IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received April 7, 2022, accepted April 14, 2022, date of publication April 18, 2022, date of current version May 3, 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3168322

Student Performance Predictions for Advanced Engineering Mathematics Course With New Multivariate Copula Models

THONG NGUYEN-HUY[1,2], RAVINESH C. DEO[3], (Senior Member, IEEE), SHAHJAHAN KHAN[4,5], ARUNA DEVI[6], ADEWUYI AYODELE ADEYINKA[2,7], ARMANDO A. APAN[8], AND ZAHER MUNDHER YASEEN[9,10,11]

[1]SQNNSW Drought Resilience Adoption and Innovation Hub, University of Southern Queensland, Toowoomba, QLD 4350, Australia
[2]Centre for Applied Climate Sciences, University of Southern Queensland, Toowoomba, QLD 4350, Australia
[3]School of Mathematics, Physics and Computing, University of Southern Queensland, Springfield, QLD 4350, Australia
[4]School of Mathematics, Physics and Computing, University of Southern Queensland, Toowoomba, QLD 4350, Australia
[5]School of Sciences and Engineering, Asian University of Bangladesh, Ashulia, Dhaka 1341, Bangladesh
[6]School of Teacher Education and Tertiary Access, University of the Sunshine Coast, Caboolture, QLD 4350, Australia
[7]Office of Research and Innovation, University of Southern Queensland, Toowoomba, QLD 4350, Australia
[8]School of Surveying and Built Environment, University of Southern Queensland, Toowoomba, QLD 4350, Australia
[9]Department of Earth Sciences and Environment, Faculty of Science and Technology, Universiti Kebangsaan Malaysia, Bangi, Selangor 43600, Malaysia
[10]USQ's Advanced Data Analytics Research Group, School of Mathematics Physics and Computing, University of Southern Queensland, Toowoomba, QLD 4350, Australia
[11]Scientific Research Center, New Era and Development in Civil Engineering Research Group, Al-Ayen University, Thi-Qar 64001, Iraq

Corresponding author: Ravinesh C. Deo (ravinesh.deo@usq.edu.au)

This work was supported in part by the University of Southern Queensland (School of Sciences *Q1* Challenge Grant 2018), and in part by the Office for Advancement of Learning and Teaching Technology Demonstrator Project Artificial Intelligence Predictive Analytics Framework under the Ethics Approval Number H18REA236.

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by the USQ Ethics Committee under Ethics Approval No. H18REA236.

ABSTRACT Engineering Mathematics requires that problem-solving should be implemented through ongoing assessments; hence the prediction of student performance using continuous assessments remains an important task for engineering educators, mainly to monitor and improve their teaching practice. This paper develops probabilistic models to predict weighted scores (*WS*, or the overall mark leading to a final grade) for face-to-face (on-campus) and web-based (online) Advanced Engineering Mathematics students at an Australian regional university over a 6-year period (2013-2018). We fitted parametric and non-parametric D-vine copula models utilizing multiple quizzes, assignments and examination score results to construct and validate the predicted *WS* in independently test datasets. The results are interpreted in terms of the probability of whether a student's continuous performance (i.e., individually or jointly with other counterpart assessments) is likely to lead to a passing grade conditional upon joint performance in students' quizzes and assignment scores. The results indicate that the newly developed D-vine model, benchmarked against a linear regression model, can generate accurate grade predictions, and particularly handle the problem of low or high scores (tail dependence) compared with a conventional model for both face-to-face, and web-based students. Accordingly, the findings advocate the practical utility of joint copula models that capture the dependence structure in engineering mathematics students' marks achieved. This therefore, provide insights through learning analytic methods to support an engineering educator's teaching decisions. The implications are on better supporting engineering mathematics students' success and retention, developing evidence-based strategies consistent with engineering graduate requirements through improved teaching and learning, and identifying/addressing the risk of failure through early intervention. The proposed methods can guide an engineering educator's practice by investigating joint influences of engineering problem-solving assessments on their student's grades.

INDEX TERMS Engineering mathematics performance prediction, D-vine copula, multivariate probability model, academic performance, education decision-making, statistical model.

The associate editor coordinating the review of this manuscript and approving it for publication was Ehab Elsayed Elattar.

I. INTRODUCTION

Over the last two years the problem of predicting students' ongoing learning using joint relationships between continuous assessments and final examinations are receiving attention from many researchers [1]–[4]. This is important as overall student outcomes define the quality of a university graduate's attributes, and is a primary factor that influences the growth of student number and ranking of a university [5]. Higher education institutions are now focusing on ways to improve student performance by provisioning early learning support through evidence-based student performance evaluation methods. Predicting and analysing performance is critical for academic progress [6] but from an educator's perspective, this issue remains a challenging task given the influence of many factors that affect a student's performance. Examples of such factors include the family background, psychological status, past schooling or academic achievements, and a learner's interaction with their peers and teachers throughout the teaching period [7]. Therefore, predictive models based on continuous assessments, that are often part of engineering education curriculum, and those that can map out an early learning phase of students in a course, can potentially yield helpful information for academics to implement strategies to improve teaching and learning [8], [9]. The qualitative and quantitative approaches employed to predict student performance are categorised in the human-based and computer-based methods. Examples of human-based methods are those that use a teacher's own judgement [10], [11] and self-reports [12] but computer-based methods often aim to apply statistical and data mining methods to predict a student's performance [3], [13], [14]. Data mining techniques use advanced statistics or machine learning methods, among others, as one of the most widely used approaches for performance prediction e.g., [1], [4], [7], [15], [16]. Whilst these approaches are contributing towards developing evidence-based teaching strategies to advance student progress in study disciplines, there is a need to develop advanced data mining methods [17] that can consider student's continuous assessments, and their joint effects with other forms of ongoing tests. This may be useful to develop early intervention plans to prevent the failure in a course.

In this paper we developed multivariate distribution models utilizing assessment (e.g. quiz, and assignment) to predict a weighted score for engineering mathematics course and determine its influence on the final grade using student performance data and copula models. Copulas have excellent capabilities to consider non-linear dependence structure among variables and have shown good predictive skills in modelling non-normally distributed data in non-education areas [18]–[20]. It is worth noting that despite some attention to copulas in modelling non-linear marginal distribution data, there appears to be a paucity of straightforward approaches that can derive joint distribution functions between the marginal distribution of a set of predictors and a target. This is especially true for the field of education despite such variables playing a key determinant role in student success. It is therefore of prime interest to estimate joint distribution effects of interacting variables, such as quizzes and assignments, and further identify how these variables influence a weighted score to assign a passing grade.

This research paper considers well-established theory of copulas [21] and further builds the original method into a new predictive framework for engineering education decision-making. We apply a new copula approach to the data from a web-based (online) and face-to-face (on-campus) engineering students' performance. Copulas are advantageous in modelling the joint distribution between variables where their marginal distributions and the data features are otherwise relatively separate. We adopt copulas based on their capability to analyse joint dependence structure, and constructing models that are assumption-free and non-parametric. Free from the influence of marginal distribution linear assumptions [22], copulas provide a distinct advantage in probabilistic or conditional estimation considering different predictors, their relative strength or joint features and conditional probability. These features provide flexibility in modelling practical data encountered in fields such as education where variables like assessments, time spent on online learning management systems, and teaching or learning activities used to determine student learning outcomes.

The choice of copula approach in this study is motivated by extensive applications of the method in many advanced modelling areas. Examples include actuarial studies and finance [23]–[25], econometrics and marketing [26]–[28], and agriculture and hydrology [18], [19], [29]–[34]. Copulas were recently used to investigate the influence of climate variability on systemic weather risks, particularly using joint models to maximise the spatial diversification portfolios in insurance industry [29]–[31]. These studies, articulated the benefits of copulas in jointly studying the dependence structure and modelling multivariate predictors/targets. It is imperative to mention that probabilistic models are potent tools that can evaluate the risk by considering the tail distribution of any data, where for example, a set of extreme values are considered realistically and within a Bayesian model framework [35]. This copula approach can therefore extend the capabilities/functionalities of conventional machine learning models that are capable of simulating single data values in a test set after taking the training parameters from a training set. It can also simulate the whole distribution, and tails (or extreme values) that machine learning may offer a limited capability to pursue. Taking the key findings from these studies, we also aver that the continuous assessments for engineering problem solving can take any value, occupy a very low, or a very high score that creates a tail distribution pattern. They could also differ in how a weighted score or a final grade is distributed; so the capability of copulas in studying tail distribution features jointly with a target variable is considerably advantageous in present research.

For the specific case of education and social sciences discipline, copulas have been rather limited, although a study by Vuolo [20] has built such models to simulate the spousal mortality with empirical examples of association between unemployment and suicide rate. That study has considered the joint dependence structure between count (i.e., the number of days of drinking alcohol) in respect to a skewed, continuous variable (i.e., grade point average) and therefore demonstrated the merits of copula models in social sciences. Similarly, copula models may be particularly beneficial in studying a student's performance based on assignments and if such models are developed, the Faculty can adopt them to mitigate risks of student failure in courses well ahead of an examination period [8]. The proposed copula models may be used to assist educators in better preparing students through their learning journeys.

Justifications to adopt the copulas are made with respect to the significant proportion of student learning data, that by the virtue of their social variability, learning patterns and causal factors, is divergent from a normal or Gaussian distribution and free from assumptions used in traditional models. By fitting an appropriate skewed distribution function to student assessment data (e.g., assignments as an input) versus a target (i.e., an examination score) or considering another outcome of interest (i.e., a grade point average as a categorical variable), educators can adopt copula functions to explore the extent of association between these variables. Most importantly, problems where multivariate predictors are used in the education area can also adopt maximum likelihood techniques for prediction without any assumption on the marginal distribution of individual data, and therefore, formulate a variety of predictive models to emulate a target that is linked to a predictor variable [20]. To the best of the authors' knowledge, no prior study has developed copula models to predict engineering mathematics performance, their grades or weighted scores through multivariate continuous assessment data.

The novelty is to develop for the first time, a D-vine quantile regression model to predict engineering mathematics student performance using the specific case of an Advanced Engineering Mathematics course result, and employing several continuous assessment marks and weighted scores used to assign a passing or a failing grade. Advancing and expanding the scope of our earlier machine learning-based study [1] and the others [2-5; 7-16], the proposed D-vine quantile regression model aims to predict the whole distribution within a probabilistic framework rather than the single- or the mean test test values predicted by a conventional machine learning model. The proposed D-vine quantile regression model (see Section IV) has enabled us to perform an accurate and fast prediction with a unique advantage over classical quantile regressions such as avoiding quantile crossing and interaction issues between the covariates. Therefore, our new copula models are constructed in such a way that the non-influential predictors are excluded to reflect a final parsimonious model.

Another major contribution is to adopt D-vine quantile regression model designed with both the parametric (PDVR) and the non-parametric (NPDVR) copula family functions. For a highly robust model, we consider six years of data, over 2013-2018, in ENM2600 (Advanced Engineering Mathematics) that are categorised in terms of face-to-face, or on-campus (ONC) and the web-based, or online study mode (ONL). This work is therefore a pioneering study employing a second-year student learning data (i.e., assignments, quizzes, examination marks and weighted scores) in an Advanced Engineering Mathematics course and aims to generalize the D-vine quantile regression modelling algorithm that typically has discrete student performance data.

As additional contribution and cross-validity of this study, we tested the developed copula models on a lower level, first year engineering mathematics course ENM1600 (Engineering Mathematics) data whose results are also summarized in Appendix A.

To ascertain the accuracy of the D-vine quantile regression model for its skill to predict engineering mathematics student performance, we evaluated this objective method (i.e., PDVR and NPDVR models) against a conventional linear regression (i.e., LR) model simulated for an independent test data. Given the nature of our numeric data, the study adopts parametric estimation skill of the newly designed D-vine copulas with an added contribution utilizing a modified algorithm that accounts for discreteness in data [36].

The rest of the paper describes the properties of copula-based models, material and methodology, results and discussion. Several challenges after the presentation of results are discussed, and a final section provides the conclusions.

II. BACKGROUND OF COPULAS

For brevity, only the most relevant details are presented here; readers may also refer to Appendix B for details. To predict Advanced Engineering Mathematics student performance, we first consider the theorem of Sklar [21] states that for a joint cumulative distribution function (JCDF) $F(x_1, \ldots, x_d)$ of a d-dimensional random variable $(X_1, \ldots, X_d)$, we have a marginal distribution $F_i(x_i)$, $i = 1,\ldots,d$ that defines a copula function C such that

$$F(x_1, \ldots, x_d) = C[F_1(x_1), \ldots, F_d(x_d)]. \quad (1)$$

The joint probability density function (JPDF) is expressed as

$$f(x_1, \ldots, x_d) = \left[\prod_{i=1}^{d} f_i(x_i)\right] c[F_1(x_1), \ldots, F_d(x_d)] \quad (2)$$

where $f_i(x_i)$ is the marginal density and

$$c = \frac{\partial^d C[F_1(x_1), \ldots, F_d(x_d)]}{\partial F_1(x_1) \ldots \partial F_d(x_d)} = \frac{\partial^d C(u_1, \ldots, u_d)}{\partial u_1 \ldots \partial u_d} \quad (3)$$

is the copula density.

The copula model data, denoted as pseudo-data, has a uniform distribution on the interval [0,1] with a conversion procedure known as univariate probability integral transformation. If $F_i(x_i)$ is continuous, the associated function $C : [0,1]^d \rightarrow [0,1]$ is unique. Otherwise, there exist many possible copulas and all would coincide over $RanF_1 \times \ldots \times RanF_d$ where $RanF_i$ denotes the range of F_i [33]. Empirical applications of the copula approach are possible for discrete

TABLE 1. The parametric bivariate copula families with the copula generator function, its parameters, the lower and upper tail dependence coefficients and the relationship between parameters and Kendall's tau (τ) coefficients.

(a) Two Elliptical Family Copulas

Copula Type	Gaussian	Student's t
Bivariate (BV)	$\Phi_\theta\left[\Phi^{-1}(u), \Phi^{-1}(v)\right]$	$T_{\theta,v}\left[T_v^{-1}(u), T_v^{BV-1}(v)\right]$
Generator	$\frac{1}{\sqrt{2\pi}}e^{-t/2}$	$\frac{\Gamma\left(\frac{v+1}{2}\right)}{\Gamma\left(\frac{v}{2}\right)\sqrt{\pi v}}\left(1+\frac{t}{v}\right)^{-\frac{2+v}{2}}$
Range	$\theta \in (-1, 1)$	$\theta \in (-1, 1)$ $\in (2, \infty)$
Tail Dependence	(0, 0)	$2t_{v+1}\left(-\sqrt{v+1}\sqrt{\frac{1-\rho}{1+\rho}}\right)$
Kendall's Tau	$\frac{2}{\pi}arcsin\theta$	$\frac{2}{\pi}arcsin\theta$

(b) Two Archimedean Family Copulas

	Independence	Clayton
Bivariate (BV)	*uv*	$\left[max\left(u^{-\theta}+v^{-\theta}-1, 0\right)\right]^{-1/\theta}$
Generator	*-ln t*	$\frac{1}{\theta}\left(t^{-\theta}-1\right)$
Range	N/A	$\theta \in (0, \infty)$
Tail Dependence	(0, 0)	$\left(2^{-\frac{1}{\theta}}, 0\right)$
Kendall's Tau	0	$\frac{\theta}{\theta+2}$

(c) Archimedean Family - Frank Copula

Bivariate (BV)	$-\frac{1}{\theta}ln\left[1+\frac{\left(e^{-\theta u}-1\right)\left(e^{-\theta v}-1\right)}{e^{-\theta}-1}\right]$
Generator	$-ln\left(\frac{e^{-\theta t}-1}{e^{-\theta}-1}\right)$
Range	$\theta \in \mathbb{R}$
Tail Dependence	(0, 0)
Kendall's Tau	$1-\frac{4}{\theta}\left[1-D_1(\theta)\right]$

(d) Archimedean Family - Gumbel Copula

Bivariate (BV)	$exp\left\{-\left[(-lnu)^\theta+(-lnv)^\theta\right]^{1/\theta}\right\}$
Generator	$(-lnt)^\theta$
Range	$\theta \in 1, \infty)$
Tail Dependence	$\left(0, 2-2^{\frac{1}{\theta}}\right)$
Kendall's Tau	$\frac{\theta-1}{\theta}$

(e) Archimedean Family - Joe Copula

Bivariate (BV)	$1-\begin{pmatrix} (1-u)^\theta+(1-v)^\theta \\ -(1-u)^\theta(1-v)^\theta \end{pmatrix}^{1/\theta}$
Generator	$-ln\left[1-(1-t)^\theta\right]$
Range	$\theta \in (1, \infty)$
Tail Dependence	$\left(0, 2-2^{\frac{1}{\theta}}\right)$
Kendall's Tau	$1+\frac{4}{\theta^2}\int_0^1 tln(t)(1-t)^{2(1-\theta)/\theta}\,dt$

(f) Archimedean Family - BB1: Clayton-Gumbel Copula

Bivariate (BV)	$\left\{1+\left[\left(u_1^{-\theta}-1\right)^\delta+\left(u_2^{-\theta}-1\right)^\delta\right]^{1/\delta}\right\}^{1/\theta}$
Generator	$\left(t^{-\theta}-1\right)^\delta$
Range	$\theta \in (0, \infty)\ \delta\ \delta \in 1, \infty)$
Tail Dependence	$\left(2^{-\frac{1}{\theta\delta}}, 2-2^{\frac{1}{\delta}}\right)$
Kendall's Tau	$1-\frac{2}{\delta(\theta+2)}$

(g) Archimedean Family - BB6 (Joe-Gumbel Copula)

Bivariate (BV)	$1-\left(1-exp\left(-\begin{pmatrix} \left(-ln\left(1-u_1^{-\theta}\right)\right)^\delta \\ +\left(-ln\left(1-u_2^{-\theta}\right)\right)^\delta \end{pmatrix}^{1/\delta}\right)\right)^{1/\theta}$
Generator	$\left\{-ln\left[1-(1-t)^\theta\right]\right\}^\delta$
Range	$\theta \in 1, \infty)\ \delta \in 1, \infty)$
Tail Dependence	$\left(0, 2-2^{\frac{1}{\theta\delta}}\right)$
Kendall's Tau	$1+\frac{4}{\theta\delta}\int_0^1\begin{pmatrix} -ln\left[1-(1-t)^\theta\right] \\ \times(1-t)\left[1-(1-t)^{-\theta}\right] \end{pmatrix}dt$

(h) Archimedean Family - BB7 (Joy-Clayton Copula)

Bivariate (BV)	$1-\left(1-\left[\begin{pmatrix} \left(1-u_1^{-\theta}\right)^{-\delta} \\ +\left(1-u_2^{-\theta}\right)^{-\delta}-1 \end{pmatrix}\right]^{-1/\delta}\right)^{1/\theta}$
Generator	$\left[1-(1-t)^\theta\right]^{-\delta}-1$
Range	$\theta \in 1, \infty)\ \delta \in (0, \infty)$
Tail Dependence	$\left(2^{-\frac{1}{\delta}}, 2-2^{\frac{1}{\theta}}\right)$
Kendall's Tau	$1+\frac{4}{\theta\delta}\int_0^1\begin{pmatrix} -\left[1-(1-t)^\theta\right]^{\delta+1} \\ \times\frac{\left[1-(1-t)^\theta\right]^\delta-1}{(1-t)^{\theta-1}} \end{pmatrix}dt$

(i) Archimedean Family - BB8 (Joe-Frank) Copula

Bivariate (BV)	$\frac{1}{\delta}\left(1-\left[\begin{pmatrix} 1-\frac{1}{1-(1-\delta)^\theta}\left(1-(1-\delta u_1)^\theta\right) \\ \times\left(1-(1-\delta u_2)^\theta\right) \end{pmatrix}\right]^{1/\theta}\right)$
Generator	$-ln\left[\frac{1-(1-\delta t)^\theta}{1-(1-\delta)^\theta}\right]$
Range	$\theta \in 1, \infty)\ \delta \in 0, 1$
Tail Dependence	(0, 0)
Kendall's Tau	$1+\frac{4}{\theta\delta}\int_0^1\left[-ln\frac{(1-t\delta)^\theta-1}{(1-\delta)^\theta-1}\right]$ $\times(1-t\delta)\left[1-(1-t\delta)^{-\theta}\right]dt$

marginal distributions that carefully consider modelling and interpreting the dependence, as highlighted in [37]. For a detailed representation of the mathematics of copulas, readers can consult papers elsewhere e.g., [38] or [39].

