Politecnico di Torino

Politecnico di Torino

MSc in Environmental and Land Engineering

Climate Change track

Satellite Data Assimilation to Improve Regional Crop Yield Estimates

Candidate: Niccolò Lanfranco

Supervisors: Prof. S. Tamea

Prof. G. De Lannoy

Co-supervisor: L. Busschaert

Academic Year 2024/2025

Acknowledgements

I would like to thank Politecnico di Torino *for the scholarship that enabled me to spend six months at the renowned* Katholieke Universiteit Leuven.

I would also like to thank my supervisor, Professor Stefania Tamea, for supporting my decision to write my thesis abroad and for recommending Professor Gabrielle De Lannoy as my supervisor in Leuven.

Words cannot express how grateful I am to Gabrielle for trusting me even before she knew me and for demonstrating her support throughout my time in Belgium. Her knowledge and willingness to help enabled me to tackle the complex topics of this thesis.

To Louise Busschaert, for your commitment and dedication in providing technical, theoretical, and emotional support. During those busy months, you always found time to help me in countless ways.

To Jonas, Devon, and Lucas (and Louise, again) for the daily waves of positive vibes in the "Coolest Christmas Office"*. It was always so pleasant to wake up and go to work, knowing you were already there sticking Post-its on the* hate wall.

To Michel, Marit, Fien, Brian, Yoojin, Hugo, Arijit, Jan, Elise, and to everyone in the Soil & Water *division for all the moments we shared. From growing crops in the* Geogarden*, to weekends on the Belgian coast, karting in Brussels, and the countless walks and runs in the forest, not to mention all the kitchen meetings to eat some sweets for whatever reason.*

I want to thank all my university friends and colleagues, especially Kiarash and Azizullah, who shared lots of memories with me during these last two years in Turin. University is nothing without community. I hope I was a good example to you as you were (and still are) to me.

Ad Annamaria, Lorenzo, Annibale, Ginevra e Sara per il vostro indispensabile supporto emotivo. Siete sempre statɜ presenti e disponibili, anche a mille chilometri di distanza.

Alla mia famiglia, per aver creduto nei miei progetti e per esservi presi cura di me nel momento del bisogno. Mi dispiace essere stato lontano quando avrei dovuto ricambiare il sostegno.

Al Kontiki e tuttɜ coloro lo animano, per essere state la mia seconda casa e famiglia, un posto in cui rifugiarmi e sentirmi accolto, in cui esprimere l'affettività in modo spontaneo e naturale.

A Teodora, per essere l'esempio di persona che desidero e voglio essere.

Abstract

The AquaCrop crop growth model of the Food and Agriculture Organization was recently integrated into NASA's Land Information System Framework. This allows unprecedented crop estimation and satellite data assimilation (DA) experiments at regional scales. Satellite DA aims to combine a model and observations to reduce the uncertainties in crop estimates. This thesis assimilates the Copernicus fraction of vegetation cover (FCOVER) product to update the canopy cover (CC) and biomass, and consequently winter wheat yield estimates, in the Piedmont Region of Italy between 2017 and 2023. After calibrating the crop parameters, testing the model, and developing the DA routines, three model ensembles (modes) were generated by perturbing the model in various ways to estimate the model forecast uncertainty. Forcing data and state variables were perturbed in all modes; mode 1 included perturbation bias correction, mode 2 did not, nor did mode 3, which included an additional variation of parameters. Next, the FCOVER observations for winter wheat were assimilated with an ensemble Kalman filter. The results were compared with other satellite products and field surveys of yield.

The ensemble mode 3 with the most degrees of freedom led to the best model-only simulations compared to reference data, and came with the strongest DA updates. The DA improved CC by design and enhanced the model's dry above-ground biomass production; however, yield estimates showed no clear improvement for all ensemble modes. The DA increments to CC were constrained by a potential upper boundary (CC_{pot}), either as a result of exceeding physical boundaries or due to the asynchrony between the model's sowing date and the observations, and this in turn limits the updates to yield. On balance, the most promising DA results were obtained for the ensemble mode 2. Further studies are needed to understand how to address the uncertainty of the planting date or crop stages in general, and investigate the joint assimilation of soil moisture retrievals to overcome the degradation due to vegetation DA updates during the information propagation within the model.

Abbreviations

Acronyms

i.c.	Initial conditions
CGLS	Copernicus Global Land Service
CLMS	Copernicus Land Monitoring Service
CREA	*Consiglio per la Ricerca in agricoltura e l'analisi dell'Economia Agraria*
DA	Data Assimilation
Det	Deterministic
EnKF	Ensemble Kalman Filter
EO	Earth Observation
ERA5	5th generation of atmospheric reanalysis from the European Center for Medium Range Weather Forecasts
FAO	Food and Agriculture Organization of the United Nations
FAPAR	Fraction of Absorbed Photosynthetically Active Radiation
GUI	Graphical User Interface
GYGA	Global Yield Gap Atlas
KF	Kalman Filter
LAI	Leaf Area Index
LDT	Land Data Toolkit
LIS(F)	Land Information System (Framework)
LSM	Land Surface Model
M1/M2/M3	Mode (experiment) 1/2/3
MERIT-DEM	Multi-Error-Removed Improved-Terrain Digital Elevation Model

NASA	National Aeronautics and Space Administration
NNT	Neural Network Technique
OL	Open Loop
OLCI	Ocean and Land Colour Instrument
PROBA-V	Project for On-Board Autonomy - Vegetation
R	Pearson correlation coefficient
RICA	*Rete di Informazione Contabile Agricola*
RMSD	Root Mean Square Difference
ROI	Region of Interest
UTM	Universal Transverse Mercator projection
VSCode	Visual Studio Code
WGS 84	World Geodetic System 1984 ellipsoid
WW	Winter Wheat

Greek Symbols

α	Model parameters	
Δt	Time (or thermal) period	day or $^oC\ day$
μ	Mean	
σ	Standard deviation	
σ^2	Variance	
θ	Volumetric water content	$\frac{m^3_{water}}{m^3_{soil}}$

Matrices

$\mathbf{Cov}(\boldsymbol{x}, \boldsymbol{y})$	Ensemble variance-covariance matrix	$[n \times m]$
$\boldsymbol{H}$	Observation operator	$[n \times m]$
$\boldsymbol{K}$	Kalman Gain	$[n \times 1]$
$\boldsymbol{P}$	State variables variance-covariance matrix	$[n \times n]$

$\boldsymbol{R}$	Observation variables variance-covariance matrix	$[n \times n]$
$\boldsymbol{u}$	Input vector	$[n \times 1]$
$\boldsymbol{x}$	State variable vector	$[n \times 1]$
$\boldsymbol{y}$	Observations vector	$[m \times 1]$

Roman Symbols

B	Dry above-ground Biomass	$\frac{ton}{ha}$
CC	Canopy Cover	–
CC_o	Initial Canopy Cover	–
CC_x	Maximum Canopy Cover	–
CDC	Canopy Decline Coefficient	$\frac{1}{^oC\ day}$
CGC	Canopy Growth Coefficient	$\frac{1}{^oC\ day}$
DMP or ΔB_+	Dry Matter Productivity (observed or modelled)	$\frac{kg}{ha \cdot day}$
ET	Evapotranspiration	mm
ET_o	Reference Evapotranspiration	mm
FCOVER	Fraction of vegetation COVER	$\frac{m^2_{veg}}{m^2_{soil}}$
f_{HI}	Harvest Index adjustment factor	–
GDD	Growing Degree Days	$^oC\ day$
HI	Harvest Index	$\frac{kg_Y}{kg_B}$
HI_o	Reference Harvest Index	$\frac{kg_Y}{kg_B}$
Kc	Crop adjustment factor	–
Ks	Stress factors	–
P	Precipitation	mm
SM	Volumetric Soil Moisture	–
SW	Shortwave incoming radiation	$\frac{W}{m^2}$

T	Temperature	^oC
Tr	Transpiration	mm
WP^*	Normalized Water Productivity	$\frac{ton}{ha}$
Y	Dry crop Yield	$\frac{ton}{ha}$

Symbols

$(.)^+$	Analysis
$(.)^-$	Forecast
$\widehat{(.)}$	Estimate
$(.)_{Adj}$	Adjusted
$(.)_{dorm}$	Dormant
$(.)_{pot}$	Potential
$(.)_{req}$	Required

List of Tables

List of Figures

Contents

Chapter 1

Introduction

As greenhouse gases continue to accumulate in the atmosphere and extreme meteorological events become stronger and more frequent (IPCC, 2021), it has become crucial to estimate their impact on crop production to implement adaptation strategies. The risk associated with the climate crisis can be tackled on three different levels, by *i)* reducing the danger through mitigation of the extreme events, *ii)* limiting the exposure of people and human activities to such events, and *iii)* decreasing the vulnerability when the occurrence is inevitable (IPCC, 2022). The first option is the most effective; however, the temporal inertia of both the policy implementation and the climate system will result in increased frequency and intensity of dangerous accidents in the upcoming years (Riahi *et al.*, 2017; Tebaldi and Friedlingstein, 2013). The extensive nature of the agricultural system leaves little space for exposure limitation, and therefore risk must be addressed on the vulnerability level.

A better understanding of the dynamics of the agricultural ecosystem can help farmers and institutions to manage and plan the use of resources more efficiently. Furthermore, mid-range (i.e. seasonal) weather forecasting, coupled with crop models, can help in the short-term adaptation strategies. Several mathematical models have been developed in an attempt to represent the field dynamics (Di Paola *et al.*, 2016). However, these models depend on a large number of parameters. They are also heavily influenced by the system's initial conditions (*i.c.*), which are the values of the state variables at the beginning of the simulations. Additionally, local impacts are difficult to model due to *i)* the coarse scale of meteorological forcings, *ii)* the heterogeneity of local soil conditions, and *iii)* the complexity of crop-specific characteristics (Stöckle and Kemanian, 2020). Therefore, while models are continuously refined to describe reality more accurately, it is useful to explore whether other ap-

proaches can overcome the fundamental limitations of this simplified representation.

The last decades have seen an exponential increase in the number of satellites. According to the UCS database (2023), more than one thousand of these are dedicated to Earth Observation (EO). Alongside the quantity, the resolution, the spectral coverage, and the data processing have also improved (Belward and Skøien, 2015). EO systems produce powerful datasets that provide "high-frequency, extensive data for tracking environmental changes, assessing ecosystem health, and supporting resource management" (Zhao and Yu, 2025). However, satellite data also comes with its downsides: it is non-continuous, both spatially and temporally, and it is still subject to uncertainty.

Both crop models and EO have their strengths and limitations. Fortunately, there is a mathematical procedure that enables these two sources of information to interact and blend to produce the best possible analysis of a system's state by optimising uncertainties and filling any gaps. This procedure is called data assimilation (DA) (Lahoz and Schneider, 2014).

1.1 Research goals

The assimilation of satellite-borne data is an innovative approach to reduce uncertainty in crop yield estimates. Over the past few decades, DA has been applied and studied in multiple contexts using different approaches (Moradkhani *et al.*, 2018). This thesis aims to explore the advantages of incorporating readily available optical satellite data into crop yield modelling. The selected case study focuses on the production of rainfed soft winter wheat (WW) in Italy's Piedmont region, between 2017 and 2023.

Agriculture covers around 35.6% of the region's surface (Cavaletto, 2025). According to the Piedmont Region Data Warehouse (also known as the *Anagrafe Agricola Unica del Piemonte*), the main grain crops are maize (20.5% of the cultivated area), rice (20.5%) and wheat (14.0%, 96.6% of which is soft). These crops are distributed heterogeneously throughout the region, with some areas, such as the province of Vercelli, where rice production is concentrated, and the Alessandria

area, which focuses more on wheat. Although WW is only the third most important grain crop in the region, the area's climate allows farmers to grow it relying purely on direct rainfall. This information influenced the crop selection for the study, since a variable dependent on human action, such as irrigation, could mask the meteorologically induced interannual and spatial variability of crop yield.

The main objective of this study is to identify and quantify the advantages of using satellite data assimilation to update crop canopy cover and biomass in a crop model at the regional scale, as opposed to the use of the model alone.

1.2 Crop modelling

Like every living being, plants are complicated. The path from seed germination to the quantification of the harvest is far from straightforward. Furthermore, different species and cultivars can behave in many different ways. Plant growth depends heavily on the environmental context, including geographical location, water availability, the presence of macro and micro nutrients in the soil, soil texture, proximity to other plants, pests and diseases, salinity, and many other variables.

In an attempt to capture the main physical drivers in this process, various models have been developed since the late 1960s (Bournan *et al.*, 1996; Di Paola *et al.*, 2016). These models can be more or less complex, for example, by including or excluding nutrient dynamics, and can be more general or focused on a particular class of crops. Two main categories can be identified based on the goal of the simulations: *i)* explanatory models, which aim to precisely describe specific crop dynamics, usually characterised by a high amount of input data and parameters, and *ii)* predictive models, where the highly detailed mechanics leave space to the need for robustness, replicability and yield prediction (Di Paola *et al.*, 2016).

Another useful classification is the spatial scale of the model application. The majority of the models have been developed at the field scale, allowing for an accurate calibration. However, their practical application may be limited, since a smaller scale leads to a lower predictive power. Therefore, large-scale models have

been developed either directly (Bondeau *et al.*, 2007; Challinor *et al.*, 2004; Deryng *et al.*, 2011; Hennicker *et al.*, 2016; Osborne *et al.*, 2015) or by converting and testing field-scale models (Balkovič *et al.*, 2013; Boogaard *et al.*, 2013; de Roos *et al.*, 2021; Liu *et al.*, 2007; Stöckle *et al.*, 2014) in order to perform climate analyses or to help regional authorities and policymakers. Regionalisation of crop models is often obtained through some coupling with Geographical Information Systems (Liu *et al.*, 2007).

The model used in this thesis is the Food and Agriculture Organization of the United Nations (FAO) crop growth model AquaCrop. It was originally developed at the field scale, but it has recently been extended to run at the regional scale (Busschaert *et al.*, 2022; De Lannoy *et al.*, in review; de Roos *et al.*, 2021). AquaCrop was primarily developed to simulate annual crops and, at its core, it utilises a simple water balance. Since its release, it has been continuously improved, while preserving a limited set of parameters and a simple setup, as the *"target users [are] water user associations, consulting engineers, irrigation and farm managers, planners and economists"* (Raes *et al.*, 2025a).

While the most important AquaCrop processes are described in Section 2.2, it is useful to introduce the concept of the general state-space model to lay the foundations for the techniques used in this thesis. The primary objective of a model is to estimate certain state variables. These can include leaf area index, soil water content, and biomass, among others. Their evolution is usually described by a system of partial or ordinary differential equations, and the general solution in discretised intervals can be described as follows:

$$\boldsymbol{x}_{i+1} \;=\; \boldsymbol{f}_{i+1,i}(\boldsymbol{x}_i, \boldsymbol{u}_{i+1}, \boldsymbol{\alpha}), \tag{1.1}$$

where the n state variables $\boldsymbol{x}_{i+1}$ (vector of size $[n \times 1]$) are computed at time step $i+1$ based on the previous time step $\boldsymbol{x}_i$ (which can also be referred to as initial conditions, *i.c.*), some fixed parameters $\boldsymbol{\alpha}$ and on external inputs $\boldsymbol{u}_{i+1}$. The evolution is described by the non-linear system function $\boldsymbol{f}_{i+1,i}$.

Process-based models are usually deterministic, i.e. they do not intrinsically

consider uncertainties. Therefore, an exact set of values for $\boldsymbol{x}_i$, $\boldsymbol{u}_{i+1}$, and $\boldsymbol{\alpha}$ would produce exact solutions. However, as the reader may appreciate, the knowledge of the agricultural system is far from exact. Thus, the model's output must be interpreted as an *estimate* (i.e. $\hat{\boldsymbol{x}}_i$) of the true values, to take into account for model error (Reichle *et al.*, 2002).

1.3 AquaCrop

AquaCrop is a crop growth model initially developed by FAO to simulate crop yield production in response to water availability, calculate irrigation requirements, and analyse management practices. It has been available with a graphical user interface (GUI) since 2009, and has only recently been converted from Delphi/Pascal to Fortran 90, expanding its applicability (De Lannoy *et al.*, in review).

The model focuses on herbaceous annual and, more recently, on perennial crops. The concepts are summarised in Figure 1.1. Plant development follows an iterative computation of the canopy cover (CC) in response to soil water availability. CC, which describes the soil fraction ($0 \leq CC \leq 1$) covered by green material, plays a role in evapotranspiration (ET), which in turn influences the water balance and simultaneously drives dry biomass production (ΔB_+). Finally, the Harvest Index (HI) quantifies the amount of accumulated biomass that constitutes the crop yield (Y).

1.3.1 Water balance

Soil is a porous medium typically filled with two fluids: air and water. It can be considered as a water reservoir, with multiple inputs and outputs. Water content in soil is generally quantified in terms of the volumetric or mass ratio of water to soil. AquaCrop is based on volumetric water content, defined as $\boldsymbol{\theta}$ $[\mathrm{m}^3_{water}\ \mathrm{m}^{-3}_{soil}]$.

Soil water content can be replenished via rainfall, irrigation, or capillary rise. From a mechanical point of view, regardless of the soil type or texture, an excess of water results in surface runoff or percolation due to the gravitational force. Conversely, the surface tension and capillary forces allow some water to be retained, a

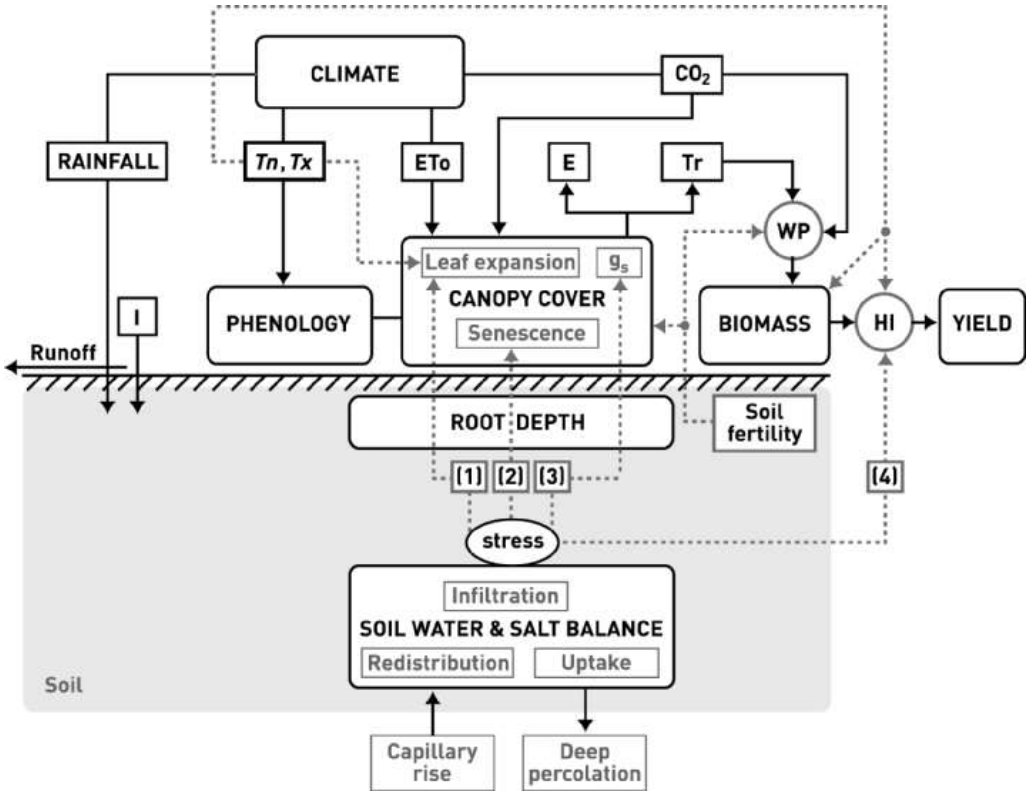

Figure 1.1: Flow chart of AquaCrop, available in the first chapter of the model's manual, with the main interactions (continuous lines) and the feedbacks (dashed lines) between the processes. All the elements present in the chart can be found in the original AquaCrop manual, Chapter 1 (Raes *et al.*, 2025a).

quantity known as *field capacity* ($\boldsymbol{\theta}_{FC}$). In addition to these processes, water can evaporate directly from the topsoil, or it can be taken up in the root zone and transpired by the vegetation. Water uptake is driven by osmosis; therefore, not all of it can be extracted from the soil. This level of depletion results in stress for the plant, and the residual water is known as the *wilting point* ($\boldsymbol{\theta}_{WP}$).

In the model, the differential flow equation is converted into a set of finite difference equations (Raes *et al.*, 2025c); that is to say, the soil profile is subdivided into multiple compartments. The link between the soil water balance and crop development occurs through a two-way pathway: transpiration (Tr [mm day^{-1}]) directly depends on CC [-], but a shortage of water can also result in stomata closure, leading to a reduction in Tr and slower crop growth. This codependence can be resolved if certain crop characteristics are known.

1.3.2 Canopy development

Before running the simulations, AquaCrop requires a crop calibration. The following description focuses on winter wheat (WW), an annual grain crop that is sown directly in the fields. As shown in Figure 1.2, CC development can be divided into different growth stages. First, the crop is sown, taking a certain amount of time to germinate. After the germination phase, described in AquaCrop as the time for the 90% of seeds to sprout, CC is set to an initial value $CC_o = cc_i \cdot d$, where cc_i is the surface area covered by a seedling $[m^2_{veg}]$, multiplied by the plant density d $[m^{-2}_{soil}]$. From that point onwards, the plant growth is determined by a combination of two functions: an exponential growth, followed by an exponential growth decay (Raes *et al.*, 2025c):

$$CC = CC_o \, e^{\Delta t \cdot CGC} \qquad \text{if} \quad CC \leq \frac{CC_x}{2}, \tag{1.2}$$

$$CC = CC_x - \frac{CC_x^2}{4CC_o} \, e^{-\Delta t \cdot CGC} \qquad \text{if} \quad CC > \frac{CC_x}{2}, \tag{1.3}$$

where CC_x [-] is the maximum canopy cover that the plant can reach without experiencing stress, and the canopy growth coefficient (CGC, [°C^{-1} day^{-1}] or [day^{-1}]) determines how quickly this value can be reached. Although the model operates at a daily resolution, crop development can also be based on thermal units, a heuristic tool called Growing Degree Day (GDD, [°C day]). Therefore, Δt can be interpreted as either days or cumulative GDD depending on the calibration criteria used.

GDD are usually computed as the cumulative amount of degrees above a certain threshold, called base temperature (T_{base} [°C]), that describes the lower limit below which the plant cannot grow. Since AquaCrop runs at a daily resolution, a simplified equation is introduced:

$$GDD_i = \frac{T_{max,i} + T_{min,i}}{2} - T_{base}, \tag{1.4}$$

using the daily maximum and minimum recorded temperatures. Note that $T_{max,i}$ [°C] is bounded by an upper limit (T_{upper} [°C]) above which crop development no longer increases with an increase in air temperature.

To prevent the exponential growth decay from continuing indefinitely, the canopy is set to CC_x as soon as $CC = 0.98 \cdot CC_x$. The moment at which the crop reaches its maximum occurs before the middle part of the flowering stage. While B [t ha^{-1}] is accumulated from the start of the growth stage, this is the moment when the plant begins to allocate resources to the harvestable product. As soon as the mid-season ends, senescence begins, following this equation:

$$CC = CC_x \left[1 - 0.05 \left(e^{\frac{3.33 \cdot CDC}{CC_x + 2.29}(\Delta t - \Delta t_{sen})} - 1 \right) \right], \tag{1.5}$$

where CDC ([°C^{-1} day^{-1}] or [day^{-1}]) is the canopy decline coefficient, which behaves similarly to CGC, and Δt_{sen} ([°C day] or [day]) is the time from sowing to the start of the senescence stage. The operator must set the length of the different stages to match the specific crop phenology within the area under investigation.

This set of equations describes the crop behaviour under perfect conditions. However, there are four main mechanisms that can interfere: *i)* extreme temperatures interrupt the growth, *ii)* water stress affects CGC and triggers early senescence, *iii)* salinity imbalances may reduce CC_x (not considered in this study), and *iv)* a lack of nutrients in the soil or management practices can impact every stage of the canopy development, so CGC, CC_x and the ability to maintain the maximum cover during the mid-season (De Lannoy *et al.*, in review).

The implementation of these stresses follows different paths. Fertility impacts are introduced as parameters, even though these can vary during the first part of the growing season. Salinity can vary due to the quality of the irrigation water used. Temperature stresses are directly determined by the forcings, thus influencing the state variable CC. The most complex effects are produced by the water balance, and the water stress must be re-computed daily during the simulation.

The actual CC is then used to compute the transpiration (Tr_i [mm day^{-1}]) at any given time i:

$$Tr_i = Ks_i \; (Kc_{WW,i} \; CC_i^*) \; ET_{o,i}, \tag{1.6}$$

where CC^* is the corrected canopy cover taking into account the interrow micro-

advection (which enhances Tr). Ks [-] ($0 \leq Ks \leq 1$) is the stress factor, including both cold and water stresses. ET_o [mm day^{-1}] is the reference evapotranspiration, computed using the FAO Penman-Monteith equation (Allen *et al.*, 1998). Kc [-] is the proportional factor, used to correct ET_o from the reference grass to WW (Raes *et al.*, 2025a). Tr is then fed back into the water balance for the next day.

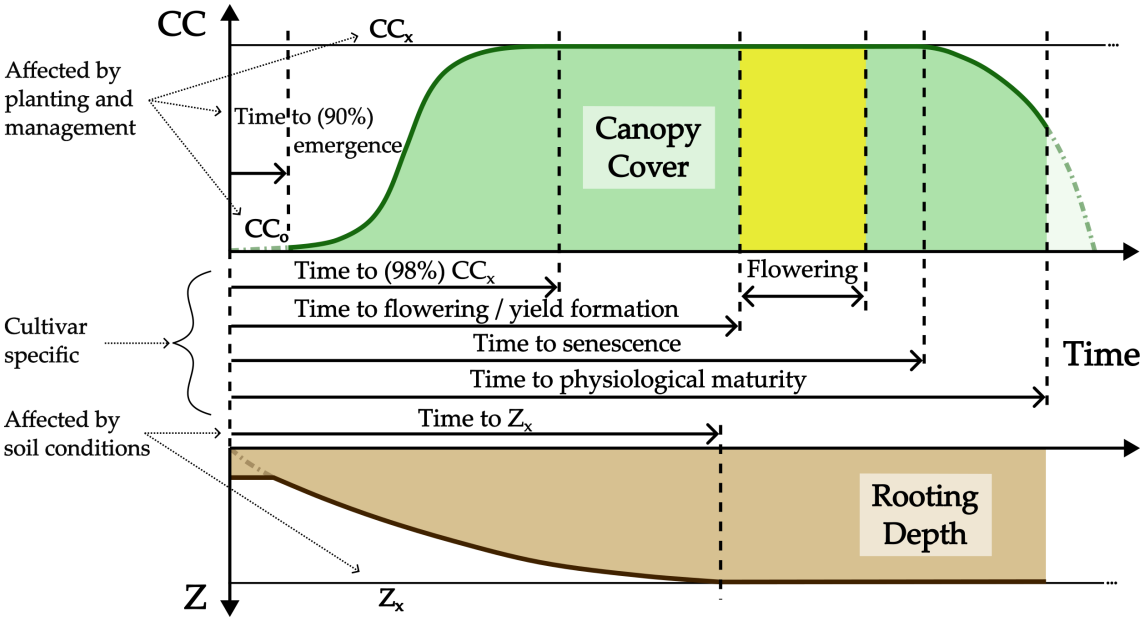

Figure 1.2: Canopy and root development along with the relative stages. Modified from figure 2.10b1 in Raes *et al.* (2025b).

1.3.3 Biomass and yield production

Transpiration is a vital process for plant life, as it drives the transport of nutrients from the root zone to the leaves, it cools down plant tissues, and allows CO_2 to be absorbed through the stomata. A strong relationship has been observed between transpiration (Tr) and biomass production (Steduto *et al.*, 2007). The accumulation of biomass in the model is therefore described by the following equation:

$$B_i = B_{i-1} + WP^* \frac{Tr}{ET_o}, \tag{1.7}$$

where B [t ha^{-1}] is the cumulative dry above-ground biomass, and WP^* [t ha^{-1}] is the normalised water productivity. As a final step, the dry yield is computed as

follows:

$$Y = f_{HI}\ HI_o\ B, \tag{1.8}$$

where HI_o [$\mathrm{kg}_y\ \mathrm{kg}_B^{-1}$] is the reference Harvest Index, which is cultivar-specific, and f_{HI} [-] is the adjustment factor that takes into account all the stresses (i.e. water, heat and cold) that the plant sustained during the critical stages of Y formation.

AquaCrop presents another indirect link between B and CC. As previously mentioned, the fertility stress is a parameter that affects CC in various ways. Its purpose is to describe the finite amount of nutrients in the soil, which are depleted as the season progresses. However, if other stresses affect plant growth more significantly, the nutrients are consumed at a slower rate, increasing their availability later in the season. This additional fertility is therefore modelled as a reduction in soil fertility stress as a function of B. If the other stresses fade away, CC would be able to grow more rapidly than normal. This example has been provided to highlight the complexity of the system and to address the precise mechanics before altering the state variables through DA.

