UDC 330.115

T. ERMOLIEVA
International Institute for Applied Systems Analysis, Laxenburg, Austria, e-mail: *ermol@iiasa.ac.at*.

P. HAVLÍK
International Institute for Applied Systems Analysis, Laxenburg, Austria, e-mail: *havlik.petr@gmail.com*.

A. LESSA-DERCI-AUGUSTYNCZIK
International Institute for Applied Systems Analysis, Laxenburg, Austria, e-mail: *augustynczik@iiasa.ac.at*.

E. BOERE
International Institute for Applied Systems Analysis, Laxenburg, Austria, e-mail: *boere@iiasa.ac.at.*

S. FRANK
International Institute for Applied Systems Analysis, Laxenburg, Austria, e-mail: *frank@iiasa.ac.at*.

T. KAHIL
International Institute for Applied Systems Analysis, Laxenburg, Austria, e-mail: *kahil@iiasa.ac.at*.

G. WANG
China Agricultural University (CAU), Beijing, China, e-mail: *gangwang@cau.edu.cn*.

J. BALKOVIČ
International Institute for Applied Systems Analysis, Laxenburg, Austria, e-mail: *balkovic@iiasa.ac.at*.

R. SKALSKÝ
International Institute for Applied Systems Analysis, Laxenburg, Austria, e-mail: *skalsky@iiasa.ac.at*.

C. FOLBERTH
International Institute for Applied Systems Analysis, Laxenburg, Austria, e-mail: *folberth@iiasa.ac.at*.

N. KOMENDANTOVA
International Institute for Applied Systems Analysis, Laxenburg, Austria, e-mail: *komendan@iiasa.ac.at*.

P.S. KNOPOV
V.M. Glushkov Institute of Cybernetics of the NAS of Ukraine, Kyiv, Ukraine, e-mail: *knopov1@yahoo.com, knopov1@gmail.com*.

A NOVEL ROBUST META-MODEL FRAMEWORK FOR PREDICTING CROP YIELD PROBABILITY DISTRIBUTIONS USING MULTISOURCE DATA[1]

Abstract. There is an urgent need to better understand and predict crop yield responses to weather disturbances, in particular, of extreme nature, such as heavy precipitation events, droughts, and heat waves, to improve future crop production projections under weather variability, extreme events, and climate change. In this paper, we develop quantile regression models for estimating crop yield probability distributions depending on monthly temperature and precipitation values and soil quality characteristics, which can be made available for different climate change projections. Crop yields, historical and those simulated by the EPIC model, are analyzed and distinguished according to their levels, i.e., mean and critical quantiles. Then, the

[1]The development of the robust quantile-based meta-model for predicting probability distributions of crop yields depending on climate scenarios, temperature and precipitation patters, and soil characteristics is supported by the joint project between the International Institute for Applied Systems Analysis (IIASA) and National Academy of Sciences of Ukraine (NASU) on "Integrated robust modeling and management of food-energy-water-land use nexus for sustainable development". The paper contributes to EU PARATUS (CL3-2021-DRS-01-03, SEP-210784020) project on "Promoting disaster preparedness and resilience by co-developing stakeholder support tools for managing systemic risk of compounding disasters".

crop yield quantiles are approximated by fitting separate quantile-based regression models. The developed statistical crop yield meta-model enables the analysis of crop yields and respective probabilities of their occurrence as a function of the exogenous parameters such as temperature and precipitation and endogenous, in general, decision-dependent parameters (such as soil characteristics), which can be altered by land use practices. Statistical and machine learning models can be used as reduced form scenario generators (meta-models) of stochastic events (scenarios), as a submodel of more complex models, e.g., Integrated Assessment model (IAM) GLOBIOM.

Keywords: extreme events, climate change, food security, crop yields projections, probability distributions, quantile regressions, robust estimation and machine learning, two-stage STO.

INTRODUCTION

Extreme weather conditions, such as high temperature, heat waves, declining rainfall, prolonged heavy precipitation, can lead to gradual crop yield declines or even abrupt crop production shocks and, therefore, to profit losses, threats to food security and market volatility. Due to climate change, frequency and severity of extreme events is projected to increase [1–6].

There is a need to better understand crop yield responses to weather disturbances, especially, of extreme nature, and relevant land use and soil management practices, i.e., conservation tillage to improve crop yields towards desirable production levels. Crop yield prediction is very challenging in the face of extreme high-impact events. Such events can happen rarely but cause major production losses and require high ex-post adaptation costs and expensive actions. Therefore, gaining earlier information about the likelihood of critical yields and the accounting for the extreme rare-high impact events is essential for farmers, crop insurers, agricultural market managers, for taking informative and timely strategic and operational (long- and short-term) ex-ante and ex-post mitigation and adaptation economic and management decisions.

The two main approaches widely used to assess the impacts of weather, climate change, and soil parameters on crop yields are:

a) process-based simulation models, e.g., EPIC [5–8], which attempt to represent key dynamic processes affecting crop yields, and

b) statistical models, which estimate functional relationships between historical observations of weather, soil characteristics and yields [9–12].

In this paper we develop a hybrid meta-model for generating stochastic crop-yield scenarios based on historical data and on input-outputs of a dynamic process-based crop yield simulation model EPIC [5, 6]. The linkage of the statistical and the process-based approaches into a meta-model for crop yields simulation combines the strengths and avoids the limitations of statistical and bio-physical models if they are used separately.

The crop yields can be distinguished according to their levels, i.e., critical quantiles (5th, 25th, 50s, 75th, and 95th) as well as mean values. Often, the critical quantiles (percentiles) can be identified by the so-called break-even analysis of the crop producer. In this case, the “break-even” yield balances out crop production costs and profits from crops sails at given crop price, which is important for establishing adequate premium and coverage policies of agricultural insurance [13–16].

The meta-model developed in this paper aims at analyzing and predicting probabilities of critical crop yield levels (scenarios) based on monthly temperature and precipitation values, different climate change scenarios (i.e., RCPs), and on soil quality characteristics, which can be altered by soil and water management practices and by weather conditions. In particular, we introduce a quantile-based regression model for the prediction of crop yield percentiles using historical data and inputs/outputs of the EPIC model. The developed crop-yield meta model can be used to generate stochastic crop yield scenarios to be included into more complex Integrated Assessment Models (IAMs), e.g., (stochastic) GLOBIOM model [17–20].

Large-scale global IAMs cannot effort direct integration/linkage of more advanced generators of stochastic events. In this case, designing a reduced form generator of pseudo-events based on a trained statistical or/and machine learning model (e.g., a neural network) offers a simple and practical solution.

The remainder of the manuscript is organized as follows. Section 1 discusses motivations for developing a robust quantile-based meta-model for estimating crop yields probability distributions, which can be non-normal, non-symmetrical, skewed and with heavy tails due to extreme weather conditions in combination with soil properties. Section 2 presents a short overview of the two main approaches, process-based simulation models and statistical models, to analyze the impacts of weather variability, climate change, soil parameters and practices on crop yields. Section 3 outlines robust statistical and machine learning approaches to predict crop yield probability distributions. Data and selected results of the studies are presented in Section 4 and Conclusions summarize main conclusions and directions of further studies.

1. CROP YIELD ANALYSIS

1.1. Non-normal probability distributions of stochastic parameters and percentile-based criteria functions. Uncertainties and risks are inherent in agricultural practices.

Stochastic variables relevant to agricultural activities, e.g., crop yields, weather parameters, costs, can be characterized by probability distributions (parametric or nonparametric) functions or by probabilistic scenarios. A probability distribution can be non-normal, heavy tailed and even multimodal.

Low yields can cause imbalances in grain supply-demand chains and thus lead to prices increase, market disturbances, trade bans, affecting food security. It is essential that such events are accounted for in planning models, which inform decision-makers, to design adequate mitigation and recovery policies to ensure policy targets, such as food security, remain feasible in the face of stochasticity. A first step towards this endogenization of risk in planning and economic models is the development of tools capable to reproduce such events in a robust manner, e.g., crop yield meta models and emulators.

