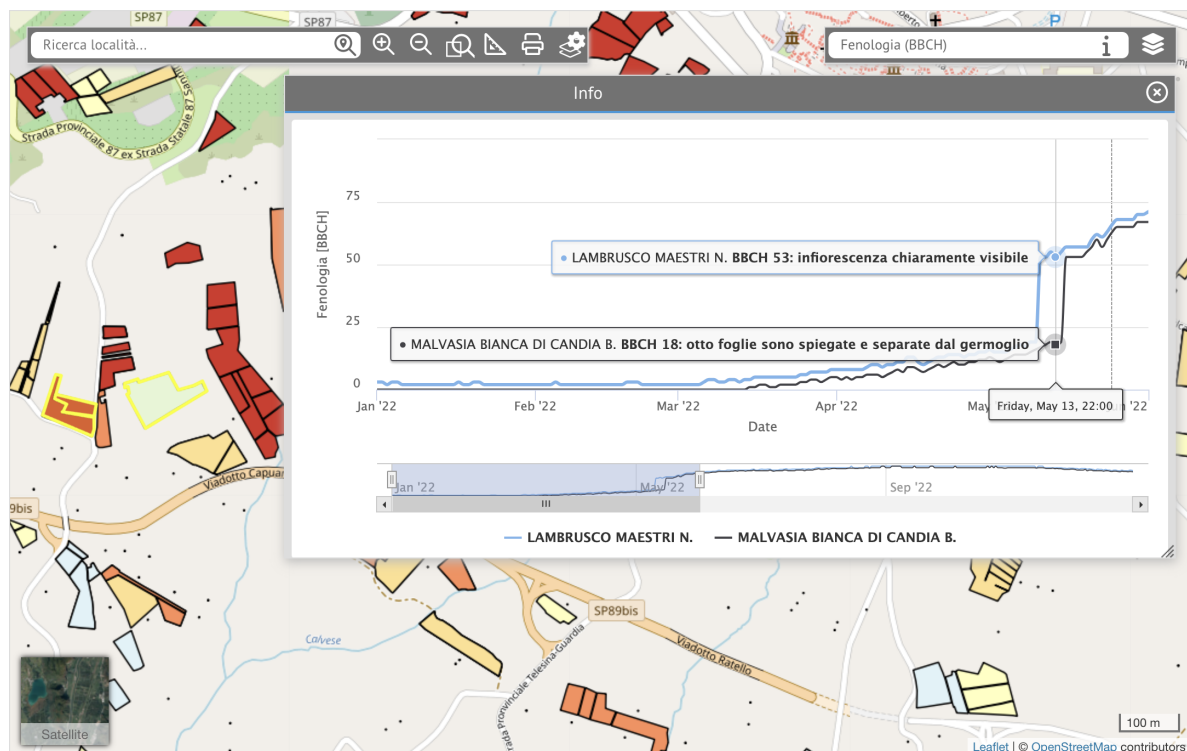
clarity, vegetative and reproductive phases are represented in two distinct colors (blue for vegetative and green for reproductive), making it easier to differentiate between the phases, especially in more complex scenarios. The phenology model generates daily BBCH estimates, depicted by a continuous gray dotted line. Overall, the sheet enables users to document both vegetative and reproductive phases, allowing for the recording of the vegetative phase on the same date or later than the reproductive phase and vice versa. By treating the two groups separately, we obtained a visualization that allows for temporal overlap without creating polygonal lines that transition between the two phases. For instance, the figure illustrates two values (one vegetative and one reproductive) sampled on the same date (May 19, 2022). The model was configured to output a single phenology, leaning towards predicting reproductive ones over vegetative ones, and it does not use past observed phenologies as input. Around the considered date, the model output phenology values for both classes on consecutive days, indicating a degree of uncertainty regarding which phase was perceived as the more relevant.

Next, the phenology DSS was integrated into the Enogis WebGIS platform. At the



**Figure 6.2:** Testing area (spanning 15 km in width and 10 km in height) located in southern Italy, with BBCH values displayed for April 21, 2022. On the right side, the layer list showcases all available DSS, with the phenology layer labeled "Fenologia BBCH". This layer includes a dropdown menu to filter by cultivar (by default, all cultivars are selected).

core of the Enogis relational system is the field, which is typically linked to geographical data that delineates its boundaries. By displaying the phenological values of all fields on a map, users can easily observe differences and assess the current status of crop development (see Figure 6.2). Fields are represented with square symbols when the zoom level is too low, while actual field contours are shown at higher zoom levels. The phenology DSS can be activated by enabling the corresponding layer, labeled "Fenologia BBCH," in the layer list. The legend displays the current phenology (updated daily) values for all the georeferenced fields in the app. Legend colors always follow the spectral (i.e., red to blue) color scheme, while displayed codes are updated dynamically throughout the season. By default, all cultivars are selected, but the user can filter only for a specific cultivar by selecting it in the dropdown menu. For example, the figure displays the resulting values for April 21, 2022, with BBCH values ranging from 9 (the bud expands but is still closed) to 15 (five leaves unfolded and spread away from the shoot) (Figure 6.2). A cluster of red fields (BBCH 9) is located towards the top of the map, mainly composed of Falanghina and Malvasia Bianca varieties. Other patterns



**Figure 6.3:** Feature info box displaying the BBCH trends for two selected fields (yellow contours, left side of the picture) for the 2022 agricultural season. The area spans 1.9 km in width and 1.2 km in height. Fields belong to the Lambrusco (blue) and Malvasia Bianca (black) varieties.

can be identified when interacting with the map.

By selecting one or more fields, an information box appears (Figure 6.3), which displays the predicted phenology for both past and future days. The prediction horizon varies depending on the specific weather forecast system utilized. Phenological data is represented through a line chart, with each line corresponding to a selected field. The example figure shows the phenology trend for the 2022 agricultural season for two Lambrusco and Malvasia Bianca fields. The bottom navigator facilitates quick navigation through the entire season, providing an overview of the general trend. The chart is interactive, allowing users to view phenological BBCH codes and descriptions by hovering over the lines. From the figure, the difference in crop development can be appreciated for the two fields: the model suggests that the Lambrusco field develops faster compared to the Malvasia Bianca one. The plot displays up to 10 phenology lines and can be used to compare fields over various dimensions (e.g., variety and geographical positions).

## 6.2 Disease coverage DSS

The disease coverage DSS provides users with valuable information and recommendations regarding the defense actions to be implemented in the field. It takes into account various factors, including historical and forecasted weather conditions, grapevine phenology, and the coverage curves of active ingredients found in pesticides. Rather than directly estimating the risk of infections or infestations, the DSS models the effectiveness of agro-pharmaceutical products and assesses when their protective effects diminish. This approach informs users about potential risks, allowing for more proactive management of crop health. One of the primary objectives of the DSS is to reduce the volume of agrochemical usage, thus enhancing both economic and environmental sustainability. By optimizing defense strategies, the system aims to promote healthier vineyards while minimizing the ecological footprint associated with agricultural practices.

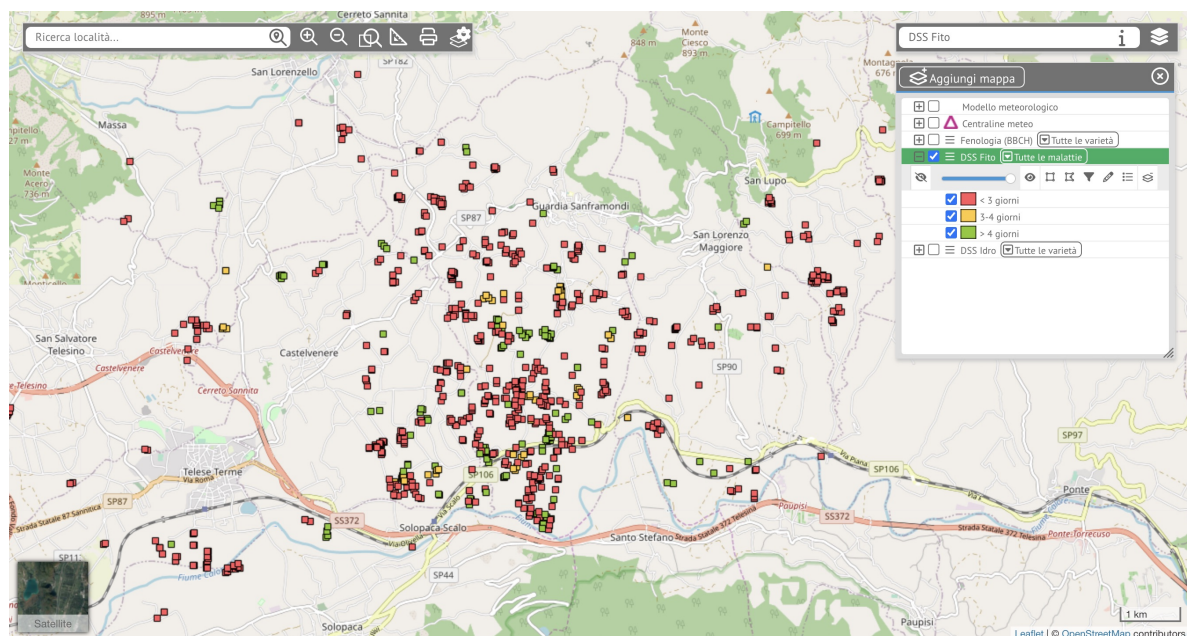
The first version of the DSS was developed as part of the EnoPhit project, an innovation effort led at Enogis before the beginning of the PhD program. The project focused on the adaptability of the DSS to the Trento province (Trentino-South Tyrol region, North-East Italy) and involved the creation of agronomically decision trees for treatment strategy planning, addressing a total of 11 diseases, with gray mold (*Botrytis cinerea*), downy mildew, and powdery mildew being the primary concerns. The EnoPhit DSS provided users with information and recommendations at the field level, featuring chart visualizations to monitor season dynamics such as crop development,



weather conditions, and pesticide coverage. During the PhD, efforts were devoted to enhancing the systems's capabilities. This included implementing the connection with the EnogisAI data platform, improving data visualization at the field scale, expanding the geographical applicability of the DSS to all the Italian territory, integrating the DSS with the WebGIS tool for comprehensive monitoring, and developing an agrochemical treatment simulator. This simulator allows for the comparison of field-implemented strategies with those suggested by the system, allowing for a more critical evaluation of the potential benefits associated with adopting more informed decision-making processes.

The disease coverage DSS was integrated into the WebGIS tool, similarly to the phenology DSS. The WebGIS interface allows users to filter data based on specific adversities and crop varieties. This functionality enables users to assess pesticide coverage for the upcoming days, providing a clear visual representation of future conditions on the map.

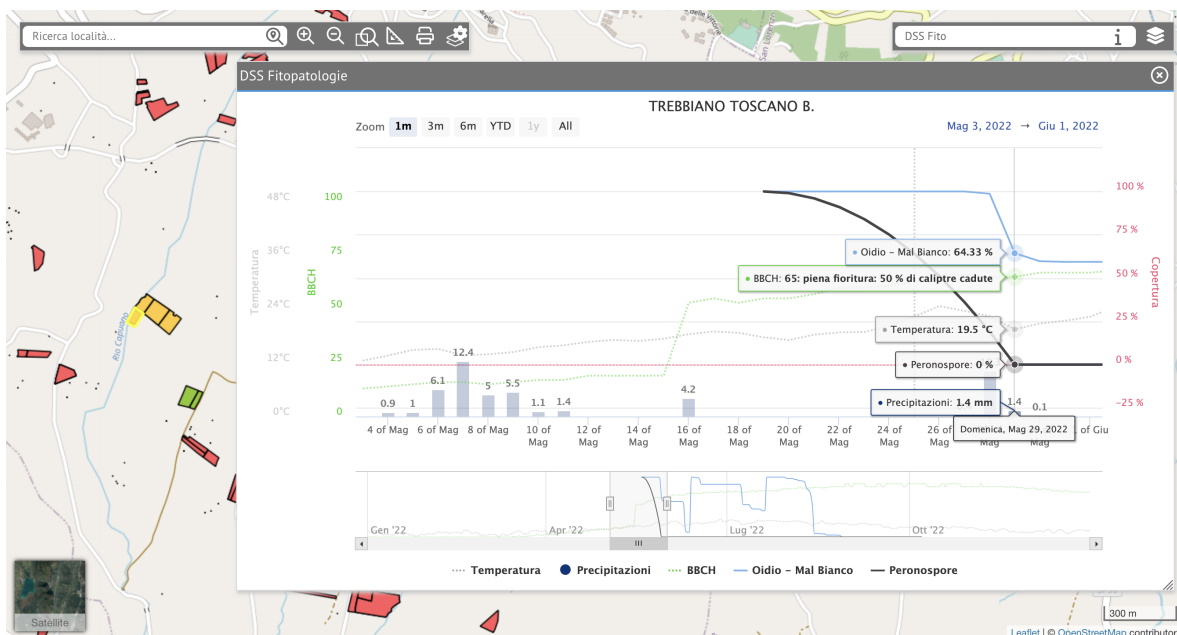
When the disease coverage feature of the DSS is activated, each field is visually represented with colors indicating the number of days it is protected against specific diseases (Figure 6.4). Red is used if the coverage is less than 3 days, yellow for 3 or 4 days, and green for more than 4 days. The WebGIS integration considers the three most



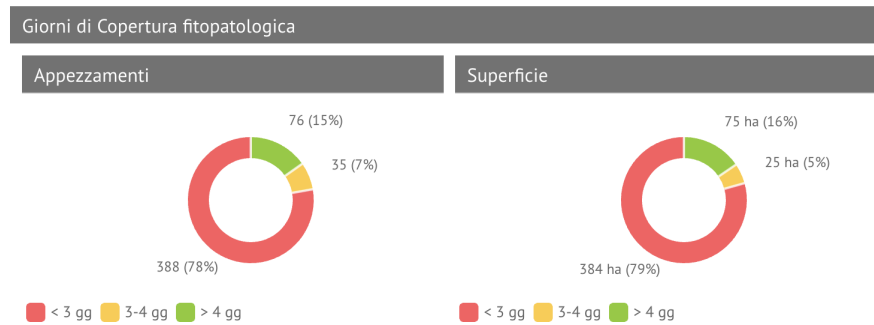
**Figure 6.4:** Testing area located in southern Italy, with coverage values displayed for May 25, 2022. The three color classes encode the number of days fields are protected against diseases: less than 3 (red), between 3 and 4 (yellow), or more than 4 days (green).

relevant diseases addressed in the EnoPhit project: gray mold (*Botrytis cinerea*), downy mildew, and powdery mildew. By default, fields are categorized into corresponding color classes based on the coverage of all three diseases, specifically identifying the disease with the least coverage. For example, a red color indicates that the coverage of at least one disease is less than three days, even if coverage for one of the other diseases extends beyond that duration. Additionally, a dropdown filter allows users to select one of three diseases, enabling the reclassification of each field according to the selected disease.