1.4 Remote sensing in agriculture

The collection of data at the field scale is a common practice in the agricultural sector. From rain gauges and thermometers to soil moisture probes and pyranometers, there are countless tools used to understand, monitor, and predict environmental variables to guarantee an optimal crop health and production. However, the extent of the cultivated surface requires equally extensive coverage for those variables, and point measurements need to be extrapolated to cover such a large area. Point data may not be fully representative of the condition of a field, due to either sensor's errors or spatial heterogeneity (Hendrickx *et al.*, 2025). Technologies and scientific procedures are being developed to tackle the lack of spatial representativeness. For example, cosmic-ray neutron sensing (CRNS) is used to expand the spatial coverage of top soil moisture to hectometric scale (McJannet *et al.*, 2017). To cover regional to global scales, the airborne and spaceborne missions are needed, with satellites for earth observation being the most optimal solution to guarantee both good spatial

and temporal coverage.

Earth observing remote sensing systems are characterised by three main elements: the sensor that collects the signal, the radiative interactions with the Earth's surface, and the processing to clean the information and deliver useful products. Observations in the agricultural sector can be collected in the visible, thermal infrared (TIR), and microwave bands of the radiative spectrum. Visible and infrared signatures can provide useful information about plant development, canopy efficiency in absorbing photosynthetically active radiation, and biomass productivity (Moulin *et al.*, 1998), through either chlorophyll fluorescence or red and near-infrared (NIR) reflectance. Passive and active microwave sensors operating in the 1-2 GHz (L-band), 4-8 GHz (C-band), and 8-12 GHz (X-band) ranges can also be used to measure vegetation water, and are very often used for the retrieval of SM data (de Roos *et al.*, 2024; Draper *et al.*, 2012; Wigneron *et al.*, 2017).

This study uses the fraction of vegetation cover (FCOVER) from the Copernicus Global Land Service (CGLS) (2017) as its fundamental satellite product. Information about green vegetation cover can be obtained from the green and red bands in the optical spectrum and processed using machine learning to produce a reliable dataset. However, the information coming from Earth's surface, through the sensor's lenses, and being converted from spectral information to a simple number is certainly affected by errors. Different land cover classes within the same pixel can contaminate the signal, and imperfections in the transmission and processing of information can introduce noise, resulting in representativeness error and increasing uncertainty in general.

1.5 Data assimilation

Kalman filter

As mentioned above, the model can only ever provide an estimate of $\boldsymbol{x}$, which is a vector of n state variables. The Kalman Filter (KF) can be used to reduce the overall state estimation error through a combination of modelling and observations.

Its general form can be written as follows:

$$\hat{\boldsymbol{x}}_i^+ \;=\; \hat{\boldsymbol{x}}_i^- + \boldsymbol{K}_i(\boldsymbol{y}_i^o - \boldsymbol{H}_i\hat{\boldsymbol{x}}_i^-), \tag{1.9}$$

where $\boldsymbol{y}^o$ are the m observations (vector of size $[m \times 1]$), and the superscripts $^-$ and $^+$ indicate the *forecast* (obtained from the model at a certain timestep i) and the *analysis* of the variables (which will be fed to the model to compute $\boldsymbol{x}_{i+1}^-$), respectively. $\boldsymbol{K}$ $[n \times m]$ is known as the Kalman gain:

$$\boldsymbol{K} = \frac{\boldsymbol{P}\boldsymbol{H}^T}{\boldsymbol{H}\boldsymbol{P}\boldsymbol{H}^T + \boldsymbol{R}}. \tag{1.10}$$

Its purpose is to allocate the optimal weights to the forecasted state variables and observations while taking into account their respective variance-covariance matrices, ($\boldsymbol{P}$ $[n \times n]$ and $\boldsymbol{R}$ $[m \times m]$). The observation operator, $\boldsymbol{H}$ $[n \times m]$, connects the state space and observation space, which would otherwise be incomparable, since they usually belong to different physical domains. To understand how $\boldsymbol{K}$ works, it is sufficient to note that if the forecasted state uncertainty $\boldsymbol{P}$ is close to zero, $\boldsymbol{K}$ would also tend to zero, and Equation 1.9 would result in $\hat{\boldsymbol{x}}_i^+ \;\approx\; \hat{\boldsymbol{x}}_i^-$. Conversely, if the observation uncertainty is low, $\boldsymbol{K}$ tends to 1 and the innovation $(y_i^o - \boldsymbol{H}_i\hat{\boldsymbol{x}}_i^-)$ would be almost entirely included in the analysis.

It should now be clear that the nature of the KF requires knowledge of the uncertainties in both the model and the observations, and the former of these are generally unknown and vary in time and space. Furthermore, the error covariance between the modelled state variables and the associated 'observation' predictions (state mapped into observation space) $\boldsymbol{P}\boldsymbol{H}^T$ and the error variance matrix of the simulated 'observation' predictions $\boldsymbol{H}\boldsymbol{P}\boldsymbol{H}^T$ need to be fully described. The most effective strategy to solve these problems is to implement an ensemble Kalman Filter (EnKF). In fact, by perturbing the state variables $\hat{\boldsymbol{x}}_i^-$, the external forcing $\boldsymbol{u}_i$ and the model parameters $\boldsymbol{\alpha}$, it is possible to generate a set of simulations that differ slightly from each other, called an *ensemble*. At each timestep, a variance-covariance matrix $\mathbf{Cov}(\hat{\boldsymbol{x}}_i^-, \hat{\boldsymbol{y}}_i^-)$ $[n \times m]$ can be computed based on the spread of the ensemble members' state $\hat{\boldsymbol{x}}_i^-$ and 'observation' predictions $\hat{\boldsymbol{y}}_i^-$. Equation 1.10 can now be

rewritten in the following form:

$$\boldsymbol{K}_i = \frac{\mathbf{Cov}(\hat{\boldsymbol{x}}_i^-, \hat{\boldsymbol{y}}_i^-)}{\mathbf{Cov}(\hat{\boldsymbol{y}}_i^-, \hat{\boldsymbol{y}}_i^-) + \boldsymbol{R}}, \tag{1.11}$$

where the only external input is $\boldsymbol{R}$, which can be retrieved from the observation provider or by estimating the uncertainty of the sensor and the retrieval process.

However, there are downsides to using a filter like the KF. Its statistical foundations do not take into account physical conservation laws, meaning that the output of what would otherwise be a continuous function may exhibit discontinuities (Janjić *et al.*, 2014), and need to be constrained to the physical bounds set by parameters. To avoid discrepancies, one could update all the state variables and parameters involved in the system, or choose a Particle Batch Filter or Smoother. During the development of this thesis, these options were considered, but the KF turned out to capture the dynamics reasonably well, and it is ideal for future real-time updating and forecasting. Sticking to the Kalman filter raised interesting science questions that could be explored further.

State-of-the-art in crop modelling DA

DA in crop modelling has mainly been focused on parameter estimation using vegetation-specific observations at the field scale. The studies that applied DA for state updating have often used an EnKF (Ines *et al.*, 2013).

To date, the main assimilated satellite-borne variable is LAI, but DA procedures have also been developed for soil moisture, vegetation indices, reflectance, above-ground biomass, canopy nitrogen accumulation, and canopy cover (Dlamini *et al.*, 2023). Satellite-based LAI data have been used in regional crop DA to improve the spatio-temporal variability of crop simulations, because a correction of absolute values is sometimes difficult due to biases and coarse spatial resolution (Chen *et al.*, 2018; Jin *et al.*, 2022). Remotely sensed canopy cover (FCOVER) has been used less because most crop models have LAI as a state variable. However, canopy cover retrievals have less long-term biases (Tenreiro *et al.*, 2021) compared to models, when the latter use canopy cover as a state variable.

Nevertheless, assimilation of FCOVER introduces additional challenges, which are discussed later in Section 3.2.2. As of the publication date of this thesis, the author is not aware of any regional crop data assimilation based on FCOVER retrievals that have already been published in the scientific literature.

Chapter 2

Data and software

2.1 Study area

The analysis was carried out in the Piedmont region of Italy. The main reasons for this choice were the presence of a considerable production of winter wheat and the extensive availability of georeferenced data, which could be easily accessed via online portals such as *Geoportale Piemonte* and the *Data Warehouse Anagrafe Agricola*. Many layers were obtained from the former. The topsoil texture map was the smallest of these layers and therefore determined the extent (44°18'07.5"N 7°07'13.9"E to 45°45'26.8"N 9°10'26.8"E) of the Region of Interest (ROI) (Regione Piemonte, 2023a). This resulted in the exclusion of mountainous areas within the region, where wheat production is negligible.

Piedmont is the far-west part of the Po Valley, where the major Italian river rises. Being surrounded by the Alpine mountain range is a benefit from an agricultural point of view, because the solid accumulation of water on top of it provides a steadier water supply to the valley. The historical, cultural, and economic development has been shaped to coexist with such environmental conditions. At the mountain foothill, a complex irrigation system was developed even before 1850 and is still in use today, allowing for the cultivation of more water-demanding crops (Baird Smith, 1852). Meanwhile, in the plains around Alessandria, rainfed agriculture of winter wheat is still the major activity.

The spatial variability in crop production is not determined by precipitation dynamics. A heterogeneous precipitation pattern indeed characterises the region, and five different clusters can be identified (Baronetti *et al.*, 2018). However, the two clusters covering the ROI have similar precipitation distributions, with an average of

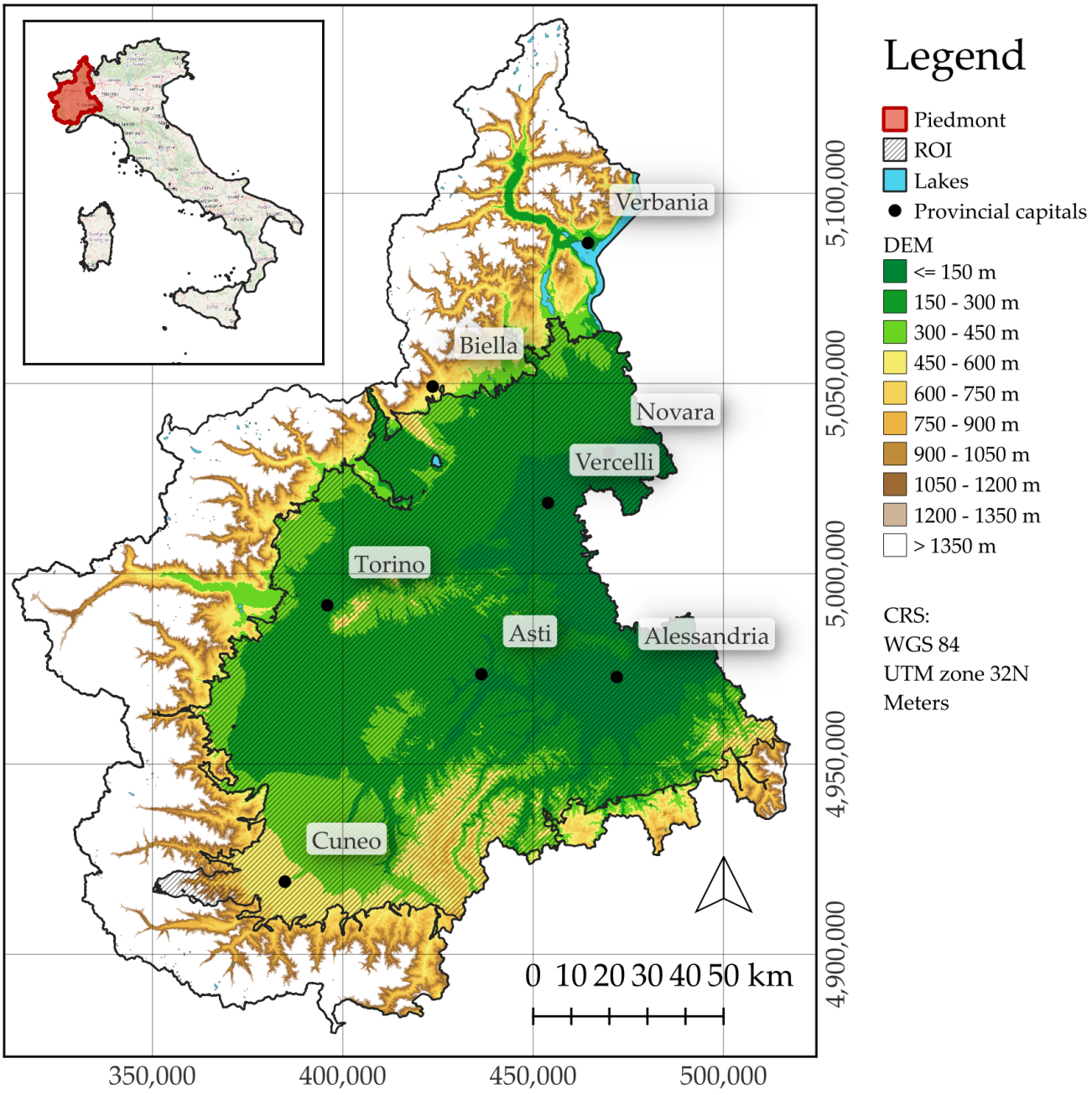

Figure 2.1: Infographic map of the Piedmont Region, Italy, with information related to elevation (Farr *et al.*, 2007), lakes (Regione Piemonte, 2017), provincial capitals and administrative boundaries (ISTAT, 2024). On top of it, the region of interest (ROI) considered for this study is displayed.

∼1000 mm of precipitation per year. This information has been used while choosing the calibration criteria explained later in Section 3.3.1.

The ROI was obtained from the borders of the vector geometries of the Soil texture map. Since the simulations were performed based on latitude and longitude, the map was transformed from the Universal Transverse Mercator projection of the World Geodetic System 1984 (WGS 84 / UTM zone 32N, EPSG:32632) to the WGS 1984 (EPSG:4326). It was then rasterised using the grid of the assimilated satellite-based CGLS FCOVER data as a reference (CGLS, 2017). This was done to ensure a perfect overlap of all subsequent layers cut with it. The grid had an angular resolution of 1/336° (∼300 m). A second ROI reference grid was obtained by reducing the resolution to 1/112° (∼900 m), while still ensuring a perfect overlap between the two. This double reference was necessary because the model would run at a resolution of 900 m to keep the computational costs limited, while the satellite products would be fed to the Land Information System without being upscaled.

2.2 Modelling tools

2.2.1 NASA - Land Information System Framework

AquaCrop has been designed to model crop development at the field scale (< 1 ha). To enable regional or global analyses, it has recently been integrated into the Land Information System Framework (LISF) (Kumar *et al.*, 2006; Peters-Lidard *et al.*, 2007), which was developed by the National Aeronautics and Space Administration (NASA) (De Lannoy *et al.*, in review). This framework is optimised for scalable runs with different Land Surface Models (LSM). It is a flexible tool that improves model estimates of land surface conditions by assimilating satellite or ground observations (Kumar *et al.*, 2006; Peters-Lidard *et al.*, 2007). It is written in Fortran 90, an efficient coding language that is widely used in high-performance computing systems.

LISF comprises a core program, the Land Information System (LIS) which hosts the models and assimilation module, a front-end processor (Land Data Toolkit, LDT), and a post-processor (Land Verification Toolkit, LVT). LDT is a key compo-

nent of data management and homogenisation. Among all the things, it is designed to *i)* read and process the *native* (or raw) data files from the format provided by their sources into a common grid, *ii)* apply quality control, and *iii)* generate the model initial conditions (Arsenault *et al.*, 2018). In this analysis, the post-processing procedure was performed outside of LISF.

2.2.2 AquaCrop 7.2 integration

AquaCrop is composed of a multitude of parameters and processes that can be adapted to the specific conditions of the analysed fields. In the recent LIS integration (referred to as LIS-AquaCrop), some processes, such as varying groundwater levels, crop rotation, perennial crops, and salinity stress, as explained by De Lannoy *et al.* (in review), have not yet been included. Future development may also improve soil information by providing the ability to define different layers with different textures and thickness.

Another important aspect of the integration of AquaCrop in LIS is that it does not consider lateral fluxes. Losses via runoff are simply taken out of the grid cell without being reallocated to a neighbouring cell. Consequently, LIS-AquaCrop simulations currently lack interaction between adjacent pixels with regard to the water balance. However, many models are characterised by these same horizontal settings, even land surface models such as Noah-MP (Yang *et al.*, 2011).

2.2.3 Additional tools

The process of collecting, interpreting, and managing raw georeferenced datasets was partly conducted using the free and open-source Quantum Geographic Information System (QGIS) in version Prizren 3.34.13. LIS simulations were run and stored on the Linux-based KU Leuven Tier-2 clusters of the *Vlaams Supercomputer Centrum* (VSC) High-Performance Computing (HPC) system. Access to its computational power was via an on-demand connection, either directly or by opening interactive sessions on the available version of Visual Studio Code. LDT and LIS outputs were managed and visualised using Python version 3.6.8.

2.3 Data

All the data useful for the analysis, and their respective sources, are listed below. Due to the heterogeneous nature of the various input sources, they must be preprocessed to ensure consistency. Spatial data has mainly been pre-processed in QGIS. They were integrated with the rest of the analysis through the Python-based VS-Code. Data management with Python and LIS was based on netCDF files, which use different georeferencing criteria compared to QGIS TIFF files: the former points at the centre, while the latter points at the top left corner of each pixel. It is important to manage the data carefully to avoid any subtle grid mismatch during the process.

2.3.1 Crop masks

A crucial aspect of the entire analysis is precisely locating when and where WW was cultivated in order to model WW (and perform the DA) in the correct locations. Years for which this information is missing cannot be taken into account. The first available products were the crop maps at the parcel level (1:2000) for the years 2021-2023 (Regione Piemonte, 2022, 2023b, 2024b). To extend this timeframe, other products were considered, such as the EU Crop Map 2018 (D'Andrimont *et al.*, 2021). To increase the number further, the possibility of generating additional ad-hoc maps was investigated using the WorldCereal "Private extraction" procedure and the "Custom crop type map" model, freely available on GitHub (Van Tricht *et al.*, 2023). However, both the EU Crop map and the custom crop maps were no longer necessary as soon as the new High Resolution Layer Croplands product of the Copernicus Land Monitoring Service was released (CLMS, 2025). This product covers the period between 2017 and 2021. For the overlapping year (2021), the parcel-level product was chosen due to its higher resolution.

The same preprocessing procedure was followed for each year. All layers were merged and transformed from WGS 84/UTM Zone 32N to WGS 1984. Then, a Boolean filter was applied to extract only the wheat fields, which were rasterised at a high resolution (10 m), and clipped to the domain. Using the 300-m reference

grid, the percentage of WW fields was computed for each pixel, and, finally, only those where at least 70% of the area was covered by wheat were kept. The same procedure applies to the Copernicus Crop Layers, except that they are distributed directly as 10-m resolution rasters.

2.3.2 Observed yield

Of all the required data, ground data plays a major role in many of the steps in this analysis. First, observed crop yield is used as a reference for crop calibration. Some simpler runs were performed using the AquaCrop GUI to minimise the discrepancy between the predicted crop yield and the actual ground yield observations. Next, soil texture, field extent, and soil depth, among other in-situ parameters, must be collected to enhance the model's accuracy and performance. Finally, at the end of the simulations, the yield is used to validate and quantify any improvement in the DA compared to the model alone. Given its importance, the more ground data, the better. Since the analysis is performed on a regional scale and over several years, it would be ideal to obtain validation data with the same spatial and temporal resolution.

Data Warehouse

One of the initial datasets that best covered the domain was the annual agricultural production for each crop type and municipality, which was provided by the Piedmont Region Data Warehouse (Regione Piemonte, 2024a). However, upon examining the distribution of the crop yield values, the modelled origin of the dataset became apparent. Almost all the values were confined to the range 5.5-6 $\frac{ton}{ha}$, or around 4.5 $\frac{ton}{ha}$, with a clear gap in between. Production estimates were actually obtained using the extent of the fields and a modelled yield for different climate zones.

GYGA

The second-best option for validating the analysis came from the Global Yield Gap Atlas (GYGA, 2021). This dataset attempts to harmonise various data sources

across the globe and is accessible at a national level, via climate zones and via stations. However, the resolution is quite low, and only one station is available across the entire ROI. Therefore, using GYGA data would only provide data for temporal validation.

RICA-CREA

Losing the spatial dimension would have reduced the quality of the validation procedure. Furthermore, if the observed data existed and were available only at a coarse scale, this would possibly mean that they were sampled and then aggregated. The GYGA researchers explained that they used the data with the best available resolution. In Italy, the source was Eurostat at the NUTS-2 level, corresponding to regional resolution.

Eurostat's data collection is based on a well-distributed randomised sampling across the EU. Samples are taken at the farm level and aggregated for privacy reasons and statistical significance. Each country has a designated entity responsible for data collection and management. In Italy, the governmental entity responsible for the collection is the *Consiglio per la Ricerca in agricoltura e l'analisi dell'Economia Agraria* (CREA), and falls under the name *Rete di Informazione Contabile Agricola* (RICA, 2023).

CREA provided the data aggregated at the municipal level, but only if there were at least five samples over a 10-year period. Higher resolution was not feasible, because it would not comply with the limits of the Italian Legislative Decree No. 196 of June 30, 2003 *"Codice in materia di protezione dei dati personali"* in terms of privacy. The sample density is low, so the values may not be representative of the entire municipality. However, they are the best approximation available for the level of coverage required by this research.

2.3.3 FCOVER and DMP

Fraction of green vegetation cover

The goal of this thesis is to study the potential of assimilating the fraction of green vegetation cover (FCOVER) product from the Copernicus Global Land Service (CGLS), which has near-real-time updates from 2014 and an angular resolution of 1°/336 (∼300 m) (CGLS, 2017). With a dekadal resolution (i.e., approximately 10 days), it is a ready-to-use product based on processed satellite images (Wolfs *et al.*, 2022). It is made available either as a 'near real time' (also called 'instantaneous') product or as 'historical' time series. The description of the full processing of satellite data is described in the Algorithm Theoretical Basis Document from Verger and Descals (2022). The main characteristics are reported here to give a panoramic view of its origins and quality.

PROBA-V Vegetation sensor was the source from 2014 to July 2020, then replaced by another pair of pushbroom sensors, Sentinel-3/OLCI A and B, with a similar spatial resolution of ∼300 m. Image processing starts with applying atmospheric distortion corrections and an algorithm to convert top-of-atmosphere (TOA) radiance to top-of-canopy (TOC) reflectance, and applying quality flags for pixels influenced by cloud coverage or covered by ice, snow, and water.

The core element in the production of FCOVER is the application of a neural network technique (NNT). NNT has been trained on a subset of SPOT Vegetation observations for PROBA-V Vegetation retrievals, and on a year of PROBA-V data for Sentinel-3/OLCI to ensure consistency between the two datasets. The training yielded the parameters needed for the near-real time processing. Outliers have been removed before both the training and the main procedure based on a multi-dimensional validity domain of spectral bands (Baret *et al.*, 2016). An exclusion of unphysical values out of range is finally performed on the output of the NNT. Areas characterised by evergreen broad-leaf forests have an independent NNT, and the historical product is further polished, filling small gaps in the data retrieval. However, this thesis involves the instantaneous product only; therefore, there are no interpolated values.

Dry matter productivity

The model validation in this thesis will employ satellite data of the Dry Matter Productivity (DMP, [kg ha^{-1} day^{-1}])(CGLS, 2018), a CGLS product with the same temporal and spatial resolution as FCOVER. There are several ways to approximate the daily dry biomass production estimates from space. Most often, it is deduced from the fraction of photosynthetically active radiation absorbed by the green elements of the canopy (FAPAR, [-]) through the Monteith's approach (Swinnen *et al.*, 2023):

$$\text{DMP} = SW \times \text{FAPAR} \times \varepsilon_c \times \varepsilon_{LUE} \times \varepsilon_T \times \varepsilon_{H_2O} \times \varepsilon_{CO_2} \times \varepsilon_{CUE} \times \varepsilon_{res} \times 20, \quad (2.1)$$

where SW [kJ m^{-2} day^{-1}] (0.2 - 3.0 μm) is the total shortwave incoming solar radiation, all the ε [-] terms are efficiencies, and the final multiplication is for the conversion to agricultural units. ε_c is the climatic efficiency, which considers the fraction of R that is useful for photosynthesis (0.4 - 0.7 μm) and is set globally to 0.48; ε_{LUE} is related to the biome-specific maximum light use, therefore the efficiency of the photosynthesis process under perfect conditions. ε_T and ε_{H_2O} are the temperature and water stresses, respectively. ε_{CO_2} is the carbon dioxide fertilization, and ε_{res} all the residual stresses like diseases and nutrients availability. An important parameter is the carbon use efficiency ($\varepsilon_{CUE} = 0.5$), which excludes the autotrophic respiration and characterises the difference between the net and gross dry matter productivity.

In the current DMP product, however, not all terms are considered yet. Both the residual efficiencies and the water stress are, in fact, excluded from the computation. It is nevertheless likely that future evolutions in the product will consider the effects of water scarcity (Swinnen *et al.*, 2023). ε_{CUE} may also need improvements to take into account the C fraction that plants use for nutrient uptake or defence mechanisms. This simplified version of the Monteith's approach makes DMP more like a 'potential' dry matter productivity than an actual observation of such a quantity, and this information must be considered while analysing the results of this study.

To summarise, DMP depends on meteorological variables (SW and T), land cover

maps and corresponding LUE look-up tables, and, most importantly, FAPAR. The latter is obtained through a similar procedure as FCOVER, where satellite retrievals pass through a NNT and get published as both 'near real time' and 'historical' products (Baret *et al.*, 2016).

Due to its incremental nature rather than cumulative nature and temporal incompatibility with the modelled B, it is not possible to use DMP directly in a DA process. Additionally, DMP also considers biomass growing below ground (Wolfs *et al.*, 2023), and does not include water and fertility stresses. However, it can still be important in the validation procedure, as it can be used to estimate the improvements in correlation with the model output for a derived variable similar to DMP: the positive dekadal increments of the above-ground biomass (ΔB_+).

Pre-processing

Even though the two products are retrieved independently, their correlation on the WW fields during the growing seasons is quite high (R = 0.80), as shown in Figure 2.2.

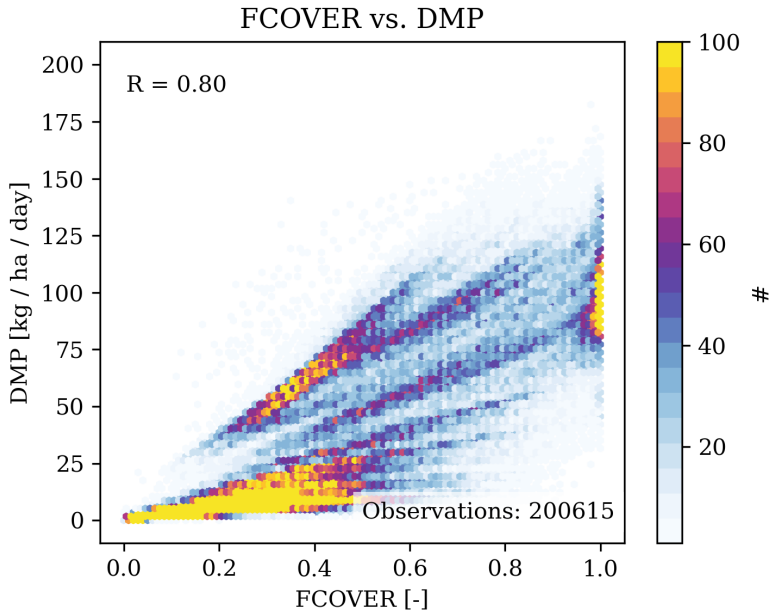

Figure 2.2: Scatter plots between the two Copernicus satellite products: Fraction of vegetation COVER (FCOVER) and Dry Matter Productivity (DMP).

Very little pre-processing has been applied to the CGLS FCOVER and DMP.

Firstly, as the DA only operates during the growing season, only images from these periods have been considered for assimilation and validation. To account for variability in growing season length, images were selected from mid-October to early August of the following year. Then, each layer was labelled with the end date of the dekad it referred to. Shifting them towards the middle of the dekad reduced the introduction of biases in the assimilation process during the growing stage. Finally, the crop masks were applied at the 300-m resolution to constrain the updates into LIS to the wheat fields only.

In addition to the data itself, FCOVER files contain supplementary information such as a quality flag and the root mean square error associated with each value. The former describes the processes that have been applied during the retrieval; in the high-resolution products, no interpolation method is used, and all outliers are filtered out. The observation error estimate has a lot of potential to be integrated into the DA procedure. However, the current LISF configuration does not include settings for a varying observation accuracy. Future research may include the varying accuracy to further optimise data assimilation.

2.3.4 Model input

Meteorological forcings

In AquaCrop, the required meteorological forcings are temperature T (daily maximum and minimum, T_{max} and T_{min}), rainfall, reference evapotranspiration ET_o and CO_2 concentration. Temperature and rainfall are obtained directly from the fifth generation of atmospheric reanalysis from the European Centre for Medium-Range Weather Forecasts (ERA5) (Hersbach *et al.*, 2020), with an angular spatial resolution of 0.25° (∼25 km) and an hourly temporal resolution.