1.2. Two-stage stochastic optimization model for robust statistical estimation and decision-making. Statistical and machine learning problems for analyzing and predicting agricultural variables, including estimation of crop yields probability distributions, can be formulated as stochastic optimization (STO) problems [19–22] optimizing some goodness of fit criteria. If stochastic parameters are non-normal with (heavy) tails, The Mean-Variance, Ordinary Least Square (OLS) or Root Mean Squared Error (RMSE)-based estimates can be misleading, and the effects can be different for different subsets of data sample as well as the estimates have to be revised when additional information about stochastic parameters arrive. For statistical estimation and machine learning problems in the presence of non-normal and possibly multimodal probability distributions it is more natural to use the median or other quantile-based criteria instead of the mathematical expectation.

Because of rare events (“observations” in the tails of probability distributions) and short periods of available historical observations, the data for statistical estimation and machine learning can be only partial or incomplete, i.e., incomplete “learning”. The ex-ante parameter estimation in the conditions of incomplete “learning” may require subsequent revisions and corrections after additional information about the uncertain parameter realization becomes known (i.e., after “learning” or partial “learning” of uncertain parameters values). Which means, that the ex-ante estimates can incur the costs for their correction, revision, or reversion. The estimates based on Mean-Variance, Ordinary Least Square (OLS) or Root Mean Squared Error (RMSE) can require considerable revisions, while quantile-based criteria provide robust estimates not very sensitive to newly acquired observations information.

Thus, the solutions of the optimization problem in the face of uncertainty are of the two types (two-stage) [18–20, 22–24]. The ex-ante solutions x correspond

to decisions taken in front of uncertainties before complete or partial observations of uncertain parameters ω may become available, whereas the ex-post solutions y describe all future actions to be taken in different time periods in response to the environment created by the chosen ex-ante x and the newly observed value of the uncertain parameter ω in that specific time period.

1.3. Basic model of a two-stage parameter estimation and decision-making. Let us illustrate the concept of the two-stage decision making and robust statistical estimation problem with an example of a simplest two-stage STO model. Assume, a random variable ω represents some stochastic parameter, for example, uncertain demand, resource availability (water, energy, land), a crop yield, a weather parameter (temperature, precipitation, solar radiation). The choice of the decision variable x, $\bar{x} \geq x \geq 0$, such that x is as close, in some sense, to stochastic ω as possible ($x \sim \omega$), can be associated with a function $f(x, \omega)$ reflecting costs of overestimation and underestimation of ω. In this case, it makes sense to introduce two types of decision variables (x, y). The variable x to meet ω can be written as $x = \omega - y^+(\omega) + y^-(\omega)$, where $y^+(\omega) = \max\{0, x - \omega\}$ and $y^-(\omega) = \max\{0, \omega - x\}$. Decision variable x defines the first-stage decisions, while decision variables $y(\omega) = (y^+(\omega), y^-(\omega))$ correspond to the second-stage correction decisions dependent on the scenario (realization, observation) of ω. Then, the function $f(x, \omega)$ is a random piecewise linear function $f(x, \omega) = \max\{\alpha y^+(\omega), \beta y^-(\omega)\} = \max\{\alpha(x - \omega), \beta(\omega - x)\}$, where α is the cost for overestimation/surplus (e.g., the cost for storage) and β is the cost for underestimation/shortage (e.g., the cost for imports).

The problem is to find the level that is "optimal", in a sense, for all foreseeable random scenarios/observations ω. The expected cost criterion leads to the minimization of the following nonsmooth function:

$$F(x) = E(x, \omega) = E \max\{\alpha(x - \omega), \beta(\omega - x)\} \tag{1}$$

subject to $\bar{x} \geq x \geq 0$ for a given upper bound x. This stochastic minimax problem is a two-stage stochastic programming with two types of decisions (x, $y(\omega)$) [24].

The STO problem (1) can be solved using an iterative solution procedure based on the stochastic quasigradient (SQG) methods [23–27]. Function $F(x)$ is convex, therefore, the SQG algorithm can be defined as the following discontinuous adaptive machine learning process:

$$x^{k+1} = \min\{\max\{0, x^k - \rho_k \xi^k\}, \bar{x}\}, \quad k = 0, 1, 2, \ldots,$$

where $\xi^k = \alpha$, if the current level of x^k exceeds the observed ω and $\xi^k = -\beta$ otherwise, ρ_k is a positive step size.

The SQG method is a convergent with probability 1 sequential estimation/solution procedure, which can be viewed as an adaptive machine learning process able to learn the optimal level x^* through sequential adjustments of its current levels $x^0, x^1, x^2, \ldots, x^k, \ldots$ to observable (or simulated) scenarios/observations $\omega^0, \omega^1, \omega^2, \ldots, \omega^k, \ldots$. It can be used as a general statistical and machine learning algorithm when addition observations of uncertain parameter ω can arrive and the solution x has to be revised accordingly. The use of quantile-based criteria leads to nonsmooth optimization problems deriving robust solutions (models), which do not deteriorate too much because of estimating and training with new or slightly different data.

Minimization of the goal function (1) provides quantile-type characteristics of the optimal solutions [15, 16, 18–22, 28]. For example, if the distribution of ω has a density, α, $\beta > 0$, then the optimal solution x minimizing $F(x)$ is the quantile determined as

$$Pr\{\omega \leq x\} = \frac{\beta}{\alpha + \beta}. \tag{2}$$

1.4. General quantile-based regression model. Section 1.3 presents the simplest quantile-based two-stage STO model for robust estimation and decision making. Let us formulate a general two-stage stochastic optimization problem representing the quantile-based regression. In traditional regression estimation problem, the conditional expectation provides a satisfactory representation of stochastic dependencies when they are well approximated by two first moments, e.g., for normal distributions. Assume that a random function $u(v)$ for each element v from a set V corresponds a random element $u(v)$ of the set U. Assume that $V \subset R^l$, $U \subset R^1$. Let P is a joint probability measure on the pairs $\theta = (u, v)$ of random variables u and v. The regression function is defined as the conditional mathematical expectation

$$r(v) = E(u \mid v) = \int uP(U \mid v). \tag{3}$$

It is easy to see that $r(v)$ minimizes the functional (providing it is well defined)

$$F(x(v)) = E(u(v) - x(v))^2, \tag{4}$$

where $Eu^2(v) < \infty$, $Ex^2(v) < \infty$.

For general (possibly, multimodal) distributions it is more natural to use the median or other quantiles instead of the expectation. Let us define the quantile regression function $r_\rho(v)$ as the maximal value y satisfying equation

$$P(u(v) \geq y \mid v) = \rho(v), \tag{5}$$

where $0 < \rho(v) < 1$. It can be shown that quantile regression function $r_\rho(v)$ minimizes the functional

$$F(x(v)) = E(\rho(v)x(v) + \max\{0, u(v) - x(v)\}). \tag{6}$$

The minimization of a more general functional (7) is similar to problem (1), where $a(v)$, $\alpha(v)$, $\beta(v)$ represent possible costs for observations v associated with over or underestimation of regression dependencies:

$$F(x(v)) = E(a(v)x(v) + \max\{\alpha(v)(u(v) - x(v)), \beta(v)(x(v) - u(v))\}). \tag{7}$$

The costs or prices $a(v)$, $\alpha(v)$, $\beta(v)$ can be a results of another, e.g., a general or partial equilibrium model. Thus, the robust estimation of $x(\cdot)$ can be considered as a part of a decision-making process when estimation model is integrated into a decision-making model. Minimization of (7) is reduced to the minimization of (6) with $\rho(v) = (a + \beta)(\alpha + \beta)^{-1}$. The median corresponds to the case when $a \equiv 0$, $\alpha = \beta$. The existence of optimal solution requires $a < \alpha$.