Clicking on a single field in the map shows season-specific trends for the selected unit (Figure 6.5). The chart displays key factors such as weather conditions, grape development, and disease coverage. Precipitation is represented by a bar chart (bottom of the main chart), while daily average temperature is depicted with a gray dotted line (first left y-axis). Phenological values are shown with a dotted green line (second left y-axis), and disease coverage lines are shown with solid lines (values are reported from 0 to 100%, right y-axis). The future scenario horizon extends up to 10-14 days, depending on the availability of weather forecasts. The chart is interactive, allowing users to enable or disable specific data series and adjust the time range for a more focused or broader view by changing the time range. Finally, dragging the mouse over the chart reveals detailed values of each series for each date. The future scenario horizon extends up to 10-14 days,



**Figure 6.5:** Feature info box displaying the disease coverage DSS as of May 25, 2022. The selected field (left side of the figure) is yellow, thus associated with 3-4 days of coverage. This is an analysis of a past season; for this reason, the timeline chart displays coverage values for the rest of the season.



**Figure 6.6:** Example of dashboard summarizing DSS output stats in terms of fields (*appezzamenti*) and surface (*superficie*). The three color classes encode the number of days fields are protected against diseases: less than 3 (red), between 3 and 4 (yellow), or more than 4 days (green).

depending on the availability of weather forecasts. The chart is interactive, allowing users to enable or disable specific data series, adjust the time range for a more focused or broader view, and access pop-ups for detailed information. This functionality enhances user engagement and facilitates a comprehensive understanding of the data presented. The example field shown in the figure is classified as covered for 3-4 days (yellow class) (Figure 6.5). The chart illustrates that coverage for downy mildew (*peronospora* in the chart, black line) is expected to reach 0 on May 29, 4 days from the considered date (May 25, vertical gray dotted line) (Figure 6.5). On the other hand, powdery mildew coverage (*oidio - mal bianco*, blue line) is expected to last longer, probably due to the nature of the utilized product. The bottom navigator and the zoom buttons at the top allow users to quickly and easily navigate the DSS's temporal dimension.

Given the amount of information shown, this DSS does not support comparing trends across multiple fields within a single chart visualization. However, users can view multiple charts simultaneously by selecting several fields. Each chart typically occupies a maximum of one-fourth of the total application window on a standard workstation, allowing for a clear and organized presentation of data.

In addition to the integration into the WebGIS, the DSS was developed to provide additional information in other tabular views of the Enogis application and specifically into the Enogis field notebook. The field notebook tool is used for recording and planning disease treatments, and the insights derived from the DSS can significantly support end users in their decision-making processes. First, the disease coverage DSS outputs are represented in a summarizing dashboard (Figure 6.6). The dashboard shows the coverage in terms of fields (*appezzamenti*) and total cultivated surface (*superficie*). The dashboard can be clicked to populate a filter in the list of all agricultural companies

## 6.2. Disease coverage DSS

|                          | Publicaz.  | Riferimento | Iniziato? | Sbloccato? | Codice | N.Certificati | Ultimo trattam. | Conforme?  | PDF | DSS Fito |
|--------------------------|------------|-------------|-----------|------------|--------|---------------|-----------------|------------|-----|----------|
|                          |            |             |           |            |        |               |                 |            |     |          |
| <input type="checkbox"/> | 31/01/2023 | attuale     |           |            | 002527 | 0             |                 | 10/08/2022 |     |          |
| <input type="checkbox"/> | 31/01/2023 | attuale     |           |            | 002410 | 0             |                 | 30/07/2022 |     |          |
| <input type="checkbox"/> | 31/01/2023 | attuale     |           |            | 002944 | 0             |                 | 08/08/2022 |     |          |
| <input type="checkbox"/> | 31/01/2023 | attuale     |           |            | 002583 | 0             |                 | 24/07/2022 |     |          |
| <input type="checkbox"/> | 31/01/2023 | attuale     |           |            | 002262 | 0             |                 | 31/07/2022 |     |          |
| <input type="checkbox"/> | 31/01/2023 | attuale     |           |            | 002544 | 0             |                 | 09/08/2022 |     |          |
| <input type="checkbox"/> | 31/01/2023 | attuale     |           |            | 002296 | 0             |                 | 30/07/2022 |     |          |
| <input type="checkbox"/> | 31/01/2023 | attuale     |           |            | 002785 | 0             |                 | 27/07/2022 |     |          |
| <input type="checkbox"/> | 31/01/2023 | attuale     |           |            | 002303 | 0             |                 | 03/08/2022 |     |          |
| <input type="checkbox"/> | 31/01/2023 | attuale     |           |            | 002896 | 0             |                 | -          |     |          |
| <input type="checkbox"/> | 31/01/2023 | attuale     |           |            | 002200 | 0             |                 | 17/07/2022 |     |          |
| <input type="checkbox"/> | 31/01/2023 | attuale     |           |            | 002331 | 0             |                 | 11/08/2022 |     |          |
| <input type="checkbox"/> | 31/01/2023 | attuale     |           |            | 001924 | 0             |                 | 24/08/2022 |     |          |
| <input type="checkbox"/> | 31/01/2023 | attuale     |           |            | 002836 | 0             |                 | 02/08/2022 |     |          |
| <input type="checkbox"/> | 31/01/2023 | attuale     |           |            | 002086 | 0             |                 | 24/07/2022 |     |          |
| <input type="checkbox"/> | 31/01/2023 | attuale     |           |            | 001627 | 0             |                 | 21/07/2022 |     |          |
| <input type="checkbox"/> | 31/01/2023 | attuale     |           |            | 002261 | 0             |                 | 12/08/2022 |     |          |

**Figure 6.7:** Enogis field notebook tool showing a selection of agricultural companies present in Enogis for the 2022 agricultural season. The last column (*DSS Fito*) shows an indicator summarizing the status of all fields belonging to each company. The three color classes encode the number of days fields are protected against diseases: less than 3 (red), between 3 and 4 (yellow), or more than 4 days (green).

in the application, for example, to select those that have coverage that is lower than three days (red class).

The DSS output can be used to improve the visualization of the summary list of all agricultural companies in the Enogis app (Figure 6.7). This list recaps general information regarding each agricultural company, which is usually comprised of multiple fields. The last column of the table (*DSS Fito*) indicates the most critical status among all farm fields, using the same color classification as used in the map. Therefore, if a farm is associated with a red indicator, it means that at least one field has low coverage (less than three days) for at least one disease. Similarly, a yellow indicator refers to 3-4 days of coverage, and a green one indicates coverage of more than 4 days. Further, by clicking on a list item, the app displays the field notebook data for all fields of the selected company. Among the available views, the list of fields integrates the DSS indicator for each field. The indicator can assist users in quickly identifying farms and fields that are potentially at risk and proactively planning field activities (i.e., treatments) to improve crop defense.

In addition to displaying actual treatments and their estimated coverage, we developed a simulator to compare real versus simulated treatment strategies (see Figure 6.8). The chart contains the coverage estimates for each disease, with actual treatment coverage represented by solid lines and simulated treatment coverage shown as dashed lines (treatments cause the coverage to reach 100%). Simulations are based on the agro-



**Figure 6.8:** Downy mildew simulation for an example field. The solid line (violet) represents the coverage associated with actual treatments, while the dashed line (fuchsia) is associated with simulated treatments. Treatments are associated with days where the coverage is maximized to 100% (e.g., on May 28<sup>th</sup>). In this case, the number of actual treatments is 7, while the number of simulated treatments is 3.

conomic rules defined in the context of the EnophitDSS project. The first treatment date corresponds to the date of the initial real treatment, while the latest coverage date is set to an arbitrary date based on the specific disease being simulated and its reference period.

Agrochemicals lose their efficacy due to run-off from rainfall or absorption by the crop. When rain occurs, it can wash away applied treatments, reducing their concentration and effectiveness in controlling pests or diseases. Additionally, crops can absorb these products, further diminishing their availability in the environment where they are needed to act. The first products used in the considered example are modeled to lose efficacy due to absorption (note the quadratic decay curve), while from the third treatment on, efficacy is mainly lost due to run-off (note the decrease in coverage associated with precipitation presence, indicated by vertical bars). The simulator tends to use the same products that farmers employ, as its objective is to suggest a more efficient use of available agrochemicals.

In the specific case, in the second part of the season, the real treatments (total of 7) outnumber the simulated ones (total of 3). This difference could be associated with a wrong estimation of the rainfall intensity either by the model or by the farmer. In the former scenario, the model increases the risk of infection, as it may not effectively manage pest and disease pressures, leading to higher susceptibility among crops. Contrarily, the latter scenario results in economic waste and contributes to additional environmental pollution. This occurs when agrochemicals are applied inefficiently, either through overuse or inappropriate timing, which not only diminishes their effectiveness but also

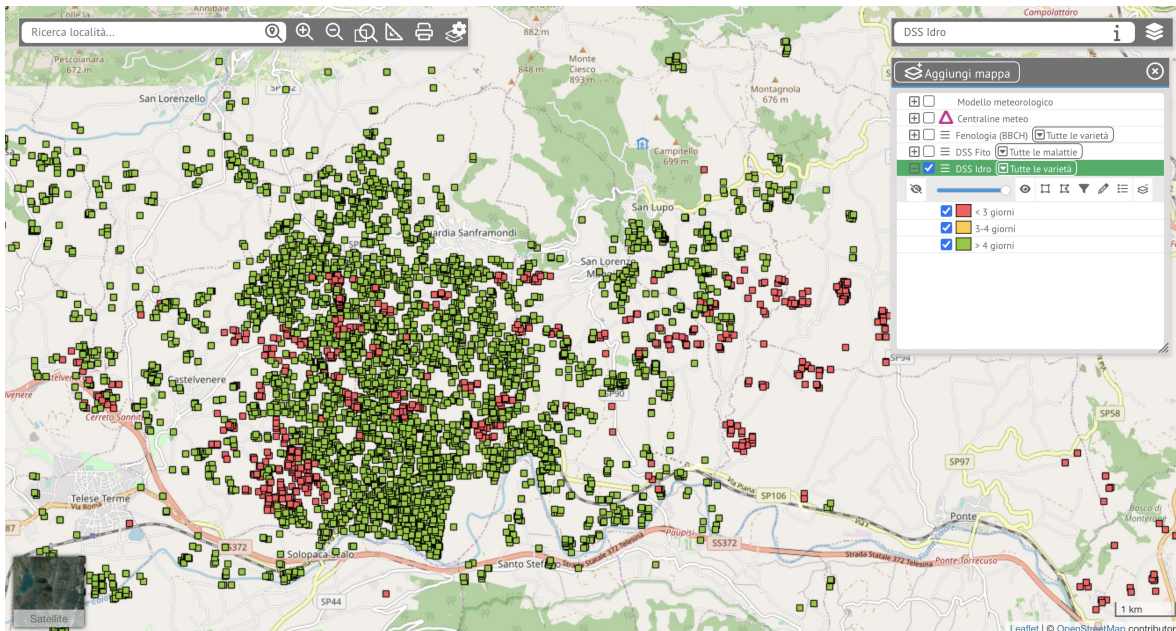


leads to runoff and contamination of surrounding ecosystems.