A. ELLIPTICAL COPULAS

Gaussian (or Normal) and Student's *t* copula derived from the density function of an elliptical distribution with mean zero and correlation matrix which is expressed as:

$$h_\varphi(x) = |\Sigma|^{1/2}\varphi\left[(x)'\Sigma^{-1}(x)\right]. \tag{4}$$

For every $x \in R^2$ where φ is a generator function and both copulas are symmetric so their lower and upper tail dependence coefficients can be the same (see Table 1).

When the margins of variables are diverse, other measures of association such as Kendall's τ and Spearman's ρ should be used because of the influence from form of the marginal distributions on the correlation. Table 1 provides such a measure, in terms of the value of τ, which is a non-parametric, robust and efficient estimator of the associations for both elliptical and non-elliptical margins [40].

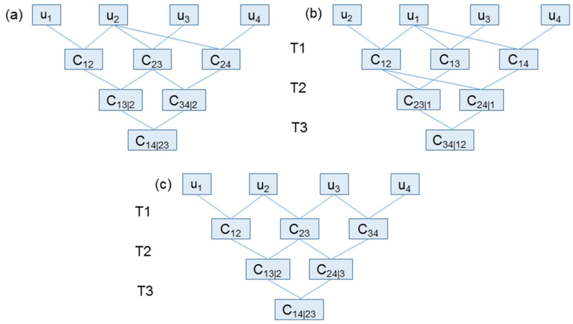


FIGURE 1. (a) Four-dimensional vine copula models constructed as a regular (R)-vine, (b) Canonical C–vine copula structure, and (c) Drawable D–vine copula structures. This study used drawable vine (D-vine). Note that the joint distribution between variables *u* is modelled by copula function *C*. For example, C_{12} is the copula function for *u*1 or quiz mark, and *u*2 or assignment mark; $C_{13|2}$ is the copula function for *C*12 and *C*23 i.e. *u*1 and *u*3 conditioned on *u*2.

B. ARCHIMEDEAN COPULAS

Archimedean copulas (ACs) have a relatively simple form for their construction and therefore resulting in a large variety of copulas within this family. Bivariate ACs are defined as follows [41]

$$C(u, v) = \varphi^{[-1]}\left[\varphi(u) + \varphi(v)\right], \tag{5}$$

where the generator function φ is a continuous strictly decreasing convex function such that $\varphi(1) = 0$ and φ^{1} being a pseudo-inversion. By inserting the generator function in Equation 5, one can derive various copula families, as shown in Table 1.

The two-parameter ACs [42] are from a mixture of two different one-parameter copulas. These mixed copulas can capture different types of dependence, i.e., lower or upper tail dependence or both. For example, the BB7 has one parameter for modelling the lower tail dependence and another for the upper (see Table 1).

C. VINE COPULAS

To apply our method for the specific cases of Advanced Engineering Mathematics, this study adopts a Vine copula method, also known as a pair-copula construction [43] based on the merits that it can overcome the aforementioned limitations. In principle, the vine method constructs joint density in Equation 2 into a sequence product of (conditional) bivariate copula densities, so-called pair-copulas, and its marginal densities so in this study, conditional copulas are used to predict student's passing grades using their continuous assessment marks. Generally, Vine copulas are expressed in three forms: regular (R)-vine, canonical (C)-vine, and drawable (D)-vine copulas. The class of R-vine is still very general and embraces a large number of possible pair-copula decomposition, i.e., $\binom{d}{2} \times (d-2)! \times 2^{\binom{d-2}{2}}$ while the C-vine and D-vine provide a specific way to decompose the density into $d(d-1)/2$ unique copulas.

Figure 1 depicts the construction of a four-dimensional vine copula that includes three trees. To interpret this, consider $T_j, j = 1, 2, 3$ with each tree T_j having 5 - j nodes and 4 - j edges; each edge corresponding to a paired-copula density U as the copula data, i.e., original data that were transformed into a uniform distribution with values in [0,1] using kernel density estimation (non-parametric method, not parametric distribution such as Weibull or Gamma). $U1$ can be Quiz 1, for example, but not necessary, and $U1$ can be any variable depending on the course and the mode of offer as a general copula model. The four-dimensional C-vine structure is generally expressed as

$$\begin{aligned}
&f(x_1, x_2, x_3, x_4)\\
&\quad = f_1(x_1).f_2(x_2).f_3(x_3).f_4(x_4)\\
&\qquad .c_{12}\left[F_1(x_1), F_2(x_2)\right].c_{13}\left[F_1(x_1), F_3(x_3)\right]\\
&\qquad .c_{14}\left[F_1(x_1), F_4(x_4)\right]\\
&\qquad .c_{23|1}\left[F_{2|1}(x_2|x_1), F_{3|1}(x_3|x_1)\right]\\
&\qquad .c_{24|1}\left[F_{2|1}(x_2|x_1), F_{4|1}(x_4|x_1)\right]\\
&\qquad .c_{34|12}\left[F_{3|12}(x_3|x_1, x_2), F_{4|12}(x_3|x_1)\right],
\end{aligned} \tag{6}$$

and the four-dimensional D-vine

$$\begin{aligned}
&f(x_1, x_2, x_3, x_4)\\
&\quad = f_1(x_1).f_2(x_2).f_3(x_3).f_4(x_4)\\
&\qquad .c_{12}\left[F_1(x_1), F_2(x_2)\right].c_{23}\left[F_2(x_2), F_3(x_3)\right]\\
&\qquad .c_{34}\left[F_3(x_3), F_4(x_4)\right]\\
&\qquad .c_{13|2}\left[F_{1|2}(x_1|x_2), F_{3|2}(x_3|x_2)\right]\\
&\qquad .c_{24|3}\left[F_{2|3}(x_2|x_3), F_{4|3}(x_4|x_3)\right]\\
&\qquad .c_{14|23}\left[F_{1|23}(x_1|x_2, x_3), F_{4|23}(x_4|x_2, x_3)\right].
\end{aligned} \tag{7}$$

In this decomposition, the selection of pairwise copula is independent of each other so such paired-copula constructions allow arbitrary types of bivariate copulas to be used in the building blocks and available for applications in high dimensional datasets.

It is imperative to mention that the vine copulas used in this paper is very flexible in modelling asymmetric distribution of data (for example, student performance marks being skewed to a certain value) and tail dependence (e.g., marks being too low, or too high rather being than uniformly distributed). Given the high-dimensional model that we require in this research work, vine copulas were considered to address the limitations of the other methods such as elliptical copulas and ACs [43].

For the case of a D-vine based regression model, $U1$ would actually become V, which is the response variable (i.e., *WS* or *EX*), and the other $U1$, $U2$, $U3$ (in this case study) correspond to $U2$, $U3$, and $U4$ in the 4-dimensional D-vine model. To predict *WS* (i.e., variable V), the variables $U1$, $U2$, $U3$ (in this case study) can be examination score, assignment 1, and quiz 2, respectively, depending on the D-vine structure (or its order) selected via maximum conditional likelihood. Therefore $C12$ for example denotes the bivariate copula of $U1$ and $U2$ (or V and $U1$) and $C23$ is the bivariate copula for $U2$ and $U3$. C13|2 is the bivariate copula for $C12$ and $C23$, i.e., $U1$ and $U3$ conditioned on $U2$.

D. FITTING MARGINAL DISTRIBUTIONS

To develop models for engineering mathematics student performance prediction, we followed a first step in developing copula models by correctly fitting the marginal distributions of student performance marks and weighted scores to attest their uniformity or dis-uniformity. This followed the notion that marginal distributions can be modelled based on parametric or non-parametric methods where parametric techniques are used to fit each variable to a proposed theoretical distribution function (e.g., Normal, Gamma, or Weibull) using maximum likelihoods [44], moment matching [45], quantile matching [46], or goodness-of-fit (GOF) [9] properties. Non-parametric methods use empirical cumulative distribution function or continuous smoothing estimator:

$$\hat{F}\left(x\right)=\frac{1}{n}\sum_{i=1}^{n}K\left(\frac{x-x^{(i)}}{h}\right), \tag{8}$$

where $K\left(x\right)=\int_{-\infty}^{x}k\left(t\right)dt$ and $k\left(\cdot\right)$ is a symmetric probability density function and $h>0$ is a parameter.

E. FITTING COPULAS

As an important consideration for developing models for engineering mathematics student performance prediction, we were mindful that copulas can be selected using several measures e.g., statistical GOF tests or the information-based criteria. GOF can be performed either based on White's information matrix equality [47], [48] or based on Kendall's process [49] that produces test statistics and p-value to reject or accept a parametric copula. Though Akaike information criteria (AIC) and Bayesian information criteria (BIC) [38] do not provide any understanding about the power of the decision rule employed, they allow for an efficient comparison of fitting between different copulas based on single numbers by correcting the log-likelihood for the number of parameters used in a model, i.e., the model with smallest AIC (or BIC) is chosen. Furthermore, these criteria-based methods take less time to compute than GOF tests. In particular, they take the relatively simple forms of $AIC=-2l\left(\theta_n\right)+2k$ and $BIC=-2l\left(\theta_n\right)+klogn$ where k and n denote the number of free parameters and the sample size, respectively. The estimation of the copula parameter θ is described in the next section. Also, the penalty for two-parameter families when using BIC is stronger than when using the AIC [50].

Since the criteria-based methods do not perform a formal GOF hypothesis test, they therefore cannot state whether the copula family with the least AIC/BIC is suitable for the particular case. If the true unknown copula is not among the series of candidates, selecting the copula with the least criteria value may be incorrect. Thus, using these criteria in combination with GOF test is preferred to avoid the misinterpretation for the copula model selection. Alternatively, to reduce computational cost for GOF tests, several graphical tools can also provide useful visual analysis supporting the copula selection such as CDF or lambda plots.

F. COPULA PARAMETER ESTIMATION

This study adopts the most common methods: full (or exact) maximum likelihood (FML) and an inference function for margins (IFM) [51] to estimate the copula parameters.The FML method accords to a method where the likelihood is maximised over the copula parameter and margin parameters simultaneously, and thus also called the one-step ML procedure. The estimated copula parameter $\hat{\theta}$ is acquired by maximising the log likelihood and the log-likelihood function, for example for a bivariate case, is defined as:

$$l\left(\theta\right)=\sum_{i=1}^{n}\log\left\{c\left[F_1\left(x_{1,i}\right),F_2\left(x_{2,i}\right)\right]f_1\left(x_{1,i}\right)f_2\left(x_{2,i}\right)\right\}. \tag{9}$$

Clearly, $\hat{\theta}$ is the global maximizer of $l\left(\theta\right)$ and the asymptotic theory can be applied to both the margins as the copula under standard regularity conditions. Hence, the maximum likelihood estimator converges to a normal distribution with mean zero, i.e., $\sqrt{n}\left(\hat{\theta}-\theta\right)\sim N\left[0,J^{-1}\left(\theta_0\right)\right]$ where θ_0 is the true value and J denotes the Fisher's information matrix. The estimates of the ML parameter can be acquired using a numerical maximisation method. However, this can be computationally difficult for high dimensional models because the parameters of margins and the dependence structure are jointly estimated.

In accordance with the proposed IFM method, the marginal distribution parameters are estimated first by optimising separately each marginal likelihood. Subsequently, the copula parameter is acquired by optimising concentrated likelihood in the second step. Thus, this method is also referred as the two-step ML procedure. Under standard regularity conditions, we also have that $\sqrt{n}\left(\hat{\theta}-\theta_0\right)$ is asymptotically normal with mean zero. The IFM method is found to be as a highly efficient estimator closed to the FML but computationally more attractive compared to the FML [42], [52].

If the margins are estimated non-parametrically using their empirical CDFs, then it results in the semi-parametric (SP) method (Genest, Ghoudi, and Rivest, 1995). Let $\hat{u}=\widetilde{F}_X\left(x_i\right)$ and $\hat{v}=\widetilde{F}_Y\left(y_i\right)$ be the pseudo-data of observations acquired by their empirical CDFs, the unknown copula parameter are estimated by the maximising the pseudo log likelihood as:

$$\hat{\theta}=\sum_{i=1}^{n}\log\sum\left(\hat{u}_i,\hat{v}_i;\theta\right). \tag{10}$$

It can be seen that the joint CDF of (X,Y), $C\left(\hat{u},\hat{v};\theta\right)$ is consistent whether the marginal distributions are known or not. Compared to the FML and IFM (with parametric margins), the SP method allows the margins to take arbitrary and unknown functional forms. A possible shortcoming of the fully parametric models (i.e., FML and IFM methods) is that the copula parameter estimation may be inconsistent even when just one of the margins is mis-specified. The SP is found to performs better than ML and IFM methods when the margins are unknown which is the most frequent case in practice [53].

III. MATERIALS AND METHODS

A. STUDENT PERFORMANCE DATA

To design and evaluate the newly PDVR and NPDVR copula-based models used in the prediction of engineering mathematics student success, this paper has analysed data from a second-year engineering mathematics course (ENM2600 Advanced Engineering Mathematics & and ENM1600 for ENM2600 Engineering Mathematics) used earlier in developing a machine learning model [1]. The data comprised of continuous internal assessments and weighted scores from 2015-2015 used to assign a passing or failing grade. The ENM2600 data had marks for 743 online (ONL) and 716 on-campus (ONC) students, whereas ENM1600 (whose results are included in the Appendix) had marks for 817 ONC and 1299 ONL students generated after a data-cleansing phase that deleted all missing rows/student records. These courses are taught and administered by the School of Mathematics, Physics, and Computing in the Faculty of Health, Engineering, and Sciences at the University of Southern Queensland (USQ) in Australia. Other than being a core component of the engineering curriculum to meet Engineer's Australia program requirements, ENM2600 plays an essential role as a service course for several programs including a Bachelor degree in Engineering, Master of Science, and others. ENM2600 is an updated course from a previous curriculum to satisfy the Australian engineering program accreditation requirements.

In the ENM2600 course, student performance is assessed using two quizzes (marked out of 50), denoted as *Q1* & *Q2*, and three assignments (marked out of 150) denoted as A1, *A2* & *A3*, including a final examination (marked out of 600), denoted as *EX* that generates a weighted score, in %, *WS* to assign a passing grade (*HD, A, B, C*, or *D*). These assessments include topics on mathematical concepts from an introduction to the advanced skills for engineering and surveying professionals (Complex Numbers, Ordinary Differential Equations, Series, Multivariable Calculus, and Linear Algebra). Areas such as Ordinary Differential Equations and Series topics include direction fields, Euler's method, first order separable ODEs, first order and second order linear ODEs with constant coefficients, Taylor and Fourier series. Multivariable Calculus includes representation of functions of several variables, surfaces and curves in space, partial differentiation, optimisation, directional derivatives, gradient, divergence and curl, line integrals of the 1-st and 2-nd kinds, iterated integrals, and Green's theorem. The assessment items (quizzes and assignments) are spread through a 13 week teaching semester and provide an ongoing evaluation of student performance.

In this study, we considered various datasets from the engineering mathematics course. As USQ is renowned for both web-based (online) and on-campus (face-to-face) teaching, in this study, the performance data for engineering mathematics students were taken from ONL or "online" and ONC or "on-campus" offers. All predictive models were built using data over 2013-2018 taking into account two teaching semesters. Before obtaining engineering mathematics students' performance data, an ethical approval (H18RE236) was applied for, and granted by the university's ethics committee in accordance with the Australian Code for Responsible Conduct of Research (2018) and National Statement on Ethical Conduct in Human Research (2017). The project was considered low-risk as it did not collect any student's identifiable information directly.

B. CONSTRUCTING D-VINE REGRESSION MODEL

We applied D-vine based regression modelling methods to predict a response Y (i.e., examination mark or weighted score in engineering mathematics) given the influence of a predictor $X_1, \ldots, X_d$ (e.g., assignment or quiz score in engineering mathematics), with $d \geq 1$. Our D-vine based regression method concurs with literature [54] so here, we present only the main steps in this technique, noting that the prediction is attained via a conditional quantile function of joint distribution of X and Y expressed as

$$q_\alpha (x_1, \ldots, x_d) = F^{-1}_{Y|X_1,\ldots,X_d} (\alpha |x_1, \ldots, x_d), \tag{11}$$

where $\alpha \in (0, 1)$ is the quantile levels of interest.