Assuming $r_\rho(v)$ in (6) is a convex function for all $v \in V$, $F(x)$ is also a convex function. If $r_\rho(v)$ is a linear function for all $v \in V$, then $F(x)$ can be minimized by linear programming methods.

2. BIO-PHYSICAL AND STATISTICAL CROP YIELD MODELS

In this paper we develop a robust meta-model based on quantile regression for generating stochastic crop-yield scenarios using historical data and input–outputs of a dynamic process-based crop yield simulation model EPIC. The development of the meta-model contributes to the research on the proper tools for linking reduced form stochastic scenario generators (e.g., of crop yields) with large scale land use and agriculture planning models, e.g., GLOBIOM. The linkage of the statistical and the process-based approaches into a hybrid meta-model for crop yields simulation combines the strengths and avoid the limitations of statistical and bio-physical models if they are applied separately. For example, process-based models may fail to properly account for extreme weather conditions, while there may be no sufficient data for statistical models reflecting crop yield responses to various soil and water management practices. Let us shortly introduce the statistical and the process-based models, in particular, the EPIC model.

2.1. Dynamic process-based crop yield simulation models. Dynamic process-based crop yield simulation models were developed primarily to simulate the growth of crops, along with the associated phenomena that influence crop growth such as temperature, water and solute movement in soils. The models build upon the physiological laws and understanding of plant and soil processes, to develop biologically meaningful equations, to predict crop yield and other phenotypes (Biophysical models in land evaluation). The models can provide explicit though often simplistic explanations of the interactions between crops yields, weather parameters, soil quality, environmental conditions in different phases of the crop growth cycle. Dynamic simulation models have been used in global gridded crop yield assessment studies (Global gridded crop models, GGCMs) to evaluate how climate change and agricultural management might impact crop yields [11, 29]. To enable model intercomparison the Agricultural Model Intercomparison and Improvement Project (AgMIP) and the Inter-Sectoral Impact Model Intercomparison Project (ISIMIP) provide common modelling protocols describing a harmonized simulation setup. From the ISIMIP conclusions, the bio-physical crop models unfortunately often are not capable to correctly model the crop response to changing weather parameters, thus, they often underestimated yield decline, and underestimation can be strong in the conditions of droughts and heat waves.

2.1.1. EPIC model. Dynamic process-based crop yield simulation model EPIC is one of the models involved in AgMIP and ISIMIP studies. Initially, it was developed by the USDA to assess how agricultural activities affect the status of US soil and water resources [7, 8, 30, 31]. The major components in EPIC are modules representing crop growth, yield and competition, weather simulation, hydrological, nutrient and carbon cycling, soil temperature and moisture, soil erosion, tillage, and plant environment control. Different soil and plant management options are available, including tillage operations, irrigation scheduling, fertilizer application rates and timing. The model offers options for simulating yields with different Potential Evapotranspiration (PET) equations, which allows sensitivity analysis of model's results with respect to alternative models' components. EPIC-IIASA runs the EPIC model for more than 120,000 spatial simulation units (SimUs) that are derived from intersecting soil and topography units, administrative borders and climate grids following a set of criteria for internal homogeneity of SimU. A large set of crop management scenarios (crop varieties, fertilization and irrigation options, and soil conservation practices) and climate change projections simulated for each SimU are a common workload for the EPIC-IIASA model.

In the present studies on the development of the crop yield meta-model we use Pan-European version of the EPIC-IIASA model with relevant data and results derived for EU countries [5, 6], which will be describe with more details in Section 4.

2.2. Statistical crop yield models. As an alternative to dynamic process-based crop yield simulation models, there are models that use observations of soil, weather, and other relevant parameters and crop yields to develop statistical models that functionally relate the former to the latter. These models become increasingly popular with the growing availability of data and various statistical and machine learning tools and software. The necessary data can come from experimental field measurements, farmer surveys, official government statistics, or some combination of these and other sources. The data may also come from process-based crop yield simulation models deriving, for example, crop yield feedbacks to possible land, water, soil managements, which have not yet been applied in practice and for which historical observations are not available.

The AgMIP studies [32] on the comparison of the process-based and statistical models at many experimental sites show satisfactory agreement between the two approaches. Statistical crop yield models can be helpful at various stages of experimental programs, to indicate what changes in the independent variables can

cause crop yields change toward the desirable outcomes. In this case, the covariates (independent variables) can be distinguished as exogenous (e.g., weather conditions) and endogenous (e.g., influenced by decisions). The later can be altered by land use practices, crops composition, water use, etc. Missing historical observations on responses of crop-yields to various soil practices under different climatic conditions motivates coupling of process-based simulation with statistical crop yield models.

2.2.1. The choice of covariates for a statistical crop yield model: temperature and precipitation parameters. Plant growth is highly dependent on precipitation and temperature. Temperature is a primary factor affecting the rate of plant development. Warmer temperatures expected with climate change and the potential for more extreme temperature events will impact plant productivity. Temperature effects are increased by water deficits and excess soil water demonstrating that understanding the interaction of temperature and water will be needed to develop more effective adaptation strategies to offset the impacts of greater temperature extreme events associated with a changing climate. Different regions of the world can benefit or suffer from the gradual climate change affecting seasonal and monthly patters of temperature and precipitation [33–35]. It is also expected that adverse effects could happen largely because of extreme climate events, such as droughts and heavy precipitation events, requiring probabilistic analysis of crop yields depending on weather parameters.

2.2.2. The choice of covariates for a statistical crop yield model: soil characteristics. The soil characteristics are very important variables in statistical models, for example, the relation between crop yields and soil organic carbon (SOC) in different soil layers. Carbon stored in the soil can help improve other soil physical properties such as infiltration rate, water-holding capacity, soil structure, and other physical properties. At the same time, carbon storage can improve the quality of soil nutrient pools and other chemical properties. Such soil practices as tillage and plant residues recycling affect SOC and soil structure, thereby influencing the physio-chemical and hydrothermal soil conditions which control the GHG fluxes and can impact GHG emission and soil carbon sequestration. Keeping plant residues is critical not only for soil nutrients, but also for soil protection from wind and water erosion [36].

2.3. Challenges in crop yield simulation modeling: data mining, harmonization, and downscaling. Crop yield simulation models are very data demanding. A constant challenge is the shortage of reliable data for model calibrations, validation, verification, and practical studies at the required spatio-temporal resolutions. In this case, various data mining, harmonization, and rescaling (e.g., downscaling) approaches are developed to fill the data gaps, recover unobservable local data and enable data harmonization from various sources and scales. Data (down and up) scaling and harmonization approaches can be considered as crop yields meta-models utilizing data and results of models at one resolution to derive results at different resolutions. For example, Folberth et al. [30] develops a model for crop yields analysis and rescaling (downscaling) based on extreme gradient boosting and random forests machine learning approaches.

The statistical quantile-based meta-model developed in this paper can operate at different resolutions and provides and effective means for scaling EPIC results.

3. ROBUST STATISTICAL ESTIMATION AND MACHINE LEARNING APPROACHES TO PREDICT CROP YIELD PROBABILITY DISTRIBUTIONS

Machine learning models have been often used for crop yield prediction, including random forest, neural networks, convolutional neural networks, recurrent neural networks, etc. However, due to the often black-box nature of these models, the tractability of the results is not straightforward. Also, the prediction accuracy is sensitive to model structure and parameter calibration, and it can be difficult to explain the accuracy or inaccuracy of the derived results.

3.1. Multivariate linear regression model. A (multiple) linear regression (MLR) can be considered as one of the machine learning algorithms, which is in fact one of the most popular models in machine learning. It is widely used because it is simple and tractable. The simplicity means it is easy to understand the responses of the dependent variables to each covariate, i.e., the regression coefficient of an independent variable reflects the change in the dependent variable as a result of a unit-change in the respective independent variables. On the other hand, the MLR assumes that the residuals are normally distributed, which means that it will fail to properly capture the effects of extreme conditions in the independent variables. It uses the method of least squares to calculate the conditional mean of the dependent variable across different values of the covariates. The linear regression model for calculating the mean takes the form

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \ldots + \beta_m x_{im}, \qquad (8)$$

where $i = 1, \ldots, n$ is a number of observations and m is a number of covariates. Coefficients of the MLR are found by minimizing the Mean Square Error "Goodness-of-Fit" function

$$MSE = (y_i - (\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_m x_{im}))^2,$$

which gives the "best regression line".