ERA5 forecasts do not include ET_o directly, but it is calculated in LIS using the FAO 56 Penman-Monteith equation (Allen *et al.*, 1998):

$$ET_o = \frac{0.408\ \Delta\ R_n + \gamma\ \frac{900}{T+273}\ u_2\ \delta e}{\Delta + \gamma\ (1 + 0.34\ u_2)}, \tag{2.2}$$

where Δ [kPa °C^{-1}] is the slope of the vapour pressure curve, R_n [MJ m^{-2} day^{-1}] is the net radiation at the crop surface, γ [kPa °C^{-1}] is the psychrometric constant, T [°C] is the mean daily temperature, u_2 [m s^{-1}] is the wind speed at a height of 2 m, and δe [kPa] is the saturation vapour pressure deficit. These variables can all be derived from a limited set of meteorological forcings, available from the ERA5 product: T_{max}, T_{min} [K], pressure [kPa], specific humidity [kg kg^{-1}] wind speed u_{10} [m s^{-1}] (corrected to 2 m), and the shortwave radiation at the Earth's surface [W $^{-2}$].

Due to the coarse resolution of the forcing data compared to that used to run the model, a downscaling procedure was necessary. Simple spatial bilinear and temporal linear interpolation was applied to all the variables. Temperature was also corrected using a lapse rate based on the 1-km Multi-Error-Removed Improved-Terrain Digital Elevation Model (MERIT-DEM) (Yamazaki *et al.*, 2017).

In addition to the above meteorological forcings, AquaCrop also requires CO_2 concentrations as input. The default yearly record of the Mauna Loa (Hawaii) station measurements was used.

Crop parameters

Crop parameters are the most difficult to select and adjust due to their high variability in terms of environmental conditions, cultivars and local practices. Wherever possible, local or regional information was used; otherwise, FAO guidelines were considered. The key crop parameters are listed in Table 2.1. Any not listed are based on the default settings of the AquaCrop 7.2 GUI or its wheat default crop file.

The first aspect to determine is the length of the growing season, which is derived from the sowing and harvesting periods. Local farmer consortia recommend sowing early enough to allow the seedlings to experience their first frost once they have developed at least three leaves, but no earlier than the second half of October, due to disease pressure (Mosca and Reyneri, 2023).

While the size of a single seedling cc_o was kept at the default value of 1.50 cm^2, the sowing density was set to 400,000 m^{-2}, as the average value in the Emilia-Romagna regional guidelines (2025). All the crop stages were determined according to FAO Irrigation and Drainage Paper No. 56 (Allen *et al.*, 1998), and were slightly tuned

Table 2.1: List of winter wheat parameters that differ from the default crop file in the AquaCrop GUI. The length of the stages was expressed in calendar days based on Table 11 of Allen *et al.* (1998) and converted into thermal units using 2020 forcing data near Alessandria. The effects of fertility stress have been calibrated based on expert guidance regarding plant density and fertilisation practices.

Crop parameter	**Value**	**Source or criterion**
Max rooting depth	1 m	Rivieccio *et al.* (2020)
Time to Maximum rooting depth	918 GDD	Rasmussen and Thorup-Kristensen (2016) Rivieccio *et al.* (2020)
Time to Emergence	167 GDD	Allen *et al.* (1998) + Calibration
Time to Senescence	2080 GDD	
Time to Maturity	2694 GDD	
Time to Flowering	1281 GDD	
Flowering length	198 GDD	
CGC	0.004843 GDD^{-1}	Internal computation
CDC	0.004739 GDD^{-1}	$0.10 < CC_{final} < 0.20$
Soil fertility		
B reduction	60%	GYGA (2021)
Fertility stress	Ks_{CC_x} 5 % Ks_{CGC} 3 % $CC_{decline}$ 0.01 % GDD^{-1} Ks_{WP*} 64 %	Calibration

in the calibration process described in Section 3.3.1. The canopy decline coefficient (CDC) was changed to keep CC at harvest in the range of 10-20%, as shown in the default wheat crop file.

Another important aspect that depends on local conditions is the volume of soil that the roots can explore. Estimates of the soil rooting depth of Italy, with a resolution of 50 m by Rivieccio *et al.* (2020), were used to calculate the average maximum rooting depth within the domain. Coupling this information with an average root growth rate of 0.9-1.2 mm °C day^{-1} (Rasmussen and Thorup-Kristensen, 2016), determined a root growth period after germination of 918 °C day.

Soil parameters

As previously mentioned at the beginning of this Chapter, the ROI was determined based on the highly detailed Topsoil Texture map, which is part of the *Carta dei suoli 1:50.000* of *Geoportale Piemonte* (Regione Piemonte, 2023a). This layer provides a fine spatial variability of soil classes based on the USDA soil texture classification

and was consequently matched with the default soil parameters in AquaCrop, i.e., *i)* saturation, *ii)* field capacity, *iii)* permanent wilting point, and *iv)* hydraulic conductivity as described by De Lannoy *et al.* (2014). The Topsoil Texture map was implemented after the homogenisation performed by the LDT toolkit.

Management and fertility

Field conditions and management rarely match the theoretical optimal requirements, so they can significantly impact both crop development and the final harvest. The causes of the yield gap (i.e. the difference between the actual and potential yields under water-stressed conditions) are difficult to understand; they can be attributed to environmental pressures, harvesting efficiency, or many other factors that depend heavily on farmers' decisions.

In AquaCrop, this problem is approached from two angles. First, a calibration procedure is performed to determine the crop's particular susceptibility to environmental stresses and the extent to which practices can influence this. For instance, a reduction in CC_x due to water deficiency may be minimal (or even negligible) if the crop is densely planted. Conversely, the CGC reduction during the growth stage or CC_x decline during the mid-season may be influenced by a lack of nutrients. However, if the fields are fertile or the farmer supplements them with additional fertilisers, this reduction can be mitigated. Secondly, it is possible to use the information contained in the yield gap to trace system inefficiencies from a lack of biomass production. This is taken into account by reducing WP^* to establish a consistent relationship between stresses and yield reduction.

In fact, the yield gap (defined as the relative difference between the water-limited potential yield and the actual yield, i.e. $Yg = (Yw - Ya)\ Yw^{-1}$) was found to be approximately 40% (GYGA, 2021), so most of the stress impacts were attributed to WP^*, as farmers in this region depend on highly efficient mechanised operations, dense planting, and external inputs.

Chapter 3

Methodology

The methodology in this chapter covers both the structural model and DA background information, and the experimental setup. The background information in Sections 3.1 and 1.5 informs decisions for the experimental setup in Section 3.3, but covers cumbersome technical details that may be skipped by the scientific reader.

The study area was already introduced in Section 2.1 and the timeline of the experiments is shown in Table 3.1. This table also summarizes the availability of key datasets in time, which determined the experiment setup. It can be seen that the main limiting factor for the temporal extent of the analyses is the availability of crop masks and yield observations over the selected area. Details will be provided below.

Table 3.1: Time coverage of the major elements of the analysis. The limiting factors that bounded the analysis between 2017 and 2023 are the availability of crop masks and the field observations.

	<	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022	2023	2024	>
Meteorology	ERA5 reanalysis																
Satellite data						PROBA-V							Sentinel-3/OLCI				
Calibration						AquaCrop GUI											
Experiments			Spin-up						Deterministic run								
			Spin-up						M1 Ensemble OL & DA								
			Spin-up						M2 Ensemble OL & DA								
			Spin-up						M3 Ensemble OL & DA								
Crop masks									CLMS Crop Type				Geoportale Piemonte				
Yield observations					RICA-CREA												
	GYGA Yw					GYGA Ya								Eurostat Ya			

3.1 Model state and parameters

A proper model setup and an understanding of the state propagation in time are crucial to successful DA experiments. In this section, we first briefly identify model

state variables and parameters, and then discuss some model challenges that arose from tracking state variables and parameters.

3.1.1 Procedure

The state variables in AquaCrop are CC, B, and the SM vector $\boldsymbol{\theta}$. These variables propagate dynamically in time, in response to input forcings and physical laws captured within the model. The nature of the state responses is determined by multiple parameters, which are by definition not varying in time and calibrated prior to the simulations. The calibration is based on the default WW crop file available in AquaCrop, and the parameters involved are listed in Table 2.1. A detailed description can be found in Section 3.3. Note that AquaCrop will transform the calibrated parameters to state-dependent versions during the model simulation.

The state variables CC and B will be perturbed and updated, and a variation criterion for the planting date and CCx (both parameters) will be included as well in one of the experiments. However, both in the calibration procedure and the creation of DA routines, the model presented some challenges that needed to be addressed.

3.1.2 Challenges

Implementing the entire ensemble DA system presented a series of challenges related to tracing model state variables and parameters. Some were strictly related to AquaCrop 7.2; others arose from the interaction with the DA procedure. Depending on the challenge, it was either directly addressed directly in this thesis or it is reported for further investigation.

It will be discussed later that DA could involve a perturbation and updating of a selection of state variables, meteorological forcings, and model parameters. To keep things simple, the analysis initially considered the CC variable only. On the other hand, there is an additional set of variables that may interfere with the propagation of the updates within the model: the dummy variables.

From a coding perspective, the model is split into different parts to represent the dynamics in self-contained routines. This leads to the creation of multiple dummy

variables, which may be a combination of other core parameters or state variables. When DA is added as an artificial step to the iterative procedure of model state propagation, it is necessary to check that all variables are physically self-consistent, incl. state variables that are not updated by the DA algorithm.

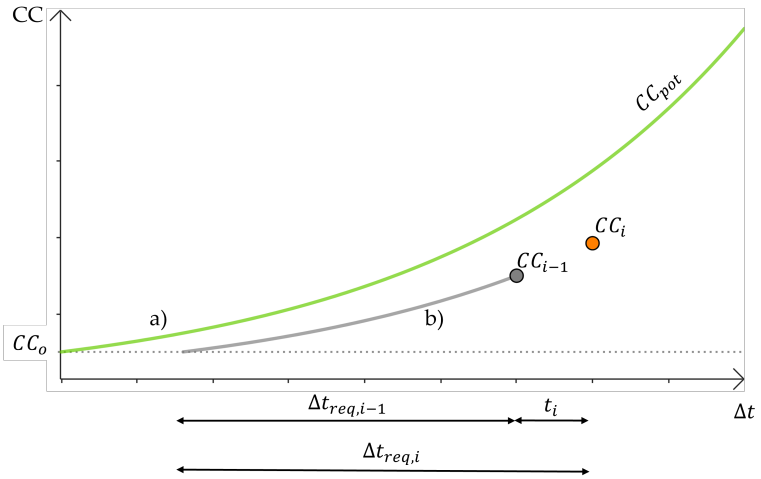

Figure 3.1: Incremental implementation of the exponential canopy growth. The green line (a) shows the function described by Equation 1.2, with soil fertility stress being the only applied stress. After suffering additional stresses, CC_{i-1} is lower than $CC_{pot,i}$. The new exponential curve (b) is forced to intercept CC_{i-1} and, together with the CGC_{Adj} for day i, determines the required growing period $\Delta t_{req,i-1}$ from CC_o to CC_{i-1}. The canopy cover on the following day (CC_i) is computed directly from the function (b). The slope of (b) may be shallower if $CGC_{Adj} < CGC$.

In order to fully understand how the information is propagated into the model, the equations describing CC development must be analysed. Equations 1.2, 1.3, and 1.5 can be used to determine evolution in the absence of stresses of any kind. However, they do not incorporate information collected during the early stages of the growing season. Therefore, incremental functions must be derived from them. For example, Equation 1.2 has been rewritten as the following set of equations:

$$\begin{cases} \Delta t_{req,i-1} = \dfrac{ln\left(\dfrac{CC_{i-1}}{CC_o}\right)}{CGC_{Adj}} & i) \\ \Delta t_{req,i} = \Delta t_{req,i-1} + t_i & ii) \\ CC_i = CC_o \; e^{\Delta t_{req,i} \; CGC_{Adj}}, & iii) \end{cases} \tag{3.1}$$

where Δt identifies cumulative temporal or thermal units (i.e., calendar days or

GDD) and t refers to a specific day (i.e., 1 for calendar days, or a certain amount of GDD_i). These are listed in Figure 3.1 to aid understanding. As the reader will notice, the growth coefficient has been substituted with $CGC_{Adj} = CGC(1 - Ks_{CGC})Ks_W$, to consider the effects of fertility and water stresses.

The same procedure can be applied to Equation 1.3, but with an additional step:

$$\begin{cases} \Delta t_{req,i-1} = \dfrac{ln\left(\dfrac{CC_{x,SF}^2}{4\ CC_o\ (CC_{x,SF} - CC_{i-1})}\right)}{CGC_{Adj}} & i) \\ \Delta t_{tot} = \Delta t_{req,i-1} + t_i + (\Delta t_{CC_x} - \Delta t_{i-1}) & ii) \\ CC_{x,Adj} = CC_{x,SF} - \dfrac{CC_{x,SF}^2}{4\ CC_o}\ e^{-\Delta t_{tot}\ CGC_{Adj}} & iii) \\ CC_i = CC_{x,Adj} - \dfrac{CC_{x,Adj}^2}{4\ CC_o}\ e^{-\Delta t_{req,i}\ CGC_{Adj}}. & iv) \end{cases} \tag{3.2}$$

Also the upper boundary CC_x is substituted by the dummy variable $CC_{x,SF} = CC_x(1 - Ks_{CC_x})$. Then, if $CC_{i-1} \ll CC_{pot,i-1}$, it means that the plant has experienced stress. In this scenario, as illustrated in Figure 3.2, these stresses can be interpreted as a "time delay" and/or a decrease of CGC of the potential development. However, the growth stages are fixed, and the time needed to reach maximum canopy cover Δt_{CC_x}, i.e. the moment when the plant cannot grow any more, is also fixed. Therefore, an adjusted upper boundary ($CC_{x,Adj}$) is calculated by intersecting the delayed potential curve at the time of maximum canopy cover, Δt_{CC_x}. $CC_{x,Adj}$ is then used as the new asymptote to which the exponential decay aims, and the next daily step in CC is computed.

CC finite differences function divergence

The aforementioned implementation of the iterative equations is flawed in AquaCrop 7.2, specifically in the computation of $CC_{x,Adj}$. As can be seen in Figure 3.2, $CC_{x,SF}$ is the asymptote of the potential curve, while $CC_{x,Adj}$ intercepts the new curve at the point corresponding to Δt_{CC_x}. While this may seem a minor difference, the problem lies in the iterative nature of the procedure. In fact, at each time step, the upper limit is underestimated, which slightly dampens canopy development. The fundamental equation is an exponential decay; if the inaccuracy occurs far from the

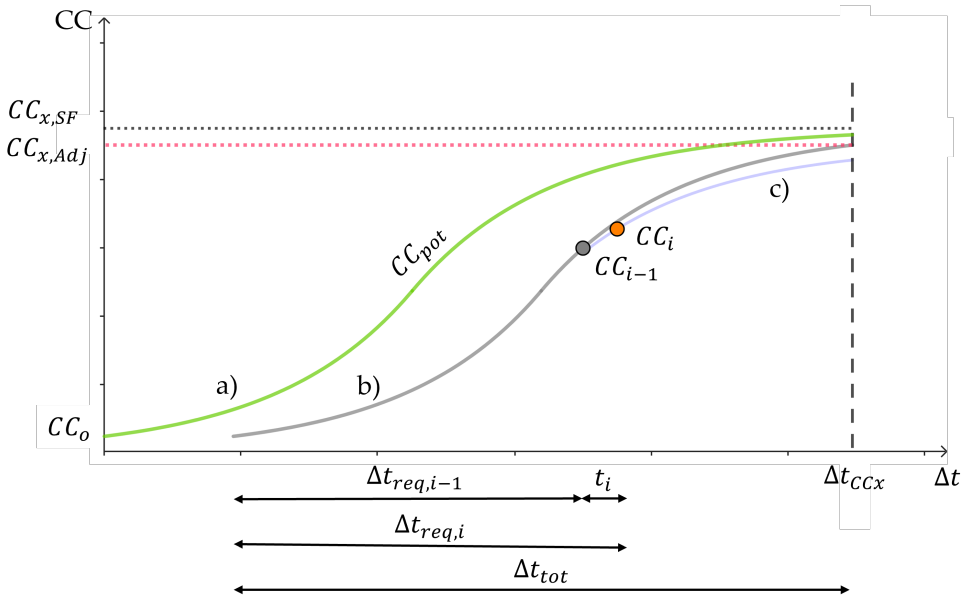

Figure 3.2: Incremental implementation of the exponentially decaying canopy growth. The green line (a) shows the potential growth in the presence of soil fertility stress. The function (b) is forced to pass through CC_{i-1}. The recalculated growth period Δt_{tot} is (usually) shorter than the initial one; therefore, the maximum reachable value is lower. The new upper limit $CC_{x,Adj}$ is defined as the value of curve (b) corresponding to the total growth period, Δt_{tot}. The new time-step is computed using the trajectory (c), which corresponds to equation *iv)* 3.2.

asymptote, the effect is negligible. However, as it approaches the asymptote, the divergence increases, reducing the achievable CC_x even without any stress being applied.

The root of the problem arises from the interpretation of Δt_{CC_x}: by definition, it is the time (or the amount of GDD) required for the plant to reach 98% of CC_x. Thus, in the iterative procedure, the new $CC_{x,Adj}$ must be divided by 98% to maintain consistency between the two versions of the growth dynamics.

While the problem can be solved in temporal units as described above, another step is required for the thermal unit settings. Specifically, the cut-off percentage in the code may not always be the same, since more than one GDD can occur per day and the time interval between the start and end of growth Δt_{CC_x} may fall in the middle of the day. The GDD in Δt_{CC_x} gets rounded up to complete the last day of growth. Therefore, dividing $CC_{x,Adj}$ by 98% may lead to an overestimation of the upper limit, resulting in worse consequences for the model: $CC_{x,Adj}$ could exceed

one, producing unphysical results. The safest way to modify the iterative equation is to introduce the following ratio:

$$\text{ratio} = \frac{CC_{pot,98}}{CC_x} = 1 - \frac{CC_x}{4\ CC_o} e^{-CGC\ T_{CC_x}} \tag{3.3}$$

in the equation *iii)* 3.2:

$$CC_{x,Adj} = \frac{CC_{x,SF} - \frac{CC_{x,SF}^2}{4\ CC_o}\ e^{-T_{tot}\ CGC_{Adj}}}{\text{ratio}}. \tag{3.4}$$

It is yet to be clarified and tested whether the ratio should be a fixed value or a function of the varying $CC_{x,SF}$ and CGC_{Adj}. Hence, this bug fix has not been included in the thesis. The correction may be applied in version 7.3 of AquaCrop.

Smoothing function

While testing and calibrating the crop development in both LIS and the GUI, a small discontinuity appeared as soon as the crop got closer to CC_x. This was caused by a change of routines in the code and resulted in an almost negligible drop in CC.

As previously mentioned, the growing stage in AquaCrop ends when $CC = 0.98 \cdot CC_x$. The following day, the canopy cover should be set to CC_x; however, this would result in a discontinuity in the curve. Thus, the developers added a "smoothing function". Based on the original function in Eq. 1.3, it does not depend on CC_{i-1}, thus it could introduce a new discontinuity by itself, which would defeat the initial purpose. Additionally, this function prevents the canopy cover from ever reaching its limit.

These last two statements are the reason why this almost negligible problem affected the implementation of DA. In fact, if CC is not computed based on the previous day in the mid-season, neither the perturbation nor the DA updates shown in Figure 3.3 would be integrated into the model. The first attempt to solve the problem was to set the dummy variable $CC_{x,Adj} = CC_{i-1}$, as this is theoretically correct. However, this resulted in a downward drift of the canopy cover, since $CC_{x,Adj}$ was used as an asymptote. The final solution involved removing the smoothing

function and reintroducing the upper discontinuity, while allowing noise and updates to be applied to the system.

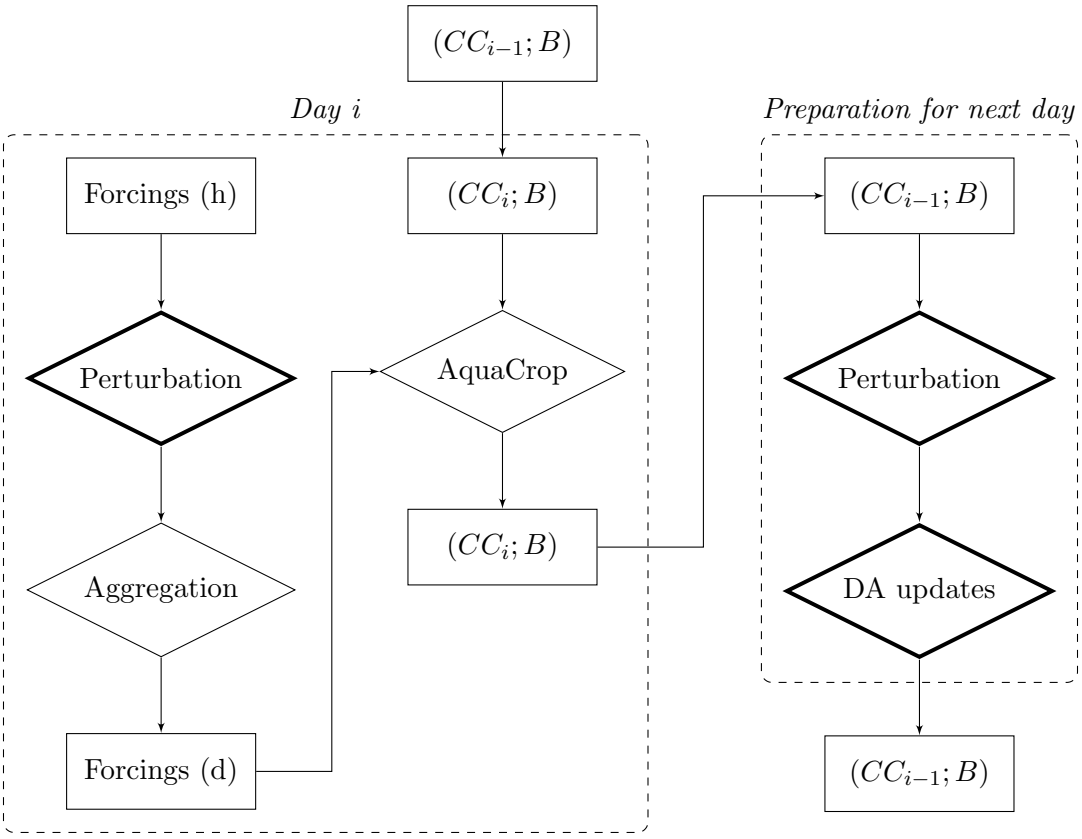

Figure 3.3: Iterative AquaCrop workflow in LIS, including the integration of DA modules (in bold). Rectangles represent data or state variables, while the diamonds highlight the main routines. Each day, the hourly (h) forcings are perturbed (only in the ensemble simulations), aggregated at the daily (d) resolution, and fed to the main AquaCrop module. Simultaneously, the CC value from the previous day is assigned to the current day's CC value and enters the model. Once the new value has been computed, the day ends, and CC_i becomes CC_{i-1}. B follows the same path, but is recomputed directly. Finally, the state variables are perturbed (in the ensemble simulations only) and get updated (in the DA simulations only).

Dormant phase

The latest model improvement relates to the yield formation. AquaCrop has interesting functionality with regard to how plants respond to prolonged water stress. As soon as there is insufficient water to sustain crop growth, CGC is affected, slowing down the development. If the stress persists, early senescence may be triggered, which actively reduces CC at a rate that varies depending on the severity of the

stress. But, when the plant reaches a certain threshold $min(CC_o, CC_{dorm})$, it enters what is called a "dormant phase". During this phase, the plant slows down its metabolism in an attempt to preserve a certain amount of water in its tissues. If the stress continues, ET_o is used as a cumulative metric to determine the critical point at which the plant dies. Otherwise, if the soil water content is replenished in time, the plant can start growing again.

Two problems in this module resulted in errors in the yield formation quantities. Firstly, the plant can enter this routine after the growth stage, but it will not be able to recover, resulting in a simple delay to its death. The second problem is code-related. In the dormant phase, the plant metabolism stops, so there should be zero production of new biomass and yield. However, when $CC_o > CC_{dorm}$, the plant continues to form yield, which leads to an overestimation of the harvestable product by 20-25%.There is more than one solution to this situation, yet the most suitable for the physiology of a determinate crop, such as winter wheat, is to remove the dormant phase after Δt_{CC_x}.

3.2 FCOVER DA

3.2.1 Procedure

The mathematics behind DA have already been introduced in Section 1.5. From a practical point of view, the model uncertainty will be approximated through ensembles in this thesis, and an ensemble Kalman filter is chosen as the DA algorithm.

As previously mentioned, an ensemble can be generated by perturbing the state variables, input, and parameters. This step must be performed carefully since, as with the DA updates, it can break the model's physical consistency. To limit incongruences, perturbation settings and constraints must be chosen. They perturbations can *i)* be additive or multiplicative, *ii)* follow a normal distribution (around a mean of zero for the additive settings and one for the multiplicative settings), but *iii)* with the exclusion of extreme values over a certain threshold, and *iv)* with temporal and/or spatial correlation, or *v)* a cross-correlation matrix with the other perturbed

variables. Finally, to force the OL ensemble mean to follow the same pattern as the deterministic run, LIS can be configured with perturbation bias correction (Ryu *et al.*, 2009). Its application consists of the conservation of an unperturbed ensemble member, which is used as a reference for the whole ensemble. Then the difference between the ensemble mean and the unperturbed member is subtracted from all the other particles.

Next, during the DA process, the state variables are updated each time there is an observation:

$$\begin{bmatrix} \widehat{CC} \\ B \end{bmatrix}^{+}_{j,i} = \begin{bmatrix} \widehat{CC} \\ B \end{bmatrix}^{-}_{j,i} + \frac{\begin{bmatrix} \sigma^2_{\widehat{CC}^-} \\ \mathrm{Cov}(\widehat{CC}^-, \widehat{B}^-) \end{bmatrix}_i}{\sigma^2_{\widehat{CC}^-,i} + \sigma^2_{FCOVER}} (\mathrm{FCOVER} - \widehat{CC}^-)_{j,i}. \tag{3.5}$$

In the aforementioned equation, j represents the ensemble member, i the time step, FCOVER is the assimilated satellite observation (described in Section 2.3.3), and σ^2 is the error variance of the simulated or observed canopy cover, i.e. for , $\sigma^2_{\widehat{CC}^-}$ and σ^2_{FCOVER}, respectively. The uncertainty of FCOVER is fixed in time, as mentioned in Section 2.3.3. The error covariance between the two state variables involved is described by the following equation:

$$\mathrm{Cov}(\widehat{CC}^-, \widehat{B}^-)_i = \frac{1}{N-1} \sum_{j=1}^{N} \left(\widehat{CC}^-_{j,i} - \overline{\widehat{CC}^-}_i \right) \left(\widehat{B}^-_{j,i} - \overline{\widehat{B}^-}_i \right), \tag{3.6}$$

where N is the number of ensemble members, and the overline $\overline{(.)}$ shows the ensemble mean.

3.2.2 Challenges

The implementation of DA to update the cumulative state variables B and CC poses some challenges. Unlike non-cumulative variables, their development in time is quite heavily dictated by model dynamics that are difficult to perturb and update. Growth stages depend heavily on sowing dates, and canopy cover is characterised by both upper and lower boundaries.

Boundaries

It is obvious to state that canopy cover is bounded between zero and one. However, additional constraints further narrow the valid range within which the physical behaviour is observed. In the early days of the season, the seedling is protected from dying ($CC_i \leq 1.25 \cdot CC_o$), and the curve cannot be altered. Then, the growth becomes sensitive to water stress; however, even if the soil is well watered, the crop cannot grow beyond the CC_{pot} curve. Furthermore, throughout the growing season, perturbations are bounded from below by CC_o to prevent the artificial death of some ensemble members. Finally, once the natural senescence is triggered, both perturbations and DA updates have little to no effect on the canopy cover trajectory, because the model forces the plant to reach the end of the growing season. In summary, perturbations and updates have therefore only been enabled within the physical boundaries of the model, and where their impact would be effective.

The presence of boundaries poses an additional problem when a variable is perturbed, as the ensemble mean begins to drift away from the unperturbed trajectory, thereby introducing a bias. This can be compensated for by applying the perturbation bias correction, but if the modelled canopy cover is already close to the boundary, the ensemble spread would reduce again to almost zero. The upper boundary identified by CC_{pot} is the one that interferes the most. In fact, it prevents many positive DA updates, leading to the introduction of another bias compared to the open-loop.

The negative effects of this limitation are exacerbated when the growing seasons of the observations and the model are not synchronised, because while one may capture the dynamics, even a small shift can prevent updates in the correct direction. The sowing date can be considered as an additional boundary, so setting the model correctly is crucial, though this is almost impossible to predict.

LIS-AquaCrop initialisation

A core aspect of how AquaCrop handles crop development lies in its initialisation procedure. Even though multiple seasons can be simulated in sequence, each one of

them is run independently, and the single link between them is the propagation of the soil water content *i.c.*, to guarantee continuity.

Before a one-year simulation begins, the full record of climat input is read and interpreted. This is used to compute the GDD for each day and to convert the length of the growing stages from thermal units to calendar units for that year. The current version of LIS-AquaCrop does not allow for perturbation during the simulation of either the input temperatures or the length of the growing stages during the simulation, as this would lead to inconsistencies in the code routines.

