Proceedings World Geothermal Congress 2026
Calgary, Canada, June 8 – 11, 2026

The Overlooked Foundation: The Crucial Role of Preprocessing in Enhancing Rock Unit Characterization

Hammoud A[1]*, Villeneuve M[1], Caspari E[1] , Mortensen A[2] and El Kasti K[3]
Technical University of Leoben, Leoben, Austria[1]*, Landsvirkjun, Iceland[2], University of Leeds, United Kingdom[3]
aliaa.hammoud@unileoben.ac.at

Keywords: Wireline logs, Drilling data, Anomalies, Data selection, Data preprocessing

ABSTRACT

As data science continues to gain ground in geothermal exploration and drilling, there is growing emphasis on automation and machine learning. However, the foundational step of data preprocessing remains underappreciated despite being the most time-consuming and critical phase. Before any meaningful interpretation or model output can be achieved, raw data must undergo a rigorous transformation process involving selection, resampling, scaling, and careful evaluation of anomalies and heterogeneities. These preprocessing stages are not just preliminary, they are essential for ensuring the quality and integrity of any subsequent analysis or automated workflow.

This paper demonstrates data preprocessing using tools from Panda and Numpy built-in libraries in Python for wireline logging and drilling data acquired while drilling production wells at two geothermal fields in North East-Iceland. A key aspect of pre-processing is setting a clear objective, which guides the selection of relevant parameters and the identification of what could affect these parameters. Given that datasets are often recorded at varying resolutions, resampling to a unified basis, typically depth, is crucial. Additionally, many machine learning algorithms are scale-sensitive, thus necessitating proper normalization. Importantly, not all outliers are errors; some anomalies reflect meaningful geological events and should be retained. Finally, we show how improvements in data quality via preprocessing enabled the development of refined ROP and WOB drilling logs which aid in enhancing bit rock interaction models and better rock unit characterization. In geothermal exploration, success begins with how we prepare our data, because high quality data at the start determines performance at the end for better geological understanding and smarter decision making.

1 INTRODUCTION

Geothermal exploration and early resource characterization increasingly rely on multi-source digital data streams from wireline logs and mud-logging to high-frequency drilling parameters. Concurrently, the sector is embracing automation and machine learning (ML) to accelerate interpretation and optimize decision-making. Yet the accuracy and transferability of any downstream model are determined upstream by how raw data are prepared. In practice, the most time-consuming and consequential work happens before modeling begins: selecting relevant parameters aligned to a clear objective, aligning parameters recorded at different resolutions, scaling parameters for algorithms that are scale-sensitive, and deciding which “outliers” represent tool noise versus geological events. Despite this impact, preprocessing choices are often under-reported, hindering reproducibility and masking sources of bias in geothermal case studies.

These challenges are amplified in volcanic geothermal systems, where sharp facies contrasts (e.g., massive lavas, hyaloclastites, intrusions, fractured zones) drive strong heterogeneity in petrophysical responses. Wireline logs can be affected by borehole conditions (washouts, breakouts, temperature), while drilling signals capture both lithology and operational effects (bit wear, stick and slip, and vibrations).

A transparent, reproducible preprocessing framework is therefore a prerequisite for credible ML-enabled or data-driven geothermal characterization. Rather than proposing a new algorithm, this paper focuses on the “front-end” steps that most strongly control downstream outcomes: (i) goal-driven parameter selection that ties every parameter to a geologic hypothesis; (ii) resampling the parameters to a unified basis, typically measured depth; (iii) anomaly triage that distinguishes correctable acquisition/operational artifacts from informative geological deviations; and (iv) scaling strategies matched to algorithm behavior and distribution shape. We demonstrate this framework using case studies from high-temperature geothermal fields in NE Iceland, including wireline logs (e.g., gamma ray, resistivity, neutron) with drilling data (e.g., rate of penetration (ROP), weight on bit (WOB), rotary speed in (RPM), and torque).

2 PARAMETER SELECTION

Selecting the parameters is the first, and often most consequential step before applying any ML or Artificial Intelligence (AI) method. The choice of the parameters should be driven by the objective of the study, availability, and significance of the parameters. Parameter selection can be part of a robust parameter optimisation workflow (e.g. Villeneuve and Moayedi-Far, in review), can be based on the availability of parameters (e.g. Hammoud et al., 2025), or can be based on known physical relevance of the parameters. We present two examples from different data sets to demonstrate parameter selection based on their availability and physical relevance.

2.1 Wireline log parameter selection

Hammoud et al. (2025) developed a Gaussian Mixture Model (GMM) framework to characterize rock properties—primarily porosity state and composition—from a limited selection of routine wireline logs. In Icelandic high-temperature wells, tool survivability or the need for cooling the borehole often limits the log suite. The motivation was to use machine learning techniques to characterize intervals in the stratigraphic log where cuttings were missing or where labels were unreliable due to sample mixing. Consequently, parameter selection for the GMM was guided by two criteria: (i) availability of logs in the studied high-temperature volcanic wells, and (ii) their physical response represented in the logs must be diagnostic of porosity and/or lithology.

