



**Politecnico
di Torino**

Politecnico di Torino

Department of Environment, Land, and Infrastructure Engineering

Master of Science in Petroleum Engineering

A. y. 2022/2023

July 2023

**Development of Supervised Machine
Learning Models for the Prediction of
Well-Logs & Application on Wells at
São Francisco and Santos Basins,
Brazil**

Tutors:

Prof. Laura Valentina Socco

Mr. Gabriel Sarantopoulos Bergamaschi

Candidate:

Vittoria De Pellegrini

Abstract

Porosity assessment is essential for reservoir characterization. While laboratory measurements and conventional well-logs have traditionally been used to estimate porosity, they may not always provide accurate results in carbonate reservoirs, due to their extremely heterogeneous pore system. This is where nuclear magnetic resonance (NMR) tools offer a valuable solution. NMR logging technology allows for more accurate quantification of different porosity types, including total, effective, and free-fluid porosity. However, acquiring NMR logs can be costly and challenging due to factors such as the activation of wireline equipment, signal-to-noise ratio, environmental factors, and the properties of the formation fluid. To overcome these challenges and provide an alternative approach, we are interested in developing predictive models, using conventional well-logs as input data.

The objective of this research is to develop a Python code, from scratch, that implements supervised machine learning (ML) algorithms, specifically Random Forest (RF) and Gradient Boosting (GB), to build ML models for accurately predicting both NMR logs and various types of conventional well-logs. The code is available as an open source on GitHub under the repository named “Well-Logs_Predictive_Models” [https://github.com/VittoDePe98/Well-Logs_Predictive_Models.git]. This open accessibility encourages wider usage and collaboration among researchers. Two separate case studies are conducted to evaluate the functionality and effectiveness of the code. In both studies, the ML models are trained on a first well (training well) and tested on a second well (test well). The first case study focuses on the São Francisco onshore Brazilian basin. It serves as a preliminary exercise for model development and data familiarization. The goal is to predict two conventional well-logs: the calculated effective porosity and the measured compressional wave slowness logs. The second case study centers around the Santos offshore Brazilian basin, particularly the deep-water pre-salt carbonate reservoir area of the Itapu Oil Field. The attention of this research is primarily directed toward this second case study. The target is to predict high technological well-logs, including NMR total, effective, and free fluid porosity logs.

The results of the research demonstrate that both models are consistent and reliable, exhibiting low regression errors (MSE, RMSE, MAE) and high accuracy (R^2) values, in predicting the calculated effective porosity, for both training and test wells. However, when it comes to predicting the measured compressional wave slowness log, the models exhibit limitations and show poor performance on the test well. The limitations become even more apparent when predicting NMR porosity logs on the evaluation well. The models exhibit significantly reduced performance, yielding negative accuracy values.

Acknowledgments

First and foremost, I would like to extend my heartfelt appreciation to my supervisor, Prof. Laura Valentina Socco, and my co-supervisor, Mr. Gabriel Sarantopoulos Bergamaschi, for their support and invaluable mentorship throughout the course of this thesis.

I would like to express my gratitude to the Brazilian National Agency for Petroleum, Natural Gas, and Biofuels (ANP, Agência Nacional de Petróleo) for granting access to the dataset and authorizing its use in this research.

I am deeply grateful to my family members, both in Italy and Brazil, for their love and encouragement. Without their guidance, I would not be the person I am today.

Lastly, I would like to express my sincere appreciation to all the professors at the DIATI department for their dedication and commitment to providing quality education.

Contents

| | |
|---|----|
| Abstract | 1 |
| Acknowledgments | 2 |
| Contents | 3 |
| List of Figures | 5 |
| List of Equations | 11 |
| List of Tables..... | 12 |
| List of Abbreviations..... | 13 |
| 1 Introduction | 15 |
| 1.1 Background & Goal Definition | 15 |
| 1.2 Machine Learning Algorithms | 17 |
| 1.2.1 Random Forest Algorithm | 18 |
| 1.2.2 Gradient Boosting Algorithm..... | 19 |
| 2 Available Data & Methodology | 20 |
| 2.1 Available Well-Log Data..... | 20 |
| 2.1.1 Well-Log Data Information | 21 |
| 2.1.2 Well-Log Data Visualization..... | 22 |
| 2.1.3 Geological Info about São Francisco and Santos Basin..... | 27 |
| 2.1.4 General Information about the Itapu Oil Field, Santos Basin | 31 |
| 2.2 Applied Methodology..... | 32 |
| 2.3 Well-Log Data Collection | 35 |
| 2.4 Well-Log Data Pre-Processing | 35 |
| 2.5 Machine Learning Workflow | 36 |
| 2.5.1 Machine Learning Algorithm Selection | 36 |
| 2.5.1.1 Random Forest Algorithm Implementation | 36 |
| 2.5.1.2 Gradient Boosting Algorithm Implementation..... | 37 |

| | | |
|-------|---|-----|
| 2.5.2 | Well-Log Data Organization..... | 37 |
| 2.5.3 | Training Phase | 40 |
| 2.5.4 | Optimization Phase | 41 |
| 2.5.5 | Test Phase | 43 |
| 2.5.6 | Prediction Phase..... | 43 |
| 2.5.7 | Evaluation Phase | 43 |
| 3 | Methodology Application & Outcomes..... | 44 |
| 3.1 | Case Study 1: Effective Porosity Log..... | 44 |
| 3.2 | Case Study 1: Compressional Wave Slowness Log | 54 |
| 3.3 | Case Study 2: NMR Porosity Logs | 64 |
| 3.3.1 | First Attempt: Training on Well 3-BRSA-1215-RJS and Validation on Well 1-BRSA-1116-RJS..... | 66 |
| 3.3.2 | Second Attempt: Training on Well 1-BRSA-1116-RJS and Validation on Well 3-BRSA-1215-RJS | 79 |
| 4 | Discussion of the Results..... | 92 |
| 4.1 | Case Study 1: Effective Porosity Log..... | 93 |
| 4.2 | Case Study 1: Compressional Wave Slowness Log | 93 |
| 4.3 | Case Study 2: NMR Porosity Logs | 94 |
| 4.3.1 | Regression Metrics for the First Attempt..... | 94 |
| 4.3.2 | Regression Metrics for the Second Attempt | 99 |
| 5 | Conclusion..... | 104 |
| 6 | References | 105 |
| 7 | Appendixes..... | 108 |
| 7.1 | Building Machine Learning Models – Complete Workflow | 108 |