We followed the notion that with $V = F_Y(Y)$ and $U_i = F_i(X_i)$, the corresponding values $v = F_Y(y)$ and $u_i = F_i(x_i)$ with a transformation using kernel density method (Gaussian kernel) and the plug-in bandwidth to minimise the asymptotic mean integrated squared error so that conditional copula function takes the form of

$$\begin{aligned} F_{Y|X_1,\ldots,X_d} & (y |x_1, \ldots, x_d) \\ &= P\left[F_Y(Y) |F_1(X_1), \ldots, F_d(X_d)\right] \\ &= C_{V|U_1,\ldots,U_d} (v |u_1, \ldots, u_d). \end{aligned} \tag{12}$$

The inverse function is therefore

$$\begin{aligned} F^{-1}_{Y|X_1,\ldots,X_d} & (\alpha |x_1, \ldots, x_d) \\ &= F_Y^{-1}\left[C_{V|U_1,\ldots,U_d} (v |u_1, \ldots, u_d)\right]. \end{aligned} \tag{13}$$

The estimated quantile of the response variable can be obtained as

$$\begin{aligned} \hat{q}_\alpha & (x_1, \ldots, x_d) \\ &= \hat{F}_Y^{-1}\left[\hat{C}^{-1}_{V|U_1,\ldots,U_d} \left(\alpha \left|\hat{u}_1, \ldots, \hat{u}_d\right.\right)\right], \end{aligned} \tag{14}$$

where $\hat{C}^{-1}_{V|U_1,\ldots,U_d} \left(\alpha \left|\hat{u}_1, \ldots, \hat{u}_d\right.\right)$ increases monotonically in α.

This computation requires one to estimate the multivariate copula first. It is noted [54] have suggested fitting a D-vine copula to data $(V, U_1, \ldots, U_d)$ with a fixed order $V - U_{l_1} - \ldots - U_{l_d}$ in such a way that V is the first node in the first tree) with $(l_1, \ldots, l_d)$ as the ordering of d-dimensional D-vine copula as an arbitrary permutation of $(1, \ldots, d)$.

The conditional distribution of the response V given the predictors (U_1, U_2, U_3) is recursively expressed in four-dimensional D-vine with order $V - U_1 - U_2 - U_3$ as

$$\begin{aligned}
&C_{V|U_1,U_2,U_3}\ (v\,|u_1, u_2\ , u_3)\\
&= h_{V|U_3;U_1,U_2}\bigl[C_{V|U_1,U_2}\ (v\,|u_1, u_2\,)\big|C_{U_3|U_1,U_2}\ (u_3\,|u_1, u_2\,)\bigr]\\
&= h_{V|U_3;U_1,U_2}\bigl\{h_{V|U_2;U_1}\\
&\quad\bigl[C_{V|U_1}\ (v\,|u_1\,)\,\big|C_{U_2|U_1}\ (u_2\,|u_1\,)\bigr]\,\big|h_{U_3|U_1;U_2}\\
&\quad\bigl[C_{U_3|U_2}\ (u_3\,|u_2\,)\,\big|C_{U_1|U_2}\ (u_1\,|u_2\,)\bigr]\bigr\}\\
&= h_{V|U_3;U_1,U_2}\ \bigl\{h_{V|U_2;U_1}\ \bigl[h_{V|U_1}\ (v\,|u_1\,)\,\big|h_{U_2|U_1}\\
&\quad(u_2\,|u_1\,)\,]\big|h_{U_3|U_1;U_2}\ \bigl[h_{U_3|U_2}\ (u_3\,|u_2\,)\,\big|h_{U_1|U_2}\ (u_1\,|u_2\,)\bigr]\bigr\}.
\end{aligned}\tag{15}$$

And thus, the conditional quantile function is defined as

$$\begin{aligned}
&C^{-1}_{V|U_1,U_2,U_3}\ (\alpha\,|u_1, u_2, u_3\,)\\
&\quad= h^{-1}_{V|U_1}\ \Bigl\{h^{-1}_{V|U_2;U_1}\ \Bigl[h^{-1}_{V|U_3;U_1,U_2}\ \bigl(\alpha\,\big|h_{U_3|U_1;U_2}\\
&\qquad\bigl(\big|h_{U_3|U_2}\ (u_3\,|u_2\,)\,\big|h_{U_1|U_2}\ (u_1\,|u_2\,)\bigr)\bigr)\,|\\
&\qquad\big|h_{U_2|U_1}\ (u_2\,|u_1\,)\bigr]\,|u_1\,\Bigr\}.
\end{aligned}\tag{16}$$

The conditional copula function has been expressed in terms of nested h-function and its inversion corresponding with the pair-copula, i.e.

$$\hat{C}_{V|U}\ = h_{V|U}\ = \frac{\partial\hat{C}_{VU}\,(v, u)}{\partial u}.\tag{17}$$

As the order of predictors can be arbitrary, it can result several D-vine copula models. Hence, to select a parsimonious model for our study, i.e., the influential predictors can be added into the model, and the order of predictors yielding the most power of predicting the response, the order of the U_i is parametrised and selected via maximum conditional likelihood [8].

The proposed algorithm in this study accords to [54], and it has many advantages in constructing a D-vine copula model as it can automatically choose the influential predictors by ranking them based on their strength of predicting response and thus ignoring any superfluous variables. The method, therefore, automatically overcomes the typical issues of regression such as collinearity, transformation, and inclusion/exclusion of predictors. Furthermore, as mentioned above, the D-vine copula allows flexible modelling of the dependence between the response and the selected predictors.

In Figure 2, we describe the steps in this study. For the case of using parametric copula families described in Table 1 we denoted the model as a parametric D-vine regression model (PDVR). Otherwise, if non-parametric copula families (independence and transformation kernel) were used, then the model was a non-parametric D-vine regression model (NPDVR).

IV. RESULTS AND DISCUSSION

A. EXPLORATORY ANALYSIS

To appraise the performance of parametric D-vine regression, PDVR and non-parametric D-vine regression, NPDVR-based models to investigate student performance and to examine its practicality in Advanced Engineering Mathematics decision-making through probabilistic prediction of student success, we explored causal relationships between continuous assessments. We therefore utilized the three assignments (i.e., A1, *A2* & *A3*), two quizzes (*Q1* & *Q2*), examination scores (*EX*) and the weighted score (*WS*) to evaluate the utility of PDVR and NPDVR, in respect to the linear regression model.

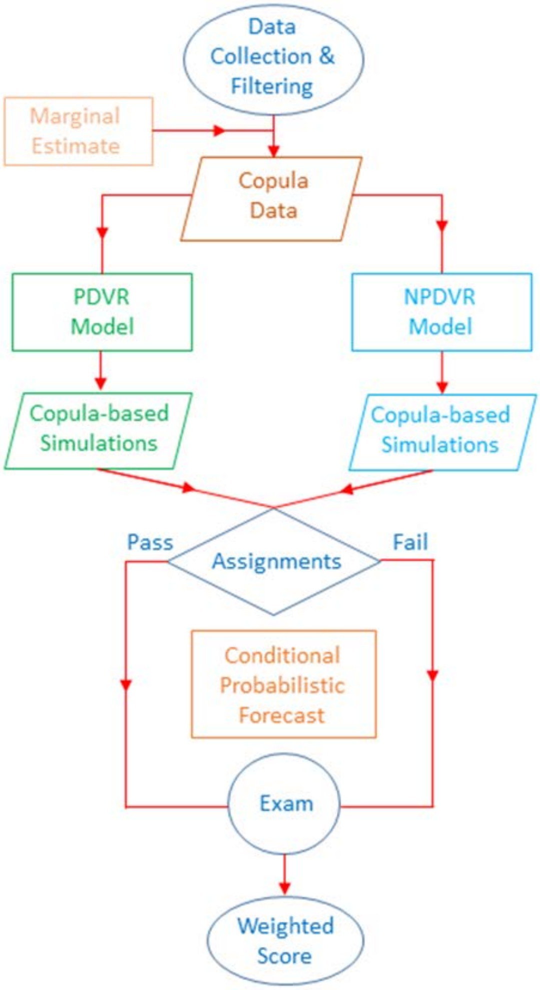


FIGURE 2. Flowchart describing the primary steps required to develop the parametric D-vine regression, PDVR and non-parametric D-vine regression, NPDVR models used to predict student performance in ENM2600 Advanced Engineering Mathematics course.

The results are shown in Figure 3, where Kendall's correlation coefficient τ and the corresponding Kendall's plots are created for the specific case of ENM2600 considering students marks for on-campus and online course modes. Next, we also explored these data in terms of a Kendall plot in accordance with [55] and [33] that attempts to generate information on bivariate copulas equivalent to a quantile-quantile plot approach.

To interpret this, we must determine whether data points lie approximately on the diagonal, and if so, then the two variables can be approximately independent. By contrast, a deviation of the data points away from the diagonal line is expected to indicate the dependence between the two variables. If this happens for the plot representing *Q1* and *WS*

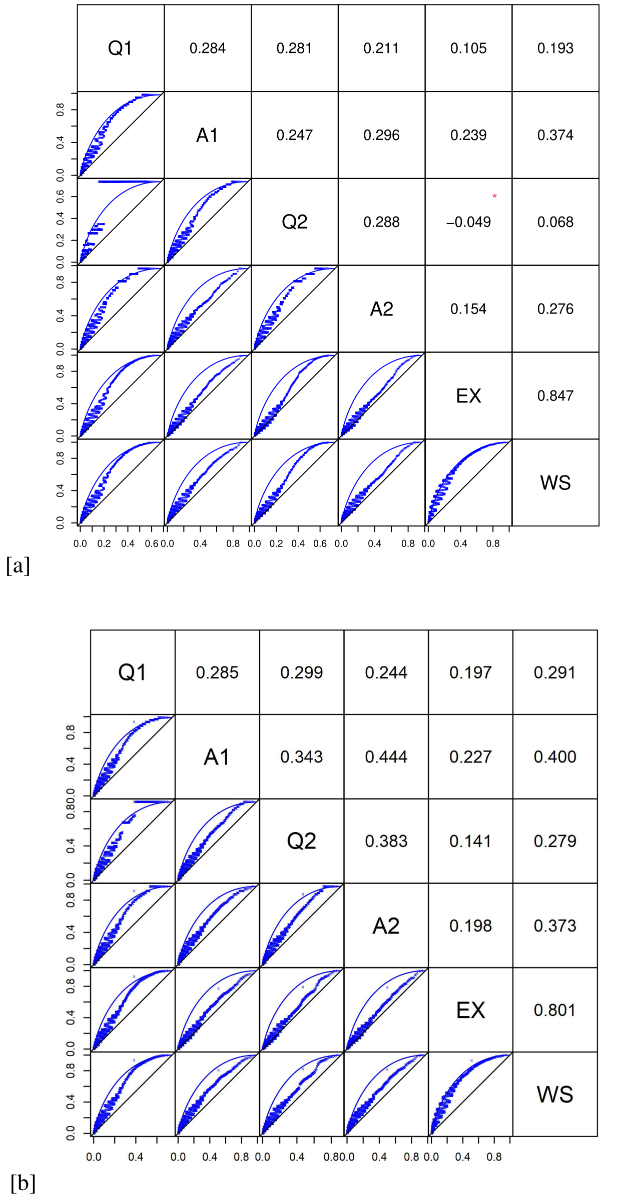


FIGURE 3. **The Kendall's correlation coefficient and Kendall's plot used to explore student performance data in ENM2600 Advanced Engineering Mathematics course for: [a] on-campus (ONC) face-to-face, and [b] online (ONL) web-based student cohorts.**

in such a way that the distance is relatively large, we would observe a stronger degree of dependency among these bivariate data. If the data however are located above the diagonal, one would expect a positive dependence, or vice versa if the data are located below the diagonal for a negative dependence result.

A closer examination of Figure 3 reaffirms the vital importance of examination score (*EX*) in predicting the weighted score (*WS*) for both on-campus and online offers of ENM2600 course. This is evident through a greater weighted proportion of *WS* (versus *Q1*, *Q2*, A1, *A2* and *A3*) required to yield a *WS* value as indicated clearly by high Kendall's correlation coefficient. This result, although not surprising, indicates that the examination mark which constitutes a bulk of course content, is the most dominant indicator of weighted score, and therefore plays a vital role in a passing grade awarded to a student.

When assessed in terms of the Kendall tau plot, the above result is further confirmed where all data points are approximately located on the curve associated with a perfect positive dependence. For example, in case of ENM2600, the degree of association between *EX* and *WS* in ONC student cohort yields a Kendall's correlation coefficient ≈ 0.847, which is only slightly larger than that of the ONL student cohort (≈ 0.80). However, the association between all continuous assessments and *WS* for ONL student cohort is generally stronger than that of ONC cohort. The association between continuous assessments and *EX* for the ONL cohort is also higher than that of ONC cohort, except for the case of A1. Furthermore, it can be construed that the influence of *Q1* and *Q2* on the values of *EX* and *WS* is relatively small in both the ONC and the ONL cohort. By contrast, for ENM1600 (see the Appendix), the degree of association between *EX* and *WS* of the ONL cohort (with Kendall's correlation coefficient ≈0.86) is slightly greater than that of the ONC cohort (≈0.84). It is interesting to see that A1 still has the highest degree of association, with *EX* and *WS* for the ONC student cohort while Q_{1-3} have a greater association with *EX* and *WS*, compared to A_{1-2} for the ONL student cohort.

B. COPULA-BASED PREDICTIVE MODEL OUTCOMES

The accuracy of a resulting copula model by non-parametric fitting of the marginal distributions was checked using graphical analysis.

Figure 4 is a histogram of assignment, A1 that has been overlaid by an empirical density and a density derived from the kernel function estimate. Evidently, the data appears to be appropriately fitted using the proposed kernel and the plug-in bandwidths of this plot that describe these data characteristics. The right side shows the histogram of the probability integral transform that reveal considerable degree of uniformity across the unit interval.

Table 2 represents the most appropriate parametric bivariate copulas selected for all pairs between the continuous assessment marks with the value of *EX* and with the value of *WS* for both study modes. This selection is based on the magnitude of the AIC and the significance level of the statistical independence test that is set to p ≈ 0.05 [50].

The result is jointly attested with a lambda-plot, as per Figure 5. To interpret this, compare the empirical and theoretical λ-functions that indicates that the BB6 and the Gumbel copulas appear to be the optimal model candidates among the various bivariate copula families. This is because these two copulas demonstrate good ability to model the dependence structure between *EX* and *WS* for ENM2600 ONC student cohort. Notably, the BB6 copula yields a smaller AIC and thus, must be selected for further modelling and analysis of student performance.

The present results show that copula-based models are relatively advanced in capturing tail dependence jointly between

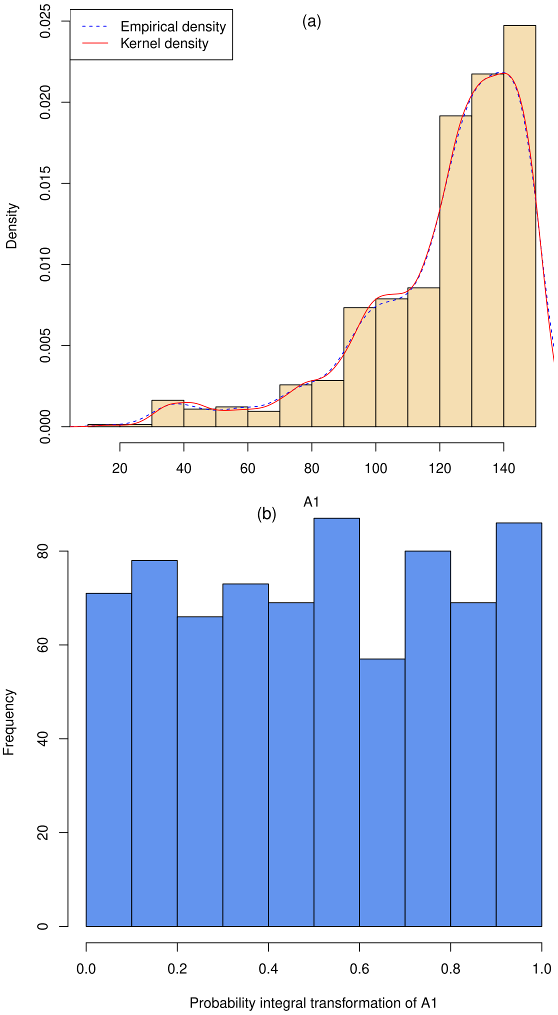


FIGURE 4. An illustrated example of kernel density estimation as required to fit marginal distribution of Assignment 1 (A1) to predict on-campus engineering mathematics ENM2600 student performance and the probability integral transformation.

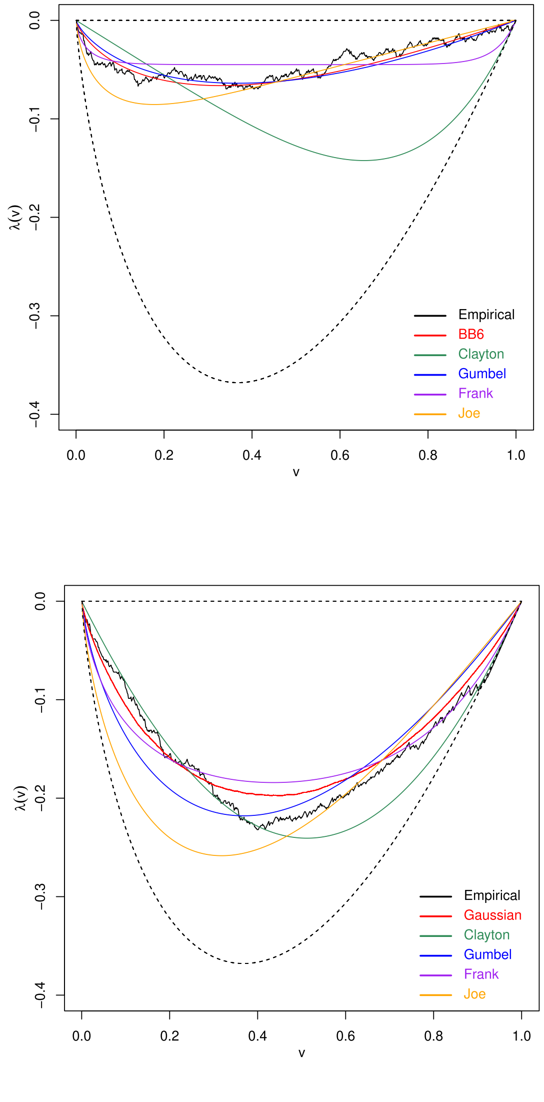


FIGURE 5. Empirical and theoretical λ-functions among different bivariate copula models employed to simulate the joint distribution of examination and weighted scores in ENM2600 Advanced Engineering Mathematics. [a] on-campus (ONC) face-to-face and [b] on-line (ONL) web-based course offers using between Assignment 1 and Assignment 2. The dashed lines represent the limits that correspond to statistical independence (i.e., $\tau = 0$) and co-monotonicity (i.e., $\tau = 1, \lambda = 0$)).

predictors and a target variable. For example, for ENM2600 ONC students, the correlation between continuous assessments and *EX*, or *WS*, are modelled well by copula functions associated with tail dependence (see Table 2). The result obtained implies that students who physically attend classes in on-campus course offer, are more likely to attain a relatively good score in continuous assessments, and as such, will have a plausible chance to obtain a high *EX* and *WS*. For example, the bivariate copula constructed between *WS* and *EX* data (ONL) attained a higher (lower) Log likelihood and AIC of logLik = 975.90 and −1947.80, whereas for ONC, these were 794.05 and −1581.54 respectively. On the contrary, students are likely to have extremely low scores for *EX* and *WS* if they attain very low outcomes in their continuous assessments.

