doi 10.26089/NumMet.v27r330

Comparison of data assimilation methods for integrating MODIS LAI into the WOFOST crop model for winter wheat yield prediction

Mikhail E. Gasanov
Lomonosov Moscow State University,
Department of Soil Science, Moscow, Russia
ORCID: 0000-0003-2938-3140, e-mail: gasanov.msu@yandex.ru

Anna Yu. Petrovskaia
Skoltech, Artificial Intelligence center, Moscow, Russia
ORCID: 0000-0002-7150-9184, e-mail: aypetrovskaia@gmail.com

Sergey A. Matveev
Lomonosov Moscow State University,
Department of Computational Mathematics and Cybernetics, Moscow, Russia
Marchuk Institute of Numerical Mathematics of RAS, Moscow, Russia
ORCID: 0000-0001-8000-8595, e-mail: sergey.matveev@gmail.com

Ivan V. Oseledets
Lomonosov Moscow State University,
Department of Artificial Intelligence, Moscow, Russia
Artificial Intelligence Research Institute (AIRI), Moscow, Russia
Skoltech, Artificial Intelligence center, Moscow, Russia
ORCID: 0000-0003-2071-2163, e-mail: i.oseledets@skoltech.ru

Abstract: This study presents a systematic comparison of three data assimilation methods for integrating MODIS Leaf Area Index observations into the WOFOST crop simulation model to predict winter wheat yield in Russia. We investigate the performance of Differential Evolution optimization, the Ensemble Kalman Filter, and the four-dimensional variational method within a unified computational pipeline, employing identical input data and evaluation metrics. A Sobol' sensitivity analysis is conducted to identify the most influential model parameters used as control variables in all three assimilation schemes. The results demonstrate both the relative strengths and limitations of each approach with respect to Leaf Area Index trajectory reproduction and predicting yield accuracy at the field scale.

Keywords: data assimilation, crop simulation model, WOFOST, MODIS LAI, Differential Evolution, Ensemble Kalman Filter, 4D-Var.

Acknowledgements: This work was supported by the Moscow Center of Fundamental and Applied Mathematics at INM RAS (Agreement with the Ministry of Education and Science of the Russian Federation No. 075–15–2025–347).

For citation: M. E. Gasanov, A. Yu. Petrovskaia, S. A. Matveev, and I. V. Oseledets, "Comparison of data assimilation methods for integrating MODIS LAI into the WOFOST crop model for winter wheat yield prediction," Numerical Methods and Programming. **27** (3), 470–488 (2026). doi 10.26089/NumMet.v27r330.

2026, **27** (3), 470–488. doi 10.26089/NumMet.v27r330

Сравнение методов ассимиляции данных MODIS LAI в модель WOFOST для прогнозирования урожайности озимой пшеницы

М. Э. Гасанов
Московский государственный университет имени М. В. Ломоносова, факультет почвоведения, Москва, Российская Федерация
ORCID: 0000-0003-2938-3140, e-mail: gasanov.msu@yandex.ru

А. Ю. Петровская
Сколтех, Центр искусственного интеллекта, Москва, Российская Федерация
ORCID: 0000-0002-7150-9184, e-mail: aypetrovskaia@gmail.com

С. А. Матвеев
Московский государственный университет имени М. В. Ломоносова, факультет вычислительной математики и кибернетики, Москва, Российская Федерация
Институт вычислительной математики имени Г. И. Марчука РАН, Москва, Российская Федерация
ORCID: 0000-0001-8000-8595, e-mail: sergey.matveev@gmail.com

И. В. Оселедец
Московский государственный университет имени М. В. Ломоносова, факультет искусственного интеллекта, Москва, Российская Федерация
Институт искусственного интеллекта (AIRI), Москва, Российская Федерация
Сколтех, Центр искусственного интеллекта, Москва, Российская Федерация
ORCID: 0000-0003-2071-2163, e-mail: i.oseledets@skoltech.ru

Аннотация: В данной работе проводится систематическое сравнение трех методов для ассимиляции наблюдений индекса листовой поверхности с радиометра MODIS в модель WOFOST с целью прогнозирования урожайности озимой пшеницы в России. Исследуется эффективность метода дифференциальной эволюции, ансамблевого фильтра Калмана и четырехмерного вариационного метода в рамках единого вычислительного подхода при идентичных входных данных и метриках оценки. Проводится анализ чувствительности методом Соболя для выявления наиболее значимых параметров модели, используемых в качестве управляющих переменных во всех трех схемах ассимиляции. Результаты демонстрируют как относительные преимущества, так и ограничения каждого подхода в отношении воспроизведения динамики индекса листовой поверхности и точности прогнозирования урожайности на масштабе отдельных полей.

Ключевые слова: ассимиляция данных, имитационная модель урожайности, WOFOST, MODIS LAI, дифференциальная эволюция, ансамблевый фильтр Калмана, 4D-Var.

Благодарности: Исследование выполнено при поддержке Московского центра фундаментальной и прикладной математики при ИВМ РАН (соглашение с Министерством науки и высшего образования Российской Федерации № 075–15–2025–347).

Для цитирования: Гасанов М.Э., Петровская А.Ю., Матвеев С.А., Оселедец И.В. Сравнение методов ассимиляции данных MODIS LAI в модель WOFOST для прогнозирования урожайности озимой пшеницы // Вычислительные методы и программирование. 2026. **27**, № 3. 470–488. doi 10.26089/NumMet.v27r330.

---

1. Introduction. Accurate prediction of crop yields at the field scale has become one of the central challenges of modern precision agriculture, driven by the increasing global food demand, accelerating climate variability, and the growing need for evidence-based site-specific agricultural management [1]. Precision agriculture frameworks rely on the ability to monitor and forecast crop development at fine spatial resolution, enabling timely interventions in irrigation, fertilization, and harvest scheduling, but achieving this at scale remains an open methodological problem [2].

Process-based crop models offer a physically and physiologically grounded approach to simulating crop growth dynamics, making them well-suited for integration into precision agriculture decision-support systems. Among the crops of greatest relevance for such systems, winter wheat (*Triticum aestivum* L.) represents a particularly important target. It is one of the most widely cultivated cereals globally and a dominant crop across the temperate agricultural zones of Eastern Europe and Central Asia, including the Russian Federation [3]. Reliable field-scale yield forecasting for winter wheat is therefore of both scientific and applied significance.

Nevertheless, the practical deployment of crop models in precision agriculture contexts faces persistent limitations. Spatial heterogeneity of soil properties, temporal variability of meteorological conditions, and the scarcity of in situ observations required for model parameterization all contribute to substantial uncertainty in simulated outputs. The diversity of management practices across fields and regions, combined with inherent uncertainty in phenological development, further constrains the predictive capacity of models calibrated on sparse ground truth data alone. These limitations motivate the integration of remote sensing observations into model workflows through data assimilation, since such observations provide spatially continuous, temporally dense, and increasingly high-resolution information on crop canopy state.

Process-based crop simulation models provide a mechanistic framework for quantitative analysis of plant growth by representing key eco-physiological processes, including phenological development, light interception, CO_2 assimilation, transpiration, respiration, and dry matter partitioning among plant organs [4, 5]. The most commonly used models of the agricultural production process, such as DSSAT [4], APSIM [5], WOFOST [6], and MONICA [7], include modules of soil water dynamics, atmospheric forcing, and crop physiology, allowing the simulation of daily biomass accumulation and final grain yield over the entire growing season. WOFOST (WOrld FOod STudies) is a single-point process-oriented simulation model with a daily time step, originally developed at Wageningen Agricultural University in the Netherlands [6]. In WOFOST, the production process is modeled on the basis of an eco-physiological approach: solar radiation drives canopy photosynthesis, and the resulting assimilates are allocated to roots, stems, leaves, and storage organs according to phenologically dependent partitioning tables [6]. Its open-source implementation via the Python Crop Simulation Environment (PCSE) has made WOFOST particularly accessible for large-scale and automated applications [8].

However, the performance of process-based models at the field scale is fundamentally limited by the uncertainty of internal model parameters, which are typically derived from experimental data collected under controlled conditions and may therefore not adequately represent specific site and season. This issue is compounded by the fact that the key meteorological and soil inputs required to drive such models are rarely available at field resolution. Gridded weather products, including reanalysis datasets and interpolated station networks, commonly operate at spatial resolutions of several kilometers, introducing systematic errors when applied to individual fields situated within heterogeneous landscapes. Similarly, publicly available soil databases such as SoilGrids or the European Soil Database provide hydraulic and textural properties at resolutions ranging from 250 m to several kilometers, which fail to capture the within-field variability that strongly governs water availability, rooting depth, and ultimately yield formation [9]. As a result, even a well-calibrated crop model may produce unreliable field-scale predictions when forced with spatially coarse inputs, highlighting the need for approaches that can compensate for this structural uncertainty through the assimilation of locally informative observations.

Satellite-based remote sensing has emerged as a practical source of spatially distributed observations of crop canopy development, offering the possibility to constrain process-based model simulations with measurements acquired throughout the growing season [10, 11]. The Moderate Resolution Imaging Spectroradiometer (MODIS) provides global Leaf Area Index (LAI) products at a temporal resolution of 4–8 days and a spatial resolution of 500 meters, which is suitable for monitoring vegetation dynamics over large agricultural regions [12]. Data assimilation, defined as the systematic integration of observational information into a numerical model to improve its trajectory and predictions, has been increasingly applied in crop science to reduce the discrepancy between simulated and remotely sensed phenological indicators such as LAI [11, 13]. For Indian agricultural conditions, a related workflow for integrating remote sensing data into crop simulation models was proposed in [14], demonstrating the practical feasibility of such approaches for field-scale yield analysis. In this framework, satellite observations serve not merely as independent validation data, but as an active forcing that corrects the model state or parameters during the simulation, thereby improving both the representation of canopy dynamics and the terminal yield estimate.

Several methodological paradigms exist for data assimilation in process-based models, each with distinct theoretical foundations and practical trade-offs. Optimization-based approaches treat the problem as parameter calibration and seek to minimize the discrepancy between modeled and observed LAI by adjusting model parameters directly. Among stochastic global optimizers, Differential Evolution (DE) is a population-based metaheuristic that explores the parameter space without requiring gradient information, making it well suited for non-smooth and multimodal cost functions typical of process-based crop models [15]. In contrast, sequential ensemble methods, notably the Ensemble Kalman Filter (EnKF), propagate an ensemble of model realizations forward in time and update model states at each observation time step using the Kalman gain matrix, thereby accounting for both model and observation uncertainties within a probabilistic framework [16]. The EnKF has been successfully applied to assimilate remote sensing observations into crop models, particularly for correcting LAI and biomass trajectories [17–19]. Variational methods, including the four-dimensional variational assimilation (4D-Var), adopt a different strategy: they formulate a cost functional that simultaneously penalizes the departure of the control variables from a prior (background) state and the misfit between the model trajectory and all available observations within the entire assimilation window [20, 21]. Unlike the EnKF, which processes observations sequentially, 4D-Var considers the full temporal distribution of observations at once and produces a single deterministic analysis, which can be advantageous when the observation density is irregular or when a globally optimal parameter set is desired. Variational data assimilation techniques have also been successfully applied in dynamic soil carbon model constructors [22], highlighting their suitability for terrestrial ecosystem applications. Although each of these methods has been applied individually to various crop models and regions [11, 13, 18], systematic comparisons of their performance for the same crop, model configuration, and observational dataset within a single experimental framework remain limited, particularly for Russian agricultural conditions.