List of Figures

| | |
|---|----|
| Fig. 1.1: Schematic representation of the constituents of a rock..... | 15 |
| Fig. 1.2: Elements of a single decision tree. | 17 |
| Fig. 1.3: Example of a single decision tree..... | 17 |
| Fig. 1.4: Architecture of the RF Decision Tree. | 18 |
| Fig. 1.5: Architecture of the GB Decision Tree. | 19 |
| Fig. 2.1: Location of well 1-BRSA-871-MG (deepskyblue), and well 1-BRSA-948-MG (orange) for São Francisco Basin. The borders of the Brazilian sub-regions are highlighted in red ('Brazilian Sub-Regions Coordinates', 2023). | 20 |
| Fig. 2.2: Location of well 1-BRSA-1116-RJS (deepskyblue), and well 3-BRSA-1215-RJS (orange) for Santos Basin. The borders of the Brazilian states are highlighted in magenta ('Brazilian States Coordinates', 2023). | 20 |
| Fig. 2.3: Composite well-log of well 1-BRSA-871-MG. Track 1: Measured well depth. Track 2: Gamma-Ray (GR) log. Track 3: Induction Electric Resistivity logs. Investigation depth of 90 (AT90) inches. Track 4: Formation Density (RHOZ), and Neutron Porosity (NPHI) logs. Track 5: Compressional Wave Slowness (DTCO) log. | 23 |
| Fig. 2.4: Composite well-log of well 1-BRSA-948-MG. Track 1: Measured well depth. Track 2: Gamma-Ray (GR) log. Track 3: Neutron Porosity (NPHI) log. Track 4: Compressional Wave Slowness (DTCO) log..... | 24 |
| Fig. 2.5: Composite well-log of well 1-BRSA-1116-RJS. Track 1: Measured well depth. Track 2: Gamma-Ray (GR) and Caliper (HCAL) logs. Track 3: Induction Electric Resistivity logs. Investigation depths of 10 (AT10), 30 (AT30), and 90 (AT90) inches. Track 4: Formation Density (RHOZ), Neutron Porosity (NPHI), and Photoelectric Factor (PEFZ) logs. Track 5: Compressional Wave Slowness (DTCO) and Shear Wave Slowness (DTSM) logs. Track 6: Nuclear Magnetic Resonance Porosity logs. Total Porosity (nmrPHIT), Effective Porosity (nmrPHIE), and Free Fluid (nmrFF). In deepskyblue, the reservoir rock. | 25 |
| Fig. 2.6: Composite well-log of well 3-BRSA-1215-RJS. Track 1: Measured well depth. Track 2: Gamma-Ray (GR) and Caliper (HCAL) logs. Track 3: Induction Electric Resistivity logs. Investigation depths of 10 (AT10), 30 (AT30), and 90 (AT90) inches. Track 4: Formation Density (RHOZ), Neutron Porosity (NPHI), and Photoelectric Factor (PEFZ) logs. Track 5: Compressional Wave Slowness (DTCO) log. Track 6: Nuclear Magnetic Resonance Porosity logs. Total Porosity (nmrPHIT), Effective Porosity (nmrPHIE), and Free Fluid (nmrFF). In deepskyblue, the reservoir rock..... | 26 |
| Fig. 2.7: Pre-salt oil and gas layer, salt layer, post-salt layer, and water depth for the Santos basin (PETROBRAS)..... | 28 |
| Fig. 2.8: Schematic lithological columns for well 1-BRSA-871-MG (first column), well 1-BRSA-948-MG (second column), well 1-BRSA-1116-RJS (third column), and well 3-BRSA-1215-RJS (fourth column) ('ANP-TERRESTRE', 2023)..... | 30 |
| Fig. 2.9: Location map of Santos basin and the main pre-salt fields. Itapu field is highlighted in red (Equinor, 2017). | 31 |
| Fig. 2.10: Schematic diagram showing the research methodology applied in this study. | 32 |

| | |
|--|----|
| Fig. 2.11 Scheme of the available wells and predicted well-logs for each basin. | 33 |
| Fig. 2.12: Notebooks present within the GitHub repository..... | 35 |
| Fig. 2.13: Schematic representation of the decision process occurring within the RF algorithm. | 36 |
| Fig. 2.14: Random Shuffling of the original dataset performed during Training and Test Split | 40 |
| Fig. 2.15: k-Fold Cross-Validation, where k=10. | 41 |
| Fig. 2.16: Random Search Grid | 42 |
| Fig. 3.1: Scatter plots of predicted versus measured effective porosity (m^3/m^3), for the RF and GB Models, applied to the Test Dataset of well 1-BRSA-871-MG. | 45 |
| Fig. 3.2: Comparing the match between the predicted and measured effective porosity, for the RF Model, across the Test Dataset Index of well 1-BRSA-871-MG. Plot 1: Predicted and Measured Effective Porosity (m^3/m^3) for the range 0 - 1900. Plot 2: Predicted and Measured Effective Porosity (m^3/m^3) for the range 500 - 750..... | 46 |
| Fig. 3.3: Comparing the match between the predicted and measured effective porosity, for the GB Model, across the Test Dataset Index of well 1-BRSA-871-MG. Plot 1: Predicted and Measured Effective Porosity (m^3/m^3) for the range 0 - 1900. Plot 2: Predicted and Measured Effective Porosity (m^3/m^3) for the range 500 - 750..... | 47 |
| Fig. 3.4: Scatter plots of predicted versus measured effective porosity (m^3/m^3), for the RF and GB Models, applied to the Entire Original Dataset of well 1-BRSA-871-MG. | 48 |
| Fig. 3.5: Comparing the match between the predicted and measured effective porosity, for the RF Model, across the Entire Original Dataset Index of well 1-BRSA-871-MG, which corresponds to the Measured Well Depth. Plot 1: Predicted and Measured Effective Porosity (m^3/m^3) for the range 0 - 9496. Plot 2: Predicted and Measured Effective Porosity (m^3/m^3) for the range 5000 - 6000. | 49 |
| Fig. 3.6: Comparing the match between the predicted and measured effective porosity, for the GB Model, across the Entire Original Dataset Index of well 1-BRSA-871-MG, which corresponds to the Measured Well Depth. Plot 1: Predicted and Measured Effective Porosity (m^3/m^3) for the range 0 - 9496. Plot 2: Predicted and Measured Effective Porosity (m^3/m^3) for the range 5000 - 6000. | 50 |
| Fig. 3.7: Scatter plots of predicted versus measured effective porosity (m^3/m^3), for the RF and GB Models, applied to the Entire Original Dataset of well 1-BRSA-948-MG. | 51 |
| Fig. 3.8: Comparing the match between the predicted and measured effective porosity, for the RF Model, across the Entire Original Dataset Index of well 1-BRSA-948-MG, which corresponds to the Measured Well Depth..... | 52 |
| Fig. 3.9: Comparing the match between the predicted and measured effective porosity, for the GB Model, across the Entire Original Dataset Index of well 1-BRSA-948-MG, which corresponds to the Measured Well Depth..... | 53 |
| Fig. 3.10: Scatter plots of predicted versus measured compressional wave slowness ($\mu s/ft$), for the RF and GB Models, applied to the Test Dataset of well 1-BRSA-871-MG. | 55 |
| Fig. 3.11: Comparing the match between the predicted and measured compressional wave slowness ($\mu s/ft$), for the RF Model, across the Test Dataset Index of well 1-BRSA-871-MG.... | 56 |