Resistivity is sensitive to pore-fluid content and the connected porosity and thus relates to effective porosity (Schön, 2015); neutron responds to hydrogen index and serves as a proxy for total porosity (Schön, 2015); gamma ray measures natural radioactivity and is commonly used as a compositional indicator (felsic units tend to exhibit higher gamma-ray counts than intermediate to mafic rocks because of the proportion of potassium feldspar) (Planke et al., 1994; Bartetzko et al., 2005, Tsutsui et al., 2024). Given this complementarity—and the absence of additional logs—the analysis used all three parameters jointly, performing clustering in a three-dimensional feature space.

Model outputs were subjected to independent geological validation to assess plausibility. Because validation must not reuse the parameters employed in the unsupervised model, we compared cluster-based interpretations against available cuttings descriptions, which provided independent evidence of composition and a qualitative porosity state.

In sum, parameter selection for unsupervised clustering should reflect (1) data availability, (2) the physical meaning of each measurement relative to the study objective, and (3) the need for an independent line of evidence for validation. This approach is particularly important when the measured log suite is limited, as in the presented case.

2.2 Drilling data parameter selection

Drilling datasets typically include multiple measured parameters, e.g. ROP, WOB, torque, RPM, standpipe pressure (SPP), flow rate, and mud temperature in/out. Selecting the optimal parameters depends on the study objective. Many studies on drilling datasets have used supervised machine-learning models to predict ROP, porosity, unconfined compressive strength (UCS) and Poisson's ratio (Alhowaish and Mezghani, 2023; Siddig et al., 2021; Gowida et al., 2021). Parameter selection in these studies was guided by correlation analysis and feature-importance analysis.

An ongoing study (Hammoud et al., in prep.) builds on the porosity states previously clustered from resistivity and neutron logs (Hammoud et al., 2025) to guide an analysis of which drilling parameters correlate most strongly with porosity. Spearman correlation (non parametric measure that measures the strength of correlation between two variables possibly non linear) and feature-importance analysis ranked WOB and ROP as having the lowest correlation or relation to resistivity and neutron, whereas, torque, and SPP showed high correlation values with resistivity (Figure 1). Upon closer examination it was clear that the high correlation for torque and SPP with resistivity was prominent because all of these parameters increase proportionally with depth. This highlights the need for a clear understanding of the factors influencing each parameter to determine whether correlations are meaningful.