In this study, we address this gap by conducting a systematic comparison of three data assimilation approaches (DE, EnKF, and 4D-Var) for integrating MODIS LAI observations into the WOFOST crop model to predict winter wheat yield in the central region of the Russian Federation. The main contributions of the present work are as follows. First, we implement all three assimilation methods within a unified computational pipeline based on the PCSE framework, ensuring identical input data, model configuration, and evaluation metrics for a fair comparison. Second, we perform a global sensitivity analysis using the Sobol' method to identify the most influential model parameters and employ them as control variables in all three assimilation schemes, thereby providing a consistent parameterization strategy. Third, we quantify the improvement in both LAI trajectory reproduction and yield prediction accuracy achieved by each method relative to the uncalibrated baseline simulation, and analyze the practical advantages and limitations of each approach for operational crop monitoring at the field scale.

The source code and data processing pipeline are made publicly available to ensure the reproducibility of the presented results.

2. Materials and methods.

2.1. WOFOST crop simulation model. Crop simulation models describe the dynamics of the atmosphere–soil–plant system and allow a user to evaluate crop productivity depending on weather conditions, soil properties, and agricultural management without conducting long-term field experiments. In our research, we utilize the WOFOST crop simulation model in its open-source implementation available in PCSE [6]. WOFOST is a single-point process-oriented model operating on a daily time step that describes dynamic growth processes, including phenological development, CO_2 assimilation, transpiration, respiration, and biomass partitioning among roots, stems, leaves, and storage organs. The fundamental point is that in WOFOST the production process of crops is modeled on the basis of an eco-physiological approach: solar radiation drives canopy photosynthesis, and the resulting assimilates are distributed according to phenologically dependent partitioning tables.

The WOFOST configuration used in this study is summarized as follows. The forcing inputs include daily meteorological variables (global radiation, minimum and maximum temperature, vapor pressure, precipitation, wind speed), soil hydraulic parameters (SMFCF, SMW, SM0, KSUB, K0, SOPE) derived from SoilGrids texture via pedotransfer functions, the Winter_wheat_104 crop parameter set (with the calibrated subset listed in Table 1), field management variables (sowing and harvest dates, initial available soil water WAV = 50 mm, initial available nitrogen NAVAILI = 25 kg/ha), and atmospheric CO_2 concentration of 360 ppm. The integrated state variables propagated daily by WOFOST include the development stage (DVS), LAI, the dry weights of leaves (WLV), stems (WST), storage organs (WSO) and roots (WRT), the total above-ground biomass (TAGP), and the rootzone soil moisture (SM, WWLOW). The principal model outputs used in this study are the LAI tra-

2026, **27** (3), 470–488. doi 10.26089/NumMet.v27r330

Table 1. WOFOST crop parameters considered for calibration

Parameter	Description	Units	Bounds
TSUM1	Temperature sum from emergence to anthesis	°C·day	[494.2, 917.8]
TSUM2	Temperature sum from anthesis to maturity	°C·day	[665.0, 1235.0]
TDWI	Initial total crop dry weight	kg/ha	[35.0, 65.0]
SPAN	Life span of leaves at 35°C	day	[21.91, 40.69]
RGRLAI	Maximum relative growth rate of LAI	1/day	[0.00574, 0.01066]
CVL	Efficiency of conversion into leaves	kg/kg	[0.4795, 0.8905]
CVO	Efficiency of conversion into storage organs	kg/kg	[0.4963, 0.9217]
CVS	Efficiency of conversion into stems	kg/kg	[0.4634, 0.8606]
RDMCR	Maximum rooting depth of the crop	cm	[87.5, 162.5]

jectory and the total weight of storage organs at maturity (TWSO, kg/ha dry matter). The observed field-level grain yield Y_{obs} is reported as dry-matter grain weight and is therefore directly comparable with the simulated TWSO. Throughout the manuscript, "simulated yield" refers to TWSO and "observed yield" refers to $Y_{\mathrm{obs}}^{\mathrm{dm}}$.

We chose the WOFOST model because it is well documented, calibrated for European crops and environmental conditions, and its rapid execution time of approximately 0.5–2 seconds per single WOFOST simulation of one field over a complete growing season (from sowing to maturity, ~ 330 daily integration steps), measured on a workstation with an Apple M2 (8 cores), 16 GB RAM, running Python 3.12 under macOS. In this study, we employ the water-limited free-drainage production mode (`Wofost71_WLP_FD`) with the Winter_wheat_104 parameterization for winter wheat.

Table 1 summarizes the crop and soil parameters considered in this study together with their default values and the calibration bounds used in the assimilation experiments. For crop parameters, the bounds are set to ±30% of the default values obtained from the PCSE crop parameter files.

2.2. Study area and field data. The study was conducted in the Oryol region of the Russian Federation, encompassing 45 agricultural fields. The study area is bounded by latitudes 51.84°–52.81° N and longitudes 36.83°–37.81° E and falls within the steppe zone characterized by a temperate continental climate with mean annual precipitation of 620–640 mm and mean annual temperature of 5.0–6.5°C.

According to the SoilGrids dataset, the soils of the study area are classified predominantly as chernozems (Haplic and Luvic Chernozems under the WRB classification). The mean textural composition of the topsoil across all 45 fields is characterized by a clay content of $26.4 \pm 2.5\%$, a sand content of $29.4 \pm 4.1\%$, and a silt content of $44.2 \pm 3.2\%$, corresponding to a silty clay loam texture. The mean bulk density was 1.09 ± 0.03 g/cm^3 and the mean soil organic carbon content was $6.2 \pm 1.0\%$, indicating high organic matter levels typical of chernozemic soils in the region.

The dataset comprises 45 agricultural fields cultivated with winter wheat during the 2019–2020 growing season. Sowing was performed between late August and mid-September 2019, and harvesting was carried out from mid-July to early September 2020. The observed grain yield across all fields ranged from 4758 to 8959 kg/ha, with a mean of 6311 kg/ha and a standard deviation of 752 kg/ha.

2.3. Weather and soil input data. The WOFOST crop model requires meteorological observations for each day of the growing season to perform simulations. We obtained daily weather data from the NASA Prediction of Worldwide Energy Resources (NASA POWER) database [23], which provides weather records from 1983 to the present with a spatial resolution of $0.5° \times 0.5°$ in latitude and longitude. The following meteorological variables were used for each simulation day: incoming global radiation (W/m^2), daily minimum and maximum temperature (°C), daily average vapor pressure (hPa), daily total precipitation (cm/day), and daily average wind speed at 2 m altitude (m/s).

Site-specific parameters were set to values consistent with the default PCSE configuration files for European winter wheat [6, 8]: initial available soil water WAV = 50 mm, atmospheric CO_2 concentration of 360 ppm (the historical baseline used in the original WOFOST formulation), and initial available nitrogen NAVAILI = 25 kg/ha. These values represent typical end-of-winter conditions for the temperate continental climate of the study region and have been adopted in previous WOFOST applications to Eastern European agricultural systems.

Soil texture data required for the WOFOST soil water balance module were obtained from the SoilGrids global dataset [9], which provides gridded predictions of sand, silt, and clay fractions at 250 m spatial resolution.

2026, **27** (3), 470–488. doi 10.26089/NumMet.v27r330

For each field, the soil texture fractions were extracted based on the geographic coordinates and used to initialize the soil hydraulic parameters through pedotransfer functions implemented in the pipeline.

The conversion of SoilGrids texture (sand, silt, clay fractions) into the soil hydraulic parameters required by the WOFOST water-balance module is performed using the Rosetta-3 pedotransfer function [24], a feed-forward artificial neural network trained on the UNSODA and related soil hydraulic databases. Given the texture fractions, Rosetta-3 predicts the van Genuchten–Mualem parameters θ_r, θ_s, $\log_{10}(\alpha/\alpha_0)$, $\log_{10} n$, and $\log_{10}(K_s/K_{s0})$ with α expressed in 1/cm, K_s — in cm/day and $\alpha_0 = 1$ 1/cm, $K_{s0} = 1$ cm/day. Before use, these values are back-transformed ($\alpha = 10^{\log_{10}\alpha}$, etc.). Using these parameters, the soil moisture retention $\theta(\psi)$ and unsaturated hydraulic conductivity $K(\psi)$ are evaluated as functions of the matric potential ψ (a pressure head in cm of water) over a grid of pF $= \log_{10}(\psi/\psi_0)$ spanning from -1 to 6 ($\psi_0 = 1$ cm), which corresponds to ψ values between 0.1 to 10^6 cm. This gives the piecewise-linear tabular inputs SMTAB and CONTAB required by WOFOST. The scalar WOFOST soil parameters are then derived from the same curves: the moisture content at saturation SM0 $= \theta(\psi = 0.1$ cm$)$, approximating near-saturated conditions; the moisture content at field capacity SMFCF $= \theta(\psi = 150$ cm$)$ (pF $= 2.2$, which corresponds to 15 kPa); the moisture content at the permanent wilting point SMW $= \theta(\psi = 15\,000$ cm$)$ (pF $= 4.2$, which corresponds to 1500 kPa); and the saturated hydraulic conductivity K0 $= K_s$.

2.4. MODIS LAI observations. Satellite-derived LAI observations were obtained from the MODIS instrument aboard the Terra and Aqua platforms. Specifically, we used the MCD15A3H.061 MODIS LAI/FPAR product [12], which provides global coverage at a spatial resolution of 500 m and a temporal composite period of 4 days, making it suitable for monitoring the seasonal dynamics of crop canopy development over agricultural regions. Field centroids were used to download LAI time series for vegetation period from Google Earth Engine [25].

The Savitzky–Golay filter [26] is applied with a symmetric window of length 13 (six observations on each side of the central point, corresponding to approximately ±24 days for the 4-day MODIS composite) and a polynomial order of 2. The window length was selected to suppress high-frequency noise while preserving the bell-shaped phenological signal of winter wheat. Because the filter is symmetric, the smoothed value at observation date t_k depends on raw observations at both $t < t_k$ and $t > t_k$.