Crop yield distributions can be negatively skewed or even multimodal because of biological constraints that limit plants growth in response to various, often cumulative, combinations of weather parameters (precipitation, temperature, pressure, etc.), which have complex temporal and spatial patters. For statistical estimation and machine learning problems in the presence of non-normal, heavy tailed and possibly multimodal probability distributions it is more natural to use the median or other quantile-based criteria instead of the mathematical expectation.

3.2. Multivariate Quantile regression model. Unlike regular linear regression, the quantile regression estimates are more robust against outliers. For these studies, the conditional quantile functions are of a major interest also for investigating and predicting the probability distributions of dependent variables based on key factors such as temperature, precipitation, and soil characteristics.

Let us first introduce a notion of a quantile (percentile) function of a random variable. Quantiles are values that divide probability distribution of a random variable into a specific number of intervals (continuous) with equal probabilities. Assume that a random variable X has a continuous and strictly monotonic cumulative distribution function $F_X : R \to [0,1]$, $F_X(x) = P(X \le x)$. The p-quantile function of X, $Q_X(p)$, returns the value x such that $F_X(x) = Pr(X \le x) = p$, which can be rewritten as the inverse of the cumulative distribution function $Q(p) = F_X^{-1}(x) = \inf\{x : F_X(x) \ge p\}$.

For a random sample $X_1, X_2, \ldots, X_n, \ldots$ with empirical distribution function $\hat{F}_X(x)$, the pth empirical quantile function can be defined as $\hat{Q}(p) = \hat{F}_X^{-1}(x) =$ $= \inf\{x : \hat{F}_X(x) \ge p\}$. The pth empirical quantile can be determined by solving the minimization problem

$$\hat{Q}(p) = \arg\min_x \left\{ \sum_{i \mid X_i \ge x} p \mid X_i - x \mid + (1-p) \sum_{i \mid X_i < x} \mid X_i - x \mid \right\}.$$

Quantile regression is an extension of linear regression that is used when the conditions of linear regression are not met (i.e., linearity, homoscedasticity, independence, or normality).

For the quantile regression we make an assumption, that the pth quantile is given as a linear function of the explanatory variables. In the case of the empirical regression and random observations of dependent and independent variables $Y_1, Y_2, \ldots, X_n$ and $X_1, X_2, \ldots, X_n, \ldots$, the coefficients $\beta(\tau)$ of the τth empirical

quantile regression can be determined by solving the minimization problem

$$\sum_i \tau \max\,(0, Y_i - \beta'(\tau)X_i) + (1-\tau)\max\,(0, \beta'(\tau)X_i - Y_i)) \tag{9}$$

or problem

$$\sum_i \max\,(\tau(Y_i - \beta'(\tau)X_i), (1-\tau)(\beta'(\tau)X_i - Y_i)), \tag{10}$$

which is similar to the problem in Section 1.3. The minimization problem can be reduced to a linear programming problem.

For quantile regression, it is possible to calculate any quantile (percentage) for particular values of the dependent variables. Solving the problem for all $\tau \in [0,1]$, it is possible to recover the entire conditional quantile function, i.e., the conditional distribution function, of Y. If $\tau = 0.5$ the minimization problem derives the median. Taking a similar structure to the linear regression model, the "best" quantile regression model equation for the τth quantile is

$$Q_\tau(y_i) = \beta_0(\tau) + \beta_1(\tau)x_{i1} + \beta_2(\tau)x_{i2} + \beta_3(\tau)x_{i3} + \ldots + \beta_m(\tau)x_{im},$$

where $i = 1, \ldots, n$ is a number of observations and m is a number of covariates (independent variables). Coefficients $\beta_m(\tau)$ are functions of the required quantile τ. They are defined as

$$\beta(\tau) = \arg\min_{\beta \in R^m}\left\{ \sum_{i \mid Y_i \geq \beta'(\tau)X_i} \tau\,|Y_i - \beta'(\tau)X_i| + (1-\tau) \sum_{i \mid Y_i < \beta'(\tau)X_i} |Y_i - \beta'(\tau)X_i| \right\}, \tag{11}$$

where Y_i are observations of dependent variables, X_i is a vector of independent variables $X_i = (x_{i1}, \ldots, x_{im})$, and $\beta(\tau)$ is a vector of coefficients $\beta(\tau) = (\beta_1(\tau), \ldots \ldots, \beta_m(\tau))$, and m is a number of observations.

4. SELECTED RESULTS

4.1. Data. In these studies, we use the Pan-European version of the EPIC-IIASA model with relevant data and results derived for EU countries [5, 6]. In this EPIC version, the daily meteorological data were obtained from the Joint Research Centre's (JRC) Crop Growth Monitoring System (CGMS) meteorological database [37] at a 50 km grid resolution for the period 1995–2007. Weather variables include daily and monthly averages of precipitation, maximum temperature, and minimum temperature, solar radiation.

Land cover information was taken from a combined CORINE 2000 and PELCOM map at 1 km resolution provided by JRC. Digital terrain information was derived from SRTM (Shuttle Radar Topographic Mission; [38]) and GTOPO sources (Global 30 Arc Second Elevation Data; http://eros.usgs.gov).

Soil data [39] can be obtained from the European Soil Bureau Database (ESBD v. 2.0), including the Soil Geographic Database of Europe, the Soil Profile Analytical Database of Europe, the Pedo-Transfer Rules Database, the Database of Hydraulic Properties of European Soils and the Map of Organic Carbon Content in topsoils in Europe. Soil data can be considered as time-invariant factors, however, they are affected by various land use and soil practices. For these, the historical data on crop yield responses to weather parameters under certain practices (and thus soil properties) may become available from EPIC simulations.

Agricultural statistics on crop yields and fertilizer consumptions were retrieved from the Statistical Office of the European Communities (EUROSTAT) and IFA/FAO datasets. Information on rainfed and irrigated crop areas were taken from the European Irrigation Map (EIM).

4.2. Selected results. The data for the statistical model were harmonized at the resolution of about 120,000 EPIC SimUs. The SimUs are represented, as a rule, by one area with "representative" characteristics for soil, topography, and present

