IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received May 26, 2021, accepted June 12, 2021, date of publication June 16, 2021, date of current version June 28, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3089870

An Amplified COCOMO-II Based Cost Estimation Model in Global Software Development Context

JUNAID ALI KHAN[1], SAIF UR REHMAN KHAN[1], TAMIM AHMED KHAN[2], AND INAYAT UR REHMAN KHAN[1]

[1]Department of Computer Science, COMSATS University Islamabad (CUI), Islamabad 45550, Pakistan
[2]Department of Software Engineering, Bahria University, Islamabad 44000, Pakistan

Corresponding author: Junaid Ali Khan (junaidalikhan17@gmail.com)

The work of Junaid Ali Khan was supported by the COMSATS University Islamabad, Pakistan.

ABSTRACT Global Software Development (GSD) projects comprise several critical cost drivers that affect the overall project cost and budget overhead. Thus, there is a need to amplify the existing model in GSD context to reduce the risks associated with cost overhead. Motivated by this, the current work aims at amplifying the existing algorithmic model with GSD cost drivers to get efficient estimates in the context of GSD. To achieve the targeted research objective, current state-of-the-art cost estimation techniques and GSD models are reported. Furthermore, the current study has proposed a conceptual framework to amplify the algorithmic COCOMO-II model in the GSD domain to accommodate additional cost drivers empirically validated by a systematic review and industrial practitioners. The main phases of amplification include identifying cost drivers, categorizing cost drivers, forming metrics, assignment of values, and finally altering the base model equation. Moreover, the proposed conceptual model's effectiveness is validated through expert judgment, case studies, and Magnitude of Relative Estimates (MRE). The obtained estimates are efficient, quantified, and cover additional GSD aspects than the existing models; hence we could overcome the GSD project's overall risk by implementing the model. Finally, the results indicate that the model needs further calibration and validation.

INDEX TERMS Global software development, cost estimation, COCOMO-II, cost overhead.

I. INTRODUCTION

As technology advances, the globalization of software companies increases. Software industries are moving towards Global Software Development (GSD) to provide a cost-effective software solution. GSD refers to the type of development which involves virtual teams from different geographical locations to carry out the development [1]. The prior studies predict that the number of GSD projects will increase over time. In the countries like India and China, global software projects are expected to increase by 20 to 30% shortly [2]. The primary purpose of adopting GSD is the low labor cost, whereas it also provides many other advantages like time to market and access to vast skills through virtual teams [3].

Besides the benefits that GSD provides, many associated challenges mainly occur due to cultural differences, time differences, language differences, and other social norms [4], [5]. One of the reported studies presented the disappointing results of the GSD in various projects [5]. Another study has conducted a survey and reported that 31.1% of the GSD projects terminated before completing the projects [6]. One of the main reasons behind the failure of GSD projects is the lack of consideration of the additional cost drivers of GSD. Each development type exhibits its particular characteristics. As in GSD, the development is carried out through remote locations with high geographic, cultural, and time zone differences. Thus, various hidden cost drivers are not considered, ultimately affecting the overall project cost [7]. The existing work lacks in considering the additional cost drivers of GSD. Most of the existing techniques lack quantifying the factors and validation of the generated results [8].

Motivated by this, current research focuses on providing a GSD-specific cost estimation model based on the additional cost drivers of GSD to provide more accurate and

The associate editor coordinating the review of this manuscript and approving it for publication was Pinjia Zhang.

realistic estimates. We believe that the proposed model can assist the practitioners in accurately computing the project's cost in the context of GSD.

The subsequent sections are structured as follows. Section 2 provides an overview of the motivation for this research. Section 3 presents the adopted research methodology. Section 4 discusses the existing cost estimation models in the GSD context. Section 5 highlights the limitations of the existing approaches. Section 6 presents the proposed approach and the amplified model. Finally, Section 7 discusses the conclusion and the future work for this research.

II. RESEARCH MOTIVATION

The success of a GSD project is measured by three parameters based on the client's satisfaction. These three parameters are quality, time, and cost [9]. Quality represents the conformity of the specifications. Moreover, it is measured by the degree of fulfillment of the requirements. In contrast, the time parameter represents the deadlines for the milestones.

Similarly, the cost parameter represents the desired budget of the project. Generally speaking, completing a project within its desired budget is a challenging task. Keeping in view of this fact, a GSD project will only be successful if it fulfills the clients' expectations. Managing cost in the GSD project is crucial because resources are distributed, and transparency is very low [10]. The reported statistics show that approximately 40% of the GSD projects fail and the primary reason behind it is the distance [9]. Some other statistics related to outsourcing extracted from empirical studies are listed below:

- Only 50% of the outsourcing in the near future will be successful [8]–[11].
- Half of the software companies that shifted their process to GSD failed to generate the expected financial benefits [11].
- 70 percent of the software companies had a significant negative experience with out-sourcing [8].
- In a survey of 50 companies, about 14% of outsourcing operations have deemed a failure [8].
- A survey found that 50% of GSD relationships world-wide fail within five years [8].

The above-mentioned facts motivated us to contribute in this research domain. Overall these statistics represent that at the beginning of project, the cost of the GSD projects seems low. Still, as we proceed with the development, the additional cost drivers of GSD like time delay, work quality, and dissatisfaction emerges and affect the project's overall cost [12]. This is the reason that many companies fail to implement the GSD methodologies at a proper manner.

The stats show that many companies failed to produce satisfactory results. In many projects, the cost saving is approximately 50 %. Whereas in many other projects, there is no cost-saving, and the companies failed to deliver the project. The main reason behind it is the preliminary cost analysis, which leads to the inaccurate estimation of the cost-saving. We can only overcome it through proper cost analysis considering all the additional cost drivers of GSD right at the time of assessment to create the most realistic estimate [13], [14].

A realistic cost estimation model is required to prevent the disappointing cost savings that should consider all the hidden cost drivers of GSD. The estimates would be fitting the characteristics of GSD, and the risk would be reduced [15]. We selected COCOMO-II as a base model to achieve the targeted purpose because COCOMO-II is the most established software cost estimation model having some built-in characteristics of GSD [16], [17]. We have identified the Critical Cost Drivers (CCDs) of GSD and amplified the model accordingly. The will help us to predict the outcome for the GSD projects by reducing the overall risk. The main contributions of this research are listed below:

- Identification and categorization of the cost driver in GSD context.
- Empirical evaluation of the identified cost drivers.
- An extensive literature review on existing GSD-specific cost estimation models and their limitations.
- Devising an amplified cost estimation model based on critical cost drivers of GSD
- Validation of the proposed conceptual model.

Based on targeted research objectives, we formulated the research methodology discussed in following section.

III. RESEARCH METHODOLOGY

To achieve the targeted objective, we adopted the Systematic Literature Review (SLR) technique to identify cost estimation factors in the GSD context. Figure 1 represents the adopted research methodology.

First of all, we focused on the problem formulation. In phase 1, a general literature review is conducted to formulate the research problem. After the formulation of the problem, the context is specified. Cost overhead in the context of GSD is selected as a base problem. In phase 2, we performed a planned SLR to extract the factors affecting cost estimation in the GSD context. The performed SLR helps us to extract the factors systematically and unambiguously [18], [19]. In addition to the cost drivers, the existing GSD-specific cost estimation models were also identified. The obtained results were then validated from the industry through a questionnaire. In phase 3, the data is compared, and the statistical tests were applied to examine the correlation between the results (Figure 1).

Notice that we have successfully accomplished the first three phases of research methodology in our previously published work [11]. For more details regarding the adopted research methodology, please refer to our published research work [11]. Phase 4 represents the amplification of the cost estimation model based on the identified critical cost drivers and the limitations of the existing cost estimation models. For the amplification, a base model is selected. We have amplified the COCOMO-II model due to its built-in distributed characteristics. The model is amplified by

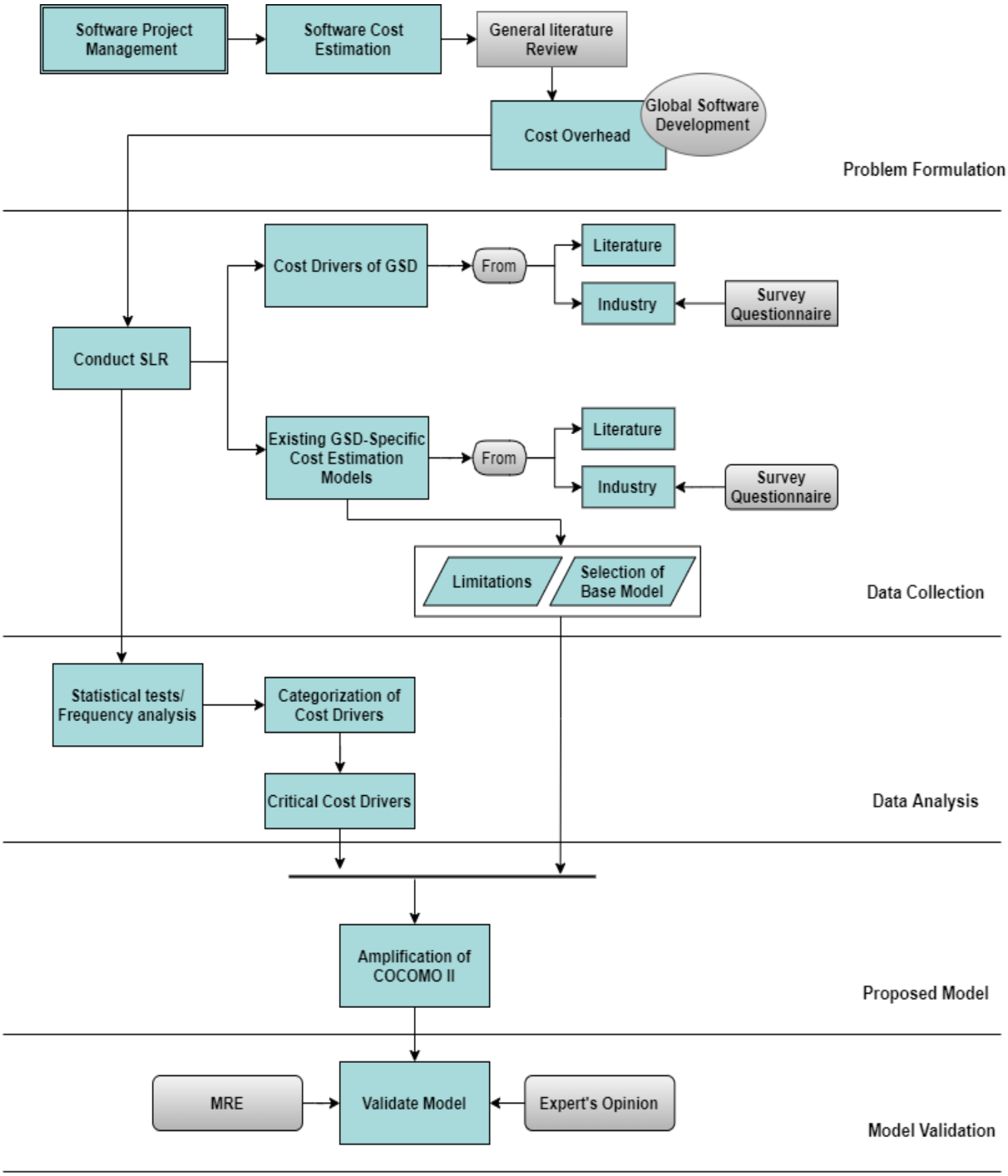


FIGURE 1. The adopted research methodology.

integrating the empirically validated additional cost drivers of GSD. The metrics are formulated, and the equation is alternated. Finally, the proposed model is validated based on different projects through performance measures including MRE. Moreover, the proposed conceptual model is also validated by industrial experts through a questionnaire. The questions covered the aspects related to the labels, the logical connection of the phases, and the identified cost drivers (Figure 1).