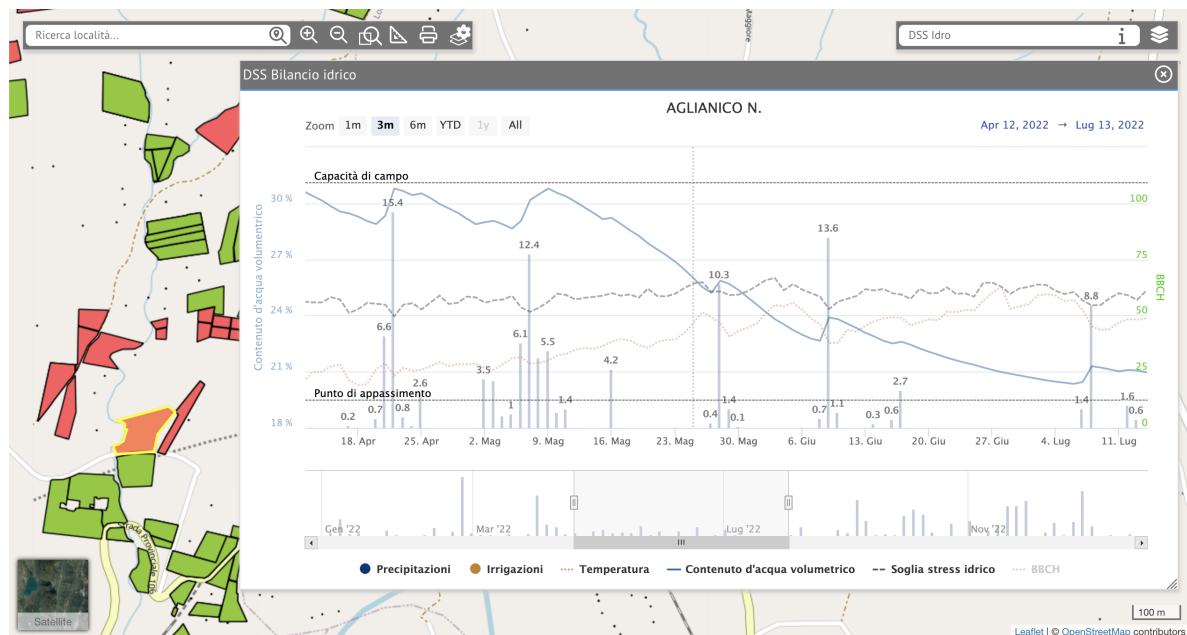
## 6.3 Soil water balance DSS

The soil water balance DSS is based on the FAO56 notebook implementation detailed in Section 5.4. The DSS provides users with valuable insights into soil water content and crop stress levels, which are influenced by various factors, including weather conditions, crop development stages, and soil hydraulic properties. The primary objective of the DSS is to optimize irrigation planning and execution, enabling better management of crop stress levels in alignment with agronomic goals and ensuring efficient water use.

The soil water balance DSS followed an integration process similar to that of the disease coverage DSS and was integrated into the Enogis WebGIS and field notebook tools. This integration enhances the accessibility and usability of the system, allowing users to effectively monitor and manage soil water dynamics within their agricultural practices. The disease coverage DSS worked by estimating agrochemical coverages after the execution of a treatment. Differently, the soil water balance DSS does not build insights on top of user inputs or agricultural activities. Instead, irrigation activities in the system are included to compute the water balance.



**Figure 6.9:** Testing area located in southern Italy, with soil water status values displayed for May 25, 2022. The three color classes encode the number of days to reach the crop stress threshold: less than 3 (red), between 3 and 4 (yellow), or more than 4 days (green).



**Figure 6.10:** Feature info box displaying the disease coverage DSS as of May 25, 2022. The selected field (left side of the figure) is red, thus associated with less than 3 days to reach the stress threshold. This is an analysis of a past season; for this reason, the timeline chart displays the soil water balance values for the rest of the season.

In terms of WebGIS integration, the DSS offers both a comprehensive overview and detailed insights for individual fields. By enabling the DSS layer (referred to as *DSS Idro*), users can visualize each field classified into three distinct color categories (Figure 6.9). The classification is based on the number of days until the available water content reaches the stress threshold, following the same scheme defined for the disease coverage DSS: less than 3 days (red), 3-4 days (yellow), and more than 4 days (green). Additionally, the DSS allows users to filter fields by variety, enabling them to focus on those varieties that may be more sensitive to soil water stress conditions.

By clicking on a specific field, users can view a seasonal chart for that field (Figure 6.10). This chart provides a comprehensive visualization that includes:

- Precipitation (blue bars),
- Irrigation events (brown bars, if present),
- Temperature (red dotted line),
- Volumetric soil water content (solid blue line),
- Stress threshold marked (black dashed line),
- Optionally, the phenology (green dotted line, when enabled).

Additionally, field capacity and wilting point are illustrated as horizontal black lines, representing the maximum and minimum values of the volumetric soil water content. Since these two parameters are solely dependent on hydraulic field properties, they remain constant throughout the seasons. The example chart displays trends of the whole agricultural season since it belongs to a past agricultural season. In an operational context, the DSS forecast horizon depends on the availability of weather forecast days (i.e., 10-14). Similar to the disease coverage DSS, the chart does not allow for direct comparisons of DSS trends between two fields due to the volume of information presented. However, users can select multiple fields within the Enogis application to open them simultaneously, enabling manual comparison of the data.

## 6.4 Interactive data exploration tool for enhanced decision-making

Finally, we present a more comprehensive approach designed to integrate DSS outputs and other generated data products into the WebGIS platform. Specifically, we developed an Enogis module that allows for classifying the territory based on manually defined rules that operate over DSS outputs and trends, as well as geographical data products produced with data analysis pipelines. Users have the option to define the classification categories based on their final goal. Examples are the identification of specific climate conditions (e.g., annual average temperature below a certain threshold) and the emergence of particular dynamics for activity planning (e.g., forecasted soil water balance below a certain threshold for specific soils or varieties). After defining the categories, users can select among several data products, such as internal (Section 3.1) and external (Section 3.2) data sources, derived data products (Section 3.3), or the DSS outputs described in this chapter. The employed data can either be static, such as environmental descriptor, or dynamic, such as continuously updated DSS or similar analysis. The primary requirement for any data is that it must include georeferenced information, meaning that features are linked to specific geographical elements, such as points or polygons. Finally, users combine the different data sources by defining decision tree rules for all or a selected subset of the considered data. The output of the decision tree is a category defined by the user for each of the geographical intersections of the input data, effectively generating a new map.

An interesting aspect of this DSS is its capability to analyze not just the areas defined by field boundaries but also entire regions, such as a designated DOC area.



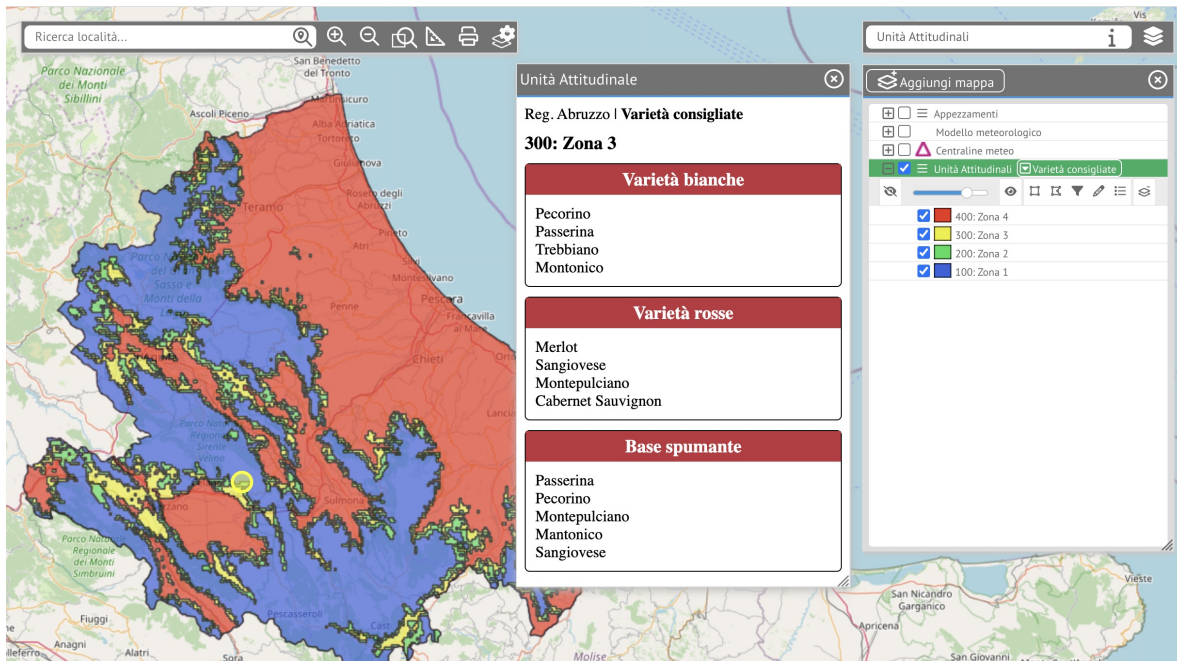
Therefore, the tool can be used to evaluate future winery development planning, for example, supporting the decision process of where to establish new vineyards or to adapt to climate change. This flexibility allows the tool to simplify and improve the evaluation of future vineyard development plans. For instance, it can support decision-making regarding the optimal locations for establishing new vineyards and adapting strategies in response to climate change. By considering broader territorial contexts, the DSS enhances strategic planning and contributes to sustainable practices within the viticulture sector.

Here, we present two applications that involve the use of the interactive exploration DSS. The first one was developed on the geographical domain of the Abruzzo region to identify and suggest the most suitable varieties to be planted for new vineyards. The rule-based system uses the Winkler climatic index as the only input variable and groups varieties into the white, red, and sparkling wine musts (referred to as *base spumante*) groups.

The Winkler index computes the growing degree days (GDDs) over a base temperature from April to October, as explained in Section 3.3.2. A total of four classes were defined over the index, and the most suitable cultivars were associated with each class (Figure 6.11). The classification is based on the following rules:

- **Winkler < 1400** - white varieties: Chardonnay, Pinot blanc, Pinot Gris, Sauvignon; red varieties: Pinot Noir;
- **1400 < Winkler < 1600** - white varieties: Chardonnay, Pinot blanc, Pinot Gris, Sauvignon, Pecorino, Passerina; red varieties: Pinot Noir, Merlot, Sangiovese; sparkling wine musts: Chardonnay, Pinot Noir, Passerina, Pecorino, Montepulciano, Mantonico, Sangiovese, Trebbiano;
- **1600 < Winkler < 1800** - white varieties: Pecorino, Passerina, Trebbiano, Montonico; red varieties: Merlot, Sangiovese, Montepulciano, Cabernet Sauvignon; sparkling wine musts: Passerina, Pecorino, Montepulciano, Mantonico, Sangiovese;
- **Winkler > 1800** - white varieties: Trebbiano, Montonico; red varieties: Sangiovese, Montepulciano, Cabernet Sauvignon

The result of the classification is then displayed interactively on the WebGIS tool (Figure 6.11). Areas belonging to different areas are easily identifiable through color coding, and the map legend allows one to select only specific classes to display on the map. By clicking on an area (yellow circle in the figure), the WebGIS tool will display

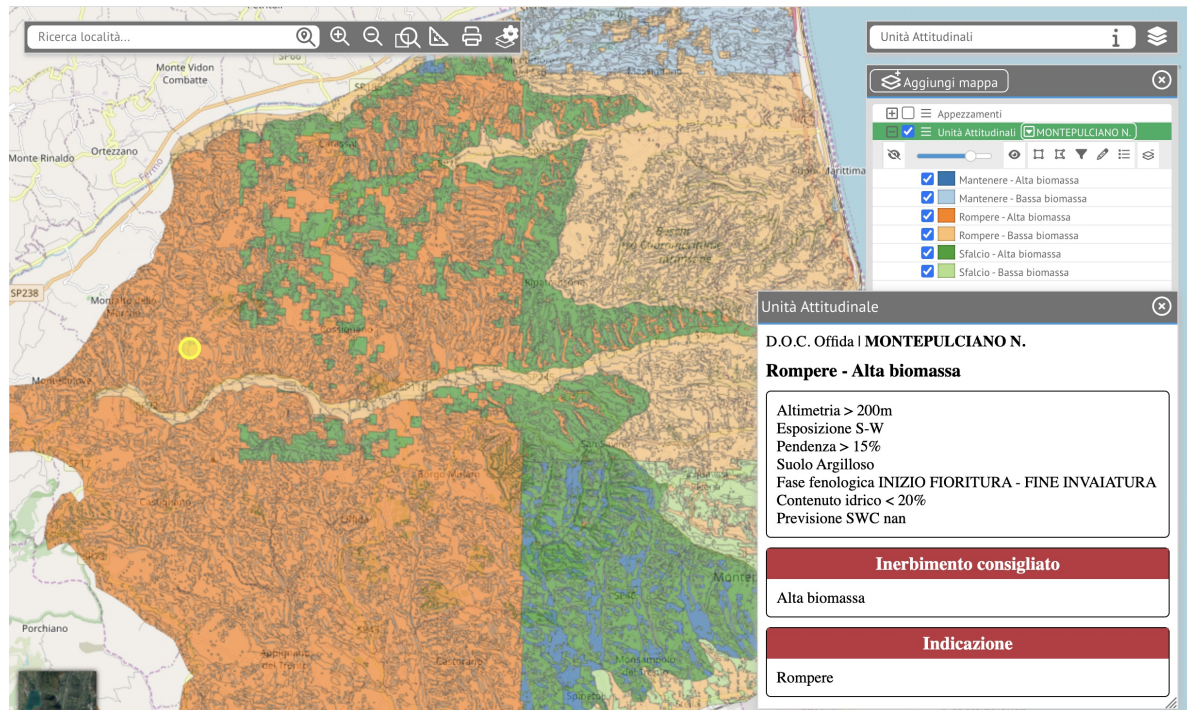


**Figure 6.11:** Classification results using the Winkler index as input. Categories and rules are defined by the users as follows: Winkler<1400 (blue), 1400<Winkler<1600 (green), 1600<Winkler<1800 (yellow), Winkler>1800 (red). The info box shows the values associated with the selected yellow area. The top part displays the input values of the DSS, and the three bottom boxes show the output of the DSS.

the most suitable varieties for that class category.