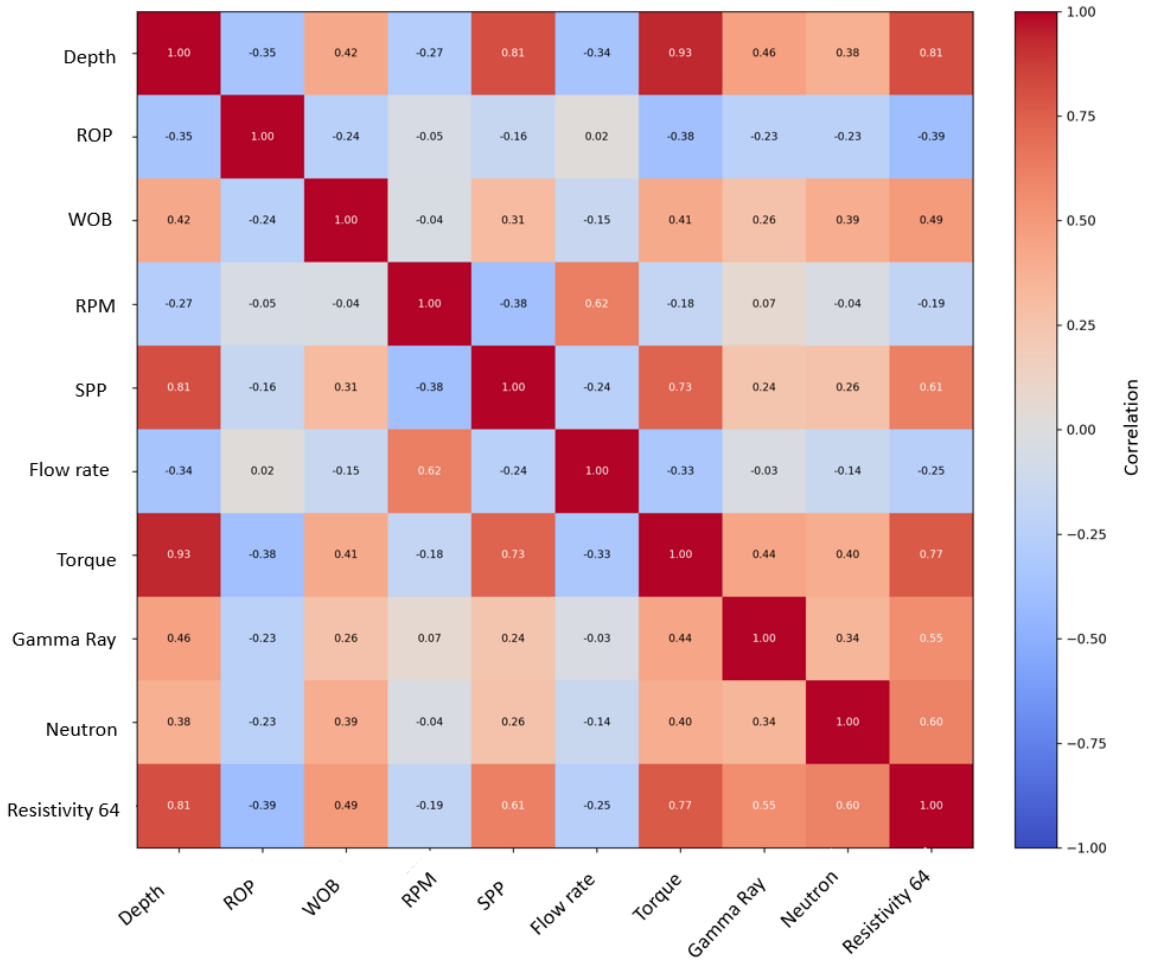


Figure1. Correlation matrix (Spearman) of depth, drilling parameters (ROP, WOB, RPM, SPP, Flow rate, Torque) and well log variables (Gamma Ray (GR), Resistivity and Neutron)

Operational and data-integration factors further affect the selection of key parameters. In a ~2 km deep well, WOB and ROP can be selected by the driller, whereas RPM tends to be held within a narrow range; For example, in well TH19, ROP was held mostly constant over a ~200 m interval while WOB fluctuated, whereas the opposite occurred in other intervals, resulting in an often non-monotonic relationship between the two parameters, as evidenced by the low Spearman value. WOB and ROP are also sometimes controlled from surface, for example due to well dropping in deviated wells, where the driller reduces ROP or WOB to limit the drop, such that at the well scale the relationship between these drilling parameters and geophysical parameters is overshadowed by operational needs. Depth matching presents an additional challenge when correlating datasets with different measured-depth references. In our datasets, we also observed issues such as a sensor recording unreasonable ROP and depth values which were removed.

The correlation matrix does not show a relation between ROP, WOB, and porosity state (from Resistivity and Neutron) for the whole processed well. This arises from the variability of the materials encountered along the entire well length, the classification of porosity using clusters of Resistivity and Neutron (see Hammoud et al., 2025), and the variability of ROP and WOB based on driller input. When examining subsets of depth intervals with similar borehole assemblies and operating conditions, relations between these parameters can be observed. The selection of ROP and WOB as indicatives of strength changes in the rock mass is, therefore, not based on a simple parameter correlation exercise, but rather based on previously conducted studies linking porosity, rock strength and drilling performance. Studies have shown that ROP and WOB respond to changing porosity through their relationship with UCS, whereby as porosity increases, UCS decreases (Heap and Violay 2021), and ROP should increase under constant WOB (Khoshnevis-zadeh et al., 2019; Li et al., 2016). At constant ROP, increasing porosity results in decreasing WOB.

3 ANOMALIES VS. HETEROGENEITIES

Anomalies are data points that deviate significantly from the overall trend of the dataset. However, not all anomalies constitute “bad” data; in many cases these deviations are informative. Depending on the purpose of

the study, anomalies may be retained or removed. In this section, we present cases from wireline-log and drilling datasets illustrating both (i) removal of non-informative anomalies and (ii) retention of anomalies that are informative.

3.1 Wireline log anomalies vs heterogeneities

Icelandic geothermal fields occur in a highly heterogeneous lithologic environment. Although predominantly basaltic, the stratigraphic column also includes rhyolitic and andesitic intrusions and lavas (Thordarson and Larsen, 2007; Gunnarsson et al., 1998). During drilling, gamma-ray is measured because it is informative of rock composition. In many wells, spikes in gamma-ray that deviate significantly from the background trend are observed (Figure 2). These spikes are not anomalous points; they indicate thin rhyolitic/andesitic layers and are therefore treated as heterogeneous data points.

Conversely, other spikes may be not useful, or they may signal the need for additional data cleaning because they can affect other parameters. A common example is borehole enlargement, identifiable on caliper logs. When enlargements are detected in the caliper log, the affected intervals are excluded from further analysis because enlargements can negatively impact the quality of other measurements such as increased Resistivity. Accordingly, we establish cutoff values representing enlargement as the original borehole diameter plus 10%. Data points where the caliper exceeds this threshold are flagged, and corresponding measurements in other logs are not used for further analysis. In summary, anomaly-detection techniques for wireline logs are important; however, to remove true anomalies without discarding high-quality data, it is essential to understand each log's purpose.

