3-2024

Improvement of flow coefficient estimation with limited well test data for real-time condition analytics of choke valve

T. Visawameteekul

Tin Seong KAM
Singapore Management University, tskam@smu.edu.sg

P. Thamvechvitee

Citation

Visawameteekul, T.; KAM, Tin Seong; and Thamvechvitee, P.. Improvement of flow coefficient estimation with limited well test data for real-time condition analytics of choke valve. (2024). *Offshore Technology Conference Asia, OTCA 2024: Kuala Lumpur, February 27 - March 1*. 1-17.
Available at: https://ink.library.smu.edu.sg/sis_research/9348

OTC Asia 2024

OTC-34759-MS

Improvement of Flow Coefficient Estimation with Limited Well Test Data for Real-time Condition Analytics of Choke Valve

T. Visawameteekul[1], T. Kam[2], P. Thamvechvitee[2]

[1]PTT Exploration and Production Public Company Limited, Bangkok, Thailand
[2]Singapore Management University, Singapore, Singapore

ABSTRACT

The study discusses a method for monitoring the internal condition of choke valves to predict sand erosion using flow coefficient (Cv). The method calculates the Cv of the choke valve by utilizing eight parameters and compares it to the newly manufactured value to generate warnings. However, the availability of spot-check well test data can significantly impact the model's efficiency if tests are performed infrequently. To address this issue, **the Extended Cv monitoring method** is proposed in this paper.

The main purpose of this study is to develop a model for estimating Cv value in the absence of well test data. This will enhance the present Cv monitoring system, which currently only monitors the valve when the well is being tested. This study aims to bridge a gap in the Cv monitoring approach by **evaluating wellhead operational data** and **dynamic well test data** instead of relying on only **static well test data**. The proposed supplemental data capture dynamic features and are collected continually, which allows us to analyze the internal status of choke valves in a continuous manner. Three representative wells from the Greater Bongkot South asset are chosen as showcases for the study. The study result indicates promising results for choke valve real-time condition monitoring.

The proposed method has been proven to enable online condition monitoring in the absence of well test data. By predicting valve condition, warnings can be generated to limit operation and prevent potential harm to plant integrity and personal safety. The Extended Cv monitoring method overcomes the limitation of the well test-based model, making it more efficient by utilizing continuously measured parameters data and employing machine learning techniques.

This paper provides a useful reference for future studies to forecast Remaining Useful Life (RUL) of choke valve. The method presented in this study has the potential for expansion to other wells and has the potential to be applied in other industries facing similar issues. Overall, this study provides a valuable contribution to the development of methods for monitoring and predicting the internal condition of choke valves to address the challenges of sand production in the oil and gas industry.

INTRODUCTION

Sand erosion on choke valves is a significant issue that affects the operation and overall safety of plants, as well as their reliability and integrity. The degradation of trim components within the valves cannot be directly observed, which makes choke valve performance monitoring challenging, especially for sand operation wells. PTTEP has identified three operating fields in the Gulf of Thailand and Gulf of Martaban as areas of concern for sand production. Among these fields, one field experiences a moderate sand rate, while the others face a more severe sand rate. It is worth noting that one of the fields has recently encountered issues with sand production. Despite the inspection program in place, all fields have experienced gas external leakage from the eroded choke valve bodies.

To address this issue, **the CV monitoring method** was introduced in 2016 in the PTTEP operating fields in the Gulf of Thailand and in 2020 in the Gulf of Martaban. This method utilizes well test data to evaluate the internal condition of choke valves by calculating the operating flow coefficient (Cv) based on deviation from the theoretical value. This method provides early warning before the choke valve body is damaged and can be implemented quickly and affordably, as it does not require any additional hardware. Only well tested data variables are used for analysis.

The primary constraint of the flow coefficient method is its reliance on well test data. This approach's application is limited by the availability of well test data for periodic checks. It permits only spot monitoring using **static well test** data, which becomes increasingly challenging as the number of wells grows. With infrequent well tests, the model's efficiency is adversely affected since it relies solely on such data, making it difficult to alert stakeholders in a timely manner.

In addition, the flow coefficient method also has a second disadvantage that can impact its effectiveness. This disadvantage relates to the method being a lag indicator, meaning that the valve condition is only measured at the time the well is routed to test. This can result in alerts being activated too late, after the valve condition has already deteriorated significantly. To address this concern, a threshold with a large contingency can be set, which provides a safer operation but comes at the cost of reduced cost effectiveness.

Paper study goal and objectives

To enhance the existing Cv monitoring method, which currently only monitors the well when it is being tested. We aim to estimate Cv value in the absence of well test data. This study aims to bridge a gap in the Cv monitoring method by allowing us to analyze the internal status of choke valves in a continuous manner. The resulting model will connect spot checks and continuously analyze the evolution of erosion on the internal components of the valve over time.

Paper structure

In the Introduction section, we provide a detailed description of the traditional CV monitoring method used to assess the erosion state of choke valves with Cv. However, we also address the limitations of this method and explain the goal of our study, which is to overcome these limitations.

Moving on to the Content section, we delve into the application of machine learning techniques for equipment monitoring. We explore various techniques employed to monitor equipment health and discuss the potential benefits of utilizing machine learning algorithms for this purpose. For the Research Designs and Methods, we provide a comprehensive explanation of the research scope, data sources, and analysis methodology employed in our study. We outline the steps taken to develop the extended Cv method for assessing choke valve erosion and elaborate on how machine learning algorithms are utilized to analyze the collected data. We then proceed to demonstrate the application of the extended Cv method to three representative wells and examine the outcomes for each well individually. Furthermore, we discuss the potential implications of these outcomes for choke valve health monitoring and the maintenance of surface facilities.

Finally, in the Conclusion section, we draw conclusions based on the current state of the study. We discuss the limitations encountered and potential future directions for this research, emphasizing the contributions this study makes to the field of choke valve erosion assessment and equipment health monitoring.

CONTENT

Data and Variables

The following section aims to provide a comprehensive overview of the various data sources that have been considered for this study. Each data source will be explained in detail to facilitate a better understanding of the study's data collection process.

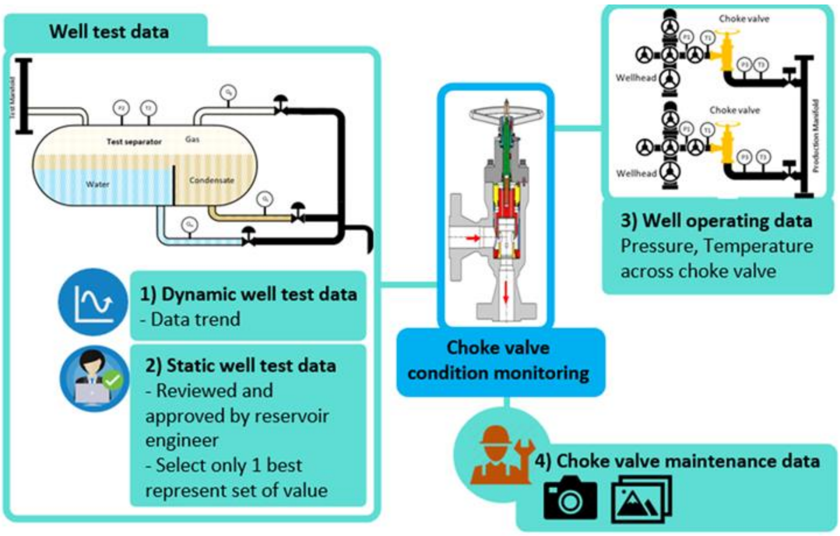
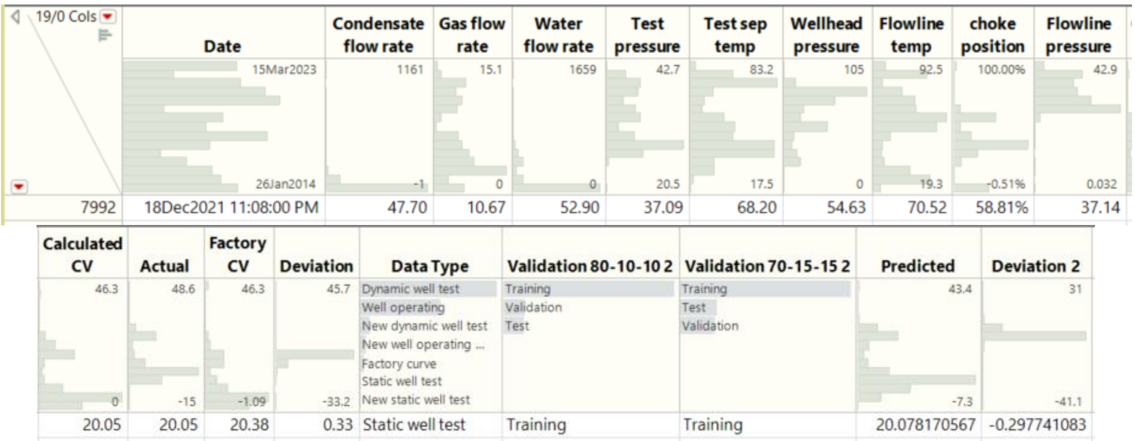


Figure 1 Data sources diagram

1) Static well test data

The ROADS system is a tool used in this study to retrieve **static well test** data. This system is particularly useful during the process of routing a well to a test separator through a test manifold. As the well flows through the separator, the well test data trend is generated, allowing reservoir engineers to monitor and analyze the well's performance.

