

**UNCERTAINTY QUANTIFICATION IN RESERVOIR
SIMULATION WITH MONTE CARLO SIMULATION
AND POLYNOMIAL CHAOS EXPANSTION**

BY
TENGKU MUHAMMAD SOFYAN ASTSAURI

A Thesis Presented to the
DEANSHIP OF GRADUATE STUDIES

KING FAHD UNIVERSITY OF PETROLEUM & MINERALS

DHAHRAN, SAUDI ARABIA

In Partial Fulfillment of the
Requirements for the Degree of

MASTER OF SCIENCE

In

PETROLEUM ENGINEERING

NOVEMBER 2024

KING FAHD UNIVERSITY OF PETROLEUM & MINERALS

DHAHRAN- 31261, SAUDI ARABIA

DEANSHIP OF GRADUATE STUDIES

This thesis, written by **Tengku Muhammad Sofyan Astsauri** under the direction of his thesis advisor and approved by his thesis committee, has been presented and accepted by the Dean of Graduate Studies, in partial fulfillment of the requirements for the degree of **MASTER OF SCIENCE IN PETROLEUM ENGINEERING.**



Dr. Yuzhu Wang
(Advisor)



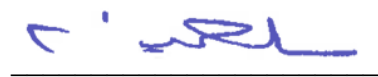
Dr. Dhafer Al-Shehri
Department Chairman



Prof. Abdulazeez Abdulraheem
(Member)



Dr. Ahmed Ibrahim
(Member)



Prof. Suliman Saleh Al-Homidan
Dean of Graduate Studies

Date



© Tengku Muhammad Sofyan Astsauri

2024

Dedicated to my parents, siblings, and everyone who love me

ACKNOWLEDGMENTS

All praise is to Almighty Allah, the Most Compassionate and Merciful, for granting me health, knowledge, and the perseverance needed to complete my thesis work. May blessings and peace be upon Prophet Muhammad (peace be upon him), his family, and his companions.

I wish to express my heartfelt appreciation to all those who provided support and guidance during the research and completion of this thesis.

I am profoundly thankful to my thesis advisor, Dr. Yuzhu Wang, for his steadfast support, encouragement, patience, and invaluable insights over the past two years. His dedication and guidance have been essential in the development of this work. I would also like to express my heartfelt gratitude to my thesis committee members, Prof. Abdulazeez Abdulraheem and Dr. Ahmed Ibrahim, for their support and insightful feedback, which has greatly improved the quality of this thesis.

I would like to convey my deepest appreciation to my friends for their ongoing encouragement and understanding throughout this academic journey. Your support provided me with constant motivation. I am also thankful to the KFUPM Indonesian community; though I cannot list each name individually, your presence has been invaluable.

Lastly, I want to extend my heartfelt thanks to my amazing family: my parents, Papa and Mama, for their constant prayers, love, and encouragement, as well as my brothers and sister, for keeping my spirits high throughout my master's journey. I am also deeply

*appreciative of all the staff, faculty, and professors at CPG-KFUPM who have guided
and supported me along this academic path.*

TABLE OF CONTENTS

ACKNOWLEDGMENTS	V
TABLE OF CONTENTS.....	VII
LIST OF TABLES.....	X
LIST OF FIGURES.....	XI
ABSTRACT	XIII
ملخص الرسالة	XIV
CHAPTER 1 INTRODUCTION.....	1
1.1 Aim of This Study.....	2
1.2 Problem Statement	2
1.3 Innovation of The Study	3
CHAPTER 2 LITERATURE REVIEW	4
2.1 Uncertainty Quantification	4
2.2 Monte Carlo Simulation.....	5
2.3 Polynomial Chaos Expansion	11

CHAPTER 3 METHODOLOGY.....	18
3.1 Reservoir Model Description.....	19
3.2 Sensitivity Analysis.....	24
3.3 Monte Carlo Simulation	25
3.4 Polynomial Chaos Expansion (PCE).....	26
3.4.1 PCE Coefficient Calculation	26
3.4.2 PCE Posterior Error Estimation	28
3.4.3 Moment of PCE: Mean and Variance	28
3.4.4 Moment of Stochastic Approach (LHS): Mean and Variance	29
CHAPTER 4 RESULT AND DISCUSSION	30
4.1 Monte Carlo Simulation	30
4.2 Polynomial Chaos Expansion	34
4.2.1 PCE Degree 5 – Varying Sample	40
4.2.2 PCE Degree 7 & 9 – Varying Sample	44
4.2.3 PCE Degree 11 – Varying Sample	51
4.3 Novel Correlation for Degree – Sample Relationship.....	56
CHAPTER 5 CONCLUSION AND FUTURE WORK	61
5.1 Conclusion	61
5.2 Future Works	62

REFERENCES.....	63
VITAE.....	69

LIST OF TABLES

Table 3.1. Brief Reservoir description PUNQ-S3.....	20
Table 3.2. Statistic information of the random input parameters used	21
Table 4.1. Summary statistic Monte Carlo simulation result.....	32
Table 4.2. Top 3 significant random parameters.....	34
Table 4.3. Coefficient and parameter number for each PCE degree.....	35
Table 4.4. Error evaluation for all PCE degree model	37
Table 4.5. Varying number of samples for each polynomial degree	40
Table 4.6. Error evaluation PCE model degree 5 for all sample variations	41
Table 4.7. Error evaluation PCE model degree 7 for all sample variation	45
Table 4.8. Error evaluation PCE model degree 9 for all sample variation	46
Table 4.9. Error evaluation PCE model degree 11 for all sample variation	52
Table 4.10. Scenario number of sample used for PCE model degree 6, 8, and 10	58

LIST OF FIGURES

Figure 2.1. Proportion of MCS for uncertainty quantification (Grenyer et al., 2021).....	6
Figure 2.2. General flowchart for MCS (Hu et al., 2016).....	7
Figure 2.3. Illustration of forward uncertainty quantification (UQ)	7
Figure 2.4. Surrogate model construction workflow procedure using PCE	12
Figure 2.5. PCE basis functions for each type distribution (Camacho et al., 2017)	13
Figure 3.1. The overall workflow of the study	18
Figure 3.2. 3D reservoir model (PUNQ-S3 model)	19
Figure 3.3. Probability density plot of random input parameters	23
Figure 4.1. Monte Carlo Simulation result for cumulative oil production	31
Figure 4.2. Convergence plot for each statistical properties of MCS	32
Figure 4.3. Sobol sensitivity analysis – main effect of the input parameters	33
Figure 4.4. Summary PCE performance for all PCE model	36
Figure 4.5. Error evaluation plot for all PCE model.....	38
Figure 4.6. Residual plot for all PCE model	39
Figure 4.7. Summary PCE degree 5 performance for all varying sample	41
Figure 4.8. Error evaluation PCE model degree 5 for all sample variations	42
Figure 4.9. Residual plot PCE model degree 5 for all sample variations	43
Figure 4.10. Summary PCE degree 7 performance for all varying sample	45
Figure 4.11. Summary PCE degree 9 performance for all varying sample	46
Figure 4.12. Error evaluation PCE model degree 7 for all sample variation	48

Figure 4.13. Error evaluation PCE model degree 9 for all sample variation	49
Figure 4.14. Residual plot PCE model degree 7 for all sample variations	50
Figure 4.15. Residual plot PCE model degree 9 for all sample variations	51
Figure 4.16. Summary PCE degree 11 performance for all varying sample	52
Figure 4.17. Error evaluation PCE model degree 11 for all sample variation.....	53
Figure 4.18. Residual PCE model degree 11 for all sample variation	55
Figure 4.19. Running time summary of PCE model vs MCS.....	56
Figure 4.20. Cross-plot PCE model degree 6 for all sample variation	59
Figure 4.21. Cross-plot PCE model degree 8 for all sample variation	60
Figure 4.22. Cross-plot PCE model degree 10 for all sample variation	60

ABSTRACT

Full Name : Tengku Muhammad Sofyan Astsauri

Thesis Title : Uncertainty Quantification in Reservoir Simulation with Monte Carlo
Simulation and Polynomial Chaos Expansion

Major Field : Petroleum Engineering

Date of Degree : November 2024

Reservoir simulations often face significant input uncertainties due to geological, measurement, and modeling variations. Understanding how these uncertainties impact the model is essential for enhancing the reliability of reservoir management decisions. Polynomial Chaos Expansions (PCE) offer an effective regression-based, data-driven approach to analyze the propagation of input uncertainties through the model. The PCE represents the model's output as a polynomial based on spectral functions derived from the probability distributions of input variables. This approach enables the calculation of essential statistical characteristics, including the average, standard deviation, percentiles, and the complete probability distribution. However, complex scenarios may involve more random variables, require higher polynomial degrees and sample size to guarantee the estimation accuracy. The increasing number of random variables, polynomial degrees, and sample size will significantly increase the computation price. To mitigate these challenges, this study uses PUNQ S-3 model as an example to examine how the sample size and polynomial degree affect the performance of the PCE model especially. In addition, sobol analysis was used to assess input parameters' sensitivity to reduce the number of the input random variables. The PCE results are validated through comparison with Monte Carlo simulations (MCS) with 17,000 realizations, leading to some key conclusions.. First, the PCE method provides an efficient alternative to traditional MCS, achieving over 95% of MCS accuracy while requiring less than 1% of the computational time. Second, balancing polynomial degree and estimation accuracy is crucial, as higher-degree polynomials require more samples, significantly increasing computational time while offering limited accuracy improvement.

ملخص الرسالة

الاسم الكامل: تنكو محمد سفيان استسوري

عنوان الرسالة: القياس الكمي لعدم اليقين في محاكاة الخزان باستخدام محاكاة مونت كارلو وتوسيع الفوضى متعددة الحدود

التخصص: هندسة البترول

تاريخ الدرجة العلمية: نوفمبر 2024

غالبًا ما تواجه عمليات محاكاة المكامن شكوكًا كبيرة في المدخلات بسبب الاختلافات الجيولوجية والقياسية والنمذجة. يعد فهم كيفية تأثير أوجه عدم اليقين هذه على النموذج أمرًا ضروريًا لتعزيز موثوقية قرارات إدارة الخزان. توفر توسيعات الفوضى متعددة الحدود (PCE) نهجًا فعالاً يعتمد على الانحدار ويعتمد على البيانات لتحليل انتشار الشكوك في المدخلات من خلال النموذج. يمثل PCE مخرجات النموذج باعتباره متعدد الحدود استنادًا إلى الوظائف الطيفية المستمدة من التوزيعات الاحتمالية لمتغيرات المدخلات. يتيح هذا النهج حساب الخصائص الإحصائية الأساسية، بما في ذلك المتوسط، والانحراف المعياري، والنسب المئوية، والتوزيع الاحتمالي الكامل. ومع ذلك، قد تتضمن السيناريوهات المعقدة المزيد من المتغيرات العشوائية، وتتطلب درجات متعددة الحدود أعلى وحجم عينة لضمان دقة التقدير. سيؤدي العدد المتزايد من المتغيرات العشوائية، ودرجات متعددة الحدود، وحجم العينة إلى زيادة كبيرة في سعر الحساب. للتخفيف من هذه التحديات، تستخدم هذه الدراسة نموذج PUNQ S-3 كمثال لفحص كيفية تأثير حجم العينة ودرجة متعدد الحدود على أداء نموذج PCE بشكل خاص. بالإضافة إلى ذلك، تم استخدام تحليل سوبول لتقييم حساسية معلمات المدخلات لتقليل عدد المتغيرات العشوائية المدخلة. يتم التحقق من صحة نتائج PCE من خلال المقارنة مع عمليات محاكاة مونت كارلو (MCS) مع 17000 عملية تحقيق، مما يؤدي إلى بعض الاستنتاجات الرئيسية. أولاً، توفر طريقة PCE بدلاً من محاكاة MCS التقليدية، حيث تحقق أكثر من 95٪ من دقة MCS بينما تتطلب أقل من 1 ٪ من الوقت الحسابي. ثانياً، يعد الموازنة بين درجة كثيرات الحدود ودقة التقدير أمرًا بالغ الأهمية، نظرًا لأن كثيرات الحدود ذات الدرجة الأعلى تتطلب المزيد من العينات، مما يزيد بشكل كبير من الوقت الحسابي مع تقديم تحسين محدود في الدقة.

CHAPTER 1

INTRODUCTION

In science and engineering, decision-making often relies on predictions generated from computational or simulation models. Although these models aim to replicate real-world scenarios, they inherently include uncertainties stemming from diverse sources of variability, randomness, and stochastic behavior. Evaluating risk and uncertainty has been extensively studied in sectors like natural resources, particularly within the petroleum industry, to enhance the likelihood of project success (Witter et al., 2019).

It is widely recognised that decisions regarding the development of petroleum fields are complex, primarily because of challenges in accurately characterising reservoirs and representing them in simulation models. Consequently, gaining an accurate and efficient understanding of how this uncertainty influences the model is vital for boosting the reliability and effectiveness of decision-making in reservoir management (Camacho et al., 2017).

As recognised, uncertainty quantification plays a crucial role in reservoir management processes. Therefore, to enhance the accuracy of uncertainty quantification, this study is structured as follows:

1.1 Aim of This Study

The objective of this study is to evaluate the performance of two popular techniques for uncertainty quantification – Monte Carlo Simulation and polynomial Chaos Expansion – in reservoir simulations, yielding valuable insight into their effectiveness and efficiency.

1.2 Problem Statement

While Monte Carlo simulation is renowned for its simplicity, it has a major limitation in terms of computational cost, as it demands a huge amount of realizations to reach convergence. Alternatively, Polynomial Chaos Expansion (PCE) provides faster convergence and delivers comparable results. However, further investigation is needed to assess the performance of PCE models. Therefore, this study will address several key research questions in this area:

1. The efficiency and accuracy of Polynomial Chaos Expansion (PCE) compared to Monte Carlo simulation in the context of uncertainty quantification.
2. How degree order and sample size impact the PCE performance.
3. What is the relationship between polynomial degree and sample size in the high-order PCE model.

1.3 Innovation of The Study

This study aims to explore the recommended number of sample for higher-order polynomials. As higher-degree polynomials create more complex surrogate models, it is expected that significantly more sample points will be needed. In complex systems, especially those with nonlinear interactions or multiple interdependent variables, lower-degree polynomials may not adequately capture the full variability or uncertainty. By investigating this relationship, the study will provide valuable insights into how the number of sample correlates with the polynomial degree, particularly in higher-order models. |

CHAPTER 2

LITERATURE REVIEW

A comprehensive introduction to Uncertainty Quantification, emphasizing its significance and widespread application across various fields will be discussed in this chapter. Additionally, It will also cover two popularly used methods for uncertainty quantification: Monte Carlo Simulation and Polynomial Chaos Expansion. Each method will be discussed in detail, including the core concepts, advantages, disadvantages, and a review of previous research related to their use

2.1 Uncertainty Quantification

Uncertainty quantification (UQ) involves characterizing, integrating, and propagating the uncertainty in computational or simulation model. While these models are intended to approximate real-world conditions, they inherently contain uncertainties due to various sources of variability, randomness, and stochastic behavior. UQ specifically utilizes techniques from computational statistics and applied mathematics to address and manage uncertainties within complex computational models and simulations (J. Zhang, 2021).

The principles of uncertainty quantification have been extensively utilized in various fields such as aerospace, manufactures, energy and environment, petroleum engineering field, etc. This approach is driven by the aim of maximizing project success, resulting in more efficient resource extraction, improved risk management, and optimized operational strategies. By integrating UQ, engineers and decision-makers are better

equipped to forecast outcomes, evaluate the feasibility of different methods, and make well-informed decisions that promote both economic and environmental sustainability.

Monte Carlo Simulation (MCS) and Polynomial Chaos Expansion (PCE) stand as two widely embraced techniques in the realm of uncertainty quantification studies. MCS relies on random sampling to deduce the distribution patterns of physical or mathematical scenarios, serving as a cornerstone and simplistic approach for uncertainty analysis. This statistical method lays the groundwork for both risk assessment and uncertainty quantification (Cremon et al., 2020; Mehana et al., 2020a). However, To thoroughly analyze how the objective function varies under these uncertainties, a large number of simulations are typically needed, which may not be practical for large-scale models due to time constraints. Polynomial Chaos Expansion (PCE) provides an alternative approach to Monte Carlo Simulation (MCS) for uncertainty quantification, addressing MCS's intensive computational requirements by creating surrogate models. These surrogate models effectively capture the nonlinear relationships between input parameters and model responses, allowing for a substantial reduction in computational time (Jia et al., 2016, 2018; Naik et al., 2019).

2.2 Monte Carlo Simulation

Monte Carlo simulation is the most widely utilized method for assessing uncertainty, appearing in 63% of the 43 quantitative studies and 52% of all 107 papers reviewed (Grenyer et al., 2021). As shown in Figure 1, Monte Carlo is a prevalent technique applied across various fields for uncertainty propagation analysis. The Monte Carlo Simulation (MCS) approach is extensively applied in design, analysis, and optimization challenges that involve uncertainty, largely due to its non-intrusive nature.