| | |
|---|----|
| Fig. 3.12: Comparing the match between the predicted and measured compressional wave slowness ($\mu\text{s}/\text{ft}$), for the GB Model, across the Test Dataset Index of well 1-BRSA-871-MG... | 57 |
| Fig. 3.13: Scatter plots of predicted versus measured compressional wave slowness ($\mu\text{s}/\text{ft}$), for the RF and GB Models, applied to the Entire Original Dataset of well 1-BRSA-871-MG. | 58 |
| Fig. 3.14: Comparing the match between the predicted and measured compressional wave slowness ($\mu\text{s}/\text{ft}$), for the RF Model, across the Entire Original Dataset Index of well 1-BRSA-871-MG..... | 59 |
| Fig. 3.15: Comparing the match between the predicted and measured compressional wave slowness ($\mu\text{s}/\text{ft}$), for the RF Model, across the Entire Original Dataset Index of well 1-BRSA-871-MG..... | 60 |
| Fig. 3.16: Scatter plots of predicted versus measured compressional wave slowness ($\mu\text{s}/\text{ft}$), for the RF and GB Models, applied to the Entire Original Dataset of well 1-BRSA-948-MG. | 61 |
| Fig. 3.17: Comparing the match between the predicted and measured compressional wave slowness ($\mu\text{s}/\text{ft}$), for the RF Model, across the Entire Original Dataset Index of well 1-BRSA-948-MG..... | 62 |
| Fig. 3.18: Comparing the match between the predicted and measured compressional wave slowness ($\mu\text{s}/\text{ft}$), for the RF Model, across the Entire Original Dataset Index of well 1-BRSA-948-MG..... | 63 |
| Fig. 3.19: NMR Porosity distribution for well 1-BRSA-1116-RJS, and well 3-BRSA-1215-RJS. Histogram 1: NMR Total Porosity distribution (m^3/m^3). Histogram 2: NMR Effective Porosity distribution (m^3/m^3). Histogram 3: NMR Free Fluid distribution (m^3/m^3)..... | 65 |
| Fig. 3.20: Scatter plots of predicted versus measured NMR porosity, for the RF Model, applied to the Test Dataset of well 3-BRSA-1215-RJS. Plot 1: NMR Effective Porosity (m^3/m^3). Plot 2: NMR Free Fluid (m^3/m^3). Plot 3: NMR Total Porosity (m^3/m^3). | 67 |
| Fig. 3.21: Comparing the match between the predicted and measured NMR porosity, for the RF Model, across the Test Dataset Index of well 3-BRSA-1215-RJS. Track 1: Predicted and Measured NMR Effective Porosity (m^3/m^3). Track 2: Predicted and Measured NMR Free Fluid (m^3/m^3). Track 3: Predicted and Measured NMR Total Porosity (m^3/m^3). | 68 |
| Fig. 3.22: Scatter plots of predicted versus measured NMR porosity, for the GB Model, applied to the Test Dataset of well 3-BRSA-1215-RJS. Plot 1: NMR Effective Porosity (m^3/m^3). Plot 2: NMR Free Fluid (m^3/m^3). Plot 3: NMR Total Porosity (m^3/m^3). | 69 |
| Fig. 3.23: Comparing the match between the predicted and measured NMR porosity, for the GB Model, across the Test Dataset Index of well 3-BRSA-1215-RJS. Track 1: Predicted and Measured NMR Effective Porosity (m^3/m^3). Track 2: Predicted and Measured NMR Free Fluid (m^3/m^3). Track 3: Predicted and Measured NMR Total Porosity (m^3/m^3). | 70 |
| Fig. 3.24: Scatter plots of predicted versus measured NMR porosity, for the RF Model, applied to the Entire Original Dataset of well 3-BRSA-1215-RJS. Plot 1: NMR Effective Porosity (m^3/m^3). Plot 2: NMR Free Fluid (m^3/m^3). Plot 3: NMR Total Porosity (m^3/m^3). | 71 |
| Fig. 3.25: Comparing the match between the predicted and measured NMR porosity, for the RF Model, across the Entire Original Dataset Index of well 3-BRSA-1215-RJS, which corresponds to the Measured Well Depth. Track 1: Predicted and Measured NMR Effective Porosity | |

| | |
|---|----|
| (m ³ /m ³). Track 2: Predicted and Measured NMR Free Fluid (m ³ /m ³). Track 3: Predicted and Measured NMR Total Porosity (m ³ /m ³)..... | 72 |
| Fig. 3.26: Scatter plots of predicted versus measured NMR porosity, for the GB Model, applied to the Entire Original Dataset of well 3-BRSA-1215-RJS. Plot 1: NMR Effective Porosity (m ³ /m ³). Plot 2: NMR Free Fluid (m ³ /m ³). Plot 3: NMR Total Porosity (m ³ /m ³)..... | 73 |
| Fig. 3.27: Comparing the match between the predicted and measured NMR porosity, for the GB Model, across the Entire Original Dataset Index of well 3-BRSA-1215-RJS, which corresponds to the Measured Well Depth. Track 1: Predicted and Measured NMR Effective Porosity (m ³ /m ³). Track 2: Predicted and Measured NMR Free Fluid (m ³ /m ³). Track 3: Predicted and Measured NMR Total Porosity (m ³ /m ³)..... | 74 |
| Fig. 3.28: Scatter plots of predicted versus measured NMR porosity, for the RF Model, applied to the Entire Original Dataset of well 1-BRSA-1116-RJS. Plot 1: NMR Effective Porosity (m ³ /m ³). Plot 2: NMR Free Fluid (m ³ /m ³). Plot 3: NMR Total Porosity (m ³ /m ³)..... | 75 |
| Fig. 3.29: Comparing the match between the predicted and measured NMR porosity, for the RF Model, across the Entire Original Dataset Index of well 1-BRSA-1116-RJS, which corresponds to the Measured Well Depth. Track 1: Predicted and Measured NMR Effective Porosity (m ³ /m ³). Track 2: Predicted and Measured NMR Free Fluid (m ³ /m ³). Track 3: Predicted and Measured NMR Total Porosity (m ³ /m ³)..... | 76 |
| Fig. 3.30: Scatter plots of predicted versus measured NMR porosity, for the GB Model, applied to the Entire Original Dataset of well 1-BRSA-1116-RJS. Plot 1: NMR Effective Porosity (m ³ /m ³). Plot 2: NMR Free Fluid (m ³ /m ³). Plot 3: NMR Total Porosity (m ³ /m ³)..... | 77 |
| Fig. 3.31: Comparing the match between the predicted and measured NMR porosity, for the GB Model, across the Entire Original Dataset Index of well 1-BRSA-1116-RJS, which corresponds to the Measured Well Depth. Track 1: Predicted and Measured NMR Effective Porosity (m ³ /m ³). Track 2: Predicted and Measured NMR Free Fluid (m ³ /m ³). Track 3: Predicted and Measured NMR Total Porosity (m ³ /m ³)..... | 78 |
| Fig. 3.32: Scatter plots of predicted versus measured NMR porosity, for the RF Model, applied to the Test Dataset of well 1-BRSA-1116-RJS. Plot 1: NMR Effective Porosity (m ³ /m ³). Plot 2: NMR Free Fluid (m ³ /m ³). Plot 3: NMR Total Porosity (m ³ /m ³). | 80 |
| Fig. 3.33: Comparing the match between the predicted and measured NMR porosity, for the RF Model, across the Test Dataset Index of well 1-BRSA-1116-RJS. Track 1: Predicted and Measured NMR Effective Porosity (m ³ /m ³). Track 2: Predicted and Measured NMR Free Fluid (m ³ /m ³). Track 3: Predicted and Measured NMR Total Porosity (m ³ /m ³). | 81 |
| Fig. 3.34: Scatter plots of predicted versus measured NMR porosity, for the GB Model, applied to the Test Dataset of well 1-BRSA-1116-RJS. Plot 1: NMR Effective Porosity (m ³ /m ³). Plot 2: NMR Free Fluid (m ³ /m ³). Plot 3: NMR Total Porosity (m ³ /m ³). | 82 |
| Fig. 3.35: Comparing the match between the predicted and measured NMR porosity, for the RF Model, across the Test Dataset Index of well 1-BRSA-1116-RJS. Track 1: Predicted and Measured NMR Effective Porosity (m ³ /m ³). Track 2: Predicted and Measured NMR Free Fluid (m ³ /m ³). Track 3: Predicted and Measured NMR Total Porosity (m ³ /m ³). | 83 |