After the well test data will be carefully examined and authorized by a qualified reservoir engineer, it is then logged as valid data in the ROADS system. However, to ensure accuracy and efficiency, only one average value from a selected period of stable fluid flow is chosen as a representative record in each well test. While this technique has proven to be effective, it has also resulted in a relatively small number of data points for this type of data.



19/0 Cols	Date	Condensate flow rate	Gas flow rate	Water flow rate	Test pressure	Test sep temp	Wellhead pressure	Flowline temp	choke position	Flowline pressure
	15Mar2023	1161	15.1	1659	42.7	83.2	105	92.5	100.00%	42.9
	26Jan2014	-1	0	0	20.5	17.5	0	19.3	-0.51%	0.032
7992	18Dec2021 11:08:00 PM	47.70	10.67	52.90	37.09	68.20	54.63	70.52	58.81%	37.14

Calculated CV	Actual	Factory CV	Deviation	Data Type	Validation 80-10-10 2	Validation 70-15-15 2	Predicted	Deviation 2
46.3	48.6	46.3	45.7	Dynamic well test Well operating New dynamic well test New well operating ... Factory curve Static well test New static well test	Training Validation Test	Training Test Validation	43.4	31
0	-15	-1.09	-33.2				-7.3	-41.1
20.05	20.05	20.38	0.33	Static well test	Training	Training	20.078170567	-0.297741083

Figure 2 Screenshot of Static well test data

2) Dynamic well test data

While **static well test** data provides crucial information regarding the well's performance at a specific moment in time, dynamic well test data offers a more comprehensive and detailed view of the well's behavior over time. **Dynamic well test data** is derived from well test activities, but instead of obtaining only one **static well test** input from a single test, the parameters are gathered every 10 minutes, providing multiple data points from a single test. This results in a significantly larger number of data points, which can be used to train models more effectively and produce more accurate predictions. The parameters used in **dynamic well test data** are the same as those used in **static well test data**. Using the ISA S75.01-1985 equation, we can calculate the Cv.

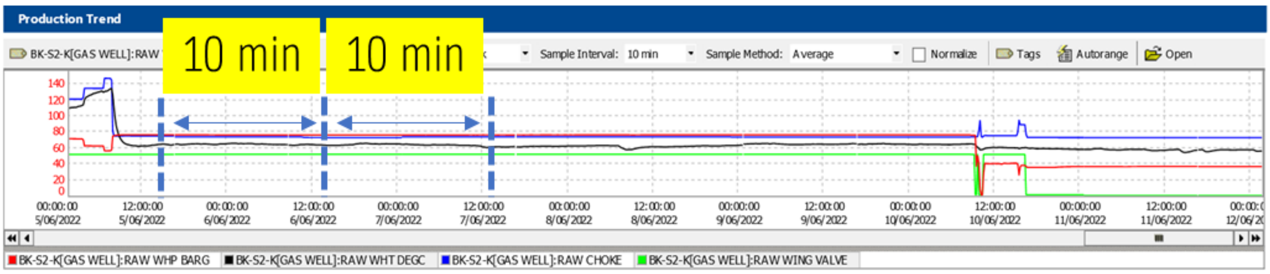
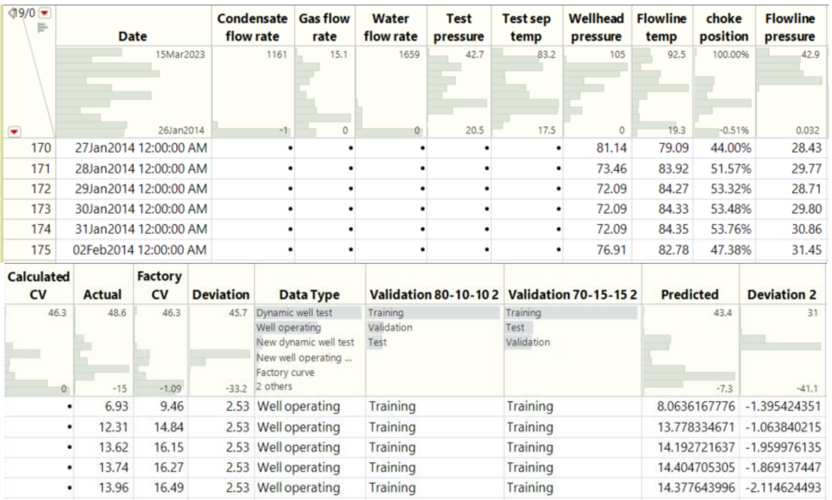


Figure 3 Sampling of Dynamic well test data

3) Well operating data

To enhance the model performance and effectiveness, we link them with the **well operational dataset**. This dataset provides valuable information even when the well is not actively being tested, as data on the well's operation is continually being gathered. The well operational dataset is obtained from the PI system once every day. However, this dataset only includes **well operating data** such as upstream/downstream pressure, temperature, and choke valve position. As a result, a physics-based method using only this data cannot be used to determine Cv. Instead, the Derived Cv is constructed using the previous deviation value of Cv from a **Dynamic** or **Static well test** as a reference condition until the next test is performed. By using this method, it is possible to estimate the Cv value accurately, even when the well is not being actively tested.



	Date	Condensate flow rate	Gas flow rate	Water flow rate	Test pressure	Test sep temp	Wellhead pressure	Flowline temp	choke position	Flowline pressure
170	27Jan2014 12:00:00 AM	•	•	•	•	•	81.14	79.09	44.00%	28.43
171	28Jan2014 12:00:00 AM	•	•	•	•	•	73.46	83.92	51.57%	29.77
172	29Jan2014 12:00:00 AM	•	•	•	•	•	72.09	84.27	53.32%	28.71
173	30Jan2014 12:00:00 AM	•	•	•	•	•	72.09	84.33	53.48%	29.80
174	31Jan2014 12:00:00 AM	•	•	•	•	•	72.09	84.35	53.76%	30.86
175	02Feb2014 12:00:00 AM	•	•	•	•	•	76.91	82.78	47.38%	31.45

Calculated CV	Actual	Factory CV	Deviation	Data Type	Validation 80-10-10 2	Validation 70-15-15 2	Predicted	Deviation 2
•	6.93	9.46	2.53	Well operating	Training	Training	8.0636167776	-1.395424351
•	12.31	14.84	2.53	Well operating	Training	Training	13.778334671	-1.063840215
•	13.62	16.15	2.53	Well operating	Training	Training	14.192721637	-1.959976135
•	13.74	16.27	2.53	Well operating	Training	Training	14.404705305	-1.869137447
•	13.96	16.49	2.53	Well operating	Training	Training	14.377643996	-2.114624493

Figure 4 Screenshot of well operating data

4) Choke valve maintenance history

To validate the output of the model, the choke valve maintenance report is utilized. This report contains data related to choke valve replacement maintenance and includes MS Word files with photographs of the choke valve interior parts. To ensure accuracy, historical maintenance reports have been compiled and summarized in a table for analysis. Domain experts will then identify the photographs from this report to justify the valve condition.

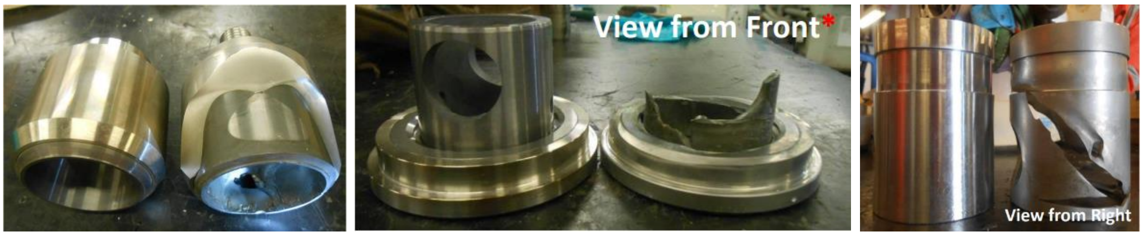


Figure 5 Photo from maintenance report

5) Manufacturer valve Cv curve

Choke valves are designed differently, and as a result, each valve has its unique Cv curve. Typically, when purchasing a new valve, the manufacturer will provide the Cv curve for that specific valve. The example of Cv curve is shown in figure 8, with the valve position as the x-axis and the Cv value as the y-axis. This curve represents the brand-new and healthy state of the valve, and it serves as a reference point for determining the valve's health during operation.