The second application focused on improving the decision-making processes relative to the adoption of grass biomass and the management of soil for the Offida DOC area in the Marche region (central Italy). Implementing grass biomass not only contributes to soil fertility but also promotes biodiversity, which is essential for maintaining a balanced ecosystem within the vineyard. By focusing on effective soil management practices, vineyard managers can improve soil health, which in turn supports better grapevine growth and resilience against environmental stresses.

The output of the DSS consists of two outputs: first, a binary classification that indicates what biomass level to adopt (low or high) for vineyard soil management; second, the indication of how to manage the biomass level among removal (*rompere*), maintain (*mantenere*), or cutting (*sfalcio*) (Figure 6.12). The DSS output is based on elevation (above or below 200 m), aspect (North-East vs. South-West), slope (steeper or gentler than 15%), soil type (clay vs sand), the phenology phase (i.e., pre-flowering, from flowering to veraison, and from ripening to end of harvesting), soil water content (lower than 20%, between 20% and 40%, and greater than 40%), and the forecasted



**Figure 6.12:** DSS example output for July 20, 2024. The map displays six classes resulting from the combination of two biomass adoption levels and three soil management suggestions. The area refers to the Offida DOC region. The application displays a feature info box for the "remove - high biomass" class (*rompere - alta biomassa* in the figure, dark orange), displaying the values of the input data (top box) and the DSS outputs (two bottom boxes).

trend of soil water content over the next week (increasing, decreasing, or equal).

In the presented figure, which refers to July 20, 2024, the "remove - high biomass" class is the prevalent one (Figure 6.12). By clicking on the map, a feature info box displays the information regarding the values of the input data, as well as the suggestion generated by the DSS. The DSS is updated on a weekly basis to reduce the amount of variability among subsequent days. Grapevine development is highly dependent on the variety, thus the DSS output maps are generated for four varieties among the most relevant for the area, namely Montepulciano, Passerina, Pecorino, and Sangiovese.

This integration streamlines data access and fosters a more holistic understanding of spatial information, empowering users to make informed decisions based on a broader context.

## 6.5 Technical implementation details

The Enogis technical ecosystem comprises multiple physical HW nodes that support multiple virtual machines (VMs). Enogis offers its services through different subscription plans, including a premium option that includes a dedicated platform installation, enhanced technical support, customized features, and advanced user management protocols. Considering these factors, we decided to develop the DSS systems as a centralized entity hosted on a dedicated VM. This setup enabled the EnogisAI DSSs to efficiently compute and serve data to all other software instances via API communication.

The first component of the DSS system is the TimescaleDB instance of the *Measurements* project described in Section 3.4.1. TimescaleDB is an efficient DBMS optimized for storing, retrieving, and computing stats of time-series data. Within the EnogisAI data platform, it collects all weather data generated by weather models or collected from weather stations. The DSS server has a direct connection to the weather database, which serves as a central ingredient powering all DSS computations.

Next, additional environmental descriptors, such as solar irradiation maps and bioclimatic indices, are accessed directly from the EnogisAI data platform. Specific field data related to irrigation, agrochemical treatments, and fertilization are synchronized nightly from all Enogis software instances configured to utilize the DSS. For the 2024 agronomic season, the system ingested 121450 fields relative to 28000 agricultural companies. The phenology DSS was activated for 118324 fields, while the disease coverage and soil water balance DSSs were activated for 37486 and 61453 fields, respectively. The disease coverage DSS was elaborated on 96496 treatments. All data synchronization and DSS computation processes are scheduled and executed in a distributed manner with Celery and Celery Beat [126]. Celery permits the writing of complex workflows to be executed in distributed and parallel environments. These workflows define the dependency order among data synchronization, field characterization with additional environmental data, model initialization, and model execution.

The DSS can be further extended with additional external computing backends. One notable example is the deployment of the phenology deep learning model. This phenology network is implemented as a standalone system to increase interoperability within the broader ecosystem. For these reasons, we adopted TorchServe [157], a flexible tool for serving and scaling PyTorch models in production. We configured TorchServe to scale the phenology model to a maximum of 10 instances, aiming to maximize prediction throughput and effectively handle the request volume generated by multiple Celery task executors.

Once the computation phase is complete, users can access the DSS results directly through the Enogis app, utilizing the map, chart, and tabular visualizations previously detailed in this chapter. To manage user authentication and authorization across different systems - specifically between Enogis deployments and the EnogisAI DSS - we implemented a robust mechanism. First, the DSS system collects and stores the authorization levels for each user across various systems, including access rights related to Enogis groups, agricultural companies, and specific fields. Second, every API call is secured with a JSON Web Token (JWT), which verifies the validity of the API request, its origin (i.e., Enogis app instance), and the permitted data scope. Communication between systems occurs in a server-to-server manner.



# Chapter 7

## Conclusions

The thesis described EnogisAI: a comprehensive framework to support and streamline the development and adoption of agronomic AI solutions in daily agricultural and agronomist activities. The framework is structured into four high-level components: data ingestion and exploration, algorithm development, model training and evaluation, and integration into DSSs.

First, we presented the data platform that integrates, describes, and links internal (i.e., company-owned and generated data) and external geographical and environmental data to support multi-modal modeling tasks. The EnogisAI data platform includes software components that provide access to data through an API interface and serve data to the entire Enogis development and production ecosystem.

Second, we developed models for various agricultural tasks, including crop development monitoring, crop management, yield prediction, soil water assessment, and time series data reconstruction and forecast. These models were the result of research activities led in collaboration with universities and the international research community. This underlines the opportunities deriving from private and public collaboration in the development of AI solutions in agronomics. The research presents several innovative aspects that significantly improve the management of crops:

- **multi-modal data utilization:** the integration of diverse data sources for phenology estimation is a key innovation. This approach leverages various types of data, such as climate, environmental, and crop-related variables, to provide a more comprehensive view of phenological changes.
- **advanced processing pipelines:** Functional Principal Component Analysis was successfully combined with clustering and regression algorithms to effectively improve crop management and yield prediction, allowing for more accurate assess-

---

ments of agricultural performance.

- **machine and deep learning applications:** we applied machine learning and deep learning techniques for reconstructing and forecasting daily pollen concentration. Further, we proposed an analytical tool to explain the dynamics of modeling predictions, enhancing our ability to understand and predict ecological interactions.

We focused on creating models that can be horizontally scaled and applied across the entire Italian context. This approach involved using data sources with extensive geographical coverage, such as European soil databases, global hydrologic variables, and national digital terrain models. One of the primary goals of EnogisAI is to establish baseline models that can be applied to national contexts, enabling later integration of higher spatial resolution data to improve performance for localized use cases. Some of the proposed methodologies were published as conference proceedings and research journal articles, following an open innovation philosophy.

Finally, a selection of the models was integrated into operational DSS solutions employed daily by farmers and agronomists. These systems enable users to make informed, data-driven decisions, enhancing agricultural practices and improving overall efficiency in the field. Specifically, we developed systems for monitoring grapevine growth, managing disease coverage, planning agrochemical applications, and assessing soil water and associated crop stress. The DSSs were integrated into the Enogis software, allowing end users to optimize their field operations and improve economic and environmental sustainability.

All the AI pipeline development phases were conducted within the Enogis ecosystem, which impacted all innovation processes on the technical side and on the research and development one. Regarding the technical side, it influenced all technological choices, as the final objective of all activities is their seamless integration with production environments. Moreover, the availability of specific agronomic datasets derived from Enogis applications, along with the interest of end users, guided EnogisAI in addressing the previously described tasks. This led to the design of solutions that are both practical and interpretable for users. By focusing on the needs of the final users, we aim to increase the adoption and impact of our solutions in real-world agricultural settings.

The industrial context in which EnogisAI was developed highlights the necessity for innovation-driven research that can ultimately provide additional value to existing applications. This innovation landscape restricts the search space of solutions to those that are compatible with the already existing ecosystem. However, adopting a deep-tech



innovation approach, such as developing and adopting new AI tools, requires a greater tolerance for risk than developing low-risk incremental features. This willingness to embrace risks is essential for driving significant advancements and changes in technology, leading to transformative shifts in companies and the broader industry. For these reasons, we stress that EnogisAI is a comprehensive AI development framework that can be replicated or extended with other private and public entities to bring additional value to the agriculture sector.

At the same time, having developed the research in an industrial and operational sector could be considered as a limitation for the overall study. The reason is that limited resources can be devoted to addressing relevant industrial tasks by experimenting with fundamental research results from largely different fields. Instead, industrial research and innovation processes focus on technologies that proved effective in similar tasks or areas, thus having a higher potential to positively influence the agricultural sector.

The validation aspects of the modeling techniques, and especially those of the soil water balance, may need further and more extended validation, considering other climatic and environmental conditions. Additionally, the degree of variability associated with open environments (e.g., agricultural fields) underscores the need to evaluate the proposed systems under new, use-case-specific conditions. Obtaining scarce validation results may require adapting the developed solutions via model calibration or fine-tuning, which may impact reusability, especially in contexts with limited historical data availability.

The final goal of the thesis is to improve the economic and environmental sustainability of agricultural production and to sustain the growing population by improving yield quality and quantity. To this end, in addition to the technical and technological innovations of EnogisAI, we underscore the process innovation associated with the framework. These can be summarized as integrating, considering, and analyzing multiple aspects of AI pipelines, particularly the importance of effectively connecting available data with end users' needs through the development of valuable and efficient smart farming tools. The role of these tools is to support farmers and agronomists in their decision-making processes by providing additional insights into current and forecasted conditions. Therefore, it is fundamental to streamline the technology transfer process from research results into production and market-ready solutions.

Future works should consider further validation of the developed models for specific use cases and contexts, focusing, for example, on micro-climatic conditions and research into explainable methods to increase model transparency. Explainable AI is a research

---

line that can increase users' adoption of smart farming tools by detailing the model decision processes and patterns that led to output suggestions, ultimately building additional trust in these systems.

Advancements in available sensing technology (e.g., higher spatiotemporal resolution of satellite constellations and climate models) will enable the development of more precise and timely applications. EnogisAI can provide the resources needed to integrate such applications into agronomic DSSs.

An opportunity resulting from developing agronomic DSSs is to compute data-driven simulations based on model advice and suggestions. This would enable the assessment of sustainability and other relevant metrics in relation to current farming practices and AI-supported practice planning and execution. These simulations and assessments could be performed considering the forecasted scenarios of the next 10 to 50 years regarding climate, socioeconomic, and cultural aspects. They would contribute to delineating potential evolving dynamics and needs of agricultural and food systems. Such applications could help with the release of supporting programs and policies for the sustainable development of future agriculture.

# Bibliography

- [1] U. DESA, “Population Division (2022), ‘World Population Prospects 2022: Summary of Results’,” tech. rep., UN DESA/POP/2022/TR, 2022.
- [2] U. DESA, “Population Division (2024), ‘World Population Prospects 2022: Summary of Results’,” tech. rep., UN DESA/POP/2024/TR, 2024.
- [3] Unicef *et al.*, “The State of Food Security and Nutrition in the World 2024,” 2024.
- [4] G. S. Malhi, M. Kaur, and P. Kaushik, “Impact of climate change on agriculture and its mitigation strategies: A review,” *Sustainability*, vol. 13, no. 3, p. 1318, 2021.
- [5] M. Crippa, E. Solazzo, D. Guizzardi, F. Monforti-Ferrario, F. N. Tubiello, and A. Leip, “Food systems are responsible for a third of global anthropogenic GHG emissions,” *Nature food*, vol. 2, no. 3, pp. 198–209, 2021.
- [6] M. Springmann, M. Clark, D. Mason-D’Croz, K. Wiebe, B. L. Bodirsky, L. Las-saletta, W. De Vries, S. J. Vermeulen, M. Herrero, K. M. Carlson, *et al.*, “Options for keeping the food system within environmental limits,” *Nature*, vol. 562, no. 7728, pp. 519–525, 2018.
- [7] R. H. Beach, J. Creason, S. B. Ohrel, S. Ragnauth, S. Ogle, C. Li, P. Ingraham, and W. Salas, “Global mitigation potential and costs of reducing agricultural non-CO<sub>2</sub> greenhouse gas emissions through 2030,” *Journal of Integrative Environmental Sciences*, vol. 12, no. sup1, pp. 87–105, 2015.
- [8] N. D. Mueller, J. S. Gerber, M. Johnston, D. K. Ray, N. Ramankutty, and J. A. Foley, “Closing yield gaps through nutrient and water management,” *Nature*, vol. 490, no. 7419, pp. 254–257, 2012.