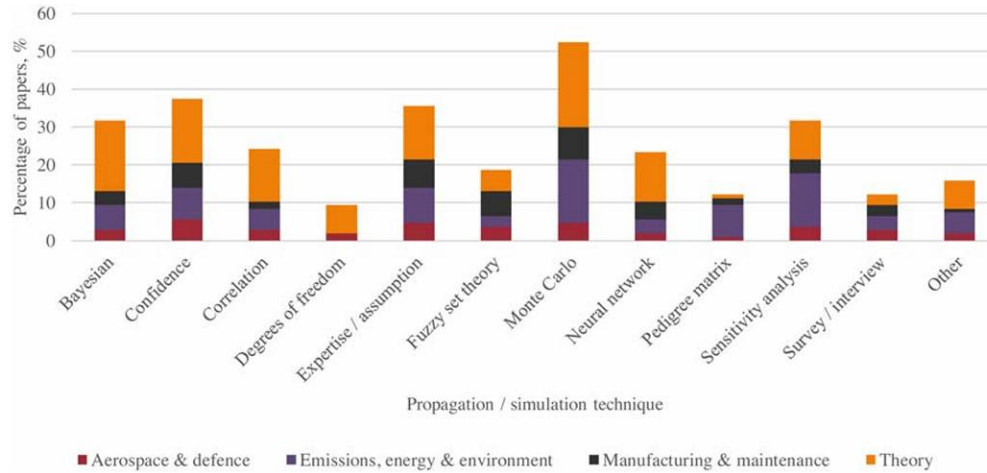


Figure 2.1. Proportion of MCS for uncertainty quantification (Grenyer et al., 2021)

Figure 2.2 depicts a standard flowchart of Monte Carlo-based Uncertainty Management and Design Optimization (UMDO). In a simple way, MCS produces multiple samples of input random variables based on their probability distributions, solving deterministic problems through simulations or computational models. The resulting output data from these samples are then aggregated to statistically evaluate the variability within the random system (Zhang, 2021). Through repeated sampling and simulations, MCS utilizes the Law of Large Numbers along with other statistical inference techniques to thoroughly describe the system's statistical behavior, thereby capturing various quantities of interest (figure 2.3). This approach allows for the analysis of system performance while accounting for the impact of uncertain variables.

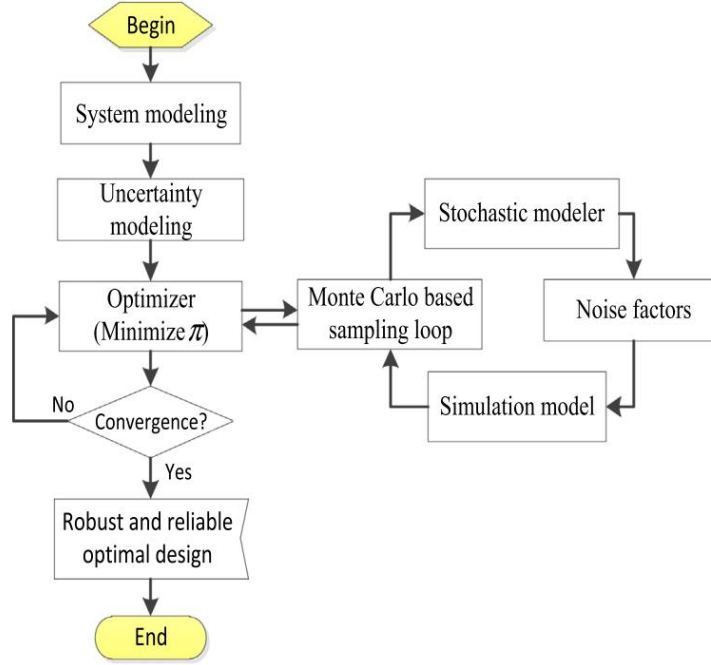


Figure 2.2. General flowchart for MCS (Hu et al., 2016)

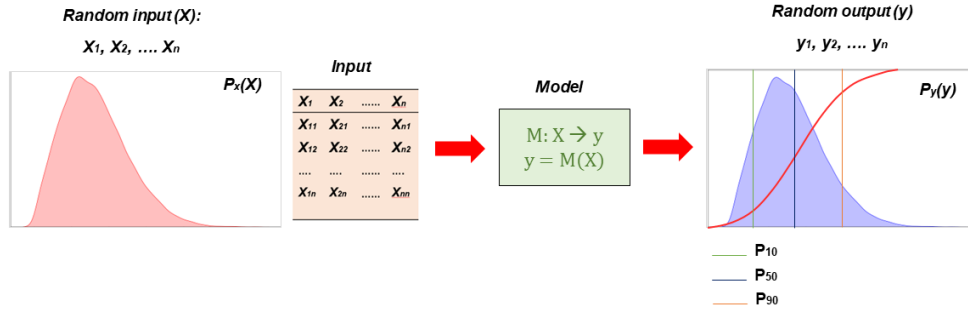


Figure 2.3. Illustration of forward uncertainty quantification (UQ)

Despite of simple and straightforward behaviour of MCS, to achieve dependable and convergent outcomes, it usually necessitates a significant number of forward simulations, due to its slow convergence rate of $\mathcal{O}(N^{-1/2})$, where N represents the number of function evaluations (Zhang, 2021). This becomes the primary disadvantage of monte

carlo simulation, as it demand high-performance numerical simulation framework and considerable computational power, especially when dealing with complex models that require extensive geological detail. According to Kroese & Rubinstein (2012) The convergence rate of this method is influenced by:

- 1) The count of stochastic series used.
- 2) The quantity of generated random values.
- 3) The probability distribution assigned to each uncertain variable.
- 4) The dependency level among the generated random variables (Mostafa, 2015)

Due to its impartiality at any accuracy level, Monte Carlo Simulation (MCS) can be used as a reference for evaluating the effectiveness of other methods (Hu et al., 2016). For instance, uncertainty analysis results are frequently expressed using the first moment, or the mean, though this generally does not convey the full structure of π . The simplest MCS algorithm entails sampling directly from the distribution function of x . An empirical estimate of the statistical properties π can be derived through ideal Monte Carlo sampling as described below:

$$\hat{\pi} = \frac{1}{N} \sum_{i=1}^N w_i f(X^{(i)}) \quad (2.1)$$

where N is random samples, $\{X^{(i)}; i = 1, \dots, N\}$, that distributed independently and generated with quadrature weights $\{w_i; i = 1, \dots, N\}$. Nearly certain convergence is achieved when $N \rightarrow \infty$.

The Monte Carlo simulation model has seen widespread use, with numerous researchers advocating for incorporating uncertain parameters in the Monte Carlo approach (figure 1). Since the initial studies on the method's relevance to the oil industry emerged in

the mid-1960s (Stoian, 1965), the Monte Carlo method has been applied across various areas of petroleum research. For instance, Dong et al (2013) combined Monte Carlo methods with an analytical reservoir simulator to evaluate the original gas in place (OGIP), technically recoverable resources (TRR), and recovery factor (RF) in shale gas reservoirs characterized by high uncertainty and risk. This model is expected to mitigate the limitations of the existing method such as decline curve analysis and volumetric method to be used in the early stage of production. Dai et al (2014) integrated a multi-phase reservoir simulator with geological and statistical models to assess reservoir heterogeneity and quantify uncertain parameters. This approach uses a Monte Carlo simulation framework in conjunction with modeling the flow and transport of CO₂, oil, and water. Using the Monte Carlo result as an input, the author successfully estimated the sensitive parameter during the CO₂-EOR implementation. Cremon et al (2020) analyzed the convergence behavior of Monte Carlo Simulation (MCS) on three-dimensional, three-phase reservoirs with significant heterogeneity, using actual field data. Their findings confirm that the convergence of target quantities adheres to a square-root law across the 10,000 realizations tested. They concluded that, as the number of realizations grows, the sample results increasingly align with the expected value, demonstrating a square-root rate of convergence. Mehana et al (2020) succeed in integrating decline curve analysis and the Monte Carlo Simulation Method to probabilistically estimate the EUR in a shale reservoir based on, as in this reservoir the Arps equation gives an unphysical exponent and leads to over-estimation. The result showed that the Monte Carlo Simulation required more than 100000 to reach the convergence state. Quosay et al (2020) introduced a novel uncertainty-based approach for optimising hydraulic fracturing in shale gas reservoirs. The authors

started by assigning a probabilistic distribution for the considered parameter and applying Monte Carlo simulation on the basic analytical equation to identify the most sensitive factors in designing the hydraulic fracturing process. Afanasyev et al (2023) aim to account for the relevant physical phenomena during CO₂ plumes in horizontal and sloping aquifers. For this purpose, the author employs the Monte Carlo method, along with several hundred detailed 3-D reservoir simulations to simulate a wide range of scenarios by introducing randomness, which helps capture the complex behaviour of CO₂ migration. Wang et al (2023) utilized stochastic Monte Carlo simulations with an extensive ensemble of 10,000 forward models to effectively evaluate the distinct effects of various uncertain parameters on the outcomes of a geothermal project in the Netherlands. The authors mentioned that the monte carlo can achieve the convergence state with approximately minimum 3000 realizations. Meanwhile, Palacios et al (2024) simulated 20000 iterations of monte carlo simulation to produce the minimum and maximum power potential/capacity (MWe) for 12 hot springs Geothermal energy in Mexico.

Although MCS is widely acknowledged as a simple approach for uncertainty quantification, the studies mentioned above consistently highlight the significant computational burden associated with Monte Carlo simulation. This high computational cost becomes a major drawback of this method, particularly when dealing with complex reservoir models

2.3 Polynomial Chaos Expansion

Among various surrogate modeling techniques, Polynomial Chaos Expansion (PCE) emerges as an alternative to Monte Carlo Simulation (MCS) for uncertainty quantification, offering a solution to the high computational demand of MCS by developing surrogate models. This method is particularly notable for its ability to model highly complex systems with minimal training costs (Zhou & Zhang, 2019), and accurately approximate the nonlinear responses of simulation models to input parameters, significantly reducing computational time (Jia et al., 2016, 2018; Naik et al., 2019)

PCE models the target problem by using a linear combination of orthogonal polynomials of different degrees, enabling the direct application of regression-based methods like least-squares for model training and PCE coefficient estimation (Filomeno Coelho et al., 2011). The process of addressing stochastic and uncertainty quantification (UQ) problems using Polynomial Chaos Expansion (PCE) involves two key steps. First, the implicit relationship between parameters and outputs is approximated by an explicit polynomial function, a process known as polynomial approximation. Next, this polynomial is used to calculate the probability distributions of the system's outputs (Shen et al., 2020). Figure 6 presents the general workflow for constructing a PCE model for uncertainty quantification.

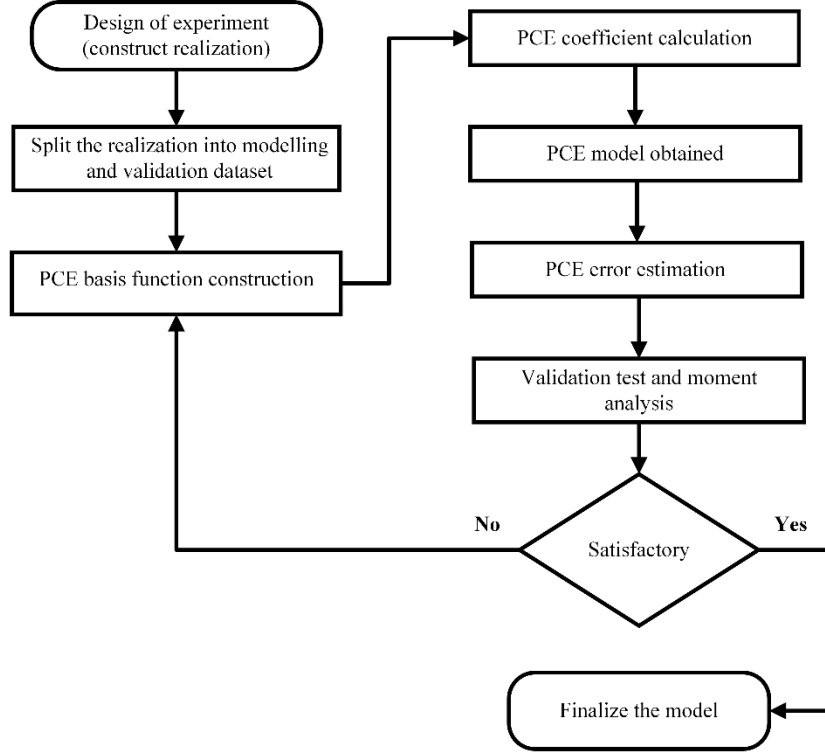


Figure 2.4. Surrogate model construction workflow procedure using PCE

Polynomial Chaos Expansion (PCE) has become a viable alternative to Monte Carlo Simulation (MCS) for uncertainty quantification, addressing the high computational demands associated with MCS. PCE achieves this by creating surrogate models that effectively approximate the nonlinear responses of simulation models to varying input parameters, thereby significantly reducing the required computational time (Jia et al., 2016, 2018; Naik et al., 2019). The statistical moments of the output are determined through the projected polynomial expansion. To develop a Polynomial Chaos Expansion (PCE) model, one begins by assuming a physical model $y = \text{model}(x)$, where $x = \{x_1, \dots, x_M\}^T \in \mathbb{R}^M, M \geq 1$ is a vector of input parameter, $y = \{y_1, \dots, y_N\}^T \in \mathbb{R}^N, N \geq 1$ is a vector of output parameters. If the parameter vector x is uncertain and described by a

probability density function (PDF), then each model output $y_i \in \mathcal{Y}$ is expected to be a random variable. Given that a particular model output y has a finite variance, it can be expressed through the following Polynomial Chaos Expansion.

$$y = M(x) = \alpha_0 B_0 + \sum_{j=1}^M \alpha_j B_1(x_j) + \sum_{j=1}^M \sum_{k=1}^j \alpha_{jk} B_2(x_j, x_k) + \sum_{j=1}^M \sum_{k=1}^j \sum_{h=1}^k \alpha_{jkh} B_3(x_j, x_k, x_h) + \dots \quad (2.1)$$

In this representation, the α 's are the coefficients and the B 's are multivariate polynomials basis functions that are orthogonal with respect to joint PDF of x .

In practice, the Polynomial Chaos Expansion (PCE) is truncated to include only a finite number of basis functions. The number of the terms N_t in the expansion depends on this truncation.

$$N_t = \frac{(M+d)!}{M!d!} \quad (1.1)$$

In this context, M indicates the number of model inputs (length of x) and d represents the maximum degree of the polynomial basis functions. The choice of polynomial basis functions is determined by the distribution of the input random variables (as shown in Figure 2.5).

<i>Z Distribution</i>	<i>Orthogonal Basis Functions</i>
Gaussian	Hermite Polynomials
Gamma	Laguerre Polynomials
Beta	Jacobi Polynomials
Uniform	Legendre Polynomials
Poisson	Charlier Polynomials
Binomial	Krawtchouk Polynomials
Hypergeometric	Hahn Polynomials

Figure 2.5. PCE basis functions for each type distribution (Camacho et al., 2017)

The foundation of Polynomial Chaos Expansion (PCE) theory was established in 1938, when Wiener introduced the concept of homogeneous chaos (Wiener, 1938). In his paper, Wiener expanded Birkhoff's ergodic theorem from a single variable to multidimensional truncated Hermite functions, demonstrating that Brownian motion theory could model stochastic processes by representing random variables with Gaussian distributions. Despite this early development, practical applications of PCE emerged much later, nearly sixty years afterward, with the groundbreaking work of Ghanem and Spanos (1991), who utilized PCE to address uncertainty quantification in mechanical engineering. Xiu & Karniadakis (2003) later broadened Wiener's approach, applying it to various families of orthogonal polynomials, thereby providing an effective method for representing non-Gaussian processes.

Polynomial Chaos Expansion (PCE) has been widely used as a surrogate model in the field of stability and control, solid mechanics, electronic circuits or computational fluid dynamics (Naik et al., 2019). Sarma et al (2005) were likely among the pioneers in applying Polynomial Chaos Expansion (PCE) for real-time optimization and uncertainty management in reservoir simulations. They introduced a closed-loop approach that utilized Karhunen-Loève (KL) expansion and Bayesian inversion techniques for real-time history matching. The methodology was enhanced in the petroleum sector by Sarma et al (2011) through non-intrusive spectral projection. In this study, Sarma introduced Non-intrusive spectral projection (NISP) as an alternative to the issues of probabilistic collocation method (PCM) – which the number of simulations needed to construct the PCE directly depends on the number of terms in the expansion, growing exponentially with both the PCE order

and the number of random variables. This significantly restricts the applicability of the PCE method for large-scale problems. Oladyshkin et al (2011) proven the effectiveness of Polynomial chaos expansion for stochastic modelling and uncertainty quantification in the existence of finite number of moments, and lack of information or even existence of the probability density function of the parameter. This result confirms that the data-driven PCE approach provides a novel method for defining parameter uncertainty in stochastic analysis, eliminating the subjectivity involved in assigning theoretical probability distributions. Hosder et al (2007) examined the accuracy and computational efficiency of Point-Collocation Non-Intrusive Polynomial Chaos (NIPC) across various sampling methods, including random, Latin hypercube, and Hammersley sampling. They observed that all three sampling methods performed similarly in terms of accuracy and computational efficiency. Furthermore, they found that using twice the minimum required number of collocation points improved the statistical approximation at each polynomial degree. Y. Zhang & Sahinidis (2013) introduced mixed-integer optimization (MIP) methods to systematically select coefficients in polynomial expansions, leading to a more straightforward and accurate surrogate model with the same level of input-output data. They argued that this approach effectively captures interactions between low- and high-order polynomial terms, setting it apart from traditional PCE techniques that rely on forward selection or backward elimination. Jia et al (2016) & Jia et al (2018) applied PCE models to reduce the computational cost and minimizing the lost of accuracy during the uncertainty quantification study for carbon sequestration project. In this study, the authors suggest that the performance of PCE reduced order model using 50 sample data gain a competitive accuracy with monte carlo result with 1000 sample used.

Many enhancements have been suggested for the traditional PCE method. In high-dimensional problems, estimating expansion coefficients is computationally intensive due to the exponential growth in the number of coefficients as the number of variables increases, Camacho et al (2017) demonstrated that Smolyak quadrature can efficiently estimate polynomial coefficients with a relatively small number of simulations. Importantly, they showed that the required number of evaluations remains independent of the expansion's degree. Naik et al (2019) firstly introduced the use of PCE-Proxy model to calibrate flow physics model of surfactant-polymer flooding process. In this study, the authors incorporate krigging model (PCE-Krigging) to estimate the probable error associated with the predicted values of the proxy model. Tarakanov & Elsheikh (2019) introduced an effective method for estimating PCE coefficients by integrating Elastic Net regularization with a data-driven approach for feature ranking. This approach tackles error minimization in calculating PCE coefficients within high-dimensional models.