2.5. Spatial and temporal data matching. WOFOST is a single-point model and therefore requires all inputs to be specified at a representative location for each field. We use the field centroid (computed from the field polygon) as the extraction point. NASA POWER weather data are queried at the centroid. The service returns the value of the enclosing $0.5° \times 0.5°$ grid cell, so fields located within the same NASA POWER cell receive identical weather forcing. SoilGrids texture fractions (sand, silt, clay) are sampled at the centroid from the native 250 m raster using nearest-neighbour extraction. MODIS MCD15A3H LAI is sampled from the 500 m pixel containing the centroid via the Google Earth Engine API, producing a 4-day composite time series. Temporally, weather forcing is provided at the native daily resolution required by WOFOST, while MODIS LAI observations are aligned to daily model time steps by matching each 4-day composite to its acquisition date. No temporal interpolation is applied between observations.

Field-level observed grain yields for the 2019–2020 winter wheat season were obtained from the operational farm management records of a commercial agricultural producer operating within the study region. The data were shared with the authors under a confidentiality agreement that permits their use for the present scientific analysis but precludes redistribution of the raw per-field values.

2.6. Sobol' sensitivity analysis. Sensitivity analysis is a methodology of qualitative and quantitative investigation of a model that helps to identify parameters most strongly affecting its output. In our research, we choose the variance-based global sensitivity analysis method developed by Sobol' [27, 28]. Previously Sobol' indices were successfully applied to sensitivity analysis of soil parameters in MONICA crop model [29].

Consider the model output as

$$Y = f(\boldsymbol{X}) = f(X_1, \ldots, X_p),$$

where f represents the WOFOST simulator, $\boldsymbol{X}$ are p varied input parameters, and Y is the model response (Crop yield or maximum LAI during season in our case). Following the approach by Sobol' [28], the multivariate function f can be represented using the Hoeffding decomposition:

$$f(X_1, \ldots, X_p) = f_0 + \sum_{i}^{p} f_i + \sum_{i}^{p} \sum_{j>i}^{p} f_{ij} + \ldots + f_{1\ldots p},$$

where $f_0 = E(Y)$ is a constant, $f_i = f_i(X_i)$ denotes main effects, and f_{ij} describes higher-order interactions. Under the assumption that the input parameters are independent, the total variance $V(Y)$ can be decomposed as:

$$V(Y) = \sum_i^p V_i + \sum_i^p \sum_{j>i}^p V_{ij} + \ldots + V_{1\ldots p},$$

where $V_i = V_{X_i}\left[E_{X_{\sim i}}(Y \mid X_i)\right]$ is the partial variance attributed to parameter X_i and the subscript "$\sim i$" denotes all input parameters except X_i. The first-order Sobol' index S_i and the total-effect index S_{T_i} are then defined as:

$$S_i = \frac{V_i}{V(Y)}, \qquad S_{T_i} = \frac{E_{X_{\sim i}}\left[V_{X_i}(Y \mid X_{\sim i})\right]}{V(Y)}.$$

The first-order index S_i quantifies the fraction of variance explained by parameter X_i alone, while the total-effect index S_{T_i} additionally accounts for all interaction terms involving X_i. Parameters with high S_{T_i} values are selected as control variables for the subsequent assimilation experiments.

In our implementation, we use the SALib library [30] to generate quasi-random Sobol' sequences with base sample size of $N = 1024$, which results in $N \times (p + 2)$ model evaluations, where p is the number of parameters. Second-order interactions are not computed to reduce the computational cost. The analysis is performed on a randomly selected reference field, and the top N_{top} parameters ranked by S_{T_i} are retained for calibration.

2.7. Data assimilation methods. The three methods compared in this study belong to two methodologically distinct families of data assimilation. DE and 4D-Var are formulated as *parameter calibration* schemes: they adjust a subset of WOFOST parameters so that the resulting model trajectory best matches the LAI observations, but they do not modify the internal state variables of the model at any intermediate time step. In contrast, the EnKF is a *state-update* scheme: at each observation time it directly corrects the state vector (LAI together with stem and storage organ dry weights), while the underlying parameters remain fixed. As a consequence, EnKF can propagate corrections from the assimilated variable (LAI) into biomass states (WST, WSO) that are mechanistically linked to terminal yield, whereas DE and 4D-Var must rely on the model's parameter sensitivity to translate the LAI fit into a yield improvement. We deliberately preserve this asymmetry to compare the methods in their canonical formulations.

§ 2.7.1. Differential Evolution optimization. The first assimilation approach treats the problem as a direct parameter calibration task, where the objective is to find the parameter vector $\boldsymbol{\theta}^*$ that minimizes the root mean square error (RMSE) between modeled and observed LAI:

$$\boldsymbol{\theta}^* = \arg\min_{\boldsymbol{\theta}\in\Theta} \sqrt{\frac{1}{N_{\text{obs}}} \sum_{k=1}^{N_{\text{obs}}} \left(\text{LAI}_k^{\text{sim}}(\boldsymbol{\theta}) - \text{LAI}_k^{\text{obs}}\right)^2},$$

where Θ denotes the bounded parameter space defined in Table 1, N_{obs} is the number of observation dates, and $\text{LAI}_k^{\text{sim}}(\boldsymbol{\theta})$ is the model LAI at the k-th observation date obtained by running WOFOST with parameters $\boldsymbol{\theta}$.

We employ the DE algorithm [15] — a population-based stochastic optimizer that maintains a population of candidate solutions and evolves them through mutation, crossover, and selection operators. DE does not require gradient information and is well suited for non-smooth, multimodal cost functions arising from process-based crop models. In our experiments, the population size is set to 30 individuals per parameter, the convergence tolerance to 10^{-2}, and the maximum number of iterations to 50.

The optimization is performed independently for each field. The set of active control variables is selected from the Sobol' sensitivity analysis (Section 2.6), while their feasible ranges are defined by the parameter bounds listed in Table 1. The initial DE population is sampled uniformly within these bounds.

§ 2.7.2. Ensemble Kalman Filter. EnKF [16] is a sequential data assimilation method that represents the probability distribution of model states through an ensemble of N_e model realizations. In our implementation, the ensemble is constructed by generating $N_e = 100$ instances of the WOFOST model with parameters sampled from normal distributions centered at the default values with a relative standard deviation of 15%.

Let $\boldsymbol{x}_t^{(i)} \in \mathbb{R}^n$ denote the state vector of the i-th ensemble member at time t, comprising the model variables subject to assimilation: LAI, WST, and WSO. At each observation time t_k, the ensemble states are collected into the state matrix:

$$\boldsymbol{A} = \left[\boldsymbol{x}_{t_k}^{(1)}, \ \ldots, \ \boldsymbol{x}_{t_k}^{(N_e)}\right] \in \mathbb{R}^{n\times N_e},$$

and the ensemble covariance is estimated as $\boldsymbol{P}_e = \operatorname{cov}(\boldsymbol{A})$. The observation vector $\boldsymbol{y}_k$ (MODIS LAI) is perturbed with Gaussian noise to form the matrix $\boldsymbol{D} \in \mathbb{R}^{m \times N_e}$, and the observation error covariance $\boldsymbol{R}_e$ is estimated from the perturbed ensemble. The observation operator $\boldsymbol{H} \in \mathbb{R}^{m \times n}$ maps the full state vector to the observed variables (in our case, selecting LAI). The Kalman gain matrix $\boldsymbol{K}$ and the updated (analysis) state matrix $\boldsymbol{A}^a$ are computed as:

$$\boldsymbol{K} = \boldsymbol{P}_e \boldsymbol{H}^T \left(\boldsymbol{H}\boldsymbol{P}_e\boldsymbol{H}^T + \boldsymbol{R}_e\right)^{-1}, \qquad \boldsymbol{A}^a = \boldsymbol{A} + \boldsymbol{K}(\boldsymbol{D} - \boldsymbol{H}\boldsymbol{A}).$$

After each update, the corrected states are imposed back into the ensemble members with enforced physical constraints (LAI $\geqslant 0.01$, WST $\geqslant 0$, WSO $\geqslant 0$). The ensemble is then propagated forward to the next observation date, and the procedure is repeated until all observations have been processed. After the final observation, the ensemble is run to the end of the growing season to produce yield estimates. The final yield prediction is taken as the ensemble mean of the TWSO.

The standard deviation of the LAI observation error is specified as $\sigma_{\text{obs}} = \max\{0.1 \cdot \text{LAI}^{\text{obs}},\ 0.1\}$, which assigns a relative error of 10% with a minimum floor of 0.1 to account for the uncertainty of MODIS retrievals.

§ 2.7.3. Four-dimensional variational method (4D-Var). The 4D-Var method [20] formulates the assimilation problem as the minimization of a cost functional that simultaneously accounts for the prior information on the model parameters and the misfit with all available observations within the assimilation window. Unlike EnKF, which processes observations sequentially, 4D-Var considers the entire temporal distribution of LAI measurements at once.

Let $\boldsymbol{x} \in \mathbb{R}^p$ denote the control vector (model parameters) and $\boldsymbol{x}_b$ stand for the background (prior) state obtained from the default PCSE parameter files. The cost functional to be minimized is:

$$J(\boldsymbol{x}) = \underbrace{\frac{1}{2}(\boldsymbol{x} - \boldsymbol{x}_b)^T \boldsymbol{B}^{-1} (\boldsymbol{x} - \boldsymbol{x}_b)}_{J_b} + \underbrace{\frac{1}{2}\big(\boldsymbol{y} - \mathcal{H}[\mathcal{M}(\boldsymbol{x})]\big)^T \boldsymbol{R}^{-1} \big(\boldsymbol{y} - \mathcal{H}[\mathcal{M}(\boldsymbol{x})]\big)}_{J_o},$$

where $\mathcal{M}(\boldsymbol{x})$ denotes the forward WOFOST run with parameters $\boldsymbol{x}$, $\mathcal{H}$ is the observation operator extracting the simulated LAI at the dates of MODIS acquisitions, $\boldsymbol{y} \in \mathbb{R}^{N_{\text{obs}}}$ is the vector of observed LAI values, and $\boldsymbol{R}$ is the observation error covariance matrix.

The background error covariance matrix is taken diagonal, $\boldsymbol{B} = \operatorname{diag}(\sigma_{b,1}^2, \ldots, \sigma_{b,p}^2)$, with $\sigma_{b,j} = 0.15 \cdot |x_{b,j}|$, corresponding to a 15% relative uncertainty on each control parameter. This value is consistent with the range of inter-cultivar and inter-site variability reported in published WOFOST parameterization studies for European winter wheat [8]. The diagonal form expresses the assumption of mutually independent prior errors on the control parameters, which is a common simplification in crop-model data assimilation given the moderate dimension of the control vector ($p = 4$ in this study).

The observation error covariance matrix is also taken diagonal, $\boldsymbol{R} = \operatorname{diag}(\sigma_{\text{obs},1}^2, \ldots, \sigma_{\text{obs},N_{\text{obs}}}^2)$, with $\sigma_{\text{obs},k} = \max\{0.1 \cdot y_k,\ 0.1\}$. The relative component of 10% corresponds to the validation uncertainty reported for the MCD15A3H LAI retrieval product, while the constant floor of 0.1 prevents the cost functional from becoming dominated by low-LAI observations during emergence and senescence, where the relative error of the retrieval grows large. Extensive studies in oceanographic 4D-Var applications have demonstrated how the analysis sensitivity and model-state accuracy depend on the observation error statistics and their spatio-temporal structure [31].