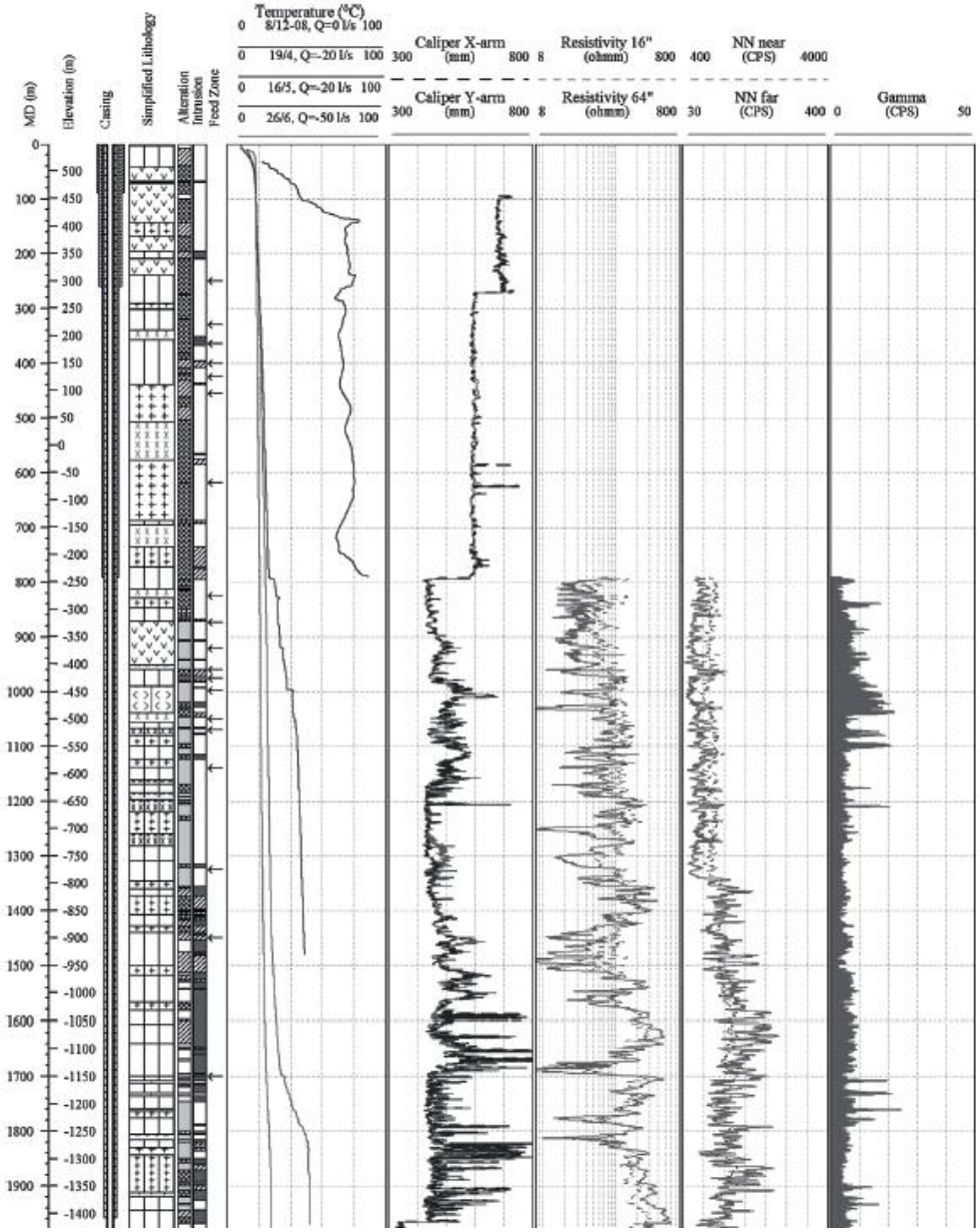


Figure 2. Lithological log with well logs for well IDDP-1. Spikes in gamma ray (e.g. at 1210 MD, and between 1700-1775 MD) and caliper (e.g. between 990-1020 MD, and between 1620 and 1690 MD) highlight an example of heterogeneities and anomalies, respectively. Adapted from Mortensen et al. (2014).

3.2 Drilling data anomalies vs heterogeneities

Drilling rig sensors record vibrations and noise generated during drilling. To ensure the data are fit for purpose, preprocessing is conducted in line with the study's objective. Common steps include (i) excluding all negative WOB values and instances with WOB = 0, and (ii) applying thresholds to exclude data outside predefined limits (Figure 3 &4 a). Our objective is to assess bit–rock interaction to infer rock properties, therefore fluctuations that do not serve this purpose should be excluded.

After defining the drilling parameters of interest here, WOB and ROP, it is important to assess factors that may affect these signals that do no reflect rock-property changes. During drilling and progressing deeper in the well, singles (new pipes) are added through the top drive. Adding singles introduces fluctuations in ROP and WOB that are operational rather than lithology related. To remove these fluctuations, peak and trough events in the top drive position log were automatically detected using the find_peaks() function from the SciPy library, with a prominence threshold of 1 to identify significant changes indicative of pipe additions. A 50-sample window was then defined around each detected top drive peak or trough, capturing the range most likely affected by the pipe handling process. Within these windows, fluctuations in WOB and ROP were further analyzed to detect anomalous signal patterns using a local prominence-based approach, again applying the find_peaks() function with a prominence threshold of 0.5 to exclude minor noise and isolate only significant anomalies. For each

identified peak or trough, the peak_widths() function was used to estimate the full width of the disturbance at 80% of the relative height. This allowed for precise characterization of the anomalous interval, and the corresponding sample span, defined by the left and right intercepts (left_ips and right_ips), was flagged and subsequently excluded from the dataset (Figure 3 &4 b).