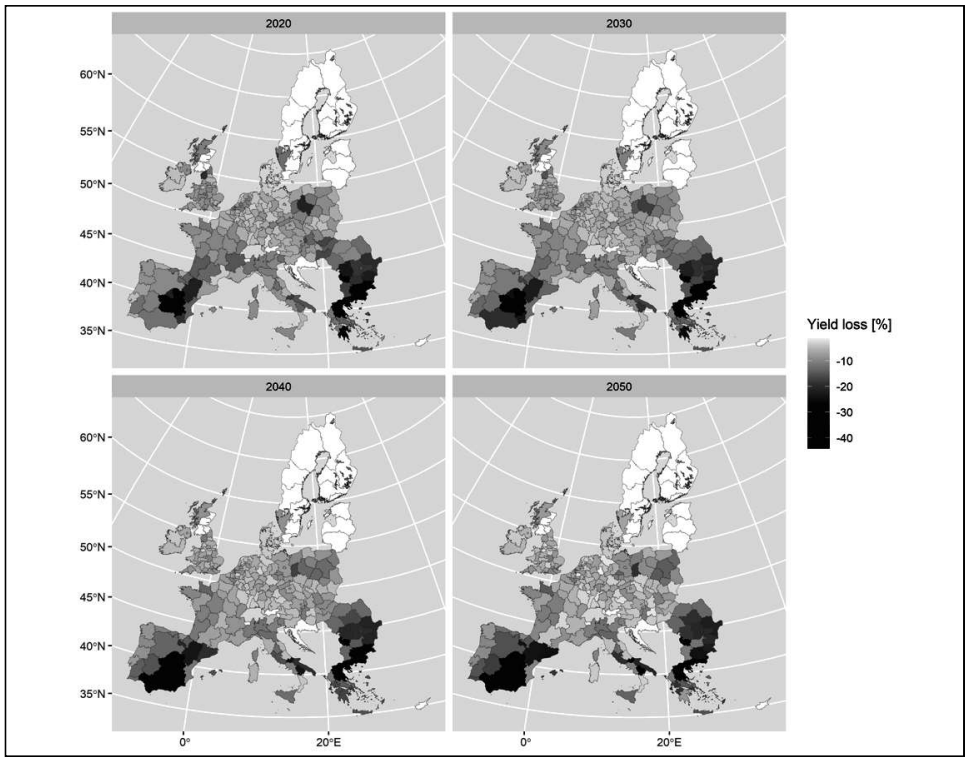

Fig. 1. Crop yield loss as the difference between the 25th and the median of the crop yield, from 2000 to 2050, in percentage terms, for rainfed maize, RCP26

weather. Since crops have different needs for water, temperature, solar radiation, and other weather parameters in different stages of their development, the weather data was aggregated to monthly resolution. If data is available, the quantile regression can be estimated at the level of SimUs. In this paper, we present the results at the level of EU NUTS administrative regions. The NUTS classification lists 92 regions at NUTS 1, 242 regions at NUTS 2 and 1166 regions at NUTS 3 level.

Probability distributions of crop yields in different years can be analyzed according to their critical quantiles (5th, 25th, 50s, 75th, and 95th) as well as mean values in these years. Often, the critical quantiles (percentile) can be defined by the so-called break-even analysis of the crop producer, e.g., specifying the costs associated with production excess and shortage (shortfalls) as in example of section 1.3.

Figures 1 and 2 present the results at NUTS level, each color characterizes the difference between the values of the 25th percentile and the 50th percentile (Fig. 1) and the values of the 25th percentile and the average crop yield (Fig. 2), in percentage terms, for years from 2000 to 2050. Here we present the results for rainfed maize yield and for the RCP26 (early response) emission scenario. The results are displayed for the years from 2000 to 2050 with a 10 years timestep. In general, the meta-model results are available for all years to 2100, for different crops and RCP scenario, to characterize probability distributions of crop yields for different climatic scenarios.

The percentage difference we define as the yield loss with respect to the reference values (mean, median, critical value, break-even yield, etc.). From Figs. 1 and 2, the negative effect of climate change on rainfed maize yield is especially visible in the southern regions of Europe, i.e., in Spain, Italy, Greece. In these countries, the yield loss to 2050 can be up to 40 percent. In different EU countries, the yield loss can attribute to different reasons, however in southern regions it is primarily due to the increased temperature and water deficits in specific crop growth periods.

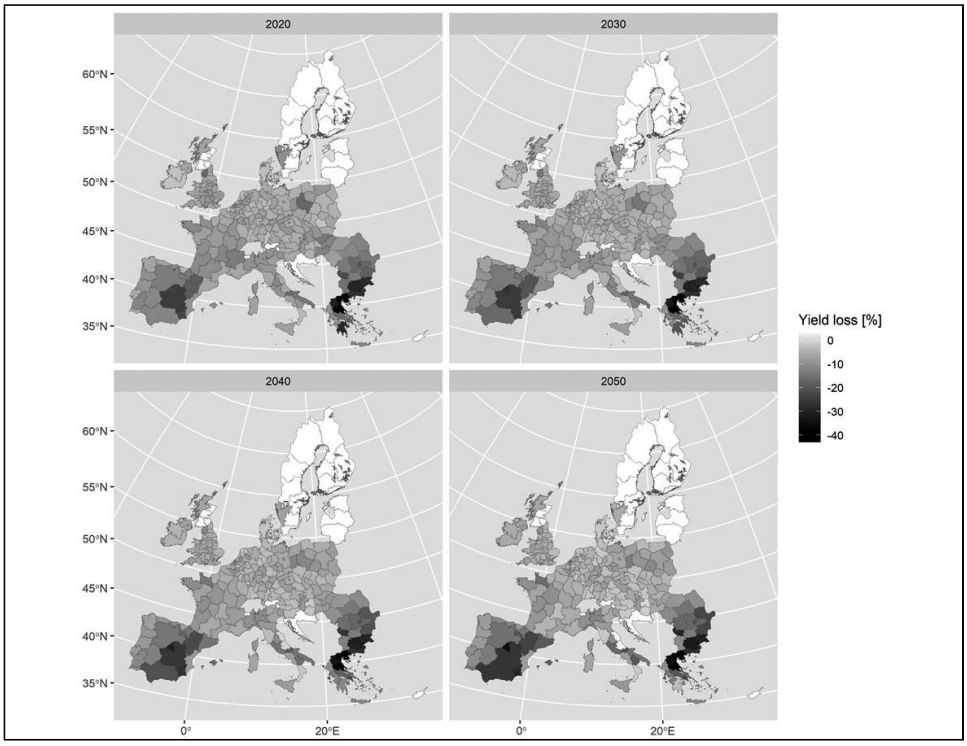

Fig. 2. Crop yield loss as the difference between the 25th and the average crop yield, from 2000 to 2050, in percentage terms, for rainfed maize, RCP26