Hosder et al (2007) already suggest that the number of sample N_p used in model simulations to develop a surrogate Non-intrusive polynomial chaos (NIPC) model should be double the number of polynomial terms N_t in the expansion. The value of N_t itself can be calculated using the equation 2. This becomes foundation for many research for defines the number of sample for the PCE model. However, to the best of our knowledge, all of the studies that has been reviewed use relatively low polynomial degree $2 < P < 5$.

In some conditions, when the model is very complex and the non-linearity is very high, higher order polynomial is need to assign in the PCE model. This situation will add the complexity in the model and more data points will be needed. Hence, in this study, the relationship between the increase in polynomial degree and used sample points will be

investigated to determine the recommended sample points that need to be used when dealing with higher order polynomial model

CHAPTER 3

METHODOLOGY

This chapter provides a detailed explanation of the methods and procedures applied in this study, including a description of the reservoir model used for the simulation, as well as the procedures and mathematical foundations of Monte Carlo simulation, Polynomial Chaos Expansion, and sensitivity analysis using the Sobol method. Figure 10 illustrates the overall workflow of the study.

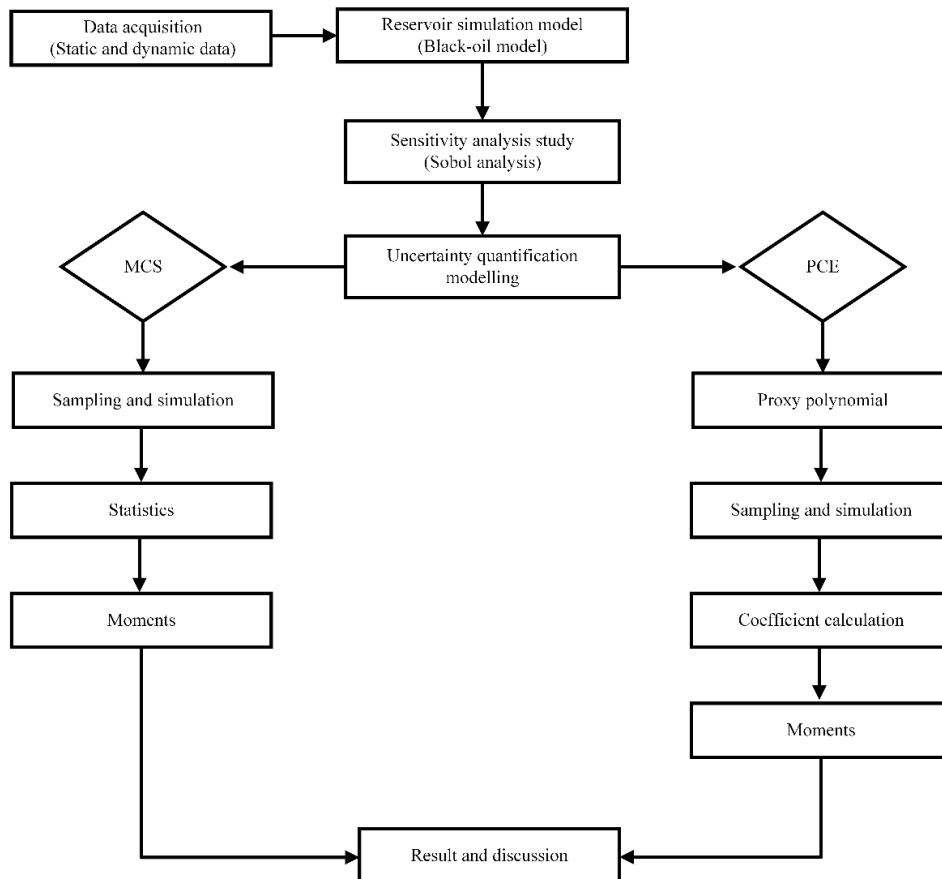


Figure 3.1. The overall workflow of the study

3.1 Reservoir Model Description

This study will use the open-source PUNQ S-3 industrial reservoir engineering model. The PUNQ (Production Uncertainty Quantification) project, financed by the EU, delved into history matching and uncertainty quantification research, involving various research institutions and oil companies throughout Europe. As part of this initiative, several reservoir engineering models were crafted to explore different history-matching methodologies. One such model, the S3, was released for usage by other research entities within this field and has seen extensive application.

The PUNQ-S3 scenario originates from an actual field study conducted by Elf Exploration Production, characterized as a compact industrial reservoir engineering model. The field features an oil reservoir with a gas cap and a functioning aquifer. The description of the reservoir is presented in figure 3.2 and table 3.1

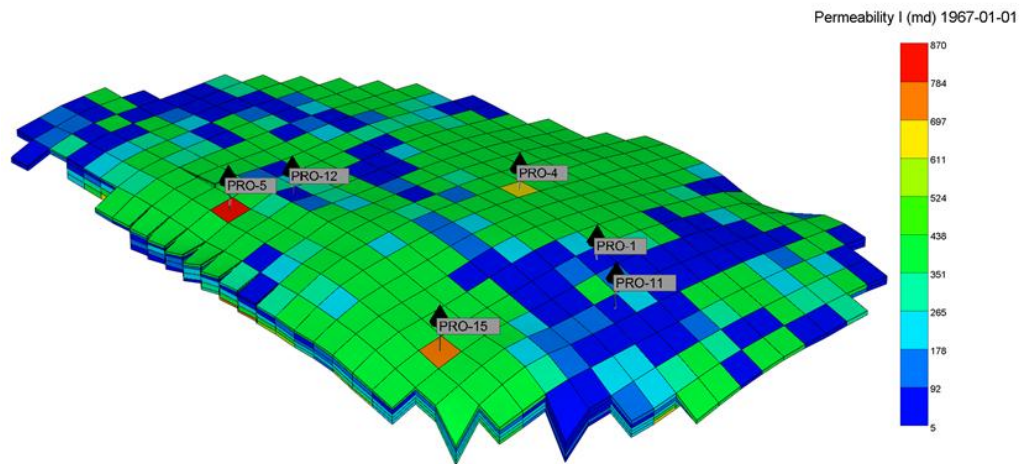


Figure 3.2. 3D reservoir model (PUNQ-S3 model)

Table 3.1. Brief Reservoir description PUNQ-S3

Parameter	Value	Unit
Grid top	2340	m
Grid bottom	2443	m
Grid structure	19×28×5	
Bulk volume	5.75×10 ⁷	m ³
Reservoir pressure	23446	kPa
HC. pore volume	25489	m ³
Reservoir oil in place	1.9972×10 ⁸	m ³
Total grid number	2660	Grid block
Grid size area	180×180	m ²

The numerical model is structured into $19 \times 28 \times 5$ grid blocks, each measuring 180×180 square meters, with 1761 of these blocks being operational. The reservoir is exploited using 10 production wells that are opened progressively every two months and are named as follows: PRO-1, PRO-4, PRO-5, PRO-11, PRO-12, PRO-15, PRO-23, PRO-24, PRO-29, and PRO-50. The reservoir is bordered by faults to the east and south while it shares connectivity with a substantial aquifer to the north and south, eliminating the need for injector wells. The spatial arrangement of these wells within the reservoir's porosity framework is illustrated in the subsequent Figure 3.2.

The reservoir simulation model is defined to have 10 uncertain parameters, each adhering to a truncated normal distribution and log normal distribution for KRG, KRO,

KRW, MULTIPERI and MULTIPERMK parameter. The specifics of these parameters are outlined in Table 2.

Table 3.2. Statistic information of the random input parameters used

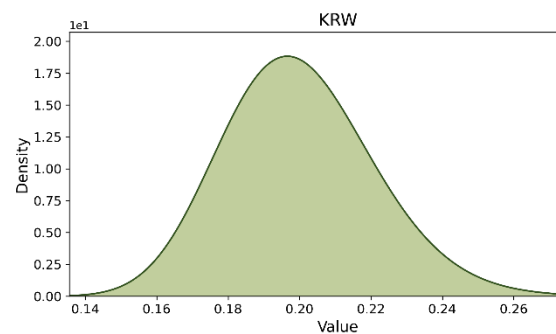
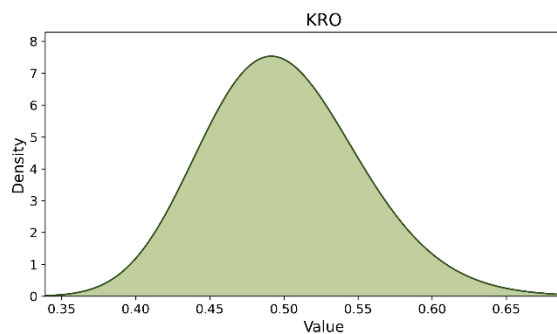
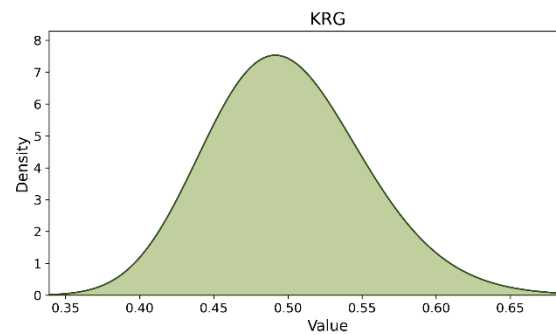
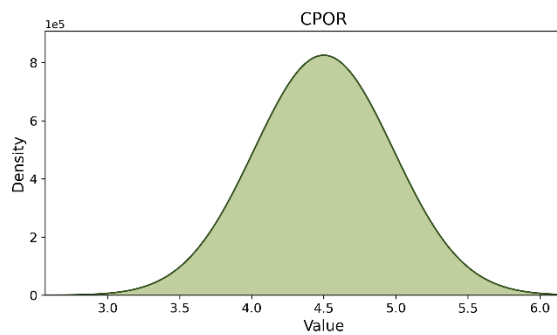
No	Parameter	Min. value	Max. value	Mean	Std
1	CPOR	3.375E-06	5.625E-05	4.5E-06	4.835E-07
2	KRG	0.375	0.625	0.5	0.0537310
3	KRO	0.375	0.625	0.5	0.0537310
4	KRW	0.15	0.25	0.2	0.0214924
5	NG	2.24904	3.7484	2.99872	0.322249
6	NW	1.64871	2.74785	2.19828	0.236239
7	MULTIPERMI	0.75075	1.25125	1.001	0.107566
8	MULTIPERMK	0.75075	1.25125	1.001	0.1075696
9	MULTIPOR	0.75075	1.25125	1.001	0.1075696
10	SORW	0.075	0.15	0.1	0.0107462

The uncertain parameters in this reservoir simulation model are described next:

CPOR *Rock compressibility, a value used in the whole model*

KRG *Maximum gas relative permeability.*

<i>KRO</i>	<i>Maximum oil relative permeability.</i>
<i>KRW</i>	<i>Maximum water relative permeability.</i>
<i>NG</i>	<i>Corey exponent of gas relative permeability</i>
<i>NW</i>	<i>Corey exponent of water relative permeability</i>
<i>MULTIPOR</i>	<i>Porosity multiplier.</i>
<i>MULTIPERMI</i>	<i>Horizontal permeability multiplier.</i>
<i>MULTIPERMK</i>	<i>Vertical permeability multiplier.</i>
<i>SORW</i>	<i>Residual oil saturation to water</i>
<i>SORW</i>	<i>Residual oil saturation to water</i>



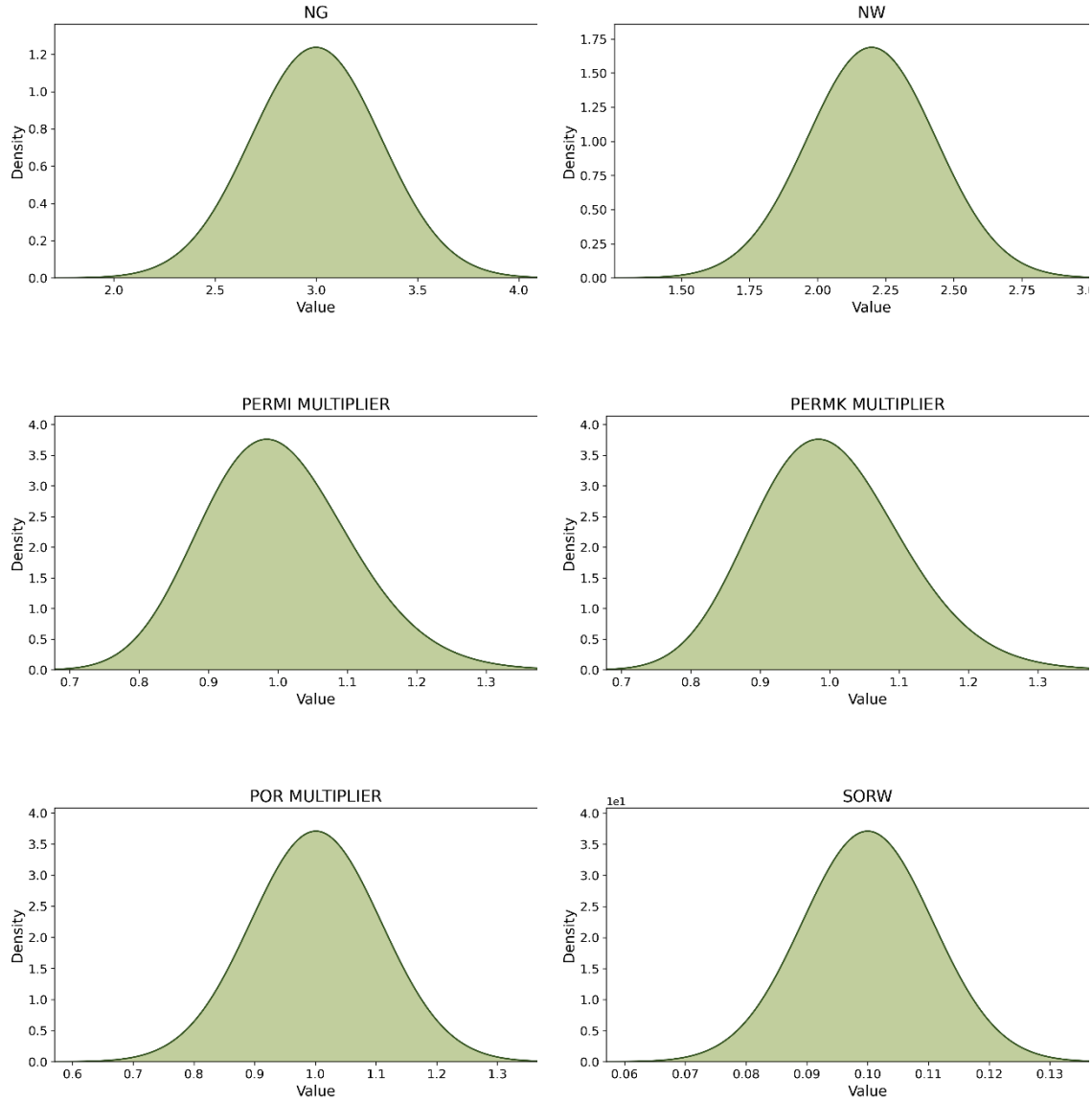


Figure 3.3. Probability density plot of random input parameters

Generally, adding more parameters can slightly enhance accuracy but at the cost of increased computational demands. In this study, we will conduct a sensitivity analysis to address this. This preliminary step is often conducted using Sobol analysis, a method introduced by (IM Sobol, 1993), which falls under the category of variance-based sensitivity analysis. The core concept of these methods involves evaluating how much each input variable contributes to the overall variance of the output. Within a probabilistic

framework, the Sobol method breaks down the output variance into segments that correspond to each input (Saltelli et al., 2010). For instance, in a model with two inputs influencing one output, it might be found that 60% of the variance in the output is due to the first input, 30% to the second, and the remaining 10% arises from their interaction. These figures are then used to gauge the sensitivity of the model to changes in its inputs.

3.2 Sensitivity Analysis

In this study, we will conduct a sensitivity analysis to address this. This preliminary step is often conducted using Sobol analysis, a method introduced by (IM Sobol', 1993), which falls under the category of variance-based sensitivity analysis. The core concept of these methods involves evaluating how much each input variable contributes to the overall variance of the output.

Suppose, we have a model function f which maps $f: X \rightarrow Y$, Where $X = (X_1, X_2 \dots X_n)^T$ is the model input vector and $Y = f(X)$ is the model output response. Sobol indices (rank) can then be calculated as

$$S_i = \frac{V_{X_i}[\mathbb{E}_{X \sim i}(Y|X_i)]}{V(Y)} \quad (3.1)$$

where the terms of functional decomposition are expressed in terms of conditional expected values $\mathbb{E}$. The sum of all Sobol indices, called total Sobol indices S_i^T , can then be estimated as follows:

$$S_i^T = \sum_{k \neq i} S_k = \frac{\mathbb{E}_{X \sim i}[V_{X_i}(Y|X_{\sim i})]}{V(Y)} \quad (3.2)$$

Within a probabilistic framework, the Sobol method breaks down the output variance into segments that correspond to each input (Saltelli et al., 2010). For instance, in a model with two inputs influencing one output, it might be found that 60% of the variance

in the output is due to the first input, 30% to the second, and the remaining 10% arises from their interaction. These figures are then used to gauge the sensitivity of the model to changes in its inputs.