Furthermore, intervals where ROP = 0 during active drilling indicates that the bit is consuming time to drill-off through a particular interval. The drill-off is seen when ROP creeps up again and fluctuates for approximately one minute. Because these fluctuations are operational rather than lithology related, they should be excluded, and this can be automated. Using Python, we compute a rolling mean of ROP with a window size of 10 (i.e., the mean of every 10 consecutive data points). For each event, if the ROP measured for 10 samples after the last ROP = 0 exceeds 20% of the rolling mean at that time, the affected data are flagged for exclusion.

Transforming drilling logs from time- to depth-index can introduce spurious fluctuations associated with operational stops (e.g., bit changes, pull-out-of-hole, sticking, and other activities). When drilling resumes, assuming WOB is restored to its pre-set level, fluctuations in ROP are often observed. These are startup transients rather than lithology-related changes. Accordingly, we apply the same rule used for post-sticking cleaning, but keyed to time gaps: if the elapsed time between consecutive samples exceeds 6 minutes, we compare the current ROP and the subsequent 10-sample window against the rolling mean (window = 10). If the ROP exceeds 20% of the rolling mean at that point, the affected data are flagged for exclusion.

Finally, axial string vibration, also known as bit bounce, manifests as fluctuations in WOB that, in turn, induce ROP fluctuations not driven by lithology. To detect bit bounce, we compute the rate of change of WOB with respect to depth:

$$\text{slope} = \frac{\Delta \text{WOB}}{\Delta \text{Depth}} \quad (1)$$

Then we calculate the percentage change between consecutive data points

$$\%\Delta WOB = \frac{WOBi - WOBi-1}{WOB\ i-1} * 100 \quad (2)$$

Finally, we flag WOB drops that are indicative of bit bounce when the data point meets all the following criteria:

A negative slope

A percentage drop greater than 15%

An absolute change in WOB greater than 1 ton

After removing non-informative operational data (Figure 3&4 c), the remaining signals can be averaged over a set depth interval, in our case 0.5 m (Figure 3&4 d), and then evaluated for heterogeneities. ROP or WOB spikes may still appear; however, these parameters can now be taken to reflect bit–rock interaction.

Drilling breaks are a sudden increase in ROP by a factor of two to three that can occur during drilling. In sedimentary basins, drilling breaks are often used to infer the presence of overpressured formations caused by under compacted, high-porosity rocks with reduced vertical effective stress, particularly in cases of compaction disequilibrium (Oloruntobi and Butt, 2019). Moreover, drilling breaks can also be triggered by abrupt lithological changes (SLB Glossary) or by operational changes, such as sudden increases in WOB or RPM, which are not always held constant during drilling operations (Oloruntobi and Butt, 2019). Given that operational changes are filtered out, drilling breaks in heterogenous volcanic environments can be the result of various geological events, for example lava flow boundaries, which can manifest as sudden increases in ROP (Millet, 2016).