Fixed growing stages

The uncertainty surrounding the planting date may not be too relevant if the growth stages could vary. However, it can be dangerous to interfere with the stages, since the crop has been specifically calibrated with them, and the propagation of the effects must be analysed in depth.

Another strategy to reduce the boundary constraints is to introduce a varying sowing date. For example, if the observations show higher canopy cover than the potential one, an ideal solution may be to redefine the growing season with an earlier planting date. In such a scenario, the model would need to be restarted from there to capture the new soil moisture dynamics.

The LIS-AquaCrop setup used in this analysis does not permit this. Hence, a simpler criterion has been developed to allow the sowing date to vary *a priori*, i.e. before the simulation starts.

3.3 Experiment setup

The experiments cover a deterministic run (Det), ensemble runs (also known as an open-loop run, or OL), and data assimilation (DA) runs. All of them are preceded by a spinup simulation, with a calibrated model discussed below.

The simplest and fastest configuration is the deterministic setup, in which the model is run with a precise set of meteorological forcings and crop parameters, and with no interference throughout the simulation. Thus, both the perturbation and

Table 3.2: Experiment setup summary, with all the additional modules turned on (Y) or off (N).

		Forcing perturbation	**State perturbation**	**Perturbation bias correction**	**Parameter variability**	**DA updates**
Det		N	N	N	N	N
M1	OL	Y	Y	Y	N	N
	DA	Y	Y	Y	N	Y
M2	OL	Y	Y	N	N	N
	DA	Y	Y	N	N	Y
M3	OL	Y	Y	N	Y	N
	DA	Y	Y	N	Y	Y

DA update modules (shown in Figure 3.3) are ignored. This is used as a reference to quantify how much the generation of an ensemble run diverges from it. Since the OL is just a practical intermediate step to obtain model forecast uncertainty estimates that can later be used in the DA, it should not influence the behaviour of the model, and the mean of the ensemble OL should, in theory, be equal (or close to) the deterministic run. To check the consistency, the deterministic run will be compared with the ensemble mean of the OL simulation to decide if there are any drifts between both simulations.

The DA and its challenges are discussed in Section 3.2. Through DA, the hope is to obtain better results than with the OL simulation, and to quantify this, the ensemble mean DA output will be compared against the ensemble mean of the OL and deterministic output. An important challenge to overcome in the DA is the CC_{pot} hard boundary. To optimize the DA design, three different setups, or modes, of ensemble perturbations have been tested, and their associated three DA simulations were evaluated. A summary of the settings used for the three modes is shown in Table 3.2 and discussed in detail below.

3.3.1 Crop calibration

Crop calibration was performed in two steps in the AquaCrop GUI. First, the FAO stages for WW must be converted from calendar days to GDD. As this procedure is highly dependent on the input data, a minimal climate analysis was performed to select an appropriate location and year from which to extract representative data from the ERA5 reanalysis. Due to the proximity to the Alpine mountain range,

there was a risk of heterogeneous climate conditions throughout the domain, which would either necessitate the use of spatially aggregated data, or the calibration of multiple crop files. However, by monitoring the variability of the maximum and minimum temperatures, it was found that this was not the case, and ERA5 data over the Alessandria province was selected. In terms of temperature patterns, 2020 was selected from the period 2014–2022 because it was the year closest to the area's average climate.

Second, after converting from temporal to thermal units, the stages were fine-tuned using the full set of selected parameters. The aim was to approximate the average GYGA actual yield and to capture visually the interannual variability. Following a simple deterministic run within LIS, the bias between the average yield from the model and the GYGA reference was confirmed to be around -0.2 t ha^{-1}.

3.3.2 Spin-up and deterministic reference

Models must be initialised with some initial conditions. In particular, AquaCrop requires an initial soil water content for each specified compartment. As it is difficult to choose the exact *i.c.* that could describe the state of the system at that specific moment, this selection is mainly determined by the operator. To minimise this impact and allow the model to stabilise around a reasonable and independent pattern, a spin-up run is performed. A 6-year spin-up was generated for each experiment in this analysis, between October 2010 and October 2016. The deterministic simulation continues thereafter and will only be evaluated from 2016 through the end of August 2023. The three OL and DA runs start from the same restart file, with the same randomiser seed per OL and DA pair, and are also evaluated through August 2023.

3.3.3 Ensemble simulations

Three modes of ensemble perturbation are introduced below. All modes use perturbed forcings (shortwave radiation SW and rainfall P) and state variables (B and CC) as listed in Table 3.3. An error cross-correlation of 0.5 was assigned to the

Table 3.3: Ensemble perturbation parameters for forcings (SW and P) and state variables (CC and B). The additive (+) perturbation follows a normal distribution around zero. Likewise, the multiplicative (×) one is based on a log-normal distribution around one.

	mode	$\boldsymbol{\mu}$	$\boldsymbol{\sigma}$	**t-corr**	**cross-correlation**			
SW	×	1	0.4	24 h	1	−0.80	-	-
P	×	1	0.6	24 h	−0.80	1	-	-
CC	+	0	0.01 [-]	-	-	-	1	0.5
B	+	0	0.01 [t/ha]	-	-	-	0.5	1

state variables due to their close relationship within the model. The modes differ in whether the parameters are also varied and whether perturbation bias correction is activated or not.

Mode 1

The first experiment involves perturbed forcings and state variables with the bias correction for perturbation turned on. It was predicted that the perturbation bias correction would constrain the spread too much to allow the updates to force the model out of its original path. This would result in a consistent OL-Det overlap, but little to no improvements in the DA.

Mode 2

The second mode involves perturbed forcings and state variables without the bias correction for perturbation. The focus is on quantifying the benefits and limitations of this setting in order to assess its necessity, although its application is more mathematically rigorous. The OL metrics may degrade, but they should recover in the DA run, possibly leading to an overall improvement.

Mode 3

As stated in Sections 1.5 and 2.2, parameters can be perturbed alongside state variables and forcing inputs. The length of the crop stages would ideally be adjusted to correct the sowing date estimates and account for calibration uncertainty. To this end, a sowing date that varies based on rainfall criteria has been introduced

in a manner similar to that available in the GUI. When a one-year simulation is initialised, the perturbed rainfall is read. Then, by defining a sowing window and the cumulative amount of rain that must fall within a certain period and the number of occurrences required, the sowing date for each ensemble member can be determined. For this analysis, two events of at least 15 mm of cumulative rain must occur within a maximum of four days each between October 4th and November 3rd. If this does not occur, the seeds are sown on the last day of the sowing window. The sowing window is centred on the sowing date of the deterministic and Mode 1 and 2 runs.

During some tests, the maximum values of FCOVER were seen to exceed the chosen CC_x, therefore falling outside the physical domain of the model. While modes 1 and 2 handled this situation by forcing the perturbations (or later DA updates) to stay below the potential line, this experiment introduces a varying CC_x around the calibrated value to soften the boundary and allow some members to get closer to the observations.

It is hypothesised that this experiment will present additional spread in the early and mid-season (via the rainfall criteria and varying CC_x, respectively), which could improve the metrics through DA without degrading OL precision.

3.4 Validation

The experiments have been evaluated from multiple sides in an attempt to gain a tomographic view of overall performance and identify areas where data assimilation has enhanced the model output, as well as aspects requiring further development.

Canopy cover and the dry above-ground biomass were compared to the FCOVER and DMP satellite products, respectively. As the former was directly assimilated, CC metrics are expected to improve in all circumstances, by design. The hope is that the trivial improvements in CC would propagate to B together with the direct updates of B to also improve the unobserved B and all other model variables. Therefore ensemble mean OL and DA output of ΔB_+ was evaluated against (not assimilated) DMP satellite observations. Finally, improvements in yield estimates were validated against the observed yield surveyed by RICA-CREA.

Table 3.4: Aggregation criteria. Spatial aggregation consists of taking the linear average of crop pixels within the coarser grid (for FCOVER and DMP) or the municipality (for yield). Temporal aggregation is performed by taking the average across the growing seasons in three different clusters each month (10 days, 10 days, and the remaining chunk).* B has been converted to daily increments (ΔB_+) prior to aggregation.

	Resolution		**Aggregation**
	Spatial	**Temporal**	
FCOVER, DMP	1/336°	∼10 d	Spatial
Modelled CC, B	1/112°	1 d	Temporal*
Observed yield	Municipality	1 y	-
Modelled yield	1/112°	1 y	Spatial

3.4.1 Dimensions management

To facilitate a comparison of model (deterministic, OL, or DA) output to the satellite or survey reference data, the spatial and temporal dimensions and the aggregation criteria are discussed. The aforementioned variables cannot be directly compared to the model output due to different resolutions and must be aggregated, as shown in Table 3.4. Furthermore, the model output was compared only during the growing seasons, and where the crop masks identified WW.

Specifically, the modelled CC and B were averaged to 10-day values to meet the temporal resolution of the satellite data. The 10-day satellite-observed $FCOVER$ and CC were aggregated from 1/336° to the model resolution of 1/112°, using only those fine-scale pixels that were covered with WW. The same procedure was applied to the modelled yield, but the aggregation was performed from the model resolution to the municipality level. At this stage, the metrics can be computed in space or time, or across the entire spatio-temporal dataset.

3.4.2 Metrics

The metrics used to evaluate the performance are the Pearson correlation coefficient (R), the root mean square difference (RMSD), and the bias:

$$\mathrm{R} = \frac{\sum_{j=1}^{N}(M_j - \overline{M})(O_j - \overline{O})}{\sqrt{\sum_{j=1}^{N}(M_j - \overline{M})^2}\sqrt{\sum_{j=1}^{N}(O_j - \overline{O})^2}}, \tag{3.7}$$

$$\mathrm{RMSD} = \sqrt{\frac{1}{N}\sum_{j=1}^{N}(M_j - O_j)^2}, \tag{3.8}$$

$$\mathrm{bias} = \frac{1}{N}\sum_{j=1}^{N}(M_j - O_j), \tag{3.9}$$

with N the sample size, M the analysed variable, and O the observations or reference values. The sample average is given by $\overline{M} = \frac{1}{N}\sum_{j=1}^{N} M_j$, and the same applies to $\overline{O}$. The sample size and the elements included in the equations vary according to the dimensions (time, space, time and space) along which the metrics are computed.

The above metrics are computed on the 'raw' data and on temporal anomalies. The latter are important to study interannual and short-term variation in CC and ΔB_+, and interannual variation in yield, because DA aims to improve this variation by compensating for the uncertainty of coarse-scale meteorological data within the model. The anomalies were computed by removing the average seasonal pattern for CC and ΔB_+ over the 2017-2023 time span, whereas yield anomalies were obtained by simply removing the yearly average after aggregating at the municipality level.

The temporal metrics of the 'raw' data are presented in maps, with a temporal metric for each location (grid cell or municipality). Spatio-temporal metrics result in a single metric and are presented in summary scatter plots and tables. For the anomalies, only spatio-temporal metrics are presented. However, for future work, a computation of time-series anomaly R, RMSD, and bias for each pixel (or municipality for yield) and a spatial aggregation of the results is recommended to focus on the temporal (interannual, short-term) performance. Likewise, a future study could consider similar metrics for spatial anomalies.

Chapter 4

Results

4.1 Deterministic simulation

The quality of the LIS-AquaCrop simulations is evaluated in terms of canopy cover CC, dry above-ground biomass production ΔB_+, and dry yield. The spatio-temporal average dry yield obtained during the deterministic run is 5.60 t ha^{-1}, which is 0.02 t ha^{-1} higher than the observed values. Figure 4.1 shows the time average yield for each municipality obtained from the surveys and simulated by AquaCrop. The modelled yield values demonstrate the rigidity of the model: the spatial variability is very limited. Conversely, the observations cover a range from 1 to 8.5 t ha^{-1}. The variability is also very limited in time as shown in Figure 4.2. This figure compares the yield per year per municipality for the observations and deterministic simulation.

The temporal performance metrics for CC, ΔB_+ and yield in Figure 4.3 show a consistently high correlation with reference data for both CC and biomass production. In contrast, yield exhibits a considerable spread, even towards negative R values. The RMSD is high for all three variables, accounting for around one-quarter of the reference values. However, the bias indicates that the mean standard difference reflects different behaviours: *i)* CC is usually overestimated by the model, *ii)* ΔB_+, instead, is underestimated, and *iii)* Y is, by design, quite unbiased, attributing the high RMSD to a symmetrical spread.

CC trajectory in the deterministic run was consistently close to the upper boundary described by CC_{pot}, anticipating low spread for Mode 1, and an ensemble mean drift for both Modes 2 and 3.

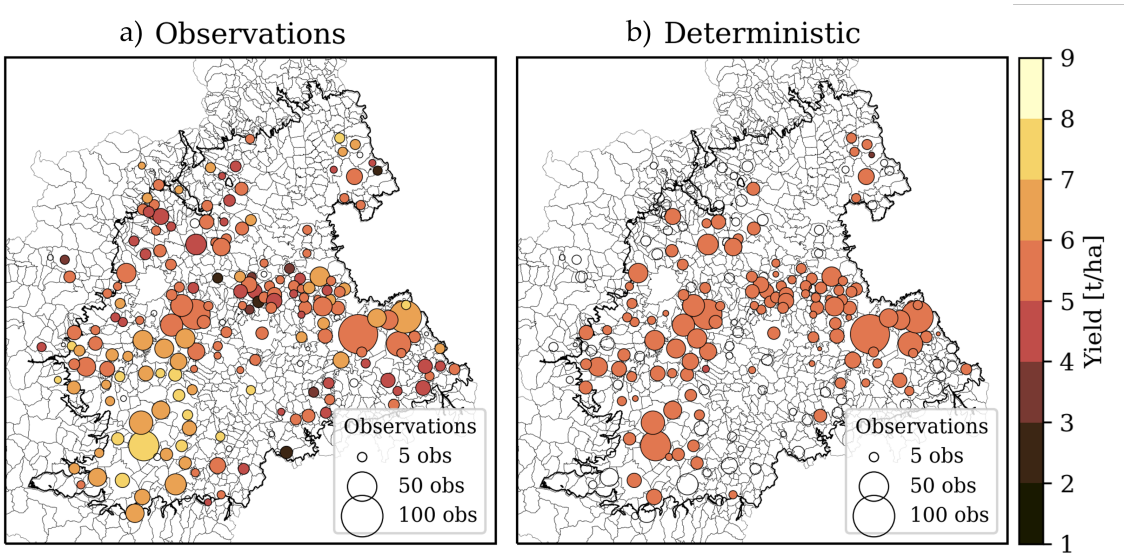

Figure 4.1: Time average of winter wheat yield per municipality. The average values of the RICA samples are shown on the left (a), while the output of the deterministic run is displayed on the right (b). There are white dots in b (no data) due to the absence of wheat fields in the crop masks for those municipalities.

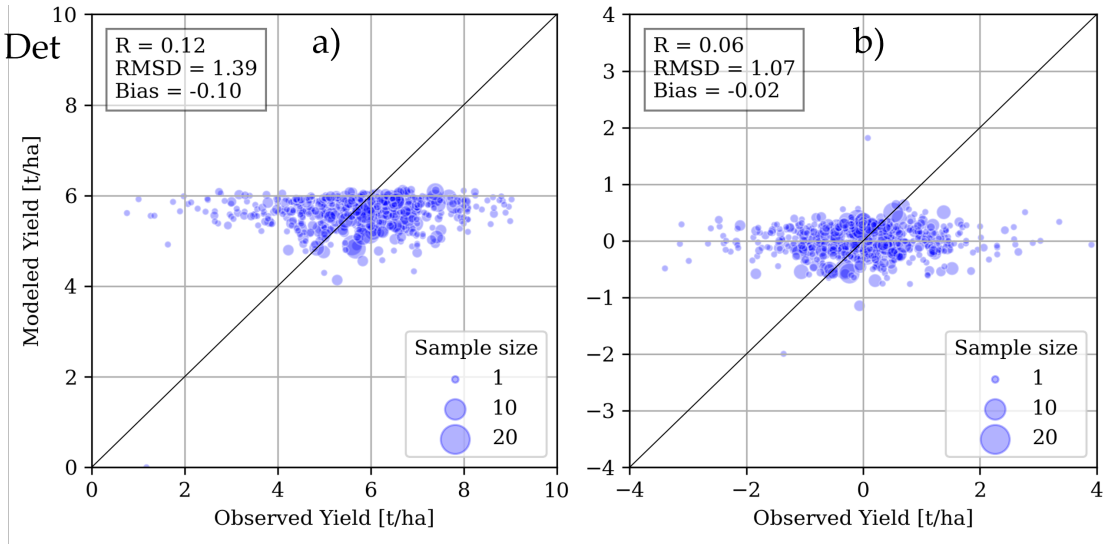

Figure 4.2: Scatter plots comparing the annual winter wheat dry yield with samples from the RICA-CREA survey. Absolute values (a) and seasonal anomalies (b) for the deterministic baseline. Sample sizes are used for visual purposes only.

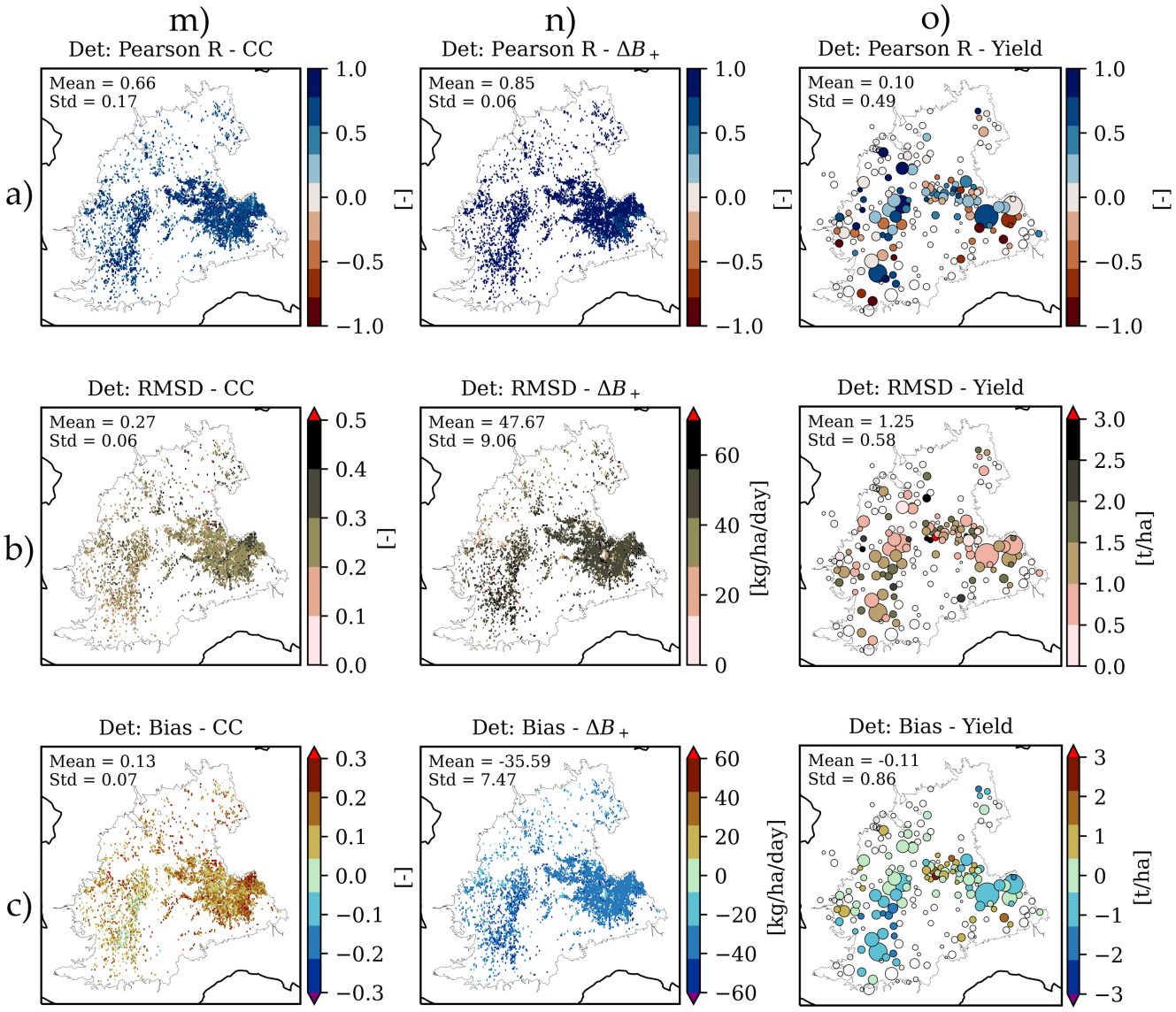

Figure 4.3: Spatial maps of temporal performance metrics for the deterministic run. The rows show the Pearson R (a), the root mean square difference (RMSD, b), and the bias (c). The columns describe the different variables: canopy cover (m), biomass production (n), and yield formation (o). The circles in the latter represent each municipality, and their size indicates the importance of wheat production, based on the number of observations. The metrics are computed only for locations with at least three years of observations. The mean and standard deviation in each map are computed using equal weights for all the locations.

4.2 Ensemble OL simulations

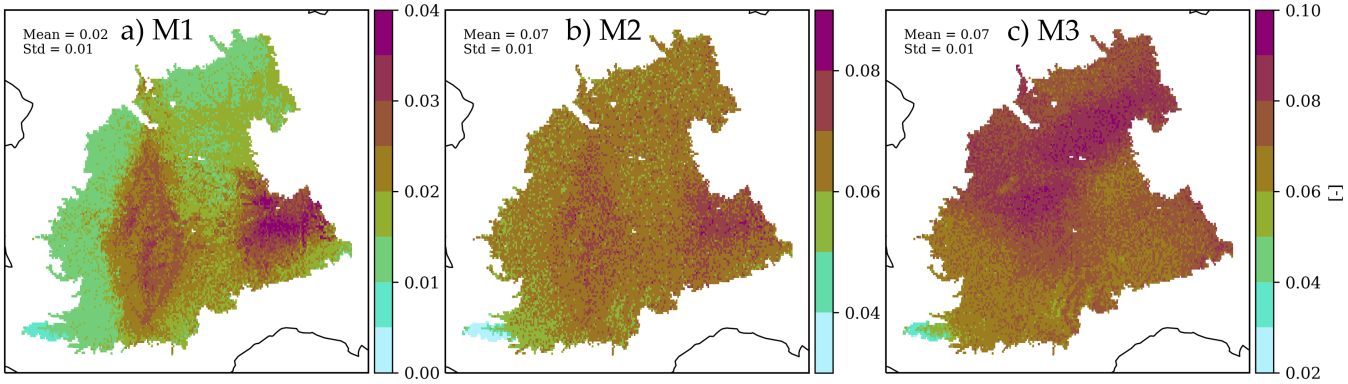

Figure 4.4: Average in-season spread of OL canopy cover generated by the perturbation of state variables and meteorological forcings. For Mode 3 (c) only, it is also generated by the variability of crop parameters. Only Mode 1 (a) had the perturbation bias correction turned on.

As mentioned in Chapter 3, canopy cover uncertainty is a fundamental aspect of DA in this analysis. The three different methods of applying the perturbation yielded three different spatial patterns of CC uncertainty, usually referred to as CC ensemble standard deviation or 'spread'. As expected, the first experiment (Mode 1) exhibits an overall lower spread of CC values among the ensemble members. The average in-season spread maps in Figure 4.4 show that Mode 3 has higher localised values ($\sim$0.10) than Mode 2 ($\sim$0.09), and Mode 1 ($\sim$0.04). However, the spatial average of Mode 3 is comparable to the settings where CC_x and planting dates do not vary in Mode 2.

An example of the ensemble members for Mode 2 is shown in Figure 4.5. Fall 2017 was the second driest autumn since 1958 in the Piedmont Region (Piemonte, 2018); in fact, the perturbation of precipitation patterns caused early water deficiency and delayed the emergence of some ensemble members.

Mode 2 has been used as a visual reference whenever it was not possible to show all experiments. Time series of ensemble means and spreads will be discussed below, along with the time series performance of the 3 modes.

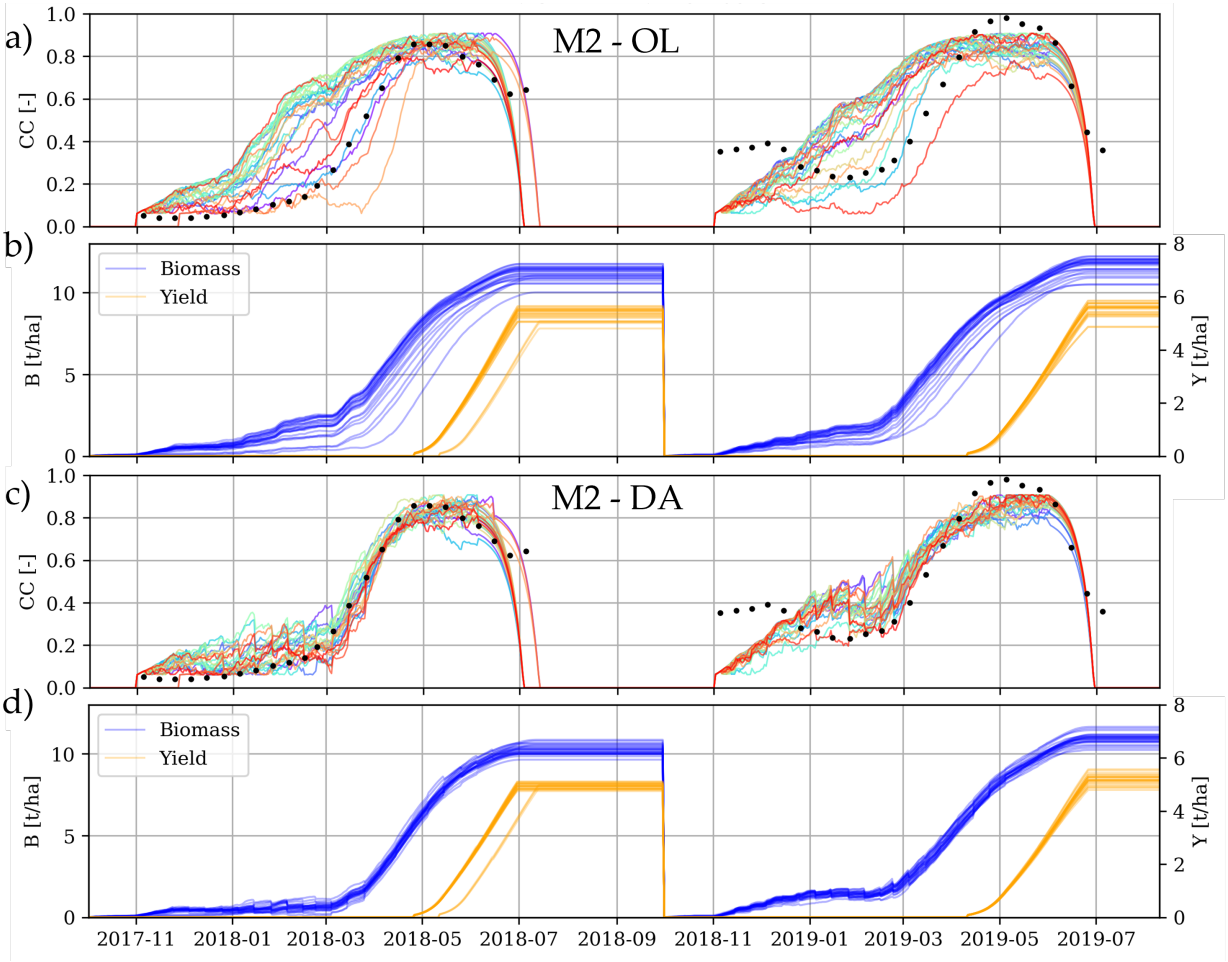

Figure 4.5: Time series of the variables CC, B, and Y for years 2018 and 2019 of a model pixel in the Municipality of Alessandria (44°56'57.3"N : 8°33'13.5"E) for Mode 2. Plots a) and b) are related to the OL, while c) and d) show the effects of the DA updates. During Fall 2017, the delayed emergence of some members was caused by water stress.

4.3 OL and DA performance

Time series of OL and DA simulations are first illustrated at the point scale to verify their correctness. Thus, CC, ΔB_+, yield, and soil moisture for a specific location (Municipality of Alessandria, at 44°56'57.3"N : 8°33'13.5"E) in Mode 2 are shown in Figure 4.6. The same time series for the other two modes can be consulted in Appendix A (Figures A.4 and A.5). To avoid biases during site selection, the location was chosen according to its importance for regional wheat production. Nevertheless, the time series have been included for informational purposes only.

For CC, the ensemble mean of the Mode 2 OL is typically lower than the deterministic run and CC growth is slightly delayed, because of perturbation bias (compare with Figure A.4 which includes perturbation bias correction). For Mode 3 (Figure A.5), the growth is further delayed due to varying growing dates, even though the sowing window was centred around the sowing date of Modes 1 and 2. CC in the deterministic and Mode 2 OL thus often experience growth too early compared to the satellite FCOVER observations. DA updates tend to delay this phase when the ensemble spread is sufficiently large. In some years, the maximum FCOVER values exceeded CC_x, and the DA is bounded by the potential line to preserve physical consistency. These effects are also visible in biomass production. Additionally, low DMP values early in the season do not always correspond to low vegetation cover observations.

