Agricultural and Forest Meteorology 263 (2018) 428–448

Contents lists available at ScienceDirect

Agricultural and Forest Meteorology

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

Cotton yield prediction with Markov Chain Monte Carlo-based simulation model integrated with genetic programing algorithm: A new hybrid copula-driven approach

Mumtaz Ali*, Ravinesh C. Deo*, Nathan J. Downs, Tek Maraseni

School of Agricultural, Computational and Environmental Sciences, Institute of Life Sciences and the Environment, University of Southern Queensland, Springfield, QLD, 4300, Australia

ARTICLE INFO

Keywords:
Crop yield prediction
Cotton yield
Climate data
Genetic programming
Markov Chain Monte Carlo based copula model

ABSTRACT

Reliable data-driven models designed to accurately estimate cotton yield, an important agricultural commodity, can be adopted by farmers, agricultural system modelling experts and agricultural policy-makers in strategic decision-making processes. In this paper a hybrid genetic programing model integrated with the Markov Chain Monte Carlo (MCMC) based Copula technique is developed to incorporate climate-based inputs as the predictors of cotton yield, for selected study regions: Faisalabad (31.4504 °N, 73.1350 °E), Multan (30.1984 °N, 71.4687 °E) and Nawabshah (26.2442 °N, 68.4100 °E), as important cotton growing hubs in the developing nation of Pakistan. Several different types of GP-MCMC-copula models were developed, each with the well-known copula families (*i.e.*, Gaussian, student t, Clayton, Gumble Frank and Fischer-Hinzmann functions) to screen and utilize an optimal cotton yield forecast model for the present study region. The results of the GP-MCMC based hybrid copula model were evaluated with a standalone GP and the MCMC based copula model in accordance with statistical analysis of the predicted yield based on correlation coefficient (r), Willmott's index (WI), Nash-Sutcliffe coefficient (NS_E), root mean squared error (RMSE) and mean absolute error (MAE) in the independent test phase. Further performance preciseness was evaluated by the Akiake Information Criterion (AIC), the Bayesian Information Criterion (BIC) and the Maximum Likelihood (Max_L) for the GP-MCMC based copula as well as the MCMC based copula model. GP-MCMC-Clayton copula model generated the most accurate result for the Multan station. For the optimal GP-MCMC-Clayton copula model, the acquired model evaluation metrics for Multan were: (LM ≈ 0.952; RRMSE ≈ 2.107%; RRMAE ≈ 1.771%) followed by the MCMC based Gaussian copula model (LM ≈ 0.895; RRMSE ≈ 4.541%; RRMAE ≈ 0.3.214%) and the standalone GP model (LM ≈ 0.132; RRMSE ≈ 23.638%; RRMAE ≈ 22.652%), indicating the superiority of the GP-MCMC-Clayton copula model in respect to the other benchmark models. The performance of GP-MCMC based copula model was also found to be superior in the case of Faisalabad and Nawabshah station as confirmed by AIC, BIC, Max_L metrics, including a larger value of the Legates-McCabe's (LM) index, utilized in conjunction with the relative percentage RRMSE and the relative mean absolute error (RMAE). Accordingly, it is averred that the developed GP-MCMC copula model can be considered as a pertinent data-intelligent tool used for accurate prediction of cotton yield, utilizing the readily available climate datasets in agricultural regions and is of relevance to agricultural yield simulation and sectoral decision-making.

1. Introduction

Timely information on the crop yield is important for agriculture-dependent nations (e.g., Pakistan), as this can generate crucial ideas for agricultural policy making, and forward planners and agricultural markets. Agriculture in Pakistan is known to contribute to about 21% of the county's GDP (Sarwar, 2014), which include cotton as an important cash crop. This is because cotton is an integral commodity for the economic development of Pakistan as the nation is highly dependent on the cotton industry and its related textile sector due to which the cotton crop has been given a principal status in the country. Cotton crop is grown from May-August as an industrial crop in 15% of the nation's available land area producing 15 million bales during 2014-15 (Reporter, 2015). Pakistan is placed at fourth position among cotton

* Corresponding authors.
E-mail addresses: Mumtaz.Ali@usq.edu.au (M. Ali), ravinesh.deo@usq.edu.au (R.C. Deo).

https://doi.org/10.1016/j.agrformet.2018.09.002
Received 21 January 2018; Received in revised form 18 July 2018; Accepted 1 September 2018

growers, third largest exporter and fourth largest consumer (Banuri, 1998). In 2013, about 1.6 million farmers (out of a total of 5 million in all sectors) engaged in cotton farming, growing more than 3 million hectares (Banuri, 1998; Reporter, 2015).

Data-intelligent models, utilizing past data can offer an accurate solutions to the problems related to the projection of future trends in agriculture, crop yield, rainfall and drought that affects agricultural productivity (Ali et al., 2018a, b; Bauer, 1975; Nguyen-Huy et al., 2017, 2018). Machine learning models, which are highly non-linear models, utilize data that has input features valued for the prediction of crop yield. In the work of Kern et al. (2018), multiple linear regression models were constructed to simulate the yield of the four major crop types in Hungary using environmental and remote sensing information. Moreover, Bokusheva et al. (2016) developed copula models for crop yields on VH indices and Craparo et al. (2015) built an ARIMA model to forecast the decline of coffee yield in Tanzania. Debnath et al. (2013) predicted area and cotton yield in India using an ARIMA model. Blanc et al. (2008) utilized a multiple regression model of the main climatic determinants of rain fed cotton yield in West Africa. Yang et al. (2014) assessed cotton yield and water demand under climate change and future adaptation measures using APSIM-OzCot model. Chen et al. (2011) studied the impact of climate change on cotton production and water consumption using COSIM model in China. Hearn (1994) design a simulation model named OZCOT for cotton crop management in Australia. Papageorgiou et al., (2011) predicted cotton yield using fuzzy cognitive maps in 2011, Greece. Jin and Xu (2012) conducted a study on the estimation of cotton yield using Carnegie Ames Stanford Approach model in China. The aforementioned models were developed to study the climate change impacts on cotton yield prediction.