- [9] S. Robinson, D. Mason-D’Croz, T. Sulser, S. Islam, R. Robertson, T. Zhu, A. Gue-  
neau, G. Pitois, and M. W. Rosegrant, “The international model for policy anal-  
ysis of agricultural commodities and trade (IMPACT): model description for ver-  
sion 3,” 2015.
- [10] M. A. Sutton, A. Bleeker, C. M. Howard, J. W. Erisman, Y. Abrol, M. Bekunda,  
A. Datta, E. Davidson, W. De Vries, O. Oenema, *et al.*, “Our nutrient world. The  
challenge to produce more food & energy with less pollution,” tech. rep., Centre  
for Ecology & Hydrology, 2013.
- [11] B. Basso and J. Antle, “Digital agriculture to design sustainable agricultural  
systems,” *Nature Sustainability*, vol. 3, no. 4, pp. 254–256, 2020.
- [12] M. Javaid, A. Haleem, I. H. Khan, and R. Suman, “Understanding the potential  
applications of Artificial Intelligence in Agriculture Sector,” *Advanced Agrochem*,  
vol. 2, no. 1, pp. 15–30, 2023.
- [13] R. Ben Ayed and M. Hanana, “Artificial intelligence to improve the food and  
agriculture sector,” *Journal of Food Quality*, vol. 2021, no. 1, p. 5584754, 2021.
- [14] F. Droulia and I. Charalampopoulos, “Future climate change impacts on Euro-  
pean viticulture: A review on recent scientific advances,” *Atmosphere*, vol. 12,  
no. 4, p. 495, 2021.
- [15] A. Alikadic, I. Pertot, E. Eccel, C. Dolci, C. Zarbo, A. Caffarra, R. De Filippi,  
and C. Furlanello, “The impact of climate change on grapevine phenology and  
the influence of altitude: A regional study,” *Agricultural and forest meteorology*,  
vol. 271, pp. 73–82, 2019.
- [16] Osservatorio Smart AgriFood, “Agricoltura 4.0: il mercato e il livello di adozione  
in Italia nel 2023.” [https://www.osservatori.net/report/smart-agrifood/  
agricoltura-40-mercato-livello-adozione-italia-2023-report/](https://www.osservatori.net/report/smart-agrifood/agricoltura-40-mercato-livello-adozione-italia-2023-report/), 2024.  
[Online; accessed 22-november-2024].
- [17] M. Vigani, T. Fellmann, A. Porcella Capkovicova, and E. Ferrari, “Harvesting  
resilience: adapting the EU agricultural system to global challenges,” *npj Sus-  
tainable Agriculture*, vol. 2, no. 1, p. 21, 2024.
- [18] Y. Vecchio, G. P. Agnusdei, P. P. Miglietta, and F. Capitanio, “Adoption of  
precision farming tools: The case of Italian farmers,” *International journal of  
environmental research and public health*, vol. 17, no. 3, p. 869, 2020.

- 
- [19] T. A. Shaikh, T. Rasool, and F. R. Lone, "Towards leveraging the role of machine learning and artificial intelligence in precision agriculture and smart farming," *Computers and Electronics in Agriculture*, vol. 198, p. 107119, 2022.
- [20] E. F. Gris, V. M. Burin, E. Brighenti, H. Vieira, and M. B. Luiz, "Phenology and ripening of *Vitis vinifera* L. grape varieties in São Joaquim, southern Brazil: a new South American wine growing region," *Ciencia e investigación agraria: revista latinoamericana de ciencias de la agricultura*, vol. 37, no. 2, pp. 61–75, 2010.
- [21] L. Leolini, S. Costafreda-Aumedes, J. A. Santos, C. Menz, H. Fraga, D. Molitor, P. Merante, J. Junk, T. Kartschall, A. Destrac-Irvine, *et al.*, "Phenological model intercomparison for estimating grapevine budbreak date (*Vitis vinifera* L.) in Europe," *Applied Sciences*, vol. 10, no. 11, p. 3800, 2020.
- [22] S. Reis, H. Fraga, C. Carlos, J. Silvestre, J. Eiras-Dias, P. Rodrigues, and J. A. Santos, "Grapevine phenology in four Portuguese wine regions: Modeling and predictions," *Applied Sciences*, vol. 10, no. 11, p. 3708, 2020.
- [23] M. S. Sirsat, J. Mendes-Moreira, C. Ferreira, and M. Cunha, "Machine Learning predictive model of grapevine yield based on agroclimatic patterns," *Engineering in Agriculture, Environment and Food*, vol. 12, no. 4, pp. 443–450, 2019.
- [24] P. Rodrigues, V. Pedroso, F. Gonçalves, S. Reis, and J. A. Santos, "Temperature-Based grapevine ripeness modeling for cv. Touriga nacional and encruzado in the dão wine region, Portugal," *Agronomy*, vol. 11, no. 9, p. 1777, 2021.
- [25] G. Blanco, F. Oldani, D. Salza, B. Basile, and C. Rossi, "Satellite-based Grapevine Phenological Stage Detection through a Deep Supervised Machine Learning Approach," in *2023 IEEE International Workshop on Metrology for Agriculture and Forestry (MetroAgriFor)*, pp. 60–65, IEEE, 2023.
- [26] W. Campillay-Llanos, S. Ortega-Farias, and L. Ahumada-Orellana, "Development and validation of phenological models for eight varieties of sweet cherry (*Prunus avium* L.) growing under Mediterranean climate condition," *Scientia Horticulturae*, vol. 326, p. 112711, 2024.
- [27] Q. Yang, L. Shi, J. Han, J. Yu, and K. Huang, "A near real-time deep learning approach for detecting rice phenology based on UAV images," *Agricultural and Forest Meteorology*, vol. 287, p. 107938, 2020.

- [28] J. Han, L. Shi, Q. Yang, K. Huang, Y. Zha, and J. Yu, “Real-time detection of rice phenology through convolutional neural network using handheld camera images,” *Precision Agriculture*, vol. 22, no. 1, pp. 154–178, 2021.
- [29] S. Ravuri, K. Lenc, M. Willson, D. Kangin, R. Lam, P. Mirowski, M. Fitzsimons, M. Athanassiadou, S. Kashem, S. Madge, *et al.*, “Skilful precipitation nowcasting using deep generative models of radar,” *Nature*, vol. 597, no. 7878, pp. 672–677, 2021.
- [30] T. D. Pham, “Time–frequency time–space LSTM for robust classification of physiological signals,” *Scientific reports*, vol. 11, no. 1, pp. 1–11, 2021.
- [31] Y. Guo, S. Chen, Y. H. Fu, Y. Xiao, W. Wu, H. Wang, and K. d. Beurs, “Comparison of multi-methods for identifying maize phenology using phenocams,” *Remote Sensing*, vol. 14, no. 2, p. 244, 2022.
- [32] A. Milella, R. Marani, A. Petitti, and G. Reina, “In-field high throughput grapevine phenotyping with a consumer-grade depth camera,” *Computers and electronics in agriculture*, vol. 156, pp. 293–306, 2019.
- [33] X. A. Wang, J. Tang, and M. Whitty, “DeepPhenology: Estimation of apple flower phenology distributions based on deep learning,” *Computers and Electronics in Agriculture*, vol. 185, p. 106123, 2021.
- [34] A. Mahmood, S. K. Singh, and A. K. Tiwari, “Pre-trained deep learning-based classification of jujube fruits according to their maturity level,” *Neural Computing and Applications*, vol. 34, no. 16, pp. 13925–13935, 2022.
- [35] L. Rodrigues, S. A. Magalhães, D. Q. da Silva, F. N. dos Santos, and M. Cunha, “Computer vision and deep learning as tools for leveraging dynamic phenological classification in vegetable crops,” *Agronomy*, vol. 13, no. 2, p. 463, 2023.
- [36] L. Shen, S. Chen, Z. Mi, J. Su, R. Huang, Y. Song, Y. Fang, and B. Su, “Identifying veraison process of colored wine grapes in field conditions combining deep learning and image analysis,” *Computers and Electronics in Agriculture*, vol. 200, p. 107268, 2022.
- [37] M. Schieck, P. Krajsic, F. Loos, A. Hussein, B. Franczyk, A. Kozierekiewicz, and M. Pietranik, “Comparison of deep learning methods for grapevine growth stage recognition,” *Computers and Electronics in Agriculture*, vol. 211, p. 107944, 2023.

- 
- [38] P. Gatou, X. Tsiara, A. Spitalas, S. Sioutas, and G. Vonitsanos, “Artificial Intelligence Techniques in Grapevine Research: A Comparative Study with an Extensive Review of Datasets, Diseases, and Techniques Evaluation,” *Sensors*, vol. 24, no. 19, p. 6211, 2024.
- [39] T. Santos, L. de Souza, A. Dos Santos, and A. Sandra, “Embrapa Wine Grape Instance Segmentation Dataset–Embrapa WGISD; Zenodo. 2019,” *The Building of the WGISD Dataset was Supported by the Embrapa SEG Project*, vol. 1, no. 09.001, pp. 05–04.
- [40] M. Sozzi, S. Cantalamessa, A. Cogato, A. Kayad, and F. Marinello, “wGrapeUNIPD-DL: An open dataset for white grape bunch detection,” *Data in Brief*, vol. 43, p. 108466, 2022.
- [41] K. Harfenmeister, S. Itzerott, C. Weltzien, and D. Spengler, “Detecting phenological development of winter wheat and winter barley using time series of sentinel-1 and sentinel-2,” *Remote Sensing*, vol. 13, no. 24, p. 5036, 2021.
- [42] T. Ha, Y. Shen, H. Duddu, E. Johnson, and S. J. Shirtliffe, “Quantifying hail damage in crops using sentinel-2 imagery,” *Remote Sensing*, vol. 14, no. 4, p. 951, 2022.
- [43] M. Gabbrielli, M. Corti, M. Perfetto, V. Fassa, and L. Bechini, “Satellite-based frost damage detection in support of winter cover crops management: A case study on white mustard,” *Agronomy*, vol. 12, no. 9, p. 2025, 2022.
- [44] J. Segarra, J. L. Araus, and S. C. Kefauver, “Farming and Earth Observation: Sentinel-2 data to estimate within-field wheat grain yield,” *International Journal of Applied Earth Observation and Geoinformation*, vol. 107, p. 102697, 2022.
- [45] K. Amankulova, N. Farmonov, P. Akramova, I. Tursunov, and L. Mucsi, “Comparison of PlanetScope, Sentinel-2, and landsat 8 data in soybean yield estimation within-field variability with random forest regression,” *Heliyon*, vol. 9, no. 6, 2023.
- [46] N. Farmonov, K. Amankulova, J. Szatmári, J. Urinov, Z. Narmanov, J. Nosirov, and L. Mucsi, “Combining PlanetScope and Sentinel-2 images with environmental data for improved wheat yield estimation,” *International Journal of Digital Earth*, vol. 16, no. 1, pp. 847–867, 2023.