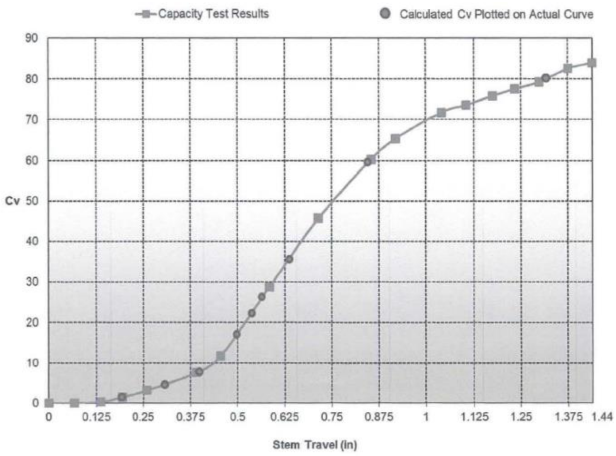


Figure 6 Example of valve Cv curve

Statement of Theory and Definitions

In this study, we utilized the off-the-shelf SAS JMP Pro 16 software as the primary tool for a range of data management tasks, including data cleaning, modification, machine learning model development, and analysis. Additionally, we have employed the same software to calibrate our models, which have been designed to predict the health status of choke valves based on the value of their flow coefficient (Cv).

To comprehend how **the Cv monitoring method** operates, it is essential to first explain what the Cv value is and how it is computed. The flow coefficient is the ability of a valve to enable the flow of gas or liquid through it. In simple terms, a valve with a larger opening will have a higher Cv value. As the valve opens, its Cv value increases until it reaches its maximum possible value, or 100% open Cv *(Kimray. (n.d.)*. This information is represented in Figure 2, where the valve opening position is plotted on the x-axis and the Cv value is plotted on the y-axis. The black curve denotes the factory Cv, which is also referred to as the theoretical Cv. The yellow and red curves in Figure 2 are alert threshold that were identified by the domain experts.

The **static well test** data is used to determine the Cv. The following variables are utilized: choke valve position, gas flow rate, condensate flow rate, water flow rate, test separator pressure, test separator temperature, wellhead pressure, and wellhead temperature. These variables are then plugged into an ISA S75.01-1985 equation, as depicted in equations 1. The ISA S75.01-1985 equation is a critical tool used to calculate the flow coefficient of valves based on various fluid parameters which is critical for optimizing the performance of the valve and the system it is a part of.

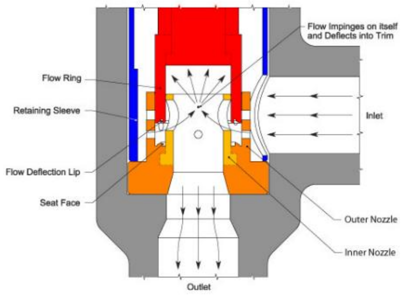


Figure 7 Choke valve design to dissipate the energy to reduce wearing

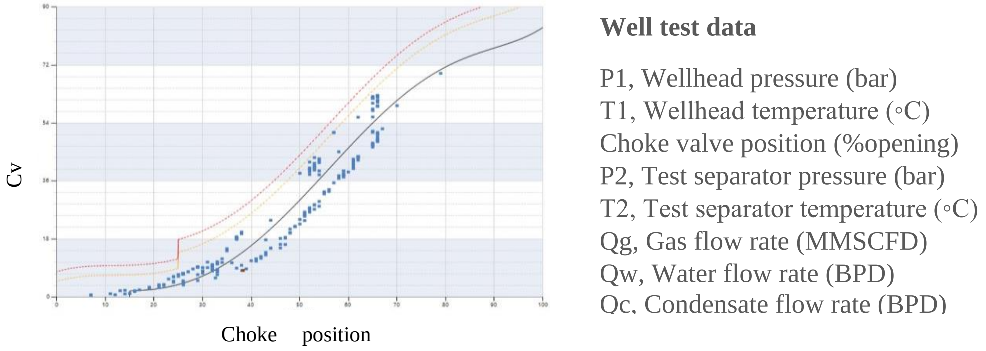


Figure 8 Flow coefficient (Cv) monitoring method

$$C_v = \frac{w_C + w_W + w_G}{63.3} \times \sqrt{\left(\frac{\left(\frac{w_G}{w_C + w_W + w_G}\right) \times \frac{T}{2.7 \times P_1 \times SG_G}}{\left(1 - \frac{P_1 - P_2}{P_1 - 3 \times F_k \times 0.68}\right)^2} + \left(\frac{w_C}{w_C + w_W + w_G}\right) \times \frac{1}{62.4 \times SG_C} + \left(\frac{w_W}{w_C + w_W + w_G}\right) \times \frac{1}{62.4 \times SG_W} \right) + \frac{1}{\Delta P}}$$

Equation 1 the Cv calculation formula

Table 1 Parameter definition inside equation 1

Symbol	Definition	Units	Formula/Value
C_v	Flow coefficient	-	Equation 1 and 2
W_C	Mass flow rate of condensate	Lb/hr	$Q_C \times SG_C \times 14.608$

W_W	Mass flow rate of water	Lb/hr	$Q_W \times SG_W \times 14.608$
W_G	Mass flow rate of gas	Lb/hr	$\frac{Q_G \times SG_G}{24 \times 13.103}$
T	Inlet temperature	◦R	◦F + 459.67
P_1	Inlet pressure	Psia	Psig + 14.7
P_2	Outlet pressure	Psia	Psig + 14.7
SG_G	Gas specific gravity	-	
SG_C	Condensate specific gravity	-	
SG_W	Water specific gravity	-	
F_k	Ratio of Specific Heats factor	-	$k/1.4$
F_L	Pressure recovery factor	-	0.92
X_T	Pressure drop ratio factor	-	0.68

To address the issue of data scarcity from well tests, we introduced two new data sources: **Dynamic well test data** and **well operating data**. Unlike the traditional approach of using **static well tests**, which only select one representative value from each test, the **Dynamic well test** data continuously measures the trend of well test parameters. This helps to increase the number of training data from each well test and provides a more comprehensive picture of the well's performance over time. Moreover, when well tests are not being performed, the well operating data will be utilized. By incorporating both **Dynamic well test** data and well operating data, we can ensure that the machine learning model is trained and tested using a diverse set of data that better represents the actual operating conditions of the well.

The output of the machine learning model will be justified with a choke valve maintenance report, which will provide insights into the performance of the choke valves and identify any potential issues. By utilizing these new data sources and providing a comprehensive report, this study aims to improve the accuracy and practicality of the machine learning model and ultimately, help to enhance the safety and efficiency of the choke valve maintenance plan.

Model Development

In this research, we present the methodology that focus on development of a unique model for each specific well rather than to pursue finding universal model that can represent all well. This approach was adopted due to early study revelation show that there were many factors which could not be included in a universal model such as the specific properties of each well, the type of sand erosion, and other variables that can vary significantly from well to well. The model is also aimed to cover a wide range of valve states and parameters affecting choke valve performance, such as flow rate, pressure, temperature, fluid type, an. Three wells with a diverse range of properties were chosen to represent various operating conditions with different properties. By selecting wells with a range of these properties, we could ensure that our model would be relevant to accurately predict the health status of choke valves across a broad spectrum of conditions.

The accuracy and applicability of a machine learning model are heavily reliant on careful selection of data. Thus, vast amount of **Well operating data** that encompasses various choke valve states was collected. The data need rigorously validation to ensure that it is suitable for machine learning model training and testing. Next, choke valve condition data was collected from **Choke valve maintenance history.** These conditions vary from healthy to broken, which are not specific to any valve manufacture or model. This ensures that the model can account for different valve states, thereby making it more comprehensive and robust.

To provide concrete examples of how the collected data is used, this study has chosen three wells from different platforms to serve as show cases in this paper. The selected wells are **BK-S2-K**, **BK-S3-N**, and **BK-S4J**. The decision to choose these wells was based on two crucial factors: The availability of diverse valve condition data and the representation of different well properties.