3.3 Monte Carlo Simulation

In the mathematical setting, the MC method approximates the expected value of a random variable such as $s = \mathbb{E}[Y]$. In general, we generate values of $y_1, \dots, y_n$ independently and randomly from the distribution of Y and take their average as the estimate of s .

$$\hat{s}_n = \frac{1}{n} \sum_{i=1}^n y_i \quad (3.3)$$

In the setting of independent and identically distributed (i.i.d.) sampling, $\hat{S}_n$ is a random variable and the mean of $\hat{S}_n$ is

$$\mathbb{E}[\hat{S}_n] = \frac{1}{n} \sum_{i=1}^n \mathbb{E}[y_i] = s \quad (3.4)$$

The MC estimator is unbiased since the expected value of $\hat{S}_n$ is equal to s . According to the strong law of large numbers, the average converges almost surely to the true expected value

$$\mathbb{P}\left(\lim_{n \rightarrow \infty} |\hat{S}_n - s| = 0\right) = 1 \quad (3.5)$$

provided that the variance of the individual terms $\mathbb{V}[Y]$ is bounded (Kroese & Rubinstein, 2012). Suppose that $\mathbb{V}[Y] = \sigma^2 < \infty$, we have

$$\mathbb{V}[\hat{S}_n] = \mathbb{E}[(s - \hat{s}_n)^2] = \frac{\sigma^2}{n} \quad (3.6)$$

and the root mean square error (RMSE) of $\hat{s}_n$ is $\sigma/\sqrt{n}$, which means the error of MC estimator is of order $n^{-1/2}$ and we typically write $\text{RMSE} = \mathcal{O}(n^{-1/2})$, where $\mathcal{O}(\cdot)$ notation allows us to focus on the rates of convergence as $n \rightarrow \infty$.

One of the strengths of the MC method is that the sample values can be used for error estimation. The most commonly used estimates of σ^2 are

$$\zeta^2 = \frac{1}{n-1} \sum_{i=1}^n (y_i - \hat{s}_n)^2 \quad (3.7)$$

$$\hat{\zeta}^2 = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{s}_n)^2 \quad (3.8)$$

and the estimator in Equation (5) is unbiased since $\mathbb{E}[\zeta^2] = \sigma^2$ for $n \geq 2$. From the central limit theorem, the error $\hat{S}_n - S$ has approximately a normal distribution with mean 0 and variance σ^2/n . Therefore, we can estimate confidence intervals around the estimate $\hat{S}_n$.

3.4 Polynomial Chaos Expansion (PCE)

3.4.1 PCE Coefficient Calculation

The coefficients for the polynomial function, as described in equation 3.9 can be ascertained through two methods: an intrusive method or a non-intrusive method. In this instance, we adopt the Least Square Minimization method to identify these coefficients. A key benefit of this technique is its flexibility, allowing the use of any set of points that accurately reflect the true probability distribution, $f(X)$, of the input variables for the determination of the coefficients.

$$y_\alpha = \arg \min_{y_\alpha \in \mathbb{R}^{p_{\text{basic}}}} \frac{1}{N_S} \sum_{i=1}^{N_S} \left(P(X^{(i)}) - \sum_{\alpha \in A} y_\alpha \Psi_\alpha(X^{(i)}) \right)^2 \quad (3.9)$$

N_s represents the aggregate count of experimental design points, which are derived from the probability distribution of the input variables, with $\Xi = (X^{(1)}, X^{(2)} \dots X^{N_s})$ constituting the input vector. The method of latin hypercube sampling (LHS) is employed for generating these sample points. Various research indicates that LHS offers a superior rate of convergence compared to the Monte Carlo Sampling (MC) technique, particularly in scenarios involving multiple dimensions (Aistleitner et al., 2012). The computational model undergoes evaluation at each of these points. To address equation 3.9 the Ordinary Least Square (OLS) method, which aims to minimize the sum of squared residuals, is applicable (Yaghoubi et al., 2017)

$$y_\alpha = (A^T A)^{-1} A^T Y \quad (3.10)$$

$Y = (y^{(1)}, y^{(2)} \dots y^{N_s})$ represents the responses of the output model, calculated at the N_s -sized input vector, while A denotes the experimental design matrix. This matrix encompasses the orthonormal bases for all N_s and is defined as $A_{ij} = \psi_j \xi^{(i)}, i = 1 \dots N_s, j = 1 \dots p - 1$.

Given that equation 3.10 presents a least square minimization challenge, augmenting the count of experimental design points, N_s , aids in reducing the truncation error, thereby enhancing the reliability and precision of the PCE-proxy model. Nonetheless, conducting simulations across an extensive array of experimental design points incurs significant computational resources, making it crucial to accurately assess errors in the PCE's subsequent predictions to ascertain the most effective number of model evaluations

3.4.2 PCE Posterior Error Estimation

The posterior error of the Polynomial Chaos Expansion (PCE) can be precisely evaluated through the error derived from least square minimization, often referred to as the relative generalization error, ε_{gen} . A specific form of this error, termed the normalized empirical error, ε_{emp} , effectively measures the PCE's capability to accurately mimic the output response at the experimental design points, Ξ , as indicated in reference (Maiy & Sudrety, 2017).

$$\varepsilon_{emp} = \frac{\sum_{i=1}^{N_S} (P(X^{(i)}) - P^{PC}(X^{(i)}))^2}{\text{Var}(Y)} \quad (3.11)$$

Where $\text{Var}(Y) = \sum_{i=1}^{N_S} \left(P(X^{(i)}) - \frac{1}{N_S} \sum_{i=1}^{N_S} P(X^{(i)}) \right)^2$ is variance of model output response of Y and the Polynomial Chaos Expansion (PCE) is formulated using every point within the experimental design set ($P^{PC}(X)$).

3.4.3 Moment of PCE: Mean and Variance

The moments of the Polynomial Chaos Expansion (PCE) can be determined analytically and are known to converge to the actual response of the output model. Specifically, the first two moments, the mean and variance, of the PCE are solely dependent on its coefficients. It is demonstrated that the mean of the PCE (μ_{PCE}) is determined by the first coefficient, which corresponds to the constant basis term $\Psi_0 = 1$ i.e.

$$\mu_{PCE} = y_1 \quad (3.12)$$

The variance of the PCE (σ_{PCE}^2) is calculated by summing the squares of all its coefficients, with the exception of the first one (y_1) i.e.

3.4.4 Moment of Stochastic Approach (LHS): Mean and Variance

For n independent and identically distributed (iid) samples $(X = (X_1, X_2 \dots X_n))$ drawn from a particular distribution, the mean of that distribution, denoted as I_n , can be estimated through the Central Limit Theorem (CLT), provided that the variance V_n of the distribution is defined. According to the CLT, as the number of samples n approaches infinity ($n \rightarrow \infty$), the sample mean of any given distribution will converge towards a normal distribution centered around the mean:

$$I_n = \frac{1}{n} \sum_{j=1}^n P(X_j) \rightarrow N \left(\mathbb{E}[P(X)], \frac{1}{n} \mathbb{V}[P(X)] \right) \quad (3.13)$$

Where $\mathbb{E}[\cdot]$ is the expectation of PCE and $\mathbb{V}[\cdot]$ is the variance of PCE.

The variance of the distribution, denoted as V_n , can subsequently be estimated as follows:

$$V_n = \frac{1}{n} \sum_{j=1}^n (P(X_j) - I_n)^2 \quad (3.14)$$

CHAPTER 4

RESULT AND DISCUSSION

4.1 Monte Carlo Simulation

This study uses the open-source PUNQ S-3 industrial reservoir engineering model with a model comprising 19x28x5 grid blocks, with cumulative oil production set to be the main objective function. Monte Carlo simulation for cumulative oil production was already computed based on the 17000 realizations of the reservoir model. The probability density curve, cumulative probability are constructed, and the statistical properties of the resulting cumulative oil production data obtained from monte carlo simulation are presented in figure 4.1 and table 4.1. Following this, a convergence plot is constructed to evaluate the stability of the Monte Carlo model as a function of the addition of new realizations. In the convergence plot, each colored line represents a moving average of the plotted data. As new data points (realizations) are added, the average is recalculated. Upon completing the analysis for one subset, the data set is shuffled randomly, and the process is repeated with a new arrangement of the data to generate another colored line. This is done 500 times for each data set, enhancing the visualization of the variability across different computations. Based on Figure 15, it appears that the Monte Carlo simulation results stabilize once the number of realizations exceeds 17,000. This indicates that obtaining stable and reliable information from Monte Carlo simulations requires a large number of realizations, which can be extremely time-consuming for complex models.

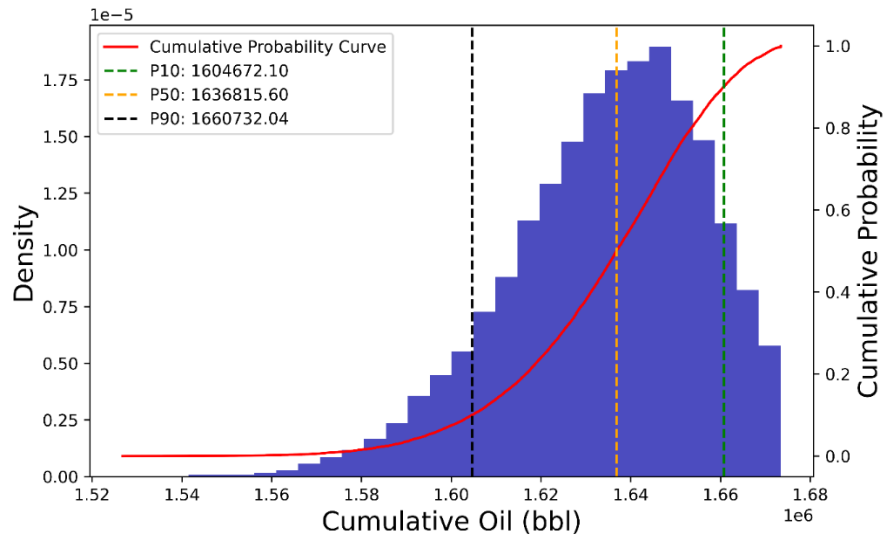
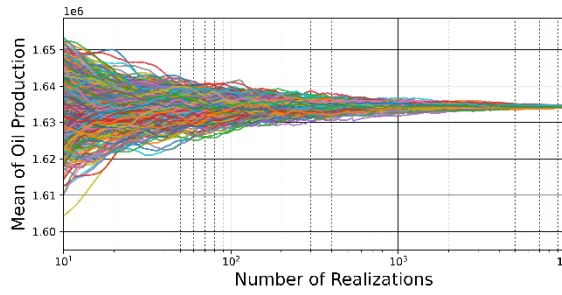


Figure 4.1. Monte Carlo Simulation result for cumulative oil production

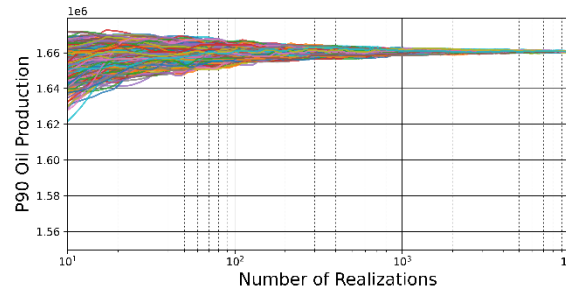
Following this, a convergence plot is constructed to evaluate the stability of the Monte Carlo model as a function of the addition of new realizations. In the convergence plot, each colored line represents a moving average of the plotted data. As new data points (realizations) are added, the average is recalculated. Upon completing the analysis for one subset, the data set is shuffled randomly, and the process is repeated with a new arrangement of the data to generate another colored line. This is done 500 times for each data set, enhancing the visualization of the variability across different computations. Based on Figure 4.2, it appears that the Monte Carlo simulation results stabilize once the number of realizations exceeds 17,000. This indicates that obtaining stable and reliable information from Monte Carlo simulations requires a large number of realizations, which can be extremely time-consuming for complex models:

Table 4.1. Summary statistic Monte Carlo simulation result

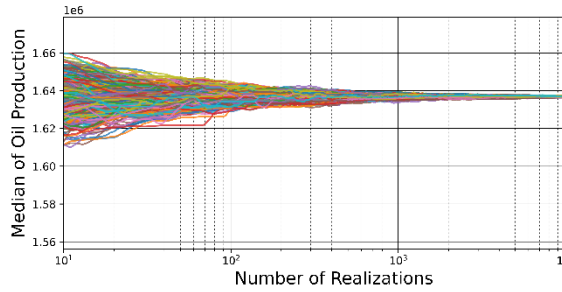
Parameter	Cumulative oil production (m ³)
Mean	1634352.984
P90	1660318.36
P50	1636817.75
P10	1604414.24
Variance	471503656.858



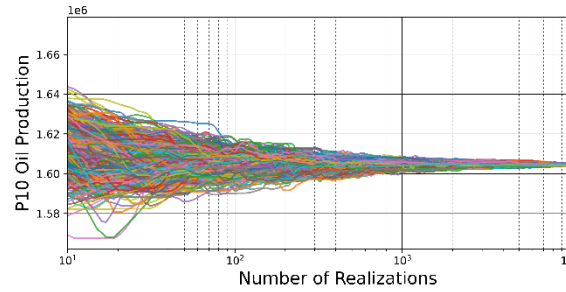
(a) Mean oil production



(b) P90 oil production



(c) P50 oil production



(d) P10 oil production

Figure 4.2. Convergence plot for each statistical properties of MCS

Furthermore, sensitivity analysis using the Sobol method was conducted on 10 random parameters, revealing that the porosity multiplier, horizontal permeability multiplier, and rock compressibility exhibit the highest sensitivity to cumulative oil production, with sensitivity factors of 57%, 40%, and 2.3% consecutively. This result is important to use for reducing the dimensionality of parameters in the PCE study. As high random input parameter will lead to the high polynomial parameter, so it is important to select and use the high correlation input parameter to the objective function (cumulative oil production).

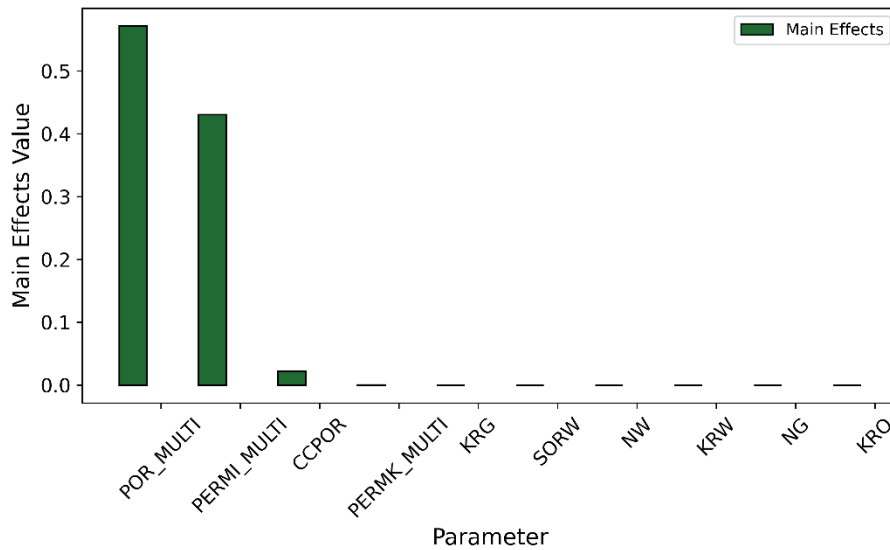


Figure 4.3. Sobol sensitivity analysis – main effect of the input parameters

Table 4.2. Top 3 significant random parameters

Parameter Name	Main Effects	Interaction Effects
Porosity Multiplier	0.571136	0.000527133
Permeability I Multiplier	0.403256	0.00117459
CCPOR	0.0220957	0.0018804

4.2 Polynomial Chaos Expansion

The Polynomial Chaos Expansion (PCE) model was constructed using the three most influential parameters: porosity multiplier, horizontal permeability multiplier, and rock compressibility. The Hermite polynomials serve as the foundation for the set of orthogonal functions. This study evaluates the PCE's performance under two key scenarios. In the first scenario, the degree of the polynomial model is varied to analyze its impact. In the second scenario, the number of samples is adjusted while maintaining a constant polynomial degree. Both scenarios aim to provide deeper insights into the model's accuracy and robustness across different parameter configurations.

Initially, the minimum number of samples for each degree is calculated based on a specific equation $N_t = \frac{(M+d)!}{M!d!}$, where M is the number of parameter used, d is the highest degree of polynomial model, and N_t is the total coefficient. The minimum recommended number of sample N_s used to build the polynomial model is twice the number of coefficient N_t Hosder et al (2007). Table 4.3 presented the number of coefficient and number of samples used for each degree polynomial

Table 4.3. Coefficient and parameter number for each PCE degree

Maximum Degree	Number of Coefficient	Number of samples
3	20	50
5	56	120
7	120	250
9	220	450
11	364	750

Table 4.3 provides the recommended sample size for each polynomial degree. For each degree, a random subset of samples is drawn from the latin hypercube 17,000 Monte Carlo Simulations (MCS) realizations, with the remaining samples set aside as the evaluation dataset to test the performance of the polynomial model. The samples used to build the surrogate model are referred to as the "training data," while those used to assess the model's accuracy are the "evaluation data." Various evaluation metrics, including the coefficient of determination, empirical error, etc, are applied to gauge the model's performance on each dataset. The uses of several evaluation metrics in this study to gain unbiased interpretation of the model performance.

There is a subtle distinction between the training and evaluation datasets in this process. When the model is used to predict outputs for the training set, it tends to overestimate because no cross-validation is applied during the construction of the surrogate model. Simply put, if a point like X_{11} is used to build the PCE model and then the same

model is used to predict the output for X_{11} , it will provide a perfect prediction. To avoid this overfitting, a leave-one-out method is employed. In this approach, the PCE is built using all data points except one ($P^{PC\sim i}(\xi)$), and the excluded point is used to evaluate the model's performance, estimating the R^2 and error called Leave-one-out (LOO) error. Then the PCE model is applied to predict outcomes in the evaluation dataset. Finally, the empirical error for the evaluation set is calculated, as shown in Equation 3.11.