- [47] F. Nutini, R. Confalonieri, L. Paleari, M. Pepe, L. Criscuolo, F. Porta, L. Ranghetti, L. Busetto, and M. Boschetti, “Supporting operational site-specific fertilization in rice cropping systems with infield smartphone measurements and Sentinel-2 observations,” *Precision Agriculture*, vol. 22, pp. 1284–1303, 2021.
- [48] B. Ortuani, A. Mayer, D. Bianchi, G. Sona, A. Crema, D. Modina, M. Bolognini, L. Brancadoro, M. Boschetti, and A. Facchi, “Effectiveness of Management Zones Delineated from UAV and Sentinel-2 Data for Precision Viticulture Applications,” *Remote Sensing*, vol. 16, no. 4, p. 635, 2024.
- [49] T. T. Nguyen, T. D. Hoang, M. T. Pham, T. T. Vu, T. H. Nguyen, Q.-T. Huynh, and J. Jo, “Monitoring agriculture areas with satellite images and deep learning,” *Applied Soft Computing*, vol. 95, p. 106565, 2020.
- [50] D. Roy and L. Yan, “Robust Landsat-based crop time series modelling,” *Remote Sensing of Environment*, vol. 238, p. 110810, 2020.
- [51] S. Pesaresi, A. Mancini, G. Quattrini, and S. Casavecchia, “Mapping mediterranean forest plant associations and habitats with functional principal component analysis using Landsat 8 NDVI time series,” *Remote sensing*, vol. 12, no. 7, p. 1132, 2020.
- [52] S. Pesaresi, A. Mancini, G. Quattrini, and S. Casavecchia, “Functional analysis for habitat mapping in a special area of conservation using sentinel-2 time-series data,” *Remote Sensing*, vol. 14, no. 5, p. 1179, 2022.
- [53] B. Kashyap and R. Kumar, “Sensing methodologies in agriculture for soil moisture and nutrient monitoring,” *IEEE Access*, vol. 9, pp. 14095–14121, 2021.
- [54] F. Viani, M. Bertolli, M. Salucci, and A. Polo, “Low-cost wireless monitoring and decision support for water saving in agriculture,” *IEEE sensors journal*, vol. 17, no. 13, pp. 4299–4309, 2017.
- [55] I. Kisekka, S. R. Peddinti, W. P. Kustas, A. J. McElrone, N. Bambach-Ortiz, L. McKee, and W. Bastiaanssen, “Spatial-temporal modeling of root zone soil moisture dynamics in a vineyard using machine learning and remote sensing,” *Irrigation science*, vol. 40, no. 4, pp. 761–777, 2022.
- [56] L. Yin, S. Yan, M. Li, W. Liu, S. Zhang, X. Xie, X. Wang, W. Wang, S. Chang, and F. Hou, “Enhancing soil moisture estimation in alfalfa root-zone using UAV-



- based multimodal remote sensing and deep learning,” *European Journal of Agronomy*, vol. 161, p. 127366, 2024.
- [57] H. Darouich, T. B. Ramos, L. S. Pereira, D. Rabino, G. Bagagiolo, G. Capello, L. Simionesei, E. Cavallo, and M. Biddoccu, “Water Use and Soil Water Balance of Mediterranean Vineyards under Rainfed and Drip Irrigation Management: Evapotranspiration Partition and Soil Management Modelling for Resource Conservation,” *Water*, vol. 14, no. 4, p. 554, 2022.
- [58] À. Puig-Sirera, G. Rallo, P. Paredes, T. A. Paço, M. Minacapilli, G. Provenzano, and L. S. Pereira, “Transpiration and water use of an irrigated traditional olive grove with sap-flow observations and the FAO56 dual crop coefficient approach,” *Water*, vol. 13, no. 18, p. 2466, 2021.
- [59] H. Darouich, R. Karfoul, T. B. Ramos, A. Moustafa, and L. S. Pereira, “Searching for Sustainable-Irrigation Issues of Clementine Orchards in the Syrian Akkar Plain: Effects of Irrigation Method and Canopy Size on Crop Coefficients, Transpiration, and Water Use with SIMDualKc Model,” *Water*, vol. 14, no. 13, p. 2052, 2022.
- [60] R. G. Allen, L. S. Pereira, D. Raes, M. Smith, *et al.*, “Crop evapotranspiration-Guidelines for computing crop water requirements-FAO Irrigation and drainage paper 56,” *Fao, Rome*, vol. 300, no. 9, p. D05109, 1998.
- [61] J. Ollerton, “Pollinator diversity: distribution, ecological function, and conservation,” *Annual review of ecology, evolution, and systematics*, vol. 48, no. 1, pp. 353–376, 2017.
- [62] D. M. Katumo, H. Liang, A. C. Ochola, M. Lv, Q.-F. Wang, and C.-F. Yang, “Pollinator diversity benefits natural and agricultural ecosystems, environmental health, and human welfare,” *Plant Diversity*, vol. 44, no. 5, pp. 429–435, 2022.
- [63] L. H. Ziska, L. Makra, S. K. Harry, N. Bruffaerts, M. Hendrickx, F. Coates, A. Saarto, M. Thibaudon, G. Oliver, A. Damialis, *et al.*, “Temperature-related changes in airborne allergenic pollen abundance and seasonality across the northern hemisphere: a retrospective data analysis,” *The Lancet Planetary Health*, vol. 3, no. 3, pp. e124–e131, 2019.

- [64] Y. Zhang and A. L. Steiner, “Projected climate-driven changes in pollen emission season length and magnitude over the continental United States,” *Nature communications*, vol. 13, no. 1, p. 1234, 2022.
- [65] E. de Gálvez-Montañez, M. M. Trigo, M. Recio, and A. Picornell, “Flowering seasonality and airborne pollen recent trends in Sierra de las Nieves, the southernmost National Park in continental Spain,” *Agricultural and Forest Meteorology*, vol. 359, p. 110295, 2024.
- [66] A. M. Chamer, D. Medan, A. I. Mantese, and N. J. Bartoloni, “Impact of pollination on sunflower yield: is pollen amount or pollen quality what matters?,” *Field Crops Research*, vol. 176, pp. 61–70, 2015.
- [67] L. Ascari, C. Siniscalco, G. Palestini, M. J. Lisperguer, E. S. Huerta, T. De Gregorio, and S. Bregaglio, “Relationships between yield and pollen concentrations in Chilean hazelnut orchards,” *European Journal of Agronomy*, vol. 115, p. 126036, 2020.
- [68] C. Suanno, I. Aloisi, D. Fernández-González, and S. Del Duca, “Pollen forecasting and its relevance in pollen allergen avoidance,” *Environmental research*, vol. 200, p. 111150, 2021.
- [69] J. M. Maya-Manzano, M. Smith, E. Markey, J. Hourihane Clancy, J. Sodeau, and D. J. O’Connor, “Recent developments in monitoring and modelling airborne pollen, a review,” *Grana*, vol. 60, no. 1, pp. 1–19, 2021.
- [70] D. Voukantsis, H. Niska, K. Karatzas, M. Riga, A. Damialis, and D. Vokou, “Forecasting daily pollen concentrations using data-driven modeling methods in Thessaloniki, Greece,” *Atmospheric Environment*, vol. 44, no. 39, pp. 5101–5111, 2010.
- [71] C. Chen, W. He, H. Zhou, Y. Xue, and M. Zhu, “A comparative study among machine learning and numerical models for simulating groundwater dynamics in the Heihe River Basin, northwestern China. Sci Rep 10 (1): 3904,” 2020.
- [72] R. Navares and J. L. Aznarte, “Predicting the Poaceae pollen season: six month-ahead forecasting and identification of relevant features,” *International journal of biometeorology*, vol. 61, no. 4, pp. 647–656, 2017.

- 
- [73] J. M. Cordero, J. Rojo, A. M. Gutiérrez-Bustillo, A. Narros, and R. Borge, “Predicting the Olea pollen concentration with a machine learning algorithm ensemble,” *International Journal of Biometeorology*, vol. 65, no. 4, pp. 541–554, 2021.
- [74] D. Voukantsis, K. Karatzas, S. Jaeger, U. Berger, and M. Smith, “Analysis and forecasting of airborne pollen-induced symptoms with the aid of computational intelligence methods,” *Aerobiologia*, vol. 29, pp. 175–185, 2013.
- [75] G. K. Zewdie, D. J. Lary, E. Levetin, and G. F. Garuma, “Applying deep neural networks and ensemble machine learning methods to forecast airborne ambrosia pollen,” *International journal of environmental research and public health*, vol. 16, no. 11, p. 1992, 2019.
- [76] J. Valencia, G. Astray, M. Fernández-González, M. Aira, and F. Rodríguez-Rajo, “Assessment of neural networks and time series analysis to forecast airborne Parietaria pollen presence in the Atlantic coastal regions,” *International journal of biometeorology*, vol. 63, pp. 735–745, 2019.
- [77] R. Navares and J. L. Aznarte, “Forecasting Plantago pollen: improving feature selection through random forests, clustering, and Friedman tests,” *Theoretical and Applied Climatology*, vol. 139, no. 1, pp. 163–174, 2020.
- [78] A. Muzalyova, J. O. Brunner, C. Traidl-Hoffmann, and A. Damialis, “Forecasting Betula and Poaceae airborne pollen concentrations on a 3-hourly resolution in Augsburg, Germany: toward automatically generated, real-time predictions,” *Aerobiologia*, vol. 37, no. 3, pp. 425–446, 2021.
- [79] Y. Lops, Y. Choi, E. Eslami, and A. Sayeed, “Real-time 7-day forecast of pollen counts using a deep convolutional neural network,” *Neural Computing and Applications*, vol. 32, pp. 11827–11836, 2020.
- [80] A. Picornell, J. Oteros, R. Ruiz-Mata, M. Recio, M. Trigo, M. Martínez-Bracero, B. Lara, A. Serrano-García, C. Galán, H. García-Mozo, *et al.*, “Methods for interpolating missing data in aerobiological databases,” *Environmental Research*, vol. 200, p. 111391, 2021.
- [81] R. Navares and J. L. Aznarte, “Geographical imputation of missing Poaceae pollen data via convolutional neural networks,” *Atmosphere*, vol. 10, no. 11, p. 717, 2019.

- [82] N. Jaques, S. Taylor, A. Sano, and R. Picard, “Multimodal autoencoder: A deep learning approach to filling in missing sensor data and enabling better mood prediction,” in *2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII)*, pp. 202–208, IEEE, 2017.
- [83] V. Fortuin, D. Baranchuk, G. Rätsch, and S. Mandt, “Gp-vae: Deep probabilistic time series imputation,” in *International conference on artificial intelligence and statistics*, pp. 1651–1661, PMLR, 2020.
- [84] Z. Pan, Y. Wang, K. Wang, H. Chen, C. Yang, and W. Gui, “Imputation of missing values in time series using an adaptive-learned median-filled deep autoencoder,” *IEEE Transactions on Cybernetics*, vol. 53, no. 2, pp. 695–706, 2022.
- [85] H. Wi, Y. Shin, and N. Park, “Continuous-time Autoencoders for Regular and Irregular Time Series Imputation,” in *Proceedings of the 17th ACM International Conference on Web Search and Data Mining*, pp. 826–835, 2024.
- [86] A. Picornell, S. Hurtado, M. L. Antequera-Gómez, C. Barba-González, R. Ruiz-Mata, E. de Gálvez-Montañez, M. Recio, M. del Mar Trigo, J. F. Aldana-Montes, and I. Navas-Delgado, “A deep learning LSTM-based approach for forecasting annual pollen curves: Olea and Urticaceae pollen types as a case study,” *Computers in Biology and Medicine*, vol. 168, p. 107706, 2024.
- [87] D. Lorenz, K. Eichhorn, H. Bleiholder, R. Klose, U. Meier, and E. Weber, “Growth Stages of the Grapevine: Phenological growth stages of the grapevine (*Vitis vinifera* L. ssp. *vinifera*)—Codes and descriptions according to the extended BBCH scale,” *Australian Journal of Grape and Wine Research*, vol. 1, no. 2, pp. 100–103, 1995.
- [88] U. Meier *et al.*, “Growth stages of mono-and dicotyledonous plants= Entwicklungsstadien mono-und dikotyler Pflanzen= Estadios de las plantas mono-y dicotiledóneas= Stades phénologiques des mono-et dicotylédones cultivées,” *Berlin [etc.]: Blackwell*, 1997.
- [89] G. Ferrara, V. Marcotuli, A. Didonna, A. M. Stellacci, M. Palasciano, and A. Mazzeo, “Ripeness prediction in table grape cultivars by using a portable NIR device,” *Horticulturae*, vol. 8, no. 7, p. 613, 2022.
- [90] I. R. Donis-González, C. Valero, M. A. Momin, A. Kaur, and D. C. Slaughter, “Performance evaluation of two commercially available portable spectrometers to

- non-invasively determine table grape and peach quality attributes,” *Agronomy*, vol. 10, no. 1, p. 148, 2020.
- [91] P. Hewage, A. Behera, M. Trovati, E. Pereira, M. Ghahremani, F. Palmieri, and Y. Liu, “Temporal convolutional neural (TCN) network for an effective weather forecasting using time-series data from the local weather station,” *Soft Computing*, vol. 24, pp. 16453–16482, 2020.
- [92] P. Hewage, M. Trovati, E. Pereira, and A. Behera, “Deep learning-based effective fine-grained weather forecasting model,” *Pattern Analysis and Applications*, vol. 24, no. 1, pp. 343–366, 2021.
- [93] P. Kumar, R. Chandra, C. Bansal, S. Kalyanaraman, T. Ganu, and M. Grant, “Micro-climate prediction-multi scale encoder-decoder based deep learning framework,” in *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pp. 3128–3138, 2021.
- [94] C. De Lucia, E. Bucchignani, A. Mastellone, M. Adinolfi, M. Montesarchio, D. Cinquegrana, P. Mercogliano, and P. Schiano, “A Sensitivity Study on High Resolution NWP ICON—LAM Model over Italy,” *Atmosphere*, vol. 13, no. 4, p. 540, 2022.
- [95] J. Muñoz-Sabater, E. Dutra, A. Agustí-Panareda, C. Albergel, G. Arduini, G. Balsamo, S. Boussetta, M. Choulga, S. Harrigan, H. Hersbach, *et al.*, “ERA5-Land: A state-of-the-art global reanalysis dataset for land applications,” *Earth system science data*, vol. 13, no. 9, pp. 4349–4383, 2021.
- [96] H. Hersbach, B. Bell, P. Berrisford, S. Hirahara, A. Horányi, J. Muñoz-Sabater, J. Nicolas, C. Peubey, R. Radu, D. Schepers, *et al.*, “The ERA5 global reanalysis,” *Quarterly Journal of the Royal Meteorological Society*, vol. 146, no. 730, pp. 1999–2049, 2020.
- [97] M. Raffa, A. Reder, G. F. Marras, M. Mancini, G. Scipione, M. Santini, and P. Mercogliano, “VHR-REA\_IT dataset: very high resolution dynamical downscaling of ERA5 reanalysis over Italy by COSMO-CLM,” *Data*, vol. 6, no. 8, p. 88, 2021.
- [98] B. Rockel, A. Will, and A. Hense, “The regional climate model COSMO-CLM (CCLM),” *Meteorologische zeitschrift*, vol. 17, no. 4, p. 347, 2008.