To investigate the case of ONL course offers, we note that elliptical copulas are dominant in modelling the association between pairwise variables, and in particular, between continuous assessments and *EX*. This reflects a weaker dependence in the upper and lower tail between the data pairs of interest. Furthermore, high coefficients of the lower tail dependence between the two important assessments (i.e., A1 and *A2*) and *WS* for both course modes imply that there is a greater probability students will have very low *WS* if they have a very low score for A1 and *A2*. On the other hand, *EX* and *WS* exhibit high upper tail dependence reflecting the fact that students probably have a very high *WS* if they have a very good result for the *EX*. The estimated τ-value (indicated in Table 1) derived from copula models is also found to be similar to the empirical values (i.e., Fig. 6). Our findings offer strong indications of the practical utility of copula models in jointly capturing the dependence structure among the student learning variables.

In this study, we also developed bivariate copula models for a probabilistic prediction of *EX* and *WS* that was

TABLE 2. Parametric bivariate copula model development parameters with each explanatory variable (Quiz Q and Assignment A) paired with response variable (i.e, Examination Score EX and Weighted Score *WS* after best copula selection. The lowest Akaike Information Criteria (AIC), in agreement with the lambda plots shown in Figure 5 were used with A1 selected as the best predictor for *EX* and the *BB*1$_{180}$ copula selected to model the pairwise *EX*-A1 *relationship* showing that A1 is a first predictor added into the proposed D-vine regression model after the response variable *EX*. Note: UTD = upper tail distribution, LTD = lower tail distribution, logLik = log-likelihood, AIC = Akaike Information Criterion, θ_1 and θ_2 = optimal copula parameters and τ = Kendall tau coefficient.

(a) θ_1, θ_2 and τ

Predictor	Copula	Parameters θ_1	 θ_2	Kendall's tau τ
	Face-to-Face (On-campus) Course Offer (Predictor vs. *EX*)			
Q1	Gaussian	0.21	n/a	0.13
A1	**$BB1_{180}$**	**0.42**	**1.07**	**0.23**
Q2	$Clayton_{90}$	-0.16	n/a	-0.07
A2	BB7	1.19	0.17	0.16
	vs. *WS*			
Q1	Clayton	0.41	n/a	0.17
A1	$BB1_{180}$	0.19	1.44	0.37
Q2	Joe_{180}	1.18	n/a	0.09
A2	BB7	1.17	0.66	0.30
EX	**BB6**	**1.63**	**4.20**	**0.82**
	Web-based (Online) Course Offer (Predictor vs. *EX*			
Q1	Gaussian	0.31	n/a	0.20
A1	**Student's *t***	**0.34**	**16.26**	**0.22**
Q2	Gaussian	0.22	n/a	0.14
A2	Gaussian	0.32	n/a	0.21
	vs. *WS*			
Q1	Gaussian	0.45	n/a	0.29
A1	$Gumbel_{180}$	1.62	n/a	0.38
Q2	Gaussian	0.44	n/a	0.29
A2	BB1	0.73	1.16	0.37
EX	**Gumbel**	**4.57**	**n/a**	**0.78**

(b) UTD, LTD, logLik and AIC.

Predictor	Copula	TailS UTD	 LTD	L-likelihood logLik	AIC AIC
	Face-to-Face (On-campus) Course Offer (Predictor vs. *EX*)				
Q1	**Gaussian**	**0.00**	**0.00**	**9.88**	**-17.75**
A1	**BB1$_{180}$**	0.21	0.09	52.09	-100.17
Q2	Clayton$_{90}$	0.00	0.00	2.69	-3.25
A2	BB7	0.21	0.02	28.47	-52.94
	vs. *WS*				
Q1	Clayton	0.00	0.18	36.01	-70.02
A1	BB1$_{180}$	0.08	0.38	138.51	-273.02
Q2	Joe$_{180}$	0.00	0.20	20.34	-38.68
A2	BB7	0.19	0.35	98.92	-193.84
EX	BB6	0.89	0.00	975.90	-1947.80
	Web-based (Online) Course Offer (Predictor vs. *EX*)				
Q1	Gaussian	0.00	0.00	26.04	-50.08
A1	Student's t	0.01	0.01	42.72	-81.45
Q2	Gaussian	0.00	0.00	15.60	-29.19
A2	Gaussian	0.00	0.00	36.73	-71.46
	vs. *WS*				
Q1	Gaussian	0.00	0.00	59.89	-117.78
A1	Gumbel$_{180}$	0.00	0.47	152.76	-303.52
Q2	Gaussian	0.00	0.00	67.90	-133.81
A2	BB1	0.18	0.44	145.03	-286.05
EX	**Gumbel**	**0.84**	**0.00**	**794.05**	**-1581.54**

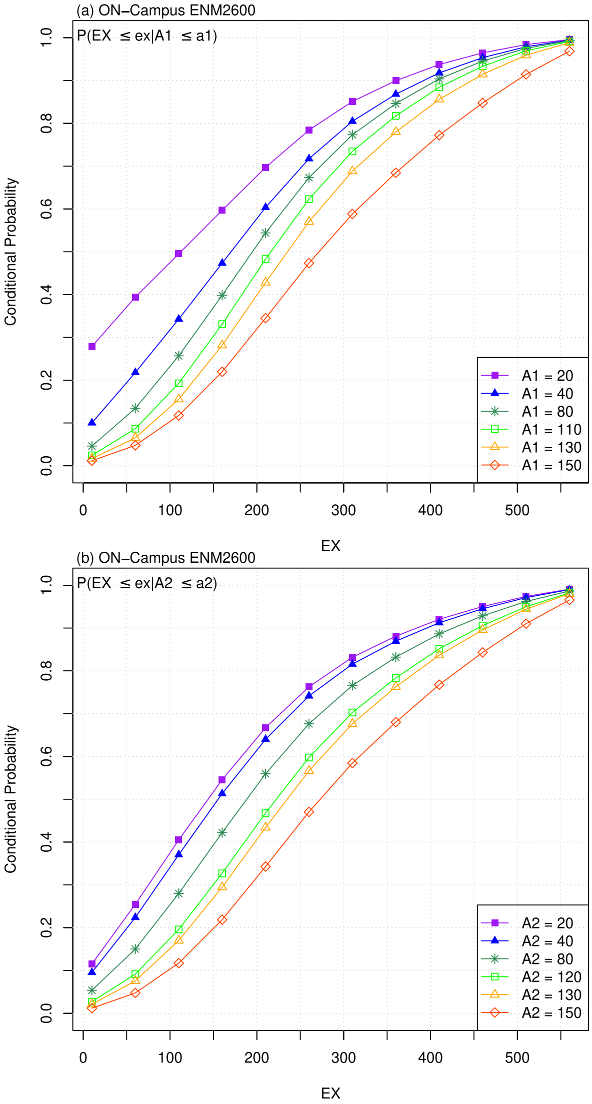


FIGURE 6. Bi-variate Copula Models: Conditional probability plot of examination score (*EX*) being less than or equal to a certain mark, *ex* given that assignment A1 is less than or equal to a certain mark, a1 in the ENM2600 course in both on-campus and online students.[*To interpret this result, consider an on-campus student who has an Assignment 1 score of 150/150 marks, is expected to have a 60% probability to score a 300/600 examination score (or a pass in the examination).*

conditional on student performance in continuous assessments by using D-vine regression model. Figure 6 and Figure 10 in Appendix A are examples of *EX* predictions given the conditional outcomes A1 for both study mode. More precisely, the figure shows the probability that the *EX* is less than or equal to a specific score given A1 is less or equal to a specific score. To observe the difference in the probability over the distribution, the values of conditioning variables (A1) are set to a wide range, from very low score to very high score (i.e., representing different quantiles).

An interpretation of these plots is relatively straightforward. For example, if a student studying ENM2600 gets a low score for *A1* = 20 (out of 150), the probability that a student has *EX* = 250 (out of 600) is approximate 77% without knowing the result of *A2* for both course offer modes. This probability is especially higher or ≈88 % if the student takes the ENM1600 (see the Appendix) but study through the

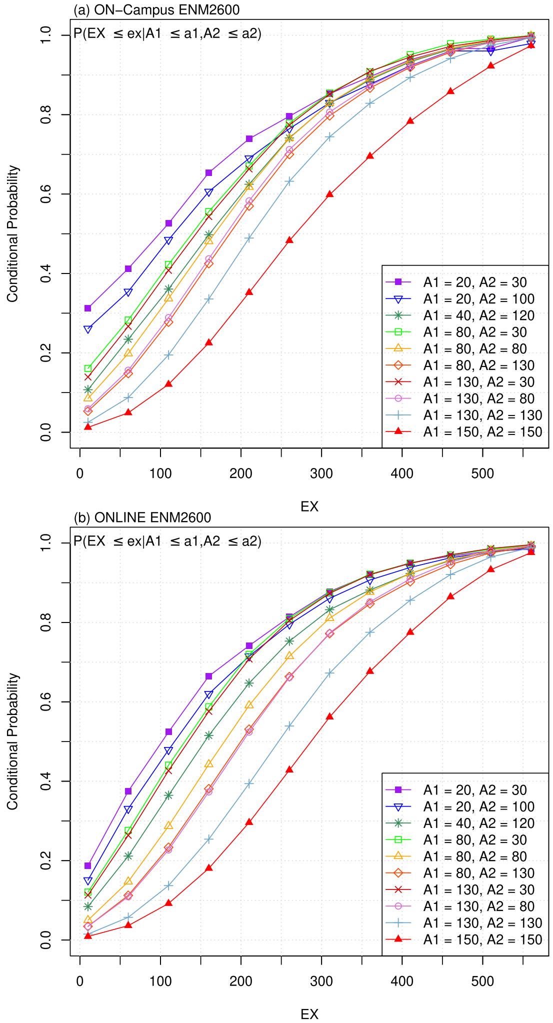


FIGURE 7. Tri-variate Copula Models: Conditional probability plot of *EX* being less than or equal to a certain mark given that A1 and *A2* are less than or equal to a certain mark (a1 and *a2*) for ENM2600 course in both on-campus and online students. [*To interpret this result, consider an on-campus student with* A1 = *20/150 marks and* A2 = *30/150 marks, expected to have 80% probability to score an* EX = *250/600 marks.*

ONL course offerings. Clearly, a higher score for continuous assessment can lead to a lower probability that the student can have *EX* lower than the average value (300/600). It is also worth pointing out that ENM1600 represents an opposite pattern to ENM2600 (see Figure 10 in Appendix A), as indicated by the fact that if a student doing ENM1600 ONL mode has a low score for $A1 = 40$, the probability that the student will have $EX = 250$ is $\approx 66\%$. This is lower than that of $\approx 74\%$ for the ONC course offer. While for ENM2600, these figures are $\approx 78\%$ for ONL mode and $\approx 70\%$ for the ONC course mode.

From an education decision-making perspective, it is of prime interest to our study to see how the predicted *EX* can vary given the joint effect of *A1* and *A2* as this information can be practically useful in investigating the relative contribution of the students' continuous learning towards their final examination. This can be done by extending the bivariate copula models to the case of multivariate copula models. For the tri-variate copula models, Figure 7 for the case of ENM2600, (and Appendix A for the case of ENM1600), are illustrations showing the probability of an *EX* being less than or equal to a specific score, *ex* given *A1* and *A2* is less than or equal to a specific score *a1* and *a2*. For example, students taking ENM2600 ONC course with low scores in both $A1 = 20$ and $A2 = 30$ are likely to have a probability of $\approx 84\%$ to get *EX* score lower than the average. This probability is slightly higher for ONL students as 87%, and these results are also confirmed for the case of ENM1600 (see Appendix A).

For the bivariate model, students studying under the ONC (ONL) course mode who have also attained, for example, $A1 = 80$ or $A1 = 60$ will have estimated probability of $\approx 76\%$ (74%) or $\approx 77\%$ (70%) to attain $EX = 300$ (i.e., a borderline pass) for ENM2600 or ENM1600 (see the Appendix), respectively. However, when lower assignment marks are considered, for example $A2 = 30$ or $A2 = 20$, these figures elevate to $\approx 84\%$ (86%) or $\approx 83\%$ (83%). These findings are expected when a student attains an average score for *A1* and also get the lowest score for *A2*, in which the probability of these students will have a low score for *EX* is higher. In addition, students who have the same performance in $A1 = 130$ for ENM2600 or $A1 = 100$ for ENM1600 can have a higher chance to attain a high *EX* score if they have better results in *A2*.

To further corroborate these findings, we note that, for both study modes in ENM2600 course, students with a good score in *A1* but a low score in *A2* (e.g., $A1 = 130$ and $A2 = 30$) are expected to have a higher probability of getting an *EX* that is lower than a specific threshold, compared to those with low score in *A1* but a high score in *A2* (e.g., $A1 = 40$ and $A2 = 120$). Furthermore, the conditional probability of the ONL course mode is more spread than that of the ONC course mode. This implies that there appears to be a larger difference in the probability of achieving an *EX* score between the student groups who have low scores and those who have good scores in both *A1* and *A2*.

In the next stage, the performance of newly developed copula-statistical predictive models was evaluated by splitting the entire dataset into two separate parts: one for training and another for testing purposes so the generalisation skill of the model can be benchmarked. To implement this strategy, $\approx 25\%$ of the data are randomly selected for testing and $\approx 75\%$ for building these models. This procedure is also repeated 100 times to account for any stochastic variations among input and target sets. Note that this newly proposed algorithm, elaborated in Figure 2, was applied to select the most parsimonious D-vine copula model in each of the training phases.

Table 3 and the material in Appendix A, summarize the most optimal PDVR and NPDVR models was built for the prediction of *EX* and *WS* based *cll* and *cAIC* using continuous assessment marks as the predictors. Evidently, the NPDVR models appear to exhibit a greater degree of parsimonious behaviour relative to the PDVR model. For example, this model utilizes fewer predictor variables to produce the same quality student performance predictions.

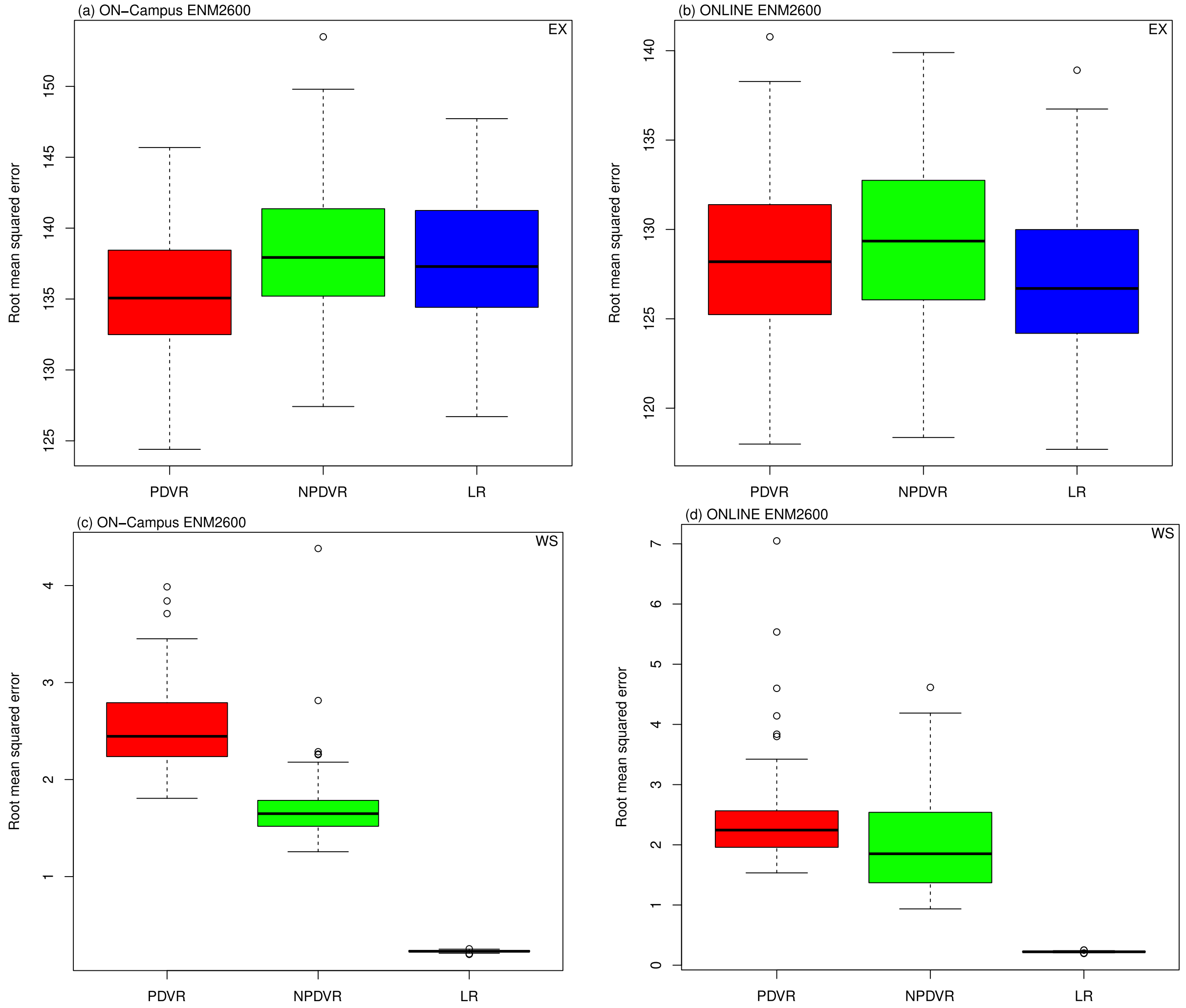


FIGURE 8. **Boxplots showing the root mean squared error produced in the prediction of *EX* and *WS* using an ensemble of 100 simulations using the proposed parametric D-vine regression (i.e., PDVR) and the non-parametric D-vine regression (i.e., NPDVR) models against a traditional linear regression (LR) method.**

In Table 4, we show the selected D-vine copula model created with joint structure *EX*-*A1*-*Q2*-*A2*-*Q1* for ENM2600 ONC student cohort. In the first tree, the edge 1, 2 denotes the pair-copula between the response *EX* and predictor *A1* (i.e., the most influential variable) and the corresponding copula function, which is constructed through the survival BB1 (BB1180) algorithm. The next edge 2, 3 is the pair-copula between *A1* and *Q2*, and so forth. The importance of each predictor added into the copula model in each step is also included for comparison purposes. The results indicate the importance of predictor *EX* to the response *WS* variable. Interestingly, the performance of assignments in ENM2600 appears to have the most influence on the *EX*, except for ENM1600 ONL (see Appendix A) where quiz scores are more important to predict *EX*. The cause of this discrepancy is not clear yet, but plausible reasons could include the difference between the two courses (i.e., advanced versus intermediate) in terms of level of complexity of problems in quiz/assignments, or others that warrant a further comparative investigation.