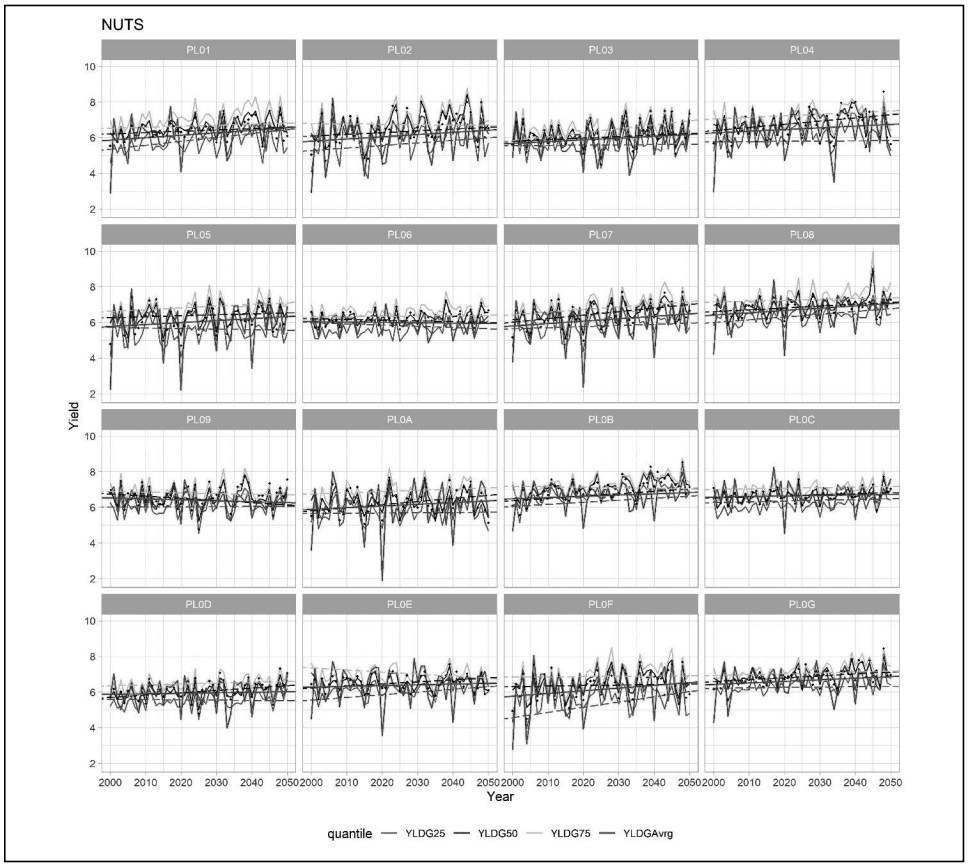

Fig. 3. Quantiles and "best" quantile regression lines, NUTS level, Poland

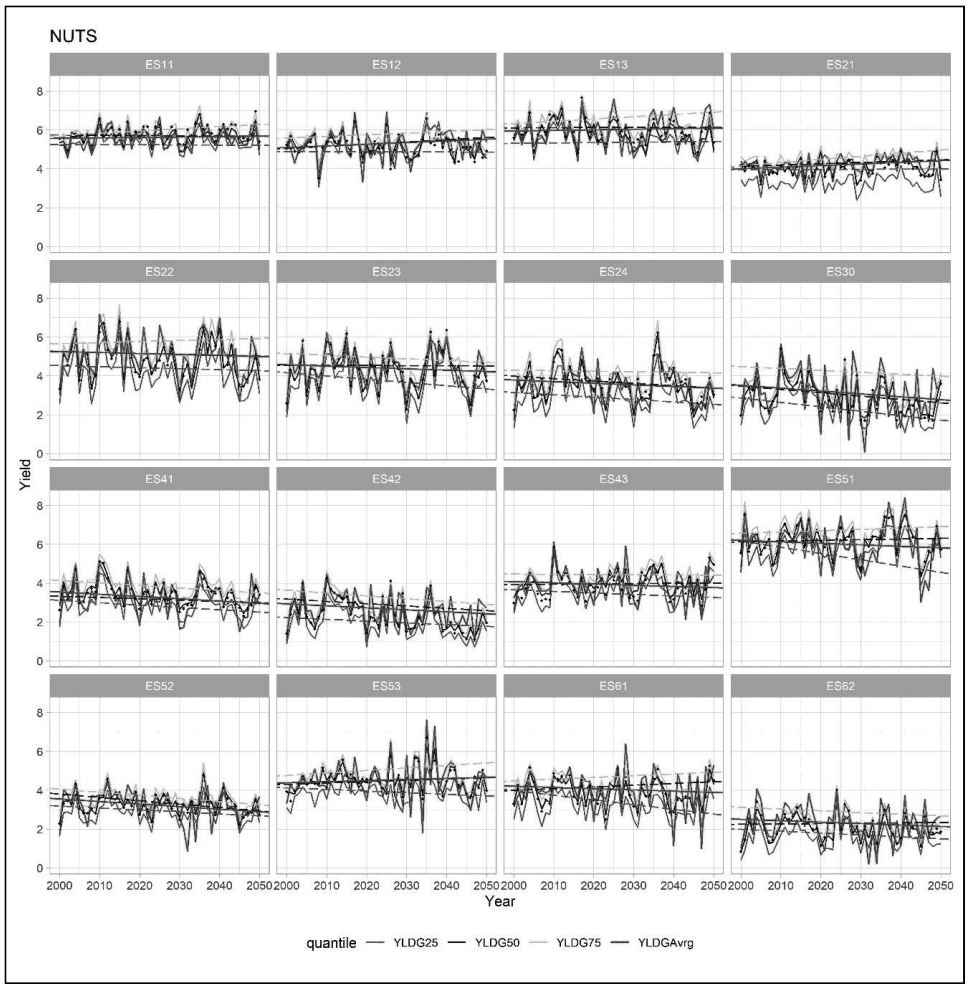

Fig. 4. Quantiles and "best" quantile regression lines, NUTS level, Estonia

Projected air temperature increases throughout the remainder of the 21st century suggests that grain yields will continue to decrease for the major crops because of the increase temperature stress on all major grain crops [3]. Beyond a certain point, higher air temperatures adversely affect plant growth, pollination, and reproductive processes [40, 41].

Figures 3–5 illustrate the results of estimating the "best" quantile regression fit line aggregated to the level of country-specific NUTS regions. As example, we take Poland, Estonia and Italy, which represent South, Middle and North of Europe.

Visualized quantiles are 25th, 50th, and 75th. In addition, these figures display the average crop yield for the years from 2000 to 2050. The results of the simulations are available till 2100, but for the reasons of clarity, the displayed values are restricted to 50 years. The estimates of the crop yields $Q_\tau(y_i)$ in each SimUs within all NUTS regions and EU countries having probability

$$Prob\{Q_\tau(y_i) \le \beta_0(\tau) + \beta_1(\tau)x_{i1} + \beta_2(\tau)x_{i2} + \beta_3(\tau)x_{i3} + \ldots + \beta_m(\tau)x_{im}\} = \tau \quad (12)$$

are calculated with coefficients $\beta_m(\tau)$ (11) for each quantile τ (percentile 100τ), where $i = 1, \ldots, n$ is a number of observations and m is a number of covariates (independent variables). Coefficients $\beta_m(\tau)$ are functions of quantile. Equation (12) means, that 100τ percent of crop yield observations is less than the value of the τ-quantile. Equation (12) provides the basis for the validation of the quantile regression model.

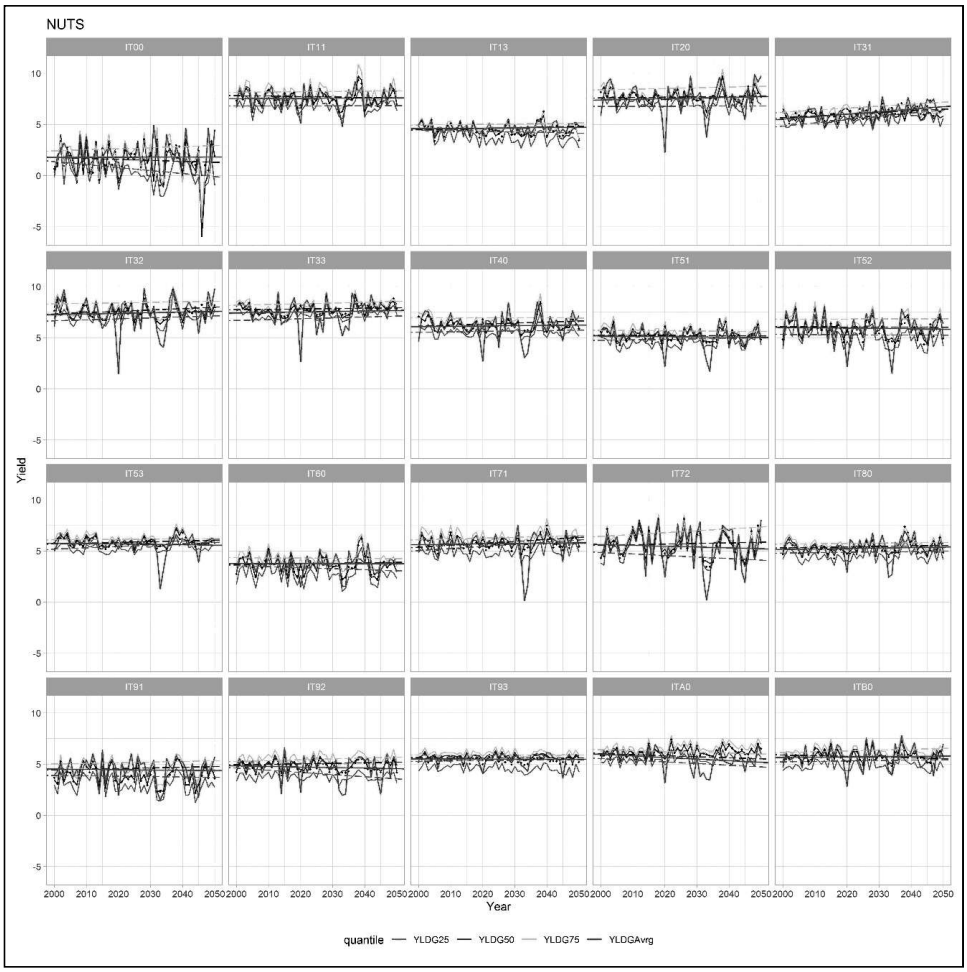

Fig. 5. Quantiles and "best" quantile regression lines, NUTS level, Italy

CONCLUSIONS

This paper develops a quantile regression model for the analysis and prediction of crop yield distributions in the presence of weather variability, climate change, altering soil properties. The developed model is important for crop insurers, farmers, authorities dealing with food security considerations, agricultural market planners, etc., as it provides insights regarding the likelihoods of various crop yield levels. The yield levels can be distinguished as 25th, 50th, and 75th percentiles, mean values, or critical values, e.g., derived from the analysis of break-even crop yield.