The background term J_b acts as a regularizer that penalizes departures from the prior parameter estimates, preventing the optimization from converging to physically implausible values. The observation term J_o drives the solution toward consistency with the satellite data across the full growing season.

Since WOFOST does not generally guarantee the existence of an analyticity, differentiability and adjointness, the gradient is approximated using first-order forward finite differences as follows:

$$\big[\nabla J(\boldsymbol{x})\big]_j \approx \frac{J(\boldsymbol{x} + \varepsilon_j \boldsymbol{e}_j) - J(\boldsymbol{x})}{\varepsilon_j},$$

here $\varepsilon_j = 10^{-3} \cdot \max\{|x_j|, 1\}$. Step was selected on the basis of a dedicated calibration experiment reported in Table A1, values of ε below 10^{-4} fall below the numerical resolution of the WOFOST integration and yield a vanishing finite-difference gradient, while values above 10^{-3} introduce truncation error. The selected step lies at the minimum of the yield RMSE response curve and ensures that the gradient reflects the genuine sensitivity of the model to the control parameters rather than floating-point round-off noise.

The cost functional is minimized using the limited-memory Broyden–Fletcher–Goldfarb–Shanno algorithm with bound constraints (L-BFGS-B) [32]. This algorithm imposes physically admissible box constraints on the control parameters requiring only the gradient approximated above.

The robustness of the 4D-Var solution to the choice of $\boldsymbol{B}$ and $\boldsymbol{R}$ was assessed by repeating the assimilation on five representative fields over a grid of $\sigma_b/|x_b| \in \{0.05,\ 0.10,\ 0.15,\ 0.25,\ 0.50\}$ and $\sigma_{\mathrm{obs}}/y \in \{0.05,\ 0.10,\ 0.20\}$.

2.8. Evaluation metrics. To quantify the agreement between model predictions and observations, we employ the following standard metrics. RMSE measures the average magnitude of the prediction error:

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(\hat{y}_i - y_i\right)^2},$$

where $\hat{y}_i$ and y_i are the simulated and observed values, respectively, and N is the number of data points. The Pearson correlation coefficient r quantifies the linear relation between simulated and observed values:

$$r = \frac{\sum\limits_{i=1}^{N}\left(\hat{y}_i - \overline{\hat{y}}\right)}{\sqrt{\sum\limits_{i=1}^{N}\left(\hat{y}_i - \overline{\hat{y}}\right)^2 \sum\limits_{i=1}^{N}\left(y_i - \overline{y}\right)^2}},$$

where $\overline{\hat{y}}$ and $\overline{y}$ denote the means of the simulated and observed values over N data points, respectively. Throughout this paper, an overbar symbol stands for arithmetic mean and in the aggregated 4D-Var results (Tables A1 and A2) it averaged over the five fields. The bias is defined as:

$$\mathrm{Bias} = \frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i).$$

These metrics are computed both for the LAI trajectory (using temporally synchronized MODIS and model LAI values) and for the final grain yield (using the observed field-level yield and the model-predicted TWSO).

2.9. Computational pipeline. The complete computational pipeline is implemented in Python 3.12 and organized as a modular package. The workflow proceeds through the following stages:

1. *Data loading and baseline simulation.* Field metadata (coordinates, sowing and harvest dates, observed yield), MODIS LAI, and SoilGrids data are loaded from parquet files prepared from Google Earth Engine data. For each field, a baseline WOFOST simulation is executed with default parameters, and the NASA POWER weather provider is initialized and cached to avoid redundant API requests.
2. *Sobol' sensitivity analysis.* A global sensitivity analysis is performed on a randomly selected reference field to rank all candidate parameters by their total-effect index S_{T_i}. The top N_{top} parameters are selected as the active control variables for calibration.
3. *Per-field assimilation.* For each field, three independent assimilation procedures are executed using the same input data: DE optimization, EnKF, and 4D-Var. Each method produces an optimized LAI trajectory and a yield estimate.
4. *Evaluation and comparison.* RMSE, Pearson correlation coefficient r, and bias are computed for each method with respect to the MODIS LAI and the observed field-level yield. Per-field and aggregated metrics are stored in parquet files, and diagnostic plots are generated for visual comparison of the methods.

All three assimilation methods operate within a unified framework: they receive identical input data (field coordinates, weather, soil, MODIS LAI), use the same WOFOST model version, and are evaluated with the same set of metrics, thereby ensuring a fair comparison. The pipeline is parallelized at the field level and executed using the `scipy` [33] optimization library for DE and 4D-Var, and a custom EnKF implementation based on PCSE model instances.

3. Results.

3.1. Sensitivity analysis. Figure 1 presents the results of the Sobol' global sensitivity analysis performed on five randomly selected fields from the study area. The heatmap displays the total-effect indices S_T for each of the nine candidate parameters listed in Table 1, computed separately for grain yield (TWSO, Figure 1a) and

2026, **27** (3), 470–488. doi 10.26089/NumMet.v27r330

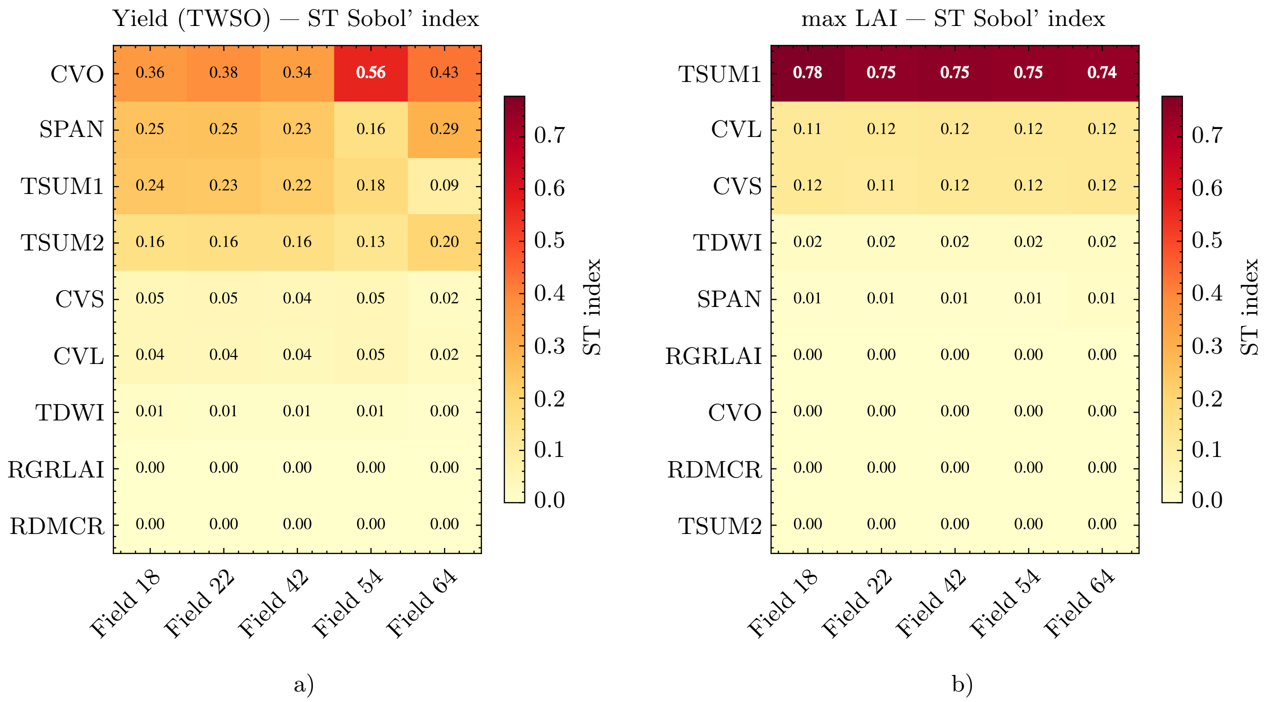

Figure 1. Sobol' total-effect sensitivity indices S_T for all candidate parameters across five fields: a) grain yield (TWSO); b) maximum LAI. Parameters are ordered by their mean S_T for yield in descending order

maximum leaf area index (LAI, Figure 1b). Each column corresponds to one field, allowing direct assessment of the spatial consistency of parameter importance.

As can be seen from Figure 1, crop yield is predominantly controlled by CVO (conversion efficiency to storage organs) across all five fields, with a mean $S_T = 0.414 \pm 0.090$, which substantially exceeds the influence of other parameters. The ranking of the remaining parameters is also consistent: SPAN (leaf lifespan) ranks second position ($\bar{S}_T = 0.236 \pm 0.048$), followed by TSUM1 (temperature sum from emergence to anthesis, $\bar{S}_T = 0.192 \pm 0.062$) and TSUM2 (temperature sum from anthesis to maturity, $\bar{S}_T = 0.164 \pm 0.025$). All other parameters contribute negligibly to yield variance ($\bar{S}_T < 0.05$).

In contrast, maximum LAI is almost exclusively determined by TSUM1 across all fields, with $\bar{S}_T = 0.753 \pm 0.015$, indicating that phenological timing is the primary and highly consistent driver of canopy development. CVL and CVS contribute marginally to LAI dynamics ($\bar{S}_T \approx 0.12$), whereas their combined effect on yield is modest.

For the dominant parameters, the first-order and total-effect indices are nearly equal: $S_1 \approx S_T$ for CVO on yield (0.407 vs. 0.414) and for TSUM1 on LAI (0.744 vs. 0.753), suggesting that these parameters act largely independently. In contrast, the larger gap observed for TSUM1 on yield ($S_1 = 0.140$, $S_T = 0.192$) indicates meaningful interaction with other phenologically-related parameters.

Based on these results, the four parameters with the highest mean total-order indices (CVO, SPAN, TSUM1, and TSUM2) were selected as the active control variables for all subsequent data assimilation experiments.

3.2. LAI trajectory improvement. Figure 2 presents the LAI dynamics for random field in a multi-panel view, comparing the MODIS observations (orange dots) with the simulated LAI from the baseline run and the three assimilation methods.