Table 1 provides a detailed summary of the dataset for each well, showcasing an interesting trend in the amount of data points per well. The number of datasets per well has increased significantly compared to predecessor modeling method in 2016 from less than 100 to several thousand. This expansion can be attributed to the collection of **dynamic well test data** that enables the gathering of parameter trends at a frequency of every10 minutes, thereby generating multiple datasets from a single well test. With a more extensive dataset, these models can better capture the complexity and variability of the actual system. Additionally, more data points provide a more diversify representation of the system's behavior, allowing for more specific predictions and better-informed decision-making.

Table 2 Number of data points for each well

Well	Static well test	Dynamic well test	Well operating	Data collected period
BK-S2-K	45	5,300	3,060	1 Jan 14 – 8 Aug 22
BK-S3-N	58	3,304	2,368	5 Jan 16 – 8 Aug 22
BK-S4-J	95	6,306	2,723	21 Jan 16 – 8 Aug 22

The health of the choke valve is evaluated using calculated Cv value based on the parameters obtained from each dataset. Figure 9 presents a simplified schematic of the parameters used in the Cv calculation. The calculated Cv value represent model's variable that serves as choke valve health indicator. In addition, calculated Cv value from continuous data can consistently be used for identifying any gradual changes in the valve's performance, which can help prevent sudden failures and associated operational disruptions.

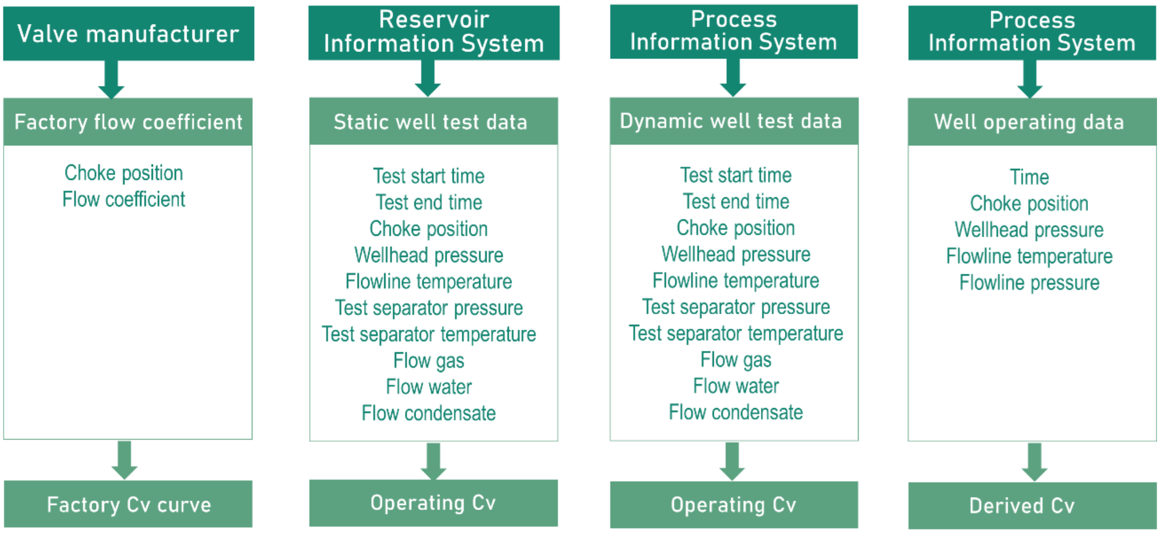


Figure 9 Data integration diagram

Explore the performance of statistical methods and machine learning methods

To determine the most effective approach for predicting the health status of choke valves, we conducted a rigorous preliminary test comparing the performance of statistical methods and machine learning methods. This test was designed to evaluate the predictive accuracy of each approach as well as their ability to handle different types of data and outlier points.

In this test, we first selected a representative sample of wellhead operating data, which included a range of predictor variables related to the behavior of the choke valve under different operating conditions. We employed a range of statistical methods, including Least Squares Regression models and generalized regression, to create predictive models based on this data. Then we utilized more advanced machine learning methods such as Random Forest, Boosted Forest, and Extreme Gradient Boosting (XGBoost) to generate additional models.

We have discovered that machine learning methods outperform statistical methods in predicting the Cv value of choke valves (Refer to detail in Appendix A). This finding is in line with the characteristics of erosion on choke valves from historical record and supports the notion that **well operating data** alone is sufficient for developing models that can accurately monitor the condition of choke valves with acceptable degree of accuracy. The results have shown significant implications for enhancing the effectiveness and timeliness of Cv monitoring methods.

In this application, the predicted value is a numeric value, which means that a regression algorithm is needed to accurately predict this value. Regression algorithms are used to predict a continuous numeric value, rather than a discrete categorical value.

Because the parameters in this application are known via domain knowledge, there is no need for dimension reduction techniques such as Principal Component Analysis (PCA) or feature selection. Instead, the label (Cv) is provided via equation, and the predicting parameter is numeric *(Almaliki, Z. A., 2022)*.

To find the best fit for this specific model, we are further exploring and comparing five different supervised regression algorithms: Decision Tree, Linear Regression, Random Forest, Gradient Boosting Tree, and Neural Network. Each of these algorithms has unique strengths and weaknesses, and by evaluating them all, we can determine which one is the best fit for this specific application. For example, Decision Trees are simple and easy to interpret *(Chauhan, N. S., 2020)*, while Random Forests are known for their ability to handle large datasets and avoid overfitting. Linear Regression is a popular choice for its simplicity and easy interpretability, while Gradient Boosting Tree is known for its excellent predictive performance. Finally, Neural Networks are known for their ability to model complex relationships between variables. By comparing and evaluating these different algorithms, we can find the best fit for our specific application and ensure accurate predictions of the numeric value for Cv.

The ratios used to split the data into training and testing sets can have a significant impact on the model's performance. There are two key factors to consider when determining the appropriate split: the variance of parameter estimates and the variance of performance statistics. Ideally, the data should be partitioned in a way that prevents either type of variance from being excessive.

In our study, we chose the split ratios of 70-15-15 and 80-10-10 as we had sufficient data and not too many hyperparameters to tune. Our dataset was of a moderate size, and our model was of moderate complexity with a manageable number of hyperparameters. Therefore, we decided that a smaller validation set would be sufficient for evaluating the model's performance. Additionally, we took care to ensure that our split was random and representative of the dataset to avoid any bias in our results.

Model Performance Evaluation Criteria

Among the various evaluation metrics, RASE and RSquare are two primary metrics used to assess the performance of models in this study *(Z., & posts by Zach, V. A., 2021)*.

- RASE measures the average distance between the predicted value and the actual value of the regression model. A lower RASE indicates that the model can predict the outcome variable more accurately.
- On the other hand, RSquare measures how well the predictor variables explain the variation in the response variable. A higher RSquare value indicates that the model can better represent the variability of the response variable and has a better fit to the data.

In the present study, the initial dataset that was used for model training spanned from the time when the well was first operated until 8 August 2022. To obtain the optimal predictive models, we split the dataset into **training, validation**, and **testing** sets using the 70-15-15 or 80-10-10 ratios. This step is crucial as it ensures that the model is trained and validated on different datasets, which helps to prevent overfitting and improve generalization.

To identify the best performing model, we process each model with the **training** data and determine the model complexity with the **validation** data. We then utilize the statistics for the test set to evaluate the model's performance. The best performing model among those constructed has the lowest RASE and the highest RSquare. Utilizing both the RASE and RSquare values is crucial since each statistic provides different insights into the model's performance.

Following the **training** and **validation** of the models, we conducted further evaluations to verify their performance. Specifically, we used newly introduced data that spanned the period from 9 August 2022 to March 2023 to test the performance of our models. This additional step of testing the model on unseen data is crucial as it provides a more realistic evaluation of the model's ability to generalize and make accurate predictions on new data.

By using this approach, we can evaluate the predictive performance of our models under real-setting conditions and determine how well they can generalize to new data. This approach also allows us to assess the robustness of our models and identify any potential weaknesses or areas for improvement. In summary, by splitting the data and testing the models on newly introduced data, we can ensure the accuracy and generalizability of our models, which is essential for practical applications.

By tuning the hyperparameters, we can improve the model's accuracy, reduce overfitting, and increase the model's robustness to unseen data. Hyperparameter tuning is an important step in building a machine learning model, as it can significantly impact the model's performance and ability to generalize to new data. This study utilized the random search method for hyper-parameter tuning, which involved manually adjusting each parameter, recording the outcome of each combination, and ultimately selecting the model with the optimal performance. The model is run 30 times with different hyperparameters (Refer to detail in Appendix C), and the table displays how each model performs on three different sets: train, validate, and test.

