

Article

A Construction-Phase Reliability Framework for Hard Rock TBM Penetration Rate Prediction Under Delayed UCS Information

Nantapol Monthanopparat *  and Tawatchai Tanchaisawat

Department of Civil Engineering, Faculty of Engineering, Chiang Mai University, Chiang Mai 50200, Thailand; tawatchai@eng.cmu.ac.th

* Correspondence: nantapol_mon@cmu.ac.th; Tel.: +66-84-041-9008

Abstract

Reliable construction-phase prediction of hard rock tunnel boring machine (TBM) rate of penetration (ROP) remains difficult because ground–machine interaction changes along the alignment and uniaxial compressive strength (UCS) is often incomplete or delayed at ring scale. This study proposes a construction-phase reliability framework that integrates sequence deep learning, inverse-distance-weighted UCS completion, chronological rolling evaluation, PassRate monitoring, and performance-triggered updating. The framework was developed from a granite-dominated TBM drive in northern Thailand and evaluated under a delayed-UCS information policy. In the Phase-2 forward deployment-style evaluation, the selected gated recurrent unit (GRU) model achieved a root mean square error (RMSE) of 0.1639 m/h, a mean absolute error (MAE) of 0.1186 m/h, and 62.63% within a symmetric $\pm 10\%$ accuracy band over 990 evaluated rings. Direct static application of representative theoretical and empirical models produced substantially lower within-band performance of 11.92–20.71%. One early reliability trigger occurred at Ring 3409, after which UCS updating, retraining, and redeployment restored the monitoring process without further intervention triggers. The results show that construction-phase TBM prediction should be managed as an auditable reliability workflow with explicit information boundaries, rather than as a single static accuracy score.

Keywords: tunnel boring machine; penetration rate; hard rock tunnelling; sequence deep learning; uniaxial compressive strength; information leakage; reliability monitoring; model updating

1. Introduction

Tunnel boring machine (TBM) performance prediction is required for construction planning, production control, resource allocation, and risk management in mechanized tunnelling. In this study, the construction phase is defined as the excavation period after the TBM has begun producing ring-scale construction records and before the analysed drive is completed. It is distinct from pre-excavation forecasting, where future operating records are unavailable, and from completed-project back-analysis, where the full geological and operational database is already known. Among the main performance indicators, rate of penetration (ROP) provides a direct measure of excavation productivity during boring and is strongly related to schedule and cost control. In hard rock tunnelling, ROP is governed by nonlinear interaction among intact rock strength, rock-mass condition, cutter–rock interaction, machine thrust, cutterhead torque, rotational speed, and operator response. Reliable construction-phase prediction can therefore support earlier recognition



Academic Editor: Yunfeng Zhang

Received: 22 May 2026

Revised: 17 June 2026

Accepted: 24 June 2026

Published: 26 June 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

of productivity loss, more transparent updating of geotechnical information, and better short-term planning of excavation resources.

Conventional theoretical and empirical TBM penetration-rate models have been widely used for design-stage and planning-stage estimation [1–8]. Theoretical models provide mechanistic interpretation and can support machine-design decisions, whereas empirical models can incorporate field-observed rock mass and machine parameters. However, their direct use for construction-phase ring-scale prediction may be limited because they are commonly static, site-specific, or dependent on geotechnical inputs that are not continuously available during excavation. In the Mae Taeng–Mae Ngad–Mae Kuang tunnel case study, eight representative conventional models produced RMSE values ranging from 0.356 to 0.893 m/h when evaluated against observed ring-scale penetration rates, as summarized in Table 1 [9]. These error levels show that direct transfer of conventional models may be inadequate for operational ring-scale decision support.

Table 1. Summary of prediction performance of existing penetration-rate models in a granite TBM drive in northern Thailand.

Model	Author or Source Model	RMSE ¹ All RMT ² (m/h)	RMSE ¹ RMT2 ³ (m/h)	RMSE ¹ RMT2F/2S ⁴ (m/h)
CSM	Rostami	0.499	0.356	0.743
MCSM	Yagiz	0.690	0.764	0.460
Yagiz Model	Yagiz	0.438	0.430	0.457
NTNU	Bruland	0.816	0.787	0.885
Modified NTNU	Macias	0.809	0.787	0.861
Gehring	Gehring	0.629	0.481	0.893
Alpine	Wilfing	0.514	0.390	0.736
Farrokh	Farrokh et al.	0.444	0.422	0.493

¹ RMSE—root mean square error, ² RMT—rock mass type, ³ RMT2—massive granite, ⁴ RMT2F/2S—highly fractured granite/faulted granite.

Machine learning and deep learning methods have improved the ability to model non-linear TBM performance from construction records. Sequence models are particularly relevant because successive rings contain local excavation history and ground–machine interaction patterns. Recent TBM studies have used long short-term memory (LSTM), gated recurrent unit (GRU), temporal convolutional network (TCN), hybrid convolutional-recurrent networks, and other real-time or rolling-window learning strategies for advance-rate or penetration-rate prediction [10–20]. These studies demonstrate the value of construction big data and sequential learning. However, most focus primarily on prediction accuracy or architecture comparison, whereas fewer studies define an auditable construction-phase reliability process that couples delayed geotechnical information, threshold calibration, trigger logic, lag handling, and redeployment decisions.

Despite advances in data-driven prediction, three barriers remain before such models can be used as construction-phase decision-support tools. First, many model evaluations do not fully preserve construction chronology; random splitting or completed-project training can mix future and past construction states, causing optimistic assessment under non-stationary conditions. Second, ring-scale UCS is often incomplete or delayed. If measured UCS inside validation or test blocks is used to estimate UCS for those same blocks, information leakage can occur. Third, conventional aggregate metrics such as RMSE and MAE do not specify how reliability should be monitored, when performance should be considered unacceptable, or when the model should be updated.

This paper addresses these barriers by presenting a construction-phase reliability framework that integrates sequence deep learning, explicit UCS information policies, rolling

evaluation, PassRate-based reliability monitoring, and reliability-based model updating. The methodological novelty is not the isolated use of GRU, IDW, or a rolling window. Instead, the novelty lies in the architecture of the workflow: (i) measured UCS availability is constrained by a documented information boundary; (ii) model selection is separated from deployment evaluation; (iii) reliability is monitored as a rolling within-band PassRate rather than only by block-average error; (iv) a trigger converts sustained reliability loss into a documented UCS-update/retraining action; and (v) a forward deployment-style event log records the decision trace. This combination is intended to make data-driven TBM prediction more usable for construction management and more transparent for engineering audit.

The significance of the study is twofold. For the research field, it provides a leakage-controlled protocol for evaluating TBM sequence models under delayed UCS information and links model accuracy to a reliability-monitoring decision rule. For construction practice, it offers a practical way to determine when a locally trained prediction model can continue to support short-term planning and when updated geological information and retraining are required. The framework therefore contributes to safer and more accountable use of machine learning in mechanized tunnelling, while maintaining the role of conventional models for design-stage interpretation and preliminary planning.

Accordingly, the objectives of this study are to: (i) formulate a construction-phase TBM penetration-rate prediction framework that preserves chronological evaluation and auditable information boundaries; (ii) demonstrate a leakage-controlled UCS completion policy under delayed UCS availability; (iii) select a practical monitoring window and PassRate trigger threshold for deployment; and (iv) validate the framework through Phase-2 forward deployment with checkpoint monitoring, trigger logic, lag handling, UCS updating, scheduled retraining, and benchmarking against representative theoretical and empirical models.

2. Materials and Methods

2.1. Case Study, Dataset Structure, and Modelling Variables

The proposed framework was developed using ring-scale TBM construction records from the Mae Taeng–Mae Ngad–Mae Kuang water transfer tunnel project in northern Thailand. The analyzed tunnel section was excavated mainly in granite-dominated hard rock, including massive granite, highly fractured granite, and faulted granite. The complete project database contained 6928 rings over approximately 9.7 km of tunnel alignment; however, the framework was developed and validated using staged ring streams selected for different analytical purposes. The dataset segmentation is shown in Figure 1. Each ring represents a 1.4 m excavation segment and one time-ordered observation row. This ring-scale structure is important because it preserves the construction sequence and allows the prediction workflow to be evaluated under alignment-wise changes in geology and machine operation.

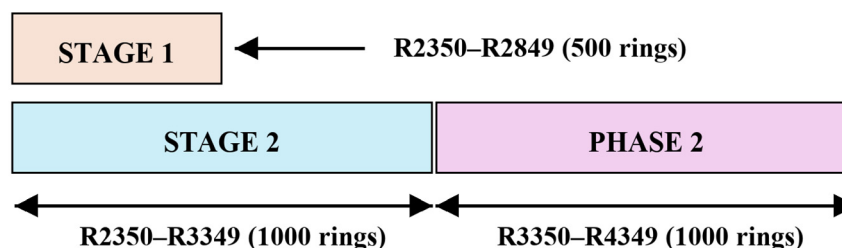


Figure 1. Staged development and deployment structure of the construction-phase TBM penetration-rate prediction framework.