Yield estimates are consistently lower when updates are applied, regardless of growth patterns throughout the years. In 2021 and 2022, a lower maximum CC was associated with higher soil moisture levels during the drier months.

4.3.1 Canopy cover

To examine the OL and DA performance at a regional scale, the results are divided into the different analysed variables. Figure 4.7 confirms that all the metrics improved in the scatter plots between CC and FCOVER, by comparing the Mode 2 DA and OL runs. The scatter plots for Mode 1 and 3 are shown in Figure A.6. Table 4.1 summarizes the metrics. It is interesting to note the relative improvements between experiments: Modes 2 and 3 achieved the highest correlations (0.87 and 0.86, respectively), but this was by improving OL runs that had already yielded better R values themselves.

CC and FCOVER anomalies were computed by removing their respective climatology, and then compared with each other (Figure 4.7, and Figure A.6). As summarised in Table 4.1, the OL metrics are similar for all experiments, and, once again, the improvements in the DA runs for modes M2 and M3 are higher than for M1, which is a consequence of the higher ensemble spread in M2 and M3 and thus stronger updates towards the FCOVER observations.

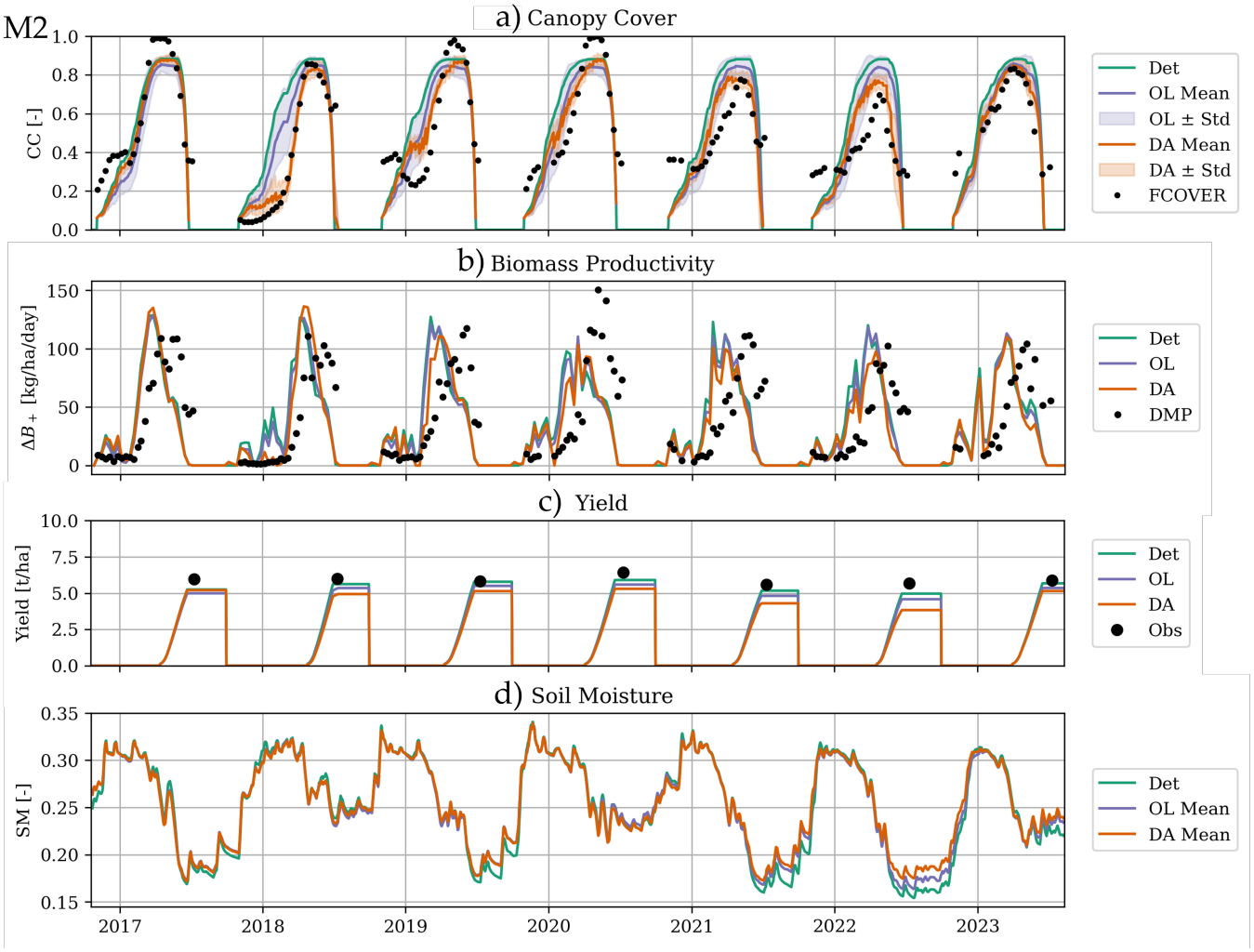

Figure 4.6: Time series showing the state variables (a - canopy cover, b - 10-day aggregated biomass production, derived from B, and d - 7-day smoothed soil moisture) and yield formation (c) in a model pixel in the Municipality of Alessandria (44°56'57.3"N : 8°33'13.5"E). The deterministic run is indicated in green, the open-loop (OL) ensemble mean for Mode 2 is illustrated in violet, and the data assimilation (DA) ensemble mean for Mode 2 is shown in orange. In graph a), the spread represented covers two standard deviations (±1). The RICA-CREA values in graph c) are provided for reference only. For the analysis, the model pixels were aggregated at the municipality level prior to any comparisons being made.

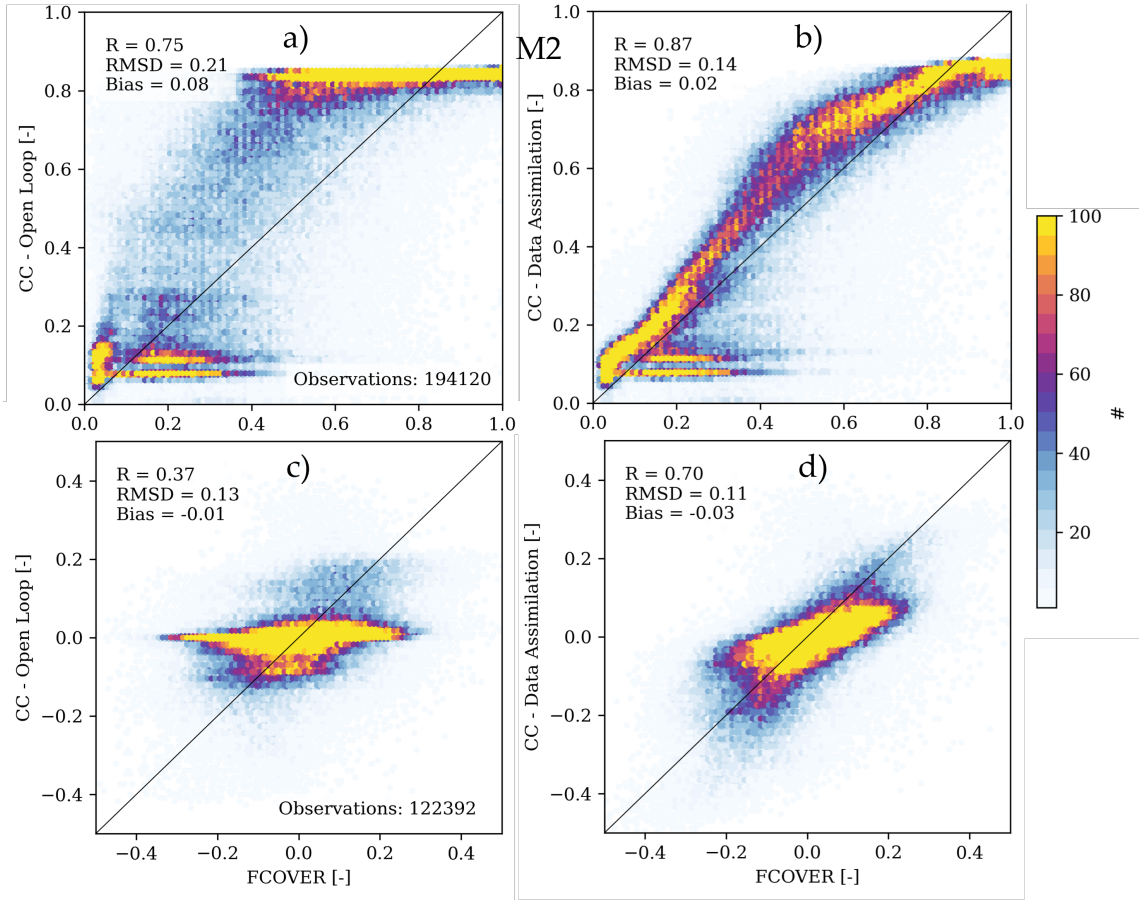

Figure 4.7: Scatter plots comparing the 10-day aggregated canopy cover with the FCOVER observations. A comparison of the absolute values from the OL (a) and DA (b) runs, and the seasonal anomalies (c and d), is shown for Mode 2. The lower number of dots in the bottom two plots is due to locations with fewer than three years of data being excluded.

4.3.2 Biomass production

The same analysis was performed on biomass production, after converting B to ΔB_+ to make it comparable with the DMP satellite product. The scatter plots in Figures 4.8 and A.7 show worse metrics than CC. The absolute values rarely fall near the 1:1 line, and the OL runs are characterised by a clustering behaviour. These patterns are consistent across all modes, with the first mode exacerbating the division between clusters in both OL and DA. Anomalies produce smoother scatter plots in all scenarios, with a high concentration of values around zero and a smaller number of values spreading over a range of around ± 50 kg ha^{-1} day^{-1} for both the model and the observations.

Table 4.1: Summary of the effects of perturbation (OL) and the performance of DA, as measured by the difference between the in-season 10-day aggregated canopy cover (CC_{10d}) and the assimilated satellite product (FCOVER). Seasonal anomalies are computed by removing the average seasonal growth pattern. The best values for each metric are in bold, and the worst are in italics.

CC_{10d} vs FCOVER		**Det**	**M1**		**M2**		**M3**	
			OL	DA	OL	DA	OL	DA
R	[-]	*0.66*	0.70	0.78	0.75	**0.87**	0.75	0.86
RMSD	[-]	*0.27*	0.25	0.19	0.21	**0.14**	0.20	0.15
Bias	[-]	0.13	*0.14*	0.07	0.08	0.02	0.04	**0.00**
Seasonal anomalies								
R	[-]	*0.30*	0.34	0.58	0.37	**0.70**	0.32	0.64
RMSD	[-]	0.14	0.14	0.13	0.13	**0.11**	*0.15*	**0.11**
Bias	[-]	**0.00**	-0.01	*-0.03*	-0.01	*-0.03*	-0.01	-0.02

The metrics summarized in Table 4.2 for the absolute values present similar patterns to those of the aforementioned CC state variable, with Mode 2 showing the best values of all the experiments. DA runs always improve their corresponding OL. Conversely, the picture is less clear when looking at the anomalies. R continues to improve regardless of the experimental setup, while the RMSD and bias are less consistent. Mode 1 DA shows the worst values for both of these metrics; likewise, the procedure has a negative impact on the other two experiments, too.

4.3.3 Yield formation

The performance of DA is finally estimated by analysing the dry yield. The metrics in Table 4.3 are computed across the entire survey dataset within the domain, in terms of both space and time. They demonstrate that the deterministic and OL simulations have a low average bias, but a low R, revealing a poor representation of the spatio-temporal variability in yield. DA does not enhance the accuracy of the model's yield estimates in terms of bias and RMSD, but it slightly improves the R. Figure 4.9 displays how the point cloud widens thanks to DA, but the R remains close to the OL value. This effect is present in all modes.

Figure 4.10 shows that intra-annual spatial variability is never matched. However, M3 DA exhibits a greater spatial variation than other experiments. Yield after

Table 4.2: A summary of the effects of perturbation (OL) and the performance of DA is provided, similar to Table 4.1, by examining the metrics between the in-season 10-day aggregated daily biomass production ($\Delta B_{+,10d}$) and the corresponding satellite product (DMP). Seasonal anomalies are computed by removing the average seasonal growth pattern. The best values for each metric are highlighted in bold, and the worst in italics.

$\Delta B_{+,10d}$ vs DMP		**Det**	**M1**		**M2**		**M3**	
			OL	DA	OL	DA	OL	DA
R	[-]	0.32	*0.31*	0.38	0.38	**0.49**	0.48	**0.49**
RMSD	[kg ha^{-1} d^{-1}]	43.59	*43.72*	40.92	41.98	**38.18**	39.65	38.83
Bias	[kg ha^{-1} d^{-1}]	*8.50*	8.33	3.58	6.18	**1.47**	5.80	2.73
Seasonal anomalies								
R	[-]	*0.25*	*0.25*	0.35	0.27	0.39	0.27	**0.40**
RMSD	[kg ha^{-1} d^{-1}]	20.17	20.15	*21.33*	**19.47**	20.40	19.72	20.73
Bias	[kg ha^{-1} d^{-1}]	-0.58	-0.55	*-2.19*	-0.49	-2.13	**-0.26**	-1.24

Table 4.3: A summary of the effects of perturbation (OL) and the performance of DA is provided by considering the metrics between annual dry yield formation (Y) and the corresponding observed values (RICA-CREA). Seasonal anomalies are computed by subtracting the interannual average. The best values for each metric are in bold, and the worst are in italics.

Y vs RICA-CREA		**Det**	**M1**		**M2**		**M3**	
			OL	DA	OL	DA	OL	DA
R	[-]	0.12	0.11	**0.14**	0.11	**0.14**	*0.09*	0.13
RMSD	[kg ha^{-1}]	**1.39**	1.40	1.52	1.44	1.65	1.49	*1.67*
Bias	[kg ha^{-1}]	**-0.10**	-0.14	-0.58	-0.39	*-0.84*	-0.46	-0.80
Seasonal anomalies								
R	[-]	0.06	0.05	**0.08**	0.06	0.02	0.02	*0.01*
RMSD	[kg ha^{-1}]	**1.07**	**1.07**	1.09	**1.07**	1.14	1.10	*1.19*
Bias	[kg ha^{-1}]	**-0.02**	**-0.02**	*-0.03*	**-0.02**	*-0.03*	**-0.02**	*-0.03*

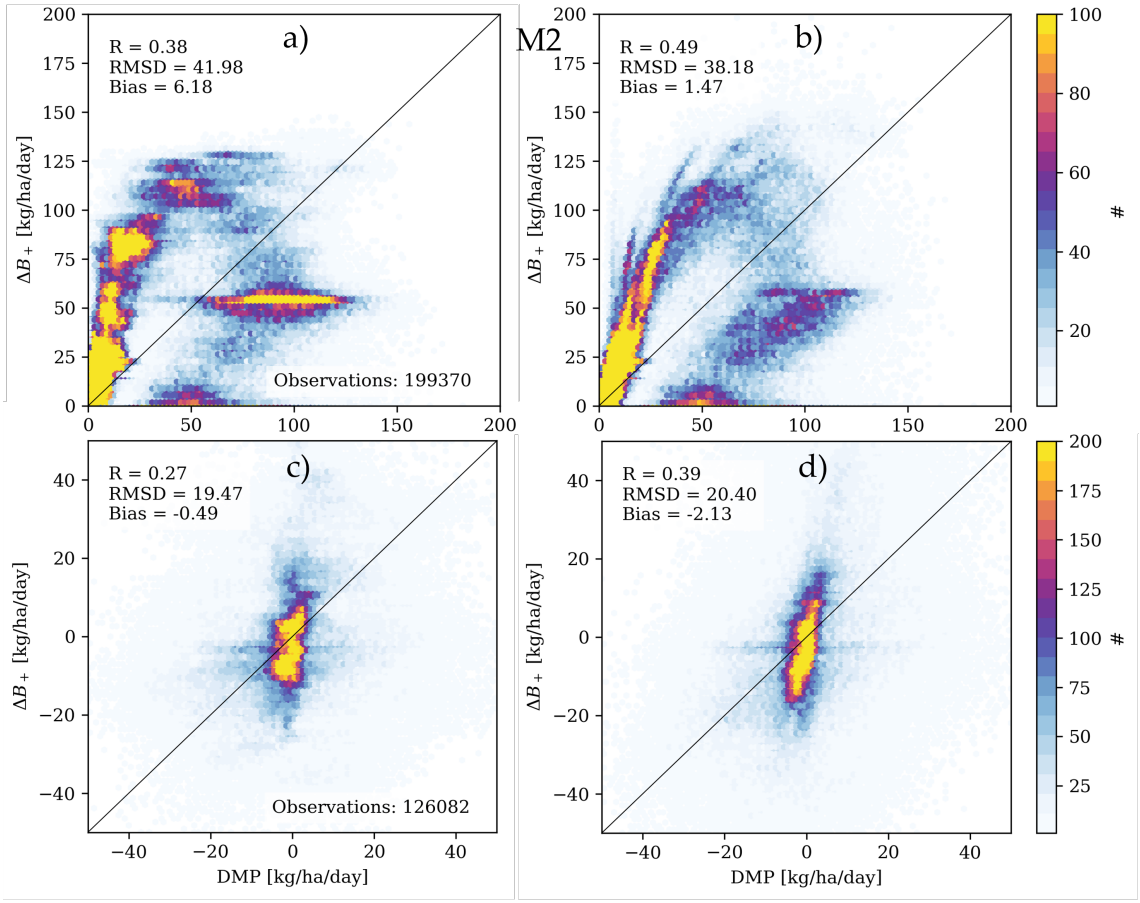

Figure 4.8: Scatter plots showing the relationship between the 10-day aggregated dry above-ground biomass and the dry matter productivity observations. A comparison of the absolute values from the OL (a) and DA (b) runs, and the seasonal anomalies (c and d), is shown for Mode 2. The lower number of dots in the bottom two plots is due to locations with fewer than three years of available data being excluded.

DA is almost always lower than the yield from the deterministic run or the OLs, regardless of whether the observations are over- or under-estimated. Rarely do any of the runs obtain yield values higher than 6 t ha^{-1}; M3 seems to exceed the threshold slightly more often.

4.4 Overall temporal performance

Whereas the performance assessment above focused on evaluating spatio-temporal variability, this section summarizes the temporal performance metrics. Figure 4.11 shows the performance metrics computed over time series at each location and then

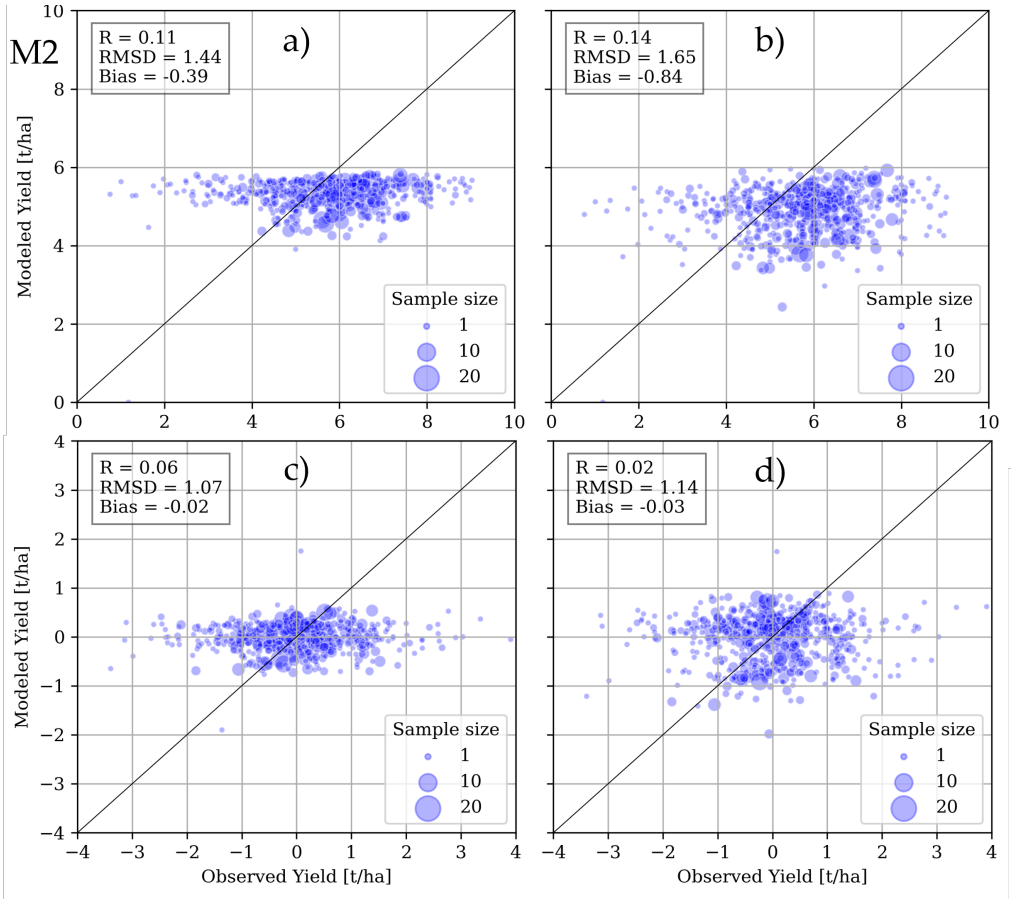

Figure 4.9: Scatter plots depicting the relationship between annual winter wheat dry yield and samples from the RICA-CREA survey. A comparison of the absolute values from the OL (a) and DA (b) runs, and the seasonal anomalies (c and d), is shown for Mode 2. Sample sizes are used for visual purposes only.

aggregated spatially to provide a final performance evaluation. Clear DA improvements are evident in the CC performance, and similar patterns emerge for ΔB_{+}. In contrast, dry yield performance after DA has remained similar to that of the deterministic run at most, but is characterised by a general degradation in all metrics and modes. The next chapter provides an interpretation of these results and makes suggestions for future research.

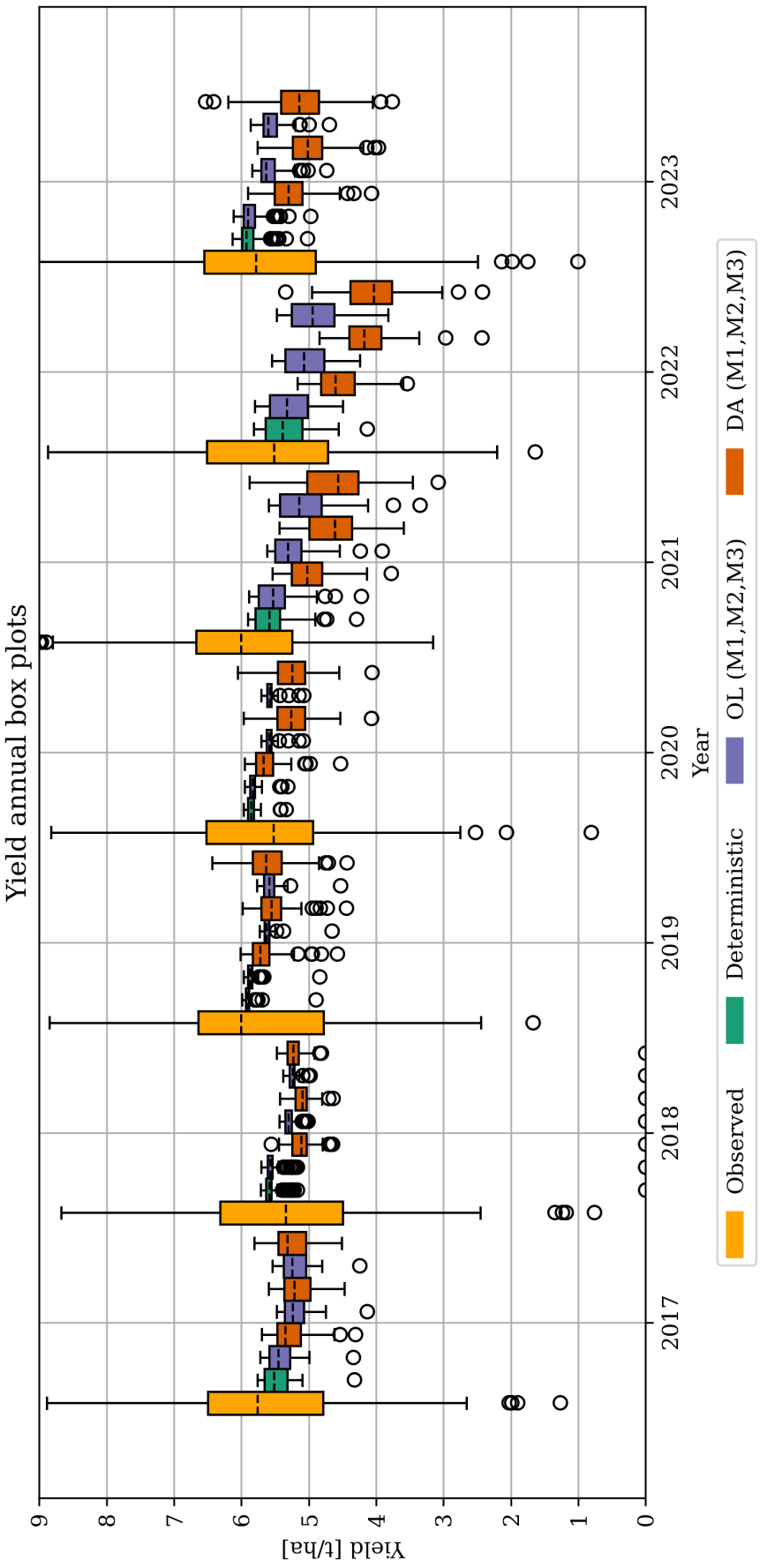

Figure 4.10: Annual box plots of the dry yield of winter wheat in the Piedmont Region. Every year is presented from left to right: the observed values are plotted in yellow, the deterministic runs are depicted in green, and the three pairs of open-loop and data assimilation are plotted in violet and orange for the three different experiments (OL and DA, respectively). The dashed line indicates the median of the sample and the coloured box includes 50% of all values, from the first to the third quartiles. The whiskers span from the largest to the smallest values that fall within 1.5 times the interquartile range, and all outliers are represented by empty circles.

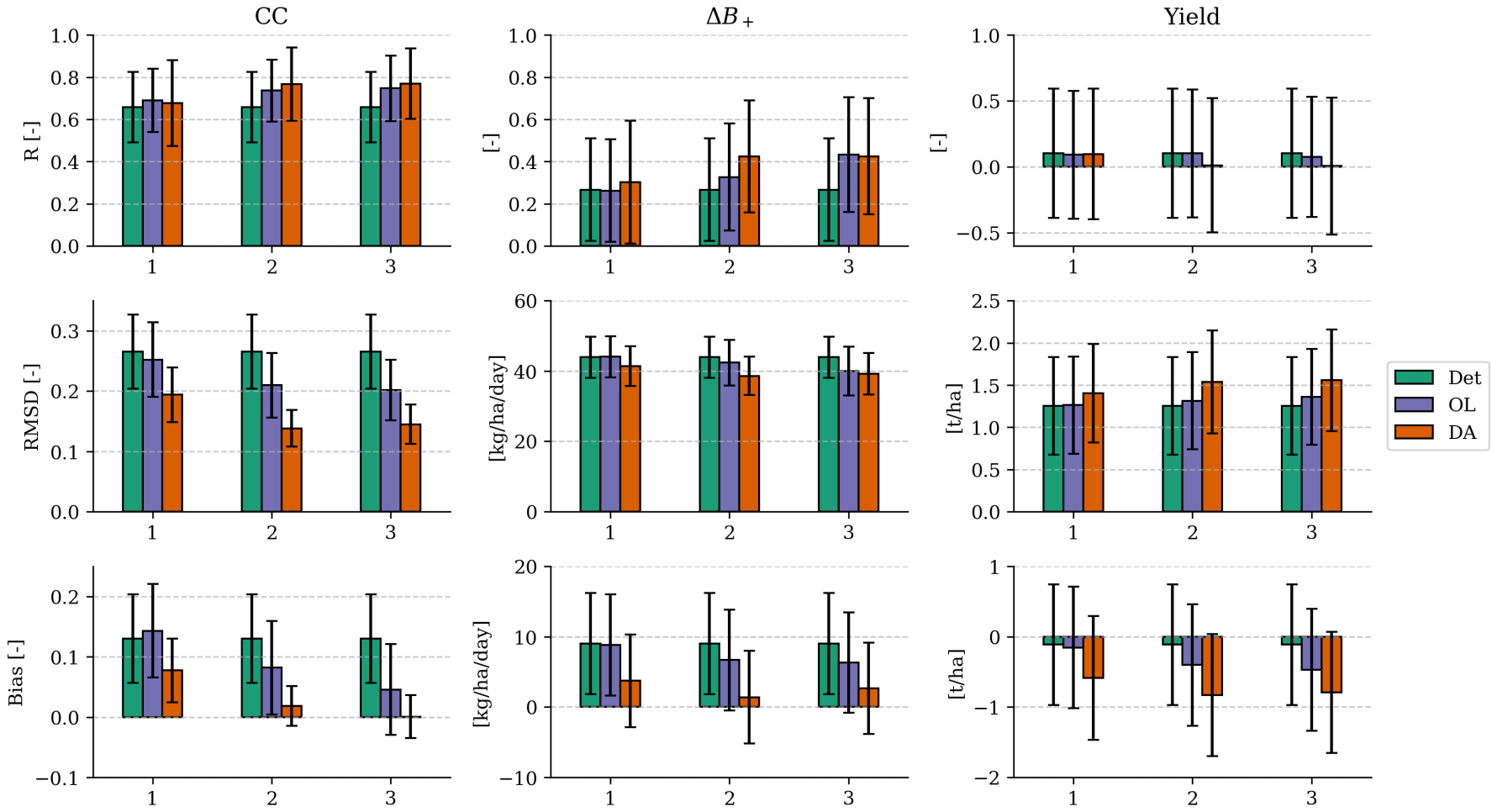

Figure 4.11: Summary of the performance of the open-loop runs (in violet) and the data assimilation (in orange) compared to the deterministic baseline (in green) for modes 1, 2, and 3. The three crop-related state variables are displayed: canopy cover on the left, dry above-ground biomass production in the middle, and annual dry yield on the right. The metrics are first computed temporally and then aggregated spatially over the valid domain. The error bars represent the standard deviation of the spatial variability.