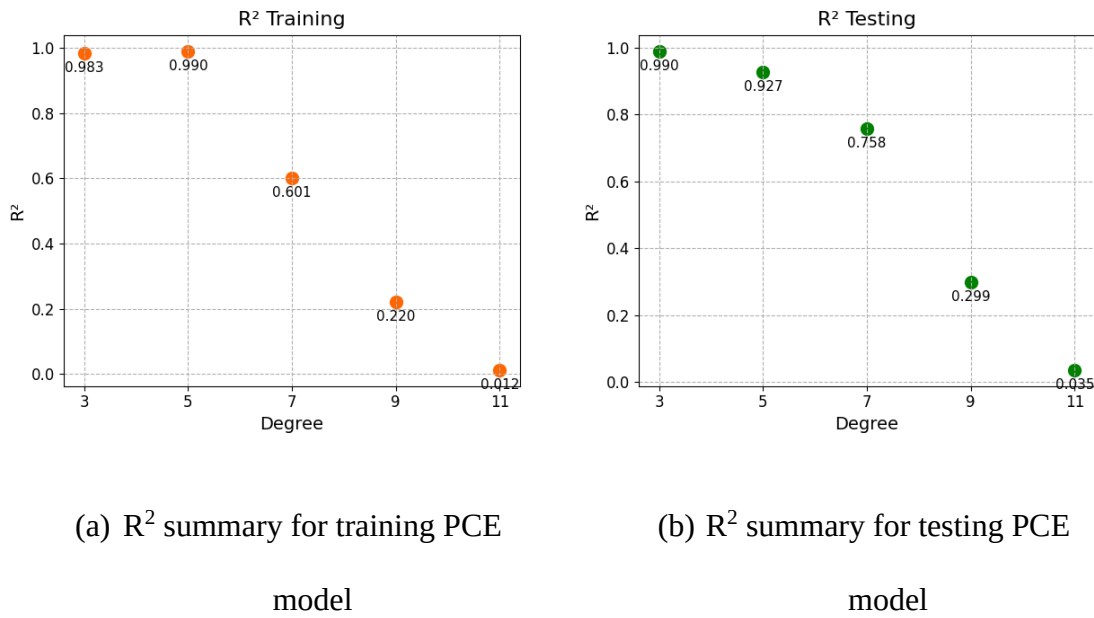


Figure 4.4. Summary PCE performance for all PCE model

Figure 4.4 illustrates the performance of PCE models with varying polynomial degrees (3, 5, 7, 9, and 11) in predicting cumulative oil production. The results show a strong coefficient of determination (R^2) as the polynomial degree increases from 3 to 5, with R^2 values exceeding 0.95 for both training and evaluation datasets. This indicates that the model fits the training data well and generalizes effectively to unseen data with the assigned number of sample. However, starting with degree 7, a significant drop in

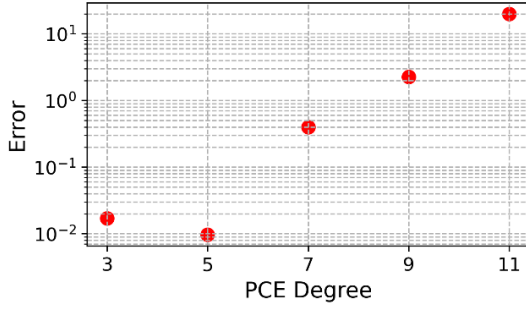
performance is observed, with R^2 values decreasing dramatically compared to lower-degree models (3 and 5). This decline continues and worsens with higher degrees, as evident in the results for degrees 9 and 11, which show very low R^2 values for both the training and evaluation sets. This poor performance is further supported by Table 4.4, Figure 4.5 and Figure 4.6 where model errors increase sharply as the polynomial degree rises. For example, the training error jumps to 24% at degree 7, 249% at degree 9, and an astonishing 4380% at degree 11.

Table 4.4. Error evaluation for all PCE degree model

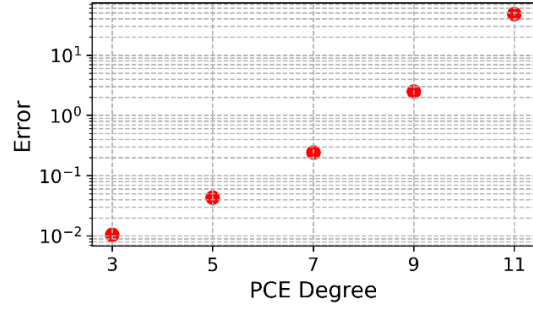
Sample	LOO	Empirical	L_2 mean	L_2 variance
3	1.71E-02	1.04E-02	7.60E-05	1.19E-02
5	9.62E-03	4.34E-02	1.89E-04	1.47E-02
7	3.97E-01	2.42E-01	1.10E-04	1.68E-01
9	2.25E+00	2.49E+00	6.16E-04	2.55E+00
11	2.00E+01	4.83E+01	5.09E-04	4.90E+01

This phenomenon exemplifies the bias-variance tradeoff, where increasing model complexity with higher-degree polynomials reduces bias but drastically raises variance. This heightened variance makes the model overly sensitive to noise, preventing it from generalizing effectively to unseen data. While higher-degree polynomials can capture more intricate patterns, they also increase the risk of overfitting, which harms the model's generalizability. Overfitting causes the model to fit noise in the data, explaining the sharp

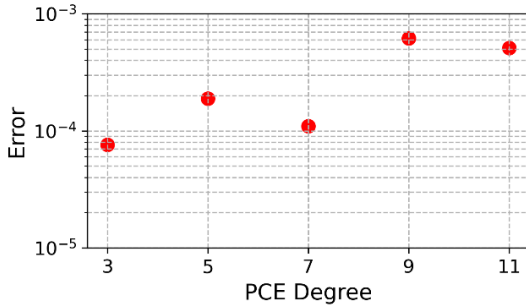
drop in R^2 and the dramatic increase in variance at higher polynomial degrees. The error variance between the MCS dataset and the PCE results escalates significantly, reaching 16% for degree 7, 250% for degree 9, and a staggering 4900% for degree 11.



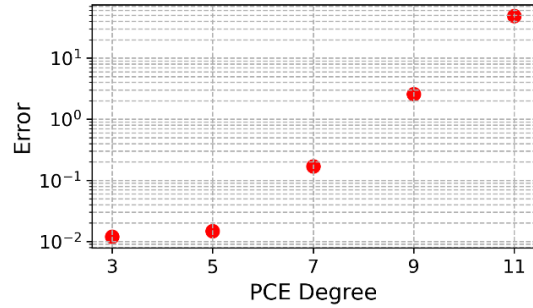
(a) Leave-one-out error



(b) Empirical error

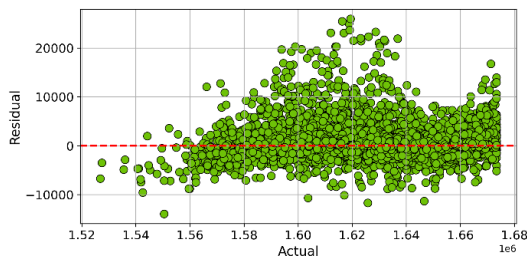


(c) Relative L_2 error in mean

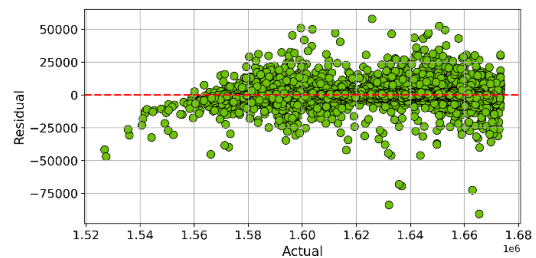


(d) Relative L_2 error in variance

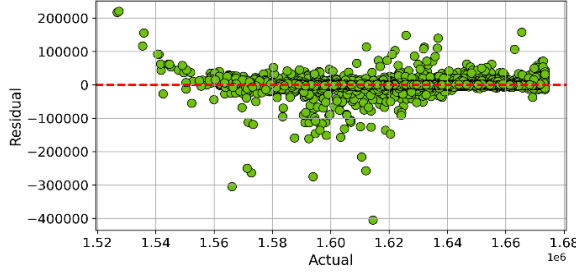
Figure 4.5. Error evaluation plot for all PCE model



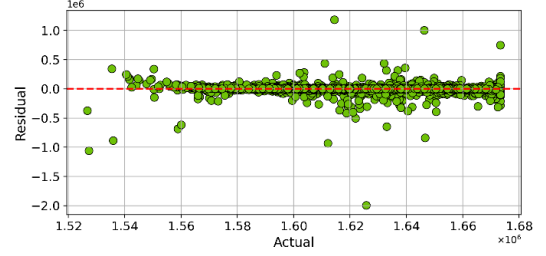
(a) Degree 3



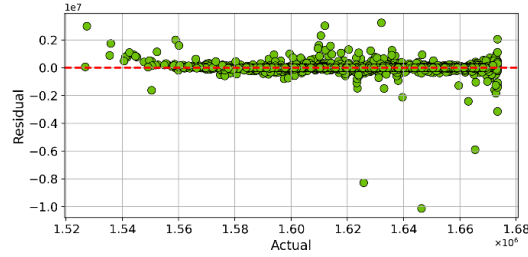
(b) Degree 5



(c) Degree 7



(d) Degree 9



(e) Degree 11

Figure 4.6. Residual plot for all PCE model

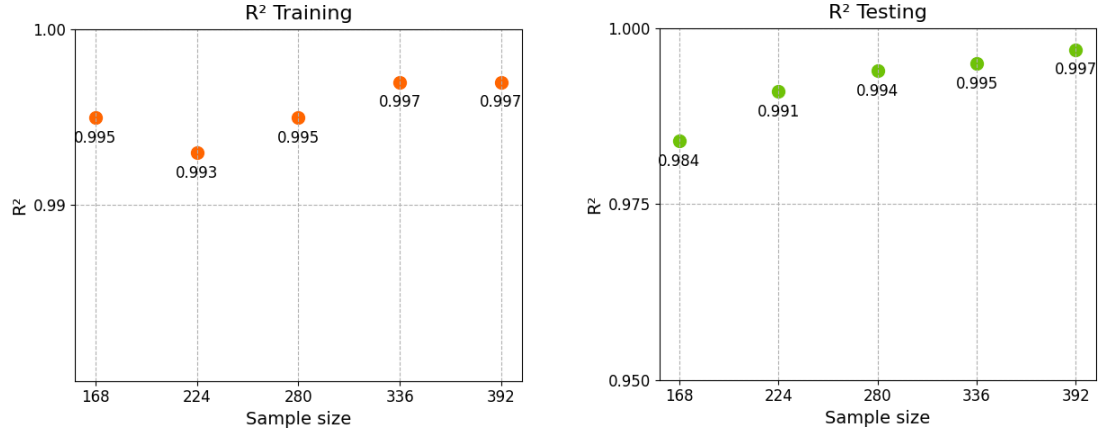
Since it has been established that higher-order polynomial models require more data samples to achieve optimal accuracy in terms of R^2 , mean and variance errors, the next section will examine the performance of polynomial degrees 5, 7, 9, and 11 under varying sample sizes. Referring to Table 4.5, several sample sizes based on the number of coefficients (N_t) for each degree will be tested. Multiplier factors (n_p) will be applied to (N_t) to determine the minimum number of samples required for optimal performance at each polynomial degree. By multiplying (N_t) by these factors, the resulting value will represent the sample size used for testing each polynomial model. The goal is to draw solid conclusions and provide insights into the relationship between polynomial degree and sample size for improved model performance.

Table 4.5. Varying number of samples for each polynomial degree

Maximum Degree	N_t	$n_p = 3$	$n_p = 4$	$n_p = 5$	$n_p = 6$	$n_p = 7$
5	56	120	168	224	280	336
7	120	250	360	480	600	720
9	220	450	660	880	1100	1320
11	364	750	1092	1456	1820	2184

4.2.1 PCE Degree 5 – Varying Sample

The evaluation of the PCE model with degree 5, as shown in figures 4.7, consistently demonstrates strong R^2 values across all sample sizes. This is expected, as the PCE model has already performed well with twice the minimum recommended number of samples. The results indicate that the model improves in both bias and variance as the sample size increases. Moreover, the relative L_2 errors for mean, variance typically decrease with larger datasets. Additional parameter such as P10, P50, and P90 are introduced to compare the MCS and PCE performance. The results also shows an improve with the PCE model as the number of sample increase. This suggests that the PCE model becomes progressively more robust and reliable in approximating the system's behavior as more training data is introduced. These findings are consistent with statistical principles, which state that increasing the sample size improves the model's generalization ability and helps reduce overfitting.



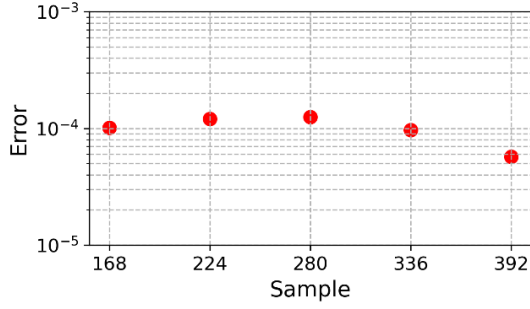
(a) R^2 summary for training PCE
model degree 5

(b) R^2 summary for testing PCE
model degree 5

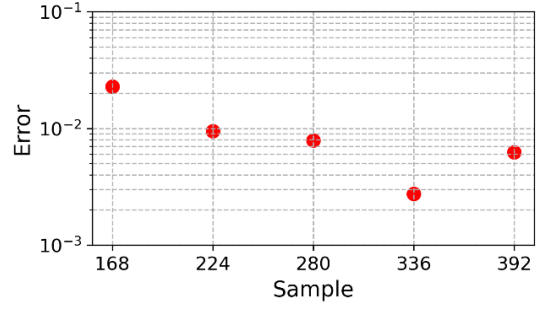
Figure 4.7. Summary PCE degree 5 performance for all varying sample

Table 4.6. Error evaluation PCE model degree 5 for all sample variations

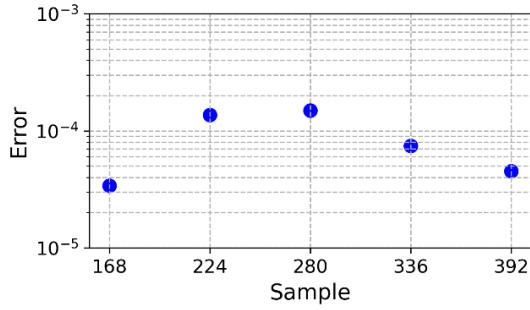
Sample	L_2 mean	L_2 variance	L_2 P90	L_2 P50	L_2 P10
168	1.02E-04	2.29E-02	1.81E-04	1.82E-04	3.41E-05
224	1.21E-04	9.49E-03	5.10E-05	1.98E-04	1.37E-04
280	1.25E-04	7.88E-03	2.59E-06	1.96E-04	1.49E-04
336	9.69E-05	2.75E-03	2.17E-06	1.59E-04	7.44E-05
392	5.73E-05	6.25E-03	5.42E-05	1.34E-04	4.51E-05



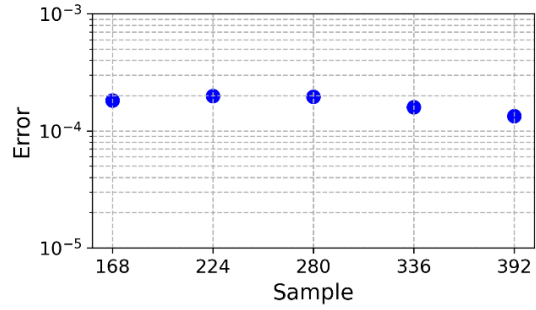
(a) L_2 error in mean



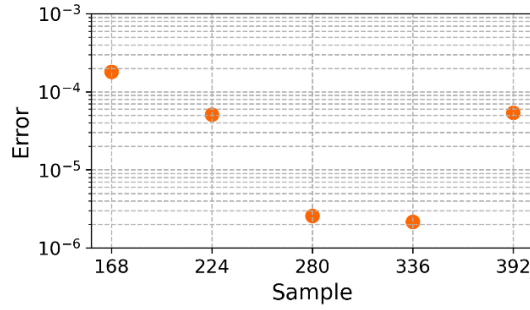
(b) L_2 error in variance



(c) L_2 error in P10



(d) L_2 error in P50

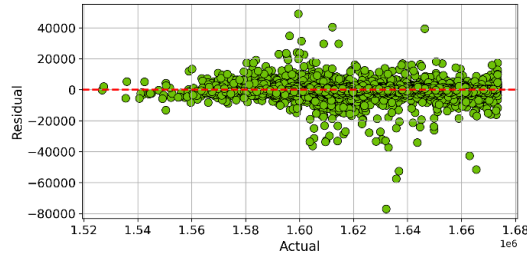


(e) L_2 error in P90

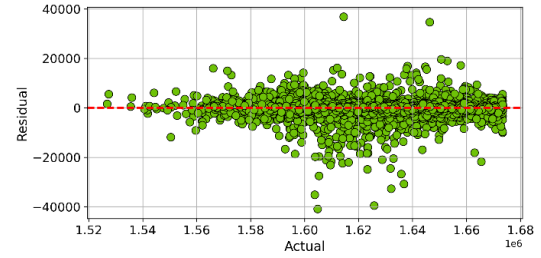
Figure 4.8. Error evaluation PCE model degree 5 for all sample variations

The residual for all sample variations are generally distributed around zero, suggesting minimal error. However, each residual plot contains some outliers. As the sample size of the PCE model increases, these outliers become less prominent, indicating an improvement in the PCE model's performance with a larger sample size. It concludes

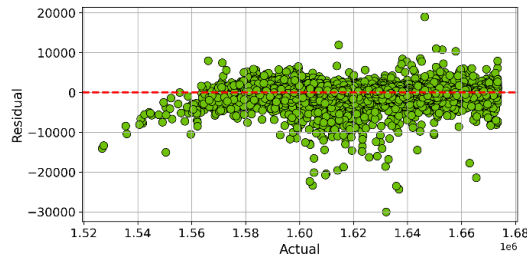
that larger sample sizes help the model to balance the trade-off between bias and variance, leading to more accurate predictions that better match the true distribution of the data.