The modelling target was boring-only rate of penetration (ROP), expressed in m/h. Boring-only ROP was used to isolate excavation performance from stoppages, maintenance, logistics, and other non-boring delays. The predictor variables included average thrust, average torque, average cutterhead speed, rock mass type (RMT), and UCS. These variables were selected because they combine machine operation, rock mass description, and intact rock strength in a form available or estimable during construction. Summary statistics for the predictor variables and ROP target across Stage 1, Stage 2, and Phase-2 are provided in Supplementary Table S1. Data quality control was applied before sequence modelling. The checks included ring continuity and ordering, consistency of the boring-only target definition, screening of non-physical operational readings and obvious sensor dropouts, and consistent treatment of missing values. Where preprocessing or scaling was required, the transformation was fitted on the training block only and then applied to validation and test blocks. The ring-scale fields used in the framework are summarized in Table 2.

Table 2. Ring-scale dataset fields used in the construction-phase prediction framework.

Variable	Description	Role in Framework
Ring number	Sequential tunnel ring identifier	Chronological ordering and fold definition
Boring-only ROP	Penetration rate during boring only (m/h)	Prediction target
Average thrust	Mean thrust during boring	Operational predictor
Average torque	Mean cutterhead torque during boring	Operational predictor
Average cutterhead speed	Mean cutterhead rotational speed during boring	Operational predictor
Rock mass type (RMT)	Categorical rock mass descriptor: RMT2, RMT2F, and RMT2S	Geological predictor
UCS	Ring-scale uniaxial compressive strength from measured and/or completed values	Geotechnical predictor

The framework is intended for construction-phase prediction and monitoring rather than for pre-excavation forecasting before machine operation is observed. Operational variables such as thrust, torque, and cutterhead speed are therefore treated as realized boring inputs available from construction records at the ring scale, whereas UCS is treated separately as the information-constrained geotechnical variable.

In the implementation, numerical sequence features were standardized using StandardScaler fitted on the training subset only, and RMT was encoded using OneHotEncoder with handle_unknown = “ignore”. The target history, operational variables, and UCS-related variable were included in the sequential feature matrix, while the one-step-ahead target ring was kept outside the input window to avoid target leakage.

The staged workflow contained three main analytical levels. Stage 1 screened sequence deep-learning model families using a 500-ring dataset. Stage 2 used 1000 consecutive rings from Ring 2350 to Ring 3349 for rolling evaluation under two UCS information conditions. Phase-2 deployment used the next 1000-ring forward stream from Ring 3350 to Ring 4349 to test checkpoint monitoring, trigger logic, UCS updating, and scheduled retraining. This paper focuses on the construction-phase framework and Phase-2 deployment, while retaining the Stage 1 and Stage 2 steps as calibration and model-selection context.

2.2. Overall Construction-Phase Reliability Framework

The overall workflow is presented in Figure 2. The framework begins with ring-scale data preparation and UCS completion. A sequence-to-one prediction model is then trained under chronological evaluation rules. Predictions are evaluated using RMSE, MAE, and a symmetric $\pm 10\%$ within-band indicator. The within-band indicator is converted into PassRate over monitoring windows. During deployment, PassRate is checked at predefined checkpoints; if reliability drops below the selected threshold, the framework triggers updating and redeployment after a practical lag period.

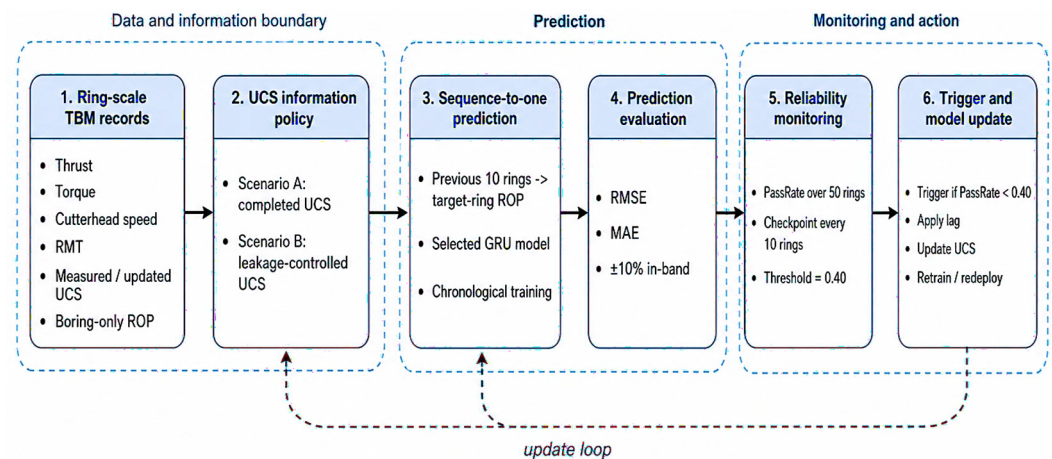


Figure 2. Proposed construction-phase framework for TBM penetration-rate prediction, reliability monitoring, and model updating.

Figure 2 emphasizes that the proposed framework is not a stand-alone neural-network model but a construction-phase procedure that links ring-scale construction records, UCS information policy, sequence deep-learning prediction, reliability evaluation, PassRate monitoring, and performance-triggered updating. This organization directly addresses the need for a reproducible and auditable methodology rather than a single static accuracy result.

2.3. UCS Completion and Information-Boundary Policies

UCS was completed at ring scale using inverse distance weighting (IDW) within a block-model representation [21,22]. Distance was defined along the ring-number axis. Because the analysed tunnel drive was evaluated as a sequential alignment stream, ring-number distance was used as a practical chainage-based distance surrogate for construction-phase UCS completion. The IDW power parameter was set a priori to $p = 2$, corresponding to a conventional inverse-square weighting. It was not tuned by fold, by deployment interval, or by Phase-2 performance. This choice intentionally avoided creating an additional calibration loop that could overfit a particular stream or introduce monitoring-parameter leakage. Sensitivity of downstream prediction to alternative IDW powers remains a topic for future work. The block model provided a practical structure for storing, updating, and extracting UCS estimates along the alignment while maintaining an explicit boundary between available and unavailable information during evaluation.

In the Stage 2 stream, only 91 of 1000 rings had directly measured UCS values, leaving 90.9% of rings without direct measured UCS. The completion procedure was therefore used to create auditable UCS inputs under delayed information availability, while Scenario B tested the model under a past-only UCS knowledge state.

Two UCS information policies were used. Scenario A represents an operational completion condition in which UCS is completed and updated within the block-model representation. Scenario B represents a conservative construction condition intended to

prevent information leakage. In Scenario B, measured UCS inside validation and test blocks are not used to construct UCS inputs for those same blocks. Instead, UCS for evaluation is generated only from information available before the relevant evaluation boundary. The conceptual information boundary is shown in Figure 3, and the corresponding experiment-level UCS rules are summarized in Table 3.

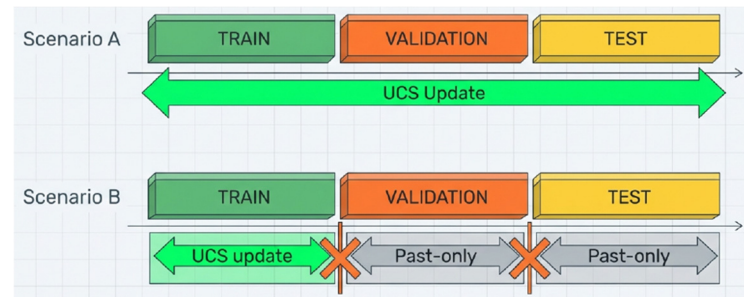


Figure 3. UCS information boundary for Scenario A and Scenario B.

Table 3. UCS handling rules and data sources by experiment.

Experiment	Block	Measured UCS Inside Block	UCS Source	Boundary for UCS
Stage 2—Scenario A	Train	Allowed	Measured and IDW *—completed	Up to end of Train
	Validation	Allowed	Measured and IDW *—completed	Up to/within Validation
	Test	Allowed	Measured and IDW *—completed	Up to/within Test
Stage 2—Scenario B	Train	Allowed	IDW *—completed	Up to end of Train
	Validation	Not allowed	Past-only IDW *	Before Validation start
	Test	Not allowed	Past-only IDW *	Before Test start
Phase 2—Deployment	Train	Allowed	Measured and IDW *—completed	Up to end of Train
	Validation	Allowed	Measured and IDW *—completed	Up to end of Validation
	Test	Not allowed	Past-only IDW	Up to end of Train + Validation

* IDW—inverse distance weighting.

Scenario B therefore represents a past-only UCS knowledge state. This distinction is important because measured UCS values obtained after excavation may be available in a completed project database, but they would not have been available for forward prediction at the time of construction.

Scenario B is central to the framework because UCS measurements can be delayed relative to excavation. Without an explicit information boundary, completed project data can unintentionally leak into forward prediction blocks and overstate deployable performance. Scenario B therefore provides the leakage-controlled condition used to select reliability settings before Phase-2 deployment.