| | |
|---|----|
| Fig. 3.36: Scatter plots of predicted versus measured NMR porosity, for the RF Model, applied to the Entire Original Dataset of well 1-BRSA-1116-RJS. Plot 1: NMR Effective Porosity (m^3/m^3). Plot 2: NMR Free Fluid (m^3/m^3). Plot 3: NMR Total Porosity (m^3/m^3)..... | 84 |
| Fig. 3.37: Comparing the match between the predicted and measured NMR porosity, for the RF Model, across the Entire Original Dataset Index of well 1-BRSA-1116-RJS, which corresponds to the Measured Well Depth. Track 1: Predicted and Measured NMR Effective Porosity (m^3/m^3). Track 2: Predicted and Measured NMR Free Fluid (m^3/m^3). Track 3: Predicted and Measured NMR Total Porosity (m^3/m^3)..... | 85 |
| Fig. 3.38: Scatter plots of predicted versus measured NMR porosity, for the GB Model, applied to the Entire Original Dataset of well 1-BRSA-1116-RJS. Plot 1: NMR Effective Porosity (m^3/m^3). Plot 2: NMR Free Fluid (m^3/m^3). Plot 3: NMR Total Porosity (m^3/m^3)..... | 86 |
| Fig. 3.39: Comparing the match between the predicted and measured NMR porosity, for the GB Model, across the Entire Original Dataset Index of well 1-BRSA-1116-RJS, which corresponds to the Measured Well Depth. Track 1: Predicted and Measured NMR Effective Porosity (m^3/m^3). Track 2: Predicted and Measured NMR Free Fluid (m^3/m^3). Track 3: Predicted and Measured NMR Total Porosity (m^3/m^3)..... | 87 |
| Fig. 3.40: Scatter plots of predicted versus measured NMR porosity, for the RF Model, applied to the Entire Original Dataset of well 3-BRSA-1215-RJS. Plot 1: NMR Effective Porosity (m^3/m^3). Plot 2: NMR Free Fluid (m^3/m^3). Plot 3: NMR Total Porosity (m^3/m^3)..... | 88 |
| Fig. 3.41: Comparing the match between the predicted and measured NMR porosity, for the RF Model, across the Entire Original Dataset Index of well 3-BRSA-1215-RJS, which corresponds to the Measured Well Depth. Track 1: Predicted and Measured NMR Effective Porosity (m^3/m^3). Track 2: Predicted and Measured NMR Free Fluid (m^3/m^3). Track 3: Predicted and Measured NMR Total Porosity (m^3/m^3)..... | 89 |
| Fig. 3.42: Scatter plots of predicted versus measured NMR porosity, for the GB Model, applied to the Entire Original Dataset of well 3-BRSA-1215-RJS. Plot 1: NMR Effective Porosity (m^3/m^3). Plot 2: NMR Free Fluid (m^3/m^3). Plot 3: NMR Total Porosity (m^3/m^3)..... | 90 |
| Fig. 3.43: Comparing the match between the predicted and measured NMR porosity, for the GB Model, across the Entire Original Dataset Index of well 3-BRSA-1215-RJS, which corresponds to the Measured Well Depth. Track 1: Predicted and Measured NMR Effective Porosity (m^3/m^3). Track 2: Predicted and Measured NMR Free Fluid (m^3/m^3). Track 3: Predicted and Measured NMR Total Porosity (m^3/m^3)..... | 91 |
| Fig. 4.1: Regression Metrics for RF and GB Models. Root Mean Squared Error (RMSE), Mean Squared Error (MSE), and Mean Absolute Error (MAE). Bar Plot 1: NMR Effective Porosity (m^3/m^3). Bar Plot 2: NMR Free Fluid (m^3/m^3). Bar Plot 3: NMR Total Porosity (m^3/m^3)..... | 95 |
| Fig. 4.2: Feature Importance for RF Model (first attempt). Bar plot 1: feature importance scores for NMR Effective Porosity (m^3/m^3). Bar plot 2: feature importance scores for NMR Free Fluid (m^3/m^3). Bar plot 3: feature importance scores for NMR Total Porosity (m^3/m^3)..... | 97 |
| Fig. 4.3: Feature Importance for GB Model (first attempt). Bar plot 1: feature importance scores for NMR Effective Porosity (m^3/m^3). Bar plot 2: feature importance scores for NMR Free Fluid (m^3/m^3). Bar plot 3: feature importance scores for NMR Total Porosity (m^3/m^3)..... | 98 |

Fig. 4.4: Feature Importance for RF Model (second attempt). Bar plot 1: feature importance scores for NMR Effective Porosity (m^3/m^3). Bar plot 2: feature importance scores for NMR Free Fluid (m^3/m^3). Bar plot 3: feature importance scores for NMR Total Porosity (m^3/m^3).....101

Fig. 4.5: Feature Importance for GB Model (second attempt). Bar plot 1: feature importance scores for NMR Effective Porosity (m^3/m^3). Bar plot 2: feature importance scores for NMR Free Fluid (m^3/m^3). Bar plot 3: feature importance scores for NMR Total Porosity (m^3/m^3).....102

List of Equations

| | |
|--------------|----|
| Eq. 2.1..... | 38 |
| Eq. 4.1..... | 92 |
| Eq. 4.2..... | 92 |
| Eq. 4.3..... | 92 |
| Eq. 4.4..... | 92 |

List of Tables

| | |
|--|----|
| Table 2.1: Available measured well-log data for each well (São Francisco basin)..... | 21 |
| Table 2.2: Available measured well-log data for each well (Santos basin)..... | 21 |
| Table 2.3: Lithology tables for São Francisco and Santos basins..... | 31 |
| Table 2.4: Predictors and Outputs for each well, with the corresponding lithology (São Francisco basin)..... | 38 |
| Table 2.5: Predictors and Outputs for each well, with the corresponding lithology (São Francisco basin)..... | 38 |
| Table 2.6: Predictors and Outputs for each well, with the corresponding lithology (Santos basin)..... | 39 |
| Table 2.7: Abbreviations and Extended Log Names..... | 39 |
| Table 2.8: Optimal couple for Effective Porosity Prediction for both RF and GB (São Francisco basin)..... | 43 |
| Table 2.9: Optimal couple for Compressional Wave Slowness Prediction for both RF and GB (São Francisco basin)..... | 43 |
| Table 2.10: Optimal couple for NMR Porosities Prediction for both RF and GB (Santos basin)..... | 43 |
| Table 4.1: Regression Metrics of RF and GB Models for estimation of Effective Porosity on the Test Dataset of well 1-BRSA-871-MG..... | 93 |
| Table 4.2: Regression Metrics of RF and GB Models for estimation of Effective Porosity on the Entire Original Dataset of well 1-BRSA-948-MG..... | 93 |
| Table 4.3: Regression Metrics of RF and GB Models for estimation of Compressional Wave Slowness on the Test Dataset of well 1-BRSA-871-MG..... | 93 |
| Table 4.4: Regression Metrics of RF and GB Models for estimation of Compressional Wave Slowness on the Entire Original Dataset of well 1-BRSA-948-MG..... | 93 |
| Table 4.5: Regression Metrics of RF and GB Models for estimation of (NMR) Total Porosity, Effective Porosity, and Free Fluid on the Test Dataset of well 3-BRSA-1215-RJS..... | 94 |
| Table 4.6: Regression Metrics of RF and GB Models for estimation of (NMR) Total Porosity, Effective Porosity, and Free Fluid on the Entire Original Dataset of well 1-BRSA-1116-RJS..... | 94 |
| Table 4.7: Regression Metrics of RF and GB Models for estimation of (NMR) Total Porosity, Effective Porosity, and Free Fluid on the Test Dataset of well 1-BRSA-1116-RJS..... | 99 |
| Table 4.8: Regression Metrics of RF and GB Models for estimation of (NMR) Total Porosity, Effective Porosity, and Free Fluid on the Entire Original Dataset of well 3-BRSA-1215-RJS..... | 99 |