All three assimilation methods substantially improve the agreement between simulated and observed LAI compared to the baseline, which exhibits a mean LAI RMSE of 0.975 ± 0.282 across all 45 fields. The EnKF achieves the lowest mean LAI RMSE of 0.638 ± 0.162 (median 0.629), benefiting from sequential state updates that directly correct the model trajectory at each observation date. The DE optimization produces a comparable mean LAI RMSE of 0.667 ± 0.177 (median 0.657); the cost function is explicitly designed for this objective, yet the stochastic nature of the evolutionary search introduces greater variability across fields compared to the

2026, **27** (3), 470–488. doi 10.26089/NumMet.v27r330

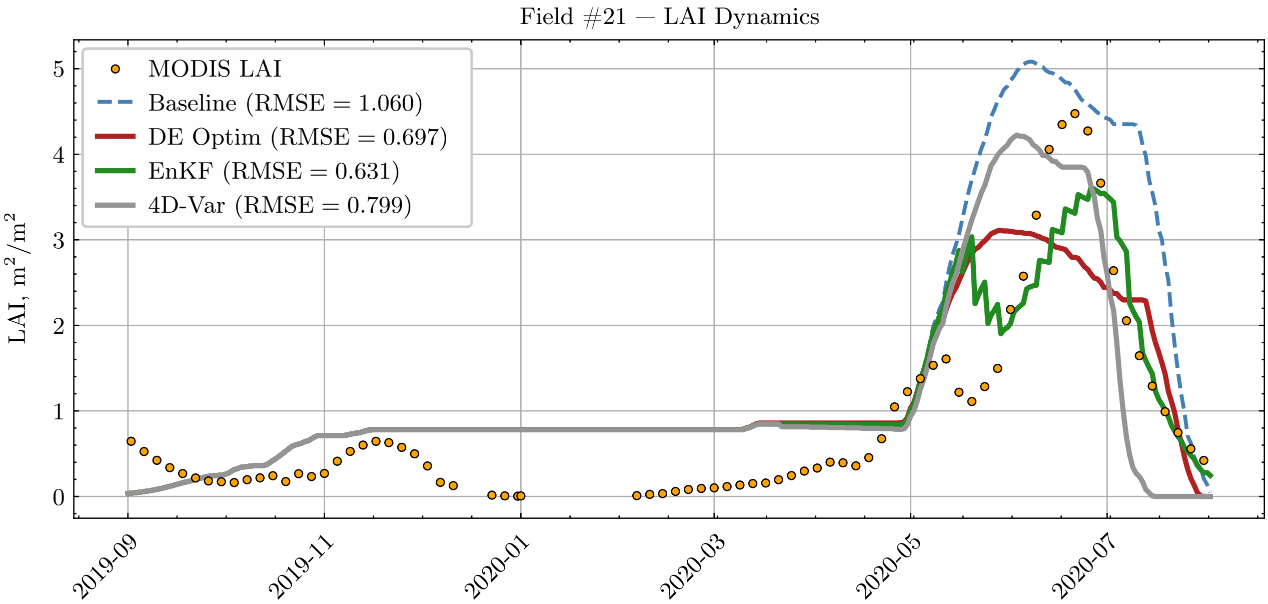

Figure 2. Detailed LAI dynamics for a representative field with RMSE values for each method indicated in the legend

EnKF. The 4D-Var method yields a mean LAI RMSE of 0.799 ± 0.200 (median 0.799), which is higher than both EnKF and DE but still represents an 18% reduction relative to the baseline. This method balances the fit to observations against departure from the background state through the J_b term, producing smooth LAI trajectories that remain physically consistent. A notable feature of the 4D-Var results is the regularizing effect of the background penalty, which prevents the parameters from attaining extreme values that might overfit the LAI observations at the expense of yield prediction accuracy.

3.3. Baseline simulation. Before applying the assimilation methods, a baseline WOFOST simulation was performed for each field using the default parameterization. Figure 3 (“Baseline”) displays the scatter plot of baseline-simulated versus observed yield. The uncalibrated model produces a systematic overestimation of yield for most fields, with a mean simulated yield of 9001 kg/ha compared to the observed mean of 6311 kg/ha, resulting in an RMSE of 2829 kg/ha and a bias of 2716 kg/ha. This overestimation can be attributed to the fact that the default winter wheat parameters were derived under Western European conditions and do not account for the specific soil–climate interactions in the study region. The baseline LAI trajectories, shown as dashed lines in Figure 2, tend to overestimate the peak LAI and exhibit a delayed senescence phase compared to the MODIS observations, further confirming the need for site-specific calibration.

3.4. Yield prediction accuracy. Figure 3 shows the scatter plots of observed versus simulated yield for all four scenarios: baseline, DE optimization, EnKF, and 4D-Var. Each panel displays the RMSE, Pearson correlation coefficient r, and bias metrics computed across all fields.

All three assimilation methods reduce the yield prediction RMSE and bias relative to the baseline simulation (RMSE = 2829 kg/ha, bias = 2716 kg/ha). The EnKF achieves the lowest bias among all methods (247 kg/ha) and a yield RMSE of 1309 kg/ha, benefiting from updating not only LAI but also the stem and storage organ dry weights, which provides a more direct pathway to improved yield estimates. The 4D-Var method attains a yield RMSE of 1307 kg/ha, virtually identical to EnKF, together with the highest Pearson correlation ($r = 0.158$), which can be attributed to the regularization provided by the background term J_b that constrains the calibrated parameters within physically plausible ranges while simultaneously accounting for all observations within the assimilation window. The DE optimization reduces RMSE to 2177 kg/ha and the bias to 1370 kg/ha. However, despite producing a better LAI fit than 4D-Var, it exhibits a near-zero Pearson correlation with observed yield ($r = -0.002$), indicating that the LAI calibration objective is not directly aligned with terminal grain yield and that excessive parameter adjustment may degrade yield prediction.

Figure 4 complements this analysis by presenting the distribution of simulated yields across methods. The observed yield has a median of 6200 kg/ha (IQR: 5775–6875 kg/ha), whereas the baseline simulation produces a

2026, **27** (3), 470–488. doi 10.26089/NumMet.v27r330

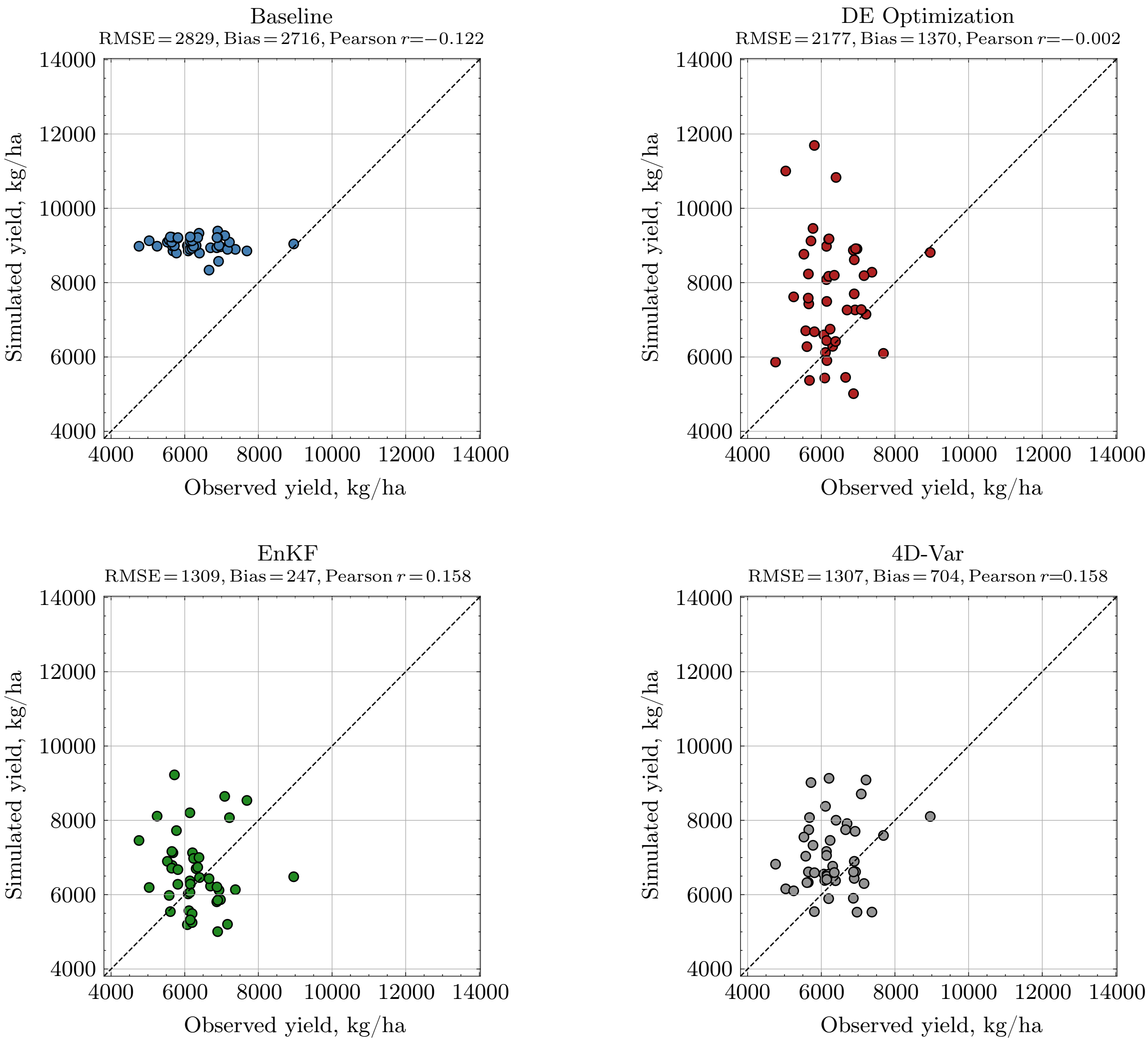

Figure 3. Scatter plots of observed versus simulated yield for the baseline and three assimilation methods. The dashed line represents the 1:1 correspondence

narrow cluster around 9001 kg/ha with a standard deviation of only 228 kg/ha, confirming the limited variability of the default parameterization. The EnKF recovers a distribution closest to the observed range, with a median of 6698 kg/ha and a standard deviation of 1173 kg/ha, while the DE optimization and 4D-Var produce medians of 7584 and 7683 kg/ha, respectively. The smallest spread (std = 626 kg/ha) among the assimilation methods is achieved by 4D-Var, in agreement with the regularizing effect of the background penalty.

3.5. Comparison of assimilation methods. Figure 5 provides a comprehensive comparison of the three methods across all fields. Figure 5a shows the box plots of per-field LAI RMSE for each method, illustrating both the central tendency and the spread of the LAI fit quality. Figure 5b presents the distribution of yield prediction errors (simulated minus observed) for each method, with the zero line indicating unbiased prediction.