The results of the predicted mean values of *EX* and *WS* in the testing phases are presented in Figure 8 and the materials in Appendix A, together with traditional method using linear regression (LR) are shown for comparison. The predicted mean values are obtained by setting the quantile level $\alpha \in (0, 1)$ in 100 repetitions of the model. The box plot reflects the stochastic property of three regression models at different quantiles represented by the values of the median, interquartile range (IQR) (i.e., from 25th percentile *Q1* to 75th percentile Q3, the minimum (*Q1* − 1.5 × IQR), maximum (Q3 + 1.5 × IQR) and the outliers.

In the above, we also show the root mean square error (RMSE) indicating that when there is a very high association (or high correlation coefficient) between response and predictors (i.e., *WS* and *EX* in this case), the LR model yields a better prediction than the vine copula-based model. However, in case that associations between response and predictors (i.e., *EX* and continuous assessments) has larger scatter or have a low correlation coefficient, the vine copula-based models provide a very competitive advantage, performing

TABLE 3. The optimal combination of predictors for the proposed parametric (PDVR) and non-parametric (NPDVR) D-vine regression models for each of the target target (i.e., *EX* & *WS*) employed to predict student performance in Advanced Engineering Mathematics ENM2600 course for the face-to-face (on-campus) and web-based (online) students. [Note that the *EX* or the *WS* target is located in the first node of the first tree and the predictors *Q* and *A* are added successively according to the conditional log-likelihood (*cll*) and the corrected Akaike Information Criteria (*cAIC*) values.]

(a) PDVR Copula Model

Predictor Variable Input Combination PDVR Method	*cll*	*cAIC*
Face-to-Face (On-campus) Students		
Target* = *EX		
EX-A1	-4615.88	9235.76
EX-A1-Q2	-4602.74	9211.47
EX-A1-Q2-A2	-4588.43	9186.86
EX-A1-Q2-A2-Q1	-4586.03	9184.07
Target* = *WS		
WS-EX	-2088.97	4181.95
WS-EX-A1	-1733.51	3473.03
WS-EX-A1-A2	-1487.12	2984.24
WS-EX-A1-A2-Q2	-1372.53	2759.07
WS-EX-A1-A2-Q2-Q1	-1342.99	2703.98
Web-Based (Online) Students		
Target* = *EX		
EX-A1	-4436.79	8877.58
EX-A1-Q1	-4427.38	8860.75
EX-A1-Q1-A2	-4421.47	8850.94
Target* = *WS		
WS-EX	-2200.63	4403.26
WS-EX-A1	-1752.53	3509.06
WS-EX-A1-A2	-1404.71	2817.43
WS-EX-A1-A2-Q2	-1288.51	2589.03
WS-EX-A1-A2-Q2-Q1	-1241.11	2498.22

(b) NPDVR Copula Model

Predictor Variable Input Combination NPDVR Method	*cll*	*cAIC*
Face-to-Face (On-campus) Students		
Target* = *EX		
EX-A1	-4605.59	9246.00
EX-A1-A2	-4582.04	9223.56
EX-A1-A2-Q2	-4554.02	9205.31
Target* = *WS		
WS-EX	-2037.07	4154.78
WS-EX-A1	-1603.74	3326.86
WS-EX-A1-A2	-1208.15	2569.48
WS-EX-A1-A2-Q2	-1046.93	2276.95
WS-EX-A1-A2-Q2-Q1	-935.10	2075.19
Web-Based (Online) Students		
Target* = *EX		
EX-A1	-4423.34	8880.96
EX-A1-A2	-4390.56	8841.46
EX-A1-A2-Q1	-4376.58	8834.61
Target* = *WS		
WS-EX	-2142.33	4356.52
WS-EX-A1	-1647.22	3405.31
WS-EX-A1-A2	-1162.12	2479.11
WS-EX-A1-A2-Q2	-889.67	1971.32
WS-EX-A1-A2-Q2-Q1	-732.19	1682.92

TABLE 4. An Illustrated example of the proposed PDVR model (see Table 3) employed to predict the *EX* values using the student assessments as the predictors for ENM2600 face-to-face (on-campus) students. Note that the copula parameters used are as per Table 2. [To interpret this result, consider Tree 1, for example, where Edge 1,2 denotes the bivariate copula between *EX* and *A1* whereas in Tree 2, the Edge 1,3;2 denotes the copula between the *EX* and the *A2* conditioned on the values of *A1*.]

(a) Tree 1

Edge	4, 5	3, 4	2, 3	1, 2
Variable	*A2*, *Q1*	*Q2*, *A2*	*A1*, *Q2*	*EX*, *A1*
Copula	Gaussian	$BB8_{180}$	$BB8_{180}$	$BB1_{180}$
θ_1	0.40	1.70	1.53	0.42
θ_2	n/a	0.97	0.98	1.07
τ	0.26	0.25	0.21	0.23
UTD	0	0	0	0.21
LTD	0	0	0	0.09

(b) Tree 2

Edge	3, 5, 4	2, 4, 3	1, 3, 2
Variable	*Q2*, *Q1*, *A2*	*A1*, *A2*, *Q2*	*EX*, *Q2*, *A1*
Copula	Joe	BB7	$Clayton_{270}$
θ_1	1.72	1.25	-0.30
θ_2	n/a	0.28	n/a
τ	0.29	0.22	-0.13
UTD	0.50	0.26	0
LTD	0	0.08	0

(c) Tree 3

Edge	2,5,3,4	1,4,2,3
Variable	*A1*, *Q1*, *Q2*, *A2*	*EX*, *A2*, *A1*, *Q2*
Copula	Gaussian	$BB7_{1}80$
θ_1	0.25	1.07
θ_2	n/a	0.16
τ	0.16	0.11
UTD	0	0.01
LTD	0	0.09

(d) Tree 4

Edge	1, 5, 2, 3, 4
Variable	*EX*, *Q1*, *A1*, *Q2*, *A2*
Copula	Gaussian
θ_1	0.10
θ_2	n/a
τ	0.06
UTD	0
LTD	0

much better than the LR model. These results reflect the distinct nature of each model. The LR model describes the best fitting by minimising the deviation between data points and the mean value while the copula model fully capturing the entire dependence structure, including the tail dependence. This dependence structure-based model, together with the conditional probabilistic-based model (Figure 6 and 7) is perhaps, a distinct advantage of the developed copula-based methods, in respect to the linear regression, or another traditional predictive model.

V. LIMITATIONS, FURTHER INSIGHTS, AND FUTURE SCOPE

Although several types of continuous assessment data were considered to evaluate the students' overall performance through their weighted scores leading to a grade, this study has some limitations that should be the subject of a further independent investigation. One such limitation was that we did not consider lurking variables, external and inter-related factors such as a student's gender (male or female), attitude, age (whether a student is mature aged, marital or school leaver status), socio-economic advantage or disadvantage (rural and urban), race (white, black, and Hispanic), household parental structure (biological parents, single parent, and other structure), first in the family to attend university, and the proper prerequisite knowledge to learn university mathematics, which potentially influence the weighted scores and the grade. There is a plausible indication that these factors can possibly act as barriers to the student's participation, access to higher education, retention and overall success at university [56], [57]. Some recent studies are showing the great relevance of such causal factors related to successful achievement of students and how these can affect the overall grades at university [58]–[60].

Based on the success of copulas as a pathway to model student performance for educational decision-making, information on external and other casual factors can be pooled into a multivariate copula model to directly explore the compound influence on the variables of interest. Such factors can be adopted to model the marginal distribution before they are actually coupled in joint distribution model. For example, Vuolo (2017) investigated the association between a GPA and student's alcohol usage by using copula methods where marginals were modelled with several predictors e.g., gender, community type and paid work. In that study, the advantage of copula methods in modelling joint probability of GPA and alcohol usage was highlighted. Contini *et al.* [58] investigated potential knowledge gaps based on mathematics students' test scores using the consideration of gender differences in schools under a STEM discipline in Italy. The findings showed the influence of gender on test results, in which the girls systematically under-performed the boys.

Another matter of concern is that the fitted marginal distribution or copula functions by parametric or non-parametric method can yield uncertainties in the final predictive model, thus influencing the results. Examples of these uncertainties are essential error sources that are derived from the data source itself, the nature of the methods used for estimations and the model selection based on a statistical approach (i.e., goodness-of-fit test). The estimate of marginal distributions or copula parameters depends on the observation period [61]. Therefore the dependence structure within univariate or multivariate distribution may vary with the data length, leading to differences in selecting margins and copulas.

It should also be noted that marginal distributions are fitted parametrically using a range of methods, e.g., maximum likelihood, moment matching, quantile matching, maximizing goodness-of-fit estimation or minimizing distance estimation [62]–[65]. Clearly, the best fitted distribution selected for any variable may be different depending on the method. On the other hand, marginal distributions can also be fitted non-parametrically, for example, using kernel density estimators as in this study. However, this method relies on the selection of the density function and plug-in bandwidth parameters, lower and upper bound and the degree of the polynomial (e.g., log-constant, log-linear or log-quadratic fitting) [66], [67]. These selections may lead to different results of the marginal fitting process, and thus, contribute to uncertainty. Copula parameter can be estimated using different approaches, such as fully parametric, semi-parametric to non-parametric methods, which potentially generates similar problems, mentioned above for fitting the marginal distributions.

The incorporation of a purely statistical approach can lead to potential issues where some of the copula parameters may equally fit the goodness-of-fit tests [61], [68], but they may carry errors due to the estimator and thus the overall accuracy of the simulated data can be confounded. This problem can have impacts on the process of finding a unique combination of parameters, which are realistically better with the others.

One combination of copula parameters may either be superior to the others based on the respective statistical goodness-of-fit tests or inferior in regard to another statistical measure. For example, when a copula family is chosen using the BIC criteria, the penalty for two-parameter copulas (e.g., Student's *t*, BB1, BB6, etc.) may be larger than that based on AIC [50]. To overcome these issues, we require further examination to reduce complications in the selection of best copula model along with the best set of parameters of the optimal copula. We thus aver that such error sources may contribute to uncertainty in copula models, so the choice of a good copula function cannot be overstated [69], [70]. Furthermore, in future studies, researchers can use copulas to generate a larger number of inputs for a machine learning model to resolve the student performance data shortage problems and the larger data in machine learning models. Hybrid copula models whereby distribution functions can be used for this purpose. Developing such types of algorithms was beyond the scope of this study and therefore could be a promising direction for future research and awaits another independent study.

Finally, as the data for this study was limited to the 2015-2018 period, a direct comparison of any 'new' student performance data after 2019 (i.e. post-COVID-19 period) with this pre-COVID-19 is impossible, but such a study would also be an interesting endeavour to pursue subject to the availability of such new data and the consistency among the assessments. While the comparison of new data by means of a confusion matrix could be a useful research, this was beyond the scope of the present study given that ethics approval for such data post 2019 is required. Furthermore, the changes in examination format being online-only after COVID-19, as opposed to invigilated exams pre-COVID-19, prohibit a direct comparison of these models acting as an obstacle in pursuing this objective, and therefore will await another separate study.

VI. CONCLUSION

This research, extending the earlier machine learning-based study [1], has incorporated a new copula-based modelling method to examine the influence of continuous assessment scores on the weighted score in engineering mathematics students that lead to a successful grade in the first- and second-year engineering mathematics courses. To advance the earlier research work, this study has built new methods to predict the whole distribution of weighted scores including the tail distribution representing the low and high scores within a probabilistic framework. To do this, a D-vine regression model was constructed and assessed with several predictor datasets for on-campus and online student cohorts in ENM2600 course offers. A cross-validation of the copula model applied for another lower level course, ENM1600 (see Appendix A) was also performed. Using the case of Advanced Engineering Mathematics, and the lower level engineering mathematics courses, the efficacy of the copula models in predicting engineering mathematics student success using continuous assessment data, over 2013–2018, was demonstrated. While this study was motivated by earlier work [1] that developed extreme learning machine, random forest and Volterra models, the added capability of copulas to predict the joint (or whole) distribution targets including the tail distribution and extreme values (see Figures 6-7) made a significant contribution to knowledge compared to the earlier work. Due to the nature of the copula method suited for probabilistic predictions (Figure 6-8) rather than point-based, single-value target predictions as shown in [1], a direct comparison between that machine learning and copula method was not possible.

The results showed that quizzes and assignment marks could be jointly modelled to produce examination scores and weighted scores. Statistical and visual analysis of predicted and real datasets indicated significant benefits of the newly developed D-vine copula models to capture the dependence structure between the predictor and target variables. Most importantly, the ability of copula-based models to correctly describe the dependence in lower and upper bounds, corresponding to very low and high scores, respectively, showed its practical usefulness in the engineering education, particularly in understanding the ongoing learning needs of future engineers that affect their assignment or other marks ahead of their examination period and to reflect with their unique learning styles and the required early interventions needed to reduce the risk of failure. With some modifications, the copula model methods generated in this study may be adopted in other discipline areas where the performance of students need to be predicted ahead of their examination times to improve teaching and learning practices.

Using prior information from internal assessments on student performance, the course instructors and the academic Faculty can develop certain remedial measures for students who secure relatively low marks in internal assessment and quizzes to prevent their failure in the final examination, and even in the overall course. It is important to mention that in the context of the present study, quiz 1 and assignment 1 are given relatively early into the semester i.e. weeks 3-6, and therefore, the remedial measures could include early interventions based on modelling performance for the final examination. Furthermore, quiz 2 and assignment 2 are normally ahead of the examination period (between weeks 7 to 13, which can be used to develop further remedial measures to prevent a poor performance in the examination. Specific examples of remedial measures could include more one-to-one support, amending and balancing the depth and the level of difficulty of the final examination, inclusion of more appropriate content that are tested early in quizzes or assignments, etc. These measures, however, would depend on the resource availability, so the copula models can act as early indicators of such resource needs and the exact remedial measures that depend on the particular course and academic institutions.

APPENDIX A
CROSS-VALIDATIONS WITH ENM1600 ENGINEERING MATHEMATICS COURSE

Using D-vine copula models developed for Advanced Engineering Mathematics student performance predictions, further testing and cross-validations were performed on another course ENM1600, which was a lower level engineering mathematics course at the University of Southern Queensland Australia.

A. MODEL DEVELOPMENT - ENM1600 ENGINEERING MATHEMATICS

In this section, the proposed model development parameters for copulas are shown in terms of the optimal combination of predictors against the target variable.

1) PARAMETRIC D-VINE COPULA

2) NON-PARAMETRIC D-VINE COPULA

B. KENDALL CORRELATION AND KENDALL PLOTS

The Kendall correlation coefficients and Kendall tau plot is used to demonstrate the association between predictors ($A1$, $A2$, $A3$, EX) versus the target (WS) in the problem of predicting student performance in ENM1600.

C. BI-VARIATE COPULA MODEL CONDITIONAL PROBABILITY PLOT

In Figure 10, as show the probabilistic prediction of examination score over [0, 600] conditional upon assignment $A1$ = [20, 40, 60, 80 & 100] out of 150 total marks.

D. TRI-VARIATE COPULA MODEL CONDITIONAL PROBABILITY PLOT

In Figure 11, as show the probabilistic prediction of examination scores over [0, 600] that are conditional upon joints effect of assignment $A1$ = [30, 60, 80, 100] & $A2$ = [10, 20, 40, 50] out of 150 total marks.

APPENDIX B
THEORY ON MULTIVARIATE COPULA MODELS

A. ELLIPTICAL COPULAS

Figure 12 shows examples of the simulated JCDF and JPDF for bivariate Gaussian copula given different levels of the association parameters where the probability is evenly distributed across all values of both marginal distributions in r the case of low relationship ($\tau = 0.15$). When the

TABLE 5. The optimal combination of predictor variables employed to simulate the target *EX* and *WS* for the ENM1600 course using the proposed parametric D-vine regression (i.e., PDVR) model. *To interpret this result, consider the target* EX *or the* WS *that is located in first node of the first tree and predictors* Q *and* A *are added successively according to conditional log-likelihood (cll) and corrected Akaike Information Criteria (cAIC).*

(a) On-campus (face-to-face) students

Predictor Input Combination	*cll*	*cAIC*
ONC Student Cohort		
Target = *EX*		
EX-A1	-5000.41	10004.82
EX-A1-Q1	-4986.20	9978.40
EX-A1-Q1-Q2	-4979.25	9966.49
Target = *EX*		
WS-EX	-2299.28	4602.55
WS-EX-A3	-1915.66	3837.32
WS-EX-A3-Q2	-1695.89	3401.79
WS-EX-A3-Q2-A1	-1588.02	3190.04
WS-EX-A3-Q2-A1-Q1	-1519.86	3055.72
WS-EX-A3-Q2-A1-Q1-A2	-1474.01	2968.01

(b) Online (web-based) students

Predictor Input Combination	*cll*	*cAIC*
ONC Student Cohort		
Target = *EX*		
EX-Q1	-7787.56	15579.12
EX-Q1-Q2	-7733.28	15472.56
EX-Q1-Q2-A1	-7713.21	15434.42
EX-Q1-Q2-A1-A3	-7700.28	15410.56
Target = *EX*		
WS-EX	-3512.44	7026.88
WS-EX-A3	-2937.78	5879.56
WS-EX-A3-Q2	-2501.88	5009.75
WS-EX-A3-Q2-Q1	-2263.64	4535.28
WS-EX-A3-Q2-Q1-A1	-2088.59	4189.18
WS-EX-A3-Q2-Q1-A1-A2	-1992.45	4000.90

relationship is relatively strong ($\tau = 0.60$), the highest probability is observed along the primary diagonal as the concordance increases, meanwhile a considerable discordance exists across the lowest and highest values. A negative association, for example, appears in a similar manner, but in the opposite corners.