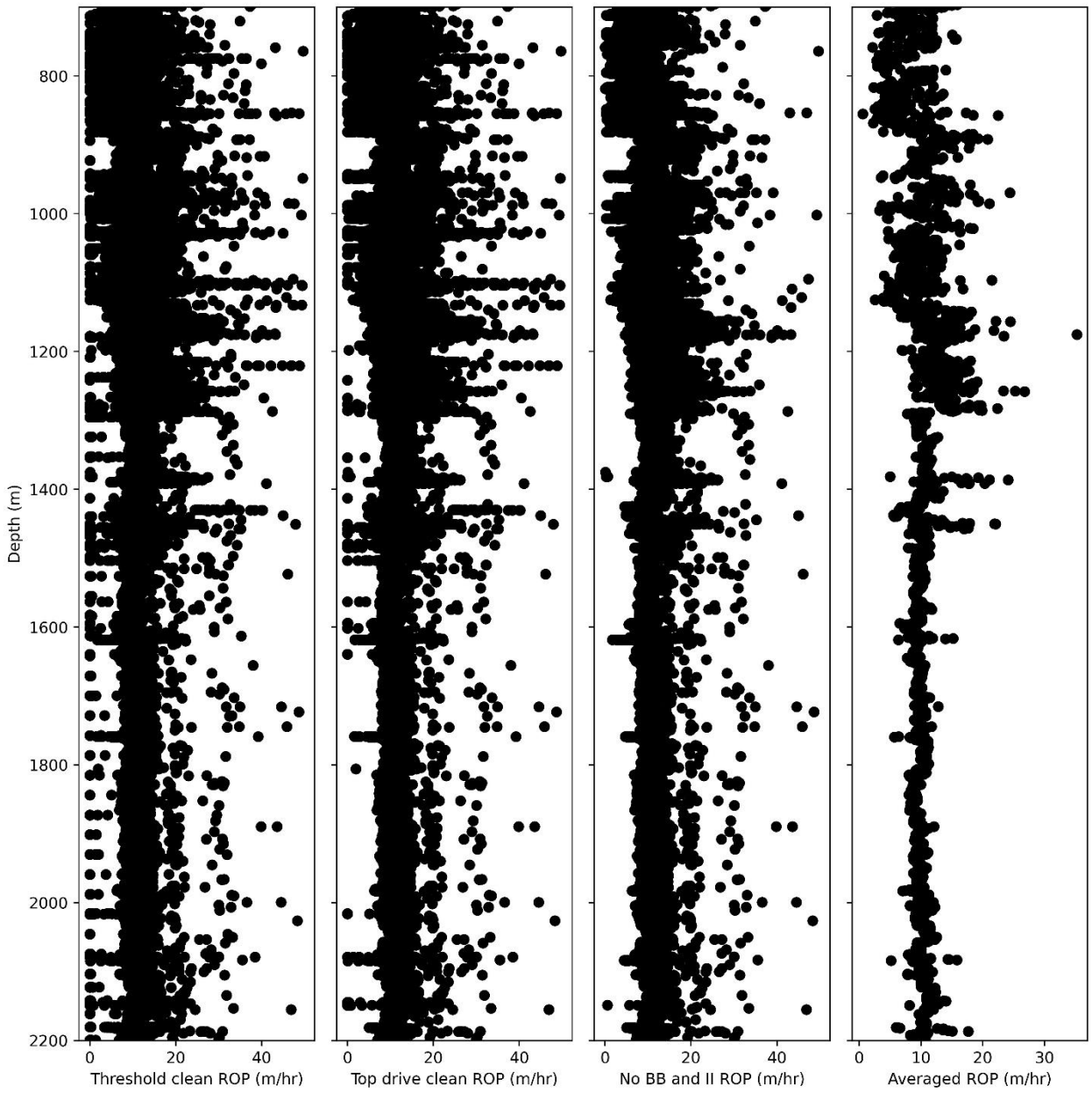


Figure 3. ROP through successive preprocessing stages: (a) threshold clean, (b) top-drive clean, (c) bit-bounce & invalid index removal, (d) averaged ROP illustrating progressive noise reduction.