Table 3 Final selected models

Well	BK-S2-K	BK-S3-N	BK-S4-J
Split ratio	70:15:15	80:10:10	80:10:10
Method	Bootstrap Forest	Bootstrap Forest	Bootstrap Forest
Number of Trees	70	40	40
No. of Terms Sampled per Split	2	3	3
Bootstrap Sample Rate	1	3	4
Minimum Splits per Tree	10	10	10
Maximum Splits per Tree	750	2000	2000
Minimum Size Split	6	8	9
Rsquare	0.953	0.884	0.932
RASE	1.907	6.767	3.793

Table 3 provides information on the model of three representative wells, all of which use the Bootstrap Forest method for regression analysis. We proceeded to further visualize the results using a bivariate fit between the predicted and actual values. Based on the bivariate fit, the performance of the bootstrap forest model is in satisfaction level.

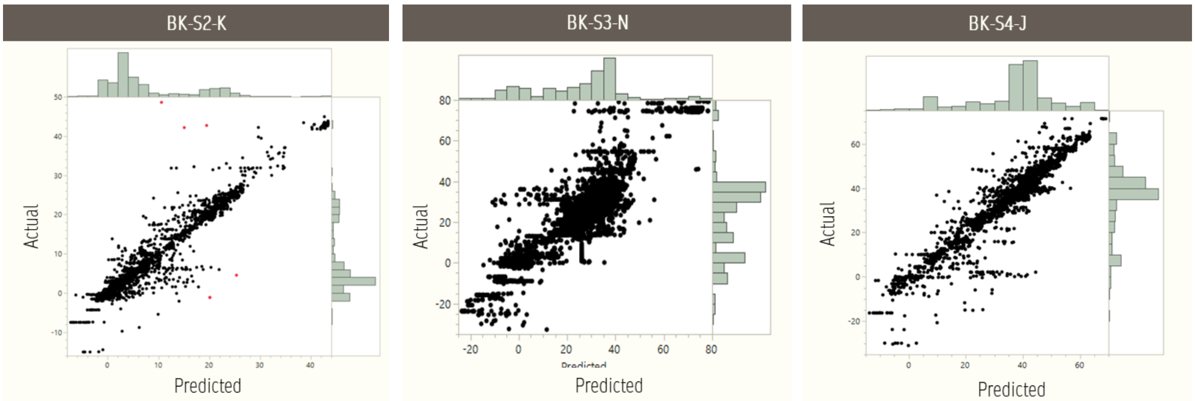


Figure 10 bivariate fit between predicted and actual Cv: 10(a) Left for BK-S2-K; (b) Middle for BK-S3-N; (c) Right for BK-S4-J

Figures 10, 11, and 12 depict bivariate plots of the actual and predicted values of the flow coefficient (Cv) for three wells **BK-S2-K**, **BK-S3-N**, and **BK-S4-J**, in sequence. The points that deviate significantly from the expected values are represented in red color. These points correspond to instances where the model exhibited poor performance in accurately predicting the target variable. The figures reveal that **BK-S3-N** exhibits relatively lower accuracy compared to the other wells. This could be attributed to the fact that **BK-S3-N** is the youngest well and has been operational for a shorter duration than the other selected wells, resulting in a smaller number of data points used for the training process. While **BK-S2-K** and **BK-S4-J** have approximately 8000-9000 data points, **BK-S3-N** has only around 5000 data points, which can contribute to lower performance. However, other factors such as operator behavior and well properties may also affect model performance.

We utilized this visualization to identify specific areas or patterns where the model was particularly ineffective. This analysis will be useful in the subsequent section where we will delve deeper into the underlying reasons for the poor performance of the model and attempt to identify potential solutions to improve its accuracy.

Analysis: when does the model poorly perform?

This section aims to identify the limitation of the method in this study for awareness and proper counter measure to be implement during application. To illustrate, Figure 13 displays the distribution histogram of predictive parameters for two wells with poorly performing cases marked as red points for further investigation. The figure reveals two distinct reasons contributing to the subpar performance of the model. In the first instance, shown in the left image, erroneous data is observed, where the flowline pressure is recorded as zero, which is physically impossible. Therefore, in such cases, the data point should be screened out, highlighting the need for data cleaning. The second instance, depicted in the right image, occurs when the flowline pressure exceeds normal operating conditions, which is credible but uncommon. Consequently, the model may not have been adequately trained to handle such instances, exposing its limitations in dealing with unfamiliar data.

Models are trained on a set of data that is assumed to represent the typical operating conditions, and the performance of the model is measured based on its ability to make accurate predictions on this data. However, when the model is presented with data that is outside of its normal operating conditions, it may not perform as well. This is because the model has not been trained on this type of data and therefore doesn't have the necessary information to make accurate predictions.

We can observe that there are outlier points that fall outside of the normal operating conditions. These points represent data that is unusual or unexpected, and therefore the model may struggle to make accurate predictions on these points. Figure 14 demonstrates that in certain instances, the introduction of dynamic well test data and operating data into the model may also result in the inclusion of erroneous data, such as those depicted in row 4580 and 4581 where the wellhead pressure is recorded as 0, which is not a feasible value. As such, it is imperative that these inaccurate data points be identified and removed from the model to improve its accuracy and reliability.

As a result, it's important to understand the limitations of our model and to recognize when it may not perform as well. By being aware of these limitations, we can take steps to improve the performance of our model or to adjust our expectations for its performance when operating outside of the normal operating conditions.

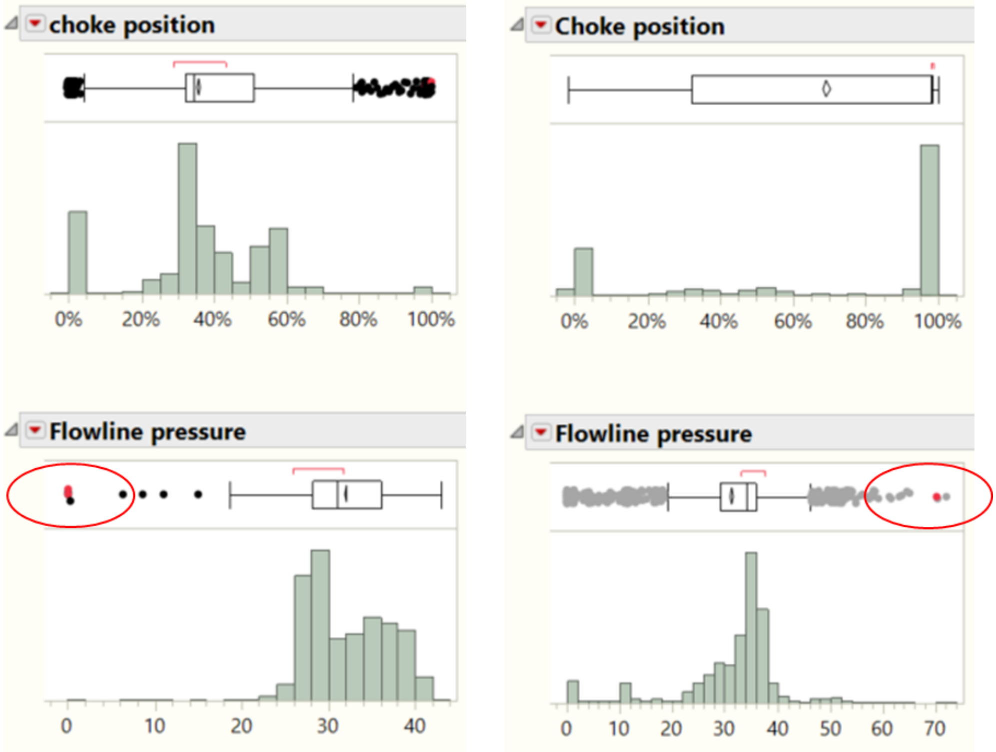


Figure 11 Analyzing the distribution of data to identify instances with subpar performance.