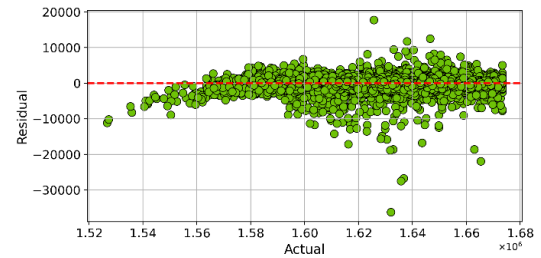
(a) 168 sample points



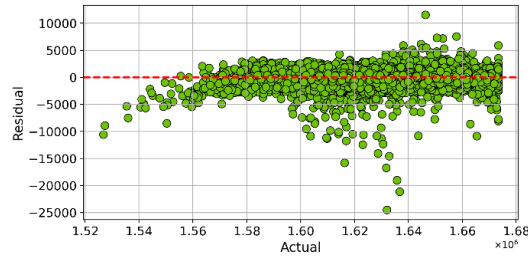
(b) 224 sample points



(c) 280 sample points



(d) 336 sample points



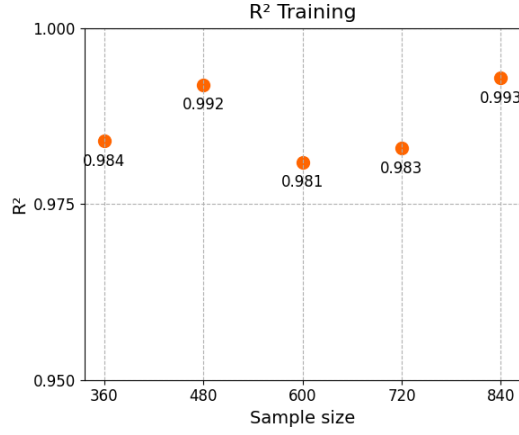
(e) 392 sample points

Figure 4.9. Residual plot PCE model degree 5 for all sample variations

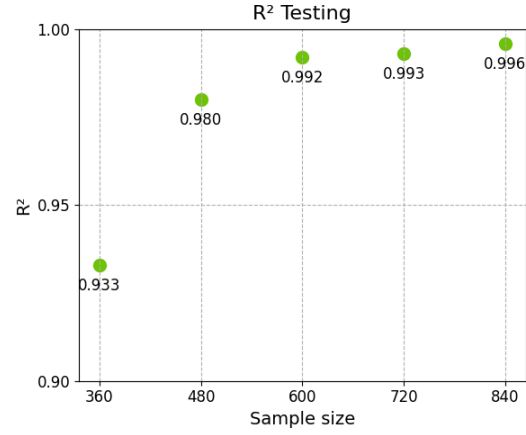
4.2.2 PCE Degree 7 & 9 – Varying Sample

Figure 4.10 and 4.11 provided the R^2 scores for both the training and evaluation datasets using a Polynomial Chaos Expansion (PCE) model of degree 7 and 9, with an increasing number of samples. The results demonstrate a significant improvement in R^2 as the number of samples increases compared to the base case, where the R^2 values were so much low. The most striking feature of these results is how the R^2 values increase as the increase the number of samples. It can be understood as a high order (etc. order 7, 9, 11,...) polynomial introduces a higher level of complexity, allowing the model to fit highly intricate nonlinear relationships in the data. This degree can be beneficial for capturing subtle patterns but also carries a significant risk of overfitting. Hence, a larger dataset is essential to ensure optimal performance, particularly for models with higher complexity like PCE.

For instance, for degree 7, initially 240 sample points ($n_p = 2$) is used. However, this sample is not sufficient to construct an optimal PCE model as it yield very bad R^2 of 0.601 and 0.758 for training and evaluation data respectively. However, when n_p increase to 3, the R^2 is significantly increase to 0.984 and 0.933 for both training and evaluation data. Though more consistent R^2 is achieved when the higher number of sample is used. This result is indicating that the minimum of sample that required to construct very good and consistent PCE model for degree 7 is at $n_p = 4$.



(a) R^2 summary for training PCE
model degree 7

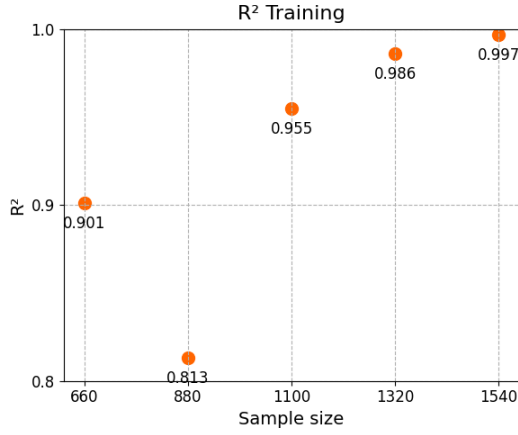


(b) R^2 summary for testing PCE
model degree 7

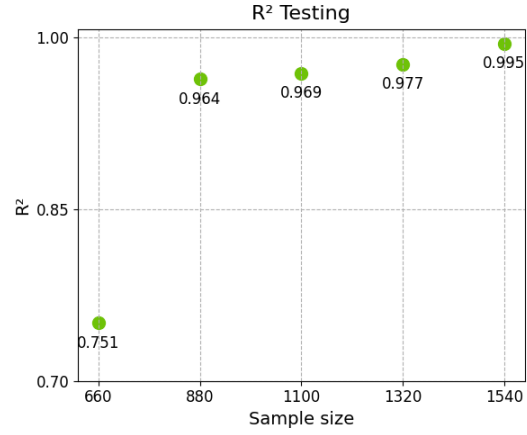
Figure 4.10. Summary PCE degree 7 performance for all varying sample

Table 4.7. Error evaluation PCE model degree 7 for all sample variation

Sample	L_2 mean	L_2 variance	L_2 P90	L_2 P50	L_2 P10
168	6.62E-06	6.03E-02	7.54E-06	2.98E-04	1.91E-04
224	1.32E-05	1.73E-02	1.12E-04	1.29E-04	6.17E-06
280	8.38E-05	1.05E-04	6.77E-05	1.46E-04	5.02E-06
336	7.83E-05	4.06E-03	4.68E-05	1.67E-04	4.21E-05
392	3.78E-05	1.27E-03	6.40E-06	4.61E-05	3.73E-05



(a) R^2 summary for training PCE
model degree 9



(b) R^2 summary for testing PCE
model degree 9

Figure 4.11. Summary PCE degree 9 performance for all varying sample

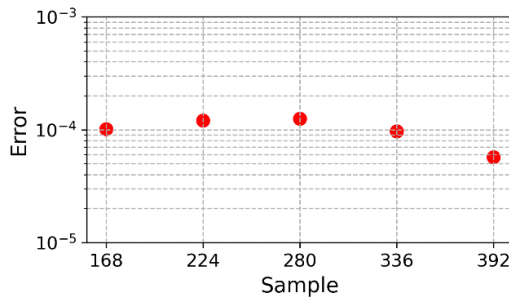
Additionally, for degree 9, it can be concluded that the 450 number of sample ($n_p = 2$) is very lack to construct good PCE model, as it yield very bad R^2 of -1.25 and -1.486 for training and evaluation data during the initial study. It start to give good R^2 when the n_p increase to 3. However, this number of sample is still considere to be less sufficient, as it gives inconsistency R^2 for training and evaluation data. It only start to give consistent good score when the $n_p = 5$ is used

Table 4.8. Error evaluation PCE model degree 9 for all sample variation

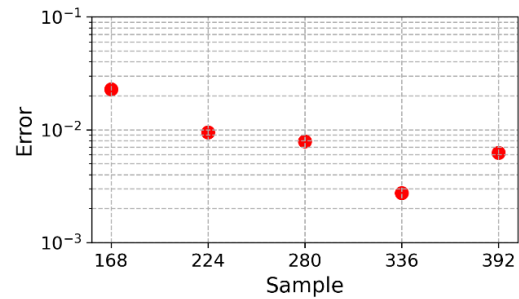
Sample	L_2 mean	L_2 variance	L_2 P90	L_2 P50	L_2 P10
168	2.37E-04	2.25E-01	2.30E-04	4.62E-05	3.64E-04
224	7.23E-05	4.26E-02	1.26E-04	1.60E-05	1.21E-04

280	7.37E-05	2.63E-02	9.64E-05	8.70E-06	1.56E-05
336	6.28E-05	1.82E-02	1.36E-04	9.54E-06	3.03E-06
392	1.25E-05	1.01E-03	4.31E-05	1.61E-05	9.67E-05

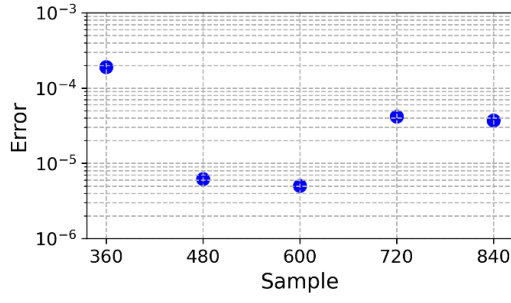
Generally all of the PCE model degree 7 and 9 for all varying sample shown very low error, indicating robust performance for both model (table 4.7 and 4.8). If we look at the figure 20, the error score is quite fluctuative for all varying sample. Same condition also occure for PCE model 9 (figure 4.13), in which the error pattern quite shown fluctuative. However, is not really becomes a concern, since the general trend is shown downward and the range of error is considerable low, especially if we compare to the base case model (PCE model 3) which result satisfstying performance for all of the aspect R^2 score, error score during the initial study.



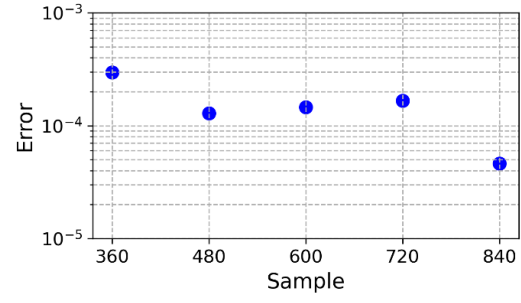
(a) L_2 error in mean



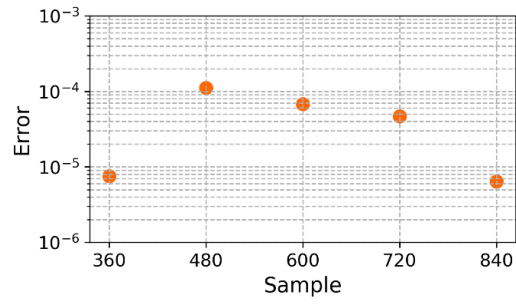
(b) L_2 error in variance



(c) L_2 error in P10

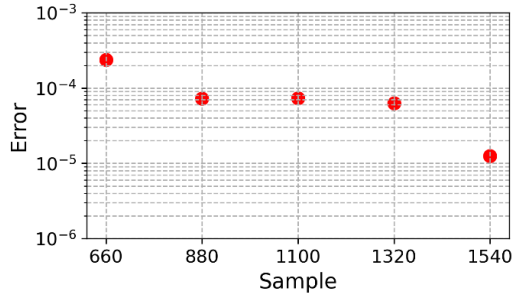


(d) L_2 error in P50

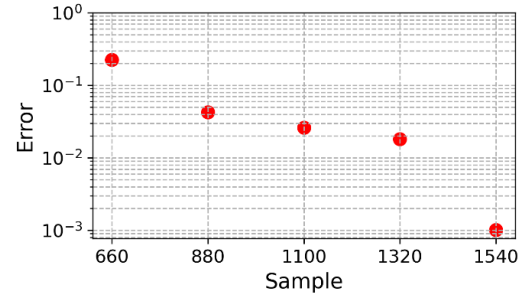


(e) L_2 error in P90

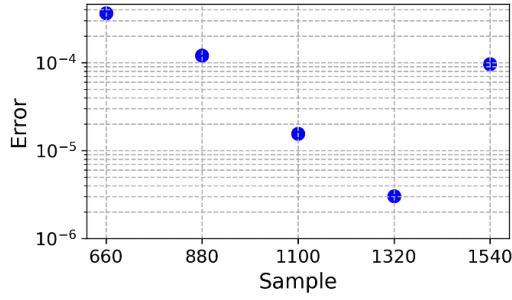
Figure 4.12. Error evaluation PCE model degree 7 for all sample variation



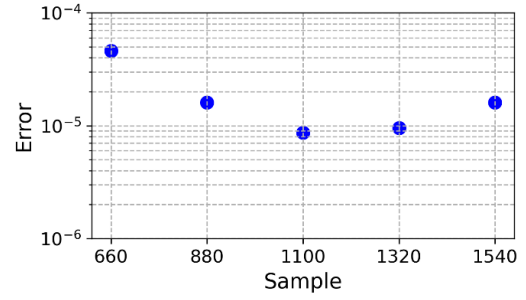
(a) L_2 error in mean



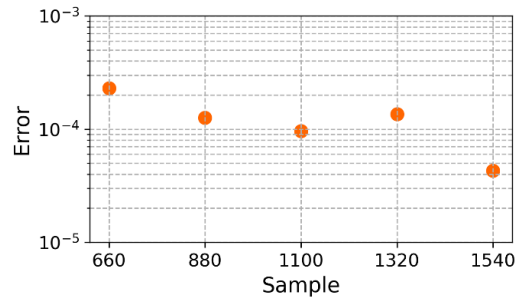
(b) L_2 error in variance



(c) L_2 error in P10



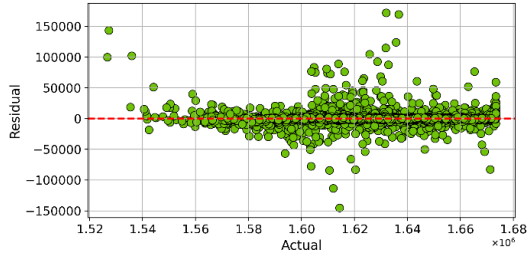
(d) L_2 error in P50



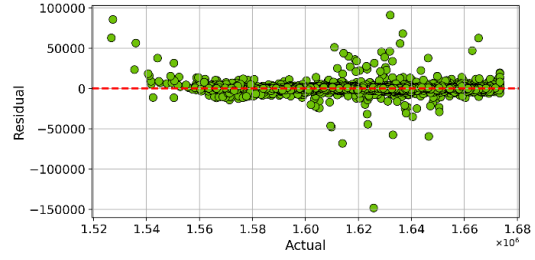
(d) L_2 error in P90

Figure 4.13. Error evaluation PCE model degree 9 for all sample variation

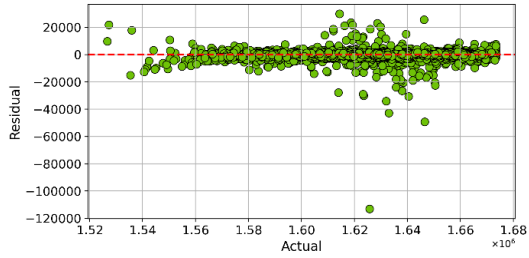
Examining the residual plots of the PCE models with degrees 7 and 9 for all sample variations confirms that the PCE models can accurately estimate cumulative oil production. This is evident from the general distribution of residuals around zero, reflecting strong model performance and low error values. Although there are some outliers, they are acceptable, as their number is small compared to the total sample size of 17,000. Moreover, the outliers consistently decrease as the sample size increases, indicating improved performance of the PCE models. These findings align with the R^2 scores for both models, where a smaller sample size results in the lowest R^2 score and the highest number of outliers and residuals.



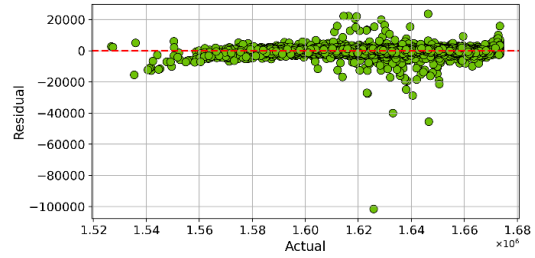
(a) 360 sample points



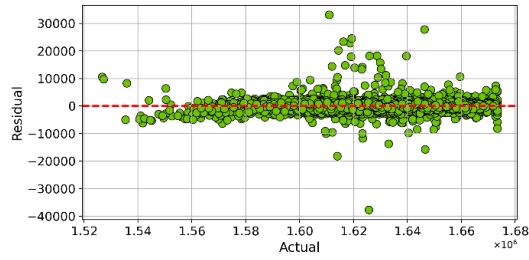
(b) 480 sample points



(c) 600 sample points

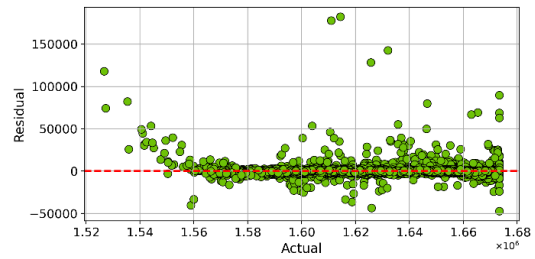
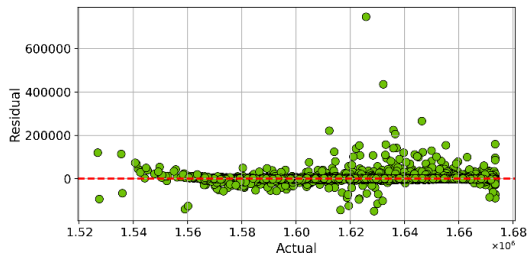