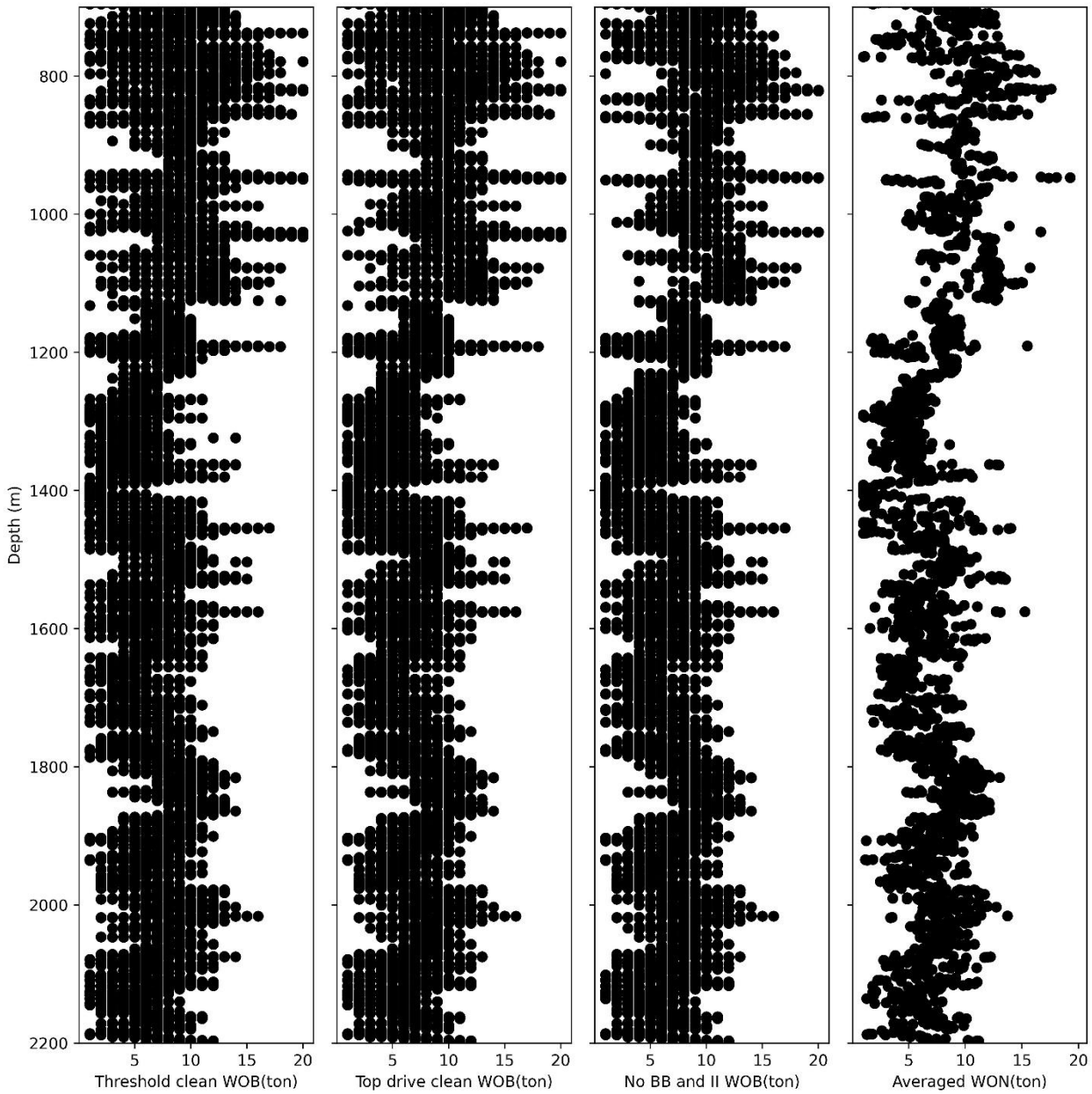


Figure 4. WOB through successive preprocessing stages: (a) threshold clean, (b) top-drive clean, (c) bit-bounce & invalid index removal, (d) averaged WOB illustrating progressive noise reduction.

4 DATA RESAMPLING

Working with diverse datasets requires understanding the basis on which parameters are measured to draw integrated conclusions. Wireline and cuttings data are depth-indexed (variables recorded versus measured depth), whereas drilling data collected by rig sensors are time-indexed. Below, we present the Python-based framework used in Hammoud et al. (2025b) to gather these datasets and resample all parameters to a common index (e.g., depth).

Dual Neutron and Gamma ray parameters were sampled at a 0.5 m spacing, whereas Resistivity was sampled at 0.25 m spacing. Before applying the Gaussian Mixture Model (GMM), we resampled the wireline data set into the same depth grid using 1D linear interpolation between the 2 nearest sampled depths.

After applying the GMM, each data point is assigned to a cluster. To validate the clustering, we cross-checked the results against cuttings descriptions of composition. This process was automated by digitizing the cuttings labels at their corresponding measured depths. For example, ISOR geology labels include: medium coarse-grained acidic, fine medium-grained acidic, acidic tuff, glassy basalt, and fine medium-grained basalts. We then created two categorical fields: (i) an exact composition–texture code (e.g., “Fine medium-grained basalts” → FMGB) and (ii) a broader composition class (e.g., “Fine medium-grained basalts,” “glassy basalts,” “basaltic

tuff" → Basalts) at 0.5 m intervals similar to the wireline data and clusters sampling. This categorisation of qualitative descriptions enabled producing histograms for each cluster, which were used to assess the compositional validity of the clustering results.

To combine drilling parameters with wireline and cuttings datasets, the time-indexed drilling records must be converted to a depth-indexed format so that all variables are referenced to measured depth. Rig sensors record drilling parameters at a 5-s interval, producing hundreds of thousands of rows per channel and an inconsistent depth spacing. To convert the drilling data from a time index to a depth index, we average values within a rolling depth window at a preferred spacing; in our case 0.5 m to match the depths from the wireline log cluster spacing. This averaging should be applied only after the drilling dataset has been properly preprocessed for the intended purpose (e.g. Section 3.2).

5 DATA SCALING

For unsupervised machine learning algorithms that cluster data points based on similarity, it is important to examine the value ranges of the selected parameters. For example, gamma ray, resistivity, and neutron logs have different ranges (e.g. 0-50 CPS, 0-500 ohm.m, 0-4000 CPS respectively); clustering on unscaled data can bias the model toward the parameter with the largest range. To avoid this, parameters should be scaled to comparable ranges. Common approaches include standardization and normalization [Belyadi and Haghighat 2021].

Before applying any scaling, it is essential to understand how the transformation affects the data to ensure that the overall data shape is preserved. In Hammoud et al. (2025), min–max normalization was applied to map each parameter to the [0,1] interval. Normalization was chosen over standardization because it is sensitive to outliers; as noted earlier, gamma "outliers" are treated as heterogeneities rather than anomalies, and normalisation preserved the data shape during scaling.

6 CONCLUSION

This paper shows why rigorous preprocessing is a necessary step in geothermal data science and provides examples of how this was undertaken in practice for geothermal well data from high-temperature, Icelandic volcanic settings. We demonstrate an objective-driven workflow that (i) selects parameters consistent with the study goals, parameter availability and physical relationships (resistivity, dual-neutron, gamma for GMM; ROP and WOB for drilling response), (ii) distinguishes heterogeneities from true anomalies (e.g., preserving gamma spikes that indicate thin felsic layers while excluding caliper-defined borehole enlargements), (iii) removes operational artifacts from drilling signals with explicit, automatable rules (pipe-addition windows, post-sticking and startup transients, and bit-bounce criteria), (iv) brings heterogeneous datasets onto a common depth index (wireline to a uniform 0.25 m grid; drilling via 0.5 m rolling-depth averaging), and (v) applies min–max normalisation to align value ranges without suppressing geologically meaningful variation. By making these choices explicit and reproducible, the study shows that preprocessing is not a preliminary chore but the core enabler to turning limited and/or noisy measurements into defensible geological insight and a foundation for trustworthy automation.