Chapter 5

Discussion

5.1 Model performance

Canopy cover

The comparison of simulated CC against satellite-observed FCOVER in the scatter plots in Figure 4.7 shows model clusters around 0.80 and some clusters around low values. The highest plateau is due to CC_x. The lowest one is at CC_o. Whereas the flat cluster around 0.10 is likely due to the nature of AquaCrop, where routines are typically characterised by hard boundaries. Additionally, canopy growth appears to be generally faster and/or earlier in the season compared to FCOVER.

Better calibration could be performed in an attempt to match the satellite product. On the other side, overestimation of CC allowed DA to update towards lower values, and updates in the opposite direction (i.e., for values below the 1:1 line in the CC scatter plots) were a rare occurrence due to the presence of the upper boundary CC_{pot}.

Biomass

The scatter plots that compare modelled dry above-ground biomass production to the DMP Copernicus product show a particular circular pattern. As the model and the satellite data are out of phase, the plants in the model produce earlier and enter the senescence stage before their satellite equivalent.

DA can smooth out the clusters, but the circular pattern remains. It is possible that this phenomenon may be reduced if CGC is lowered during the calibration process. Likewise, other parameters may need to be updated alongside the state variables to get better B estimates. For example, the Ks factor in eq. 1.6 can be

corrected to raise the cold stress threshold. This suggests that DA can eliminate some model inaccuracies, but a better crop calibration is still required for the core state variable, CC.

Furthermore, the observed and simulated variables do not represent the same quantities exactly. As mentioned in Subsection 2.3.3, the DMP should always be higher than the model's dry above-ground biomass production since the former also considers biomass stored in the root zone, and does not include both water stress and fertility effects (Swinnen *et al.*, 2023). Therefore, the ΔB_+ values of the earliest part of the seasons (i.e., the arched clusters above the 1:1 line) differ drastically from the expected outcome.

Yield

The regional implementation of AquaCrop within NASA's LIS provides insight into how crop growth responds to different meteorological conditions and soil textures. Even though the Piedmont study domain is composed of a heterogeneous soil texture map and spatially varying forcing data, the deterministic run on a regional scale is characterised by a narrow spread in yield formation. This may be due to the coarse resolution of the ERA5 reanalysis ($0.25^o \times 0.25^o$) and the aggregation of the model output at the municipality level, which partially evens out the spatial variability. Nevertheless, it is evident that LIS-AquaCrop exhibits a degree of spatial rigidity in itself. Likewise, as can be seen in Figures 4.9 and A.8, the LIS-AquaCrop Y variations in space and time are narrow. One probable reason for this is the calibration of the fertility stress. The GYGA reference values describe the annual average yield gap from the potential rainfed scenario in the region, and are consistent with the fertility stress definition used in this analysis. However, applying an average fertility parameter across the entire domain can result in a hardly-bounded range of values, which can limit the model's performance if metrics are computed using absolute yield. For example, a municipality with higher overall fertility than average may experience a consistent bias between surveyed production and model output.

The lack of spatial variation due to the assumption of homogenous fertility can be eliminated by only focusing on temporal metrics for raw and anomaly estimates.

After DA, the hope is that the anomaly variation in yield would have improved compared to their corresponding OL values, but this is not the case in Table 4.3.

5.2 Ensemble design

There are noticeable differences between the three pairs of OL and DA experiments (modes). As expected, Mode 1 preserves an unbiased CC ensemble mean, whereas Modes 2 and 3 typically exhibit lower CC values due to the proximity of the deterministic run to the potential curve. In fact, while the perturbation is normally distributed around the mean, the presence of the upper boundary increases the skewness of the ensemble distribution, making the mean diverge from the median towards smaller CC. The presence of the potential upper boundary helps to maintain physical consistency, particularly when the observations are contaminated by other land covers, as in the early part of the seasons in the Alessandria example (see Figure 4.6). However, reasonable increments are also limited when the sowing date and growth stages are out of phase with the observations.

This mechanism probably drastically limited the benefits of model updates in Modes 2 and 3. The growth curve was usually dampened, and both flowering and yield formation were occurring too early in the model. This left the maximum FCOVER values in the senescence phase, or even when the modelled plants had dried out and nutrient accumulation had ceased. It is probably the main reason why the updated yield is consistently lower than in the open-loop runs, especially in Mode 3.

Varying the sowing date and maximum canopy cover did not significantly increase the ensemble spread compared to M2; therefore, the additional complexity may not be required. On the other hand, the planting criteria could be modified to ensure a more consistent distribution of members across the sowing window. Using strict rainfall thresholds results in low variability. Furthermore, CC_x variability was partially limited due to the diverging finite differences implementation of the exponential growth equation mentioned in Subsection 3.1.2.

The third experiment probably experienced a flaw in the rainfall criterion. In

order to allow for a varying sowing date based on precipitation, this forcing variable must be read before the simulation begins. However, the perturbation routine for ensemble generation occurs later in the information flow. Introducing the criterion requires an earlier perturbation that may not match the rainfall experienced by the same particle during the simulation.

Given all the evidence, the settings used for Mode 2 strike the right balance between effectiveness, mathematical correctness, and simplicity. Yet the variation of model parameters preserves potential benefits that may have been overlooked due to the limitations presented.

5.3 Model propagation

This thesis aims to exploit the dependence of Y on CC and use FCOVER observations to update Y. The reduction in the added value of the DA results in the step between the computation of B and Y needs to be understood and solved. Similar to the comparison between satellite products in Figure 2.2, the Pearson correlation has been computed between the maximum in-season FCOVER value (FCOVER_x) and the corresponding surveyed yield values in Figure 5.1 to see if there is any relationship between the interannual variation in FCOVER and yield. Although there may be more effective methods of quantifying their alignment, and even if it is known that the soil water balance interferes with the relationship between Y and CC, it is nevertheless beneficial to recognise the absence of a direct connection between observed FCOVER and observed yield.

Furthermore, canopy cover and soil moisture are coupled. SM is affected by updating the CC using satellite observations. In the model, a negative CC increment results in a lower B, and therefore less water is taken up from the root zone. This increases its availability after the state variables are updated, resulting in growth stimulation. Conversely, an observed decrease in CC may be caused by earlier water stress, i.e., lower SM. These outcomes are completely opposite to each other and impossible to predict with the current setup.

In addition to the ambiguous interaction between vegetation cover and soil mois-

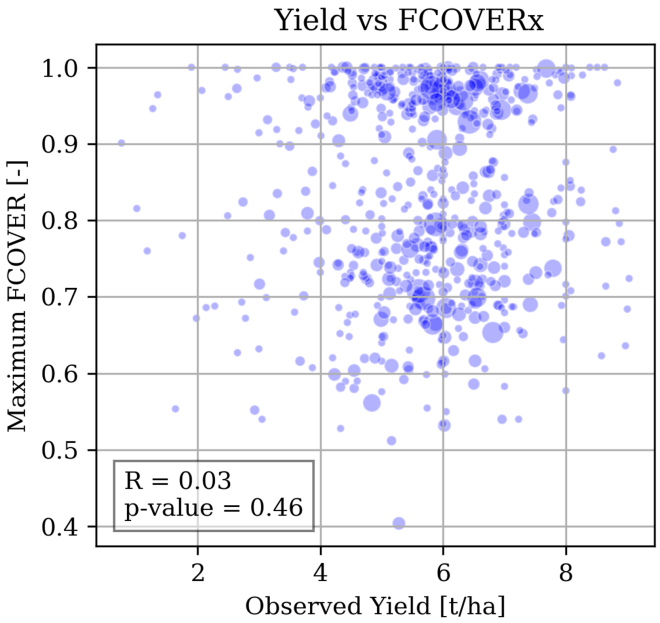

Figure 5.1: Relationship between the observed maximum in-season value of FCOVER and the surveyed yield from RICA-CREA. The correlation is not significant ($p > 0.05$).

ture, the timing of water stress affects the allocation of resources to the plant's productive elements, e.g., the kernel in the case of wheat. For example, reducing biomass before flowering or during yield formation can result in fewer nutrients being taken up from the soil and allocated to the green material, leaving a higher concentration for the kernel and subsequently increasing f_{HI} in Equation 1.8. Meanwhile, water stress during flowering is detrimental to pollination efficiency, with obvious effects on production that season. Therefore, precise quantification and timing of water availability may be essential for yield estimation.

5.4 Validation

The model output has been analysed in many different ways in an attempt to capture any useful information, i.e. to see if the DA could improve the temporal or spatial (anomaly) variation of the various variables. For example, the overall metrics shown in Figure 4.11 estimate the accuracy with which DA captures the temporal variability for each municipality (yield) or grid cell (CC, ΔB_+) with WW. The summary metric

is displayed as the spatial mean and standard deviation over the entire domain. By contrast, an evaluation of both the temporal and spatial variability is performed in the Tables 4.1, 4.2 and 4.3. The corresponding Figure A.9 depicts the results for the full dataset and therefore the spatio-temporal performance. Furthermore, the metrics were also computed by taking a spatial average and calculating the temporal dynamics in Figure A.10. Similarly, Figure A.11 shows the ability of DA to capture variability across the domain after taking the average in every model pixel (for CC and ΔB_+) and each municipality (for Y) over all years.

Opposite results were obtained for the Pearson correlation when examining the spatial and temporal behaviours, suggesting that data assimilation can adapt the model to some extent to correct for spatial patterns, on average. Nevertheless, DA requires general improvements to enhance its performance regardless of the analysed dimensions.

5.5 Future improvements

Advancements can be made from both a modelling and DA perspective. Within the regional LIS application of AquaCrop, researchers will continue to activate features that are currently only available in the field-scale AquaCrop version, or to adapt them to vary spatially. One such feature is fertility stress, which may help to capture spatial patterns in crop production if it becomes non-uniform. Spatially varying water table depth will also be introduced to enhance model performance in specific regions. Following up on the challenges detailed in Section 3.1.2, future versions of the core model will present a more accurate incremental canopy cover growth curve, enabling the crop to reach slightly higher maximum CC values, and removing a possible bias in relation to the observations. Crop calibration is a time-consuming process that limits the large-scale application of crop models. DA for state updating is a promising approach that can reduce uncertainty in general and have a positive impact on crop growth simulations, but the results suggest that parameter updating could likely further improve the results.

DA within crop modelling systems is still in its infancy (de Roos *et al.*, 2024),

and the DA procedure of this study can benefit from further development. As this study shows, the potential CC upper boundary creates a trade-off between physical consistency and model uncertainty that requires further exploration. Dependency on sowing timing not only affects the amplitude of vegetative growth but also constrains physically sound out-of-phase curves. The ability to decouple these mechanisms could greatly enhance the efficacy of data assimilation. This can be achieved through the implementation of varying growth stages or by perturbing temperatures in the forcing data.

While all the aforementioned options will likely improve DA, further progress is needed to improve regional crop yield estimates with a view towards a stable mid-range yield forecast. The assimilation of satellite-borne soil moisture observations alongside vegetation cover can resolve the ambiguity of the effects of CC updates on the soil water balance, in line with earlier studies that showed the benefit of joint DA for land surface models (Heyvaert *et al.*, 2024).

Chapter 6

Conclusions

The climate crisis is already leading to extreme meteorological events, which might pose risks (IPCC, 2022) to food prediction in the future. Predictive tools are essential for farmers to adapt their techniques both in the long term and during the growing season. The joint benefits of mid-range weather forecast models and crop models enriched with satellite data assimilation are potentially strategic and warrant further study and testing.

This thesis contributes to the advancement of high-resolution crop forecasting by assessing which DA techniques are the most suitable and effective in reducing model uncertainties in crop yield estimates. Experiments were performed using AquaCrop within NASA's Land Information System, forced with reanalysis meteorological forcing data to estimate winter wheat production in Italy's Piedmont Region between 2017 and 2023. The Copernicus satellite product "Fraction of vegetation COVER" (FCOVER) has been assimilated to update the model vegetation states, and the propagation of the updates has been investigated to understand its impact on yield formation. Evaluation was performed by computing a series of statistical metrics, comparing the model-only and DA biomass estimates with the Dry Matter Productivity (DMP) Copernicus product, and the yield estimates with observed yield values from the RICA-CREA survey.

After adapting the AquaCrop code and its surrounding LISF routines, the DA successfully updated canopy cover and biomass estimates in the model. However, its propagation within the model was short-lived, resulting in small and inconsistent updates to the dry yield. Future research could consider a series of adjustments to improve DA efficacy by addressing three main elements. Firstly, LIS-AquaCrop should be enhanced by implementing additional spatially varying parameters such as

soil fertility stress, and by correcting some model inaccuracies. Secondly, the effects of the bounded nature of canopy cover on the ability to apply DA increments must be thoroughly examined. This would involve attributing the limit to unphysical values only and finding solutions to the temporal stage mismatch. One possible solution would be to introduce varying growing stages or to perturb input temperatures, and to update crop parameters along with state variables during the DA. Finally, implementing near-surface soil moisture satellite DA will be essential to complement the techniques described in this thesis.

This thesis has advanced our understanding of the limits and potential of regional crop DA. These promising advancements can now be used as a basis for future work towards a safer, more reliable, and more predictable future for human nutrition.

Bibliography

Allen, R. G., Pereira, L. S., Raes, D., and Smith, M. (1998). Crop evapotranspiration: guidelines for computing crop water requirements. FAO. ISBN 9251042195.

Arsenault, K. R., Kumar, S. V., Geiger, J. V., Wang, S., Kemp, E., Mocko, D. M., Beaudoing, H. K., Getirana, A., Navari, M., Li, B., Jacob, J., Wegiel, J., and Peters-Lidard, C. D. (2018). The Land surface Data Toolkit (LDT v7.2) - A data fusion environment for land data assimilation systems. *Geoscientific Model Development*, vol. 11:pp. 3605–3621. ISSN 19919603. doi:10.5194/gmd-11-3605-2018.

Baird Smith, R. (1852). Italian Irrigation: a Report on the Agricultural Canals of Piedmont and Lombardy, addressed to the honourable the court of Directors of the East India Company., vol. 1. William Blackwood and Sons.

Balkovič, J., van der Velde, M., Schmid, E., Skalský, R., Khabarov, N., Obersteiner, M., Stürmer, B., and Xiong, W. (2013). Pan-European crop modelling with EPIC: Implementation, up-scaling and regional crop yield validation. *Agricultural Systems*, vol. 120:pp. 61–75. ISSN 0308521X. doi:10.1016/j.agsy.2013.05.008.

Baret, F., Weiss, M., Verger, A., and Smets, B. (2016). Implementing Multi-scale Agricultural Indicators Exploiting Sentinels. Algorithm Theorethical Basis Document for LAI, FAPAR and FCOVER from PROBA-V products at 300m resolution (GEOV3). Issue 1.73. Tech. rep., CGLS. URL `https://land.copernicus.eu/en/technical-library/algorithm-theoretical-basis-document-fraction-of-green-vegetation-cover-333-m-version-1.0/@@download/file`.

Baronetti, A., Acquaotta, F., and Fratianni, S. (2018). Rainfall variability from a dense rain gauge network in north-western Italy. *Climate Research*, vol. 75:pp. 201–213. ISSN 16161572. doi:10.3354/cr01517.

Belward, A. S. and Skøien, J. O. (2015). Who launched what, when and why; trends in global land-cover observation capacity from civilian earth observation satellites. *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 103:pp. 115–128. ISSN 09242716. doi:10.1016/j.isprsjprs.2014.03.009.

Bondeau, A., Smith, P., Zaehle, S., Schaphoff, S., Lucht, W., Cramer, W., Gerten, D., Lotze-Campen, H., Müller, C., Reichstein, M., and Smith, B. (2007). Modelling the role of agriculture for the 20th century global terrestrial carbon balance. *Global Change Biology*, vol. 13:pp. 679–706. ISSN 1354-1013. doi: 10.1111/j.1365-2486.2006.01305.x.

Boogaard, H., Wolf, J., Supit, I., Niemeyer, S., and van Ittersum, M. (2013). A regional implementation of WOFOST for calculating yield gaps of autumn-sown wheat across the European Union. *Field Crops Research*, vol. 143:pp. 130–142. ISSN 03784290. doi:10.1016/j.fcr.2012.11.005.

Bournan, B. A. M., van Keulen, H., van Laar, H. H., and Rabbinge, R. (1996). The 'School of de Wit' Crop Growth Simulation Models: A Pedigree and Historical Overview. *Agricultural Systems*, vol. 52:pp. 171–198.

Busschaert, L., de Roos, S., Thiery, W., Raes, D., and De Lannoy, G. J. M. (2022). Net irrigation requirement under different climate scenarios using AquaCrop over Europe. *Hydrology and Earth System Sciences*, vol. 26:pp. 3731–3752. ISSN 16077938. doi:10.5194/hess-26-3731-2022.

Cavaletto, S. (2025). Il settore agricolo e rurale piemontese. URL `https://www.regione.piemonte.it/web/temi/agricoltura/settore-agricolo-rurale-piemontese`.

CGLS (2017). Fraction of Vegetation Cover 2014-present (raster 300 m), global, 10-daily - version 1. doi:https://doi.org/10.2909/09578c73-4f5d-4d2c-90ff-4e17fb7dbf69. URL `https://land.copernicus.eu/en/products/vegetation/fraction-of-green-vegetation-cover-v1-0-300m`.

CGLS (2018). Dry Matter Productivity 2014-present (raster 300 m), global, 10-daily - version 1. doi:10.2909/67797662-7edc-4a29-b93b-a58af384b137. URL `https://land.copernicus.eu/en/products/vegetation/dry-matter-productivity-v1-0-300m`.

Challinor, A., Wheeler, T., Craufurd, P., Slingo, J., and Grimes, D. (2004). Design and optimisation of a large-area process-based model for annual crops. *Agricultural and Forest Meteorology*, vol. 124:pp. 99–120. ISSN 01681923. doi:10.1016/j.agrformet.2004.01.002.

Chen, Y., Zhang, Z., and Tao, F. (2018). Improving regional winter wheat yield estimation through assimilation of phenology and leaf area index from remote sensing data. *European Journal of Agronomy*, vol. 101:pp. 163–173. ISSN 11610301. doi:10.1016/j.eja.2018.09.006.

CLMS (2025). High Resolution Layer - Croplands. Product User Manual 1.0. Tech. rep., Copernicus Land Monitoring Service.

D'Andrimont, R., Verhegghen, A., Lemoine, G., Kempeneers, P., Meroni, M., and van der Velde, M. (2021). EUCROPMAP 2018.

De Lannoy, G. J. M., Busschaert, L., Bechtold, M., Lanfranco, N., de Roos, S., Heyvaert, Z., Mortelmans, J., Scherrer, S. A., Van den Bossche, M., Kumar, S., Mocko, D. M., Kemp, E., Heng, L., Steduto, P., and Raes, D. (in review). Advancing Crop Modeling and Data Assimilation Using AquaCrop v7.2 in NASA's Land Information System Framework v7.5. *EGUsphere*, vol. 2025:pp. 1–35. doi:10.5194/egusphere-2025-4417. URL `https://egusphere.copernicus.org/preprints/2025/egusphere-2025-4417/`.

De Lannoy, G. J. M., Koster, R. D., Reichle, R. H., Mahanama, S. P. P., and Liu, Q. (2014). An updated treatment of soil texture and associated hydraulic properties in a global land modeling system. *Journal of Advances in Modeling Earth Systems*, vol. 6:pp. 957–979. ISSN 19422466. doi:10.1002/2014MS000330.

de Roos, S., Bechtold, M., Busschaert, L., Lievens, H., and De Lannoy, G. J. M. (2024). Assimilation of Sentinel-1 Backscatter to Update AquaCrop Estimates

of Soil Moisture and Crop Biomass. *Journal of Geophysical Research: Biogeosciences*, vol. 129. ISSN 2169-8953. doi:10.1029/2024JG008231.

de Roos, S., De Lannoy, G. J. M., and Raes, D. (2021). Performance analysis of regional AquaCrop (v6.1) biomass and surface soil moisture simulations using satellite and in situ observations. *Geoscientific Model Development*, vol. 14:pp. 7309–7328. ISSN 19919603. doi:10.5194/gmd-14-7309-2021.

Deryng, D., Sacks, W. J., Barford, C. C., and Ramankutty, N. (2011). Simulating the effects of climate and agricultural management practices on global crop yield. *Global Biogeochemical Cycles*, vol. 25:pp. n/a–n/a. ISSN 08866236. doi:10.1029/2009GB003765.

Di Paola, A., Valentini, R., and Santini, M. (2016). An overview of available crop growth and yield models for studies and assessments in agriculture. doi:10.1002/jsfa.7359.

Dlamini, L., Crespo, O., van Dam, J., and Kooistra, L. (2023). A Global Systematic Review of Improving Crop Model Estimations by Assimilating Remote Sensing Data: Implications for Small-Scale Agricultural Systems. doi:10.3390/rs15164066.

Draper, C. S., Reichle, R. H., De Lannoy, G. J. M., and Liu, Q. (2012). Assimilation of passive and active microwave soil moisture retrievals. *Geophysical Research Letters*, vol. 39. ISSN 0094-8276. doi:10.1029/2011GL050655.

Farr, T. G., Rosen, P. A., Caro, E., Crippen, R., Duren, R., Hensley, S., Kobrick, M., Paller, M., Rodriguez, E., Roth, L., Seal, D., Shaffer, S., Shimada, J., Umland, J., Werner, M., Oskin, M., Burbank, D., and Alsdorf, D. E. (2007). The shuttle radar topography mission. *Reviews of Geophysics*, vol. 45. ISSN 87551209. doi:10.1029/2005RG000183.

GYGA (2021). Global Yield Gap and Water Productivity Atlas. URL `http://www.yieldgap.org/`.

Hendrickx, M. G. A., Vanderborght, J., Janssens, P., Bombeke, S., Matthyssen, E., Waverijn, A., and Diels, J. (2025). Pooled error variance and covariance estimation of sparse in situ soil moisture sensor measurements in agricultural fields in Flanders. *SOIL*, vol. 11:pp. 435–456. ISSN 2199-398X. doi:10.5194/soil-11-435-2025.

Hennicker, R., Janisch, S., Kraus, A., and Ludwig, M. (2016). DANUBIA: A Web-Based Modelling and Decision Support System to Investigate Global Change and the Hydrological Cycle in the Upper Danube Basin, pp. 19–27. Springer International Publishing. doi:10.1007/978-3-319-16751-0_2.

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., Nicolas, J., Peubey, C., Radu, R., Schepers, D., Simmons, A., Soci, C., Abdalla, S., Abellan, X., Balsamo, G., Bechtold, P., Biavati, G., Bidlot, J., Bonavita, M., De Chiara, G., Dahlgren, P., Dee, D., Diamantakis, M., Dragani, R., Flemming, J., Forbes, R., Fuentes, M., Geer, A., Haimberger, L., Healy, S., Hogan, R. J., Hólm, E., Janisková, M., Keeley, S., Laloyaux, P., Lopez, P., Lupu, C., Radnoti, G., de Rosnay, P., Rozum, I., Vamborg, F., Villaume, S., and Thépaut, J. N. (2020). The ERA5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, vol. 146:pp. 1999–2049. ISSN 1477870X. doi:10.1002/qj.3803.

Heyvaert, Z., Scherrer, S., Dorigo, W., Bechtold, M., and De Lannoy, G. (2024). Joint assimilation of satellite-based surface soil moisture and vegetation conditions into the Noah-MP land surface model. *Science of Remote Sensing*, vol. 9:p. 100129. ISSN 26660172. doi:10.1016/j.srs.2024.100129.

Ines, A. V., Das, N. N., Hansen, J. W., and Njoku, E. G. (2013). Assimilation of remotely sensed soil moisture and vegetation with a crop simulation model for maize yield prediction. *Remote Sensing of Environment*, vol. 138:pp. 149–164. ISSN 00344257. doi:10.1016/j.rse.2013.07.018.

IPCC (2021). Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change [Masson-Delmotte, V., P. Zhai, A. Pirani, S.L. Connors, C. Péan, S. Berger, N. Caud, Y. Chen, L. Goldfarb, M.I. Gomis, M. Huang, K. Leitzell, E. Lonnoy, J.B.R. Matthews, T.K. Maycock, T. Waterfield, O. Yelekçi, R. Yu, and B. Zhou (eds.)]. Cambridge University Press. ISBN 9781009157896, 2391 pp. doi:10.1017/9781009157896.

IPCC (2022). Climate Change 2022: Impacts, Adaptation and Vulnerability. Contribution of Working Group II to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change [H.-O. Pörtner, D.C. Roberts, M. Tignor, E.S. Poloczanska, K. Mintenbeck, A. Alegría, M. Craig, S. Langsdorf, S. Löschke, V. Möller, A. Okem, B. Rama (eds.)]. Cambridge University Press. ISBN 9781009325844, 3056 pp. doi:10.1017/9781009325844.

ISTAT (2024). Confini delle unità amministrative a fini statistici. URL `https://www.istat.it/notizia/confini-delle-unita-amministrative-a-fini-statistici-al-1-gennaio-2018-2/`.

Janjić, T., McLaughlin, D., Cohn, S. E., and Verlaan, M. (2014). Conservation of mass and preservation of positivity with ensemble-type Kalman filter algorithms. *Monthly Weather Review*, vol. 142:pp. 755–773. ISSN 00270644. doi:10.1175/MWR-D-13-00056.1.

Jin, N., Tao, B., Ren, W., He, L., Zhang, D., Wang, D., and Yu, Q. (2022). Assimilating remote sensing data into a crop model improves winter wheat yield estimation based on regional irrigation data. *Agricultural Water Management*, vol. 266:p. 107583. ISSN 03783774. doi:10.1016/j.agwat.2022.107583.

Kumar, S. V., Peters-Lidard, C. D., Tian, Y., Houser, P. R., Geiger, J., Olden, S., Lighty, L., Eastman, J. L., Doty, B., Dirmeyer, P., Adams, J., Mitchell, K., Wood, E. F., and Sheffield, J. (2006). Land information system: An interoperable framework for high resolution land surface modeling. *Environmental Modelling and Software*, vol. 21:pp. 1402–1415. ISSN 13648152. doi:10.1016/j.envsoft.2005.07.004.

Lahoz, W. A. and Schneider, P. (2014). Data assimilation: Making sense of Earth Observation. doi:10.3389/fenvs.2014.00016.

Liu, J., Williams, J. R., Zehnder, A. J., and Yang, H. (2007). GEPIC – modelling wheat yield and crop water productivity with high resolution on a global scale. *Agricultural Systems*, vol. 94:pp. 478–493. ISSN 0308521X. doi:10.1016/j.agsy.2006.11.019.

McJannet, D., Hawdon, A., Baker, B., Renzullo, L., and Searle, R. (2017). Multiscale soil moisture estimates using static and roving cosmic-ray soil moisture sensors. *Hydrology and Earth System Sciences*, vol. 21:pp. 6049–6067. ISSN 1607-7938. doi:10.5194/hess-21-6049-2017.

Moradkhani, H., Nearing, G., Abbaszadeh, P., and Pathiraja, S. (2018). Fundamentals of Data Assimilation and Theoretical Advances, pp. 1–26. Springer Berlin Heidelberg. doi:10.1007/978-3-642-40457-3_30-1.

Mosca, G. and Reyneri, A. (2023). Coltivazioni erbacee. Cereali e colture industriali, vol. 1. Edagricole Calderini. ISBN 978-8850656219.

Moulin, S., Bondeau, A., and Delecolle, R. (1998). Combining agricultural crop models and satellite observations: From field to regional scales. doi:10.1080/014311698215586.

Osborne, T., Gornall, J., Hooker, J., Williams, K., Wiltshire, A., Betts, R., and Wheeler, T. (2015). JULES-crop: a parametrisation of crops in the Joint UK Land Environment Simulator. *Geoscientific Model Development*, vol. 8:pp. 1139–1155. ISSN 1991-9603. doi:10.5194/gmd-8-1139-2015.