(d) 720 sample points



(e) 840 sample points

Figure 4.14. Residual plot PCE model degree 7 for all sample variations



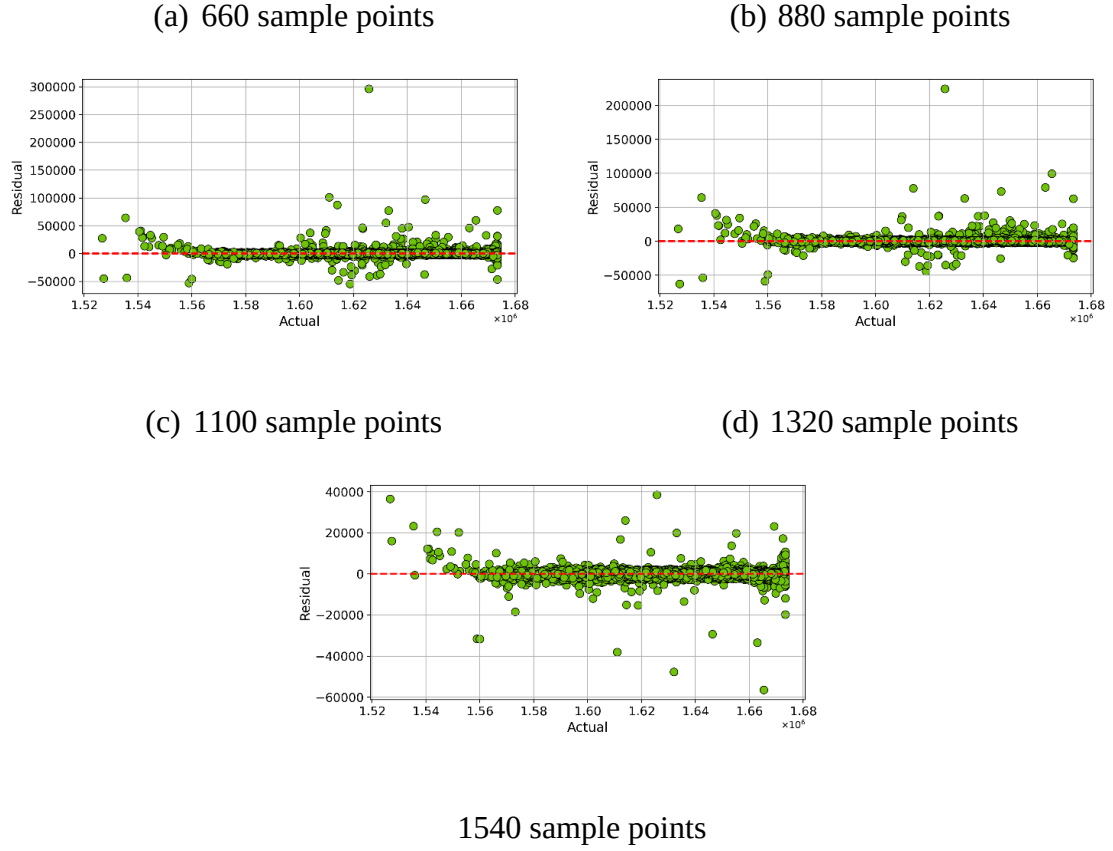
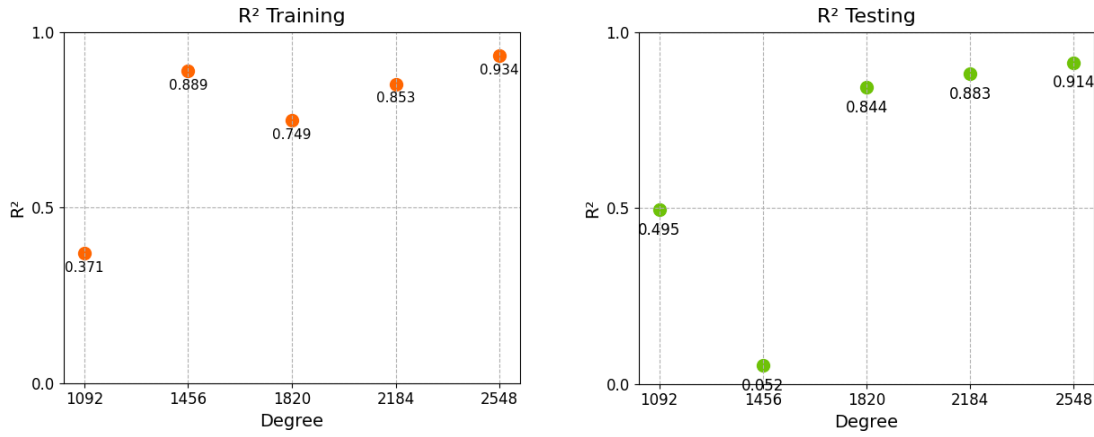


Figure 4.15. Residual plot PCE model degree 9 for all sample variations

4.2.3 PCE Degree 11 – Varying Sample

The results presented in Figure 4.16 provide an analysis of the R^2 scores for the PCE model with degree 11, evaluated on both the training and evaluation datasets using different sample sizes. At sample 1092 ($n_p = 3$), the result indicating massive improvement from the initial study when 728 ($n_p = 2$) of sample were used. Despite of this increase, the model still need to be refine as the R^2 still yield negative score for both training and evaluation. As the others high polynomial orders, the performance of this PCE model improve as the sample number increase. The satisfying score R^2 can be only achieve when the 2548 ($n_p = 7$) sample used, as this give consistent score above 0.9 for both

training and evaluation data. Additionally, all of the error metrics also show decreasing as the sample numbers increase. It is again indicated that a larger dataset is essential to ensure optimal performance, particularly for models with higher order which posses high complexity.



(a) R^2 summary for training PCE

model degree 11

(b) R^2 summary for testing PCE

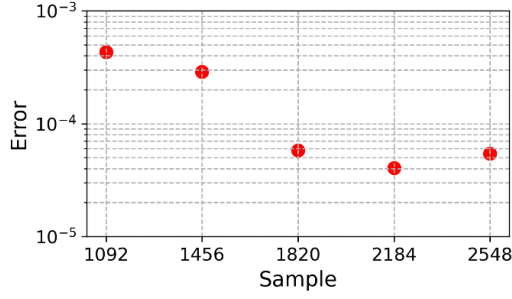
model degree 11

Figure 4.16. Summary PCE degree 11 performance for all varying sample

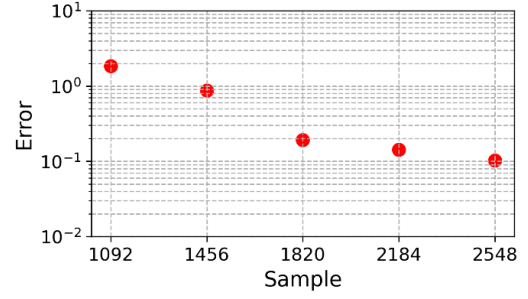
Table 4.9. Error evaluation PCE model degree 11 for all sample variation

Sample	L_2 mean	L_2 variance	L_2 P90	L_2 P50	L_2 P10
1092	4.32E-04	1.85E+00	7.33E-05	4.40E-05	4.52E-05
1456	2.88E-04	8.79E-01	7.71E-05	1.58E-06	9.07E-05
1820	5.79E-05	1.93E-01	8.59E-05	5.77E-05	1.37E-04
2184	4.04E-05	1.43E-01	3.75E-05	6.40E-05	8.61E-05

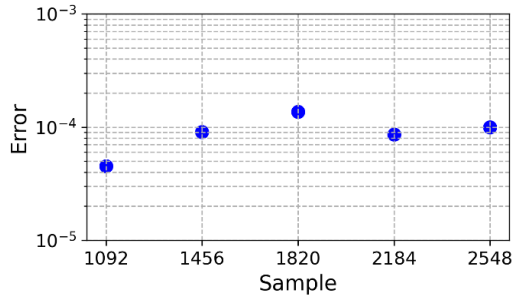
2548	5.44E-05	1.02E-01	4.32E-05	9.53E-06	1.00E-04
------	----------	----------	----------	----------	----------



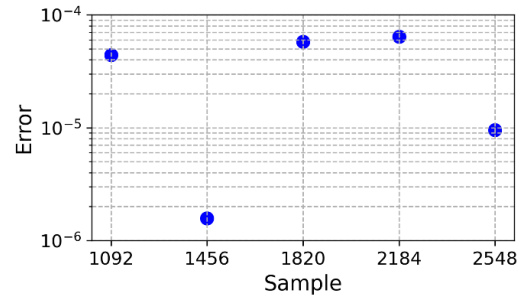
(a) L_2 error in mean



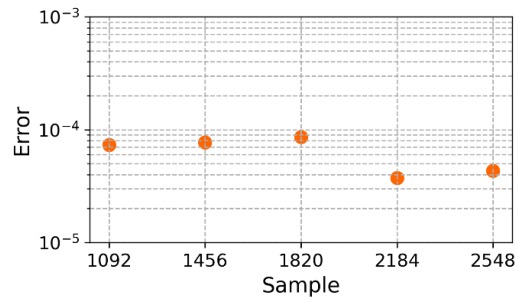
(b) L_2 error in variance



(c) L_2 error in P10



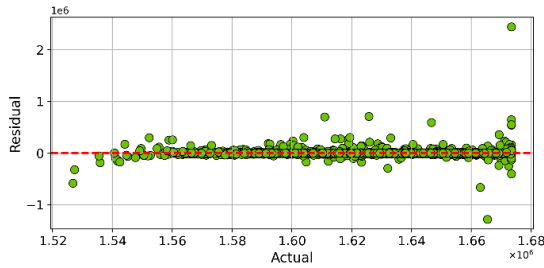
(d) L_2 error in P10



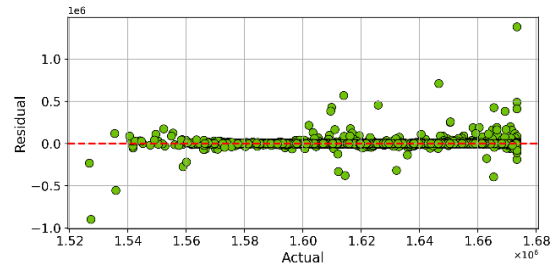
(e) L_2 error in P90

Figure 4.17. Error evaluation PCE model degree 11 for all sample variation

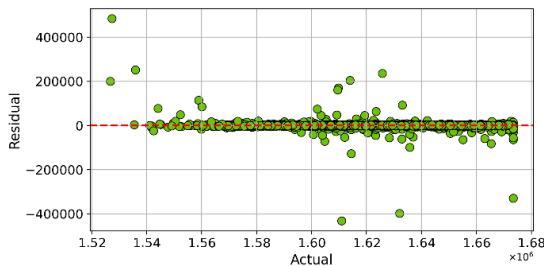
The residual plot reveals that the magnitude of the residuals is extremely high. For instance, in the residual plots for 1,092 and 1,456 sample points, the residual magnitudes reach 2 million and 1 million basis points, respectively, indicating very poor performance from the PCE model. This result aligns with the very low R^2 scores from this model. However, as the number of samples increases, the outliers consistently decrease, suggesting an improvement in the PCE model's performance. Although the residual plot for 2,548 samples ($n_p = 7$) still shows high residual magnitudes compared to other models (PCE degree 5, 7, 9), it is sufficient to conclude that the minimum sample size needed for PCE degree 11 to yield good results is around 2,548 samples ($n_p = 7$). Therefore, the investigation concludes here, as the objective of determining the minimum sample size has been achieved.



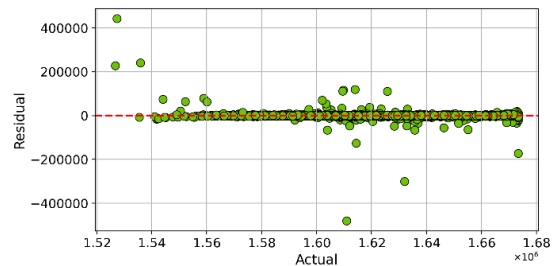
(a) 1092 sample points



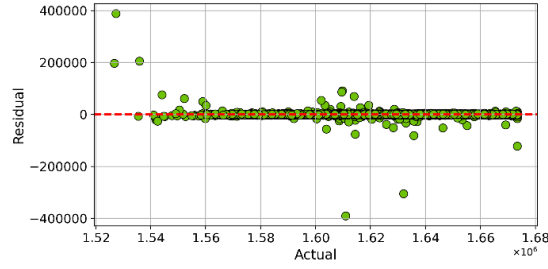
(b) 1456 sample points



(c) 1820 sample points



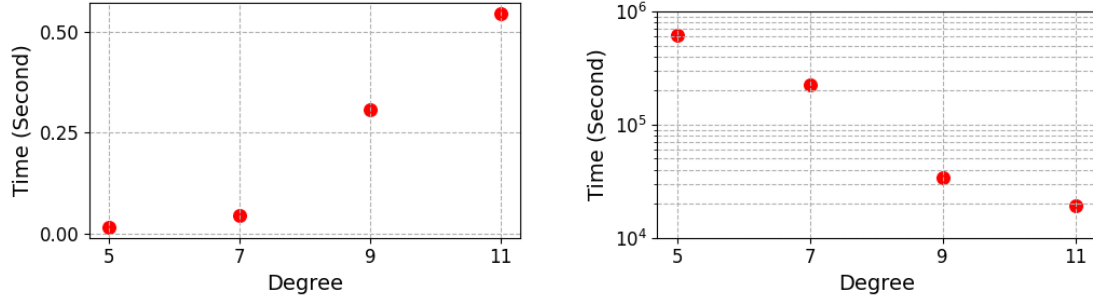
(d) 2184 sample points



(e) 2548 sample points

Figure 4.18. Residual PCE model degree 11 for all sample variation

The analysis reveals that, regardless of the coefficient determination score across all sample variations tested for each degree of the PCE model (5, 7, 9, and 11), all models deliver results that are closely comparable to the outcomes of a converged Monte Carlo simulation with 17,000 realizations. This conclusion is supported by evaluating the mean, variance, and P10, P50, and P90 errors, all of which exhibit minimal discrepancies. Furthermore, as depicted in Figure 4.19, the average runtime for each PCE model is remarkably efficient, taking less than one second. In contrast, achieving convergence with the MCS model requires approximately 10,493 seconds. These findings suggest that PCE models are a viable alternative to MCS models, offering competitive accuracy while dramatically reducing computational time.



(a) Elapsed time for each PCE Model

(b) Elapsed time (%) PCE model
compared to MCS

Figure 4.19. Running time summary of PCE model vs MCS

4.3 Novel Correlation for Degree – Sample Relationship

Based on the finding above, it is important to note that the sample size plays a crucial role in mitigating the adverse effects of increasing polynomial degree. The study reveals that higher-degree polynomials require much larger datasets to maintain their performance and reduce overfitting. For instance, while a degree 3 and 5 PCE model performs well with a sample size roughly 2–3 times the number of coefficients N_t . Higher-degree models, such as degree 7, 9 and 11, require at least 4–7 times N_t to reach acceptable accuracy. Without this increase in sample size, these higher-order models produce poor R^2 values and fail to align their predicted histograms with the true data distribution.

Looking into the Polynomial Chaos Expansion (PCE) model performance in this study, it is clear that the reference-based recommendation for the minimum number of samples—typically twice the number of coefficients—does not always correct, especially for higher-degree polynomials. Our results demonstrate that the required sample size increases significantly as the polynomial degree rises. For instance, while lower-degree

models (such as degree 3 and 5) perform well with twice the number of coefficients, higher-degree models like degree 7, 9, and 11 require larger multipliers, ranging from three to seven times the number of coefficients. This correlation suggests that more complex models need a greater sample size to avoid overfitting and capture the data distribution accurately.

Hence, a generalized formula to estimate the minimum number of samples for any polynomial degree (higher than 5) is proposed as follows:

$$N_s = n_p \times N_t \quad (2)$$

In which the $n_p = (1 + \left\lceil \frac{d}{2} \right\rceil)$, is the multiplier factors, and $N_t = \frac{(M+d)!}{M!d!}$, is the total number of coefficients. An addition of interger “1” in n_p is to ensures the sample size is always more conservative by always using slightly more samples.

Hence the final formula will be written as:

$$N_s = \left(1 + \left\lceil \frac{d}{2} \right\rceil\right) \times \frac{(M+d)!}{M!d!} \quad (4.1)$$

This formula provides a more flexible and accurate guideline for determining the appropriate sample size for PCE models, ensuring robust performance even at higher degrees.

To validate these recommendations, we proceeded to construct several new PCE models with degrees 6, 8, and 10. The sample size for each polynomial model is determined based on our earlier recommendation (equation 20). Thus, the corresponding n_p values for constructing the PCE models are $n_p = 4$, $n_p = 5$, and $n_p = 6$, respectively. For comparison, we also investigated how the model's performance is affected when n_p is

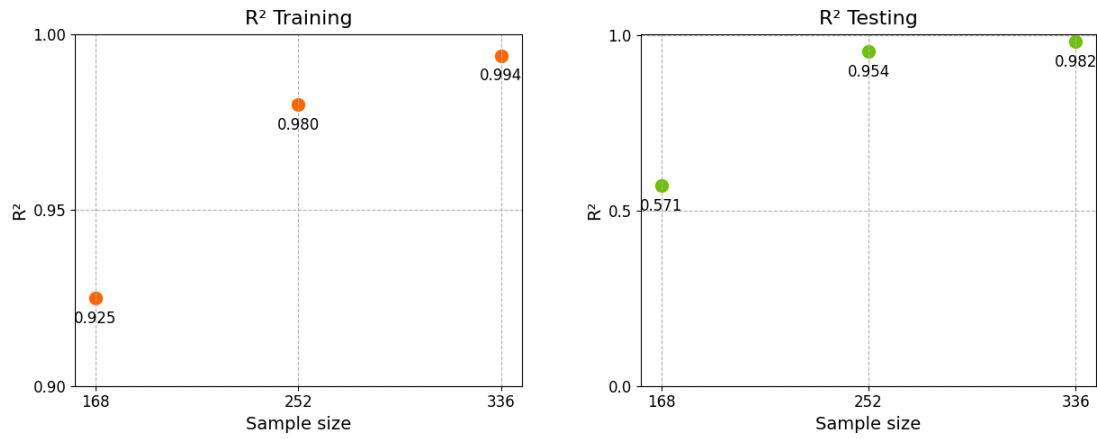
decreased by 1 to 2 factor point for each degree. The summary of sample that will be used for each PCE model is presented in table 4.10.