A (multiple) linear quantile regression (MLQR) is one of the simplest machine learning algorithms because it is easy to understand the responses of the dependent variables to each covariate.

The analysis of probability distribution of "crop yield loss" can be useful for farmers, insurers, other stakeholders, as a possible trigger of systemic risks in agricultural markets, interconnected food-water-energy systems, insurance networks, etc., the earlier information about the likelihood of critical crop yields and crop yield losses is essential for securing food production.

The validation procedure for the estimated quantile regression models is based on equation (12), i.e., the percentage of the crop yield observations is below the respective estimated quantile.

Machine learning approaches (utilizing hardly interpretable neural networks) can be effectively combined with disciplinary or interdisciplinary models, e.g., agricultural, environmental, energy, etc., for effective decision-making in the conditions of uncertainty, increasing interdependencies and complex analytically intractable systemic ("unknown risks").

Statistical or machine learning models can be used as reduced form scenario generators (meta-models) of stochastic events (scenarios), as a submodel of a more complex e.g., Integrated Assessment model (IAM) GLOBIOM [17–20].

REFERENCES

1. Clarke B., Otto F., Stuart-Smith R., Harrington L. Extreme weather impacts of climate change: An attribution perspective. *Environmental Research: Climate*. 2022. Vol. 1, Iss. 1. 012001. https://doi.org/10.1088/2752-5295/ac6e7d.
2. Hatfield J.L., Boote K.J., Kimball B.A., Ziska L.H., Izaurralde R.C., Ort D., Thomson A.M., Wolfe D.W. Climate impacts on agriculture: Implications for crop production. *Agron. J*. 2011. Vol. 103, Iss. 2. P. 351–370.
3. Hatfield J.L., Prueger J.H. Temperature extremes: Effect on plant growth and development. *Weather and Climate Extremes.* 2015. Vol. 10, Part A. P. 4–10.
4. Pareek N. Climate change impact on soils: adaptation and mitigation. *MOJ Eco. Environ. Sci*. 2017. Vol. 2, Iss. 3. P. 136–139. https://doi.org/10.15406/mojes.2017.02.00026.
5. Balkovič J., van der Velde M., Skalský R., Xiong W., Folberth C., Khabarov N., Smirnov A., Mueller N.D., Obersteiner M. Global wheat production potentials and management flexibility under the representative concentration pathways. *Global and Planetary Change*. 2014. Vol. 122. P. 107–121. https://doi.org/10.1016/j.gloplacha.2014.08.010.
6. Balkovič J., Velde M., Schmid E., Skalský R., Khabarov N., Obersteiner M., Stürmer B., Xiong W. Pan-European crop modelling with EPIC: Implementation, up-scaling and regional crop yield validation. *Agric. Syst.* 2013. Vol. 120. P. 61–75.
7. Jones C.A., Dyke P.T., Williams J.R., Kiniry J.R., Benson V.W., Griggs R.H. EPIC: An operational model for evaluation of agricultural sustainability. *Agric. Syst.* 1991. Vol. 37, Iss. 4. P. 341–350.
8. Williams J.R. The erosion productivity impact calculator (EPIC) model: A case history. *Philosophical Transactions: Biological Sciences*. 1990. Vol. 329, Iss. 1255. P. 421–428.
9. Drummond S.T., Sudduth K.A., Joshi A., Birrell S.J., Kitchen N.R. Statistical and neural methods for site-specific yield prediction. *Transactions of the ASAE*. 2003. Vol. 46, Iss. 1. P. 5–14. https://doi.org/10.13031/2013.12541.
10. Van Klompenburg T., Kassahun A., Catal C. Crop yield prediction using machine learning: A systematic literature review. *Computers and Electronics in Agriculture.* 2020. Vol. 177. 105709. https://doi.org/10.1016/j.compag.2020.105709.
11. Müller C., Elliott J., Chryssanthacopoulos J., Arneth A., Balkovic J., Ciais P. et al. Global gridded crop model evaluation: benchmarking, skills, deficiencies and implications. *Geosci. Model Dev*. 2017. Vol. 10, Iss. 4. P. 1403–1422. https://doi.org/10.5194/gmd-10-1403-2017.
12. Van der Goot E., Supit I., Boogaard H. L., van Diepen K., Micale F., Orlandi S., Otten H., Geuze M., Schulze D. Methodology of the MARS crop yield forecasting system. Vol. 1: Meteorological data collection, processing and analysis. Luxembourg: European Commission (EC), 2004. 100 p.
13. Li X., Ren J., Niu B., Wu H. Grain area yield index insurance ratemaking based on time-space adjustment in China. *Sustainability*. 2020. Vol. 12, Iss. 6. 2491. https://doi.org/10.3390/su12062491.
14. Okhrin O., Odening M., Xu W. Systemic weather risk and crop insurance: The case of China. *J. Risk Insur.* 2013. Vol. 80, N 2. P. 351–372.
15. Ermoliev Y.M., Ermolieva T.Y., MacDonald G.J., Norkin V.I. Insurability of catastrophic risks: the stochastic optimization model. *Optimization*. 2000. Vol. 47, N 3. P. 251–265.