List of Abbreviations

1. **AdaBoost**: Adaptive Boosting
2. **ANN**: Artificial Neural Network
3. **AT10**: Array induction resistivity log, investigation 10 inches (ohm.m)
4. **AT30**: Array induction resistivity log, investigation 30 inches (ohm.m)
5. **AT90**: Array induction resistivity log, investigation 90 inches (ohm.m)
6. **BVI**: Bulk volume immovable (m^3/m^3)
7. **BVM**: Bulk volume movable (m^3/m^3)
8. **CastBoost**: Categorical Boosting
9. **CBW**: Clay-bound water (m^3/m^3)
10. **DTCO**: Delta T compressional wave, compressional slowness log ($\mu s/ft$)
11. **DTSM**: Delta S shear wave, shear slowness log ($\mu s/ft$)
12. **ESPEC**: Undefined elemental spectroscopy log (-)
13. **FL**: Fuzzy Logic
14. **GB**: Gradient Boosting
15. **GR**: Gamma-ray log (API)
16. **HCAL**: Caliper log. Borehole diameter (inches)
17. **MAE**: Mean absolute error (m^3/m^3)
18. **ML**: Machine Learning
19. **MLPNN**: Multilayer Perceptron Neural Network
20. **MSE**: Mean squared error (m^3/m^3)²
21. **NMR**: Nuclear magnetic resonance
22. **NMRFF**: Nuclear magnetic resonance free fluid log (m^3/m^3)
23. **NMRPHIE**: Nuclear magnetic resonance effective porosity log (m^3/m^3)
24. **NMRPHIT**: Nuclear magnetic resonance total porosity log (m^3/m^3)
25. **NPFI**: Thermal neutron porosity log (m^3/m^3)
26. **PEFZ**: Photo-electric log (unitless)
27. **PHIE_HILT**: HILT effective porosity (m^3/m^3)
28. **R²**: Coefficient of determination (%)
29. **RES**: Undefined resistivity log (ohm.m)
30. **RF**: Random Forest

31. **RHGX_HILT**: HILT Grain Density (g/cm^3)
32. **RHOZ**: Formation density log (g/cm^3)
33. **RMSE**: Root mean squared error (m^3/m^3)
34. **XGBoost**: Extreme Gradient Boosting

1 Introduction

1.1 Background & Goal Definition

Porosity is a crucial feature in reservoir characterization and plays a significant role in various aspects of hydrocarbon exploration and production. It is commonly used to estimate the volume of hydrocarbons present in a reservoir and evaluate the productive potential of a well, etc. (Mustafa *et al.*, 2023). In the context of porosity, the term “total porosity” encompasses the combined contribution of different components within a rock formation. This includes clay-bound water, capillary-bound water, free water, and hydrocarbon. Fig. 1.1 likely provides a graphical representation demonstrating the various components contributing to total porosity.

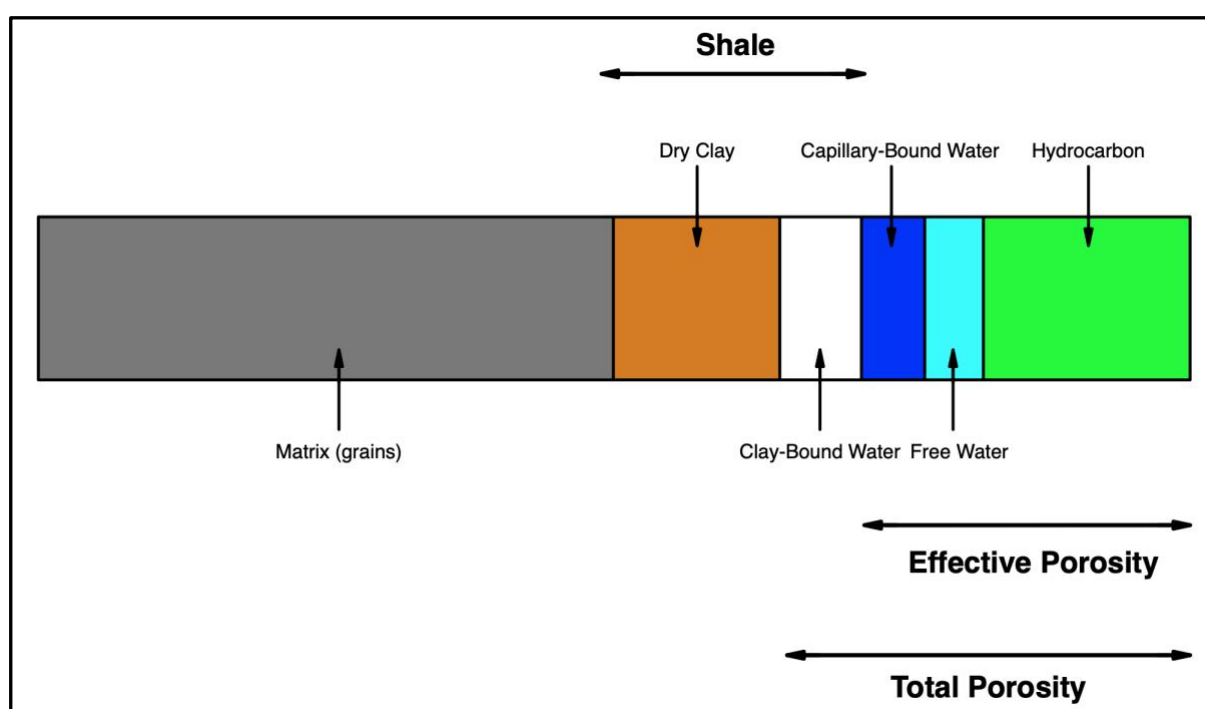


Fig. 1.1: Schematic representation of the constituents of a rock.

Conventional logging tools, including neutron porosity, density, and sonic logs, are commonly employed for total porosity estimation. However, estimating total porosity in carbonate reservoirs poses challenges due to the complex and diverse nature of their porous systems, influenced by various geological processes (Rocha *et al.*, 2019). Indeed, traditional porosity tools are less accurate in carbonate reservoirs, as lithology strongly affects their measurements (Farmanov *et al.*, 2023).

Among available techniques, the nuclear magnetic resonance (NMR) tools have been proven to be the most reliable and accurate method for total porosity determination, especially in carbonate reservoirs. Unlike conventional tools, NMR is not influenced by lithology and quantifies total porosity by detecting the response of the hydrogen nuclei in fluids (water and hydrocarbons), under an applied artificial magnetic field (Mustafa *et al.*, 2023).

However, acquiring NMR well-logs can be challenging, time-consuming, and expensive, leading to their limited availability for all drilled wells. As NMR measurements provide valuable insights, machine learning (ML) models can serve as powerful tools for predicting NMR porosity in carbonate reservoirs where direct NMR logs are not obtained (Tamoto, Gioria and Carneiro, 2023). By utilizing a single conventional well-log dataset, the benefits of NMR porosity determination can be extended to a broader range of wells within the reservoir, even without direct NMR measurements.