The information boundary defines which UCS measurements are allowed to influence model inputs at the time of prediction. It does not describe the complete information available after excavation, but the restricted information state that would be available during construction-phase deployment.

2.4. Sequence Formulation and Selected Model Configuration

The penetration-rate prediction task was formulated as a sequence-to-one problem, as illustrated in Figure 4. For a target ring r , the model predicted the boring-only ROP one

ring ahead using only the preceding $W = 10$ rings. No operational, geological, RMT, or UCS input from the target ring itself was included in the input window. This formulation allows the model to use local excavation history while avoiding target-ring input leakage.

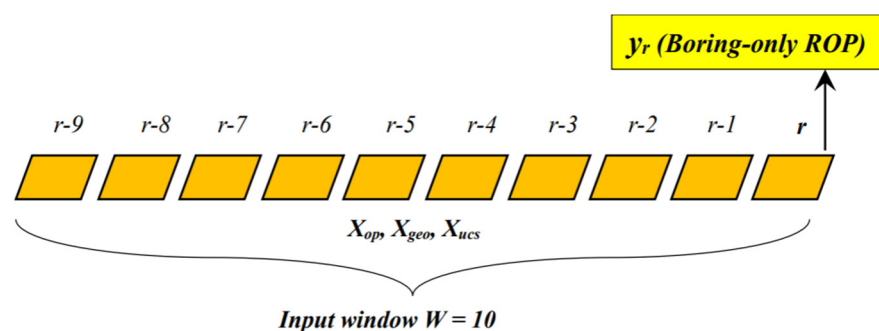


Figure 4. Ring-scale sequence-to-one windowing with $W = 10$ rings.

Four sequence model families were screened in Stage 1: LSTM, GRU, TCN, and Transformer [10–13]. Stage 1 used a 500-ring dataset as an initial sequence-family screen. LSTM and GRU produced the strongest test metrics, TCN remained acceptable for further rolling evaluation, and Transformer produced the weakest Stage-1 test performance; therefore, Transformer was removed after Stage 1. The LSTM, GRU, and TCN families were then carried forward to Stage 2, where TCN was eliminated after rolling evaluation and the final frozen GRU configuration was selected. A concise Stage-1 screening table and the full Stage-1 metrics are provided in Supplementary Table S2.

For auditability, each model run retained the random seed, model family, fixed train-validation-test ring ranges, sequence length, selected hyperparameters, and fold or deployment-loop identifier. The final deployment results were generated with the frozen GRU configuration, so that reported differences are attributable to the UCS information policy, chronological evaluation, and updating workflow rather than post hoc model retuning.

The frozen GRU implementation used an Adam optimizer with learning rate 1×10^{-3} , mean squared error as the training loss, and mean absolute error as a monitored metric. The network consisted of a GRU layer with 32 units, dropout = 0.2, a dense hidden layer with 16 ReLU units, and a single linear output neuron. Training used a maximum of 200 epochs and batch size = 16, with early stopping on validation loss (patience = 15, restore best weights) and ReduceLROnPlateau on validation loss (factor = 0.5, patience = 7, minimum learning rate = 1×10^{-5}).

To further support reproducibility, the implementation set PYTHONHASHSEED = 42, NumPy and TensorFlow random seeds to 42, and TensorFlow deterministic operation flags when available. The run manifest recorded the TensorFlow version, Python version, platform, available devices, deterministic-operation status, input files, loop boundaries, and output file paths.

2.5. Rolling Evaluation Design for Stage 2

Stage 2 rolling evaluation used 1000 consecutive rings from Ring 2350 to Ring 3349. The dataset was divided into four chronological folds, each with fixed 100-ring validation and 100-ring test blocks. The training block expanded forward as construction progressed. This design preserves the chronological structure of construction and provides repeated evaluation across successive tunnel intervals. The fold definitions are summarized in Table 4.

Table 4. Stage 2 rolling evaluation design.

Fold ID	Training Ring Range	Validation Ring Range	Test Ring Range
1	2350–2849 (500 rings)	2850–2949 (100 rings)	2950–3049 (100 rings)
2	2350–2949 (600 rings)	2950–3049 (100 rings)	3050–3149 (100 rings)
3	2350–3049 (700 rings)	3050–3149 (100 rings)	3150–3249 (100 rings)
4	2350–3149 (800 rings)	3150–3249 (100 rings)	3250–3349 (100 rings)

The 1000-ring Stage 2 stream contained RMT2 = 766 rings (76.6%), RMT2F = 223 rings (22.3%), and RMT2S = 11 rings (1.1%). This distribution shows that massive granite dominated the rolling evaluation stream, while fractured and faulted granite conditions were less frequent and should be interpreted with the corresponding class imbalance in mind.

Scenario A was used for operational UCS completion and model selection. Scenario B used the same rolling structure but applied the leakage-controlled UCS rule in Table 3. The purpose of Scenario B was to test whether the selected model remained reliable when future UCS information was excluded from the evaluation horizon. The Scenario B predictions were then used to screen reliability settings for deployment.

2.6. Accuracy Metrics and PassRate Reliability Indicator

Model accuracy was evaluated using root mean square error (RMSE) and mean absolute error (MAE), which are standard regression error metrics [23]. Equation (1) defines RMSE as the square root of the mean squared difference between predicted and measured boring-only ROP values. This metric gives greater weight to larger prediction errors and is therefore useful for identifying large deviations in ring-scale prediction.

$$\text{RMSE} = \sqrt{[(1/n) \sum_{r=1}^n (\hat{y}_r - y_r)^2]} \quad (1)$$

Equation (2) defines MAE as the mean absolute difference between predicted and measured boring-only ROP values. MAE provides a direct average-error measure in m/h and complements RMSE by being less sensitive to isolated large errors.

$$\text{MAE} = (1/n) \sum_{r=1}^n |\hat{y}_r - y_r| \quad (2)$$

In Equations (1) and (2), y_r is the measured boring-only ROP at ring r and $\hat{y}_r$ is the predicted value. For construction-phase reliability monitoring, Equation (3) defines a binary within-band indicator based on a symmetric $\pm 10\%$ accuracy band. This band was not assumed to be a universal TBM standard for all projects. It was selected as a strict project-level operational tolerance for ring-scale monitoring, conceptually aligned with the use of accuracy bands in construction cost and production-control practice [24], and its effect was checked through tolerance sensitivity in Supplementary Table S3.

$$\text{in_band}(r) = 1 \text{ if } |\hat{y}_r - y_r| \leq 0.10|y_r|; \text{ otherwise } 0 \quad (3)$$

Equation (4) defines PassRate as the proportion of evaluated rings whose predictions fall inside the $\pm 10\%$ accuracy band. This converts individual ring-level prediction outcomes into a reliability indicator that can be interpreted over a block or monitoring window.

$$\text{PassRate} = (1/n) \sum_{r=1}^n \text{in_band}(r) \quad (4)$$

During deployment, PassRate was monitored within a rolling window of length w at predefined checkpoints. Equation (5) defines the checkpoint PassRate as the mean of the most recent w binary within-band indicators before or at checkpoint i . This rolling

formulation allows recent reliability degradation to be detected during deployment rather than only after a full test block has ended.

$$\text{PassRate}_w(i) = (1/w) \sum_{k=i-w+1}^i \text{in_band}(k) \quad (5)$$

Three reliability tiers were considered: $\text{PassRate} < 0.40$ was treated as an intervention state, $0.40 \leq \text{PassRate} < 0.50$ as an acceptable/watch state, $0.50 \leq \text{PassRate} < 0.80$ as a target state, and $\text{PassRate} \geq 0.80$ as an excellent state. The 0.40 threshold was not treated as a target accuracy level; it was used to detect sustained degradation under a strict $\pm 10\%$ ring-scale accuracy band. Scenario B reliability screening evaluated candidate monitoring windows from 10 to 100 rings and threshold tiers of 0.40, 0.50, and 0.80. The selected threshold was calibrated only from Scenario B predictions. Phase-2 records were held forward for deployment-style evaluation and were not used to tune the monitoring threshold.

At the selected 50-ring window, the Scenario B threshold workbook showed mean passed-fold rates of 67.16%, 48.53%, and 21.57% for PassRate trigger thresholds of 0.40, 0.50, and 0.80, respectively. A tolerance-band sensitivity check also showed that Scenario B within-band performance changed from 28.25% at $\pm 5\%$ to 50.25% at $\pm 10\%$ and 65.00% at $\pm 15\%$ (Supplementary Table S3). These results confirm that the tolerance band materially affects PassRate values. The $\pm 10\%$ band was therefore retained as a strict central monitoring band, while the 0.40 tier was retained as an intervention threshold rather than a claim of high-precision prediction.

For every prediction row, the in-band indicator was recomputed directly from measured and predicted ROP as $|\hat{y} - y| \leq 0.10|y|$. Checkpoint PassRate was then calculated from the most recent 50 recomputed in-band indicators, rather than from stored model outputs, to keep the reliability trigger auditable.