IV. CURRENT STATE-OF-THE-ART COST ESTIMATION MODELS

Software cost estimation has been focused for many years in software engineering research, and different estimation techniques have been proposed. However, most of the estimation techniques were presented before introducing the advent of GSD [17]–[20]. Before the emergence of GSD, the software was developed locally, and the adopted estimation techniques considered the local development of

the software products. Hence, the traditional cost estimation techniques lack in considering the global characteristics of GSD. GSD is different from the local development type because of the significant overhead due to the additional cost drivers like language, cultural, and geographic factors [21]. Each site in GSD exhibits its characteristics, and their productivity varies. Recently, the researchers have focused on this context and proposed several cost estimation models applicable in GSD context [22], [23]. Notice that some of the proposed models are algorithmic, while others are non-algorithmic [24]. The following section discusses existing GSD-specific cost estimation models.

A. ALGORITHMIC MODELS

Algorithmic models are the formal models that involve mathematical computation for estimation. These are the mathematical models with the pre-assigned parameters and their values. The different variants of algorithmic-based cost estimation models in the GSD context are discussed as follows:

COCOMO-II is a widely used algorithmic model for collocated projects. However, it does not fit the GSD context because it lacks in distributed characteristics of GSD [17]. So, we made necessary refinements in the COCOMO-II model to make it executable for GSD projects. The amplified variants to COCOMO-II are discussed in the next section.

Cost Xpert is a different model proposed by Madachy [25]. The author selected COCOMO-II as a base model and adapted it according to the distributed environment. The differentiating element is that the model considered phases instead of functions or modules for the estimation. The identified effort multipliers were phase-sensitive and related to people working in different teams. Still, some of the major cost drivers like coordination and collaboration, which significantly impact the global environment, were not considered in this model. Therefore, the model is only suitable for projects where work allocation is based on phases rather than specified functions. Moreover, there is no validation and quantification in this model, which seems to be a drawback for its reliability.

Another variant of COCOMO-II has been proposed by Keil *et al.* [26]. The authors tailored the model to fit it for the GSD context. The model introduced the additional cost drivers of GSD. The focus of the model was on collaboration and communication between the multiple sites. Thus, the cost drivers related to these two factors were identified. The identified factors were related to product, personnel, and project. The proposed model's main objective was to provide a framework for decision-making to calculate the tradeoff between the collocated and global distributed environment. The model still requires further verification and validation as it is at a very early stage of development. Furthermore, the factors have been derived only through the author's experience and the literature, so it does not seem to be a systematic approach for investigating factors as it lacks empirical support.

Betz and Makio [16] proposed a most comprehensive model based on COCOMO-II. The authors focused on the Post Architecture Variant of COCOMO-II because it contains an extensive list of cost drivers. The model is amplified through three steps. In step 1, the cost drivers were identified through a qualitative survey based on semi-structured interviews with German software companies. 11 new effort multipliers were identified and were named as Effort multipliers of Outsourcing (EMO). In step 2, the identified cost drivers were categorized based on theoretical thoughts and author experiences. The identified cost drivers were grouped into four categories. In step 3, the identified factors were assigned values for the quantification. However, the proposed model is not validated due to the actual missing data of offshoring. Moreover, quantification is not performed systematically, and the derivation of values is unclear. So, the model still needs further calibration and research.

B. NON-ALGORITHMIC MODELS

These types of models are also known as Non-parametric models. Non-algorithmic models depend upon soft computing approaches like fuzzy logic, analogy neural networks, and expert judgment. These methods perform the analysis on the chronological data of the previous projects [27]. GSD-specific non-algorithmic cost estimation models are discussed below:

Use Case Points (UCP) is a non-algorithmic model designed to perform cost estimation based on research and historical data. UCP is a new and relatively simple approach [28] for measuring the project's size, but it is not widely accepted in the software industries. The accuracy of the UCP is associated with the level of details incorporated in the use case diagram. Various steps are performed to calculate the use case point. Initially, actor types are categorized from simple to complex, and then weightage value is assigned to calculate the adjusted and unadjusted use case points. Different transactions are identified to categorize the types of actors. However, there is no adoption of UCP in GSD due to the consideration of limited cost drivers of GSD, the author presented initial research, and the work is still in the data capturing phase and lacks validation.

The functionality of the Analogy-based model is based on the measures and similarity functions of a project. Analysis of the previous datasets is performed to obtain accurate estimates of the desired project. For estimating the cost based on the analogy model [29], the characterizing attributes are selected and based on these selected attributes, the estimation is performed. Thus, the complexity metrics are designed for the input values, the database is updated, and new values are generated. The model contains several limitations as the work is still in progress. The model does not include an exhaustive list of characteristics for the distributed environment. A company working on new technology would lack a similar dataset, and the model would not be applicable without maintaining a dataset of sufficient related projects.

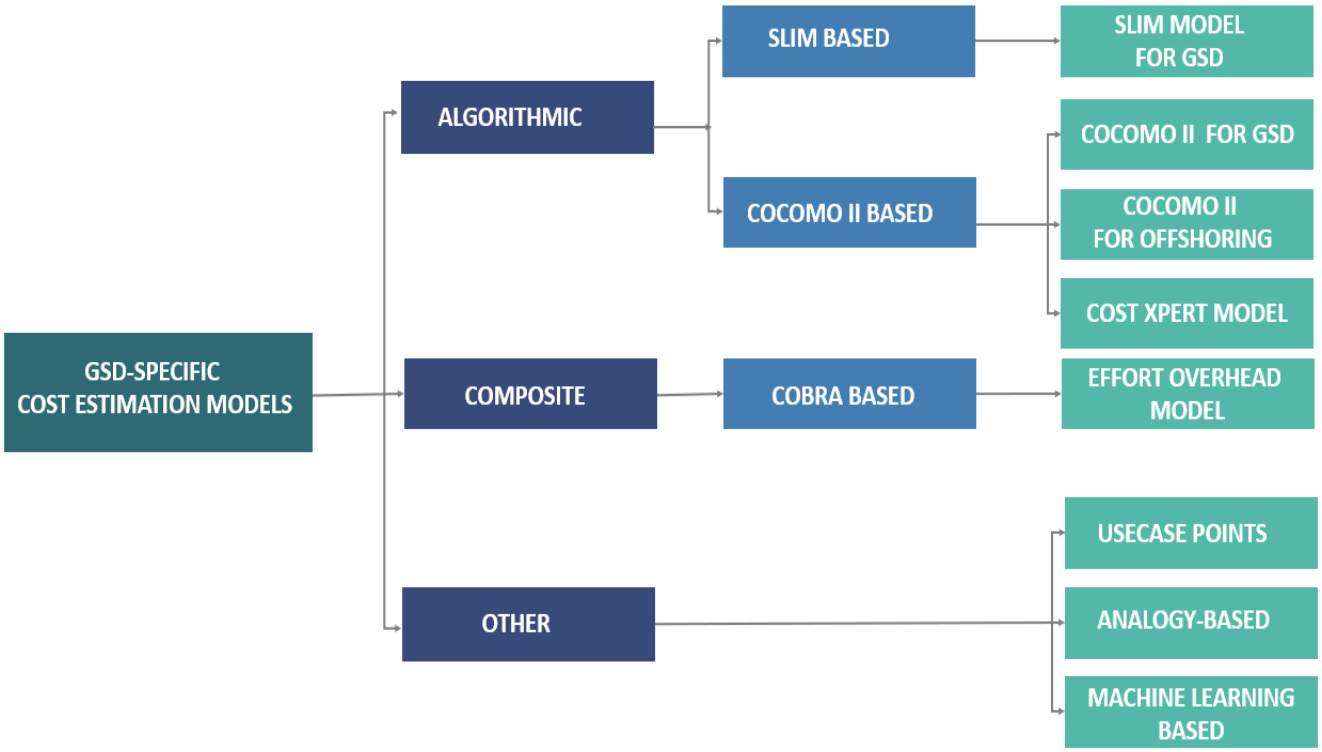


FIGURE 2. Current-state-of-the-art GSD-specific cost estimation models.

With technological advancements, the practitioners are trying to incorporate Machine Learning (ML) techniques in estimation models for accurate and precise estimates. Different ML mechanisms are being used in estimation models like artificial neural networks, regression trees, and genetic algorithms. Humayun and Gang [30] compared the ML techniques to check their applicability in different situations. The authors claimed that ML models could be used for accurate and adequate estimates. There are several limitations to developing machine learning-based estimation models as their performance is associated with the data on which they are trained. The applicability of ML-based models will be challenged if pertinent project data is not available. The work on ML-based models is still in progress and need further attention for improvement and accurate result.

In [31], the authors worked with industrial partners and presented a unique model to better estimate GSD projects. The model is presented for the environment where work is allocated through phases rather than the specific functionality. Their proposed model's distinct factors are the groups' environmental characteristics, labor cost, currency fluctuation, and compensation rates. The model is developed using spreadsheets and is in the early phase of development. However, it still requires further calibration and validation for optimized results.

In [32], the authors presented a cost estimation model for a Spanish GSD-based company. The model amplified the Cost estimation benchmarking and risk assessment (COBRA) model to fit the GSD context. Initially, the cost drivers influencing the cost overhead were identified and ranked according to the experts. Then, a causal model is developed based on the direct and indirect influences of the factors. Furthermore, the experts quantified the cost drivers, and finally, the past projects were analyzed. Based on the results generated from past projects, the experts categorized the cost drivers accordingly.

V. LIMITATIONS OF EXISTING COST ESTIMATION MODELS

The existing cost estimation models are either at the preliminary stage of development or lack additional cost drivers of GSD. Some models like the Analogy-based model [29] and Machine learning-based models [30] are based on theoretical approaches. The authors just presented an initial idea, but these approaches are not implemented and quantified.