- [99] G. C. Rodrigues and R. P. Braga, “Estimation of reference evapotranspiration during the irrigation season using nine temperature-based methods in a hot-summer Mediterranean climate,” *Agriculture*, vol. 11, no. 2, p. 124, 2021.
- [100] P. Droogers and R. G. Allen, “Estimating reference evapotranspiration under inaccurate data conditions,” *Irrigation and drainage systems*, vol. 16, pp. 33–45, 2002.
- [101] A. Berti, G. Tardivo, A. Chiaudani, F. Rech, and M. Borin, “Assessing reference evapotranspiration by the Hargreaves method in north-eastern Italy,” *Agricultural Water Management*, vol. 140, pp. 20–25, 2014.
- [102] S. Khanal, K. Kc, J. P. Fulton, S. Shearer, and E. Ozkan, “Remote sensing in agriculture—accomplishments, limitations, and opportunities,” *Remote Sensing*, vol. 12, no. 22, p. 3783, 2020.
- [103] S. Tarquini, I. Isola, M. Favalli, F. Mazzarini, M. Bisson, M. T. Pareschi, and E. Boschi, “TINITALY/01: a new triangular irregular network of Italy,” *Annals of Geophysics*, 2007.
- [104] S. Tarquini and L. Nannipieri, “The 10 m-resolution TINITALY DEM as a trans-disciplinary basis for the analysis of the Italian territory: Current trends and new perspectives,” *Geomorphology*, vol. 281, pp. 108–115, 2017.
- [105] B. Tóth, M. Weynants, L. Pásztor, and T. Hengl, “3D soil hydraulic database of Europe at 250 m resolution,” *Hydrological Processes*, vol. 31, no. 14, pp. 2662–2666, 2017.
- [106] B. Tóth, M. Weynants, A. Nemes, A. Makó, G. Bilas, and G. Tóth, “New generation of hydraulic pedotransfer functions for Europe,” *European journal of soil science*, vol. 66, no. 1, pp. 226–238, 2015.
- [107] W. Boughton, “A review of the USDA SCS curve number method,” *Soil Research*, vol. 27, no. 3, pp. 511–523, 1989.
- [108] V. M. Ponce and R. H. Hawkins, “Runoff curve number: Has it reached maturity?,” *Journal of hydrologic engineering*, vol. 1, no. 1, pp. 11–19, 1996.
- [109] R. H. Hawkins, T. J. Ward, D. E. Woodward, and J. A. Van Mullem, “Curve number hydrology: State of the practice,” American Society of Civil Engineers, 2008.

- 
- [110] H. H. Jaafar, F. A. Ahmad, and N. El Beyrouthy, “GCN250, new global gridded curve numbers for hydrologic modeling and design,” *Scientific data*, vol. 6, no. 1, p. 145, 2019.
- [111] E. E. S. Agency, “CCI Land Cover Product User Guide version 2.4,” *ESA CCI LC project*, <https://www.esa-landcover-cci.org/?q=node/164>, 2018.
- [112] B. Poulter, N. MacBean, A. Hartley, I. Khlystova, O. Arino, R. Betts, S. Bon-temps, M. Boettcher, C. Brockmann, P. Defourny, *et al.*, “Plant functional type classification for earth system models: Results from the European Space Agency’s Land Cover Climate Change Initiative,” *Geoscientific Model Development*, vol. 8, no. 7, pp. 2315–2328, 2015.
- [113] USDA, “Hydrologic Soil Groups,” in *National Engineering Handbook: Part 630 - Hydrology*, 2009.
- [114] M. Neteler, M. Bowman, M. Landa, and M. Metz, “GRASS GIS: a multi-purpose Open Source GIS,” *Environmental Modelling & Software*, vol. 31, pp. 124–130, 2012.
- [115] E. S. Agency, “Copernicus DEM—global and European digital elevation model (COP-DEM),” 2018.
- [116] F. Nex, F. Remondino, G. Agugiaro, R. De Flippi, M. Poletti, C. Furlanello, S. Menegon, G. Dallago, and S. Fontanari, “3D Solarweb: A solar cadaster in the Italian Alpine landscape,” *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. 40, pp. 173–178, 2013.
- [117] A. Hall and G. V. Jones, “Spatial analysis of climate in winegrape-growing regions in Australia,” *Australian Journal of Grape and Wine Research*, vol. 16, no. 3, pp. 389–404, 2010.
- [118] H. Bai, G. A. Gambetta, Y. Wang, J. Kong, Q. Long, P. Fan, W. Duan, Z. Liang, and Z. Dai, “Historical long-term cultivar  $\times$  climate suitability data to inform viticultural adaptation to climate change,” *Scientific Data*, vol. 9, no. 1, p. 271, 2022.
- [119] A. Winkler, *General viticulture*. University of California Press, 1974.

- [120] M. Huglin, “Nouveau mode d’évaluation des possibilites heliothermiques d’un milieu viticole [climatologie].,” *Comptes rendus des Séances de l’Académie d’Agriculture de France*, vol. 64, 1978.
- [121] J. Gladstones, “Viticulture and environment: Winetitles,” *Adelaide, Australia*, 1992.
- [122] E. De Martonne, “Aridité et indices d’aridité,” *Académie des Sciences. Comptes Rendus*, vol. 182, no. 23, pp. 1935–1938, 1923.
- [123] P. Bourgault, D. Huard, T. J. Smith, T. Logan, A. Aoun, J. Lavoie, É. Dupuis, G. Rondeau-Genesse, R. Alegre, C. Barnes, *et al.*, “xclim: xarray-based climate data analytics,” *Journal of Open Source Software*, vol. 8, no. 85, p. 5415, 2023.
- [124] Timescale, “TimescaleDB.” <https://github.com/timescale/timescaledb/>. [Online; accessed 11-november-2024].
- [125] Timescale, “About compression methods.” <https://docs.timescale.com/use-timescale/latest/compression/compression-methods/>. [Online; accessed 11-november-2024].
- [126] Celery, “Celery.” <https://github.com/celery/celery/>. [Online; accessed 11-november-2024].
- [127] L. Coviello, F. M. Martini, L. Cesaretti, S. Pesaresi, F. Solfanelli, and A. Mancini, “Clustering of Remotely Sensed Time Series Using Functional Principal Component Analysis to Monitor Crops,” in *2022 IEEE Workshop on Metrology for Agriculture and Forestry (MetroAgriFor)*, pp. 141–145, IEEE, 2022.
- [128] P. Tiwari and P. Shukla, “Artificial neural network-based crop yield prediction using NDVI, SPI, VCI feature vectors,” in *Information and Communication Technology for Sustainable Development: Proceedings of ICT4SD 2018*, pp. 585–594, Springer, 2019.
- [129] X.-P. Song, P. V. Potapov, A. Krylov, L. King, C. M. Di Bella, A. Hudson, A. Khan, B. Adusei, S. V. Stehman, and M. C. Hansen, “National-scale soybean mapping and area estimation in the United States using medium resolution satellite imagery and field survey,” *Remote sensing of environment*, vol. 190, pp. 383–395, 2017.



- 
- [130] K. Zhang, X. Ge, P. Shen, W. Li, X. Liu, Q. Cao, Y. Zhu, W. Cao, and Y. Tian, “Predicting rice grain yield based on dynamic changes in vegetation indexes during early to mid-growth stages,” *Remote sensing*, vol. 11, no. 4, p. 387, 2019.
- [131] H. L. Shang, “A survey of functional principal component analysis,” *ASTA Advances in Statistical Analysis*, vol. 98, pp. 121–142, 2014.
- [132] S. Bai, J. Z. Kolter, and V. Koltun, “An empirical evaluation of generic convolutional and recurrent networks for sequence modeling,” *arXiv preprint arXiv:1803.01271*, 2018.
- [133] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, “Distributed representations of words and phrases and their compositionality,” *Advances in neural information processing systems*, vol. 26, 2013.
- [134] W. A. Falcon, “Pytorch lightning,” *GitHub*, vol. 3, 2019.
- [135] O. Yadan, “Hydra - A framework for elegantly configuring complex applications.” Github, 2019.
- [136] K. Simonyan, “Very deep convolutional networks for large-scale image recognition,” *arXiv preprint arXiv:1409.1556*, 2014.
- [137] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.
- [138] A. Mancini, F. Solfanelli, L. Coviello, F. M. Martini, S. Mandolesi, and R. Zanolì, “Time Series from Sentinel-2 for Organic Durum Wheat Yield Prediction Using Functional Data Analysis and Deep Learning,” *Agronomy*, vol. 14, no. 1, p. 109, 2024.
- [139] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, “Mobilenetv2: Inverted residuals and linear bottlenecks,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4510–4520, 2018.
- [140] R. G. Allen, L. S. Pereira, M. Smith, D. Raes, and J. L. Wright, “FAO-56 dual crop coefficient method for estimating evaporation from soil and application extensions,” *Journal of irrigation and drainage engineering*, vol. 131, no. 1, pp. 2–13, 2005.

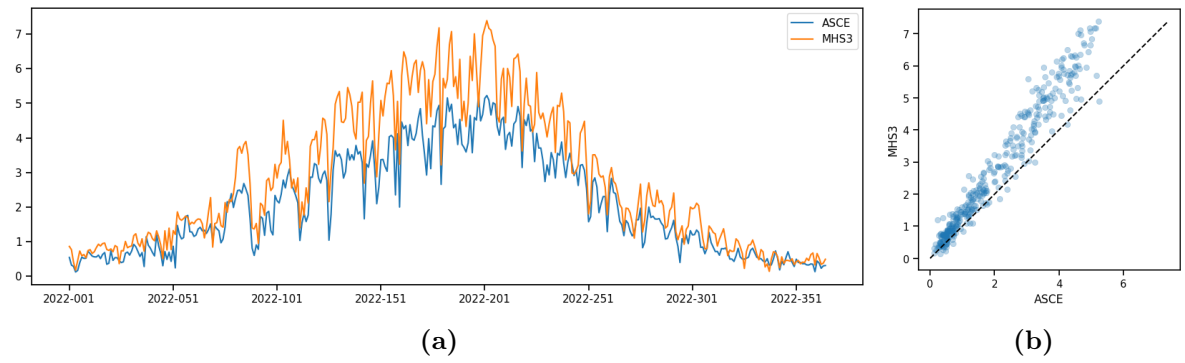
- [141] G. Rallo, T. Paço, P. Paredes, À. Puig-Sirera, R. Massai, G. Provenzano, and L. Pereira, “Updated single and dual crop coefficients for tree and vine fruit crops,” *Agricultural Water Management*, vol. 250, p. 106645, 2021.
- [142] H. Ritzema, *Drainage principles and applications*. No. 16, ILRI, 2006.
- [143] K. R. Thorp, J. Brekel, and K. C. DeJonge, “Version 1.2. 0—pyfao56: FAO-56 evapotranspiration in Python,” *SoftwareX*, vol. 24, p. 101518, 2023.
- [144] I. A. Walter, R. G. Allen, R. Elliott, M. E. Jensen, D. Itenfisu, B. Mecham, T. A. Howell, R. Snyder, P. Brown, S. Echings, *et al.*, “ASCE’s standardized reference evapotranspiration equation,” *Watershed management and operations management*, vol. 2000, pp. 1–11, 2000.
- [145] L. Makra, I. Matyasovszky, G. Tusnády, L. H. Ziska, J. J. Hess, L. G. Nyúl, D. S. Chapman, L. Coviello, A. Gobbi, G. Jurman, *et al.*, “A temporally and spatially explicit, data-driven estimation of airborne ragweed pollen concentrations across Europe,” *Science of the Total Environment*, vol. 905, p. 167095, 2023.
- [146] K. S. Bakar and S. K. Sahu, “spTimer: Spatio-temporal Bayesian modeling using R,” *Journal of Statistical Software*, vol. 63, pp. 1–32, 2015.
- [147] R. C. Cornes, G. van der Schrier, E. J. van den Besselaar, and P. D. Jones, “An ensemble version of the E-OBS temperature and precipitation data sets,” *Journal of Geophysical Research: Atmospheres*, vol. 123, no. 17, pp. 9391–9409, 2018.
- [148] L. Makra, L. Coviello, A. Gobbi, G. Jurman, C. Furlanello, M. Brunato, L. H. Ziska, J. J. Hess, A. Damialis, M. P. P. Garcia, *et al.*, “Forecasting daily total pollen concentrations on a global scale,” *Allergy*, vol. 79, no. 8, pp. 2173–2185, 2024.
- [149] L. Prokhorenkova, G. Gusev, A. Vorobev, A. Dorogush, A. Gulin, S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, *et al.*, “Advances in neural information processing systems,” *NeurIPS*, vol. 2018, pp. 6638–6648, 2018.
- [150] C. Bentéjac, A. Csörgő, and G. Martínez-Muñoz, “A comparative analysis of gradient boosting algorithms,” *Artificial Intelligence Review*, vol. 54, pp. 1937–1967, 2021.
- [151] L. McInnes, J. Healy, and J. Melville, “Umap: Uniform manifold approximation and projection for dimension reduction,” *arXiv preprint arXiv:1802.03426*, 2018.