B. ARCHIMEDEAN COPULAS

Table 1 introduces the most common one- and two-parameter ACs tested in this study. In Figure 13, we show the JCDF and JPDF for Clayton, Gumbel, Frank and Joe copulas and also their association parameters corresponding to the same τ value of 0.50. Notably, each copula clearly represents different dependence structures over the joint distribution representation.

The Clayton copula appears to be the most useful in modelling the lower dependence while Gumbel and Joe copulas capture the upper dependence. Frank, like the Gaussian, is a symmetric copula accounting for positive and negative associations, as well as the concordant parts, however, they are lighter in upper tails. The Independence copula can be used to check the independence between variables where copula do not rely on associated parameters as well as τ values.

TABLE 6. As in Table 5 but for non-parametric D-vine regression (i.e., PDVR) model.

(a) On-campus (face-to-face) students

Predictor Input Combination	*cll*	*cAIC*
ONC Student Cohort		
Target = *EX*		
EX-A1	-4970.75	9976.04
EX-A1-Q1	-4933.38	9931.72
EX-A1-Q1-Q2	-4905.84	9907.01
EX-A1-Q1-Q2-A3	-4877.75	9896.20
Target = *EX*		
WS-EX	-2277.63	4636.69
WS-EX-A3	-1841.13	3803.07
WS-EX-A3-Q2	-1524.05	3205.57
WS-EX-A3-Q2-A1	-1295.14	2775.49
WS-EX-A3-Q2-A1-Q1	-1143.12	2500.62
WS-EX-A3-Q2-A1-Q1-A2	-1007.72	2258.31

(a) Online (web-based) students

Predictor Input Combination	*cll*	*cAIC*
ONC Student Cohort		
Target = *EX*		
EX-Q1	-7773.44	15593.83
EX-Q1-A3	-7683.42	15450.37
EX-Q1-A3-Q2	-7636.12	15391.41
EX-Q1-A3-Q2-A1	-7605.14	15358.88
Target = *EX*		
WS-EX	-3480.07	7047.45
WS-EX-A3	-2814.29	5768.53
WS-EX-A3-Q2	-2316.23	4813.70
WS-EX-A3-Q2-A1	-2015.58	4245.81
WS-EX-A3-Q2-A1-A2	-1771.87	3789.82
WS-EX-A3-Q2-A1-A2-Q1	-1558.91	3395.56

APPENDIX C
CONSTRUCTION OF MULTIVARIATE COPULA MODELS

In this paper, the multivariate elliptical copula, extended from bivariate function, takes the form of an inversion of the Sklar's theorem. The multivariate Gaussian copula is defined as:

$$(u_1, \ldots, u_d; \rho) = \phi_\rho\left[\phi^{-1}(u_1), \ldots, \phi^{-1}(u_d)\right], \quad (18)$$

and the multivariate Student's t copula:

$$C(u_1, \ldots u_d; \rho) = T_{\rho,v}\left[T_v^{-1}(u_1), \ldots, T_v^{-1}(u_d)\right] \quad (19)$$

where ρ denotes a symmetric, positive definite matrix with elements in the diagonal equal to one. ϕ_ρ and $T_{\rho,v}$ are the standardised multivariate normal Student's t distribution, respectively, with correlation matrix ρ and v degrees of freedom. Although these elliptical members are generally possible to capture a wide range of dependence, including

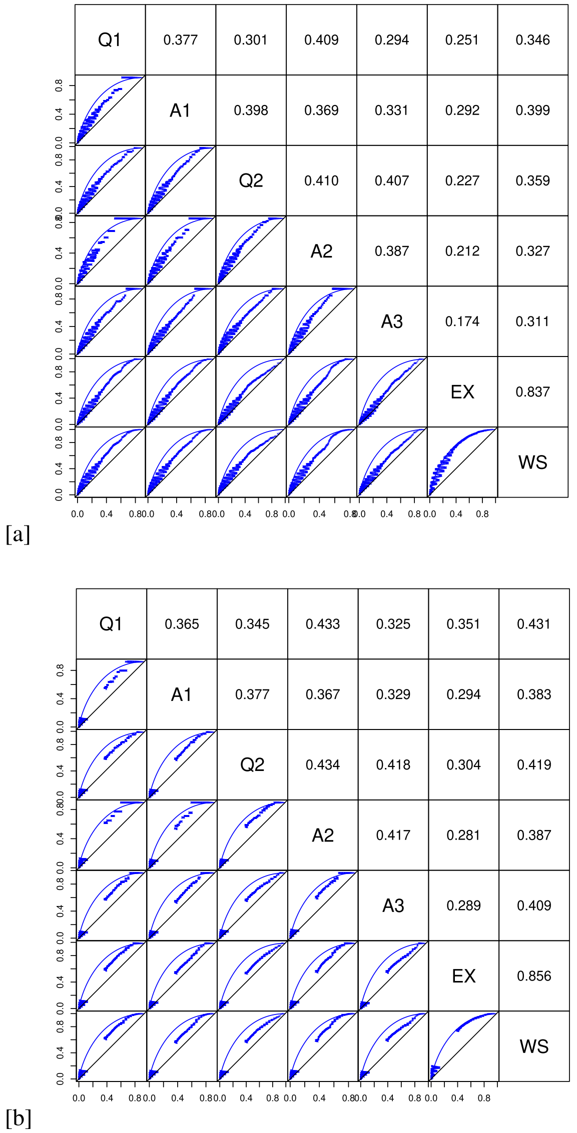


FIGURE 9. The ranked Kendall correlation and Kendall tau plot for ENM1600 Engineering Mathematics performance for [a] on-campus (ONC) face-to-face and [b] online (ONL) web-based students.

heavy tails, they are not appropriate when there is asymmetric dependence structures [71]. Further, the elliptical copula in most cases cannot be given explicitly because the distribution F and the corresponding marginal distributions are usually represented in integral forms [72].

APPENDIX D DETAILS OF MULTIVARIATE COPULAS AND THEIR APPLICATIONS

A. ELLIPTICAL

Generally including multivariate Gaussian and Student's t copulas. The multivariate Gaussian copula is defined as:

$$C\left(u_1,\ldots,u_d;\rho\right)=\Phi_\rho\left[\Phi^{-1}\left(u_1\right),\ldots,\Phi^{-1}\left(u_d\right)\right],\qquad(20)$$

and the multivariate Student's t copula:

$$C\left(u_1,\ldots,u_d;\rho\right)=T_{\rho,v}\left[T_v^{-1}\left(u_1\right),\ldots,T_v^{-1}\left(u_d\right)\right],\qquad(21)$$

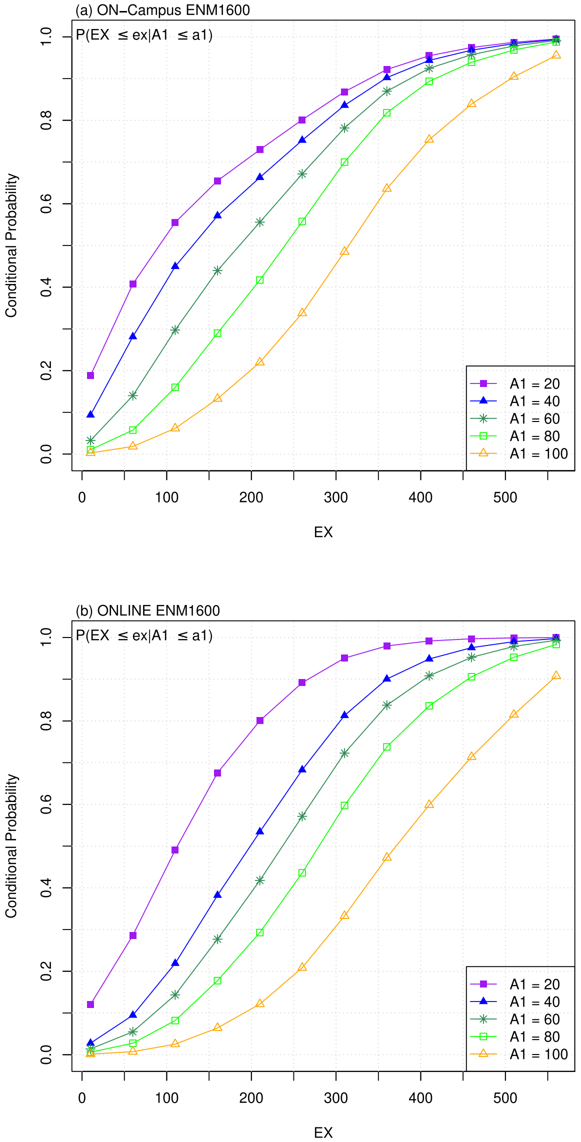


FIGURE 10. Bi-Variate Copula Model for ENM1600: Conditional probability plot showing the probability of an examination score, *EX* being less than or equal to a threshold mark, *ex* conditional upon Assignment 1, *A1* being less than or equal to a threshold mark, *a1*. *To interpret this result, consider an on-campus student who has an Assignment 1 score of 80/150 marks, is expected to have a 70% probability to score a 300/600 examination score (or a pass in the examination).*

where ρ denotes a symmetric, positive definite matrix with elements in the diagonal equal to one. Φ and $T_{\rho,v}$ are the standardised multivariate normal Student's t distribution, respectively, with correlation matrix ρ and v degrees of freedom.

1) APPLICATIONS

The Gaussian copula exhibits tail independence meanwhile the Student's t copula is symmetric dependence in the lower and upper tail. Although these elliptical members are generally possible to capture a wide range of dependence including heavy tails, they are not appropriate when there is asymmetric dependence structures [71]. Further, the elliptical copula in most case cannot be given explicitly because the distribution

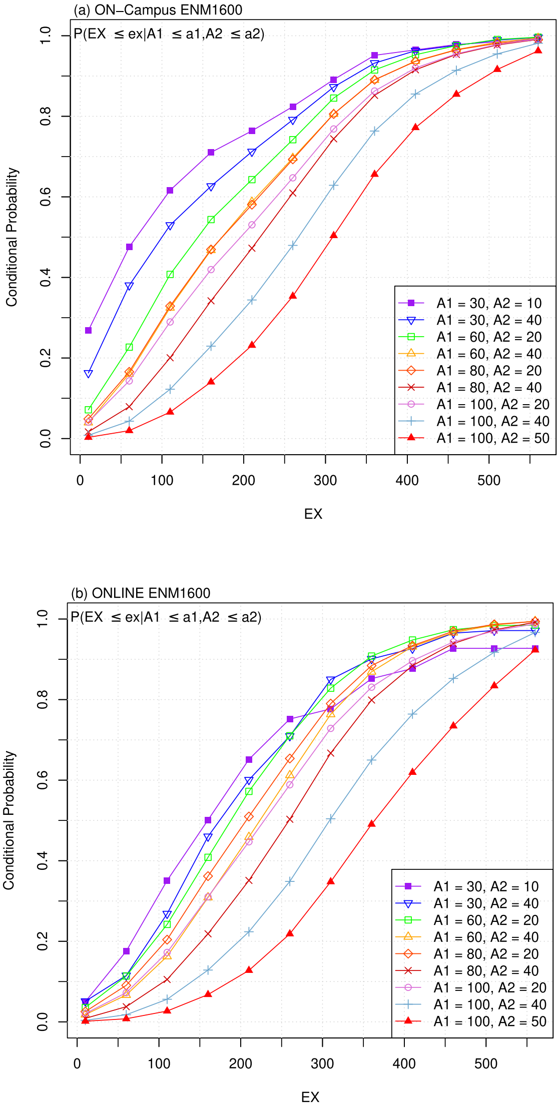


FIGURE 11. Tri-Variate copula model results for ENM1600: Conditional probability plot showing the probability of *EX* being less than or equal to a threshold mark, *ex* given that the Assignment 1, *A1* and Assignment 2, *A2* are less than or equal to threshold marks *a1* and *a2*. *To interpret this result, consider an on-campus student with* A1 = *100/150 marks and* A2 = *50/150 marks, expected to have a 50% probability to score an* EX = *300/600 marks (or a pass in the examination).*

F and the corresponding marginal distributions are usually represented in integral forms [72].

B. ARCHIMEDEAN

Generally including exchangeable and non-exchangeable Archimedean copulas (ACs). Exchangeable ACs is a classical construction where bivariate ACs are extended to the d-dimensional case given a strict generator:

$$\varphi : [0, 1] \rightarrow [0, \infty] . \tag{22}$$

The associated function C of a d-dimensional AC has close form representation defined as:

$$C (u_1, \ldots, u_d) = \varphi^{-1} [\varphi (u_1) + \ldots + \varphi (u_d)] , \tag{23}$$

if and only if φ^{-1} is completely monotonic on $\mathbb{R}$. Non-exchangeable ACs is asymmetric generalization, hierarchical ACs (HACs) [42] also known as nested ACs (NACs). The HAC comprised ACs belonging to the same family may be named as a homogeneous HAC, otherwise, a heterogeneous HAC [73]. There are two special forms of HACs, namely the fully nested ACs (FNACs) and the partially nested ACs (PNACs) [42], [74]. The FNACs takes a relatively simple form where u_1 and u_2 is coupled first by a bivariate copula function C_1 with the parameter θ_1. Then that copula C_1 is coupled with u_3 by a new copula C_2 and the parameter θ_2, and so on.

The pair-copula is derived from the corresponding generator described in Table 1. The PNAC is a mixture of ordinary ACs and FNACs. The HAC has been thoroughly investigated in the literature [40], [75]–[77]. There are also multiplicative ACs proposed by Liebscher [78] and Morillas [79].

1) APPLICATIONS

Archimedean copulas (ACs) can overcome the limitations of the elliptical copulas. In exchangeable ACs, the rendered dependence is symmetric in respect of the permutation of variables, which means that the distribution is exchangeable [72]. The multivariate ACs is very restricted in high-dimensional cases because the multivariate dependence structure relies on a single parameter of the generator function. Non-exchangeable ACs can improve flexibility and allow for non-exchangeable dependence structures. However, one of the restrictions of the HACs is that only AC families are used in the building block. Further, parameter restrictions require the parameters estimated for higher levels to be smaller than those for lower levels, which may reduce the flexibility for modelling dependence structures.

C. EXCHANGEABLE ARCHIMEDEAN COPULAS

Archimedean copulas overcome the limitations of elliptical class. Bivariate ACs can be extended to the d-dimensional case, given a strict generator $\varphi : [0, 1] \rightarrow [0, \infty]$. The associated function C of a d-dimensional AC has close form representation defined as:

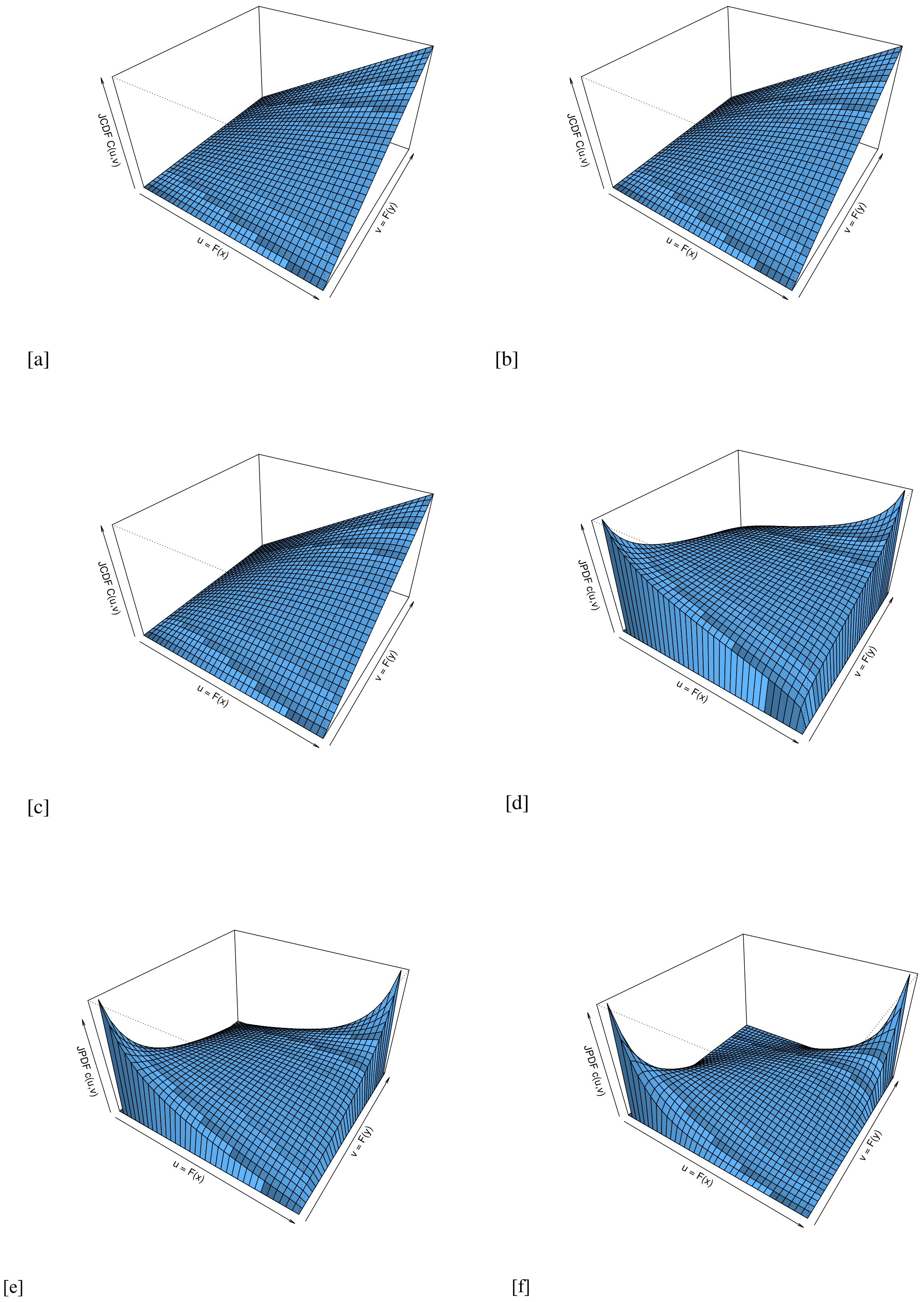
$$C (u_1, \ldots, u_d) = \varphi^{-1} [\varphi (u_1) + \ldots + \varphi (u_d)] \tag{24}$$

if and only if φ^{-1} is completely monotonic on R^+.