19/0 Cols	Date	Condensate flow rate	Gas flow rate	Water flow rate	Test pressure	Test sep temp	Wellhead pressure	Flowline temp	choke position	Flowline pressure
	15Mar2023	1161	15.1	1659	42.7	83.2	105	92.5	100.00%	42.9
	26Jan2014	-1	0	0	20.5	17.5	0	19.3	-0.51%	0.032
4572	13Jan2019 12:00:00 AM	•	•	•	•	•	21.96	27.23	0.35%	21.98
4573	14Jan2019 12:00:00 AM	•	•	•	•	•	22.25	27.01	0.35%	22.25
4574	15Jan2019 12:00:00 AM	•	•	•	•	•	22.78	26.79	0.35%	22.75
4575	16Jan2019 12:00:00 AM	•	•	•	•	•	23.36	27.07	0.35%	23.39
4576	17Jan2019 12:00:00 AM	•	•	•	•	•	24.13	27.17	0.35%	24.16
4577	18Jan2019 12:00:00 AM	•	•	•	•	•	24.80	27.04	0.32%	24.74
4578	19Jan2019 12:00:00 AM	•	•	•	•	•	24.51	27.43	0.35%	24.51
4579	20Jan2019 12:00:00 AM	•	•	•	•	•	25.28	26.85	0.34%	25.31
4580	21Jan2019 12:00:00 AM	•	•	•	•	•	0.00	26.75	0.38%	0.03
4581	22Jan2019 12:00:00 AM	•	•	•	•	•	0.00	26.88	0.35%	0.03
4582	31Jan2019 12:00:00 AM	•	•	•	•	•	26.91	26.53	1.22%	26.94

Figure 12 Example of data at a point where the model performs poorly

Test performance of the model on new data

To validate the effectiveness and reliability of our developed model, we assess its performance by applying it to new data from a more recent period, beyond the original data collection period that was used to train the model. This approach of testing the model on a time window outside the original training data allows us to evaluate its generalizability and ability to adapt to new data. By doing so, we can ensure that our model is not only accurate and effective for the original data, but also remains robust and reliable in predicting outcomes for new data from a more recent period. Table 4 provides a summary of the data utilized for training purposes for each well, with the data collection period. On the other hand, Table 5 presents a summary of newly introduced data for each well, encompassing the time spanning till March 2023. This additional data provides a more recent and updated representation of the well's condition.

Table 4 Data used for model training

Well	Static well test	Dynamic well test	Well operating	Data collected period
BK-S2-K	45	5,300	3,060	1 Jan 14 – 8 Aug 22
BK-S3-N	58	3,304	2,368	5 Jan 16 – 8 Aug 22
BK-S4-J	95	6,306	2,723	21 Jan 16 – 8 Aug 22

Table 5 New data used for model test

Well	Static well test	Dynamic well test	Well operating	Data collected period
BK-S2-K	9	366	217	9 Aug 22 – 22 Mar 23
BK-S3-N	7	267	226	9 Aug 22 – 22 Mar 23
BK-S4-J	19	226	678	9 Aug 22 – 22 Mar 23

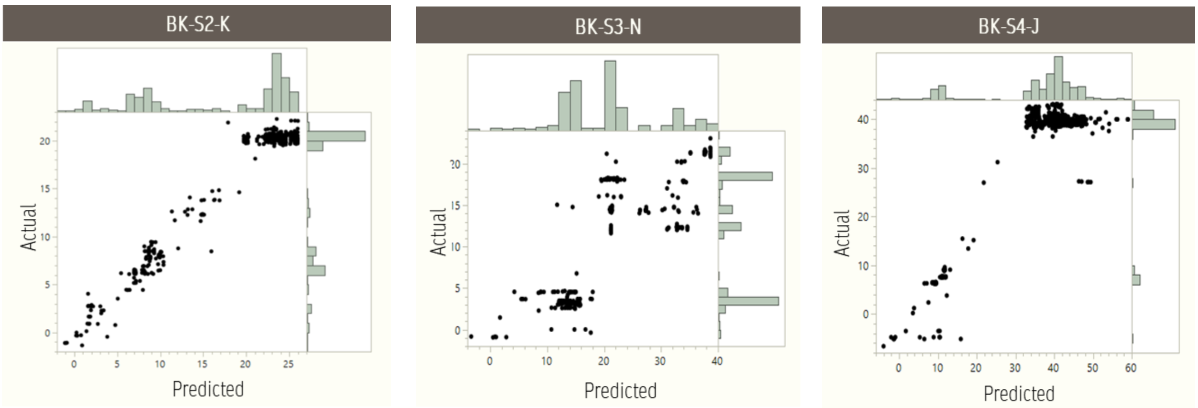


Figure 13 Bivariate fit between predicted and actual Cv on test data (a) Left for BK-S2-K; (b) Middle for BK-S3-N; (c) Right for BK-S4-J

Table 6 Evaluation of the model's performance on new data

Well	RSquare	RASE	N
BK-S2-K	0.995	0.506	592
BK-S3-N	0.991	0.650	500
BK-S4-J	0.999	0.319	919

The performance of the model for **BK-S2-K**, **BK-S3-N**, and **BK-S4-J** on a new dataset not used in training is presented in Figure 15, 16, 17 and Table 6. **BK-S2-K** and **BK-S4-J** exhibit a similar trend, indicating that the model's accuracy is relatively good even with new data, as the relationship between predicted and actual values is clear. However, we noticed that the model tends to slightly over-predict at higher Cv values. Conversely, the bivariate plot for **BK-S3-N** indicates poorer performance compared to the other two wells overall. The table confirms this observation, showing that **BK-S3-N** has the lowest R-squared value and the highest RASE.

Test performance of the model on the past event

To determine the speed at which the model can respond to a specific incident that occurred in the past, we applied our model to the case at **BK-S2-K** where a valve body leak occurred on 8 June 2015. The aim was to observe how early the model could detect the anomaly prior to the incident. Figure 18 displays the model's predicted values at each time stamp from January 2015 to June 2015, which encompasses the date when the incident occurred. The first instance where the predicted flow coefficient showed an anomaly above the set threshold was in January 2015, which was 5 months prior to the incident. This highlights the effectiveness of the model in detecting potential anomalies well in advance, which can aid in preventing incidents before they occur.

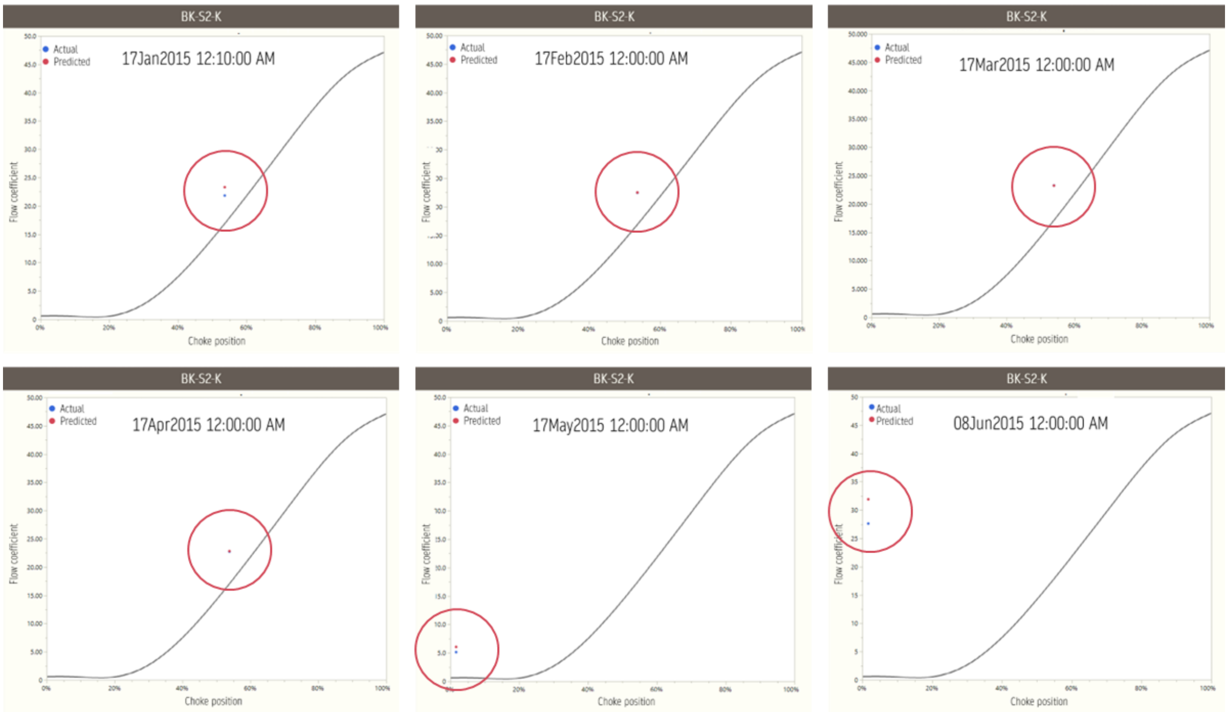


Figure 14 Predicted flow coefficient over time at BK-S2-K on the event of valve body leak

CONCLUSIONS

In conclusion, this study addresses the significant challenge of sand production in the oil and gas industry, particularly regarding the impact on choke valve operation and safety. The proposed method for monitoring the internal condition of choke valves utilizing well test data and continuous operating data is a promising approach to predicting sand erosion and preventing potential harm to plant integrity and personal safety.