The widespread distribution of carbonate rocks across various regions of Brazil, such as Central-West, North, Southeastern, and Northeast Brazil (Pinheiro Junior *et al.*, 2021), presents challenges in acquiring direct NMR well-logs. However, recent studies have showcased the success of employing ML techniques to generate synthetic NMR logs for these carbonate areas. For instance, (González Carrasquilla and Tapia Briones, 2019) developed unsupervised and supervised ML models, specifically the Fuzzy Logic (FL), and Artificial Neural Network (ANN), for the offshore carbonate area of Campos Basin in Southeastern Brazil. These models utilized conventional well-logs as input data and were trained on a consistent dataset and tested on a single well. Their findings showed that the ANN model outperformed the FL model, highlighting the utility of this alternative methodology when direct NMR well-logs are unavailable.

Another study by (Tamoto, Gioria and Carneiro, 2023) focused on the offshore carbonate area of Santos Basin, also in Southeastern Brazil. They developed four supervised ML models, namely Multilayer Perceptron Neural Network (MLPNN), Adaptive Boosting (AdaBoost), Extreme Gradient Boosting (XGBoost), and Categorical Boosting (CatBoost), employing conventional well-logs as input data. In this case, the ML models were trained on a wide training dataset of over 30,000 data points and tested on three different wells. Remarkably, these ML models achieved impressive results, with differences between real NMR well-logs and the ML outputs being less than 5% for most of the well-logging interval. This research further emphasizes the significance of supervised ML predictive models in predicting NMR porosities for carbonate zones.

Building upon these previous studies, the objective of this research is to develop supervised ML models, starting from scratch, to predict NMR well-logs in the Santos Basin, utilizing conventional well-logs as input data. Given the complexity of predicting NMR well-logs in carbonate areas, the models are initially developed to predict conventional well-logs for the non-carbonate area of the São Francisco Basin, Brazil. Consequently, two separate case studies are conducted. Case study 1 serves as a preliminary exercise to familiarize with the functionality of these models. The focus is on predicting two conventional well-logs: the calculated effective porosity and measured compressional wave slowness logs. Case study 2 represents the primary application of the developed models. The target is to predict NMR total, effective, and free fluid porosity logs for the pre-salt carbonate area of the Itapu Oil Field, in the Santos Basin, Brazil. By conducting these two case studies, the research aims to demonstrate the effectiveness and applicability of the supervised ML models in predicting both conventional well-logs and NMR well-logs in carbonate zones.

1.2 Machine Learning Algorithms

ML algorithms are mainly divided into four categories: supervised learning, unsupervised learning, semi-supervised learning, and reinforcement learning (Sarker, 2021). As already mentioned, we are interested in the supervised learning category, and we briefly discuss its scope. Supervised learning learns a function that correlates an input to an output, and it uses labeled training data. The two most typical supervised tasks are classification, when the target variable is discrete, and regression when it is continuous (Sarker, 2021). In this study, the regression problem is of our interest.

The Ensemble-Based Decision Tree algorithms belong to the family of supervised learning (Farmanov *et al.*, 2023). The two most popular ensemble learning algorithms are bagging, which means “to combine”, and boosting, which means “to improve”, regressors. An example of bagged and boosted decision tree algorithms is given by the Random Forest (RF) and the Gradient Boosting (GB), respectively (Mahesh, 2020). Although the mentioned previous studies have not explored these specific algorithms in the context of predicting NMR well-logs, our research focuses on evaluating their capability in this domain.

Before describing the RF and GB algorithms, we briefly recap the concept of the Decision Tree. A Decision Tree is a tree-like graph that shows decisions and their outcomes. The elements of a decision tree are decision nodes, leaf nodes, and branches. The decision node represents a choice; the leaf node gives an outcome or decision, and a branch is an alternative. In Fig. 1.2 and Fig. 1.3 we provide a schematic representation of a single decision tree:

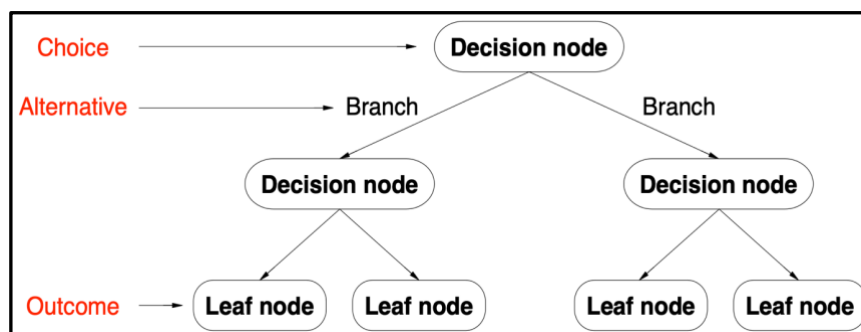


Fig. 1.2: Elements of a single decision tree.

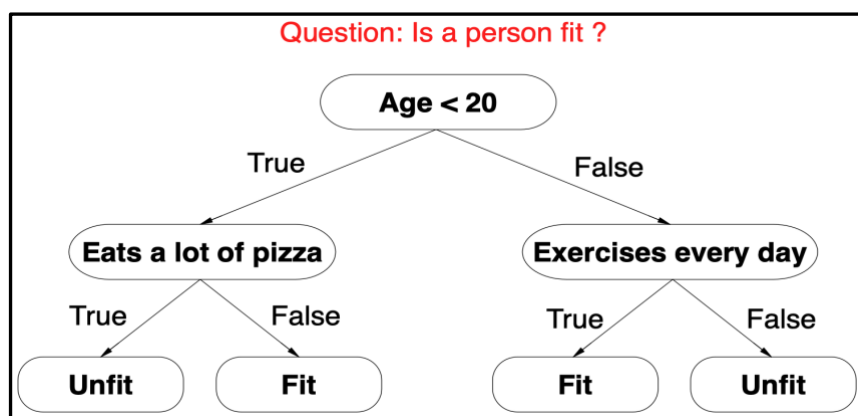


Fig. 1.3: Example of a single decision tree.

1.2.1 Random Forest Algorithm

The mechanism of the RF method involves building, simultaneously, N -independent decision trees, where the data used to build each tree, and the features selected at each node are randomly chosen. For regression problems, the output, which is our prediction, is given by averaging the results from each regression tree (Farmanov *et al.*, 2023). The final RF model is an average model. As follows, a schematic representation of the RF Decision Tree (Fig. 1.4). The dataset is the training dataset.