2.7. Phase-2 Deployment Design

Phase-2 deployment was designed as a forward deployment-style evaluation using historical construction records rather than a retrospective random split or live real-time site trial. The deployment stream covered 1000 consecutive rings from Ring 3350 to Ring 4349. The same sequence-to-one formulation was retained, and the selected GRU configuration defined in Section 2.4 was used as the deployment model. Each deployment loop consisted of an available train/validation basis and a forward 100-ring test block. The monitoring threshold and window length were fixed before Phase-2 evaluation using Scenario B only.

The practical 10-ring lag was defined to represent the construction interval required to collect and check new UCS information, update the block-model UCS field, retrain the selected sequence model, and redeploy the updated prediction file. In the triggered Phase-2 interval, Rings 3410–3419 represented 14.0 m of tunnel excavation and approximately 23.5 h of production time, equivalent to about 10.2 rings/day. This order of magnitude provided a realistic short operational buffer for data updating and model redeployment in the project workflow. The lag calculation is documented in Supplementary Table S8.

Checkpoint monitoring was conducted within each 100-ring forward test block. Because the selected monitoring window was 50 rings, the first checkpoint was evaluated after 50 predicted rings, followed by checkpoints at 10-ring intervals. Thus, checkpoint evaluation was performed at 50, 60, 70, 80, 90, and 100 rings within each deployment block. At each checkpoint, PassRate was calculated over the most recent 50 rings and compared with the selected trigger threshold of 0.40. If the checkpoint PassRate was lower than 0.40, the corresponding ring was defined as a trigger point. A 10-ring practical lag was then applied to represent the time required for UCS updating, model retraining, and redeployment preparation. After the lag interval, the training basis was updated using newly available information, and the next deployment loop was initiated. For auditability, the trigger ring

was defined as the first checkpoint ring where the rolling PassRate fell below the selected threshold. The pre-trigger ring was the checkpoint immediately preceding the trigger event, and the lag interval represented the practical delay required for UCS updating, data consolidation, retraining, and redeployment preparation. The redeployment start ring was then defined as the trigger ring plus the lag length plus one ring.

The Phase-2 loop runner used scheduled retraining every 100 rings when no trigger occurred. When a trigger occurred, the current loop was terminated for reporting, a 10-ring lag was applied, and the next loop started at the redeployment ring. For test-window construction, the last $W = 10$ rows from the train-validation stream were prepended to the test stream so that the first test ring could be predicted with a complete input sequence.

The deployment outputs were exported per loop as prediction records, checkpoint event logs, range summaries, metric summaries, and JSON run manifests. Prediction records included an event tag identifying the pre-trigger checkpoint, trigger ring, and lag-end or end-of-block ring. Metric summaries were reported at the pre-trigger ring, trigger ring when present, and lag-end or end-of-block ring, which enables reconstruction of the operational decision timeline.

To improve reproducibility, the deployment and updating procedure is summarized in Algorithm 1.

Algorithm 1. Phase-2 deployment and updating procedure as follows:

1. Define the available train and validation data for the current deployment loop;
 2. Complete UCS inputs using the information-boundary rule in Table 3;
 3. Train the selected GRU model using the fixed configuration defined in Section 2.4;
 4. Predict boring-only ROP for the next 100-ring forward test block;
 5. Starting after the first 50 evaluated rings, compute PassRate over the most recent 50 rings at 10-ring checkpoint intervals;
 6. If checkpoint PassRate is lower than 0.40, define the corresponding ring as a trigger point;
 7. Apply a 10-ring practical lag to represent UCS updating, model retraining, and redeployment preparation;
 8. Update the training basis using newly available information and begin the next deployment loop;
 9. If no trigger occurs, continue to the end of the 100-ring block and proceed to scheduled retraining for the next loop.
-

The deployment design therefore embeds prediction, reliability monitoring, triggering, lag handling, UCS updating, and scheduled retraining into a repeated construction-phase workflow. All loop boundaries, checkpoint results, trigger points, lag intervals, UCS update boundaries, and redeployment start rings were logged for auditability and are reported in the Section 3. This structure allows the framework to be evaluated not only by overall prediction accuracy, but also by whether it provides a traceable decision process for maintaining model reliability during forward excavation.

3. Results

3.1. Scenario B Rolling Prediction Under Delayed UCS Availability

Scenario B was evaluated after the Stage 2 model-selection process had been completed. The selected setting was the expanding-window rolling protocol with the frozen GRU configuration defined in Section 2.4. The main purpose of Scenario B was to determine

whether the prediction engine remained reliable when the UCS information boundary was closed within validation and test blocks. In other words, measured UCS values inside the evaluation blocks were not used to generate UCS inputs for those same blocks.

The fold-wise measured and predicted boring-only ROP series are shown in Figure 5. Across the four rolling folds, the selected GRU configuration produced a mean RMSE of 0.2368 m/h and a mean MAE of 0.1675 m/h. This result is close to the best Scenario A configuration obtained during model selection, indicating that the leakage-controlled UCS policy did not cause a material collapse in predictive accuracy. This is important for construction-phase deployment because a model that relies on future UCS information cannot be considered operationally valid.

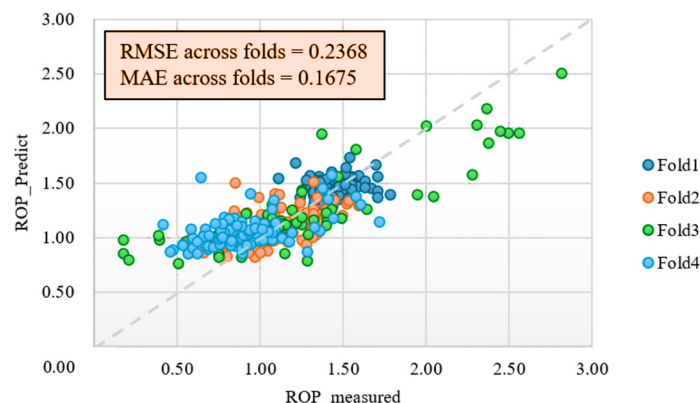


Figure 5. Scenario B fold-wise comparison of measured and predicted boring-only ROP using the selected GRU configuration under the construction-condition UCS policy.

Although the Scenario B error metrics indicate stable predictive behavior, they should not be interpreted as sufficient evidence of operational readiness. RMSE and MAE summarize point error over a block, but they do not show whether recent predictions remain consistently within a construction-tolerance band. Therefore, the Scenario B predictions were further processed through PassRate-based reliability screening.

3.2. Reliability Screening and Deployment-Setting Selection

Scenario B predictions were transformed into reliability indicators by calculating PassRate within a symmetric $\pm 10\%$ band. The screening considered monitoring windows from 10 to 100 rings and threshold tiers of 0.40, 0.50, and 0.80. The results are shown in Figure 6. This screening answers an operational question: if the site monitors the most recent w rings, how often does the model remain above the acceptable, target, or excellent reliability tier? Tolerance sensitivity was also checked at $\pm 5\%$, $\pm 10\%$, and $\pm 15\%$ to confirm that the selected band was strict and that PassRate values should not be interpreted independently of the selected tolerance definition.

Based on this screening, the deployment setting was fixed at a 50-ring monitoring window and a 0.40 PassRate threshold before Phase-2 evaluation. Under this setting, the Scenario B reliability screen produced PassRate = 67.16%. When stricter thresholds were used, PassRate decreased to 48.53% at the 0.50 tier and 21.57% at the 0.80 tier. These values show why the 0.40 tier was selected as an action threshold rather than as a claim of high-precision prediction. The selected setting defines a practical trigger rule for Phase-2 deployment-style evaluation and was not recalibrated using Phase-2 outcomes.

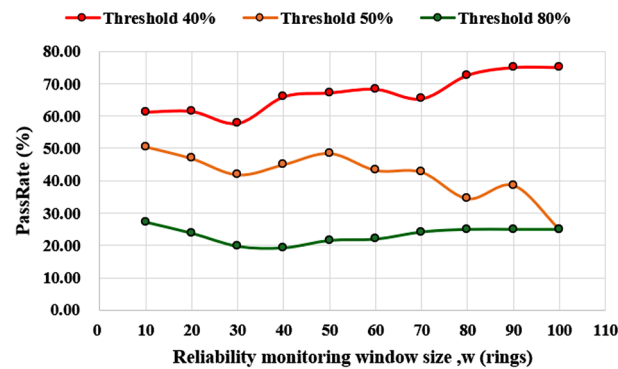


Figure 6. Scenario B reliability screening results: PassRate across monitoring windows $w = 10$ – 100 evaluated under threshold tiers of 0.40, 0.50, and 0.80.