Table 4.10. Scenario number of sample used for PCE model degree 6, 8, and 10

Maximum Degree	N_t	Sample 1 $(n_p - 2)$	Sample 2 $(n_p - 1)$	Sample 3 $(n_p = n_p)$
6	64	168	252	336
8	165	495	660	825
10	286	1144	1430	1716

Figures 4.19 to 4.20 show the R^2 scores for the PCE models with degrees 6, 8, and 10. As anticipated, the best results for all PCE models come from using the n_p values derived from the recommended equation. For example, in the PCE model with degree 6 (Figure 26), equation 19 indicates that $n_p = 4$ (336 samples) is the minimum required sample size to achieve good model performance. When n_p is reduced by 1 or 2 points, such as $n_p - 1$ (252 samples) and $n_p - 2$ (168 samples), the PCE model yields lower R^2 scores. While $n_p - 1$ (252 samples) seems to perform well for PCE degree 6 and 8, the same does not hold for PCE degree 10, where $n_p - 1$ (1,430 samples) results in an R^2 score below 0.9 for both training and testing. Due to this inconsistency, the previous formula includes the addition of "1" to n_p to ensure the sample size is always slightly larger and more conservative.

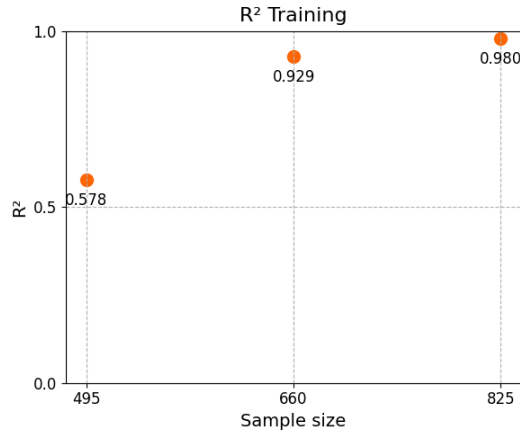
The proposed recommendation for achieving good performance for each PCE degree has been successfully applied to all new PCE models with degrees 6, 8, and 10. Using the sample size derived from equation 20 consistently provides satisfying results for each PCE model.



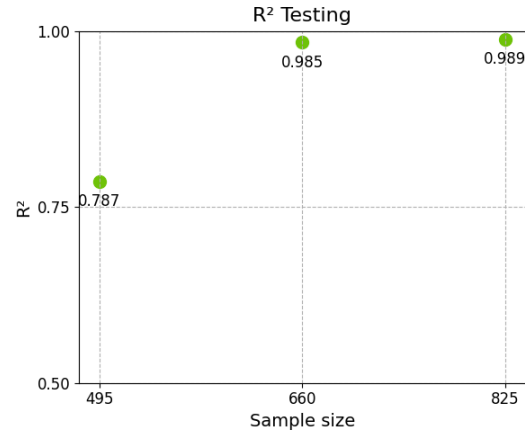
(a) R^2 summary for training PCE
model degree 6

(b) R^2 summary for testing PCE
model degree 6

Figure 4.20. Cross-plot PCE model degree 6 for all sample variation

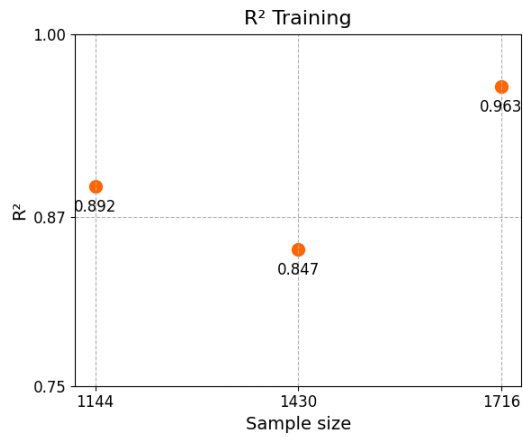


(a) R^2 summary for training PCE
model degree 8

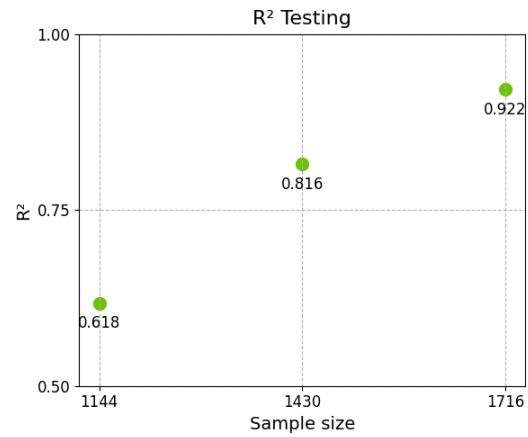


(b) R^2 summary for testing PCE
model degree 8

Figure 4.21. Cross-plot PCE model degree 8 for all sample variation



(a) R^2 summary for training PCE
model degree 10



(b) R^2 summary for testing PCE
model degree 10

Figure 4.22. Cross-plot PCE model degree 10 for all sample variation

CHAPTER 5

CONCLUSION AND FUTURE WORK

5.1 Conclusion

This study investigates the performance of Polynomial Chaos Expansion (PCE) models in predicting cumulative oil production using various polynomial degrees (3, 5, 7, 9, and 11) and different sample sizes. Several key insights emerge from the analysis, highlighting the importance of balancing both polynomial degree and sample size in constructing robust and accurate PCE models.

1. The Polynomial Chaos Expansion (PCE) method offers a highly efficient alternative to traditional Monte Carlo Simulation (MCS) for uncertainty quantification, achieving comparable accuracy with significantly fewer samples than converged MCS which required 17000 realization. For example, a degree 3 and 5 PCE model achieves an R^2 of above 0.95 and error below 1% for the first two moments (mean and variance) using only 50 and 120 samples respectively. Even the higher degree used in this study ($d = 11$) only required 2548 samples to get $R^2 > 0.9$, whereas MCS requires 17000 of realizations to reach the convergence. Additionally, The PCE model achieves convergence in less than a second, whereas Monte Carlo simulation requires over two hours to complete\
2. PCE model performance is strongly influenced by both sample size and polynomial degree. Lower-degree polynomials (e.g., degree 3 and 5) can achieve high accuracy with fewer samples, but higher-degree models (e.g., degrees 7, 9, and 11) require a

much larger sample size to avoid overfitting and capture the data's complexity. For example, degree 7 requires at least 3 times N_t for acceptable accuracy. Similarly, degree 9 and 11 needs 5 times and 7 times N_t respectively for good R^2 scores.

3. The recommended equation from this study, $N_s = \left(1 + \left\lfloor \frac{d}{2} \right\rfloor\right) \times \frac{(M+d)!}{M!d!}$, is found to be valid to search the minimum number of sample that should be used to yield good performance on PCE model, balancing the number of sample and PCE degree properly.

5.2 Future Works

This study successfully presents a novel approach to identify the minimum sample size needed to optimize the performance of a higher-degree Polynomial Chaos Expansion (PCE) model. As the PCE degree and the input random variables increase, the model's coefficients grow exponentially, which in turn requires more samples and raises computational costs. Therefore, it is essential to develop a new method that can automatically determine the sensitivity of PCE coefficients at higher degrees to enhance the model's efficiency.

REFERENCES

- Afanasyev, A., Vedeneeva, E., & Mikheev, I. (2023). Monte Carlo simulation of the maximum migration distance of CO₂ in a sloping aquifer. *Gas Science and Engineering*, 117. <https://doi.org/10.1016/j.jgsce.2023.205078>
- Aistleitner, C., Hofer, M., & Tichy, R. (2012). A CENTRAL LIMIT THEOREM FOR LATIN HYPERCUBE SAMPLING WITH DEPENDENCE AND APPLICATION TO EXOTIC BASKET OPTION PRICING. *https://doi.org/10.1142/S021902491250046X*, 15(7). <https://doi.org/10.1142/S021902491250046X>
- Camacho, A., Talavera, A., Emerick, A. A., Pacheco, M. A. C., & Zanni, J. (2017). Uncertainty quantification in reservoir simulation models with polynomial chaos expansions: Smolyak quadrature and regression method approach. *Journal of Petroleum Science and Engineering*, 153, 203–211. <https://doi.org/10.1016/J.PETROL.2017.03.046>
- Cremon, M. A., Christie, M. A., & Gerritsen, M. G. (2020). Monte Carlo simulation for uncertainty quantification in reservoir simulation: A convergence study. *Journal of Petroleum Science and Engineering*, 190. <https://doi.org/10.1016/j.petrol.2020.107094>
- Dai, Z., Viswanathan, H., Fessenden-Rahn, J., Middleton, R., Pan, F., Jia, W., Lee, S. Y., Mcpherson, B., Ampomah, W., & Grigg, R. (2014). Uncertainty quantification for

- CO₂ sequestration and enhanced oil Recovery. *Energy Procedia*, 63, 7685–7693.
<https://doi.org/10.1016/j.egypro.2014.11.802>
- Dong, Z., Holditch, S. A., & McVay, D. A. (2013). Resource evaluation for shale gas reservoirs. *SPE Economics and Management*, 5(1), 5–16.
<https://doi.org/10.2118/152066-pa>
- Filomeno Coelho, R., Lebon, J., & Bouillard, P. (2011). Hierarchical stochastic metamodels based on moving least squares and polynomial chaos expansion : Application to the multiobjective reliability-based optimization of space truss structures. *Structural and Multidisciplinary Optimization*, 43(5), 707–729.
<https://doi.org/10.1007/S00158-010-0608-5>
- Grenyer, A., Erkoyuncu, J. A., Zhao, Y., & Roy, R. (2021). A systematic review of multivariate uncertainty quantification for engineering systems. In *CIRP Journal of Manufacturing Science and Technology* (Vol. 33, pp. 188–208). Elsevier Ltd.
<https://doi.org/10.1016/j.cirpj.2021.03.004>
- Hosder, S., Walters, R. W., & Balch, M. (2007). Efficient sampling for non-intrusive polynomial chaos applications with multiple uncertain input variables. *Collection of Technical Papers - AIAA/ASME/ASCE/AHS/ASC Structures, Structural Dynamics and Materials Conference*, 3, 2946–2961. <https://doi.org/10.2514/6.2007-1939>
- Hu, X., Chen, X., Parks, G. T., & Yao, W. (2016). Review of improved Monte Carlo methods in uncertainty-based design optimization for aerospace vehicles. In *Progress in Aerospace Sciences* (Vol. 86, pp. 20–27). Elsevier Ltd.
<https://doi.org/10.1016/j.paerosci.2016.07.004>

- IM Sobol'. (1993). Sensitivity estimates for nonlinear mathematical models. *Math. Model. Comput.* <https://cir.nii.ac.jp/crid/1370009142503384324>
- Jia, W., McPherson, B. J., Pan, F., Xiao, T., & Bromhal, G. (2016). Probabilistic analysis of CO₂ storage mechanisms in a CO₂-EOR field using polynomial chaos expansion. *International Journal of Greenhouse Gas Control*, 51, 218–229. <https://doi.org/10.1016/j.ijggc.2016.05.024>
- Jia, W., McPherson, B., Pan, F., Dai, Z., & Xiao, T. (2018). Uncertainty quantification of CO₂ storage using Bayesian model averaging and polynomial chaos expansion. *International Journal of Greenhouse Gas Control*, 71, 104–115. <https://doi.org/10.1016/j.ijggc.2018.02.015>
- Kroese, D. P., & Rubinstein, R. Y. (2012). Monte Carlo methods. *Wiley Interdisciplinary Reviews: Computational Statistics*, 4(1), 48–58. <https://doi.org/10.1002/wics.194>
- Maiy, C. V., & Sudrety, B. (2017). Surrogate models for oscillatory systems using sparse polynomial chaos expansions and stochastic time warping. *SIAM-ASA Journal on Uncertainty Quantification*, 5(1), 540–571. https://doi.org/10.1137/16M1083621/SUPPL_FILE/MAISUDRET_SUPPLEMENT.PDF
- Mehana, M., Callard, J., Kang, Q., & Viswanathan, H. (2020). Monte Carlo simulation and production analysis for ultimate recovery estimation of shale wells. *Journal of Natural Gas Science and Engineering*, 83. <https://doi.org/10.1016/j.jngse.2020.103584>

- Naik, P., Aramideh, S., & Ardekani, A. M. (2019). History matching of surfactant-polymer flooding using polynomial chaos expansion. *Journal of Petroleum Science and Engineering*, 173, 1438–1452. <https://doi.org/10.1016/j.petrol.2018.09.089>
- Oladyshkin, S., Class, H., Helmig, R., & Nowak, W. (2011). A concept for data-driven uncertainty quantification and its application to carbon dioxide storage in geological formations. *Advances in Water Resources*, 34(11), 1508–1518. <https://doi.org/10.1016/j.advwatres.2011.08.005>
- Palacios, A., Koon-Koon, R., Miranda-Flores, G., Lewis, S., Castro-Olivera, P. M., Cordova-Lizama, A., & Hernandez-Garcia, A. I. (2024). A Monte Carlo-based approach for the evaluation of Mexico's geothermal potential. *Renewable Energy*, 223. <https://doi.org/10.1016/j.renene.2024.120006>
- Quosay, A. A., Knez, D., & Ziaja, J. (2020). Hydraulic fracturing: New uncertainty based modeling approach for process design using Monte Carlo simulation technique. *PLoS ONE*, 15(7 July). <https://doi.org/10.1371/journal.pone.0236726>
- Saltelli, A., Annoni, P., Azzini, I., Campolongo, F., Ratto, M., & Tarantola, S. (2010). Variance based sensitivity analysis of model output. Design and estimator for the total sensitivity index. *Computer Physics Communications*, 181(2), 259–270. <https://doi.org/10.1016/J.CPC.2009.09.018>
- Sarma, P., Etc, C., & Xie, J. (2011). *SPE 141963 Efficient and Robust Uncertainty Quantification in Reservoir Simulation with Polynomial Chaos Expansions and Non-intrusive Spectral Projection*.

- Shen, D., Wu, H., Xia, B., & Gan, D. (2020). Polynomial Chaos Expansion for Parametric Problems in Engineering Systems: A Review. In *IEEE Systems Journal* (Vol. 14, Issue 3, pp. 4500–4514). Institute of Electrical and Electronics Engineers Inc. <https://doi.org/10.1109/JSYST.2019.2957664>
- Stoian, E. (1965). Fundamentals and Applications of the Monte Carlo Method. *Journal of Canadian Petroleum Technology*, 4(03), 120–129. <https://doi.org/10.2118/65-03-02>
- Tarakanov, A., & Elsheikh, A. H. (2019). Regression-based sparse polynomial chaos for uncertainty quantification of subsurface flow models. *Journal of Computational Physics*, 399. <https://doi.org/10.1016/j.jcp.2019.108909>
- Wang, Y., Voskov, D., Daniilidis, A., Khait, M., Saeid, S., & Bruhn, D. (2023). Uncertainty quantification in a heterogeneous fluvial sandstone reservoir using GPU-based Monte Carlo simulation. *Geothermics*, 114. <https://doi.org/10.1016/j.geothermics.2023.102773>
- Wiener, N. (1938). The Homogeneous Chaos. *American Journal of Mathematics*, 60(4), 897. <https://doi.org/10.2307/2371268>
- Xiu, D., & Karniadakis, G. E. (2003). Modeling uncertainty in flow simulations via generalized polynomial chaos. *Journal of Computational Physics*, 187(1), 137–167. [https://doi.org/10.1016/S0021-9991\(03\)00092-5](https://doi.org/10.1016/S0021-9991(03)00092-5)
- Yaghoubi, V., Marelli, S., Sudret, B., & Abrahamsson, T. (2017). Sparse polynomial chaos expansions of frequency response functions using stochastic frequency

transformation. *Probabilistic Engineering Mechanics*, 48, 39–58.
<https://doi.org/10.1016/J.PROBENGMECH.2017.04.003>

Zhang, J. (2021). Modern Monte Carlo methods for efficient uncertainty quantification and propagation: A survey. In *Wiley Interdisciplinary Reviews: Computational Statistics* (Vol. 13, Issue 5). John Wiley and Sons Inc. <https://doi.org/10.1002/wics.1539>

Zhou, Y., & Zhang, X. (2019). Natural frequency analysis of functionally graded material beams with axially varying stochastic properties. *Applied Mathematical Modelling*, 67, 85–100. <https://doi.org/10.1016/J.APM.2018.10.011>

VITAE

Name : *Tengku Muhammad Sofyan Astsauri*

Nationality : *Indonesia*

Date of Birth : *6/14/1998*

Email : *tengkusofyan41@gmail.com*

Address : *H. M Yunus Street, Cemerlang Asri Complex, Building B,
No. 10, Tanjungbalai, North Sumatera, 21369*

Academic Background : *B. Eng in Petroleum Engineering from Islamic University
of Riau*

List of Journal Publications (Thesis – Related):

- Astsauri, T., Wang, Y. (2024). Balancing Degree and Sample Size of Polynomial Chaos Expansion Method for Optimum Uncertainty Quantification. Natural Resource Research. **Submitted**

List of Journal Publications (Non Thesis – Related):

- Astsauri, T., Habiburrahman, M., Ibrahim, A. F. (2024). Evaluation of Reservoir Conditions on CO₂ Trapping Dynamics on Depleted Gas Reservoir for CCUS. **Drafted**
- Utilizing Machine Learning for Flow Zone Indicators Prediction and Hydraulic Flow Unit Classification, Scientific Reports <https://doi.org/10.1038/s41598-024-54893-1>

- Machine Learning Application for Facies Distribution Using Combined K-Means and Random Forest Algorithm: Case Studies from Central Sumatra Basin, Lecture Notes in Networks and Systems (Springer) https://doi.org/10.1007/978-981-97-0573-3_46
- Applied Random Forest for Parameter Sensitivity of Low Salinity Water Injection (LSWI) Implementation on Carbonate Reservoir, Alexandria Engineering Journal (Scopus Q1) <https://doi.org/10.1016/j.aej.2021.06.096>
- Applied Fractional Factorial Design for CO₂ Immiscible Huff and Puff Technique in Sumatra Light Oil Reservoir, Presented on 3rd I-COSINE 2021, <https://doi.org/10.1063/5.0114449>