The Extended Cv monitoring method is a significant improvement over the well test-based model, as it is more efficient and utilizes continuously measured parameters from **well operating data**. The study compares and evaluates different machine learning algorithms and data split ratios, providing a useful reference for future studies to determine the best-fit model for each well.

The Bootstrap Forest algorithm is identified as the best approach for all three representative wells from the Greater Bongkot South asset, with the lowest RASE and the highest Rsquare, demonstrating the potential for expansion to other wells.

However, it is crucial to exercise caution when introducing **well operating data** and **dynamic well test data** to avoid including invalid data into the model. Furthermore, the method presented in this study has the potential for application in other industries facing similar challenges, indicating the importance of continued research and development in this area.

Overall, this study makes a valuable contribution to the development of methods for monitoring and predicting the internal condition of choke valves, improving the efficiency and safety of choke valve operation, and reducing the impact of sand production in the oil and gas industry. Future studies can build upon the findings presented here, leading to further advancements in the field of sand erosion prediction and prevention.

ACKNOWLEDGEMENTS

We express our gratitude and appreciation to all the individuals and teams who have contributed to the success of this research project. Their unwavering support, dedication, and cooperation have been invaluable in making this endeavor a reality. First and foremost, we extend our sincerest thanks to the Maintenance Team for their tireless efforts in ensuring the smooth functioning and operation of the equipment and facilities. Their prompt response to provide the maintenance record of choke valves and share field experience instrumental in the progress of this study. We are immensely grateful to the Reservoir Team for their assistance in providing data and information related to reservoir characteristics and performance. Their expertise and insights have significantly pivotal in drawing meaningful conclusions. Our sincere appreciation goes out to the Production Team for their support in furnishing essential data related to production processes and operations. Their willingness to collaborate and share valuable insights have been crucial in shaping the outcomes of this study. We also extend our heartfelt thanks to the Inspection Team for their meticulous attention to detail and commitment to quality assurance. Their assistance in ensuring accuracy and reliability in data collection and analysis has been of great significance to the overall reliability of this research. Additionally, we would like to acknowledge the guidance and academic support provided by associate professor Kam Tin Seong from Singapore Management University. His wisdom and expertise have been instrumental in directing the course of this research and shaping it into its final form. The successful completion of this research project would not have been possible without the commitment and collaboration of each team and individual mentioned above. Their collective efforts have not only enriched this paper but have also contributed to the advancement of knowledge in our field.

REFERENCES

Almaliki, Z. A. (2022, October 21). Do you know how to choose the right machine learning algorithm among 7 different types? Medium. Retrieved December 8, 2022, from https://towardsdatascience.com/do-you-know-how-to-choose-the-right-machine-learning-algorithm-among-7-different-types-295d0b0c7f60

Animah, I., & Shafiee, M. (2018, May). Condition assessment, remaining useful life prediction and life extension decision making for offshore oil and gas assets. Journal of Loss Prevention in the Process Industries, 53, 17–28. https://doi.org/10.1016/j.jlp.2017.04.030

Atif Qureshi, M., Miralles-Pechuán, L., Payne, J., O'Malley, R., & Namee, B. M. (2020). Valve Health Identification Using Sensors and Machine Learning Methods. Communications in Computer and Information Science, 45–60. https://doi.org/10.1007/978-3-030-66770-2_4

Baheti. (2022, October 21). Train Test Validation Split: How To & Best Practices [2022]. Train Test Validation Split: How to & Best Practices [2022]. Retrieved December 8, 2022, from https://www.v7labs.com/blog/train-validation-test-set#h2

Chauhan, N. S. (2020 C.E., February 9). Decision Tree Algorithm, Explained - KDnuggets. KDnuggets. Retrieved December 8, 2022, from https://www.kdnuggets.com/decision-tree-algorithm-explained.html

Date. (2021, June 8). An Overview of Generalized Linear Regression Models. Time Series Analysis, Regression and Forecasting. Retrieved December 8, 2022, from https://timeseriesreasoning.com/contents/generalized-linear-regression-models/

Gopaluni, R. B., Tulsyan, A., Chachuat, B., Huang, B., Lee, J. M., Amjad, F., Damarla, S. K., Kim, J. W., & Lawrence, N. P. (2020). Modern Machine Learning Tools for Monitoring and Control of Industrial Processes: A Survey. IFAC-PapersOnLine, 53(2), 218–229. https://doi.org/10.1016/j.ifacol.2020.12.126

Kudelina, K., Vaimann, T., Asad, B., Rassõlkin, A., Kallaste, A., & Demidova, G. (2021, March 19). Trends and Challenges in Intelligent Condition Monitoring of Electrical Machines Using Machine Learning. Applied Sciences, 11(6), 2761. https://doi.org/10.3390/app11062761

Nguyen, Q. H., Ly, H. B., Ho, L. S., Al-Ansari, N., Le, H. V., Tran, V. Q., Prakash, I., & Pham, B. T. (2021, February 5). Influence of Data Splitting on Performance of Machine Learning Models in Prediction of Shear Strength of Soil. Mathematical Problems in Engineering, 2021, 1–15. https://doi.org/10.1155/2021/4832864

Park, M. (n.d.). (2020, Jan) A Monitoring System for Machine Learning Models in a Large-Scale Context | TU Delft Repositories. A Monitoring System for Machine Learning Models in a Large-Scale Context | TU Delft Repositories. Retrieved December 8, 2022, from https://repository.tudelft.nl/islandora/object/uuid%3A42f9cb1d-18fa-4dd0-9436-39d4d202c2e3

Saeid, N. H. (2018, December 27). Numerical predictions of sand erosion in a choke valve. Journal of Mechanical Engineering and Sciences, 12(4), 3988–4000. https://doi.org/10.15282/jmes.12.4.2018.01.0347

Sydorenko. (2021, May 4). How to Choose a Machine Learning Algorithm. How to Choose a Machine Learning Algorithm: A Simple Step-By-Step Guide. Retrieved December 8, 2022, from https://labelyourdata.com/articles/how-to-choose-a-machine-learning-algorithm

Wong, T. K., Mun, H. K., Phang, S. K., Lum, K. L., & Tan, W. Q. (2021). Real-time Machine Health Monitoring System using Machine Learning with IoT Technology. MATEC Web of Conferences, 335, 02005. https://doi.org/10.1051/matecconf/202133502005

What is Valve Flow Coefficient (Cv)? | Kimray. (n.d.). What Is Valve Flow Coefficient (Cv)? | Kimray. Retrieved December 8, 2022, from https://kimray.com/training/what-valve-flow-coefficient-cv

Z., & posts by Zach, V. A. (2021, June 22). RMSE vs. R-Squared: Which Metric Should You Use? Statology. Retrieved December 8, 2022, from https://www.statology.org/rmse-vs-r-squared/