3.3. Phase-2 Deployment Event Log and Loop Execution

The selected framework was then applied to the forward 1000-ring deployment stream from Ring 3350 to Ring 4349. As summarized in the condensed event log in Table 5, the deployment was executed as a sequence of loop-based prediction and updating cycles. Each loop used the available train/validation basis to predict a forward test block, followed by checkpoint monitoring using the selected 50-ring PassRate window and the 0.40 trigger threshold. The full checkpoint-level event log is provided in Supplementary Table S7 to preserve the audit trail without overloading the main text.

Table 5. Condensed Phase-2 deployment event log showing loop-level trigger status, minimum checkpoint PassRate, and deployment action. The full checkpoint-level event log is provided in Supplementary Table S7.

Loop	Range	Min WB50	Trigger	Trigger Ring	Action	End/Trigger PassRate (%)
1	3350–3419	0.36	Yes	3409	UCS update + retrain after 10-ring lag	38.33
2	3420–3519	0.48	No		Scheduled retraining	48.00
3	3520–3619	0.60	No		Scheduled retraining	65.00
4	3620–3719	0.70	No		Scheduled retraining	74.00
5	3720–3819	0.44	No		Scheduled retraining	56.00
6	3820–3919	0.42	No		Scheduled retraining	55.00
7	3920–4019	0.50	No		Scheduled retraining	58.00
8	4020–4119	0.64	No		Scheduled retraining	76.00
9	4120–4219	0.60	No		Scheduled retraining	72.00
10	4220–4319	0.44	No		Scheduled retraining	63.00
11	4320–4349	0.87	No		Phase end	86.67

Loop 1 illustrates the trigger mechanism. The model was trained using Rings 2350–3249 and validated using Rings 3250–3349 before being deployed to predict Rings 3350–3449. At the first checkpoint, corresponding to Ring 3399 after 50 predicted rings, the 50-ring PassRate was 0.44 and therefore remained above the 0.40 trigger threshold. The deployment was allowed to continue. At the second checkpoint, corresponding to Ring 3409 after 60 predicted rings, the 50-ring PassRate decreased to 0.36. Because this value was below the selected threshold, Ring 3409 was identified as the first trigger point. A 10-ring practical lag was then applied from Ring 3410 to Ring 3419 to represent UCS updating,

model retraining, and redeployment preparation. The next deployment loop therefore began from Ring 3420.

After the first trigger event, the redeployed framework maintained checkpoint reliability above the intervention threshold. Loop 2 produced PassRate values between 0.48 and 0.58 across its checkpoints and completed the 100-ring test block without a new trigger. Loops 3 and 4 showed stronger reliability, with checkpoint PassRate values ranging from 0.60 to 0.70 in Loop 3 and from 0.70 to 0.80 in Loop 4. In Loop 4, the checkpoint at Ring 3689 reached the excellent tier with PassRate = 0.80. These results indicate that the update following the first trigger restored the short-term reliability of the deployment stream.

The later loops also remained above the trigger threshold, although some local reductions were observed. Loop 5 produced checkpoint PassRate values from 0.44 to 0.64, while Loop 6 decreased to a minimum of 0.42 at Ring 3909. Although these values approached the intervention threshold, no checkpoint fell below 0.40, and no trigger was activated. Loop 7 recovered to a final checkpoint PassRate of 0.64. Loops 8 and 9 were the most stable portions of the deployment, with PassRate values mostly within the target tier and occasional excellent-tier checkpoints. Loop 8 ended with PassRate = 0.80 at Ring 4119, while Loop 9 reached PassRate = 0.80 at Ring 4189.

Loop 10 showed another temporary reduction in reliability, with PassRate values of 0.46 and 0.44 at intermediate checkpoints. However, the values remained above the 0.40 threshold and recovered to 0.68 by the end of the loop. The final Loop 11 covered the remaining 30 rings from Ring 4320 to Ring 4349 and ended the Phase-2 stream with PassRate = 0.87, classified as excellent. Because the deployment stream ended at Ring 4349, this final value was calculated over the remaining 30-ring segment rather than a full 50-ring monitoring window.

Overall, Table 5 and Supplementary Table S7 demonstrate that the proposed framework does not treat the 1000-ring deployment stream as a single retrospective test set. Instead, it converts reliability loss into traceable construction-phase decisions. The first reliability drop below the 0.40 threshold triggered UCS updating, retraining, and redeployment after a 10-ring practical lag, while later reductions remained above the intervention threshold and were managed through scheduled retraining. The event record provides an auditable connection between checkpoint reliability, trigger decisions, UCS information updates, and redeployment timing. The resulting loop execution forms the operational basis for the Phase-2 prediction performance evaluated in the next section.

Table 5 is condensed to one row per deployment loop because the complete checkpoint-level event log is provided in Supplementary Table S7. This preserves auditability while improving readability of the main manuscript.

3.4. Phase-2 Prediction Performance

The predictive performance of the Phase-2 deployment is summarized in Table 6 and visualized in Figure 7. The loop-wise metrics show that the framework maintained practical prediction accuracy over the forward deployment stream while adapting to changes in rock-mass and operating conditions. Across the 990 evaluated rings, the deployment achieved an overall RMSE of 0.1639 m/h, MAE of 0.1186 m/h, and 62.63% within the symmetric $\pm 10\%$ accuracy band. These values indicate that the framework retained acceptable prediction quality after the leakage-controlled reliability setting from Scenario B was carried forward into forward deployment.

Table 6. Phase-2 loop-wise prediction performance summary using RMSE, MAE, and PassRate within the $\pm 10\%$ accuracy band.

Loop ID	Segment	At End Rings	n.	Avg. ROP Measure (m/h)	RMSE (m/h)	MAE (m/h)	Within (%)	Remarks
1	CP1	3399	50	0.902	0.1697	0.1272	44.00	Trigger Lag 10 rings
	Trigger Ring	3409	60	0.870	0.1712	0.1333	38.33	
	End block	3419	70	-	-	-	-	
2	CP1	3469	50	0.839	0.1263	0.0982	48.00	No Trigger
	End block	3519	100	0.882	0.1388	0.1090	48.00	
3	CP1	3569	50	0.904	0.1267	0.0968	60.00	No Trigger
	End block	3619	100	0.952	0.1185	0.0874	65.00	
4	CP1	3669	50	1.074	0.1037	0.0755	76.00	No Trigger
	End block	3719	100	1.053	0.1314	0.0897	74.00	
5	CP1	3769	50	0.987	0.1976	0.1107	62.00	No Trigger
	End block	3819	100	0.917	0.1899	0.1209	56.00	
6	CP1	3869	50	1.146	0.1279	0.1048	66.00	No Trigger
	End block	3919	100	1.235	0.2087	0.1473	55.00	
7	CP1	3969	50	1.494	0.1929	0.1607	52.00	No Trigger
	End block	4019	100	1.604	0.1862	0.1561	58.00	
8	CP1	4069	50	1.813	0.1804	0.1339	72.00	No Trigger
	End block	4119	100	1.786	0.1712	0.1233	76.00	
9	CP1	4169	50	1.493	0.1373	0.1055	74.00	No Trigger
	End block	4219	100	1.500	0.1779	0.1244	72.00	
10	CP1	4269	50	1.155	0.1502	0.1222	58.00	No Trigger
	End block	4319	100	1.221	0.1413	0.1134	63.00	
11	End block	4349	30	1.403	0.0927	0.0753	86.67	Phase End
Summary			990	1.221	0.1639	0.1186	62.63	

The loop-runner event log identified one trigger event in Loop 1 at Ring 3409, where WB50 decreased to 0.36. The subsequent lag interval covered Rings 3410–3419, and redeployment started at Ring 3420. Later loops completed without additional trigger events under the selected 0.40 threshold and were handled through scheduled retraining.

Loop-level behavior was not constant along the alignment. This variability is expected because the deployment stream contained changing ground conditions and operating regimes. The early part of the deployment showed a reliability reduction that triggered updating in Loop 1, as described in Section 3.3. After the trigger, subsequent loops generally maintained checkpoint PassRate above the 0.40 intervention threshold. RMT composition varied substantially by loop: for example, Loop 1 and Loop 8 contained high proportions of RMT2F, whereas Loops 3, 4, and 9 were dominated by RMT2. Figure 7 therefore overlays loop-wise performance with RMT composition to avoid interpreting metric changes independently of geological context.

The deployment results should be interpreted together with the event log in Table 5. Table 5 shows when the monitoring rule triggered or allowed the model to continue, whereas Table 6 quantifies the predictive performance achieved by each loop after deployment. This distinction is important because the proposed framework is not evaluated only as a retrospective prediction model. It is evaluated as a construction-phase workflow in which prediction, reliability monitoring, triggering, lag handling, UCS updating, and retraining operate as connected steps.