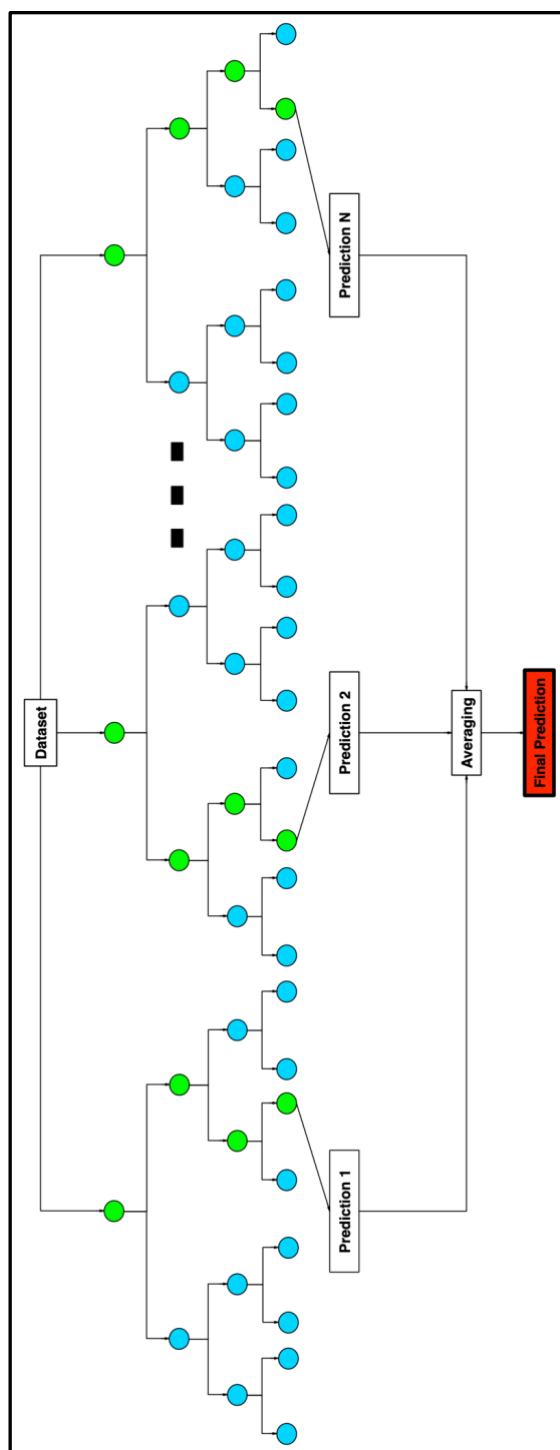


Fig. 1.4: Architecture of the RF Decision Tree.

1.2.2 Gradient Boosting Algorithm

Unlike the RF method, the GB technique builds N-dependent regression trees, called the weak learners, additively. This means that each tree is built one after another (Mahesh, 2020). The general idea is that the subsequent models should be able to correct the mistakes committed by the older ones, which means minimizing their errors (Tamoto, Gioria and Carneiro, 2023). As follows, a schematic representation of the GB Decision Tree (Fig. 1.5):

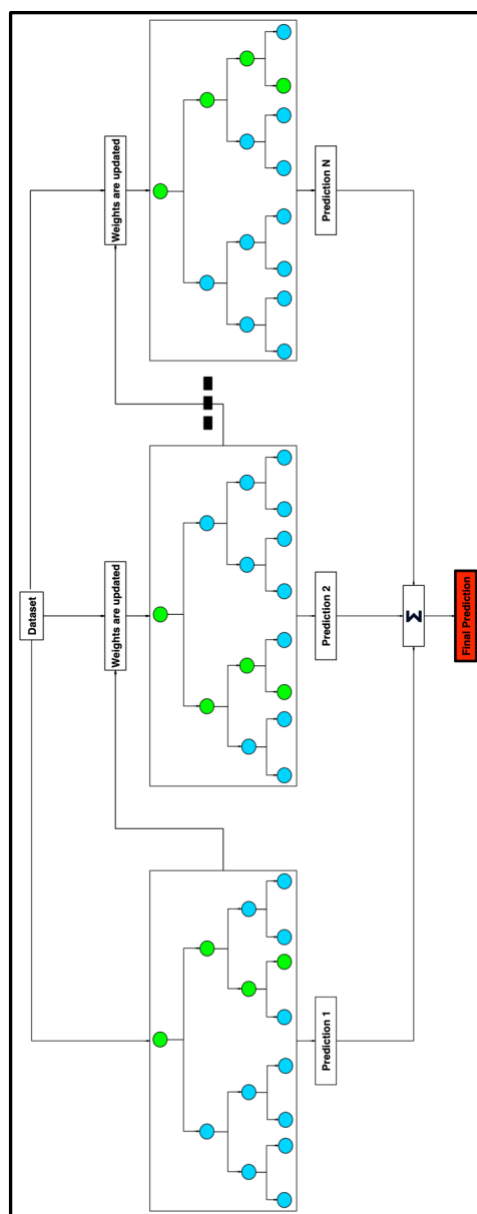


Fig. 1.5: Architecture of the GB Decision Tree.

From Fig. 1.5 we can understand that, at each iteration, the values of the weights applied to each of the input variables, used to predict the target variable, are adjusted. The final prediction is the sum of all the results coming from each weak learner. The final GB model is a strong learner model.

2 Available Data & Methodology

2.1 Available Well-Log Data

In Fig. 2.1 and Fig. 2.2, we display the geographical mapping of the wells of interest, providing a visual representation of their locations.

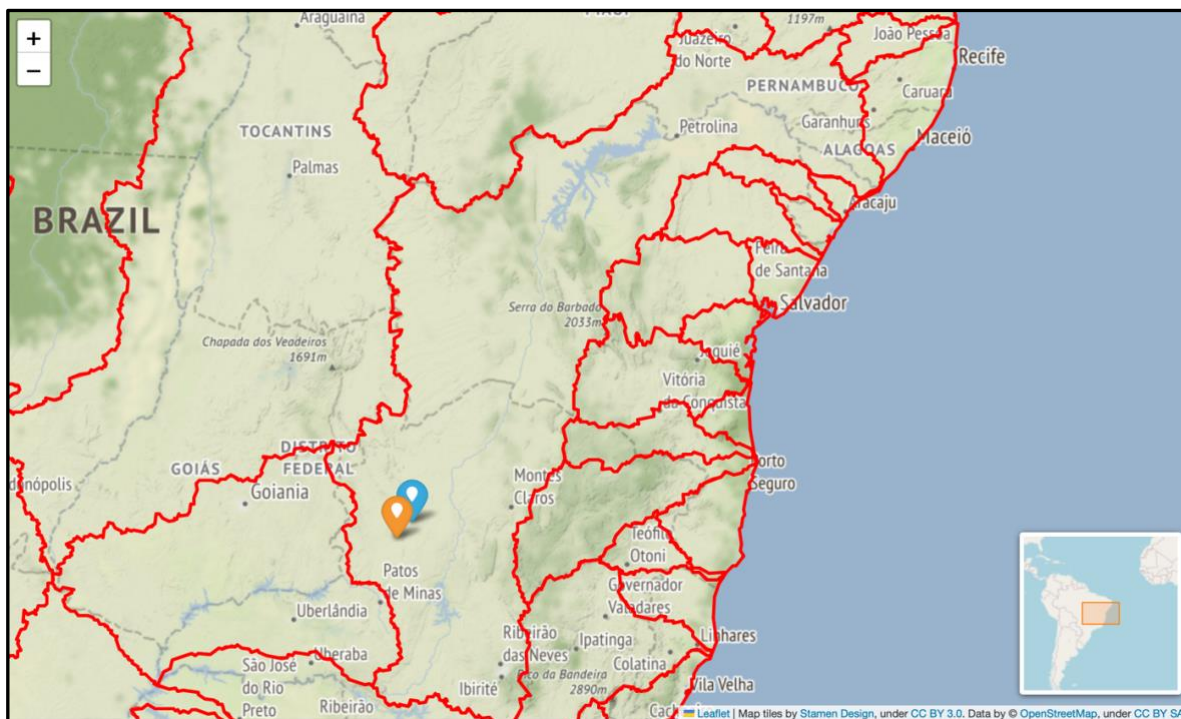


Fig. 2.1: Location of well 1-BRSA-871-MG (deepskyblue), and well 1-BRSA-948-MG (orange) for São Francisco Basin. The borders of the Brazilian sub-regions are highlighted in red ('Brazilian Sub-Regions Coordinates', 2023).