The results indicate that EnKF achieves the lowest median LAI RMSE of 0.629, closely followed by DE at 0.657, while 4D-Var attains a median of 0.799. In terms of yield prediction, the mean absolute bias is lowest for EnKF (247 kg/ha), followed by 4D-Var (704 kg/ha) and DE (1370 kg/ha). Furthermore, 4D-Var exhibits the smallest standard deviation of yield errors (1114 kg/ha compared to 1711 kg/ha for DE and 1300 kg/ha for EnKF), indicating more stable predictions across fields. The EnKF and 4D-Var thus achieve comparable

2026, **27** (3), 470–488. doi 10.26089/NumMet.v27r330

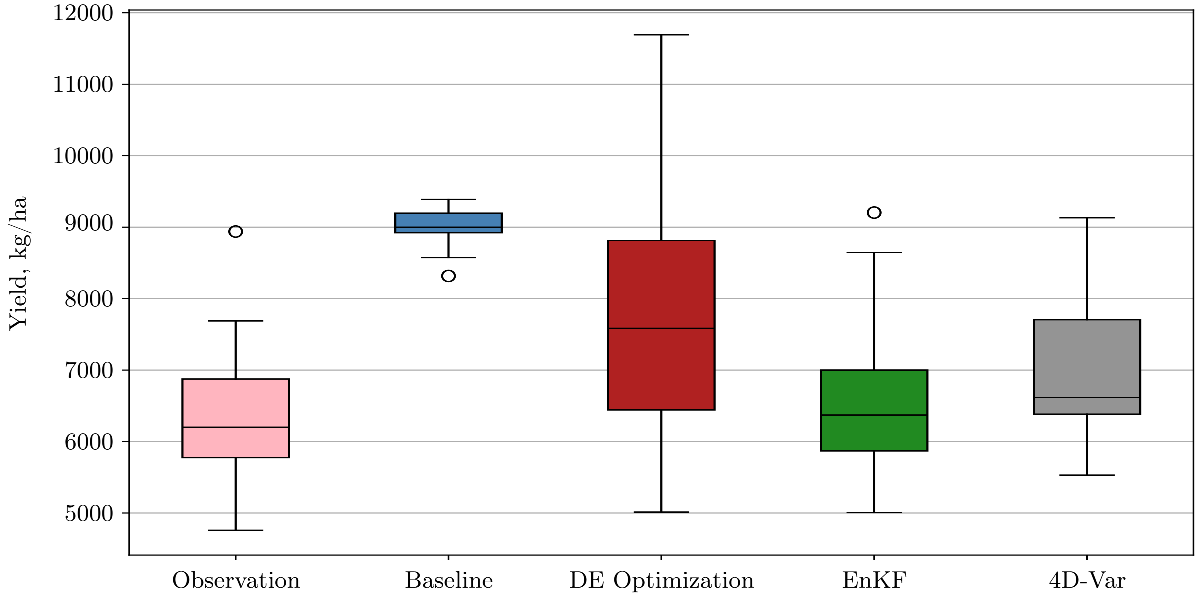

Figure 4. Distribution of simulated yield by methods: observed yield and predictions from baseline, DE optimization, EnKF, and 4D-Var

yield RMSE (1309 and 1307 kg/ha, respectively), with EnKF offering a substantially lower bias and 4D-Var providing the most stable error distribution through background regularization. Despite a competitive LAI fit, DE demonstrates the weakest yield prediction performance among the three methods.

The relatively poor yield prediction accuracy of the DE optimization, despite its competitive LAI RMSE, warrants further discussion. The DE algorithm minimizes a cost function defined solely in terms of LAI discrepancy, without any constraint on the realism of the adjusted parameters. Consequently, the optimization may converge to parameter combinations that reproduce the observed LAI trajectory but are physically inconsistent: for instance, crop-specific parameters controlling light use efficiency or assimilate partitioning may be shifted to extremes that fit the observed LAI during the vegetative phase while distorting grain-filling dynamics. Furthermore, the stochastic nature of the evolutionary search, combined with the relatively small population size and a moderate number of generations, introduces variability in the final solution, as reflected in the large standard deviation of yield errors (1711 kg/ha). In contrast, EnKF sequentially updates not only LAI but also biomass state variables (stem and storage organ dry weights) that are more directly linked to yield formation, and the 4D-Var method constrains the parameter space through the background term J_b, which penalizes large departures from physically meaningful prior values. These

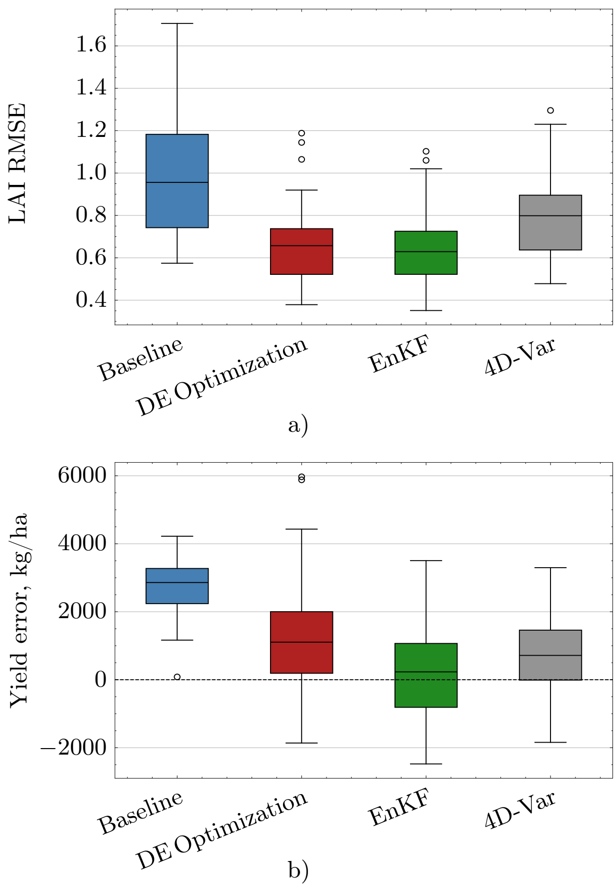

Figure 5. Comparison of assimilation methods across all fields: a) LAI RMSE distribution; b) yield error distribution

2026, **27** (3), 470–488. doi 10.26089/NumMet.v27r330

structural differences explain the superior yield prediction accuracy of EnKF and the greater stability of the 4D-Var approach compared to the DE optimization.

Figure 6 provides a complementary view of the yield improvement achieved by each method. Each panel plots the absolute baseline yield error on the horizontal axis against the absolute error of the corresponding assimilation method on the vertical axis. Points below the diagonal indicate fields where the method improved upon the baseline.

As can be seen from Figure 6, the majority of fields fall below the diagonal for all three methods: EnKF improves yield predictions for 42 out of 45 fields (93.3%) and 4D-Var for 41 out of 45 fields (91.1%), while the DE optimization improves 35 fields (77.8%). The mean error reduction relative to the baseline amounts to 1664 kg/ha for EnKF, 1672 kg/ha for 4D-Var, and 1075 kg/ha for DE, confirming that sequential and variational assimilation methods provide a more reliable improvement across fields than the parameter-only evolutionary approach. The improvement is particularly pronounced for fields where the baseline error is large, suggesting that data assimilation is most beneficial when the default model parameterization is poorly suited to local conditions. For fields with small baseline errors, the assimilation methods provide modest additional improvement, and in a few cases the yield error may marginally increase due to the indirect relationship between LAI fit and yield accuracy.

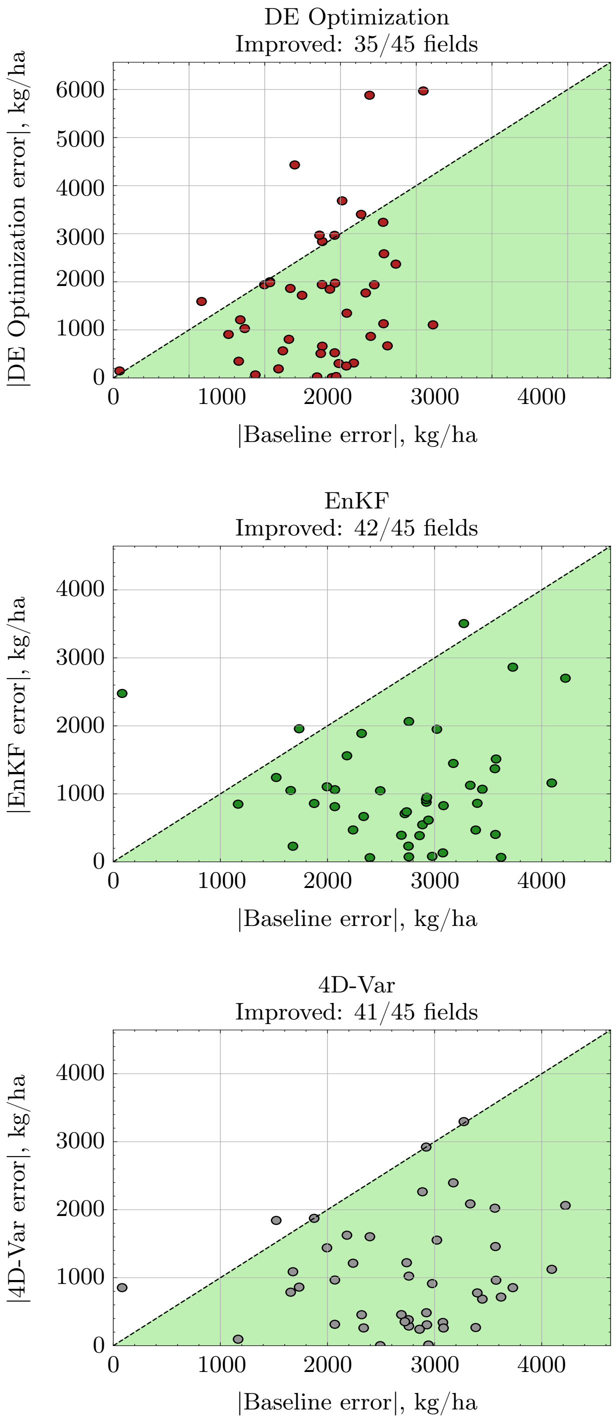

Figure 6. Yield error comparison: absolute baseline error versus absolute assimilation error for each method. Points below the diagonal indicate improvement over the baseline

3.6. Robustness of 4D-Var to covariance specification. To assess the sensitivity of the 4D-Var solution to the specification of the background and observation error covariance matrices, the assimilation procedure was repeated on five representative fields, selected to span the yield quantiles of the dataset, over a 5×3 grid of relative uncertainties: $\sigma_b/|x_b| \in \{0.05,\ 0.10,\ 0.15,\ 0.25,\ 0.50\}$ for the diagonal of $\boldsymbol{B}$ and $\sigma_{\mathrm{obs}}/y \in \{0.05,\ 0.10,\ 0.20\}$ for the relative component of $\boldsymbol{R}$. The complete grid (15 configurations, 75 4D-Var runs) is reported in Appendix 1.

The yield prediction RMSE varied between 3177 and 3406 kg/ha across the full grid, corresponding to a relative variation of ±3.5% around the ensemble mean. In the regime $\sigma_b/|x_b| \geqslant 0.10$ and $\sigma_{\mathrm{obs}}/y \leqslant 0.10$, RMSE differed by less than 20 kg/ha across configurations, confirming that the solution reported in the previous subsections lies on a plateau of stability rather than on the slope of a hyperparameter-sensitive optimum. The combination of a tight prior ($\sigma_b/|x_b| = 0.05$) and a wide observation tolerance ($\sigma_{\mathrm{obs}}/y = 0.20$) produced the largest RMSE (3406 kg/ha), indicating that excessive shrinkage toward the background prevents the model from exploiting the available observational information.