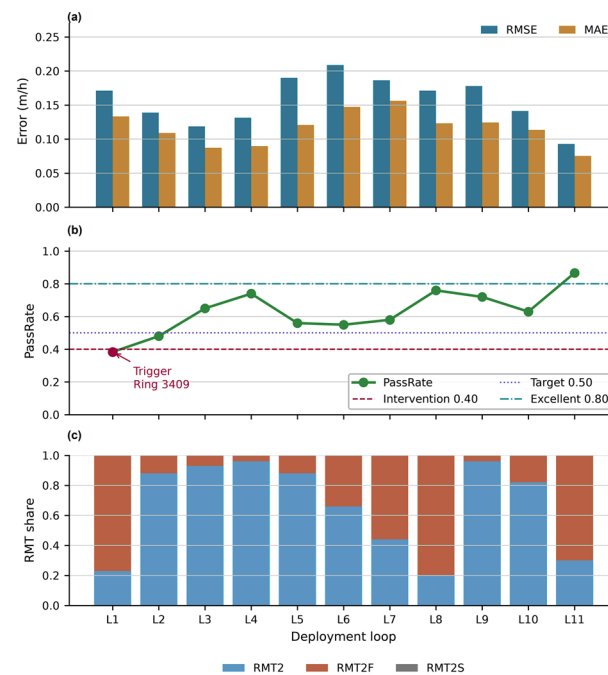


Figure 7. Phase-2 deployment performance timeline: (a) loop-wise RMSE and MAE; (b) PassRate with intervention, target, and excellent thresholds; and (c) loop-wise RMT composition.

Figure 7 further illustrates the evolution of loop-wise RMSE, MAE, PassRate, and RMT composition. The performance trend confirms that prediction reliability varied along the alignment but remained manageable after the initial trigger event. Per-class evaluation over the Phase-2 prediction rows gave RMSE = 0.1543 m/h and PassRate = 63.75% for RMT2, and RMSE = 0.1848 m/h and PassRate = 58.56% for RMT2F. No independent Phase-2 reliability claim is made for RMT2S because this class was sparse in the Stage-2 stream and absent from the evaluated Phase-2 prediction rows (Supplementary Tables S5 and S6).

Overall, the Phase-2 results demonstrate that the proposed framework can maintain deployable prediction performance over a forward construction stream rather than relying on a single static train–test assessment. The overall within-band performance of 62.63% exceeds the target-tier proportion when interpreted at the deployment-summary scale, while the event log shows that reductions below the 0.40 trigger threshold can be detected and converted into update actions. The combined results from Tables 5 and 6 therefore support the use of the framework as an auditable prediction and monitoring procedure for construction-phase TBM performance control.

The 62.63% within-band value should therefore be interpreted as an operational reliability level for a forward monitoring workflow, not as a deterministic accuracy claim for every ring. Under the same Phase-2 stream, direct static application of representative theoretical and empirical models achieved only 11.92–20.71% within-band performance, so the main value of the framework is the combination of improved within-band frequency with a trigger rule that identifies when performance degradation requires update action.

3.5. Benchmark Comparison with Theoretical and Empirical Models

The Phase-2 deployment framework was benchmarked against representative theoretical and empirical TBM penetration-rate models applied to the same Phase-2 dataset as a direct static application baseline, not as a recalibrated or locally optimized benchmark for each conventional formulation. The comparison is summarized in Table 7 and illustrated in Figure 8. The purpose of this benchmark was not to reinterpret, invalidate, or optimize

the original formulations, but to evaluate how their direct application compares with the proposed construction-phase workflow under the same ring-scale deployment data.

Table 7. Benchmark comparison between the proposed Phase-2 deployment framework and representative theoretical and empirical TBM penetration-rate models on the Phase-2 dataset.

Model/Operation	n.	RMSE	MAE	Within $\pm 10\%$
Phase2_Deployment Framework	990	0.1639	0.1186	62.63
Existing Models				
Farrokhi et al. model [8]	990	0.4700	0.3929	19.70
Gehring model [6]	990	0.7213	0.5576	11.92
Alpine/Wilfing model [7]	990	0.5603	0.4214	20.71

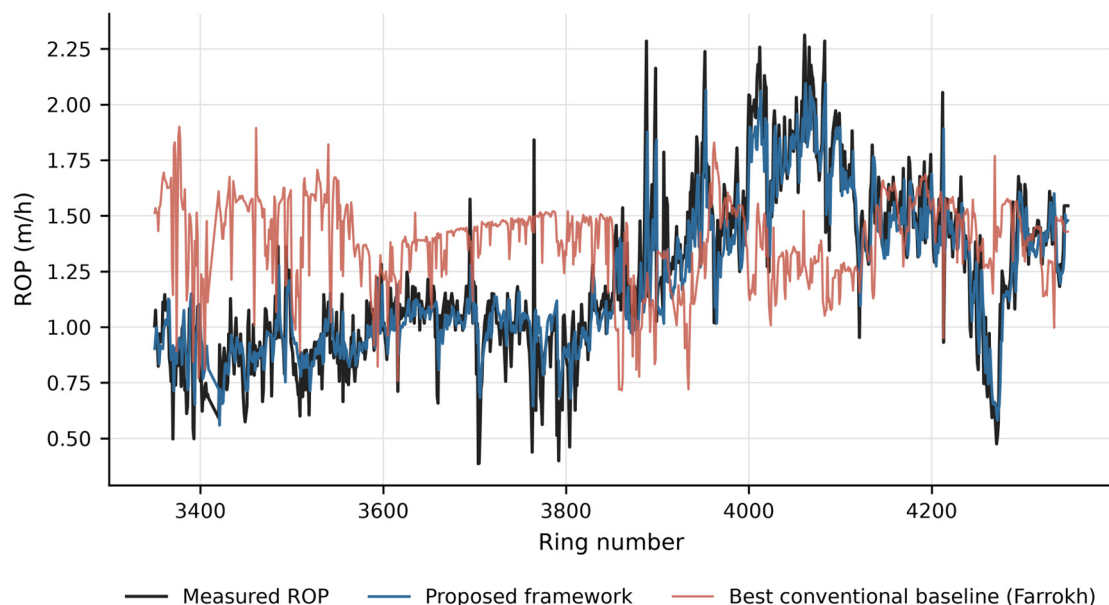


Figure 8. Time-series comparison of measured ROP, the proposed framework prediction, and the best-performing zero-shot conventional baseline during Phase-2 deployment.

The proposed framework achieved an overall RMSE of 0.1639 m/h, MAE of 0.1186 m/h, and 62.63% within the symmetric $\pm 10\%$ accuracy band over 990 evaluated rings. In contrast, the representative theoretical and empirical models produced substantially larger errors, with RMSE values ranging approximately from 0.4700 to 0.7213 m/h and within-band rates from 11.92% to 20.71%. This comparison indicates that direct application of conventional models was insufficient to provide construction-phase prediction quality for the present ring-scale deployment dataset.

The difference was especially clear when evaluated using PassRate. While RMSE and MAE quantify average point error, PassRate expresses how often predictions remain within an operational tolerance band. The proposed framework exceeded the 0.50 target tier at the overall deployment scale, whereas the conventional models remained far below the 0.40 intervention threshold. This result shows that the benefit of the proposed framework is not limited to lower average error; it also improves the frequency with which predictions remain within a practically interpretable accuracy band.

The benchmark should be interpreted in context. Theoretical and empirical models remain valuable for design-stage estimation, mechanistic interpretation, and preliminary planning. However, they are generally static and were not designed to incorporate sequential construction history, delayed UCS information boundaries, checkpoint monitoring,

trigger logic, or scheduled retraining. The comparison is therefore intentionally framed as a zero-shot direct static conventional baseline versus a locally trained learning-based construction workflow. This asymmetry is a limitation of the benchmark, but it also reflects the practical difference between applying published equations directly and maintaining a site-specific construction-phase model.

As shown in Figure 8, the best-performing existing baseline captured some broad variation in the Phase-2 ROP signal but showed larger systematic deviations from measured ROP than the proposed deployment framework. This visual behavior is consistent with the lower within-band rates reported in Table 7. The benchmark results therefore support the central argument of this study: construction-phase TBM prediction should be evaluated as an updateable workflow with explicit information boundaries and reliability monitoring, rather than as a one-time static prediction exercise.

4. Discussion

The results demonstrate that construction-phase TBM penetration-rate prediction should be treated as a monitored workflow rather than as a one-time model fitting exercise. The proposed framework combines chronological evaluation, explicit UCS information boundaries, PassRate-based reliability screening, and deployment updating into a single operational procedure. This structure is important because hard rock TBM performance evolves along the alignment as rock-mass condition, fracture intensity, machine operation, and available geotechnical information change during excavation.

4.1. Effect of Delayed UCS Information and Information Leakage Control

A central feature of the framework is the explicit control of UCS information availability. In completed-project analysis, UCS values measured within validation or test intervals can easily be incorporated into completed UCS surfaces or block models, even though such information would not have been available at the time of forward prediction. This can create information leakage and overstate deployable model performance. Scenario B was therefore introduced as a leakage-controlled condition in which measured UCS inside validation and test blocks was excluded from UCS completion for those same blocks.