Notice that this work is the extension of our previously conducted SLR [11], in which we have extracted the additional cost drivers related to the GSD domain and empirically validated the identified cost drivers. Moreover, we have provided an abstract view of the proposed framework in our published work [11]. To extend our previous work, we have considered COCOMO-II as a base model. This is due to the fact that COCOMO-II is the most established cost estimation model with built-in GSD characteristics that are not considered in any other presented model [16]. Regarding the COCOMO-II applicability, it is used when the top-level design and the detailed information about the project can be extracted [17]. There are various considerable advantages

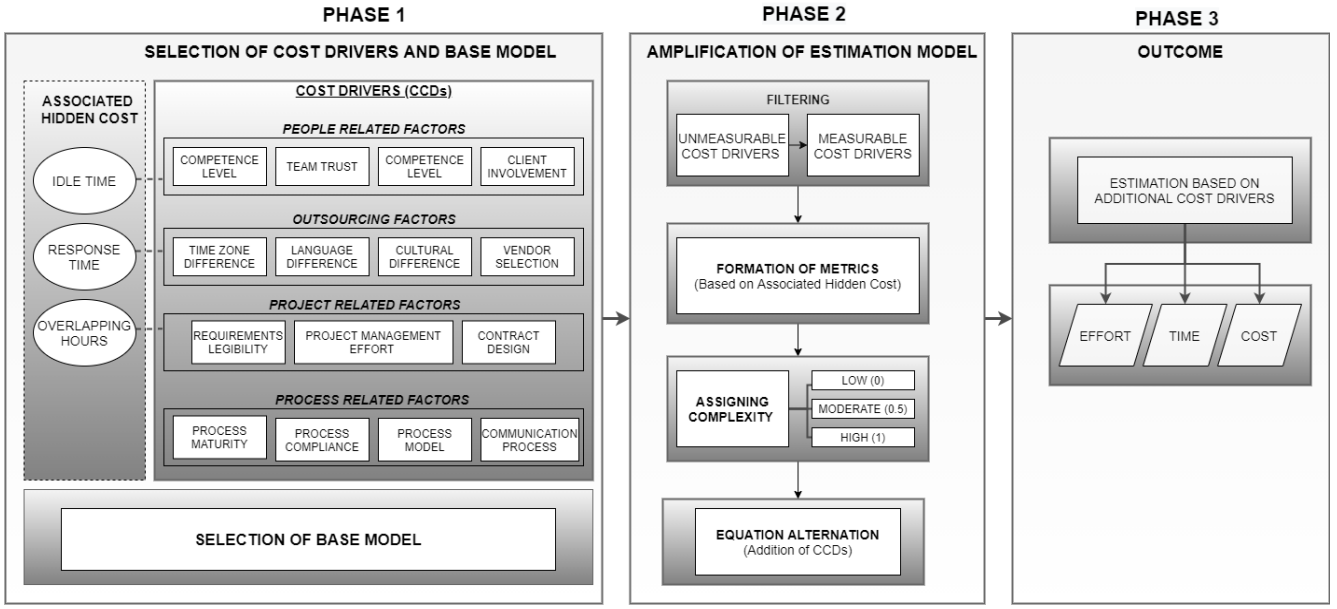


FIGURE 3. The proposed conceptual model.

FIGURE 4. Main phases of proposed conceptual model.

associated with the applicability of the COCOMO-II. Multiple studies reported that COCOMO-II supports a calibration process in an effective manner. In other words, it means that primary metrics used in COCOMO-II model are clearly defined, and the variables are presented in a more detailed manner [33]. Another noticeable advantage of using COCOMO-II is its ability to function more effectively due to dynamically adjusting according to the reported changes. Thus, COCOMO-II has been widely adopted as an industry-standard model [34]. Based on the reasons mentioned above and the distributed nature of the GSD projects, the current study recommended using an algorithmic model where the characteristics of the projects are less known [34].

On the other hand, Machine Learning (ML) models, also known as learning-oriented models, can learn from the previous data and predict the outcome based on the historical data [34]. However, in the GSD context, the machine learning-based cost estimation models are only presented at a conceptual level [30]. Chirra and Reza [33] reported that when a model is created with different historical project datasets, there is a variation in the predicted accuracy. While using machine learning models in the GSD context, it is difficult to determine which technique will give more accurate results on which type of dataset. [30]. The authors mentioned that the ML models based on regression trees lack in effectively modeling the complexities involved in software development projects [30]. Moreover, the less applicability of ML-based estimation models is mainly due to the unavailability of the guidelines or instructions for necessary designing artificial neural-based estimation models. In addition, a large amount of training data is also required. [34]. Shekhar and Kumar [34] concluded that the estimation accuracy is relatively low when the estimations are carried out through the rule induction method. In contrast, when association rules are used, accuracy is comparatively better. However, it cannot be guaranteed that the model will be able to classify all new projects [35]. Furthermore, Bibi and Stamelos [35] reported the limitations of the ML models based on Artificial Neural Networks (ANN) and Case-Based Reasoning (CBR). The authors concluded that the CBR technique is sensitive to the similarity function and also lacks in dealing with the missing values. Similarly, ML models based on ANN exhibit weak explanatory ability. In addition, the ANN technique is prone to get the overfitting to the training dataset and requires plentiful data for training. Comparing the COCOMO-II model with other cost estimation models like (Analogy-based, ML-based, UCP), it serves several distinct characteristics that are missing in any other proposed model [33]. However, we intend to use the ML-based models in our future work subject to the availability of the required guidelines and the instructions [30].

Figure 3 depicted the generalized limitations of the GSD-specific cost estimation models. There are different parameters that each estimation technique lacks in considering while performing cost estimation in the GSD context.

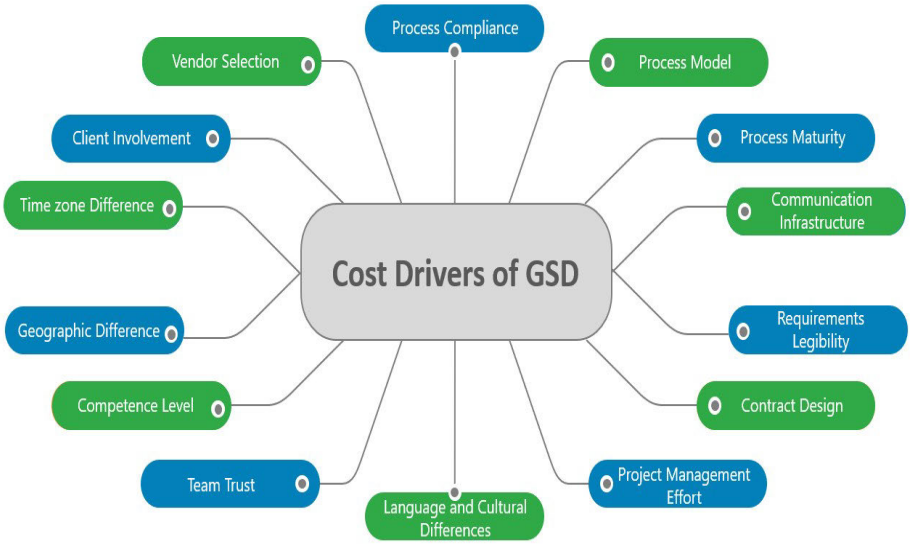


FIGURE 5. Identified cost drivers of GSD.

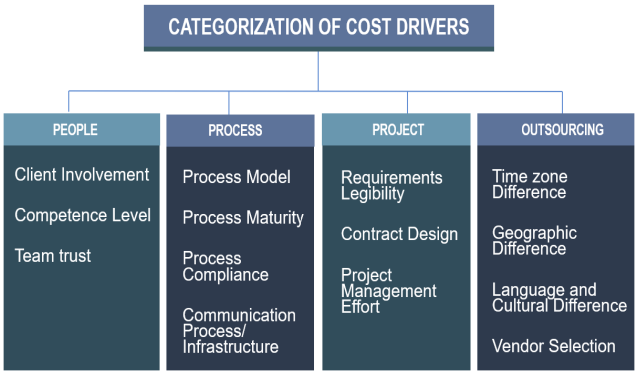
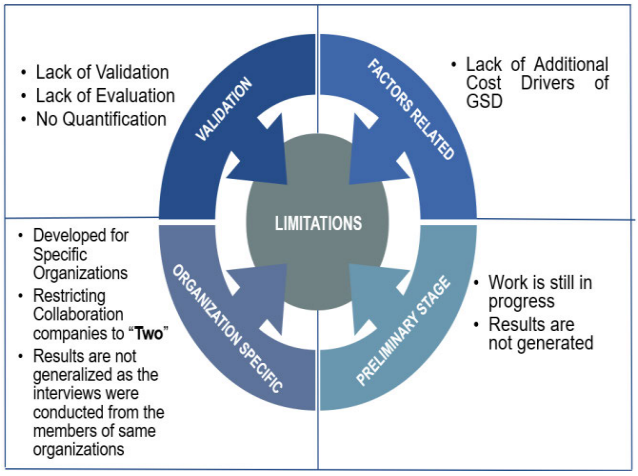


FIGURE 6. Categorization of identified cost drivers.

Estimation Models	No. of Responses (n=175)	Percentage of Usage
Expert Judgement	107	61.1%
Analogy Based	77	44%
Pay as Go	53	30.3%
Algorithmic	27	15.4%
Hybrid Model	22	12.6%
Machine Learning	12	6.9%

FIGURE 7. Industrial perspective of the cost estimation models.

The major limitation is the quantification of the proposed model. The majority of the models are proposed based on theoretical aspects, but they are not quantified. Moreover,



the metrics support for the existing cost estimation models is not available. A detailed review of the existing cost estimation models is presented in Table 1. The labels include the author name, model or technique name, the basic mechanism of the proposed model, findings based on the proposed approach, its limitations, and the evaluation measures through which the proposed models are evaluated (Table 1).

VI. PROPOSED MODEL

Based on the conducted research, we have developed a conceptual model of cost estimation in the GSD context. Figure 4 presents the proposed conceptual model.

The proposed model consists of three main components. In phase 1, the critical cost drivers (CCD's) of the GSD context are selected along with a base cost estimation model. The rationale behind choosing these CCDs is to include the hidden cost associated with the GSD projects counted when estimation is performed. In phase 2, the amplification of

TABLE 1. Review matrix of GSD-specific cost estimation models.