- [152] M. Ester, H.-P. Kriegel, J. Sander, X. Xu, *et al.*, “A density-based algorithm for discovering clusters in large spatial databases with noise,” in *kdd*, vol. 96, pp. 226–231, 1996.
- [153] M. Engen, E. Sandø, B. L. O. Sjølander, S. Arenberg, R. Gupta, and M. Goodwin, “Farm-scale crop yield prediction from multi-temporal data using deep hybrid neural networks,” *Agronomy*, vol. 11, no. 12, p. 2576, 2021.
- [154] N. Darra, E. Anastasiou, O. Kriezi, E. Lazarou, D. Kalivas, and S. Fountas, “Can yield prediction be fully digitized? A systematic review,” *Agronomy*, vol. 13, no. 9, p. 2441, 2023.
- [155] J. Gomis-Cebolla, V. Rattayova, S. Salazar-Galán, and F. Francés, “Evaluation of ERA5 and ERA5-Land reanalysis precipitation datasets over Spain (1951–2020),” *Atmospheric Research*, vol. 284, p. 106606, 2023.
- [156] M. Adinolfi, M. Raffa, A. Reder, and P. Mercogliano, “Investigation on potential and limitations of ERA5 Reanalysis downscaled on Italy by a convection-permitting model,” *Climate Dynamics*, vol. 61, no. 9, pp. 4319–4342, 2023.
- [157] PyTorch, “TorchServe.” <https://github.com/pytorch/serve>. [Online; accessed 11-november-2024].

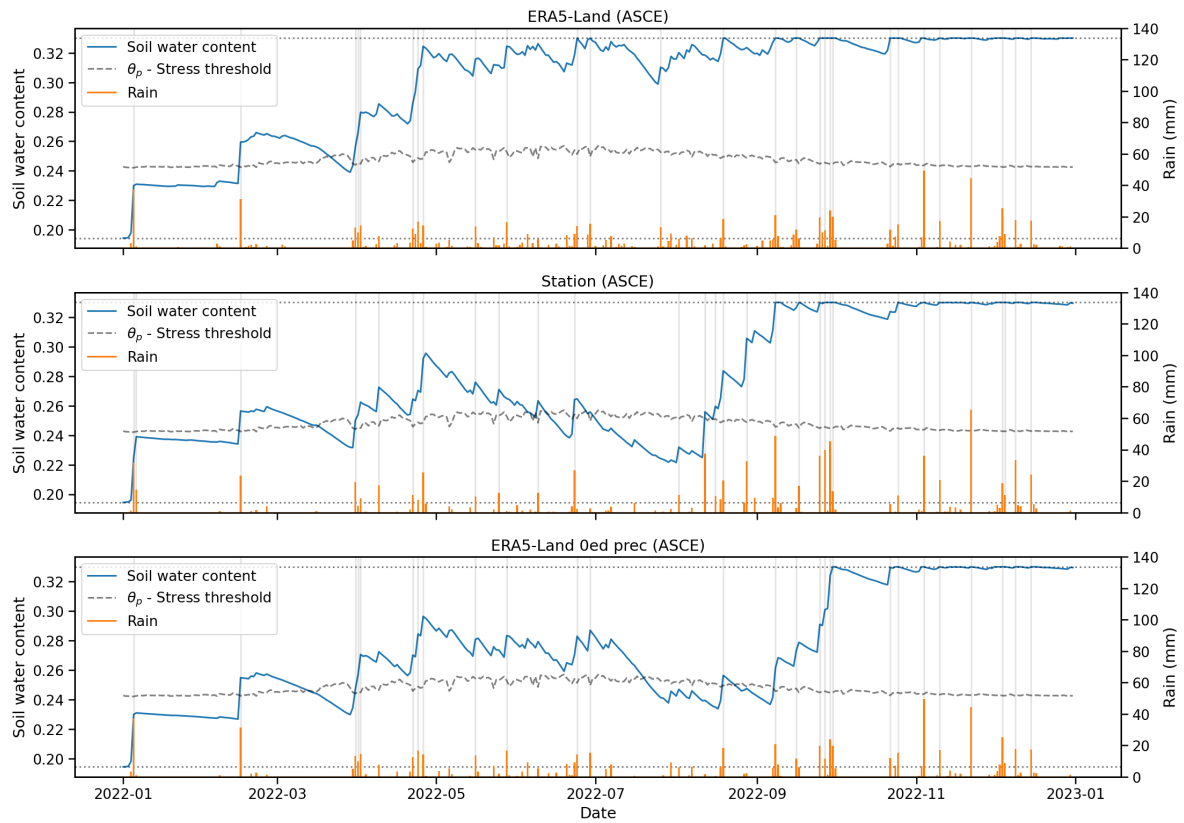


# Appendix A

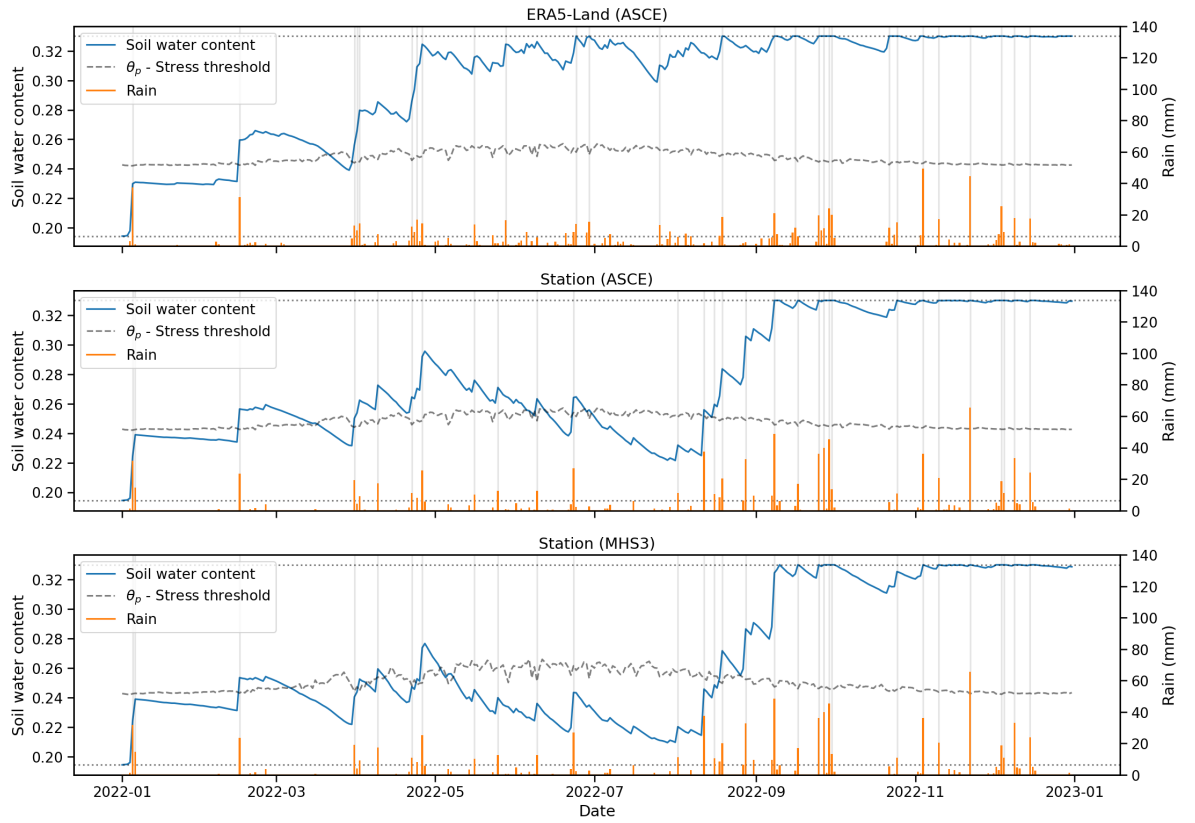
## Additional figures



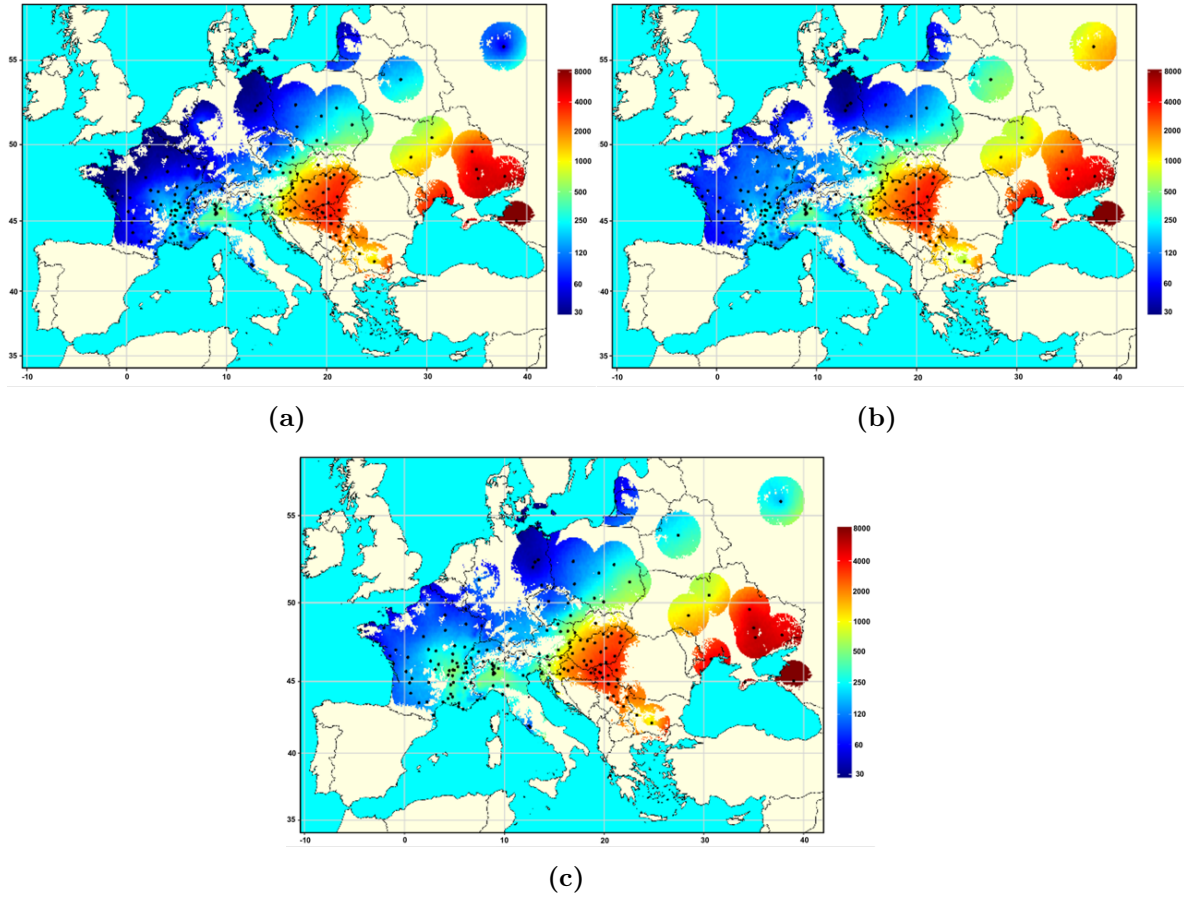
**Figure A.1:** Comparison between ASCE and MHS3 methods for estimating  $ET_0$  computed on the reference plot data for year 2022. Both methods used as input the same ERA5-Land variables.



**Figure A.2:** Soil available water simulation on 2022 using ASCE to compute  $ET_0$  with ERA5-Land (top), ERA5-Land substituting the total precipitation variable with station recorded precipitation (center), and ERA5-Land settings to 0 the precipitation for days that precipitation values were not recorded by the station (bottom).



**Figure A.3:** Simulation on 2022 using Soil available water simulation on 2022 using ASCE to compute  $ET_0$  with ERA5-Land (top), ERA5-Land substituting the total precipitation variable with station recorded precipitation (center), and using MHS3 on station temperature and precipitation data (bottom).



**Figure A.4:** Total annual ragweed pollen concentrations for Europe on base-10 logarithmic color scale, with the geographical locations of the 162 aerobiological stations, using the (a) raw (a), GM restored (b), and (c) DL restored data sets, July 15 - October 15, 2010, (pollen grains  $\text{m}^{-3}$ ).