Figure 14 illustrates an example of a five-dimensional AC in a classical copula construction process. Clearly, we note that the rendered dependence is symmetric with respect of the permutation of variables, which means that the distribution is exchangeable [72] and the multivariate AC is relatively restricted in high-dimensional cases as the multivariate dependence structure relies on a single parameter of generator function.

D. NON-EXCHANGEABLE ARCHIMEDEAN COPULAS

We can construct multivariate ACs in alternative ways to improve the flexibility of modelling of student performance data and allow for non-exchangeable dependence structures



[a]

[b]

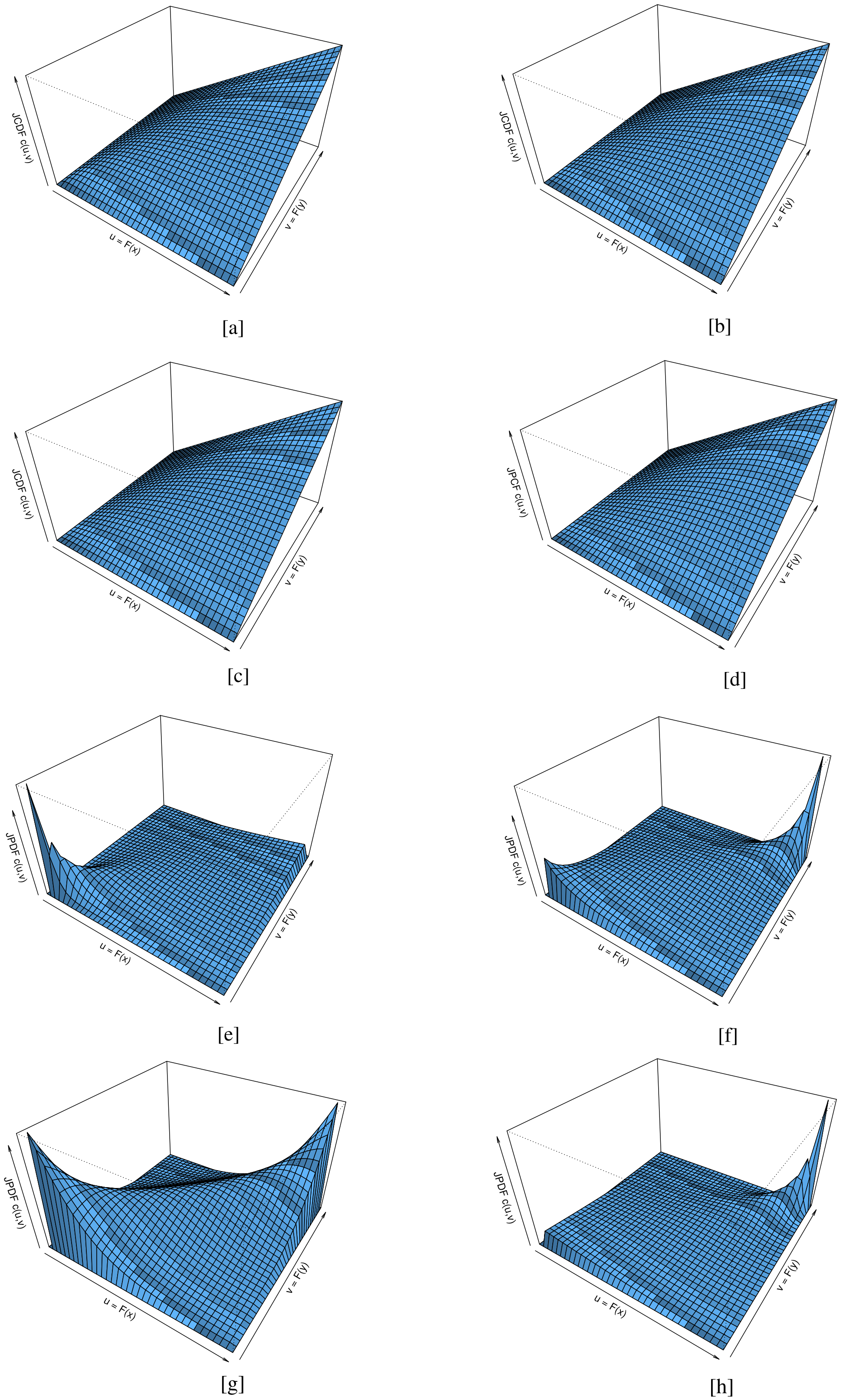
[c]

[d]

[e]

[f]

FIGURE 12. The simulated joint cumulative distribution function (JCDF) and the probability density function (JPDF) of bivariate Elliptical (Gaussian) copula models with differently parameters, θ and τ. For JCDF, see plots [a-c] [a] Gaussian: $\theta = 0.23$; $\tau = 0.15$, [b] Gaussian: $\theta = 0.45$; $\tau = 0.30$, [c] Gaussian: $\theta = 0.71$; $\tau = 0.60$. For JPDF, see plots [d-f] [a] Gaussian: $\theta = 0.23$; $\tau = 0.15$, [b] Gaussian: $\theta = 0.45$; $\tau = 0.30$, [c] Gaussian: $\theta = 0.71$; $\tau = 0.60$.



[a] [b]

[c] [d]

[e] [f]

[g] [h]

FIGURE 13. **The simulated joint cumulative distribution function (JCDF) and the probability density function (JPDF) of bivariate Archimedian (Gaussian) copula models with differently parameters, θ and τ. For JCDF, see plots [a-d] [a] Clayton: $\theta = 2.00$, $\tau = 0.50$. [b] Gumbel: $\theta = 2.00$, $\tau = 0.50$. [c] Frank: $\theta = 2.00$, $\tau = 0.50$. [d] Joe: $\theta = 2.86$, $\tau = 0.50$. For JPDF, see plots [e-h] [e] Clayton: $\theta = 2.00$, $\tau = 0.50$. [f] Gumbel: $\theta = 2.00$, $\tau = 0.50$. [g] Frank: $\theta = 2.00$, $\tau = 0.50$. [h] Joe: $\theta = 2.86$, $\tau = 0.50$.**

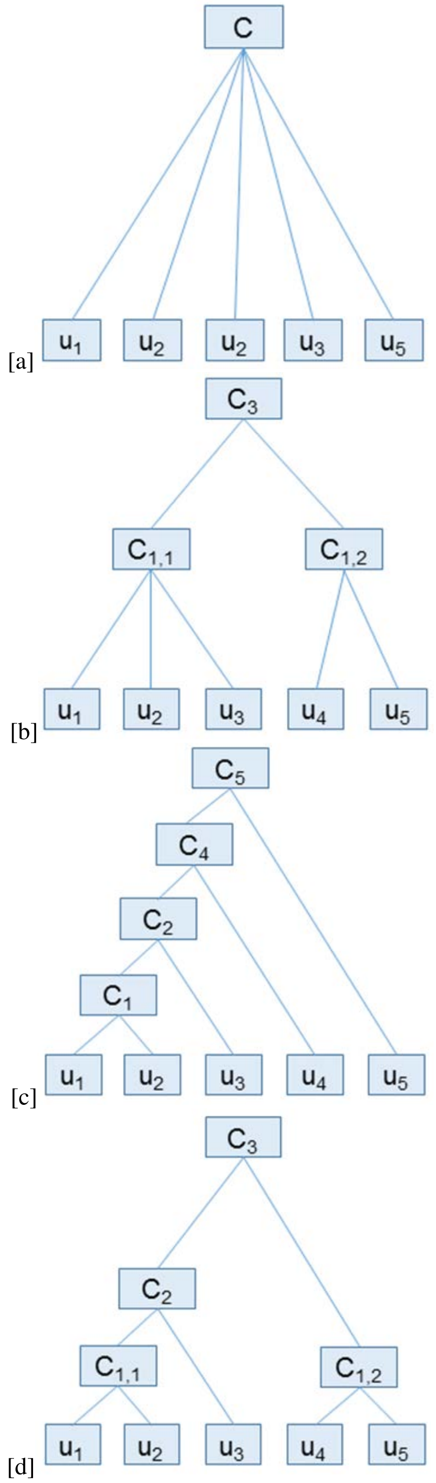


FIGURE 14. (a) Five-dimensional symmetric Archimedean copulas, (b–d) Hierarchical Archimedean copulas (HAC) constructed and partial nested, (c) Fully nested structure. The joint distribution between variables *u* is modelled by the copula function *C*. For example, C_1 is the copula function for *u*1 and *u*2 and *C*2 the copula function for *C*1 and *u*3.

to be considered. Asymmetric generalization, hierarchical ACs (HACs) [42] also known as the nested ACs (NACs), is the most popular approach due to their flexibility so under sufficient nesting conditions [80], this structure constructs of a hierarchy of ACs with different levels. At the first level, variables are grouped into distinct multivariate ACs with all copulas in the first level again grouped into copulas at level two, etc. This procedure continues until the top level contains only a single HAC is achieved. The HAC comprises of ACs belonging to the same family, or homogeneous HAC, otherwise, a heterogeneous HAC [73] is achieved.

Two special forms of HACs, namely the fully nested ACs (FNACs) and the partially nested ACs (PNACs) [42], [74] are considered with FNACs taking a relatively simple form where $U1$ and $U2$ is coupled first by a bivariate copula C_1 with parameter θ_1.The copula C_1 is coupled with $u3$ by a new copula C_2 and the parameter θ_2, and so forth.

Figure 14c describes one possible structure for a five-dimensional FNA copula model. The pair-copula is derived from the corresponding generator described in Table 1 where the PNAC is a mixture of the ordinary ACs and FNACs.

A possible structure of a five-dimensional PNAC is depicted in Figure 14(b-d). It should be noted that the HAC has thoroughly been investigated e.g., [75] with multiplicative ACs in Liebscher [78] and Morillas [79].

However, one of their restrictions is that only the AC families are used in building blocks and so the parameter restrictions require the parameters estimated for higher levels to be smaller than those for lower levels, which may reduce the flexibility to model dependence structures.

E. VINE

Vine copulas are generally expressed in three forms: regular (R)-vine, canonical (C)-vine, and drawable (D)-vine copulas. For vine copulas, please see Part C in section II for more details.) It is important to mention that vine copulas, also known as pair-copula constructions ([43]), are able to overcome limitations mentioned above.

APPENDIX E
DESCRIPTIVE STATISTICS OF STUDENT PERFORMANCE DATA-SET

To better understand the data features used in the modelling process, in Tables 7 and 8, we show the descriptive statistics of the data-set used to construct the proposed multivariate copula models. It is evident that there is little difference between the skewness, flatness and standard deviations of the online (ONL) and on-campus (ONC) student performance among both subjects under investigation. For example, the α value for weighted score (WS) is ≈ 17.0 vs 16.9 for ENM2600 ONL and ONC students, respectively whereas it is 17.4 and 17.0 for ENM1600. However, the characteristics of the predictor and target data-set based on their distribution indicators for each course and student cohort occupies disparate values to suggest that the nature of each predictor is different from the other. For example, in terms of quiz 1 and quiz 2, we have $\alpha = -1.8$ vs -1.0 for ENM2600 ONL data-set, whereas in terms of the exam score, it is 0.1 and in terms of weighted score, it is -0.1. Similar differences are noted for ENM1600.

TABLE 7. Descriptive statistics of Advanced Engineering Mathematics (ENM2600) student performance data (2015-2018) used to develop multivariate copula models. Note: α = skewness factor, β = flatness factor, γ standard deviation of the data.

	ENM2600 ONL					
	Q1/50	A1/150	Q2/50	A2/150	EX/600	WS/100%
α	-1.8	-1.2	-1.0	-1.2	0.1	-0.1
β	3.8	1.1	0.6	1.0	-0.8	-0.4
γ	6.9	28.7	10.0	29.3	136.5	17.0
	ENM2600 ONC					
α	-2.8	-1.7	-2.1	-2.4	0.3	0.1
β	9.9	3.3	4.3	7.2	-0.9	-0.6
γ	6.0	26.8	8.1	24.1	146.9	16.9

TABLE 8. Descriptive statistics for engineering mathematics ENM1600.

	ENM1600 ONL						
	Q1/50	A1/150	Q2/50	A2/150	A3/150	EX/600	WS/100%
α	-1.3	-2.1	-1.1	-1.6	-1.1	-0.2	-0.4
β	1.8	5.5	1.2	3.3	0.7	-0.7	-0.3
γ	8.0	11.6	16.0	8.4	21.1	133.4	17.4
	ENM1600 ONC						
α	-1.3	-2.2	-1.4	-1.8	-1.1	0.1	-0.1
β	2.5	7.0	2.1	3.6	0.5	-0.6	-0.4
γ	7.9	13.0	17.2	8.2	22.1	137.2	17.0

It is imperative to note that the copula models in this paper provided a distinct advantage to handle the different statistical properties of the predictor and target variables whereby joint distributions with different shape, scale, or other statistical factors for the quiz, assignment, exam score, weighted score are considered through a unique set of copula parameters. This indicates that in spite of the diverse (or disparate) features provided by each predictor towards modelling the target variable, copula models are precisely tailored to incorporating such differences in the statistical properties represented in Tables 7 and 8, and are therefore considered robust in predicting engineering mathematics student performance data.

ACKNOWLEDGMENT

The authors thank Prof. Yury Stepanyants and Dr. Trevor Langlands for his initial discussions or feedback, University of Southern Queensland's School of Sciences (now called the School of Mathematics, Physics, and Computing) and the Office for Advancement of Learning and Teaching who seeded the project through a Technology Demonstrator Project Initiative. They also thank the three reviewers and the Journal Editor for their support in reviewing the article.

REFERENCES

[1] R. C. Deo, Z. M. Yaseen, N. Al-Ansari, T. Nguyen-Huy, T. A. M. Langlands, and L. Galligan, "Modern artificial intelligence model development for undergraduate student performance prediction: An investigation on engineering mathematics courses," *IEEE Access*, vol. 8, pp. 136697–136724, 2020.

[2] A. J. Fernández-García, R. Rodríguez-Echeverría, J. C. Preciado, J. M. C. Manzano, and F. Sánchez-Figueroa, "Creating a recommender system to support higher education students in the subject enrollment decision," *IEEE Access*, vol. 8, pp. 189069–189088, 2020.

[3] H. A. Mengash, "Using data mining techniques to predict student performance to support decision making in university admission systems," *IEEE Access*, vol. 8, pp. 55462–55470, 2020.

[4] A. Nabil, M. Seyam, and A. Abou-Elfetouh, "Prediction of students' academic performance based on Courses' grades using deep neural networks," *IEEE Access*, vol. 9, pp. 140731–140746, 2021.

[5] S. Hossain, D. Sarma, F. Tuj-Johora, J. Bushra, S. Sen, and M. Taher, "A belief rule based expert system to predict student performance under uncertainty," in *Proc. 22nd Int. Conf. Comput. Inf. Technol. (ICCIT)*, Dec. 2019, pp. 1–6.

[6] Ryan S. J. D. Baker *et al.*, "Data mining for education," *Int. Encyclopedia Educ.*, vol. 7, no. 3, pp. 112–118, 2010.

[7] H. Agrawal, H. Mavani, and K. J. Somaiya, "Student performance prediction using machine learning," *Int. J. Eng. Res.*, vol. 4, no. 3, pp. 111–113, Mar. 2015.

[8] J. R. Dai, M. Y. Li, W. W. Li, Z. Lu, and Z. G. Zhang, "Setting of academic warning based on multivariate copula functions," *Appl. Mech. Mater.*, vols. 571–572, pp. 156–163, Jun. 2014.

[9] A. Luceño, "Fitting the generalized Pareto distribution to data using maximum goodness-of-fit estimators," *Comput. Statist. Data Anal.*, vol. 51, no. 2, pp. 904–917, 2006.

[10] A. J. Gabriele, E. Joram, and K. H. Park, "Elementary mathematics teachers' judgment accuracy and calibration accuracy: Do they predict students' mathematics achievement outcomes?" *Learn. Instruct.*, vol. 45, pp. 49–60, Oct. 2016.

[11] K. W. Thiede, J. L. Brendefur, R. D. Osguthorpe, M. B. Carney, A. Bremner, S. Strother, S. Oswalt, J. L. Snow, J. Sutton, and D. Jesse, "Can teachers accurately predict student performance?" *Teach. Teacher Educ.*, vol. 49, pp. 36–44, Jul. 2015.

[12] L. L. Baird, "Using self-reports to predict student performance. Research monograph no. 7," Dept. College Entrance Examination Board, New York, NY, USA, Tech. Rep. CEEB-BM-7, 1976.

[13] M. Stapel, Z. Zheng, and N. Pinkwart, "An ensemble method to predict student performance in an online math learning environment," *9th Int. Conf. Educ. Data Mining*. North Carolina, USA, Jul. 2016, pp. 231–238.

[14] T. Tanner and H. Toivonen, "Predicting and preventing student failure–using the k-nearest neighbour method to predict student performance in an online course environment," *Int. J. Learn. Technol.*, vol. 5, no. 4, pp. 356–377, 2010.

[15] S. Kotsiantis, C. Pierrakeas, and P. Pintelas, "Predicting students performance in distance learning using machine learning techniques," *Appl. Artif. Intell.*, vol. 18, no. 5, pp. 411–426, 2004.

[16] J. Xu, K. H. Moon, and M. van der Schaar, "A machine learning approach for tracking and predicting student performance in degree programs," *IEEE J. Sel. Topics Signal Process.*, vol. 11, no. 5, pp. 742–753, Aug. 2017.

[17] R. Alamri and B. Alharbi, "Explainable student performance prediction models: A systematic review," *IEEE Access*, vol. 9, pp. 33132–33143, 2021.