The Scenario B results indicate that the selected GRU model retained stable predictive performance even under restricted UCS availability. The fold-average RMSE and MAE of 0.2368 and 0.1675 m/h, respectively, were close to the best Scenario A setting. This suggests that the combination of sequential operational parameters, RMT, and past-only UCS estimates preserved useful predictive structure when forward UCS measurements were unavailable. From an engineering perspective, this result is important because it supports the feasibility of using the model during construction, where complete ring-scale UCS information is rarely available in real time.

However, the Scenario B results should not be interpreted as proof that UCS uncertainty is unimportant. Rather, they show that prediction can remain operationally useful when UCS is handled under an auditable information policy. The value of Scenario B is therefore methodological as much as numerical: it provides a conservative test of whether the prediction workflow can operate under construction-realistic information constraints.

4.2. Role of PassRate Monitoring Beyond RMSE and MAE

The study also shows that RMSE and MAE alone are insufficient for construction-phase decision support. These metrics summarize average point error over a block, but they do not directly indicate whether recent predictions remain acceptable for operational monitoring. A model can achieve a reasonable average error while still producing clusters of unacceptable predictions over short intervals. For this reason, the proposed framework

converts prediction error into a symmetric $\pm 10\%$ within-band indicator and then evaluates PassRate over rolling monitoring windows.

The selected 50-ring monitoring window and 0.40 trigger threshold should be interpreted as an intervention rule rather than a claim of high-precision prediction. The 0.40 threshold defines a minimum acceptable reliability state below which continued unmodified deployment is considered unsuitable. In contrast, the 0.50 and 0.80 tiers represent target and excellent reliability levels. This tiered interpretation is important because it avoids presenting a single pass/fail criterion as a universal standard. Instead, the threshold can be calibrated according to project tolerance, ground variability, and the consequences of prediction error.

In Phase-2 deployment, PassRate monitoring provided information that RMSE and MAE could not provide on their own. The event log showed exactly when short-term reliability fell below the intervention threshold, when retraining was triggered, and when scheduled retraining was sufficient. This demonstrates the practical value of PassRate as a construction-phase monitoring signal that links model performance to operational action.

4.3. Deployment Behavior and Model Updating

The Phase-2 deployment results confirm the need for update logic in non-stationary tunnelling conditions. The first deployment loop fell below the 0.40 PassRate threshold at Ring 3409, triggering UCS updating, retraining, and redeployment after a 10-ring practical lag. After this trigger, subsequent loops remained above the intervention threshold and were managed through scheduled retraining. This behavior indicates that the framework can detect early degradation and restore short-term reliability through updating.

The overall Phase-2 performance, with RMSE = 0.1639 m/h, MAE = 0.1186 m/h, and 62.63% within the $\pm 10\%$ accuracy band over 990 evaluated rings, supports the effectiveness of the loop-based procedure. The improvement is not only reflected in lower error values, but also in the ability to maintain reliability above the target tier at the overall deployment scale. The deployment event log and loop-wise performance table should therefore be interpreted together: the event log explains when the model was allowed to continue or required updating, while the performance metrics quantify the resulting predictive accuracy.

The final deployment stream also shows that prediction reliability is not uniform along the alignment. Several loops remained within target or excellent tiers, while others approached the intervention threshold. This variability is expected in hard rock tunnelling because local geology, fracture intensity, cutter–rock interaction, and operating strategy can change over relatively short distances. The framework therefore does not assume that one model remains valid indefinitely. Instead, it provides a repeatable mechanism for monitoring and updating the model as construction progresses.

The local reliability reductions are consistent with the engineering expectation that fractured granite intervals can introduce greater short-range heterogeneity, blocky breakage behavior, variable cutter–rock interaction, muck variability, and operator adjustments. Nevertheless, the Phase-2 loop results do not show a one-to-one relationship between RMT2F proportion and lower reliability; Loop 8 contained a high RMT2F proportion but maintained strong PassRate after prior updating. This indicates that reliability is influenced jointly by rock-mass condition, operating regime, UCS updating, and the amount of relevant recent training information. Because the dataset was dominated by RMT2 and contained limited RMT2S, the results should be interpreted as strongest for a granite-dominated construction stream, with only limited class-specific inference for faulted granite. Condition-specific reporting is also consistent with broader infrastructure-materials research, where

mechanical performance is interpreted by material condition rather than only by a global metric [25].

4.4. Comparison with Theoretical and Empirical Models

The benchmark comparison shows that direct application of representative theoretical and empirical models was less effective for the Phase-2 ring-scale deployment dataset. These models produced RMSE values of approximately 0.4700–0.7213 m/h and within-band rates of 11.92–20.71%, whereas the proposed framework achieved RMSE = 0.1639 m/h and 62.63% within-band performance. The difference was especially clear when evaluated using PassRate, because the conventional models remained far below the 0.40 intervention threshold.

This comparison should be interpreted carefully. Theoretical and empirical models remain valuable for design-stage estimation, machine selection, mechanistic interpretation, and preliminary feasibility assessment. Their weaker performance in this study does not imply that the original formulations are invalid. Rather, it indicates that direct static application is not sufficient for construction-phase ring-scale prediction under evolving ground–machine interaction and delayed UCS information. The proposed framework adds a construction-phase layer that incorporates sequential learning, information-boundary control, monitoring, and updating.

Therefore, the contribution is not simply that a GRU model outperformed conventional equations. The more important contribution is the workflow that defines how prediction should be evaluated and maintained during excavation. This distinction is critical for practical use because construction teams require not only a prediction value, but also a rule for determining when the prediction process remains reliable and when it should be updated.

4.5. Practical Implications and Limitations

The proposed framework has several practical implications for TBM construction management. First, it allows prediction performance to be monitored in a form that is interpretable to project teams through the $\pm 10\%$ PassRate. Second, it separates available and unavailable UCS information, reducing the risk of overestimating model performance through data leakage. Third, it provides an auditable event log that records checkpoint reliability, trigger decisions, lag intervals, UCS update boundaries, and redeployment timing. These features make the framework more suitable for construction-phase decision support than a static model accuracy report.

The framework is also intentionally practical. It relies on common TBM operational records, routine geological classification, and UCS values that can be measured or estimated during construction. It does not require advanced real-time instrumentation beyond typical construction and geology records. However, transferability should be understood as procedural rather than numerical. The model configuration, monitoring-window length, tolerance band, PassRate thresholds, UCS completion method, and retraining cadence must be recalibrated before use in another geology, machine class, or data environment.

Several limitations should also be acknowledged. The framework was developed and validated using a single hard rock TBM project in granite-dominated conditions and a double shield TBM context. Its performance may differ in other geological settings, machine classes, or projects with different data quality. Because RMT2 dominated the Stage-2 stream and RMT2S accounted for only 1.1%, the results should not be interpreted as an independent validation for faulted granite. The UCS completion method used IDW with $p = 2$ as a fixed a priori operational setting; alternative interpolation methods and IDW powers should be tested in future work. The GRU results were generated with a fixed

deterministic seed for auditability, while repeated-seed analysis, bootstrap uncertainty intervals, and cross-project validation remain important future work.

4.6. Transferability of the Framework

Although the numerical thresholds selected in this study were calibrated to the case project, the framework can be transferred only as a procedure. The general workflow consists of chronological model evaluation, explicit geotechnical information boundaries, PassRate-based reliability screening, checkpoint monitoring, trigger logic, lag handling, UCS updating, and scheduled retraining. Other projects should recalibrate the sequence model, monitoring-window length, tolerance band, and reliability thresholds according to project-specific geology, TBM class, construction rate, data latency, and risk tolerance. Without such recalibration or independent project evidence, the numerical settings reported here should not be treated as universal recommendations.

This distinction between a transferable workflow and project-specific parameter settings is central to the proposed framework. The objective is not to prescribe a universal GRU configuration or a universal PassRate threshold. Instead, the framework provides a structured method for converting ring-scale prediction into a construction-phase reliability process that can be calibrated, audited, and updated during excavation.

5. Conclusions

This study developed and evaluated a construction-phase reliability framework for hard rock TBM ROP prediction under delayed UCS information. The main contribution is an auditable workflow that combines sequence prediction, UCS information boundaries, rolling evaluation, PassRate monitoring, trigger logic, practical lag handling, UCS updating, and scheduled retraining.

Under leakage-controlled Scenario B, the selected GRU configuration remained operationally viable when measured UCS inside validation and test blocks was excluded from UCS completion. The tolerance sensitivity confirmed that the $\pm 10\%$ band was strict, producing lower within-band rates than a $\pm 15\%$ band and higher rates than a $\pm 5\%$ band.

Scenario B reliability screening fixed the 50-ring monitoring window and 0.40 PassRate trigger before Phase-2 evaluation. The 0.40 tier is an intervention threshold for detecting sustained degradation, not a universal acceptance standard.

In Phase-2 forward deployment-style evaluation over Rings 3350–4349, the framework achieved RMSE = 0.1639 m/h, MAE = 0.1186 m/h, and 62.63% within the $\pm 10\%$ band over 990 evaluated rings. One early trigger occurred at Ring 3409; after UCS updating, retraining, and redeployment following the 10-ring lag, later loops remained above the intervention threshold.