Author(s)	Model or Technique	Model's Description	Finding(s)	Limitation(s)	Evaluation Measures
Betz et al. [16]	Model-Based on COCOMO-II	The model is based on the effort multipliers of GSD.	The proposed model is the most comprehensive as compared to other models because of the 11 additional effort multipliers.	The collaboration companies are restricted to "two," and the resulted numerical values were unclear as no systematic approach is followed for quantification.	The model is not evaluated.
Lamersdorf [32]	CoBRA-based Cost Estimation Model	The influencing factors are identified and presented as a cost overhead-based causal model.	The inter-relationship of the factors presented through the causal model can be of great importance to identify the direct and indirect relationships.	The model is organization-specific because the interviews are conducted within one organization, and the results are not generalized.	Goal Question Metric (GQM) is used for the evaluation of the proposed model.
Mamoona Humayun [30]	Machine Learning-Based Cost Estimation Model	Discussed the mechanism and the overview of various machine learning-based cost estimation models in the GSD context.	The distributed characteristics of GSD are addressed in the corresponding techniques.	These models depend upon the data set used for training so that unsuitable data can result in inaccurate estimates. Moreover, the handling of cost drivers is not discussed.	Canadian data set is used for the evaluation.
Manal et al. [23]	Cost Estimation Model based on UCP	Presented an estimation model based On Use Case Points where the estimation is carried out using case diagrams.	As the proposed technique contains simple concepts, so technical understanding is not required for its execution.	The identified factors are application-specific and depend upon the type of application. Detailed use cases are required for these types of models.	The model is at the stage of data extraction. Therefore, it is not evaluated, and the results are not generated.
Manal et al. [29]	Analogy based cost estimation model	A case-based reasoning model is presented in which similar projects are assigned at the same cost. In the case of GSD, the cost is calculated by identifying the similarity attributes.	The model is applicable if we have done similar projects in the past.	The model is not suitable for the companies working on new technology and does not have any historical data for comparison.	The proposed approach is still in progress, so the results are not evaluated.
Ramacharan [36]	Cost Estimation Model Based on Scheduling Mechanism	Estimating the effort by using the scheduling and productivity parameters, the size measure used in this technique is a line of code.	The comparison of the models is made to highlight the effectiveness of the proposed model.	The model lacks in considering the variation of the record.	The model is evaluated by comparing it to the other models.
Jamshed Ahmed [37]	SOCEM: Software Outsourcing Cost Estimation Model	The challenges of vendor organization regarding software cost estimation are identified in the context of GSD.	The working of the model is based on the challenges of cost estimation.	The estimation model is presented abstractly. Identification and handling of cost drivers are not discussed. Moreover, the model is organization-specific.	Validation is mentioned as future work.
Madachy [25]	Cost Xpert Model	A model is presented that is based on phase-sensitive cost drivers. The cost drivers are related to the professionals working in a different virtual team.	Phase-sensitive effort multipliers are used, and the model is proposed by collaborating with the industry.	The cost drivers like communication and coordination are not considered. Quantification is also missing.	No validation or evaluation is performed.
Keil et al. [26]	COCOMO-II Based Cost Estimation Model	Presented a decision-making framework to resolve the tradeoff between the estimation in collocated and GSD environments.	Factors related to multi-site collocation and multi-site communication are introduced in the model.	The complexity factors are not quantified, and no systematic approach is adopted for acquiring the factors.	No validation or evaluation is performed.

the base model is performed where standardize modification activities are selected. Initially, the immeasurable cost drivers are converted into measurable cost drivers, then criteria are developed, and the formulation of metrics occurs. Once the metrics are formed, then these cost drivers are assigned values considering their level of occurrence. Finally, these values are added to the base model equation, and the results are generated. In phase 3, we estimate the additional cost drivers that were not considered in the traditional cost estimation model. Figure 5 illustrates the main phases of model amplification.

The main phases of the proposed model further consist of sub-phases. The sub-phases represent the task that is being performed inside a module. In phase 1, based on the hidden associated cost, the cost drivers are identified and categorized. Moreover, a base model is selected to integrate the identified cost drivers. Similarly, in the amplification phase, the measurable cost drivers are filtered and assigned complexities once the metrics are formulated. Finally, based on the amplified equation, the estimates are obtained in Phase 3. The estimates are based on time and effort corresponding to the overall cost of the project (Figure 5).

The phases and the sub-phases of the proposed model (Figure 4) are discussed explicitly in the subsequent sections:

A. PHASE 1

This phase presents the list of factors affecting cost estimation in the GSD context, categorizing the identified cost drivers, and selecting the base model. Moreover, it gives the rationale for the selection of the base model (COCOMO-II).

1) IDENTIFICATION OF COST DRIVERS

The categorized factors are extracted through the performed SLR and are validated through an empirical study [11]. The identified factors had a moderate or critical impact on the cost estimation of GSD projects. The extracted cost drivers of GSD are depicted in Figure 6. The only reason for the variance of some cost drivers' percentages is the practitioners' mindset; they are diverted more toward the measurable factor, whereas literature considers the non-measurable factors. But in the end, these are only measurable cost drivers that the cost estimation models require.

Seven factors are common that we extracted and used in the COCOMO-II model [16]. In contrast, we provided an extensive list of cost drivers by adding seven missing factors in the existing model [16]. The missing factors were related to the process, project, and people.

2) CATEGORIZATION OF COST DRIVERS

We added 14 cost drivers with moderate or critical effects on GSD projects. Furthermore, the identified cost drivers are categorized using a customized taxonomy based on 4P'S and Outsourcing factors. The designed categories are illustrated in Figure 7.

The categories with the corresponding factors represented in Figure 7 are discussed in the following sections:

a: PEOPLE-RELATED FACTORS

These factors are based on personal attributes. As the development type varies, the personal attributes also vary with the change in development type. The factors included in this category are; client involvement, the team's competence level, and team trust. Client involvement in GSD projects is important because due to the distributed nature of development, the client is usually unaware of the project. The client should be involved through the communication medium to overcome the transparency of fulfilling the project's requirements. Similarly, in GSD, we do not have a face-to-face meeting in offices, so measuring an employee's competence level is also difficult, but it is essential for accurate estimation. Competence level is a part of skill management where the proficient and highly competent team is selected from remote locations [11], [38]. The estimations should be made considering the team's competence level and developing it in the scheduled timeframe. Team trust should be formed between the members working at different sites. When we lack formal face-to-face meetings, the members hesitate to coordinate and cooperate with the concerned personnel. They don't have the opportunities to develop interpersonal skills, which might negatively impact the project.

b: PROCESS-RELATED FACTORS

Process-related factors refer to the cost drivers that are associated with the development methodologies and processes. Effective processes are key to the productive growth of a software company. Due to GSD development's distributed nature, the operations also vary from those used in the in-house development. The factors included in this category are process model, process maturity, process compliance, and communication process. Choosing the right process model is the base to be in the right direction in any project. In GSD, we have different approaches and methodologies at multiple sites. Integrating these processes is essential because if they do not interoperate, they can lead to data loss or rework, decreasing the product's quality [11]–[17]. Similarly, if we talk about the communication mechanisms, it also varies in the GSD context due to the lack of face-to-face meetings. The study [11] highlighted the importance of communication in the GSD context as it plays a vital role in the success of a project. If we lack effective communication mechanisms, it can lead to delays increasing the project's overall effort [17].

c: PROJECT-RELATED FACTORS

Project-related factors refer to the cost drivers associated with the overall success of the project. These factors are symmetric and are directly linked with the project. In the context of GSD, the project-related factors include requirements legibility, contact design, and project management effort. Requirements

are the core of any project and even more crucial in distributed development, where formal meetings occur through virtual communication channels. It is challenging to understand the actual need of the client in a virtual environment.

Similarly, the project's contact design also plays a vital role in initiating the project with a collaboration company [16]. An intricate contract design will lead to the complexities in the projects. Comparing project management in distributed development with in-house development is more difficult due to the resources' dispersion. Maintaining the project management effort with distributed characteristics is a challenging task.

d: OUTSOURCING FACTORS

Outsourcing factors should be considered where collaborating with an international partner [16]. The factors included in this category are; time zone difference, geographic difference, language or cultural difference, and vendor selection. While working on a GSD project, outsourcing factors should be considered because they can indirectly affect a project's performance. These factors are also known as ''hidden cost drivers'' because these factors are usually not considered during the estimation. Still, they impact the overall project in terms of delays, additional effort, or rework. Time zone difference is a crucial factor presented in several studies [8], [11], [23]. Its impact on a project is asymmetric; an increase in a time zone difference can decrease the virtual teams' overlapping hours, resulting in poor communication and coordination. The effort associated with time zone difference is idle time when a member cannot proceed because he is waiting for a virtual team [11].

Furthermore, choosing the right outsourcing partner is of great importance in distributed development. So, a factor ''Vendor selection'' is added in this category. Based on the experience, the right outsourcing partner for development should be selected.

e: SELECTION OF BASE MODEL

For selecting the base model for the proposed approach, we have performed a critical analysis of the literature. To analyze the cost estimation models presented in the literature, we performed an SLR and extracted all the relevant GSD context's relevant models. 75% of the extracted models were based on Algorithmic Cost Estimation Modeling Technique [8]. Within Algorithmic Models, COCOMO-II is the most established model used for the amplification in the GSD context as it contains the built-in characteristics of a distributed environment. In this work, we selected algorithmic model compared to other estimation models especially ML based estimation models. This is mainly due to the fact that ML models are autonomous; however, ML models are highly susceptible to the estimation errors [30]. Moreover, the other main challenge of ML based estimation models is to determine that which the ML technique provides more accurate results on which dataset [30].

Driver	Sym	VL	L	N	H	VH	XH
PREC	SF_1	0.05	0.04	0.03	0.02	0.01	0.0
FLEX	SF_2	0.05	0.04	0.03	0.02	0.01	0.0
RESL	SF_3	0.05	0.04	0.03	0.02	0.01	0.0
TEAM	SF_4	0.05	0.04	0.03	0.02	0.01	0.0
PMAT	SF_5	0.05	0.04	0.03	0.02	0.01	0.0
RELY	EM_1	0.75	0.88	1.00	1.15	1.40	
DATA	EM_2		0.94	1.00	1.08	1.16	
CPLX	EM_3	0.75	0.88	1.00	1.15	1.30	1.65
RUSE	EM_4		0.89	1.00	1.16	1.34	1.56
DOCU	EM_5	0.85	0.93	1.00	1.08	1.17	
TIME	EM_6			1.00	1.11	1.30	1.66
STOR	EM_7			1.00	1.06	1.21	1.56
PVOL	EM_8		0.87	1.00	1.15	1.30	
ACAP	EM_9	1.5	1.22	1.00	0.83	0.67	
PCAP	EM_{10}	1.37	1.16	1.00	0.87	0.74	
PCON	EM_{11}	1.26	1.11	1.00	0.91	0.83	
AEXP	EM_{12}	1.23	1.10	1.00	0.88	0.80	
PEXP	EM_{13}	1.26	1.12	1.00	0.88	0.80	
LTEX	EM_{14}	1.24	1.11	1.00	0.9	0.82	
TOOL	EM_{15}	1.20	1.10	1.00	0.88	0.75	
SITE	EM_{16}	1.24	1.10	1.00	0.92	0.85	0.79
SCED	EM_{17}	1.23	1.08	1.00	1.04	1.10	

FIGURE 8. Values of scaling factors and effort multipliers [43].

COCOMO-II model contains three main stages including application estimation composition model, early design estimation model, and post architecture estimation model. In this work, we have selected COCOMO-II post architecture as a base model due to the following characteristics [39]

- Availability of the tools
- Parameter's coverage
- Built-in distributed characteristics
- Different factors are available according to the situation

Furthermore, we extracted the information regarding cost estimation models through an empirical study [11]. We designed a questionnaire and distributed it among the 175 project managers of global software companies to identify the cost estimation models used in the industries. Figure 8 depicts the responses of the project managers corresponding to the mentioned estimation models.