APPENDIX

Appendix A Model screening

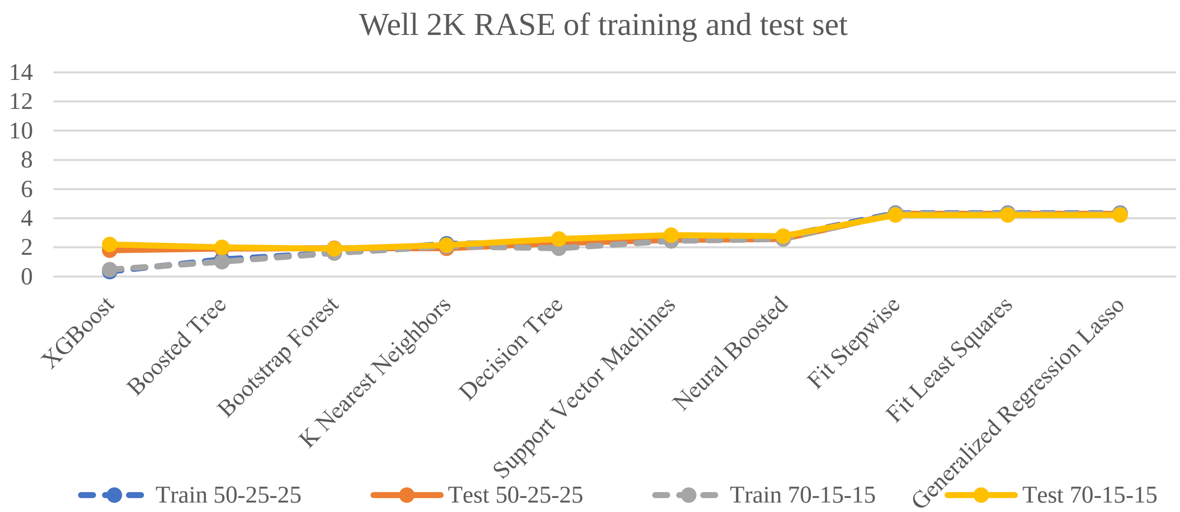


Figure 15 Model screening of well 2K

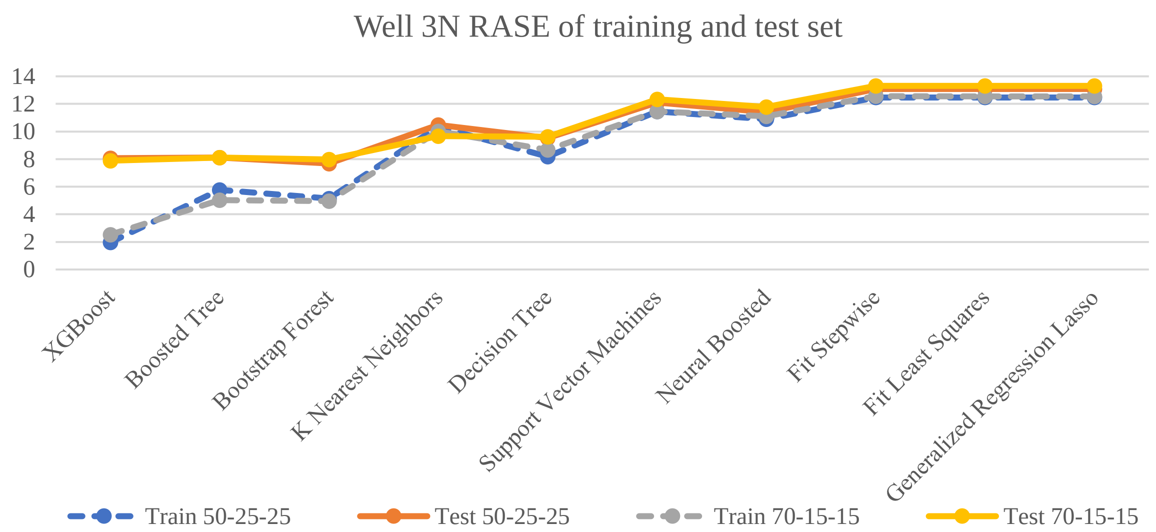


Figure 16 Model screening of well 3N

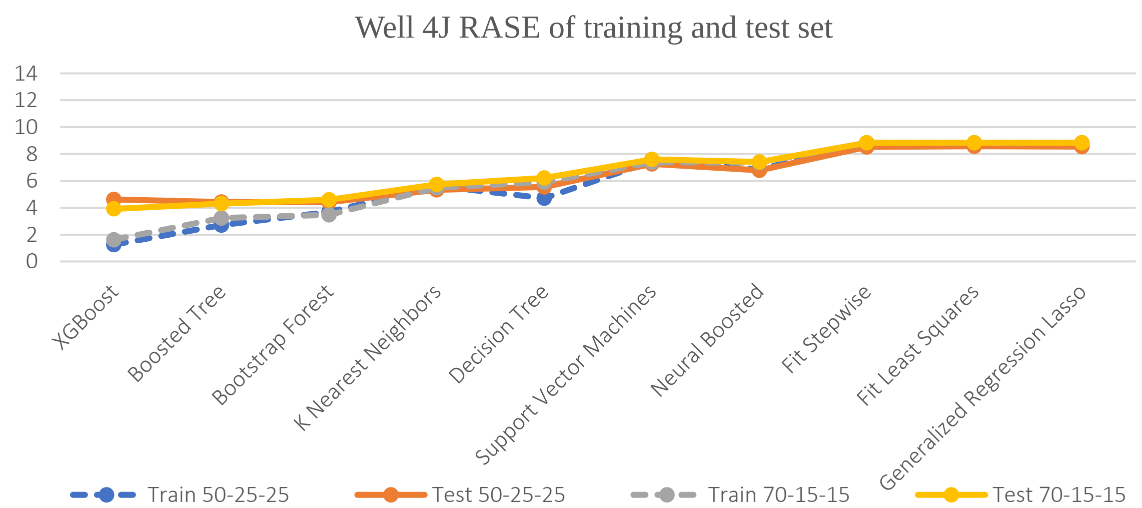


Figure 17 Model screening of well 4J

Appendix B Well historical data

1) BK-S2-K

BK-S2-K is the oldest well of the three and has been in production since 2013. The well witnessed a gas external leak on June 8, 2015, due to valve body damage. This event allows us to study and collect data for the entire life cycle of the choke valve, from when it was newly made until erosion began and the trim reached the end of its life, and ultimately to the top event of gas external leaking. This information is presented in Figures 24 and 25. The data for this well starts from January 1, 2014, before the occurrence of erosion, and extends to the present day. We used a dataset of 8,405 samples to develop the model.

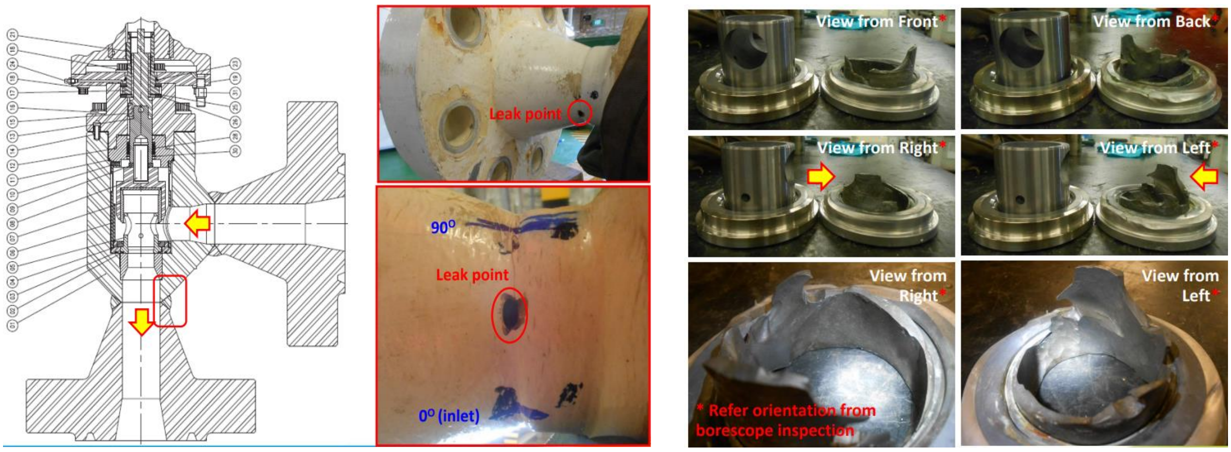


Figure 18 Well 2K choke valve body damage condition

The well was classified as a high-sand operating well between 2014 and June 2015. However, after the valve body damage incident, a zone change was made, which resulted in a change in the well's characteristics and a decrease in sand production. **BK-S2-K** well provides a unique opportunity to study the life cycle of a choke valve and its degradation over time, and the data collected from this well will be crucial in developing and testing models to predict the performance of similar valves in the future.

2) BK-S3-N

The well **BK-S3-N** has been in production with a moderate sand rate since 2015, and two maintenance documents exist detailing the inspection of internal valve parts in March 2016 and May 2018. In March 2016, the trim's condition was deemed to have reached its end of life, and this point is used as a reference for the Cv deviation rationale for the Remaining Useful Life (RUL) *(Park, M. (n.d.), 2020)*. In May 2018, the valve body began to degrade, indicating that the trim was used beyond its expiration date. **Error! Reference source not found.** shows that the seal surface surrounding the outflow port eroded through the valve body wall.

Figure 19 Well 3N choke valve trim damage condition

This well's data began on January 5, 2016, and it continues to operate until this day. The original database of 5,730 samples from **BK-S3-N** was used, and after removing invalid datapoints, 5,724 samples remained to build the model. This well was classified as a medium-sand operating well.

3) BK-S4-J

This particular well has been operating since 2014 and has undergone maintenance six times. The extent of erosion that has taken place in this well is considered to be severe. Based on maintenance records, it was found that the 4J choke valve trim had sustained damage in November 2017, April 2018, February 2019, December 2019, February 2020, and again in December 2020.

Photos of the trim condition found are provided in Figure 30 and Figure 31. Figure 30 shows that the flow collar, nozzle, and seat have been severely damaged. After the replacement of the trim, it was discovered that the new trim had eroded again in February of the following year, which is less than a year of operation. In September of the same year, the same type of erosion was observed on the valve body, as depicted in Figure 31.

Figure 20 Trim condition of choke valve from well 4J on April 2018

Figure 21 Trim condition of choke valve from well 4J on September 2019

The production data for **BK-S4-J** spans over a period of six and a half years, from January 21, 2016, to August 11, 2022, providing a substantial sample size of 9,124 data points for model building.