The decomposition of the cost functional reveals that the observation term J_o exceeds the background term J_b by two to three orders of magnitude in all configurations. With 15–25 MODIS LAI observations per field constraining four control parameters, the information content of the observations dominates the prior, and

2026, **27** (3), 470–488. doi 10.26089/NumMet.v27r330

the 4D-Var problem is effectively data-driven. The mean relative parameter adjustment $\overline{|\Delta x|/|x_b|}$ remained within the narrow range 0.11–0.12 across the entire grid, further confirming that the optimizer converges to the same neighbourhood of the parameter space regardless of the strength of the prior. The absolute value of J_o ($\sim 10^3$) substantially exceeds the χ-squared expectation $N_{\text{obs}}/2 \sim 10$, which indicates a residual structural mismatch between WOFOST and MODIS LAI that cannot be eliminated by parameter calibration alone.

A systematic positive yield bias of approximately 2000 kg/ha was observed in every configuration, including the most permissive ones. This invariance demonstrates that the over-estimation of yield is structural in nature and is not attributable to the specification of $\boldsymbol{B}$ or $\boldsymbol{R}$. Possible contributing factors include the use of the water-limited production mode, which omits nitrogen, pest, and disease limitations and the indirect link between the assimilated variable (LAI) and the harvest-index parameters that control the partitioning of biomass into storage organs.

3.7. Computational considerations. Table 2 summarizes the per-field cost of the three assimilation methods. The methods differ by roughly an order of magnitude in both the number of WOFOST evaluations and the resulting wall-clock time, with EnKF being the cheapest in both metrics.

Two practical considerations qualify the raw cost comparison. First, the methods parallelize differently. The EnKF is embarrassingly parallel at the ensemble level: members are independent and scale near-linearly with available cores. DE parallelizes within each generation but synchronizes between them, as reflected in the 6× gap between wall and CPU time in Table 2. 4D-Var is essentially sequential, since each iteration depends on the previous gradient, which explains why its wall time exceeds that of DE despite requiring fewer evaluations. Second, EnKF returns a full ensemble of state and parameter trajectories and therefore provides yield uncertainty (e.g. mean ± standard deviation) at no additional cost, whereas DE and 4D-Var produce point estimates that require bootstrap resampling or Hessian inversion to obtain comparable uncertainty bounds.

Table 2. Computational cost of the three assimilation methods per field (mean ± standard deviation). DE was parallelized across 6 CPU workers via Python `multiprocessing` and EnKF and 4D-Var were run single-threaded

Method	WOFOST evaluations	Wall time, s	CPU time, s
DE	1333 ± 258	68 ± 12	≈ 409
EnKF	100	32 ± 6	32 ± 6
4D-Var	367 ± 146	83 ± 34	83 ± 34

4. Conclusion. In this study, we presented a systematic comparison of three data assimilation methods (Differential Evolution, Ensemble Kalman Filter, and 4D-Var) for integrating MODIS LAI observations into the WOFOST crop simulation model to predict winter wheat yield in the central region of Russia. All three methods were implemented within a unified computational pipeline using identical input data, model configuration, and evaluation metrics.

The results demonstrate that all three assimilation approaches consistently improve both the LAI trajectory reproduction and the yield prediction accuracy compared to the uncalibrated baseline simulation. The EnKF achieves the lowest mean LAI RMSE and the smallest absolute yield bias, benefiting from sequential updates applied to multiple model state variables simultaneously. The 4D-Var method attains a yield RMSE virtually identical to EnKF offering a deterministic analysis with regularization through the background error covariance that produces physically consistent parameter adjustments and the most stable error distribution across the study area. The DE optimization, despite producing a competitive LAI RMSE gives the weakest yield prediction, as unconstrained parameter search optimizes the LAI cost function at the expense of physically consistent biomass partitioning.

From a practical perspective, the choice of assimilation method depends on the specific requirements of the application. When accurate LAI reconstruction is the primary objective (e.g., for monitoring vegetation stress), the DE approach provides a straightforward global optimization with competitive LAI performance, though its yield prediction reliability is limited. When probabilistic yield forecasts with uncertainty bounds are required, the EnKF is the natural choice, offering the lowest bias, the highest field-level improvement rate, and built-in uncertainty quantification through ensemble spread. When a balanced optimization of both LAI and yield is desired with physically interpretable and stable parameter adjustments, the 4D-Var method offers a compelling alternative with the most consistent error distribution across the study area.

Furthermore, for winter wheat crop, data assimilation may be restricted to the second half of the growing season, as LAI observations during the first half are of limited informational value due to persistent snow cover. Another potential improvement involves the joint assimilation of LAI and soil moisture data, the latter of which can be indirectly estimated from satellite remote sensing or derived from hydrological models such as HYDRUS.

A further improvement could be achieved by directly incorporating pixel-wise uncertainty estimates provided with MODIS LAI products. These uncertainties are derived internally during the retrieval process using radiative transfer model lookup tables and quality assessment layers, and therefore may better reflect the actual retrieval confidence under varying atmospheric and surface conditions. Using such physically-based error estimates could improve the specification of the observation error covariance and enhance the robustness of the assimilation results.

Several limitations of the present study should be noted. First, the observation error statistics for MODIS LAI were estimated empirically as a fixed fraction of the observed value, which may not fully capture the spatial and temporal variability of retrieval errors. Second, the background error covariance in 4D-Var was specified as diagonal, neglecting potential correlations among parameters. Third, the study was limited to a single growing season and a specific set of fields, which may limit the generalizability of the relative method performance. We plan to address these limitations in future work by incorporating multi-year datasets, exploring adaptive observation error models, and testing the methods on additional crop types and regions.

Data availability. The source code required to reproduce all figures and tables of this study are publicly available at `https://github.com/mishagrol/CropModelAssimilation`. The MODIS MCD15A3H LAI, NASA POWER weather data, and SoilGrids soil texture data used as model inputs are freely available from their respective providers. The raw per-field observed grain yields cannot be redistributed due to the confidentiality terms of the data-sharing agreement with the data provider. In the preparation of this paper, LLM was employed to assist with improving the clarity, grammar, and overall quality of the English language. Following its use, the authors thoroughly reviewed and edited the content and independently verified all facts to ensure accuracy and alignment with the intended meaning. The tool was not used to generate scientific content, interpret data, or make intellectual contributions. The responsibility for the scientific content and conclusions of this paper lies entirely with the authors.

References

1. C. Rosenzweig, J. W. Jones, J. L. Hatfield, et al., “The Agricultural Model Intercomparison and Improvement Project (AgMIP): Protocols and pilot studies,” Agric. For. Meteorol. **170**, 166–182 (2013). doi 10.1016/j.agrformet.2012.09.011.
2. S. Wolfert, L. Ge, C. Verdouw, and M.-J. Bogaardt, “Big Data in Smart Farming — a review,” Agric. Syst. **153**, 69–80 (2017). doi 10.1016/j.agsy.2017.01.023.
3. F. Schierhorn, D. Müller, T. Beringer, et al., “Post-Soviet cropland abandonment and carbon sequestration in European Russia, Ukraine, and Belarus,” Glob. Biogeochem. Cycles **27** (4), 1175–1185 (2013). doi 10.1002/2013GB004654.
4. J. W. Jones, G. Hoogenboom, C. H. Porter, et al., “The DSSAT cropping system model,” Eur. J. Agron. **18** (3–4), 235–265 (2003). doi 10.1016/S1161-0301(02)00107-7.
5. D. P. Holzworth, N. I. Huth, P. G. deVoil, et al., “APSIM — Evolution towards a new generation of agricultural systems simulation,” Environ. Model. Softw. **62**, 327–350 (2014). doi 10.1016/j.envsoft.2014.07.009.
6. A. de Wit, H. Boogaard, D. Fumagalli, et al., “25 years of the WOFOST cropping systems model,” Agric. Syst. **168**, 154–167 (2019). doi 10.1016/j.agsy.2018.06.018.
7. C. Nendel, M. Berg, K. C. Kersebaum, et al., “The MONICA model: Testing predictability for crop growth, soil moisture and nitrogen dynamics,” Ecol. Model. **222** (9), 1614–1625 (2011). doi 10.1016/j.ecolmodel.2011.02.018.
8. H. Boogaard, J. Wolf, I. Supit, et al., “A regional implementation of WOFOST for calculating yield gaps of autumn-sown wheat across the European Union,” Field Crops Res. **143**, 130–142 (2013). doi 10.1016/j.fcr.2012.11.005.
9. L. Poggio, L. M. de Sousa, N. H. Batjes, et al., “SoilGrids 2.0: producing soil information for the globe with quantified spatial uncertainty,” Soil **7** (1), 217–240 (2021). doi 10.5194/soil-7-217-2021.

10. W. A. Dorigo, R. Zurita-Milla, A. J. W. de Wit, et al., "A review on reflective remote sensing and data assimilation techniques for enhanced agroecosystem modeling," Int. J. Appl. Earth Obs. Geoinf. **9** (2), 165–193 (2007). doi 10.1016/j.jag.2006.05.003.

11. J. Huang, J. L. Gómez-Dans, H. Huang, et al., "Assimilation of remote sensing into crop growth models: Current status and perspectives," Agric. For. Meteorol. **276–277**, Article Number 107609 (2019). doi 10.1016/j.agrformet.2019.06.008.

12. R. B. Myneni, S. Hoffman, Y. Knyazikhin, et al., "Global products of vegetation leaf area and fraction absorbed PAR from year one of MODIS data," Remote Sens. Environ. **83** (1–2), 214–231 (2002). doi 10.1016/S0034-4257(02)00074-3.

13. X. Jin, L. Kumar, Z. Li, et al., "A review of data assimilation of remote sensing and crop models," Eur. J. Agron. **92**, 141–152 (2018). doi 10.1016/j.eja.2017.11.002.

14. V. Badenko, D. Eremenko, A. Topaj, and M. Gasanov, "A Method for Application of Remote Sensing Data in Crop Simulation Models," in *Proc. XV Int. Sci. Conf. Agricultural Machinery Industry "Interagromash 2022"*. Lecture Notes in Networks and Systems. Vol. 574. (Springer, Cham, 2022), pp. 1596–1604. doi 10.1007/978-3-031-21432-5_171.

15. R. Storn and K. Price, "Differential Evolution — A Simple and Efficient Heuristic for Global Optimization over Continuous Spaces," J. Glob. Optim. **11** (4), 341–359 (1997). doi 10.1023/A:1008202821328.