Figure 8 represents the usage percentages of different types of cost estimation models as employed in global software-oriented industries. The high percentages of expert judgment, Analogy-based models, and Pay as go' models over other estimation models represent that the GSD industries are still relying on non-algorithmic estimation models. This is mainly due to lack of proposed formal models in the GSD context. Moreover, the formal models developed in GSD context still lack in additional cost drivers of GSD. Generally, they are at the early stage of development, and lack in quantification and validation. Due to this reason, GSD organization lacks in using the formal models for effort and cost estimation purposes. The obtained results served as a baseline in devising a formal model.

TABLE 2. Categorization and value assignment of cultural and geographic difference.

Categories	Criteria (Variation in Number of Countries)	Value
Low	Collaboration companies are from the same country and same geographic location	1.00
Medium	Collaboration companies are from the same country and different geographic location	1.10
High	Collaboration companies are from different countries and different geographic locations	1.25

TABLE 3. Categorization and value assignment of time zone difference.

Categories	Criteria (Overlapping Hours)	Value
Low	Overlapping hours > 8 hours	1.00
Medium	If overlapping hours are between 4-8 hours	1.10
High	Overlapping hours < 4 hours	1.25

TABLE 4. Categorization and value assignment of client involvement.

Categories	Criteria (Number of Session/Meetings)	Value
Low	The number of sessions with client < 2 per month	1.25
Medium	The number of sessions with the client is 2-4/month	1.10
High	Number of Sessions > 4 per month	1.00

TABLE 5. Categorization and value assignment of vendor selection.

Categories	Criteria (Standard or Different Wr.t Each site)	Value
Low	Have not outsourced any project to the vendor before	1.25
Medium	Have done some projects with the vendor before	1.10
High	Outsourced projects with the vendor regularly	1.00

B. PHASE 2

In this phase, we amplified the base model of COCOMO-II post architecture. This phase is divided into sub-phases: (i) metrics formulation, (ii) values assignment, and (iii) equation alternation. The sub-phases are discussed in the subsequent sections:

1) METRICS FORMULATION

The metric formulation is a sub-phase of quantification where identified cost drivers are categorized in terms of complexities and values that have been assigned. The metrics are developed based on the literature support [16] and assistance from the experts.

Table 2 indicates the categories, criteria, and value assignment for the ''Cultural or Geographic Difference.'' With the geographic location, the culture of the virtual teams also changes. The ''Low'' category is presumed if both collaboration companies are from the same countries and same geographic regions. In this case, the cultural difference would be low, so the nominal value ''1.00'' has been assigned. The ''Medium'' category is presumed when the collaboration companies are from the same country, but the groups are from different geographic regions. Value ''1.10'' has been assigned due to the variation in the locations of the groups. Finally, the ''High'' category represents that the collaboration companies are from different countries, and the virtual groups are also from different geographic locations. In this case, the cultural and geographic difference among the groups would be high, so a value of ''1.25'' has been assigned considering the high difference.

Table 3 indicates the categories, criteria, and value assignment for the cost driver ''Time zone Difference.'' The category ''Low'' is presumed if the overlapping office hours between the virtual groups are more than 8 hours; this indicates a low time zone difference. The category ''Medium'' is presumed if the groups' overlapping office hours lie between 4 to 8 hours. In this case, the time zone difference would be medium. Furthermore, the ''High'' category represents the high time zone difference. The criteria presumed is if the overlapping hours between the groups are less than 4 hours, then a higher value ''1.25'' would be assigned.

Table 4 indicates the categories, criteria, and value assignment for the cost driver ''Client Involvement.'' The criteria used to measure the client involvement are the number of meetings or sessions with the client. The higher the number of meetings with the client, the more involvement would be increased, and the nominal value ''1.00'' is assigned. If the number of meetings with the client is less, this corresponds to the low client involvement, and a higher value ''1.25'' is assigned in this case.

Table 5 indicates the categories, criteria, and value assignment for the cost driver ''Vendor Selection.'' The criteria used to measure vendor selection are the vendor's outsourcing experience, whether the vendor is new or we have any experiences with the vendor before. Suppose we have done any projects with the vendor. In that case, the chances of risk are low, and if the vendor is selected without having any experience, then the factor of risk should be counted. The cost should be estimated considering the uncertainties that may occur.

TABLE 6. Categorization and value assignment of competence level.

Categories	Criteria (Experience)	Value
Low	If the experience of the team < 2 years	1.25
Medium	If the experience of the team lies between 2-4 years	1.10
High	If the experience of the team > 4 years	1.00

Table 6 indicates the categories, criteria, and value assignment for the cost driver ''Competence Level.'' The criteria used to demonstrate the competence level of the team is ''the experience.'' The higher the number of experiences in the relevant domain, the higher the competence level will be. The experience involves the work on similar projects. The individuals with a higher number of projects correspond to the high competence level.

Once the metrics are formulated, the next step is the equation alternation, where the fundamental equation of COCOMO-II is altered by incorporating cost drivers' values.

2) EQUATION ALTERNATION

This section provides a brief overview of the COCOMO-II model, its equation, and the alternated equation according to the GSD context.

COCOMO is a formal model used for the estimation of software projects. Barry Boehm formed its theoretical basis in the 1970s, and the earliest version was introduced in 1981 [40]. But with the technological advancements in the software world, many changes emerged, and these changes were incorporated in COCOMO and introduced with the new version COCOMO-II in the year 2000. COCOMO-II is a renowned model and is widely used in software industries because it is calibrated with the actual data of 161 projects, and the measurements are on more than 250 projects. The model can be calibrated through the historical projects, and if not available, we can use standard values to calibrate it. Depending on the phases, different COCOMO-II models can be applied. The available variants are early prototyping, early design, and post architecture models [41], [42]. We considered the post architecture model for the current research as it contains an exhaustive list of cost drivers that could be used in the GSD context. Equation 1 represents the COCOMO-II model [40]:

$$PM = A * Size^{E} * \prod EM \tag{1}$$

where ''PM'' represents ''Person month,'' ''A'' is the constant, whose value is 2.94 for COCOMO-II, but ''A'' value depends on the software company's historical data. The scale factor (E) depends upon five factors; development flexibility, risk resolution, process maturity, team cohesion, and precedence. Scaling factors have a direct influence on the effort multipliers. The Effort Multipliers (EM) or Cost drivers are the projects' characteristics that can directly or indirectly affect the project's effort. These characteristics are assigned numerical values ranging from low to high. A cost driver with a higher value represents the increase in the project's effort. In contrast, the cost driver with a low value represents that it has a low effect on the project's effort, and hence the deviation would be nominal. The corresponding values of the scaling factors and effort multipliers are presented in Figure 9. It represents five scaling factors and 17 effort multipliers with their corresponding values [43]. The formal model's direct use is not possible as it does not consider the GSD context's explicit characteristics. Therefore, we have identified the cost drivers of GSD and used the modular composition to integrate the identified cost drivers into the equation. The obtained cost drivers are named critical cost drivers (CCDs) because they have a moderate or crucial effect on the project. The amplified equation is as follows:

TABLE 7. Comparative results of case study.

COCOMO-II (In-house)	COCOMO-II (Distributed)	Amplified Model (Best Case)	Amplified Model (Worst Case)
39	110	182	986

$$PM = A * Size^{E} * \prod EM * \prod CCD \tag{2}$$

We added 14 cost drivers to the model and named them ''Critical Cost Drivers of GSD.'' The identified cost drivers are categorized as people-related, process-related, project-related, and outsourcing factors. Subsequently, these cost drivers were quantified based on the metrics presented in Phase 2. The estimates obtained through the amplified model are discussed in Phase 3.

C. PHASE 3

In this phase, we discuss the estimates based on the amplified model. Their variation with the COCOMO-II model is discussed. An illustrative example has been taken and applied in our context to achieve the targeted objective.

1) ESTIMATES BASED ON ADDITIONAL COST DRIVERS

This section provided a simplified cost estimation example in the GSD context and applied the proposed approach to it. The illustrative example is adopted from [21], where a company wants to develop software through offshore development. The estimation KSLOC of the project is 50, and for simplification, the value of the constant ''A'' is not calibrated.

The results indicate that after the addition of Critical Cost Drivers of GSD, the project's effort increased 50% in the best case than the COCOMO-II [16]. The effort is increased around eight times in the worst case assumed that the wages are eight times higher in the USA than in Pakistan or India. The comparison of SF and EM value w.r.t existing COCOMO-II and Amplified COCOMO-II are represented in Figure 11. The variation in the values depicts the distinct characteristics of GSD. The traditional cost estimation models are not applicable for the GSD context, and their values vary in the context of GSD.

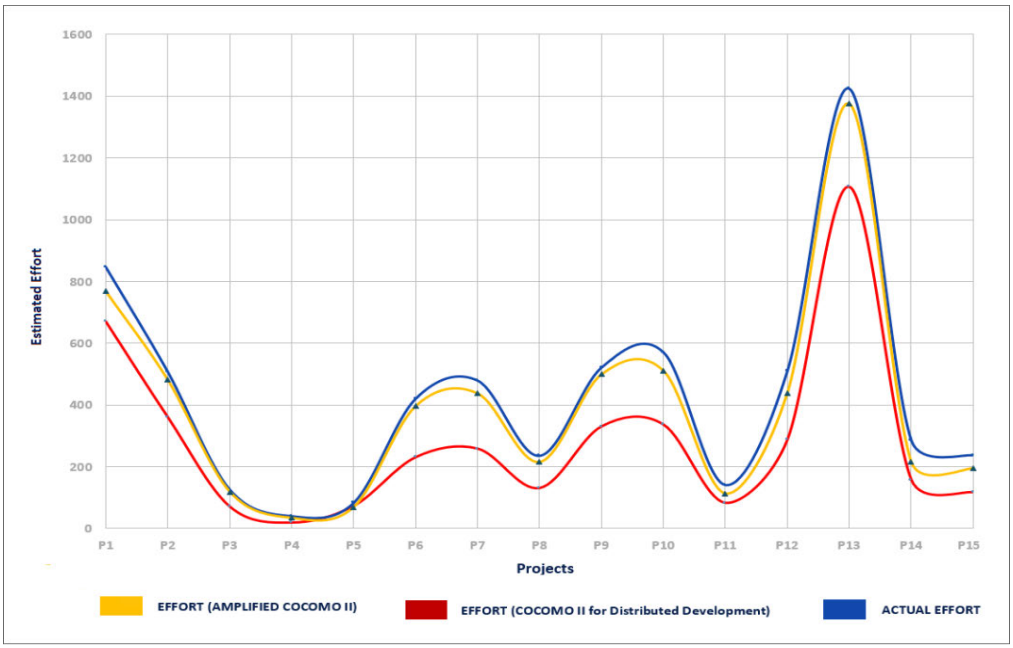


FIGURE 9. Comparison of estimated effort.