Peters-Lidard, C. D., Houser, P. R., Tian, Y., Kumar, S. V., Geiger, J., Olden, S., Lighty, L., Doty, B., Dirmeyer, P., Adams, J., Mitchell, K., Wood, E. F., and Sheffield, J. (2007). High-performance Earth system modeling with NASA/GSFC's Land Information System. *Innovations in Systems and Software Engineering*, vol. 3:pp. 157–165. ISSN 16145046. doi:10.1007/s11334-007-0028-x.

Piemonte, A. (2018). Il Clima in Piemonte - Autunno 2017. Tech. rep., ARPA Piemonte Sistemi Previsionali. URL `https://www.arpa.piemonte.it/sites/default/files/media/2023-12/Autunno2017.pdf`.

Raes, D., Steduto, P., Hsiao, T. C., and Fereres, E. (2025a). AquaCrop version 7.2 - Reference Manual, Chapter 1 FAO crop-water productivity model to simulate yield response to water, Technical report. Tech. rep., FAO. URL `https://ees.kuleuven.be/en/aquacrop/resources/reference-manual/chapter1aquacropversion7-2.pdf`.

Raes, D., Steduto, P., Hsiao, T. C., and Fereres, E. (2025b). AquaCrop version 7.2 - Reference Manual, Chapter 2 Users guide, Technical report. Tech. rep., FAO. URL `https://ees.kuleuven.be/en/aquacrop/resources/reference-manual/chapter2aquacropversion7-2.pdf`.

Raes, D., Steduto, P., Hsiao, T. C., and Fereres, E. (2025c). AquaCrop version 7.2 - Reference Manual, Chapter 3 Calculation procedures, Technical report. Tech. rep., FAO. URL `https://ees.kuleuven.be/en/aquacrop/resources/reference-manual/chapter3aquacropversion7-2.pdf`.

Rasmussen, I. S. and Thorup-Kristensen, K. (2016). Does earlier sowing of winter wheat improve root growth and N uptake? *Field Crops Research*, vol. 196:pp. 10–21. ISSN 03784290. doi:10.1016/j.fcr.2016.05.009.

Regione Emilia Romagna (2025). Disciplinari di Produzione Integrata. URL `https://agricoltura.regione.emilia-romagna.it/produzioni-agroalimentari/agricoltura-sostenibile/agricoltura-integrata/Collezione-dpi/dpi_2025`.

Regione Piemonte (2017). Ppr - Laghi (tav. P2). URL https://www.geoportale.piemonte.it/geonetwork/srv/ita/catalog.search#/metadata/r_piemon:4fb63c5b-813b-4c6d-9fed-6ce92df6ca64.

Regione Piemonte (2022). Uso del suolo agricolo su mosaicatura catastale di riferimento regionale 2021. URL https://www.geoportale.piemonte.it/geonetwork/srv/eng/catalog.search#/metadata/r_piemon:5f3b4327-41e2-4fa3-b7de-ccc66f9cf3ce.

Regione Piemonte (2023a). Carta dei suoli 1:50.000. URL https://www.geoportale.piemonte.it/geonetwork/srv/eng/catalog.search#/metadata/r_piemon:37c6413b-b07f-4f4c-9344-f2e43ea52bbd.

Regione Piemonte (2023b). Uso del suolo agricolo su mosaicatura catastale di riferimento regionale 2022. URL https://www.geoportale.piemonte.it/geonetwork/srv/eng/catalog.search#/metadata/r_piemon:3d164c06-6539-4298-ad56-f8c4161b659a.

Regione Piemonte (2024a). Anagrafe Agricola Unica - Data Warehouse e Open Data. URL http://www.sistemapiemonte.it/fedwanau/filtri.jsp.

Regione Piemonte (2024b). Uso del suolo agricolo su mosaicatura catastale di riferimento regionale 2023. URL https://www.geoportale.piemonte.it/geonetwork/srv/eng/catalog.search#/metadata/r_piemon:7573bb81-0c2c-46d9-b3f6-609d4e64e34e.

Reichle, R. H., Mclaughlin, D. B., and Entekhabi, D. (2002). Hydrologic Data Assimilation with the Ensemble Kalman Filter. *Monthly Weather Review*, vol. 130:pp. 103–114. doi:https://doi.org/10.1175/1520-0493(2002)130%3C0103:HDAWTE%3E2.0.CO;2.

Riahi, K., van Vuuren, D. P., Kriegler, E., Edmonds, J., O'Neill, B. C., Fujimori, S., Bauer, N., Calvin, K., Dellink, R., Fricko, O., Lutz, W., Popp, A., Crespo Cuaresma, J., KC, S., Leimbach, M., Jiang, L., Kram, T., Rao, S., Emmerling, J., Ebi, K., Hasegawa, T., Havlik, P., Humpenöder, F., Aleluia Da Silva, L., Smith, S., Stehfest, E., Bosetti, V., Eom, J., Gernaat, D., Masui, T., Rogelj, J., Strefler, J., Drouet, L., Krey, V., Luderer, G., Harmsen, M., Takahashi, K., Baumstark, L., Doelman, J. C., Kainuma, M., Klimont, Z., Marangoni, G., Lotze-Campen, H., Obersteiner, M., Tabeau, A., and Tavoni, M. (2017). The Shared Socioeconomic Pathways and their energy, land use, and greenhouse gas emissions implications: An overview. *Global Environmental Change*, vol. 42:pp. 153–168. ISSN 09593780. doi:10.1016/j.gloenvcha.2016.05.009.

RICA (2023). Italian survey data of winter wheat yield in Piemonte. URL https://rica.crea.gov.it/modulo_richiesta_dati.php.

Rivieccio, R., Di Bene, C., Paolanti, M., Marchetti, M., and Napoli, R. (2020). Soil rooting depth of Italy. *Journal of Maps*, vol. 16:pp. 36–42. ISSN 17445647. doi:10.1080/17445647.2019.1690595.

Ryu, D., Crow, W. T., Zhan, X., and Jackson, T. J. (2009). Correcting Unintended Perturbation Biases in Hydrologic Data Assimilation. *Journal of Hydrometeorology*, vol. 10:pp. 734–750. ISSN 1525-7541. doi:10.1175/2008JHM1038.1.

Steduto, P., Hsiao, T. C., and Fereres, E. (2007). On the conservative behavior of biomass water productivity. *Irrigation Science*, vol. 25:pp. 189–207. ISSN 03427188. doi:10.1007/s00271-007-0064-1.

Stöckle, C. O. and Kemanian, A. R. (2020). Can Crop Models Identify Critical Gaps in Genetics, Environment, and Management Interactions? doi:10.3389/fpls.2020.00737.

Stöckle, C. O., Kemanian, A. R., Nelson, R. L., Adam, J. C., Sommer, R., and Carlson, B. (2014). CropSyst model evolution: From field to regional to global scales and from research to decision support systems. *Environmental Modelling & Software*, vol. 62:pp. 361–369. ISSN 13648152. doi:10.1016/j.envsoft.2014.09.006.

Swinnen, E., Toté, C., and Van Hoolst, R. (2023). Copernicus Global Land Operations "Vegetation and Energy". Algorithm Theoretical Basis Document. Dry Matter Productivity (DMP), Gross Dry Matter Productivity (GDMP), Net Primary Production (NPP), Gross Primary Production (GPP). Collection 300m. Version 1.1. Issue 1.30. Tech. rep., CGLS.

Tebaldi, C. and Friedlingstein, P. (2013). Delayed detection of climate mitigation benefits due to climate inertia and variability. *Proceedings of the National Academy of Sciences of the United States of America*, vol. 110:pp. 17229–17234. ISSN 00278424. doi:10.1073/pnas.1300005110.

Tenreiro, T. R., García-Vila, M., Gómez, J. A., Jiménez-Berni, J. A., and Fereres, E. (2021). Using NDVI for the assessment of canopy cover in agricultural crops within modelling research. *Computers and Electronics in Agriculture*, vol. 182:p. 106038. ISSN 01681699. doi:10.1016/j.compag.2021.106038.

UCS (2023). UCS Satellite Database. URL `https://www.ucs.org/resources/satellite-database`.

Van Tricht, K., Degerickx, J., Gilliams, S., Zanaga, D., Battude, M., Grosu, A., Brombacher, J., Lesiv, M., Bayas, J. C. L., Karanam, S., Fritz, S., Becker-Reshef, I., Franch, B., Mollà-Bononad, B., Boogaard, H., Pratihast, A. K., Koetz, B., and Szantoi, Z. (2023). WorldCereal: a dynamic open-source system for global-scale, seasonal, and reproducible crop and irrigation mapping. *Earth System Science Data*, vol. 15(12):pp. 5491–5515. ISSN 1866-3516. doi:10.5194/essd-15-5491-2023. URL `https://essd.copernicus.org/articles/15/5491/2023/`.

Verger, A. and Descals, A. (2022). Copernicus Global Land Operations "Vegetation and Energy". Algorithm Theoretical Basis Document. Leaf Area Index (LAI), Fraction of Absorbed Photosynthetically Active Radiation (FAPAR) Fraction of green Vegetation Cover (FCover). Collection 300m. Version 1.1. Issue I1.10. Tech. rep., CGLS. URL `https://land.copernicus.eu/en/technical-library/algorithm-theoretical-basis-document-fraction-of-green-vegetation-cover-333-m-version-1.1/@@download/file`.

Wigneron, J.-P., Jackson, T., O'Neill, P., De Lannoy, G., de Rosnay, P., Walker, J., Ferrazzoli, P., Mironov, V., Bircher, S., Grant, J., Kurum, M., Schwank, M., Munoz-Sabater, J., Das, N., Royer, A., Al-Yaari, A., Al Bitar, A., Fernandez-Moran, R., Lawrence, H., Mialon, A., Parrens, M., Richaume, P., Delwart, S., and Kerr, Y. (2017). Modelling the passive microwave signature from land surfaces: A review of recent results and application to the L-band SMOS & SMAP soil

moisture retrieval algorithms. *Remote Sensing of Environment*, vol. 192:pp. 238–262. ISSN 00344257. doi:10.1016/j.rse.2017.01.024.

Wolfs, D., Swinnen, E., Van Hoolst, R., and Toté, C. (2023). Copernicus Global Land Operations "Vegetation and Energy", Product User Manual, Dry Matter Productivity (DMP), Gross Dry Matter Productivity (GDMP) Net Primary Production (NPP) Gross Primary Production (GPP), Collection 300m, Version 1.1, Issue I1.30. Tech. rep., Copernicus. URL `https://land.copernicus.eu/en/products/vegetation/dry-matter-productivity-v1-0-300m`.

Wolfs, D., Verger, A., Van der Goten, R., and Sánchez-Zapero, J. (2022). Copernicus Global Land Operations "Vegetation and Energy", Leaf Area Index (LAI), Fraction of Absorbed Photosynthetically Active Radiation (FAPAR) Fraction of green Vegetation Cover (FCover), Collection 300m, Version 1.1, Issue I1.20. Tech. rep., Copernicus. URL `https://land.copernicus.eu/en/products/vegetation/fraction-of-green-vegetation-cover-v1-0-300m`.

Yamazaki, D., Ikeshima, D., Tawatari, R., Yamaguchi, T., O'Loughlin, F., Neal, J. C., Sampson, C. C., Kanae, S., and Bates, P. D. (2017). A high-accuracy map of global terrain elevations. *Geophysical Research Letters*, vol. 44:pp. 5844–5853. ISSN 19448007. doi:10.1002/2017GL072874.

Yang, Z.-L., Niu, G.-Y., Mitchell, K. E., Chen, F., Ek, M. B., Barlage, M., Longuevergne, L., Manning, K., Niyogi, D., Tewari, M., and Xia, Y. (2011). The community Noah land surface model with multiparameterization options (Noah-MP): 2. Evaluation over global river basins. *Journal of Geophysical Research*, vol. 116:p. D12110. ISSN 0148-0227. doi:10.1029/2010JD015140.

Zhao, Q. and Yu, L. (2025). Advancing Sustainable Development Goals through Earth Observation Satellite Data: Current Insights and Future Directions. doi:10.34133/remotesensing.0403.

Appendices

Appendix A

Additional figures

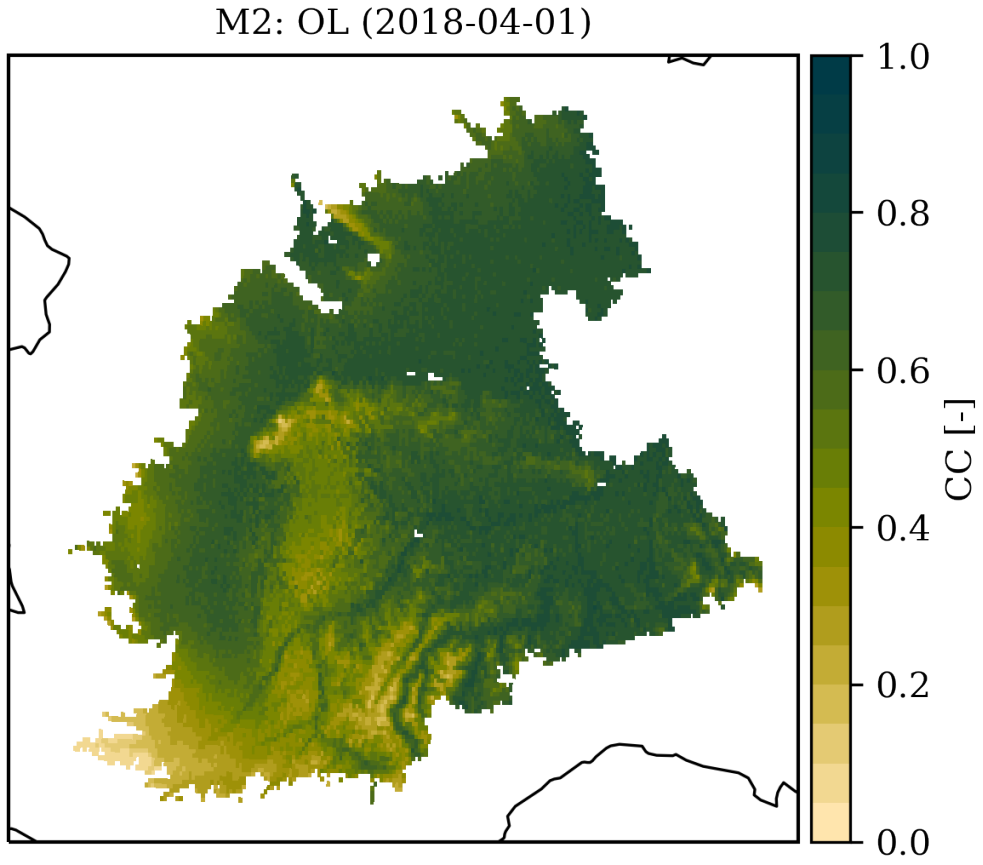

Figure A.1: Snapshot of model's ensemble mean canopy cover in the middle of the 2018 season for Mode 2. Visible patterns due to soil texture heterogeneity and spatial variation of forcing data.

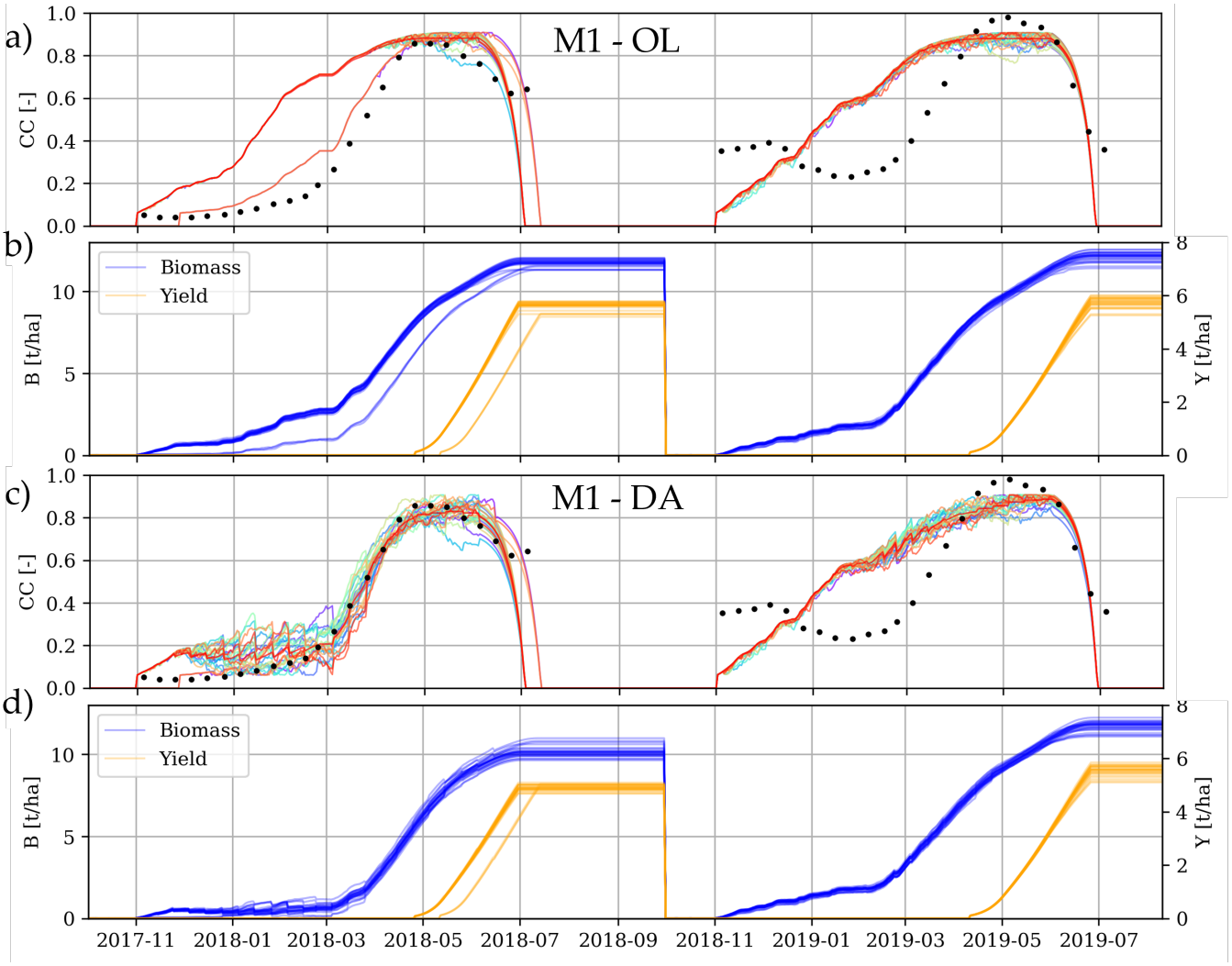

Figure A.2: Time series of the variables CC, B, and Y for years 2018 and 2019 of a model pixel in the Municipality of Alessandria (44°56'57.3"N : 8°33'13.5"E) for Mode 1. Plots a) and b) are related to the OL, while c) and d) show the effects of the DA updates. During Fall 2017, the delayed emergence of some members was caused by water stress.

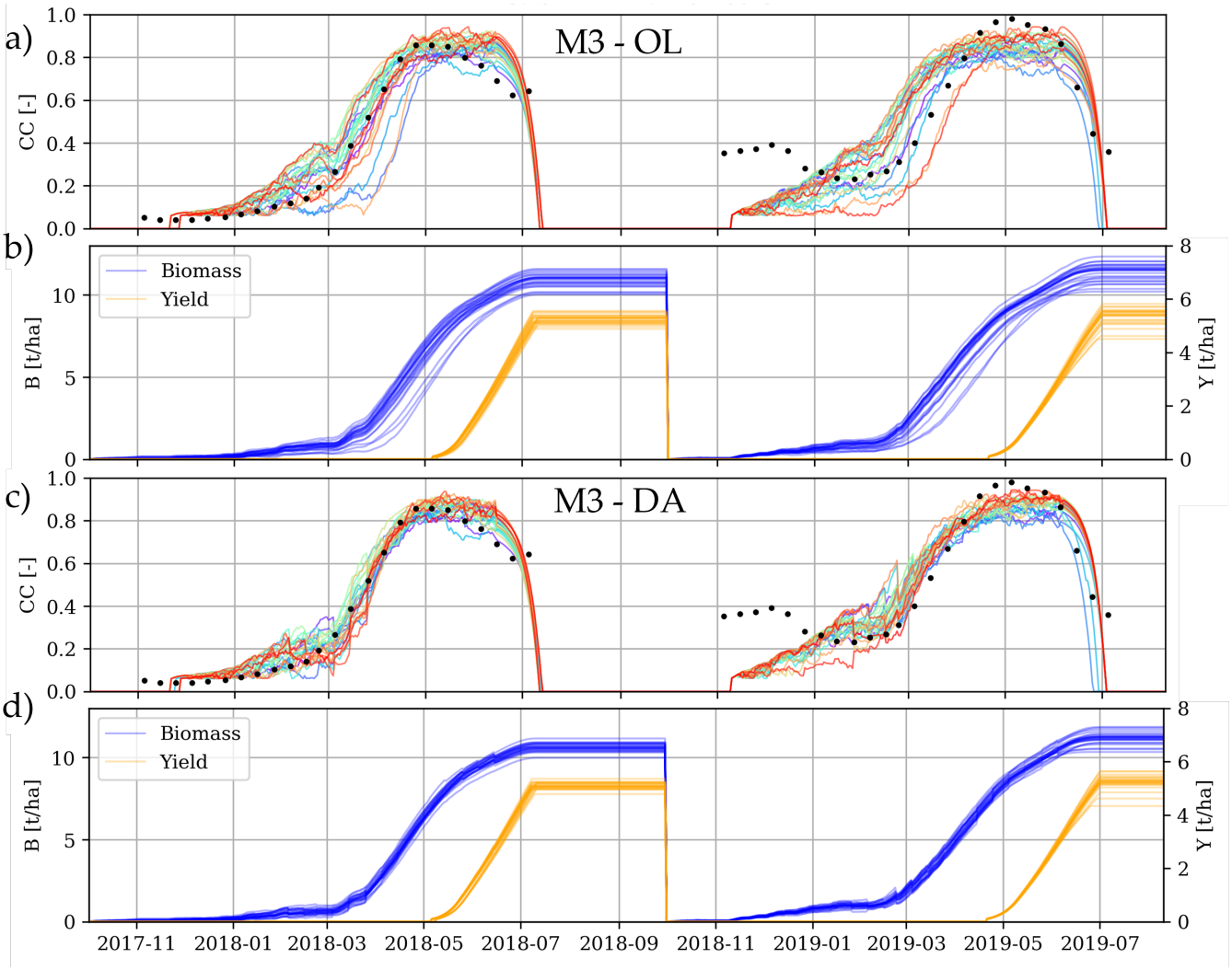

Figure A.3: Time series of the variables CC, B, and Y for years 2018 and 2019 of a model pixel in the Municipality of Alessandria (44°56'57.3"N : 8°33'13.5"E) for Mode 3. Plots a) and b) are related to the OL, while c) and d) show the effects of the DA updates.

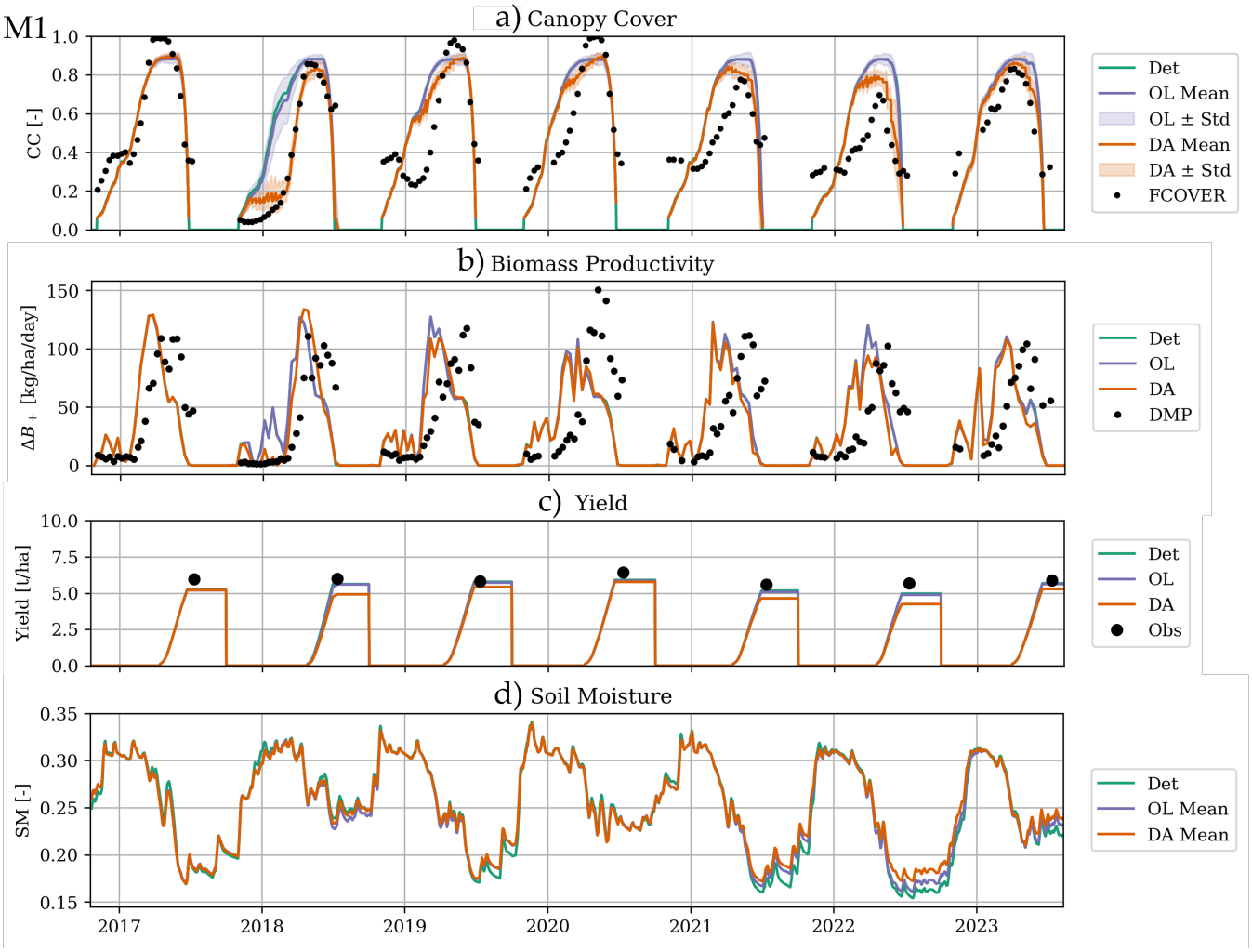

Figure A.4: Time series showing the state variables (a - canopy cover, b - 10-day aggregated biomass production, derived from B, and d - 7-day smoothed soil moisture) and the yield formation (c) in a model pixel in the Municipality of Alessandria (44°56'57.3"N : 8°33'13.5"E). Deterministic run in green, open-loop (OL) ensemble mean for Mode 1 in violet, and data assimilation (DA) ensemble mean for Mode 1 in orange. In graph a), the represented spread covers two standard deviations (± 1). The RICA-CREA values in graph c) are provided for reference only. For the analysis, the model pixels were aggregated at the municipality level prior to any comparisons.

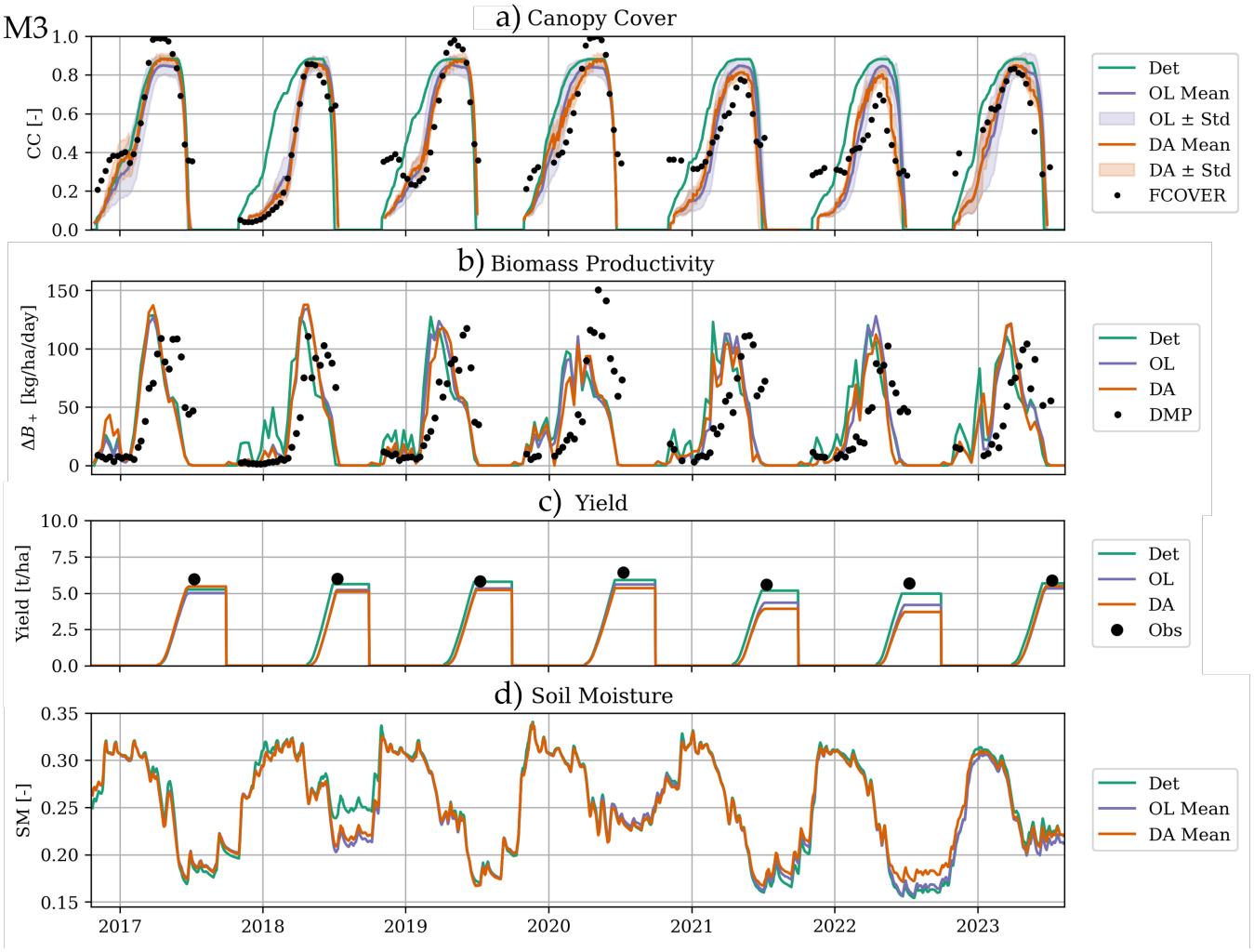

Figure A.5: Time series showing the state variables (a - canopy cover, b - 10-day aggregated biomass production, derived from B, and d - 7-day smoothed soil moisture) and the yield formation (c) in a model pixel in the Municipality of Alessandria (44°56'57.3"N : 8°33'13.5"E). Deterministic run in green, open-loop (OL) ensemble mean for Mode 3 in violet, and data assimilation (DA) ensemble mean for Mode 3 in orange. In graph a), the represented spread covers two standard deviations (± 1). The RICA-CREA values in graph c) are provided for reference only. For the analysis, model pixels were aggregated at the municipality level prior to any comparisons.