16. Ermoliev Y.M., Ermolieva T.Y., MacDonald G.J., Norkin V.I. Stochastic optimization of insurance portfolios for managing exposure to catastrophic risks. *Annals of Operations Research*. 2000. Vol. 99, N 1. P. 207–225.
17. Havlík P., Schneider U.A., Schmid E., Boettcher H., Fritz S., Skalský R., Aoki K., de Cara S., Kindermann G., Kraxner F., Leduc S., McCallum I., Mosnier A., Sauer T., Obersteiner M. Global land-use implications of first and second generation biofuel targets. *Energy Policy*. 2011. Vol. 39. P. 5690–5702.
18. Ermolieva T., Havlík P., Ermoliev Y., Mosnier A., Obersteiner M., Leclere D., Khabarov N., Valin H., Reuter W. Integrated management of land use systems under systemic risks and security targets: A Stochastic Global Biosphere Management Model. *Journal of Agricultural Economics.* 2016. Vol. 67, Iss. 3. P. 584–601.
19. Ermolieva T., Havlik P., Ermoliev Y., Khabarov N., Obersteiner M. Robust management of systemic risks and food-water-energy-environmental security: Two-stage strategic-adaptive GLOBIOM model. *Sustainability*. 2021. Vol. 13, Iss. 2. 857. https://doi.org/10.3390/su13020857.
20. Ermolieva T., Havlik P., Frank S., Kahil T., Balkovič J., Skalský R., Ermoliev Y., Knopov P.S., Borodina O.M., Gorbachuk V.M. A risk-informed decision-making framework for climate change adaptation through robust land use and irrigation planning. *Sustainability*. 2022. Vol. 14, Iss. 3. 1430. https://doi.org/10.3390/su14031430.
21. Ermolieva T., Ermoliev Y., Obersteiner M., Rovenskaya E. Two-stage nonsmooth stochastic optimization and iterative stochastic quasigradient procedure for robust estimation, machine learning and decision making. In: Resilience in the Digital Age. Roberts F.S., Sheremet I.A. (Eds). Cham: Springer, 2021. P. 45–74 https://doi.org/10.1007/978-3-030-70370-7_4.
22. Ermolieva T., Ermoliev Y., Havlik P., Lessa-Dersi-Augustynczik A., Komendantova N., Kahil T., Balkovic J., Skalsky R., Folberth C., Knopov P.S., Wang G. Connections between robust statistical estimation, robust decision making withtwo-stage stochastic optimization, and robust machine learning problems. *Cybernetics and Systems Analysis.* 2023. Vol. 59, N 3. P. 385–397. https://doi.org/10.1007/s10559-023-00573-3.
23. Ermoliev Y. Stochastic quasigradient methods. In: Encyclopedia of Optimization. Pardalos P.M. (Ed.). New York: Springer Verlag, 2009. P. 3801–3807.
24. Ermoliev Y. Two-stage stochastic programming: Quasigradient method. In: Encyclopedia of Optimization. Pardalos P.M. (Ed.). New York: Springer Verlag, 2009. P. 3955–3959.
25. Ermoliev Y. Stochastic quasigradient methods in minimax problems. In: Encyclopedia of Optimization. Pardalos P.M. (Ed.). New York: Springer Verlag, 2009. P. 3813–3818.
26. Ermoliev Y.M., Wets R.J.-B. Numerical Techniques for Stochastic Optimization. Heidelberg: Springer Verlag, 1988. URL: https://pure.iiasa.ac.at/id/eprint/3065/.
27. Ermoliev Y.M., Gaivoronski A.A. Stochastic quasigradient methods for optimization of discrete event systems. *Annals of Operation Research*. 1992. Vol. 39. P. 1–39. https://doi.org/10.1007/BF02060934.
28. Ermoliev Y., Hordijk L. Global changes: Facets of robust decisions. In: Coping with Uncertainty: Modeling and Policy Issue. Marti K., Ermoliev Y., Makowski M., Pflug G. (Eds.). Berlin: Springer Verlag, 2003.
29. Jägermeyr J., Gerten D., Heinke J., Schaphoff S., Kummu M., Lucht W. Water savings potentials of irrigation systems: global simulation of processes and linkages. *Hydrol. Earth Syst. Sci*. 2015. Vol. 19, Iss. 7. P. 3073–3091.
30. Folberth C., Baklanov A., Balkovic J., Skalsky R., Khabarov N., Obersteiner M. Spatio-temporal downscaling of gridded crop model yield estimates based on machine learning. *Agricultural and Forest Meteorology*. 2019. Vol. 264. P. 1–15. https://doi.org/10.1016/j.agrformet.2018.09.021.
31. Williams J.R., Jones C.A., Dyke P.T. A modelling approach to determining the relationship between erosion and soil productivity. *Trans. ASAE*. 1984. Vol. 27, Iss. 1. P. 129–144.
32. Rosenzweig C., Jones J.W., Hatfield J.L., Ruane A.C., Boote K.J., Thorburn P., Antle J.M., Nelson G.C., Porter C., Janssen S., Asseng S., Basso B., Ewert F., Wallach D., Baigorria G., Winter J.M. The Agricultural Model Intercomparison and Improvement Project (AgMIP): Protocols and pilot studies. *Agric. For. Meteorol.* 2013. Vol. 170. P. 166–182.

33. Kahil M.T., Dinar A., Albiac J. Modeling water scarcity and droughts for policy adaptation to climate change in arid and semiarid regions. *Journal of Hydrology*. 2015. Vol. 522. P. 95–109.
34. Kahil M.T., Connor J.D., Albiac J. Efficient water management policies for irrigation adaptation to climate change in Southern Europe. *Ecol. Econ*. 2015. Vol. 120(C). P. 226–233.
35. Golodnikov A.N., Ermoliev Y.M., Ermolieva T.Y., Knopov P.S., Pepelyaev V.A. Integrated modeling of food security management in Ukraine. II. Models for structural optimization of agricultural production under risk. *Cybernetics and Systems Analysis*. 2013. Vol. 49, N 2. P. 217–228. https://doi.org/10.1007/s10559-013-9503-6.
36. Zhu K., Ran H., Wang F., Ye X., Niu L., Schulin R., Wang G. Conservation tillage faclitated soil carbon sequestration through diversified carbon conversions. *Agriculture, Ecosystems and Environment.* 2022. Vol. 337. 108080.
37. Genovese G., Bettio M. (Eds.). Methodology of the MARS Crop Yield Forecasting System. Vol. 4 Statistical Data Collection, Processing and Analysis. Luxembourg: European Commision (EC).
38. Werner M. Shuttle Radar Topography Mission (SRTM), Mission overview. *J. Telecom (Frequenz)*. 2001. Vol. 55. P. 75–79.
39. Skalský R., Tarasovičová Z., Balkovič J., Schmid E., Fuchs M., Moltchanova E., Kindermann G., Scholtz P. GEO-BENE Global Database for Bio-Physical Modeling v. 1.0 (Concepts, Methodologies and Data). 2008. URL: https://geo-bene.project-archive.iiasa.ac.at/files/Deliverables/Geo-BeneGlbDb10(DataDescription).pdf.
40. Klein J.A., Harte J., Zhao X.-Q. Experimental warming, not grazing, decreases rangeland quality on the Tibetan plateau. *Ecol. Appl.* 2007. Vol. 17, Iss. 2. P. 541–557.
41. Sacks W.J., Kucharik C.J. Crop management and phenology trends in theU.S. corn belt: Impacts on yields, evapotranspiration and energy balance. *Agric. For. Meteor*. 2011. Vol. 151, Iss. 7. P. 882–894.

Т. Єрмольєва, П. Гавлик, А. Лесса-Дерсі-Аугустинчик, Е. Бор, С. Франк, Т. Кахіл, Г. Ванг, Ю. Балкович, Р. Скальські, К. Фолберт, Н. Комендантова, П.С. Кнопов

НОВА НАДІЙНА СТРУКТУРА МЕТАМОДЕЛІ ДЛЯ ПРОГНОЗУВАННЯ РОЗПОДІЛУ ЙМОВІРНОСТІ ВРОЖАЙНОСТІ З ВИКОРИСТАННЯМ ДАНИХ ІЗ БАГАТЬОХ ДЖЕРЕЛ

Анотація. Зазначено, що є нагальна потреба у кращому розумінні та прогнозуванні впливу погодних явищ (зокрема, екстремального характеру, як-от, сильних опадів, посух та теплових хвиль) на врожайність сільськогосподарських культур. Це дасть змогу покращити майбутні прогнози виробництва сільськогосподарських культур в умовах мінливості погоди, екстремальних явищ та зміни клімату. Розроблено квантильні регресійні моделі для оцінювання розподілу ймовірності врожайності сільськогосподарських культур залежно від місячних значень температури та опадів та якісних характеристик ґрунту, які можуть бути доступні для різних прогнозів зміни клімату. Врожайність сільськогосподарських культур, історичну та синтезовану моделлю EPIC, аналізують та розрізняють відповідно до їхніх рівнів, тобто середніх та критичних квантилів. Далі квантилі врожайності сільськогосподарських культур апроксимують, налаштовуючи окремі моделі регресії на основі квантилів. Розроблена статистична метамодель врожайності сільськогосподарських культур дає змогу аналізувати врожайність сільськогосподарських культур залежно від таких екзогенних параметрів, як температура та опади, а також ендогенних параметрів (наприклад, характеристик ґрунту), які можуть змінюватися у результаті практик землекористування. Статистичні та машинні моделі навчання можна використовувати як генератори сценаріїв зменшеної розмірності та структури (метамоделі) стохастичних подій (сценаріїв), як підмодель більш складних моделей, наприклад, моделі інтегрованої оцінки (IAM) GLOBIOM.

Ключові слова: екстремальні події, зміна клімату, продовольча безпека, прогнози врожайності, розподіл імовірностей, квантильні регресії, надійна оцінка та машинне навчання, двоетапна задача стохастичної оптимізації (STO).

Надійшла до редакції 10.04.2023