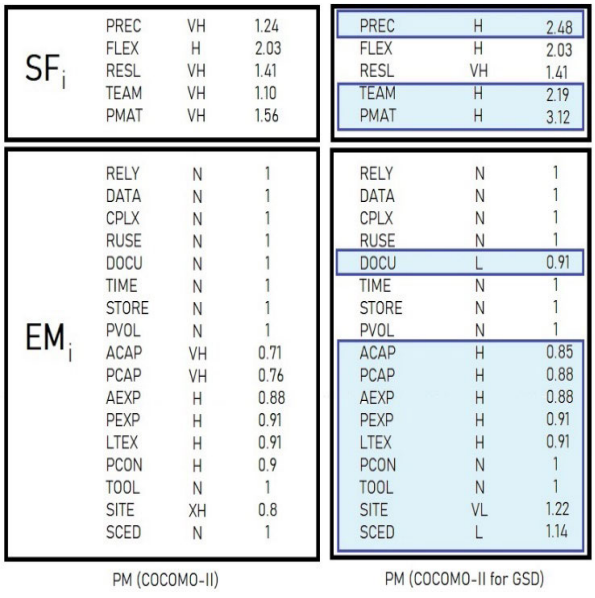


FIGURE 10. Comparison of SF and EM values.

2) COMPARISON OF PROPOSED MODEL

To compare the proposed model with the existing model, various projects were gathered by distributing a questionnaire. The actual effort of the projects was extracted through the questionnaire [27]. The estimated effort of the corresponding models was derived based on the equations mentioned in the above sections. The values of the estimated effort are depicted in Table 8.

The main reason behind the variation in the results is considering the additional cost drivers of GSD. The existing techniques lack quantifying the cost drivers as it is regarded as a difficult task. The comparison of the values of estimated effort is presented in Table 8. Figure 10 represents the visualization of the estimated values through a chart where key points represent different projects. It represents the projects' actual effort and the estimated effort through

TABLE 8. The comparison of the actual and estimated effort through the considered models.

Project	Actual Effort	Estimated Effort through COCOMO-II (Existing)	Estimated Effort through Amplified COCOMO-II for GSD (Proposed)
1	670	846.83	768.12
2	360	507.17	480.69
3	70	125.51	118.10
4	18	39.36	33.22
5	72	82.49	68.14
6	230	420.11	395.61
7	258	479.06	436.92
8	130	235.10	214.46
9	330	521.21	498.73
10	336	570.66	510.11
11	82	140.41	112.24
12	287	510.09	438.83
13	1107	1425.12	1375.95
14	157	285.06	214.28
15	116	236.19	193.51

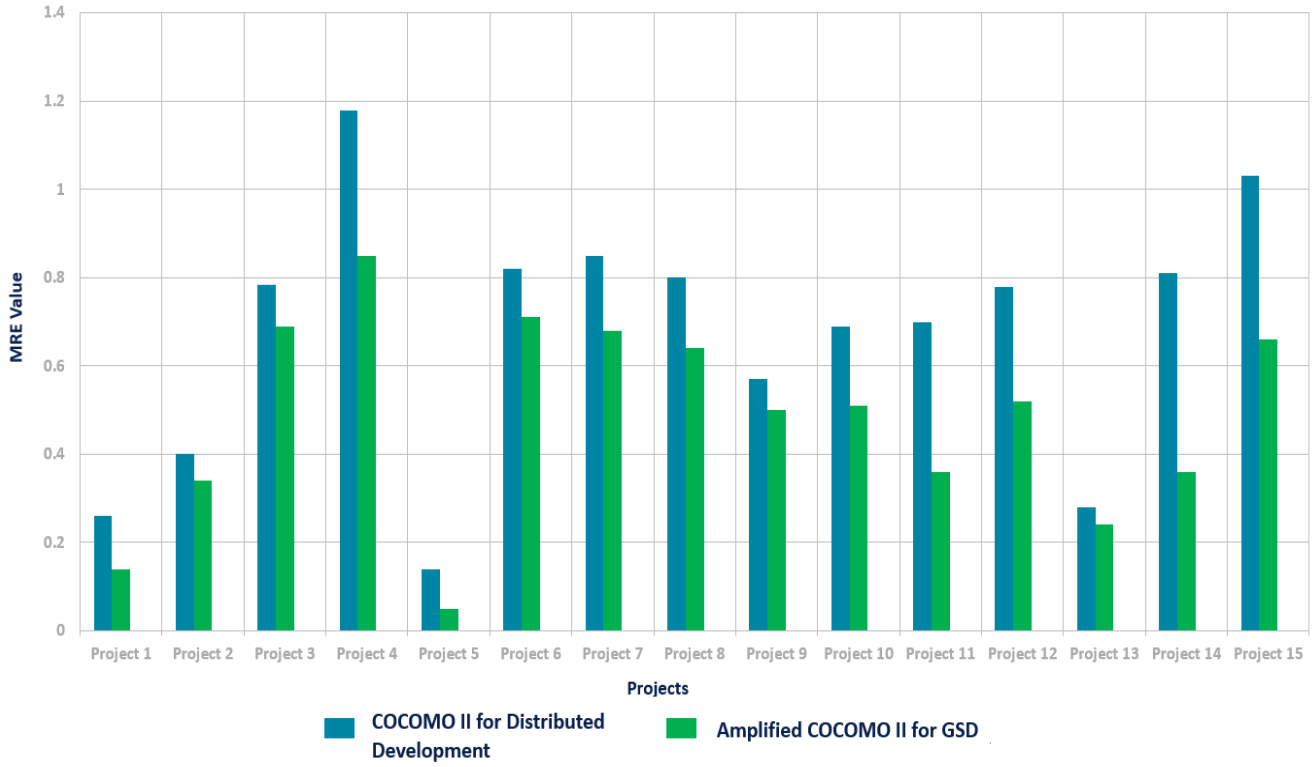


FIGURE 11. Comparison of MRE values.

COCOMO-II for distributed development and its amplified version. The estimation line shows the improvement in the estimated effort through the consideration of CCDs.

We adopted the hypothesis testing technique to validate the proposed estimation model. To achieve the targeted objective, the null hypothesis is formulated. The null hypothesis and alternative hypothesis designed in this research context are as follows:

Null Hypothesis (H_0): *There is no significant difference between the MRE values of the existing and the proposed models.*

Alternative Hypothesis (H_1): *There is a significant difference between the MRE values of the existing and the proposed models.*

Moreover, we adopted the accuracy measures to calculate the cost estimation models' performance and conducted parametric testing to evaluate the above-mentioned null hypothesis. Notice that the standard measurement used to measure the cost estimation model's performance is the Magnitude of Relative Error (MRE) [27]. In the literature, several studies [44], [45], [27] have adopted MRE measures to check the accuracy of their proposed estimation models. The formulae of MRE is represented by Equation 3.

$$\text{MRE} = \frac{|\text{Actual Effort} - \text{Estimated Effort}|}{\text{Actual Effort}} \tag{3}$$

The MRE values of the projects calculated through the above equation are presented in Table 9.

From Table 9, it can be observed that there is a gradual decrease in the Amplified model's MRE values compared to the existing COCOMO-II model for the distributed development environment. The decline in MRE values assures the increased accuracy of the proposed model. The variation in the MRE values indicates the change in the estimated effort of the projects. The estimated effort of the project changes due to the incorporated critical cost drivers of GSD. Thus, the values presented in Tables 8 and 9 indicate that the critical cost drivers significantly impact the estimated effort and MRE values. Ultimately, it rejects the null hypothesis. Finally, we selected the alternate hypothesis, i.e., the incorporated critical cost drivers significantly impact the GSD projects' estimated effort.

3) PARAMETRIC TESTING FOR VALIDATION

Considering the normal distribution of the obtained values, parametric testing is performed. Moreover, to validate the null hypothesis, we performed a paired t-test on the obtained MRE values (Table 9). Note that we have adopted the validation mechanism from a similar previously conducted work [46], as it is performed to evaluate the significant difference between two measurements [46]. The null hypothesis (H_0) states that there is no difference between the MRE values of the existing and proposed models. In contrast, the alternative hypothesis represents that there is a difference between the MRE values of the existing and the proposed models. The independent variables are represented by the estimated

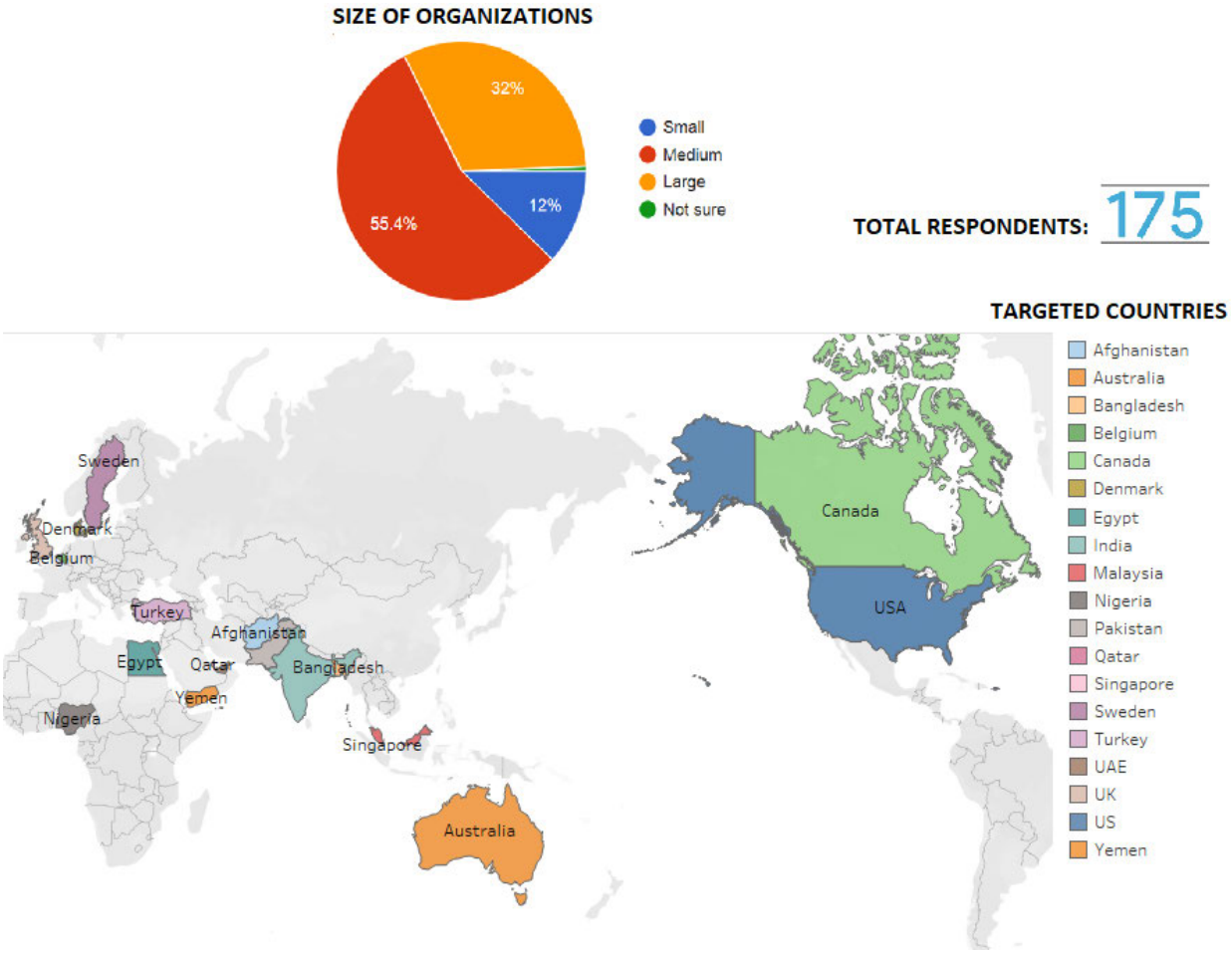


FIGURE 12. Demographics of respondents.

and actual effort of the existing and proposed models, whereas the MRE values represent the dependent variables. Table 10 presents the obtained results of the conducted paired t-test.