Acknowledgements

The authors gratefully acknowledge Landsvirkjun for access to data and for invaluable logistical and on-site support during fieldwork at the Krafla and Theistareykir geothermal fields. Their assistance was instrumental in the successful completion of this study. A. Hammoud was supported by a scholarship from the Rector's Office of Montanuniversität Leoben. Scientific mobility grants to A. Hammoud from the COST Actions DHC (CA12119) and FOLIAGE (CA21156) facilitated international collaboration and exchange that advanced this work.

ChatGPT (GPT-5) was used solely for language editing (improving grammar and clarity); it did not generate results or influence the scientific content. The authors remain responsible for all content.

REFERENCES

Alhowaish, H. A., Mezghani, M. M., and A. Shakriov.: "Predicting Porosity from Drilling Data Using Machine Learning – Challenges and Solutions." Paper presented at the ADIPEC, Abu Dhabi, UAE, October 2023.

Bartetzko, A., Delius, H., & Pechnig, R.: Effect of compositional and structural variations on log responses of igneous and metamorphic rocks. I: mafic rocks. Geological Society, London, Special Publications, **240(1**), (2005), 255–278.

Belyadi, H., & Haghighat, A.: Machine learning guide for oil and gas using Python: A step-by-step breakdown with data, algorithms, codes, and applications. Gulf Professional Publishing, (2021).

Gowida, A., Elkatatny, S. & Gamal, H.: Unconfined compressive strength (UCS) prediction in real-time while drilling using artificial intelligence tools. Neural Comput & Applic **33**, 8043–8054 (2021).

Gunnarsson, B., Marsh, B. D., & Taylor, H. P., Jr.: Generation of Icelandic rhyolites: Silicic lavas from the Torfajökull central volcano. Journal of Volcanology and Geothermal Research, **83(1–2)**, (1998), 1–45.

Hammoud, A., Villeneuve, M., Caspari, E., & Mortensen, A. K.: From logs to layers: Framework for interpreting subsurface facies in Iceland geothermal reservoirs. Geothermics, (2025).

Heap, M.J., Violay, M.E.: The mechanical behaviour and failure modes of volcanic rocks: a review. Bull Volcanol **83, 33** (2021).

Khoshnevis-zadeh, R., Soleimani, B. & Larki, E. Using drilling data to compare geomechanical parameters with porosity (a case study, South Pars gas field, south of Iran). Arab J Geosci **12, 611**, (2019).

Li, G., Yang, M., Meng, Y., Liu, H., Han, L., Zhou, F., & Zhang, H.: The assessment of correlation between rock drillability and mechanical properties in the laboratory and in the field under different pressure conditions. Journal of Natural Gas Science and Engineering, **30**, (2016), 405–413.

Millett, J. M., Wilkins, A. D., Campbell, E., Hole, M. J., Taylor, R. A., Healy, D., Jerram, D. A., Jolley, D. W., Planke, S., Archer, S. G., & Blischke, A.: The geology of offshore drilling through basalt sequences: Understanding operational complications to improve efficiency. Marine and Petroleum Geology, **77**, (2016), 1177–1192.

Mortensen, A. K., Egilson, Þ., Gautason, B., Árnadóttir, S., & Guðmundsson, Á.: Stratigraphy, alteration mineralogy, permeability and temperature conditions of well IDDP-1, Krafla, NE-Iceland. Geothermics, **49**, (2014), 31–41.

Oloruntobi, O., & Butt, S.: Energy-based formation pressure prediction. Journal of Petroleum Science and Engineering, **173**, (2019), 955–964.

Planke, S.: Geophysical response of flood basalts from analysis of wire line logs: Ocean Drilling Program Site 642, Vøring volcanic margin. Journal of Geophysical Research: Solid Earth, **99(B5),** (1994), 9279–9296.

Schön, J. H.: Physical properties of rocks: fundamentals and principles of petrophysics. Elsevier, (2015).

Siddig, O., Gamal, H., Elkatatny, S. et al. Real-time prediction of Poisson's ratio from drilling parameters using machine learning tools. Sci Rep **11, 12611** (2021).

Thordarson, T., & Larsen, G.: Volcanism in Iceland in historical time: Volcano types, eruption styles and eruptive history. Journal of Geodynamics, **43(1)**, (2007), 118–152.

Tsutsui, K., Holford, S., Schofield, N., Bunch, M., King, R., & McClay, K.: Petrophysical characteristics of igneous rocks in the outboard Browse Basin, North West Shelf of Australia: Implications for predicting igneous sequences prior to exploration drilling. Geological Society, London, Special Publications, **547(1),** (2024).

Villeneuve, M., Moayedi-Far, A.: Feature Optimization: Identifying Critical Parameters for TBM Performance Assessment, Chapter in: Performance of Tunnel Boring Machines (TBM) in Rock, eds. Hassanpour, J., Rostami, J. (in review).