[18] M. Ali, R. C. Deo, N. J. Downs, and T. Maraseni, "Cotton yield prediction with Markov chain Monte Carlo-based simulation model integrated with genetic programing algorithm: A new hybrid copula-driven approach," *Agricult. Forest Meteorol.*, vol. 263, pp. 428–448, Dec. 2018.

[19] M. Ali, R. C. Deo, N. J. Downs, and T. Maraseni, "Multi-stage hybridized online sequential extreme learning machine integrated with Markov chain Monte Carlo copula-bat algorithm for rainfall forecasting," *Atmos. Res.*, vol. 213, pp. 450–464, Nov. 2018.

[20] M. Vuolo, "Copula models for sociology: Measures of dependence and probabilities for joint distributions," *Sociol. Methods Res.*, vol. 46, no. 3, pp. 604–648, Aug. 2017.

[21] M. Sklar, "Fonctions de repartition an dimensions et leurs marges," *Publ. Inst. Statist. Univ. Paris*, vol. 8, pp. 229–231, 1959.

[22] L. Zhang and V. P. Singh, "Trivariate flood frequency analysis using discharge time series with possible different lengths: Cuyahoga river case study," *J. Hydrologic Eng.*, vol. 19, no. 10, Oct. 2014, Art. no. 05014012.

[23] E. W. Frees, P. Shi, and E. A. Valdez, "Actuarial applications of a hierarchical insurance claims model," *ASTIN Bull., J. IAA*, vol. 39, no. 1, pp. 165–197, May 2009.

[24] E. W. Frees and P. Wang, "Credibility using copulas," *North Amer. Actuarial J.*, vol. 9, no. 2, pp. 31–48, Apr. 2005.

[25] C. Genest, M. Gendron, and M. Bourdeau-Brien, "The advent of copulas in finance," *Eur. J. Finance*, vol. 15, nos. 7–8, pp. 609–618, Dec. 2009.

[26] A. Patton, "Copula methods for forecasting multivariate time series," in *Handbook Economic Forecasting*, vol. 2, B. V. Elsevier, Ed. Amsterdam, The Netherlands: North Holland, 2013, pp. 899–960.

[27] P. K. Trivedi and D. M. Zimmer, *Copula Modeling: An Introduction for Practitioners*. Boston, MA, USA: Now, 2007.

[28] C.-C. Wu, H. Chung, and Y.-H. Chang, "The economic value of co-movement between oil price and exchange rate using copula-based garch models," *Energy Econ.*, vol. 34, no. 1, pp. 270–282, 2012.

[29] T. Nguyen-Huy, R. C. Deo, S. Mushtaq, J. Kath, and S. Khan, "Copula-based agricultural conditional value-at-risk modelling for geographical diversifications in wheat farming portfolio management," *Weather Climate Extremes*, vol. 21, pp. 76–89, Sep. 2018.

[30] T. Nguyen-Huy, R. C. Deo, S. Mushtaq, J. Kath, and S. Khan, ''Copula statistical models for analyzing stochastic dependencies of systemic drought risk and potential adaptation strategies,'' *Stochastic Environ. Res. Risk Assessment*, vol. 33, no. 3, pp. 779–799, Mar. 2019.

[31] T. Nguyen-Huy, J. Kath, S. Mushtaq, D. Cobon, G. Stone, and R. Stone, ''Integrating el Niño-southern oscillation information and spatial diversification to minimize risk and maximize profit for Australian grazing enterprises,'' *Agronomy Sustain. Develop.*, vol. 40, no. 1, pp. 1–11, Feb. 2020.

[32] A.-C. Favre, S. El Adlouni, L. Perreault, N. Thiémonge, and B. Bobée, ''Multivariate hydrological frequency analysis using copulas,'' *Water Resour. Res.*, vol. 40, no. 1, pp. 1–12, Jan. 2004.

[33] C. Genest and A.-C. Favre, ''Everything you always wanted to know about copula modeling but were afraid to ask,'' *J. Hydrol. Eng.*, vol. 12, no. 4, pp. 347–368, Jul. 2007.

[34] J.-T. Shiau, S. Feng, and S. Nadarajah, ''Assessment of hydrological droughts for the yellow river, China, using copulas,'' *Hydrol. Processes, Int. J.*, vol. 21, no. 16, pp. 2157–2163, 2007.

[35] T. Nguyen-Huy, R. C. Deo, Z. M. Yaseen, R. Prasad, and S. Mushtaq, ''Bayesian Markov chain Monte Carlo-based copulas: Factoring the role of large-scale climate indices in monthly flood prediction,'' in *Intelligent Data Analytics for Decision-Support Systems in Hazard Mitigation*. Singapore: Springer, 2021, pp. 29–47.

[36] N. Schallhorn, D. Kraus, T. Nagler, and C. Czado, ''D-vine quantile regression with discrete variables,'' 2017, *arXiv:1705.08310*.

[37] C. Genest, A.-C. Favre, J. Béliveau, and C. Jacques, ''Metaelliptical copulas and their use in frequency analysis of multivariate hydrological data,'' *Water Resour. Res.*, vol. 43, no. 9, Sep. 2007.

[38] H. Joe, *Dependence Modeling With Copulas*. Boca Raton, FL, USA: CRC Press, 2014.

[39] R. B. Nelsen, *An Introduction to Copulas*. New York, NY, USA: Springer, 2007.

[40] S. T. Rachev, *Handbook of Heavy Tailed Distributions in Finance: Handbooks in Finance, Book 1*. Amsterdam, The Netherlands: Elsevier, 2003.

[41] E. Brechmann and U. Schepsmeier, ''CDVine: Modeling dependence with C-and D-vine copulas in R,'' *J. Stat. Softw.*, vol. 52, no. 3, pp. 1–27, 2013.

[42] H. Joe, *Multivariate Models and Multivariate Dependence Concepts*. Boca Raton, FL, USA: CRC Press, 1997.

[43] K. Aas, C. Czado, A. Frigessi, and H. Bakken, ''Pair-copula constructions of multiple dependence,'' *Insurance, Math. Econ.*, vol. 44, no. 2, pp. 182–198, Apr. 2009.

[44] B. D. Ripley, *Modern Applied Statistics With S*. New York, NY, USA: Springer, 2002.

[45] M. Evans, N. Hastings, B. Peacock, and C. Forbes, *Statistical Distributions*. Hoboken, NJ, USA: Wiley, 2011.

[46] S. A. Klugman, H. H. Panjer, and G. E. Willmot, *Loss Models: From Data to Decisions*, vol. 715. Hoboken, NJ, USA: Wiley, 2012.

[47] W. Huang and A. Prokhorov, ''A goodness-of-fit test for copulas,'' *Econ. Rev.*, vol. 33, no. 7, pp. 751–771, Oct. 2014.

[48] H. White, ''Maximum likelihood estimation of misspecified models,'' *Econometrica, J. Econ. Soc.*, vol. 50, pp. 1–25, Jan. 1982.

[49] C. Genest, J.-F. Quessy, and B. Rémillard, ''Goodness-of-fit procedures for copula models based on the probability integral transformation,'' *Scandin. J. Statist.*, vol. 33, no. 2, pp. 337–366, Jun. 2006.

[50] U. Schepsmeier, J. Stoeber, E. C. Brechmann, B. Graeler, T.Nagler, T. Erhardt, C. Almeida, A. Min, C. Czado, M. Hofmann, M. Killiches, H. Joe, T. Vatter, ''Package 'vinecopula,''' *R Package Version*, vol. 2, no. 5, 2015.

[51] U. Cherubini, E. Luciano, and W. Vecchiato, *Copula Methods in Finance*. Hoboken, NJ, USA: Wiley, 2004.

[52] E. Kole, K. Koedijk, and M. Verbeek, ''Testing copulas to model financial dependence,'' Dept. Financial Manage., RSM Erasmus Univ., Rotterdam, The Netherlands, Working Paper, 2005.

[53] G. Kim, M. J. Silvapulle, and P. Silvapulle, ''Comparison of semiparametric and parametric methods for estimating copulas,'' *Comput. Statist. Data Anal.*, vol. 51, no. 6, pp. 2836–2850, Mar. 2007.

[54] D. Kraus and C. Czado, ''D-vine copula based quantile regression,'' *Comput. Statist. Data Anal.*, vol. 110, pp. 1–18, Jun. 2017.

[55] C. Genest and J.-C. Boies, ''Detecting dependence with Kendall plots,'' *Amer. Statistician*, vol. 57, no. 4, pp. 275–284, Nov. 2003.

[56] M. Devlin and J. McKay, ''Reframing' the problem': Students from low socioeconomic status backgrounds transitioning to university,'' in *Universities in Transition: Foregrounding Social Contexts of Knowledge in the First Year Experience*, H. Brook, D. Fergie, M. Maeorg, D. Michell, and R. Burton, Eds. Adelaide, NSW, Australia: Univ. Adelaide Press, 2014, pp. 97–125.

[57] M. Niederle and L. Vesterlund, ''Explaining the gender gap in math test scores: The role of competition,'' *J. Econ. Perspect.*, vol. 24, no. 2, pp. 44–129, 2010.

[58] D. Contini, M. L. D. Tommaso, and S. Mendolia, ''The gender gap in mathematics achievement: Evidence from Italian data,'' *Econ. Educ. Rev.*, vol. 58, pp. 32–42, Jun. 2017.

[59] K. P. Mongeon, S. W. Ulrick, and M. P. Giannetto, ''Explaining university course grade gaps,'' *Empirical Econ.*, vol. 52, no. 1, pp. 411–446, Feb. 2017.

[60] S. Scherer, C. P. Talley, and J. E. Fife, ''How personal factors influence academic behavior and GPA in African American STEM students,'' *SAGE Open*, vol. 7, no. 2, pp. 1–14, 2017.

[61] M. Sadegh *et al.*, ''Multi-hazard scenarios for analysis of compound extreme events,'' *Geophys. Res. Lett.*, vol. 45, no. 11, pp. 5470–5480, 2018.

[62] A. C. Cullen, H. C. Frey, and C. H. Frey, *Probabilistic Techniques in Exposure Assessment: A Handbook for Dealing With Variability and Uncertainty in Models and Inputs*. New York, NY, USA: Springer, 1999.

[63] M. L. Delignette-Müller and C. Dutang, ''Fitdistrplus: An R package for fitting distributions,'' *J. Statist. Softw.*, vol. 64, no. 4, pp. 1–34, 2015.

[64] W. N. Venables and B. D. Ripley, *Modern Applied Statistics With S-PLUS*. New York, NY, USA: Springer, 2013.

[65] D. Vose, *Risk Analysis: A Quantitative Guide*. Hoboken, NJ, USA: Wiley, 2008.

[66] T. Nagler, ''Asymptotic analysis of the jittering kernel density estimator,'' *Math. Methods Statist.*, vol. 27, no. 1, pp. 32–46, Jan. 2018.

[67] T. Nagler, ''A generic approach to nonparametric function estimation with mixed data,'' *Statist. Probab. Lett.*, vol. 137, pp. 326–330, Jun. 2018.

[68] J. A. Vrugt, H. V. Gupta, W. Bouten, and S. Sorooshian, ''A shuffled complex evolution metropolis algorithm for optimization and uncertainty assessment of hydrologic model parameters,'' *Water Resour. Res.*, vol. 39, no. 8, pp. 1–19, Aug. 2003.

[69] R. Garcia and G. Tsafack, ''Dependence structure and extreme comovements in international equity and bond markets,'' *J. Banking Finance*, vol. 35, no. 8, pp. 1954–1970, Aug. 2011.

[70] J. D. Woodard, N. D. Paulson, D. Vedenov, and G. J. Power, ''Impact of copula choice on the modeling of crop yield basis risk,'' *Agricult. Econ.*, vol. 42, pp. 101–112, Nov. 2011.

[71] L. Hua and H. Joe, ''Tail order and intermediate tail dependence of multivariate copulas,'' *J. Multivariate Anal.*, vol. 102, no. 10, pp. 1454–1471, Nov. 2011.

[72] O. Okhrin and A. Ristig, ''Hierarchical Archimedean copulae: TheHAC-Package,'' *J. Stat. Softw.*, vol. 58, no. 4, pp. 1–20, 2014.

[73] J. Górecki, M. Hofert, and M. Holeňa, ''On structure, family and parameter estimation of hierarchical Archimedean copulas,'' *J. Stat. Comput. Simul.*, vol. 87, no. 17, pp. 3261–3324, Nov. 2017.

[74] A. J. McNeil, ''Sampling nested Archimedean copulas,'' *J. Stat. Comput. Simul.*, vol. 78, no. 6, pp. 567–581, Jun. 2008.

[75] M. Hofert and M. Mächler, ''Nested Archimedean copulas meet R: The nacopula package,'' *J. Stat. Softw.*, vol. 39, no. 9, pp. 1–20, 2011.

[76] C. Savu and M. Trede, ''Hierarchies of Archimedean copulas,'' *Quant. Finance*, vol. 10, no. 3, pp. 295–304, Mar. 2010.

[77] N. Whelan, ''Sampling from Archimedean copulas,'' *Quant. Finance*, vol. 4, no. 3, p. 339, 2004.

[78] E. Liebscher, ''Modelling and estimation of multivariate copulas,'' Univ. Appl. Sci., Merseburg, Germany, Working Paper, 2006.

[79] P. M. Morillas, ''A method to obtain new copulas from a given one,'' *Metrika*, vol. 61, no. 2, pp. 169–184, Apr. 2005.

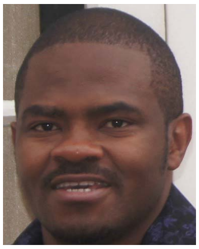
[80] M. Fischer, C. Köck, S. Schlüter, and F. Weigert, ''An empirical analysis of multivariate copula models,'' *Quant. Finance*, vol. 9, no. 7, pp. 839–854, Oct. 2009.

THONG NGUYEN-HUY is currently a Postdoctoral Researcher at the SQNNSW Drought Resilience Adoption and Innovation Hub, University of Southern Queensland, Australia. He works closely with researchers, governments, insurers, and financial institutions on projects focusing on agricultural resilience, climate risks, and alternative risk transfer systems. He has published widely in the fields of climate, agriculture, environment, hydrology, energy, and risk management. His research interests include modeling and data analysis, developing and applying novel statistical models, AI algorithms, and remote sensing techniques.

RAVINESH C. DEO (Senior Member, IEEE) currently leads the USQ's Advanced Data Analytics Laboratory as a Professor at the University of Southern Queensland, Australia. He is a Clarivate Highly Cited Researcher with publications ranking in top 1% by citations for field and publication year in the Web of Science citation index and is among scientists and social scientists who have demonstrated significant broad influence, reflected in the publication of multiple papers frequently cited by peers. He leads cross-disciplinary research in deep learning and artificial intelligence. He is supervising more than 20 Ph.D./M.Sc. degrees and has supervised more than 30 Ph.D./M.Sc. degrees/postdoctoral researchers. He has published more than 250 articles, 150 journals, and seven books with a cumulative citation that exceed 9,200 and an H-index of 54. He has received the Employee Excellence Awards, the Elsevier Highly Cited Paper Awards, and the Publication Excellence and Teaching Commendations, including $2.6 million research funding.

SHAHJAHAN KHAN is currently a Professor of statistics at the University of Southern Queensland, Australia. He is the Leader of evidence-based decision-making in public health. His 2020 book on *Meta-Analysis: Methods for Health and Experimental Studies* (Springer Nature) has over 14k downloads. He has supervised more than 16 Ph.D. and three M.Phil. students. He has published over 250 research articles in systematic review, meta-analysis, predictive inference, pre-test and shrinkage estimations, linear models, and robust tests areas. He is an Expatriate Fellow (elected) of the Bangladesh Academy of Sciences. As the President of the Islamic Countries Society of Statistical Sciences, he has organized international statistical conferences in Malaysia, Egypt, Qatar, Indonesia, Bangladesh, and Pakistan. He has presented 21 research workshops and 25 keynote and plenary addresses in international conferences. He has received the prestigious Q. M. Hossain Gold Medal of Bangladesh Statistical Association. He is the Founding Chief Editor of *Journal of Applied Probability and Statistics* (JAPS).

ARUNA DEVI received the Graduate Diploma degree in education from The University of Adelaide, the Bachelor of Secondary Education degree from The University of the South Pacific, Fiji, the master's degree (Hons.) in inclusive education from The University of Queensland, and the Ph.D. degree (Hons.) from the University of Southern Queensland. Her research has focused on "Preparing Teachers to Instruct Students with Autism in Inclusive Settings: Australian Pre-Service Teachers" and Recent Graduates' Perspectives—An Exploratory Case Study." She is currently an Associate Lecturer with the School of Education and Tertiary Access, University of the Sunshine Coast, Australia. Her research interests include learning difficulties, student learning and development, autism, special and inclusive education, mathematics, and teacher education. She is interested in socio-cognitive theories for self-efficacy belief, including qualitative and quantitative educational research topics.



ADEWUYI AYODELE ADEYINKA is currently an eResearch Analyst at the Office of Research, University of Southern Queensland, Australia. He specializes in the application of index-based risk transfer products in the context of agricultural risk management. He has presented the outcomes of his research at the Actuaries Summit in Australia, U.K., and USA. He has published in prestigious journals.

ARMANDO A. APAN received the B.Sc. degree in forestry from the University of the Philippines Los Baños, the M.Sc. degree in natural resources from the Asian Institute of Technology, Thailand, and the Ph.D. degree in geography and environmental science from Monash University, Australia. He is currently a highly accomplished Professor of remote sensing and GIS at the University of Southern Queensland, Australia. He has over 180 articles published in international refereed journals, book chapters, and conference proceedings, with over 2,900 citations and an H-index of 26. His research interests include the application of geospatial technologies and spatial modeling. In 2006, he has received the Queensland Spatial Science Excellence Award.

ZAHER MUNDHER YASEEN received the master's and Ph.D. degrees from the National University of Malaysia (UKM), Malaysia, in 2012 and 2017, respectively. He is currently an Adjunct Research Fellow at the University of Southern Queensland, Australia; and a Senior Researcher of civil engineering. He was named as the 2021 Clarivate Highly Cited Researcher with publications ranking in top 1% by citations for field and publication year in the Web of Science citation index. He has published over 150 articles in international journals, with a Google Scholar H-Index of 27 and a total of 2400 citations. His research interests include hydrology, water resources engineering, hydrological process modeling, environmental engineering, and the climate. In addition, his interests include machine learning and advanced data analytics.

• • •