The paired t-test results depict t = 5.721 and a p-value of 0.000053, which is significantly less than 0.05. As the resultant alpha (p-value) is less than the adjusted statistical level (0.05). Thus, we can conclude that the value of the mean is significantly different in the two datasets. Hence, by conventional means ($p < 0.05$), i.e., the null hypothesis is rejected. Figure 12 represents the Barchart of the comparative values of MRE. The horizontal axis represents the considered projects, while the vertical axis denotes the MRE values of the projects (Figure 12). Notice that the low MRE value of a project represents the higher accuracy of the estimated value in this research context. Through low MRE values, it could be observed that the estimated effort resides near to the actual effort of the project.

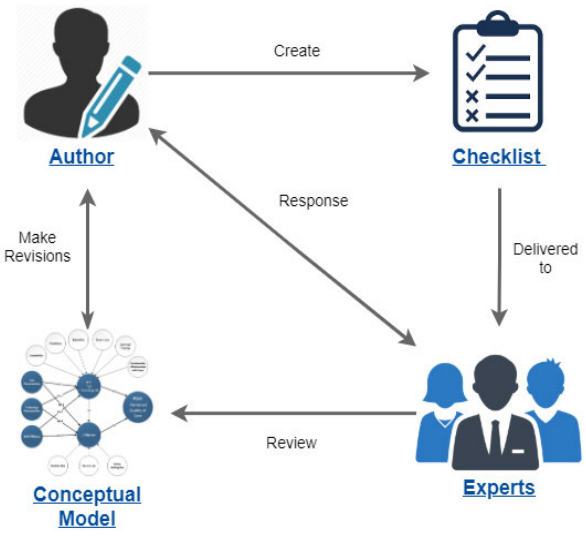


FIGURE 13. Expert validation.

4) VALIDATION THROUGH EXPERT OPINION

We have also adopted expert validation to validate the proposed conceptual model, also known as expert opinion. In expert validation, the industry experts are selected to validate the model; then, the experts decide to reject, accept,

FIGURE 14. Expert validation.

TABLE 9. The obtained MRE results.

Project	MRE of COCOMO II for GSD (Existing)	MRE of Amplified COCOMO-II for GSD (Proposed)
1	0.26	0.14
2	0.40	0.34
3	0.79	0.69
4	1.18	0.85
5	0.14	0.05
6	0.82	0.71
7	0.85	0.68
8	0.80	0.64
9	0.57	0.50
10	0.69	0.51
11	0.70	0.36
12	0.78	0.52
13	0.28	0.24
14	0.81	0.36
15	1.03	0.66

or review it. To find valid experts, we have followed criteria, and the criteria are listed below. The expert must be working as a ''Project Manager, and the expert must have at least ten years of experience. Figure 13 represents the adopted expert validation process.

As a result, five experts meet the inclusion criteria. The first expert Mr. Jamil Ahmed has 20 years of experience as a Software Project Manager. Secondly, Mr. Agha Hassan Afzaal khan has 11 years of experience as a Software Project Manager. Thirdly, Mr. Waris Mirza, who worked as a project manager for the last 18 years. He is currently working at Virtual Remittance Gateway (VRG). Fourthly, Dr. Khizar Mahmood has 13 years of experience as a project manager and currently works as a senior project manager in a DAR Middle East technology company.

We have attained the expert's response through a questionnaire. In the questionnaire following aspects were covered related to design, logical relations, labeling, and identified cost drivers:

- The overall design is good enough or needs some improvements?
- Do the phases present in this proposed model are relevant to each other?
- The order of the phases is correct or wanted to change the order.
- The components presented in each phase are related to the phases?
- The information presented in different components is enough?
- Do we have correctly labeled the phases?
- We have identified the cost drivers of GSD and presented them in the proposed conceptual model. Let us know

TABLE 10. Paired T-test statistics.

	Paired Differences					t	df	Sig. (2-tailed)
	Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference Lower	Upper			
Pair [Existing – Proposed]	0.19000	0.12862	.03321	0.11877	0.26123	5.721	14	0.000053

if these identified cost drivers are correct or need refinement?

The experts reviewed the conceptual model based on the checklist presented above. The reviews of the experts were accommodated for the improvement of the model. The dashboard for the obtained responses of experts is represented in Figure 14.

VII. RESEARCH IMPLICATIONS

The practical implication of our research is targeted to the researchers and the practitioners stated as follows:

- The extensive review of current state-of-the-art cost estimation techniques could help researchers understand the cost estimation process in the GSD context from various perspectives.
- The proposed conceptual model could be helpful to be served as a guideline for presenting a new model in the GSD context as it contains all the primary phases of amplification.
- The identified hidden cost drivers could be used to estimate the overhead of the GSD projects
- The mathematical model of cost estimation could be helpful for practitioners, particularly working on GSD projects.

VIII. CONCLUSION AND FUTURE WORK

In this paper, we addressed the issues of cost estimation in the GSD context. We presented an approach based on COCOMO-II post architecture to estimate the effort in a distributed environment to achieve the targeted objective. The empirically validated cost drivers were integrated into the model for the amplification. The additional cost drivers were extracted and validated in our previously published work [11]. Current work extends our previous work and quantifies the identified factors. The main phases of amplification include identifying cost drivers, categorizing cost drivers, forming metrics, assignment of values, and finally altering the equation of the base model. Moreover, for the validation of the proposed model, expert judgment and MRE measures were used. The work is still at an early stage and needs further calibration and validation. But even in traditional development models, it is impossible to get accurate and precise estimates [16]. For simplicity, the proposed model considers the values in ranges, not exact values, which is a limitation of this approach. But it provides estimation based on the additional critical cost drivers of GSD by reducing the overall risk of the project. The future work of this research is to develop a mathematical tool based on the formulated equation.

APPENDIX (DEMOGRAPHICS OF RESPONDENTS)

Dataset (Published on Mendeley): https://data.mendeley.com/datasets/m2zr2ns5k7/1

ACKNOWLEDGMENT

The authors are sincerely thankful to the Software Reliability Engineering Group (SREG) members at COMSATS University Islamabad, who have provided them with their feedback and critical analysis of the research. Moreover, they would like to extend their appreciation to the project managers and the experts who participated in their survey and provided them with a timely response.

REFERENCES

[1] S. McCarthy, P. O'Raghallaigh, C. Fitzgerald, and F. Adam, "Understanding and shared commitment in agile distributed ISD project teams," Tech. Rep., 2019.

[2] E. Ó. Conchúir, P. J. Ågerfalk, H. H. Olsson, and B. Fitzgerald, "Global software development: Where are the benefits?" *Commun. ACM*, vol. 52, no. 8, pp. 127–131, Aug. 2009, doi: 10.1145/1536616.1536648.

[3] J. D. Herbsleb, "Global software engineering: The future of socio-technical coordination," in *Proc. FoSE Futur. Softw. Eng.*, May 2007, pp. 188–198, doi: 10.1109/FOSE.2007.11.

[4] A. Tariq and A. A. Khan, "Framework supporting team and project activities in global software development (GSD)," in *Proc. Int. Conf. Emerg. Technol.*, Oct. 2012, pp. 334–339, doi: 10.1109/ICET.2012.6375435.

[5] A. A. Khan, J. Keung, M. Niazi, S. Hussain, and M. Shameem, "GSEPIM: A roadmap for software process assessment and improvement in the domain of global software development," *J. Softw., Evol. Process*, vol. 31, no. 1, pp. 1–12, 2019, doi: 10.1002/smr.1988.

[6] G. Isern, "Intercultural project management for IT: Issues and challenges," *J. Intercultural Manage.*, vol. 7, no. 3, pp. 53–67, Sep. 2015, doi: 10.1515/joim-2015-0021.

[7] M. El Bajta, A. Idri, J. N. Ros, J. L. Fernandez-Aleman, J. M. C. de Gea, F. Garcia, and A. Toval, "Software project management approaches for global software development: A systematic mapping study," *Tsinghua Sci. Technol.*, vol. 23, no. 6, pp. 690–714, Dec. 2018, doi: 10.26599/TST.2018.9010029.

[8] D. Wickramaarachchi and R. Lai, "Effort estimation in global software development—A systematic review," *Comput. Sci. Inf. Syst.*, vol. 14, no. 2, pp. 393–421, 2017, doi: 10.2298/CSIS160229007W.

[9] J. Mäkiö and S. Betz, "A system dynamics perspective into offshore software outsourcing—Uncovering correlations between critical success factors," Tech. Rep., 2004, pp. 1–9.

[10] E. Ó Conchúir, H. Holmström Olsson, P. J. Ågerfalk, and B. Fitzgerald, "Benefits of global software development: Exploring the unexplored," *Softw. Process, Improvement Pract.*, vol. 14, no. 4, pp. 201–212, Jul. 2009, doi: 10.1002/spip.417.

[11] J. A. Khan, S. U. R. Khan, J. Iqbal, and I. U. Rehman, "Empirical investigation about the factors affecting the cost estimation in global software development context," *IEEE Access*, vol. 9, pp. 22274–22294, 2021, doi: 10.1109/access.2021.3055858.

[12] A. M. Majanoja, L. Linko, and V. Leppänen, "Developing offshore outsourcing practices in a global selective outsourcing environment—The IT supplier's viewpoint," *Int. J. Inf. Syst. Proj. Manag.*, vol. 5, no. 1, pp. 27–43, 2017, doi: 10.12821/ijispm050102.

[13] S. M. A. Suliman and G. Kadoda, "Factors that influence software project cost and schedule estimation," in *Proc. Sudan Conf. Comput. Sci. Inf. Technol. (SCCSIT)*, Nov. 2017, pp. 1–9, doi: 10.1109/SCCSIT.2017.8293053.

[14] R. Jain and U. Suman, "A project management framework for global software development," *ACM SIGSOFT Softw. Eng. Notes*, vol. 43, no. 1, pp. 1–10, Mar. 2018, doi: 10.1145/3178315.3178329.