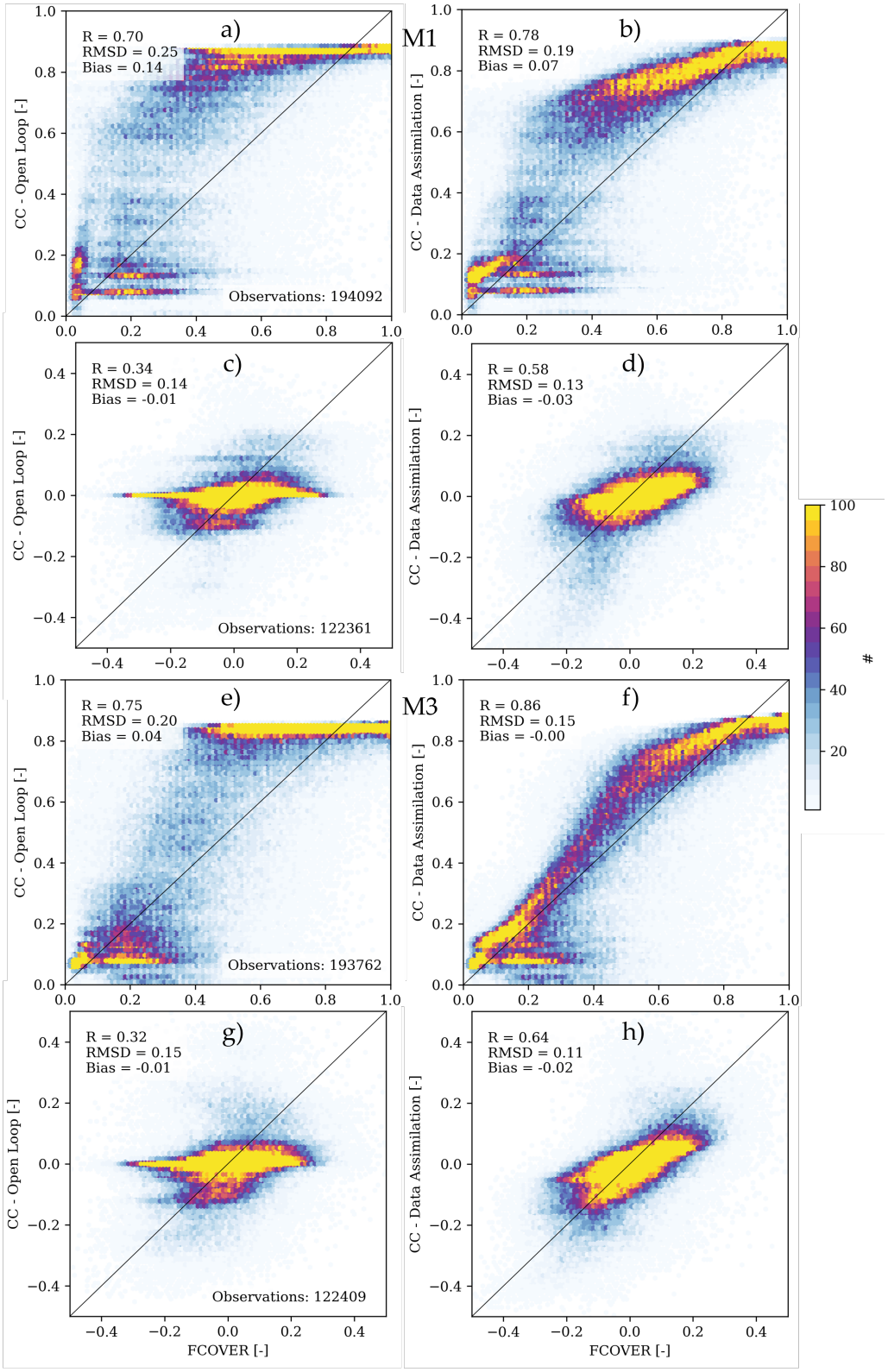

Figure A.6: Scatter plots display the 10-day aggregated canopy cover versus the FCOVER observations. A comparison between the absolute values from the OL (a) and DA (b) runs, and the seasonal anomalies (c and d), is shown for Mode 1, and the same is done for Mode 3 below (e, f, g, h).

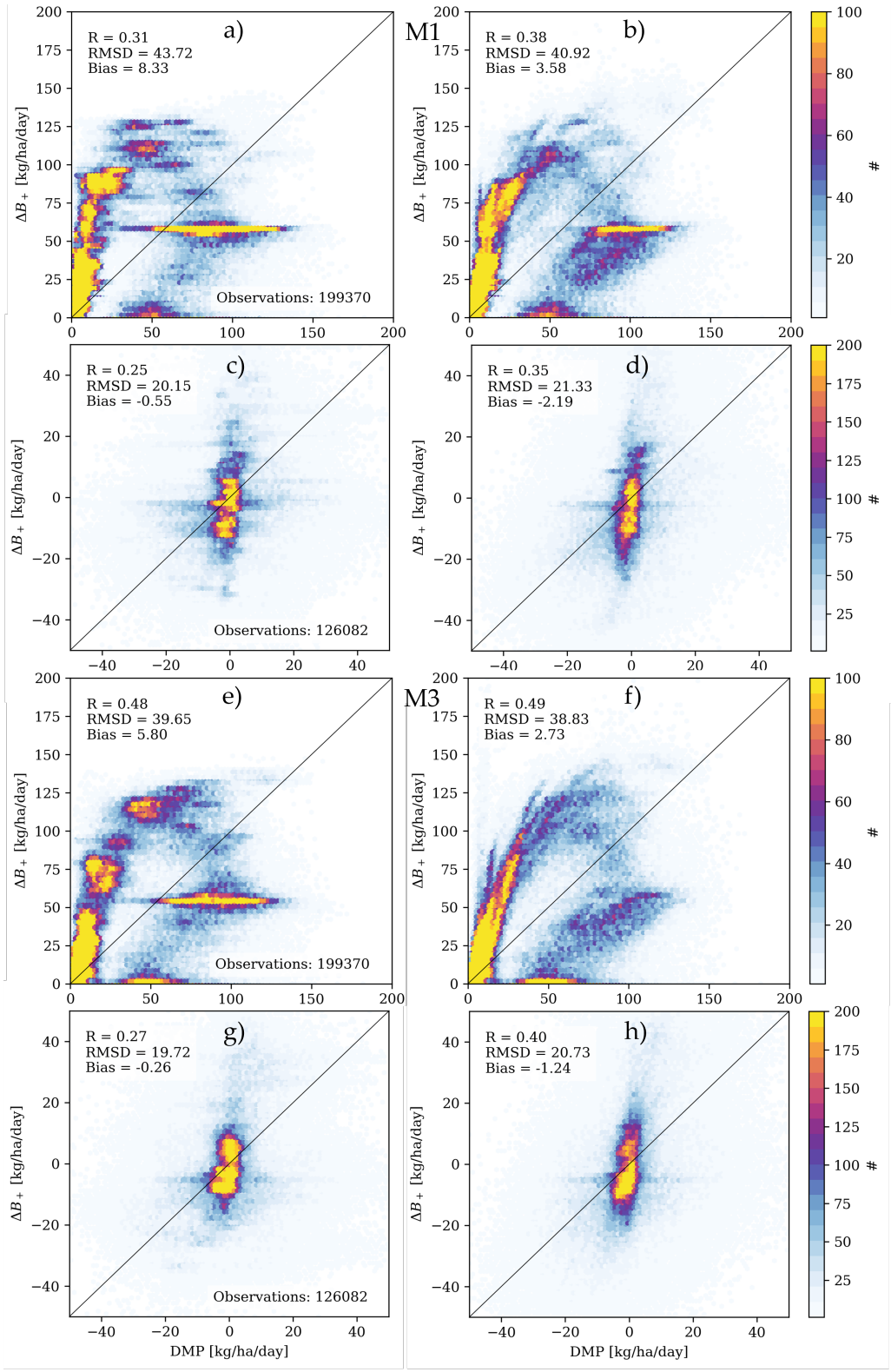

Figure A.7: Scatter plots display the 10-day aggregated above-ground biomass production versus the DMP observations. A comparison between the absolute values from the OL (a) and DA (b) runs, and the seasonal anomalies (c and d), is shown for Mode 1, and the same is done for Mode 3 below (e, f, g, h).

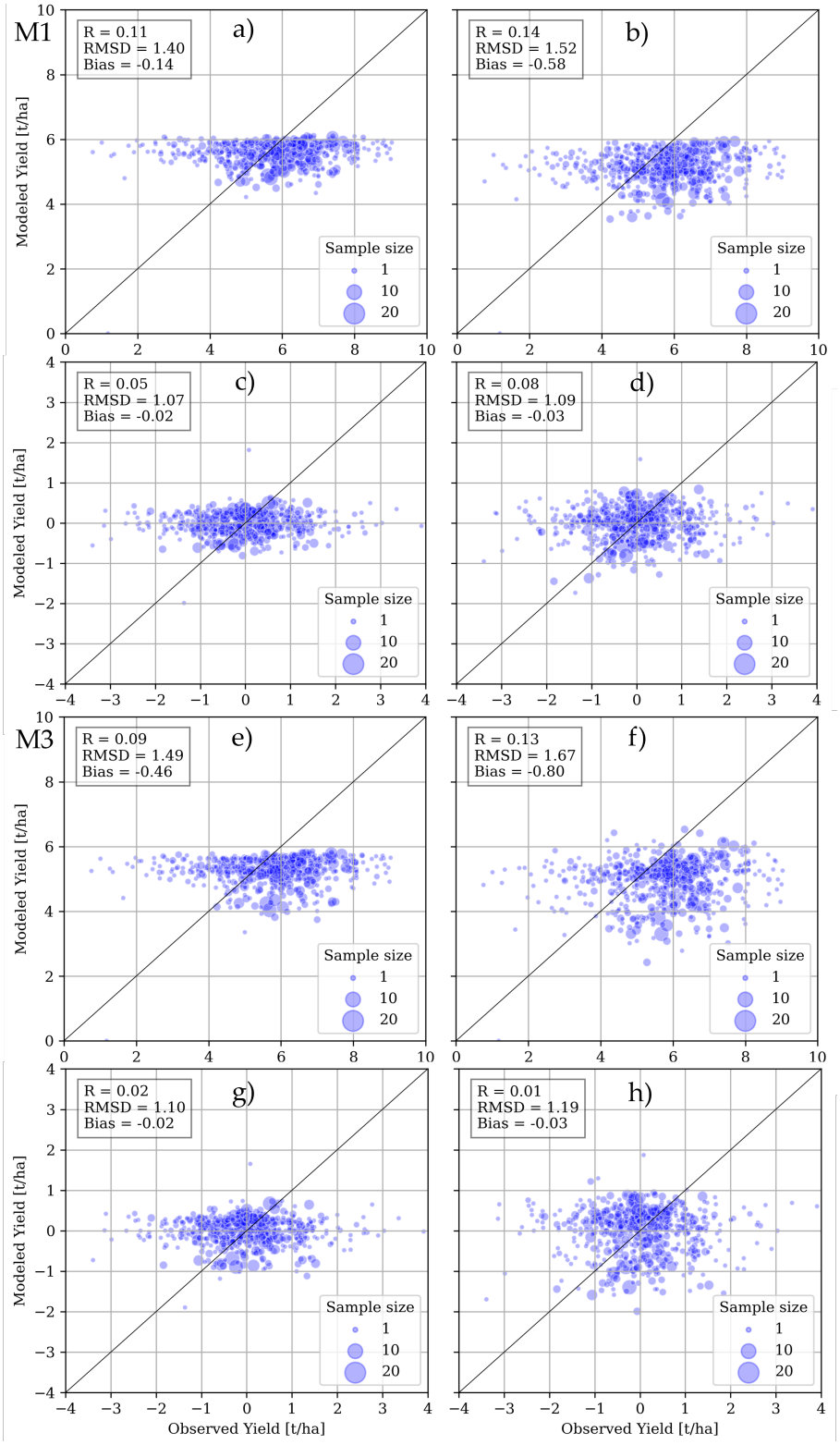

Figure A.8: Scatter plots display the annual dry yield formation versus the RICA-CREA survey. A comparison between the absolute values from the OL (a) and DA (b) runs, and the seasonal anomalies (c and d), is shown for Mode 1, and the same is done for Mode 3 below (e, f, g, h).

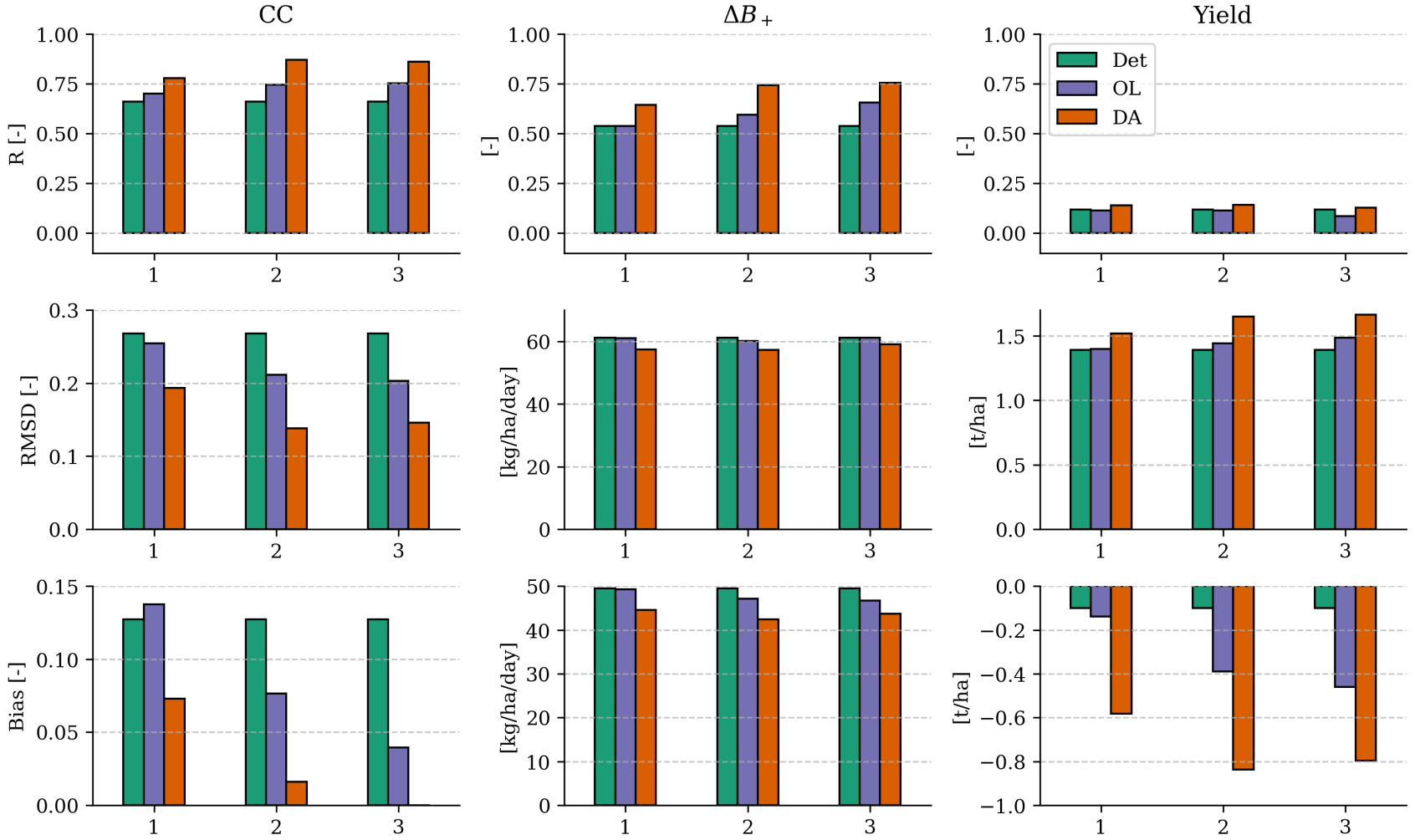

Figure A.9: Overall spatio-temporal performance of the open-loop runs (in violet) and data assimilation (in orange), compared to the deterministic baseline (in green) for the different modes (1, 2, and 3). The three crop-related state variables are displayed (canopy cover on the left, dry above-ground biomass production in the middle, and annual dry yield on the right). The metrics are computed over the whole dataset, within the growing seasons.

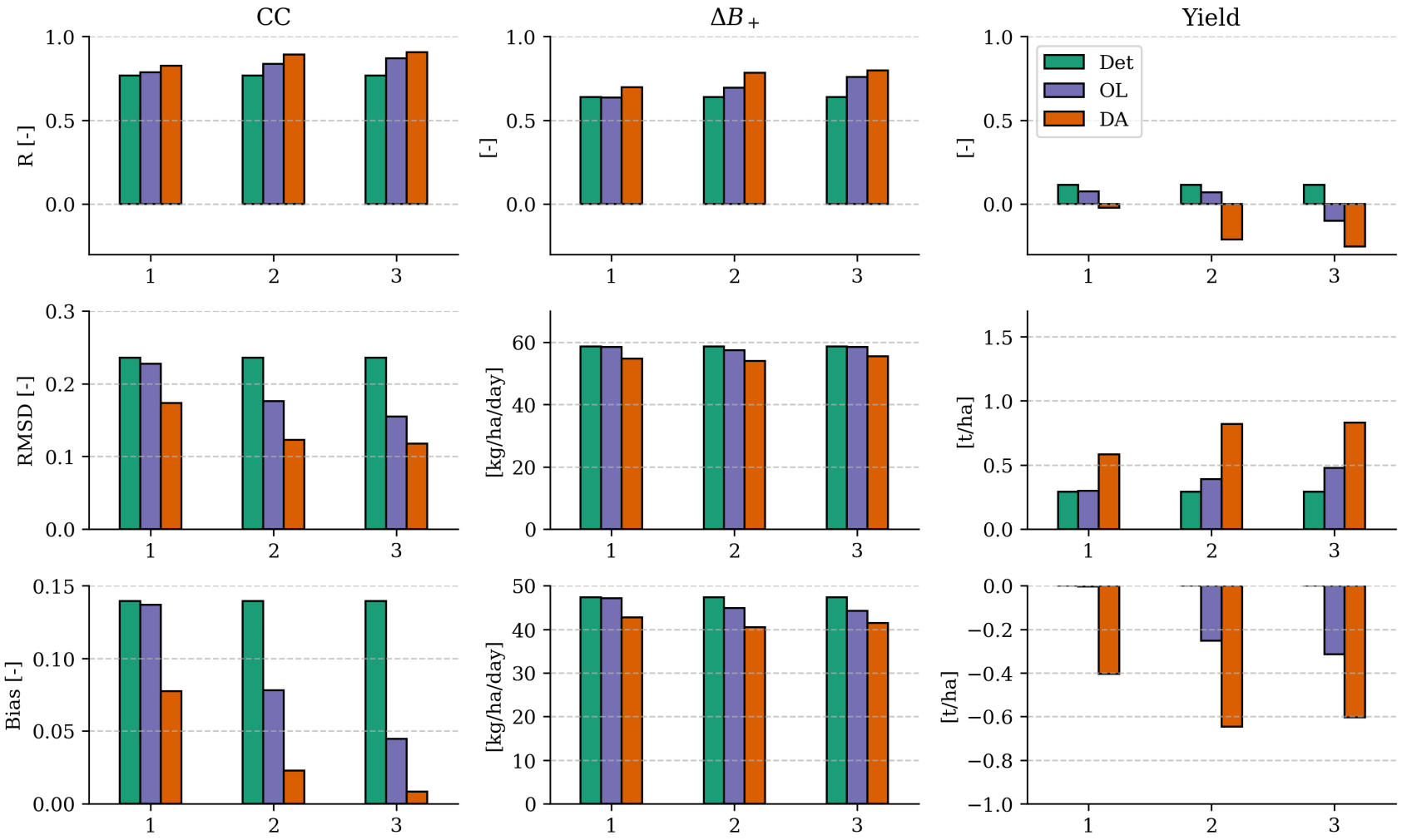

Figure A.10: Overall temporal performance of the open-loop runs (in violet) and data assimilation (in orange), compared to the deterministic baseline (in green) for the different modes (1, 2, and 3). The three crop-related state variables are displayed (canopy cover on the left, dry above-ground biomass production in the middle, and annual dry yield on the right). The values are first aggregated over the whole domain, and then the metrics are computed temporally. It is important to note that, since the duration of the analysis is only 7 years, the yield metrics are far from robust.

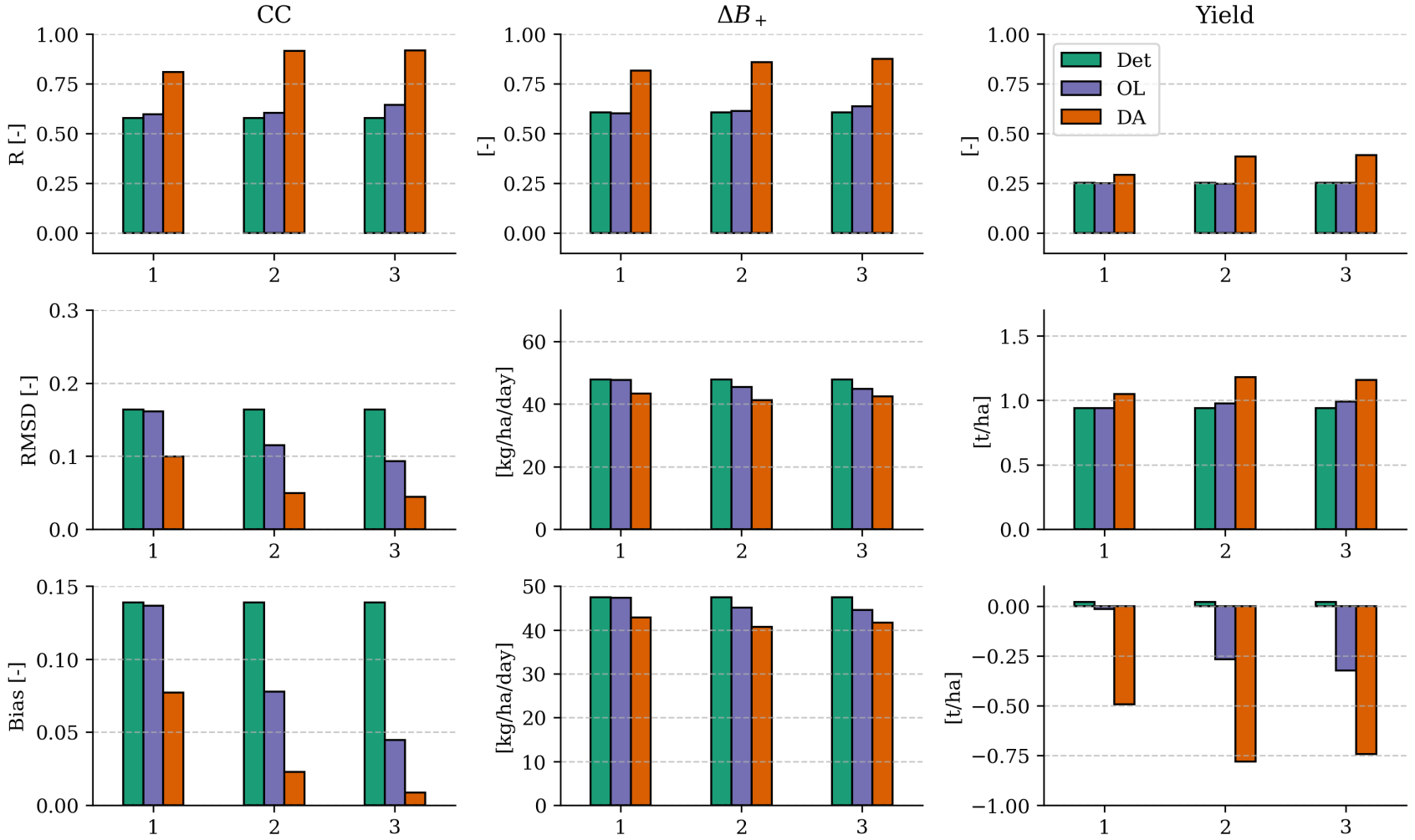

Figure A.11: Overall spatial performance of the open-loop runs (in violet) and data assimilation (in orange), compared to the deterministic baseline (in green) for the different modes (1, 2, and 3). The three crop-related state variables are displayed (canopy cover on the left, dry above-ground biomass production in the middle, and annual dry yield on the right). The values are first aggregated over time for each municipality, and then the metrics are computed spatially.

Summary (Italian)

AquaCrop, il modello di crescita delle colture dell'Organizzazione per l'alimentazione e l'agricoltura (FAO), è stato recentemente integrato nel *Land Information System Framework* della NASA (LISF, quadro del sistema di informazione territoriale). Ciò consente di effettuare stime sulle colture senza precedenti ed esperimenti di assimilazione dei dati satellitari (*data assimilation*, DA) su scala regionale. La DA satellitare mira a combinare un modello e le osservazioni per ridurre le incertezze nelle stime sulla dinamica di crescita delle colture. Questa tesi assimila il prodotto Copernicus relativo alla frazione di copertura vegetale (FCOVER) per aggiornare la copertura verde del suolo (*canopy cover*, CC) e la biomassa e, di conseguenza, le stime sulla resa del frumento invernale in Piemonte tra il 2017 e il 2023. Dopo aver calibrato i parametri delle colture, testato il modello e sviluppato le routine DA, sono stati generati tre insiemi (*ensemble*) di modelli (*modes*, modalità) perturbando il modello in vari modi per stimare l'incertezza delle previsioni del modello. I dati meteorologici e le variabili di stato sono stati perturbati in tutte le modalità. La modalità 1 includeva la correzione del bias di perturbazione; la modalità 2 no, così come la modalità 3, che includeva invece un'ulteriore variazione di parametri. Successivamente, le osservazioni FCOVER per il frumento sono state assimilate con una tecnica chiamata *ensemble Kalman filter* (EnKF). I risultati sono stati confrontati con altri prodotti satellitari e indagini sul campo relative alla resa.

La modalità 3, che presentava il maggior numero di gradi di libertà, ha portato alle migliori simulazioni basate esclusivamente sul modello rispetto ai dati di riferimento, e ha anche fornito gli aggiornamenti DA più significativi. L'assimilazione di dati ha migliorato CC (da design) e la produzione di biomassa secca fuori terra del modello; tuttavia, le stime di resa non hanno mostrato un chiaro miglioramento per tutte le modalità di generazione di *ensemble*. Gli incrementi della DA su CC sono stati limitati da un limite potenziale superiore (CC_{pot}), sia che questi valori fossero oltre i limiti fisici, sia a causa dell'asincronia tra la data di semina del modello e le

osservazioni, e questo a sua volta limita gli aggiornamenti alla resa. I risultati DA più promettenti sono stati ottenuti per la modalità *ensemble* 2. Sono necessari ulteriori studi per comprendere come affrontare l'incertezza della data di semina o delle fasi di crescita delle colture in generale. L'assimilazione congiunta dei dati relativi alla copertura vegetale e all'umidità del suolo dev'essere studiata al fine di superare il degrado dovuto agli aggiornamenti DA della vegetazione durante la propagazione delle informazioni all'interno del modello.