In summary, existing literature shows that there are few studies in Pakistan that have developed methods for the prediction of cotton yield, despite its relevance as a world leader in cotton production. Ali et al. (2015) used a forecasting ARIMA model for the production of sugarcane and cotton crops of Pakistan from 2013–2030. Hina Ali et al. (2013) also analyzed production forecasting of cotton in Pakistan. Ahmad et al. (2017) developed an ARIMA model to forecast area, production and yield of major crops in Pakistan in 2017. Raza and Ahmad (2015) studied the impact of climate change on cotton productivity in Punjab and Sindh, Pakistan using fixed effect models. Ayaz et al. (2015) studied weather effect on cotton crop in Sindh, Pakistan. Carpio and Ramirez (2002) used yield and acreage models to forecast cotton yield in India, Pakistan and Australia. Ahmad (1975) designed a time series prediction for the supply response of cotton in Punjab, Pakistan in 1975.

All the previous studies indicate that the prediction of cotton yields have been based primarily on the effect of climate change with the adoption of ARIMA model only. In addition to that, all these studies have been conducted for a large area, either for a whole province, or national region, but not for a small locality. Moreover, there is a limitation of applying advanced data-intelligent algorithms for more accurate prediction models at a micro scale which can provide help for decision-making in precision agriculture and farming systems which may be the way future farming trends are analyzed. To address these mentioned issues, there is an apparent need for data intelligent models to predict cotton yield more accurately and at a much finer scale than attempted previously. In this study, for the first time, a hybrid genetic programing integrated with a Markov Chain Monte Carlo (GP-MCMC) based copula model has been developed for the prediction of cotton yield in Faisalabad, Multan and Nawabshah in Pakistan. The novelty of this study is to utilize as yet untested accurate GP-MCMC based copula models for the prediction of cotton yield in Pakistan.

To advance the application of copula models, especially in agriculture where they have been relatively scarcely applied the present study aims to address four primary objectives. (1) To apply GP and MCMC based copula, MCMC based copula models and a standalone GP model to determine which is of these models is the most accurate data-intelligent tool for predicting cotton yield in the developing nation of Pakistan. (2) To model influence of climate dataset (*i.e.*, temperature, rainfall and humidity) to predict effectively the cotton yield in the proposed districts of Punjab and Sindh, the primary agricultural hubs in Pakistan. (3) To develop and optimize the copula-based models by tuning the GP and the MCMC techniques as well as to evaluate their performances in comparison with MCMC based copula and standalone GP model. (4) To validate the predictive ability of each model with respect to cotton yield in Pakistan, making a major contribution to the use of data-driven models for agricultural yield estimation.

2. Theoretical framework

In this section an overview of the proposed predictive GP-MCMC based copula models with its comparative counterparts, MCMC based copula models and GP are presented.

2.1. Genetic programming (GP) model

Genetic programming is a heuristic evolutionary algorithm which has the potential to offer solutions of any form without the user specifying the problem (Deo and Samui, 2017; Koza, 1992). Evolutionary principles are utilized to acquire the persistent patterns in the structure of data without requiring prior knowledge. According to McPhee et al. (2008), an organized domain-independent method is used to genetically breed a population in genetic programming for getting computers to solve the problem that is starting from a high-level statement of what needs to be done. Fig. 1(a) demonstrates the basic structure of a GP model. More specifically, a population is transformed iteratively to produce successive new generations of programs by using similar genetic operations that occur naturally. These genetic operations are divided into five components: crossover (sexual recombination), mutation; reproduction; gene duplication; and gene deletion. Huang et al. (2006) showed that a GP model has the skill of self-parameterizing to extract the features bypassing the user, tuning the model, and due to this capability resembles to some extent the Extreme Learning Machine model.

In a GP model, the input data goes through a number of routes where (1) analyzation of attributes occurs; (2) selection of the best fitness functions are made for the purpose of minimizing the mean-squared error; (3) generation of functional and terminal sets; and (4) parameterization of genetic operations (Sreekanth and Datta, 2011). The GP model is optimized by the emulation of an evolutionary process to an adequate agreement between the response and input variable. A functional node performs the arithmetic operations (þ, −, ×, ÷), Boolean logic functions (AND, OR, NOT), conditionals (IF, THEN, ELSE), or any other functions (SIN, EXP) that may be used. A random tree structure is developed using these functions (Deo and Samui, 2017; Mehr et al., 2013). GP is developed in this paper by (1) randomly creating the initial population (i.e., computer program); (2) performing the execution of the program with the best fitness values; (3) based on reproduction, mutation, and crossover, generation of a new population of computer programs; (4) comparison and evaluation of fitness; and (5) finally the selection of the best program by the evolutionary process (Mehr et al., 2013). For this purpose, a randomly equated population is being created and best fitness is determined where “parents” are chosen individually and the “off-springs” are developed from the parents through the process of reproduction, mutation, and crossover (Deo and Samui, 2017).

2.2. Markov Chain Monte Carlo (MCMC) based copula models

In this study we hybridize the MCMC-copula models used previously (*e.g.*, (Ali et al., 2018b)) with the GP algorithm. Basically, a copula model, which has recently found important applications in the agricultural sector, is a powerful mathematical tool that has the ability to connect two or more time-independent variables (Nelsen, 2003; Nguyen-Huy et al., 2017, 2018). A copula function is basically a