[15] R. Kraus, "IT offshoring—A cost-oriented analysis," Tech. Rep., Jul. 2021.

[16] S. Betz and J. Mäkiö, "Amplification of the COCOMO II regarding offshore software projects," in *Offshoring of Software Development*. Munich, Germany: OUTSHORE, 2008, p. 33. [Online]. Available: https://books.google.com.pk/books?hl=en&lr=&id=81LP4M0iv5oC&oi=fnd&pg=PA33&dq=Amplification+of+the+COCOMO+II+regarding+offshore+software+projects&ots=p3f4o8Q0YM&sig=sQbSqNkH8NntlOdE1PXMaxUVqSQ#v=onepage&q=AmplificationoftheCOCOMO IIregarding offshore

[17] J. Koskenkyla, *Cost Estimation in Global Software Development—Review of Estimation Techniques*, vol. 109, 2012.

[18] B. Kitchenham, *Procedures for Performing Systematic Reviews*, vol. 33. Keele, U.K.: Keele Univ., 2004, pp. 1–26, 2004.

[19] B. Kitchenham, O. Pearl Brereton, D. Budgen, M. Turner, J. Bailey, and S. Linkman, "Systematic literature reviews in software engineering—A systematic literature review," *Inf. Softw. Technol.*, vol. 51, no. 1, pp. 7–15, Jan. 2009, doi: 10.1016/j.infsof.2008.09.009.

[20] C. E. L. Peixoto, J. L. N. Audy, and R. Prikladnicki, "Effort estimation in global software development projects: Preliminary results from a survey," in *Proc. 5th IEEE Int. Conf. Global Softw. Eng.*, Aug. 2010, pp. 123–127, doi: 10.1109/ICGSE.2010.22.

[21] T. F. C. Tait and E. H. M. Huzita, "Software project management in distributed software development context," in *Proc. 15th Int. Conf. Enterp. Inf. Syst.*, vol. 2, 2013, pp. 216–222, doi: 10.5220/0004442402160222.

[22] T. A. Khalid and E.-T. Yeoh, "Early cost estimation of software reworks using fuzzy requirement-based model," in *Proc. Int. Conf. Commun., Control, Comput. Electron. Eng. (ICCCCEE)*, Jan. 2017, pp. 1–5, doi: 10.1109/ICCCCEE.2017.7866082.

[23] M. El Bajta, A. Idri, J. L. Fernández-Alemán, J. Nicolas Ros, and A. Toval, "Software cost estimation for global software development—A systematic map and review study," in *Proc. 10th Int. Conf. Eval. Novel Approaches Softw. Eng.*, 2015, pp. 197–206, doi: 10.5220/0005371501970206.

[24] R. Britto, V. Freitas, E. Mendes, and M. Usman, "Effort estimation in global software development: A systematic literature review," in *Proc. IEEE 9th Int. Conf. Global Softw. Eng.*, Aug. 2014, pp. 135–144, doi: 10.1109/ICGSE.2014.11.

[25] R. Madachy, "Distributed global development parametric cost modeling," in *Software Process Dynamics and Agility (ICSP)* (Lecture Notes in Computer Science), vol. 4470. Berlin, Germany: Springer, 2007, pp. 159–168, doi: 10.1007/978-3-540-72426-1_14.

[26] P. Keil, D. J. Paulish, and R. S. Sangwan, "Cost estimation for global software development," in *Proc. Int. Workshop Econ. Driven Softw. Eng. Res. (EDSER)*, 2006, pp. 7–10, doi: 10.1145/1139113.1139117.

[27] L. V. Patil, R. M. Waghmode, S. D. Joshi, and V. Khanna, "Generic model of software cost estimation: A hybrid approach," in *Proc. IEEE Int. Advance Comput. Conf. (IACC)*, Feb. 2014, pp. 1379–1384, doi: 10.1109/IAdCC.2014.6779528.

[28] M. Azzeh, "Software cost estimation based on use case points for global software development," in *Proc. 5th Int. Conf. Comput. Sci. Inf. Technol.*, Mar. 2013, pp. 214–218, doi: 10.1109/CSIT.2013.6588782.

[29] M. E. Bajta, "Analogy-based software development effort estimation in global software development," in *Proc. IEEE 10th Int. Conf. Global Softw. Eng. Workshops*, Jul. 2015, pp. 51–54, doi: 10.1109/ICGSEW.2015.19.

[30] M. Humayun and C. Gang, *111-K0003*, vol. 2, no. 3, 2012.

[31] S. Ramacharan and K. V. G. Rao, "Parametric models for effort estimation for global software development," *Lect. Notes Softw. Eng.*, vol. 1, no. 2, pp. 178–182, 2013, doi: 10.7763/lnse.2013.v1.40.

[32] A. Lamersdorf, J. Münch, A. F.-D. V. Torre, C. R. Sánchez, and D. Rombach, "Estimating the effort overhead in global software development," in *Proc. 5th Int. Conf. Glob. Softw. Eng. (ICGSE)*, Aug. 2010, pp. 267–276, doi: 10.1109/ICGSE.2010.38.

[33] S. M. R. Chirra and H. Reza, "A survey on software cost estimation techniques," *J. Softw. Eng. Appl.*, vol. 12, no. 6, pp. 226–248, 2019, doi: 10.4236/jsea.2019.126014.

[34] S. Shekhar and U. Kumar, "Review of various software cost estimation techniques," *Int. J. Comput. Appl.*, vol. 141, no. 11, pp. 31–34, May 2016, doi: 10.5120/ijca2016909867.

[35] S. Bibi and I. Stamelos, "Selecting the appropriate machine learning techniques for the prediction of software development costs," *IFIP Int. Fed. Inf. Process.*, vol. 204, pp. 533–540, 2006, doi: 10.1007/0-387-34224-9_62.

[36] S. Ramacharan and K. V. G. Rao, "Scheduling based cost estimation model: An effective empirical approach for GSD project," in *Proc. 13th Int. Conf. Wireless Opt. Commun. Netw. (WOCN)*, Jul. 2016, pp. 1–4, doi: 10.1109/WOCN.2016.7759881.

[37] J. Ahmad, A. W. Khan, and I. Qasim, *Software Outsourcing Cost Estimation Model (SOCEM). A Systematic Literature Review Protocol*, vol. 2, no. 1, 2018.

[38] R. Britto, "Knowledge classification for supporting effort estimation in global software engineering projects," Tech. Rep., 2015.

[39] S. Ramacharan and K. V. G. Rao, "Software effort estimation of GSD projects using calibrated parametric estimation models," in *Proc. 2nd Int. Conf. Inf. Commun. Technol. Competitive Strategies (ICTCS)*, 2016, pp. 1–8, doi: 10.1145/2905055.2905177.

[40] Z. T. Abdulmehdi, M. S. S. Basha, M. Jameel, and P. Dhavachelvan, "A variant of COCOMO II for improved software effort estimation," *Int. J. Comput. Electr. Eng.*, vol. 6, no. 4, pp. 346–350, 2014, doi: 10.7763/ijcee.2014.v6.851.

[41] P. R. Raj, P. R. Haji, F. Rizvi, and N. Khan, "COCOMO and COCOMO II model—A case study," Tech. Rep., May 2020, pp. 2665–2668.

[42] N. Yadav, N. Gupta, M. Aggarwal, and A. Yadav, "Comparison of COSYSMO model with different software cost estimation techniques," in *Proc. Int. Conf. Issues Challenges Intell. Comput. Techn. (ICICT)*, Sep. 2019, pp. 1–5, doi: 10.1109/ICICT46931.2019.8977686.

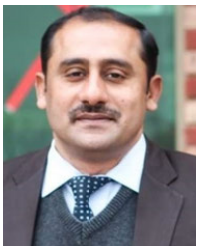
[43] B. Clark, S. Devnani-Chulani, and B. Boehm, "Calibrating the COCOMO II post-architecture model," in *Proc. 20th Int. Conf. Softw. Eng.*, Apr. 1998, pp. 477–480, doi: 10.1109/icse.1998.671610.

[44] A. Trendowicz, J. Heidrich, J. Münch, Y. Ishigai, K. Yokoyama, and N. Kikuchi, "Development of a hybrid cost estimation model in an iterative manner," in *Proc. 28th Int. Conf. Softw. Eng.*, May 2006, pp. 331–340, doi: 10.1145/1134285.1134332.

[45] W. L. Du, L. F. Capretz, A. B. Nassif, and D. Ho, "A hybrid intelligent model for software cost estimation," *J. Comput. Sci.*, vol. 9, no. 11, pp. 1506–1513, Nov. 2013, doi: 10.3844/jcssp.2013.1506.1513.

[46] K. Rak, Ž. Car, and I. Lovrek, "Effort estimation model for software development projects based on use case reuse," *J. Softw., Evol. Process*, vol. 31, no. 2, pp. 1–17, 2019, doi: 10.1002/smr.2119.

JUNAID ALI KHAN received the B.E. degree in software engineering from COMSATS University Islamabad, Wah Campus, Pakistan, in 2017, where he is currently pursuing the M.S. degree. His research interests include software project management, software process improvement, and their application in global software development context.



SAIF UR REHMAN KHAN received the Ph.D. degree in software engineering from the University of Malaya, Kuala Lumpur, Malaysia, in 2018. He is currently serving with the Department of Computer Science, COMSATS University Islamabad (CUI), Islamabad, Pakistan. His research interests include software engineering include verification and validation, search-based software engineering, cyber-physical systems, requirements engineering, and software project management. He has been in several expert review panels, both locally and internationally. He was a recipient of Best Paper Presentation Award with the Faculty of Computer Science and Information Technology, UM, in 2014, and a Certificate of Outstanding Contribution in Reviewing (Future Generation Computer Systems), in 2018.

TAMIM AHMED KHAN received the B.E. degree (Hons.) in software engineering from The University of Sheffield, U.K., in 1995, the M.B.A. degree in finance and accounting from Preston University, Islamabad, Pakistan, in 1997, the M.S. degree in computer engineering from CASE, Taxila University, Pakistan, in 2006, and the Ph.D. degree in software engineering from the University of Leicester, U.K., in 2012. He is currently serving as a Professor with the Department of Software Engineering, Bahria University, Islamabad. His research interests include service oriented architectures, e-learning, and software quality assurance.

INAYAT UR REHMAN KHAN received the Ph.D. degree in computer science (e-learning) from COMSATS University Islamabad (CUI), Islamabad, Pakistan, in 2017. He has over 18 years of teaching experience, and he is currently working as an Assistant Professor with the Department of Computer Science, CUI, Islamabad. His research interests include software engineering include software project management, and software reusability. In the e-learning domain, his research interests also include computer-assisted core interests are education, designing learning tools, computer animations for learning, HCI for design in learning tools, cognitive learning, and the use of educational psychology for e-learning applications. He is extensively involved in conducting training for basic computing courses for national and multinational companies.