16. G. Evensen, "The Ensemble Kalman Filter: theoretical formulation and practical implementation," Ocean Dyn. **53** (4), 343–367 (2003). doi 10.1007/s10236-003-0036-9.

17. G. S. Nearing, W. T. Crow, K. R. Thorp, et al., "Assimilating remote sensing observations of leaf area index and soil moisture for wheat yield estimates: An observing system simulation experiment," Water Resour. Res. **48** (5), Article Number W05525 (2012). doi 10.1029/2011WR011420.

18. Y. Li, Q. Zhou, J. Zhou, et al., "Assimilating remote sensing information into a coupled hydrology–crop growth model to estimate regional maize yield in arid regions," Ecol. Model. **291**, 15–27 (2014). doi 10.1016/j.ecolmodel.2014.07.013.

19. W. Zhuo, H. Huang, X. Gao, et al., "An Improved Approach of Winter Wheat Yield Estimation by Jointly Assimilating Remotely Sensed Leaf Area Index and Soil Moisture into the WOFOST Model," Remote Sens. **15** (7), Article Number 1825 (2023). doi 10.3390/rs15071825.

20. P. Courtier, J.-N. Thépaut, and A. Hollingsworth, "A strategy for operational implementation of 4D-Var, using an incremental approach," Q. J. R. Meteorol. Soc. **120** (519), 1367–1387 (1994). doi 10.1002/qj.49712051912.

21. V. P. Shutyaev and E. I. Parmuzin, "Some algorithms for studying solution sensitivity in the problem of variational assimilation of observation data for a model of ocean thermodynamics," Russ. J. Numer. Anal. Math. Model. **24** (2), 145–160 (2009). doi 10.1515/RJNAMM.2009.010.

22. S. K. Shangareeva, V. M. Stepanenko, G. M. Faykin, et al., "Variational Data Assimilation in the Constructor of Dynamic Soil Carbon Models," Supercomput. Front. Innov. **12** (4), 88–100 (2025). doi 10.14529/jsfi250406.

23. A. H. Sparks, "nasapower: A NASA POWER Global Meteorology, Surface Solar Energy and Climatology Data Client for R," J. Open Source Softw. **3** (30), Article Number 1035 (2018). doi 10.21105/joss.01035.

24. Y. Zhang and M. G. Schaap, "Weighted Recalibration of the Rosetta Pedotransfer Model with Improved Estimates of Hydraulic Parameter Distributions and Summary Statistics (Rosetta3)," J. Hydrol. **547**, 39–53 (2017). doi 10.1016/j.jhydrol.2017.01.004.

25. N. Gorelick, M. Hancher, M. Dixon, et al., "Google Earth Engine: Planetary-scale geospatial analysis for everyone," Remote Sens. Environ. **202**, 18–27 (2017). doi 10.1016/j.rse.2017.06.031.

26. A. Savitzky and M. J. E. Golay, "Smoothing and Differentiation of Data by Simplified Least Squares Procedures," Anal. Chem. **36** (8), 1627–1639 (1964). doi 10.1021/ac60214a047.

27. I. M. Sobol', "Global sensitivity indices for nonlinear mathematical models and their Monte Carlo estimates," Math. Comput. Simul. **55** (1–3), 271–280 (2001). doi 10.1016/S0378-4754(00)00270-6.

28. I. M. Sobol', "On Sensitivity Estimation for Nonlinear Mathematical Models," Matem. Modelirovanie **2** (1), 112–118 (1990).

29. M. Gasanov, A. Petrovskaia, A. Nikitin, et al., "Sensitivity Analysis of Soil Parameters in Crop Model Supported with High-Throughput Computing," in *Proc. Int. Conf. Computational Science (ICCS 2020)*. Lecture Notes in Computer Science. Vol 12143. (Springer, Cham, 2020), pp. 731–741. doi 10.1007/978-3-030-50436-6_54.

30. J. Herman and W. Usher, "SALib: An open-source Python library for Sensitivity Analysis," J. Open Source Softw. **2** (9), Article Number 97 (2017). doi 10.21105/joss.00097.

31. V. Shutyaev, V. Zalesny, V. Agoshkov, et al., "Four-Dimensional Variational Data Assimilation and Sensitivity of Ocean Model State Variables to Observation Errors," J. Mar. Sci. Eng. **11** (6), Article Number 1253 (2023). doi 10.3390/jmse11061253.

32. R. H. Byrd, P. Lu, J. Nocedal, and C. Zhu, "A Limited Memory Algorithm for Bound Constrained Optimization," SIAM J. Sci. Comput. **16** (5), 1190–1208 (1995). doi 10.1137/0916069.

33. P. Virtanen, R. Gommers, T. E. Oliphant, et al., "SciPy 1.0: fundamental algorithms for scientific computing in Python," Nat. Methods **17**, 261–272 (2020). doi 10.1038/s41592-019-0686-2.

Received	*Accepted*	*Published*
April 3, 2026	*June 30, 2026*	*August 4, 2026*

Information about the authors

Mikhail E. Gasanov — PhD Student; Lomonosov Moscow State University, Department of Soil Science, Leninskie Gory, 1, building 12, 119234, Moscow, Russia.

Anna Yu. Petrovskaia — Research Scientist; Skoltech, Artificial Intelligence center, Bolshoy Bulvar, 30, 121205, Moscow, Russia.

Sergey A. Matveev — Ph.D., Associate Professor; 1) Lomonosov Moscow State University, Department of Computational Mathematics and Cybernetics, Leninskie Gory, 1, 119991, Moscow, Russia; 2) Marchuk Institute of Numerical Mathematics of RAS, Gubkina ulitsa, 8, 119333, Moscow, Russia.

Ivan V. Oseledets — Dr. Sci., Professor; 1) Lomonosov Moscow State University, Department of Artificial Intelligence, Leninskie Gory, 1, 119991, Moscow, Russia; 2) Artificial Intelligence Research Institute (AIRI), Gubkina ulitsa, 8, 105064, Moscow, Russia; 3) Skoltech, Artificial Intelligence center, Bolshoy Bulvar, 30, 121205, Moscow, Russia.

Appendix 1

Selection of the finite-difference step and sensitivity analysis of 4D-Var to background and observation covariances

The choice of the finite-difference step ε for the gradient approximation has a substantial impact on the 4D-Var solution because the WOFOST integration is piecewise-smooth. As a result, sufficiently small perturbations of the control parameters turn out to be lower than the numerical resolution of the model and lead to a vanishing finite-difference gradient. To reasonably select ε, the 4D-Var assimilation was repeated on five representative fields for $\varepsilon \in \{1.5 \times 10^{-8},\ 10^{-6},\ 10^{-4},\ 10^{-3},\ 10^{-2}\}$. Results presented in Table A1.

The 4D-Var assimilation was repeated on five representative fields, selected to span the yield quantiles of the dataset, for each combination of relative background uncertainty $\sigma_b/|x_b| \in \{0.05,\ 0.10,\ 0.15,\ 0.25,\ 0.50\}$

Table A1. Sensitivity of the 4D-Var solution to the relative step ε of the forward finite-difference gradient approximation: yield RMSE and bias, decomposition of the cost functional, mean relative parameter adjustment, and mean number of L-BFGS-B iterations, aggregated over 5 representative fields.

ε	Yield RMSE, kg/ha	Yield bias, kg/ha	$\overline{J_b}$	$\overline{J_o}$	$\overline{\lvert\Delta x\rvert/\lvert x_b\rvert}$	$\overline{n_{\text{iter}}}$
1.5×10^{-8}	3178	2001	2.68	1101.12	0.114	5.8
1.0×10^{-6}	3178	2001	2.68	1101.12	0.114	5.8
1.0×10^{-4}	3178	2001	2.68	1101.12	0.114	6.2
$\mathbf{1.0 \times 10^{-3}}$	**2580**	**1409**	**3.00**	**1139.82**	**0.136**	**7.2**
1.0×10^{-2}	2620	1738	2.78	1134.59	0.126	6.0

2026, **27** (3), 470–488. doi 10.26089/NumMet.v27r330

Table A2. Sensitivity of the 4D-Var solution to the specification of $\boldsymbol{B}$ and $\boldsymbol{R}$: yield RMSE, decomposition of the cost functional, and mean relative parameter adjustment, aggregated over five representative fields

$\sigma_b/\lvert x_b\rvert$	σ_{obs}/y	Yield RMSE, kg/ha	$\overline{J_b}$	$\overline{J_o}$	$\overline{\lvert\Delta x\rvert/\lvert x_b\rvert}$
0.05	0.05	3214	22.57	1557.11	0.113
0.05	0.10	3245	22.90	1101.47	0.110
0.05	0.20	3406	22.62	790.24	0.110
0.10	0.05	3196	5.91	1556.40	0.116
0.10	0.10	3178	6.03	1100.62	0.113
0.10	0.20	3295	6.25	788.82	0.118
0.15	0.05	3195	2.65	1556.36	0.117
0.15	0.10	3178	2.68	1100.62	0.114
0.15	0.20	3241	2.89	788.64	0.121
0.25	0.05	3194	0.96	1556.34	0.118
0.25	0.10	3177	0.97	1100.62	0.114
0.25	0.20	3206	1.07	788.59	0.122
0.50	0.05	3193	0.24	1556.34	0.118
0.50	0.10	3177	0.24	1100.62	0.114
0.50	0.20	3206	0.27	788.59	0.122

and relative observation uncertainty $\sigma_{\text{obs}}/y \in \{0.05,\ 0.10,\ 0.20\}$. The aggregated results are summarized in Table A2. For each grid point, the yield prediction RMSE, the means of J_b, J_o, and the relative parameter adjustment $\overline{|\Delta x|/|x_b|}$ are reported (averaged over the five fields). All other parameters of the 4D-Var procedure ($\pm 30\%$ parameter bounds, L-BFGS-B with forward finite-difference gradient, $10^{-6}\times$ function tolerance, 100 maximum iterations) were held fixed at the values used in the main experiments.

Three structural features of the response surface are worth highlighting. First, the relative variation of yield RMSE over the full grid is $\pm 3.5\%$ around the ensemble mean of 3225 kg/ha, and for $\sigma_b/|x_b| \geqslant 0.10$ and $\sigma_{\text{obs}}/y \leqslant 0.10$ the RMSE varies by less than 20 kg/ha. This identifies a stable plateau that contains the operating point used in the main results. Second, the observation term J_o exceeds the background term J_b by two to three orders of magnitude in every configuration, which indicates that the 4D-Var problem is data-dominated for the present density of MODIS LAI observations and that the prior acts essentially as a soft physical-plausibility constraint rather than as an active regularizer. Third, the mean relative parameter adjustment $\overline{|\Delta x|/|x_b|}$ is essentially independent of the covariance specification, confirming that the optimizer converges to the same neighbourhood of the parameter space irrespective of the strength of the prior. The systematic positive yield bias persisting across all 15 configurations is therefore attributable to structural limitations of the model formulation rather than to the choice of the covariance hyperparameters.