Compared with representative theoretical and empirical models applied as zero-shot direct static baselines, the proposed locally trained workflow produced substantially higher within-band performance. The result supports the value of treating construction-phase prediction as a monitored and updateable reliability process rather than a static model-comparison exercise.

The framework is transferable as a workflow, but not as a universal numerical setting. Future work should test the method across independent projects, evaluate alternative UCS completion methods and IDW powers, quantify prediction uncertainty across repeated seeds, and extend the monitoring rule to mixed-face or highly heterogeneous ground conditions.

Supplementary Materials: The following supporting information can be downloaded at: <https://www.mdpi.com/article/10.3390/geotechnics6030061/s1>, Table S1: Descriptive statistics for predictor variables and ROP across Stage 1, Stage 2, and Phase-2; Table S2: Stage-1 model screening and full metrics; Table S3: Scenario B tolerance sensitivity; Table S4: Scenario B RMT-specific metrics; Table S5: Phase-2 RMT-specific metrics; Table S6: Phase-2 loop-wise RMT composition; Table S7: Full Phase-2 deployment event log; Table S8: 10-ring lag justification.

Author Contributions: Conceptualization, N.M. and T.T.; methodology, N.M. and T.T.; software, N.M.; validation, N.M. and T.T.; formal analysis, N.M.; investigation, N.M.; resources, N.M. and T.T.; data curation, N.M.; writing—original draft preparation, N.M.; writing—review and editing, N.M. and T.T.; visualization, N.M.; supervision, T.T.; project administration, N.M. and T.T. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available on request from the corresponding author due to privacy.

Acknowledgments: The authors gratefully acknowledge Right Tunnelling Public Company Limited for providing access to the TBM construction records and project information used in this study. The authors also thank the TBM operation team and geology team involved in the Mae Taeng–Mae Ngad–Mae Kuang water transfer tunnel project for their support and contribution to data collection during construction. The authors sincerely appreciate the guidance and support received from the Department of Civil Engineering, Faculty of Engineering, Chiang Mai University, throughout the development of this research.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

Abbreviation	Definition
AACE	Association for the Advancement of Cost Engineering
AI	Artificial intelligence
CSM	Colorado School of Mines
DL	Deep learning
GRU	Gated recurrent unit
IDW	Inverse distance weighting
LSTM	Long short-term memory
MAE	Mean absolute error
MCSM	Modified Colorado School of Mines model
NTNU	Norwegian University of Science and Technology
RMT	Rock mass type
RMT2	Massive granite
RMT2F	Highly fractured granite
RMT2S	Faulted granite
ROP	Rate of penetration
RMSE	Root mean square error
TBM	Tunnel boring machine
TCN	Temporal convolutional network
UCS	Uniaxial compressive strength

References

1. Rostami, J. Development of a Force Estimation Model for Rock Fragmentation with Disc Cutters through Theoretical Modeling and Physical Measurement of Crushed Zone Pressure. Ph.D. Thesis, Colorado School of Mines, Golden, CO, USA, 1997.
2. Yagiz, S. Development of Rock Fracture and Brittleness Indices to Quantify the Effects of Rock Mass Features and Toughness in the CSM Model Basic Penetration for Hard Rock Tunnelling Machines. Ph.D. Thesis, Colorado School of Mines, Golden, CO, USA, 2002.
3. Yagiz, S. TBM performance prediction based on rock properties. In *Proceedings of the International Symposium of the International Society for Rock Mechanics, EUROCK 2006, Liege, Belgium, 9–12 May 2006*; Taylor & Francis/Balkema: London, UK, 2006; pp. 663–670.
4. Bruland, A. *Hard Rock Tunnel Boring, Vol. 7: The Boring Process*; Norwegian University of Science and Technology, Department of Building and Construction Engineering: Trondheim, Norway, 2000.
5. Macias, F.J. Hard Rock Tunnel Boring: Performance Predictions and Cutter Life Assessment. Ph.D. Thesis, Norwegian University of Science and Technology, Trondheim, Norway, 2016.
6. Gehring, K.H. Leistungs- und Verschleissprognosen im maschinellen Tunnelbau. *Felsbau* **1995**, *13*, 439–448.
7. Wilfing, L. The Influence of Geotechnical Parameters on Penetration Prediction in TBM Tunnelling in Hard Rock. Ph.D. Thesis, Technische Universitat Munchen, Munich, Germany, 2016.
8. Farrokh, E.; Rostami, J.; Laughton, C. Study of various models for estimation of penetration rate of hard rock TBMs. *Tunn. Undergr. Space Technol.* **2012**, *30*, 110–123. [[CrossRef](#)]
9. Monthanopparat, N.; Tanchaisawat, T. Advance Tunnel Boring Machine Performance Prediction in Massive and Highly Fractured Granite: Integrating Innovative Deep Learning and Block Model Technique. *Geotech. Eng. J. SEAGS AGSSEA* **2024**, *55*, 26–34.
10. Hochreiter, S.; Schmidhuber, J. Long short-term memory. *Neural Comput.* **1997**, *9*, 1735–1780. [[CrossRef](#)] [[PubMed](#)]
11. Cho, K.; van Merriënboer, B.; Gulcehre, C.; Bahdanau, D.; Bougares, F.; Schwenk, H.; Bengio, Y. Learning phrase representations using RNN encoder-decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, Doha, Qatar, 25–29 October 2014*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2014; pp. 1724–1734.
12. Bai, S.; Kolter, J.Z.; Koltun, V. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. *arXiv* **2018**, arXiv:1803.01271.
13. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017*; pp. 5998–6008.
14. Feng, S.; Chen, Z.; Luo, H.; Wang, S.; Zhao, Y.; Liu, L.; Ling, D.; Jing, L. Tunnel boring machines (TBM) performance prediction: A case study using big data and deep learning. *Tunn. Undergr. Space Technol.* **2021**, *110*, 103636. [[CrossRef](#)]
15. Gao, B.; Wang, R.; Lin, C.; Guo, X.; Liu, B.; Zhang, W. TBM penetration rate prediction based on the long short-term memory neural network. *Undergr. Space* **2021**, *6*, 718–731. [[CrossRef](#)]
16. Liu, Z.; Wang, Y.; Li, L.; Fang, X.; Wang, J. Realtime prediction of hard rock TBM advance rate using temporal convolutional network (TCN) with tunnel construction big data. *Front. Struct. Civ. Eng.* **2022**, *16*, 401–413. [[CrossRef](#)]
17. Mahmoodzadeh, A.; Nejati, H.R.; Mohammadi, M.; Ibrahim, H.H.; Rashidi, S.; Rashid, T.A. Forecasting tunnel boring machine penetration rate using LSTM deep neural network optimized by grey wolf optimization algorithm. *Expert Syst. Appl.* **2022**, *209*, 118303. [[CrossRef](#)]
18. Shan, F.; He, X.; Armaghani, D.J.; Sheng, D. Effects of data smoothing and recurrent neural network (RNN) algorithms for real-time forecasting of tunnel boring machine (TBM) performance. *J. Rock. Mech. Geotech. Eng.* **2024**, *16*, 1538–1551. [[CrossRef](#)]
19. Li, L.; Liu, Z.; Fang, X.; Qi, W. Multi-step real-time prediction of hard-rock TBM penetration rate combining temporal convolutional network and squeeze-and-excitation block. *Sci. Rep.* **2024**, *14*, 14326. [[CrossRef](#)] [[PubMed](#)]
20. Khatti, J.; Mishra, S. Estimating shield tunnel boring machine penetration rate in mixed face conditions: Feature selection and multicollinearity effects on machine and deep learning models. *Front. Built Environ.* **2025**, *11*, 1699466. [[CrossRef](#)]
21. Rezaei, M.; Fallahi, S. Block model optimization and resource estimation of the Angouran Mine by transferring the exploratory data from the local coordinate system to the UTM. *Miner.-Geol.-Pet. Eng. Bull.* **2023**, *38*, 1–17. [[CrossRef](#)]
22. Ruffhead, A.C. Enhancement of inverse-distance-weighting 2D interpolation using accelerated decline. *Rep. Geod. Geoinform.* **2023**, *116*, 9–14. [[CrossRef](#)]
23. Chai, T.; Draxler, R.R. Root mean square error (RMSE) or mean absolute error (MAE)? Arguments against avoiding RMSE in the literature. *Geosci. Model Dev.* **2014**, *7*, 1247–1250. [[CrossRef](#)]

24. AACE International. *TCM Framework: 7.3-Cost Estimating and Budgeting*; Cost Estimate Classification System as Applied in Engineering, Procurement, and Construction for the Process Industries; AACE International: Morgantown, WV, USA, 2005.
25. Imoh, U.U.; Apata, A.C.; Raji, S.O.; Adeboje, A.O.; Babafemi, A.J.; Movahedi Rad, M. Mechanical and microstructural performances of hot-mix asphalt modified with recycled polyethylene terephthalate. *Results Eng.* **2026**, *29*, 109572. [[CrossRef](#)]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.