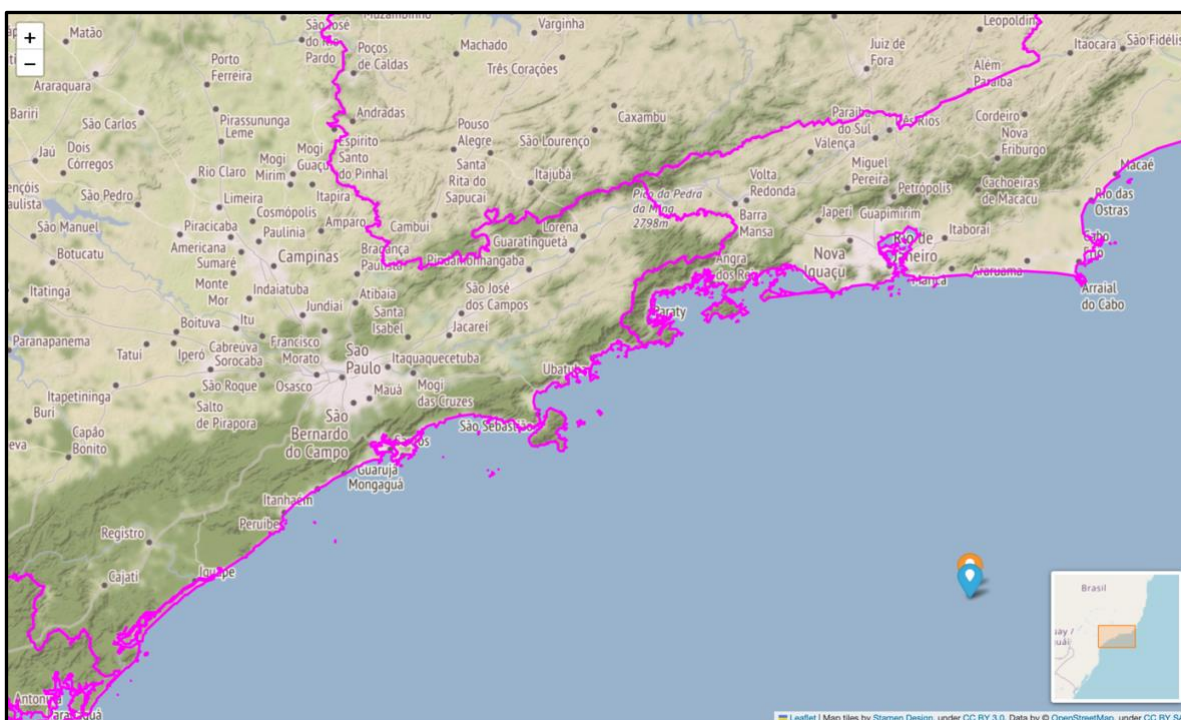


Fig. 2.2: Location of well 1-BRSA-1116-RJS (deepskyblue), and well 3-BRSA-1215-RJS (orange) for Santos Basin. The borders of the Brazilian states are highlighted in magenta ('Brazilian States Coordinates', 2023).

The available well-log data for the São Francisco basin (onshore), and Santos basin (offshore) are provided by the ANP (Agência Nacional de Petróleo). However, while the onshore well-log data are publicly published, the offshore well-log data are delivered physically in hard drive form. Table 2.1 summarizes the measured well-log data for the São Francisco basin, while Table 2.2 provides a summary of the available data for Santos basin, as sourced from ('ANP-TERRESTRE', 2023).

| WELL NUMBER | WELL LOGS |
|---------------------|--|
| 1 - BRSA - 871 - MG | DTCO - DTSM - ESPEC - GR - HCAL - NMR - NPHI - PEFZ - RES - RHOZ |
| 1 - BRSA - 948 - MG | DTCO - DTSM - ESPEC - GR - HCAL - NMR - NPHI - PEFZ - RES - RHOZ |

Table 2.1: Available measured well-log data for each well (São Francisco basin).

| WELL NUMBER | WELL LOGS |
|-----------------------|--|
| 1 - BRSA - 1116 - RJS | DTCO - DTSM - ESPEC - GR - HCAL - NMR - NPHI - PEFZ - RES - RHOZ |
| 3 - BRSA - 1215 - RJS | DTCO - DTSM - ESPEC - GR - HCAL - NMR - NPHI - PEFZ - RES - RHOZ |

Table 2.2: Available measured well-log data for each well (Santos basin).

Additionally, calculated parameters are provided for each well in both basins. These calculated parameters include Effective Porosity (PHIE_HILT), and the Grain Density (RHGX_HILT).

2.1.1 Well-Log Data Information

As follows, we briefly explain the importance of each Wireline Logging Curve used in this work. While we provide a brief explanation here, more comprehensive information about these logs and their applications can be found in open sources such as the glossary provided by (Schlumberger, 2023) or in books like "The log analysis handbook" written by (Crain and Ganz, 1986). The order of explanation follows the presentation in Table 2.1, and Table 2.2.

The **Sonic log**, in our case **Compressional Wave Slowness (DTCO)** and **Shear Wave Slowness (DTSM)**, provides measurements of the slowness of a refracted elastic wave that reaches the wellbore wall with a specific inclination angle, known as the critical angle. It is primarily used to derive the porosity of a formation.

The **Elemental Spectroscopy log (ESPEC)** measures the elemental concentrations in the formation, such as calcium, magnesium, and silicon. The ESPEC tool emits gamma-rays into the formation, which interact with the atomic nuclei of the surrounding elements. Consequently, gamma-rays undergo energy changes that are indicative of the elemental composition of the formation. These measurements allow for more accurately defining the clay content, mineralogy, and matrix properties in each potential zone.

The **Gamma-Ray log (GR)** provides a reading of the natural radioactivity emitted by the reservoir formation. It is useful for identifying shaly zones, which are characterized by high

radioactivity. It also helps calculate shale volume and identify the type of clay mineral present in the shale.

The **Caliper log (HCAL)** represents a measurement of the diameter of the well, in inches, and it gives information on the borehole shape. It is important for quality control of most of the logs since they are sensitive to variations in the borehole diameter.

The **Nuclear Magnetic Resonance log (NMR)** is the most important acquired log. An NMR returns the porosity as partitioned in clay-bound water (CBW), bulk volume immovable (BVI) which represents the irreducible water, in our case the capillary-bound water, and bulk volume movable (BVM) that is the volume occupied by the mobile fluids, like hydrocarbons, and water. For a detailed description of this powerful tool, we can refer to (Westphal *et al.*, 2005).

The **Neutron Porosity log (NPHI)** provides a reading of the formation's response when it is subjected to a radioactive bombardment of fast neutrons. Its response is an indicator of the porosity of the formation.

The **Photo-Electric Factor log (PEFZ)** gives a reading of the so-called photo-electric absorption by the formation, which refers to the absorption of gamma rays under a certain threshold. It is an extremely good indicator of the lithology.

The **Resistivity log (RES)** measures the electrical resistivity of the formation, specifically the apparent formation resistivity. In this research, investigation depths of 10-30-90 inches (AT10, AT30, and AT90) are considered. The resistivity curves are used as a water saturation indication; we have low resistivity values for water-bearing zones and high resistivity values for hydrocarbon-bearing zones.

The **Formation Density log (RHOZ)** measures the response of the formation to a radioactive bombardment of gamma rays. Its response is an indicator of the bulk density of the formation.

2.1.2 Well-Log Data Visualization

To gain a comprehensive understanding of the well-logging features, and their correlation with the measured depth, Fig. 2.3, Fig. 2.4, Fig. 2.5, and Fig. 2.6 provide an example of a complete well-logging feature analysis.

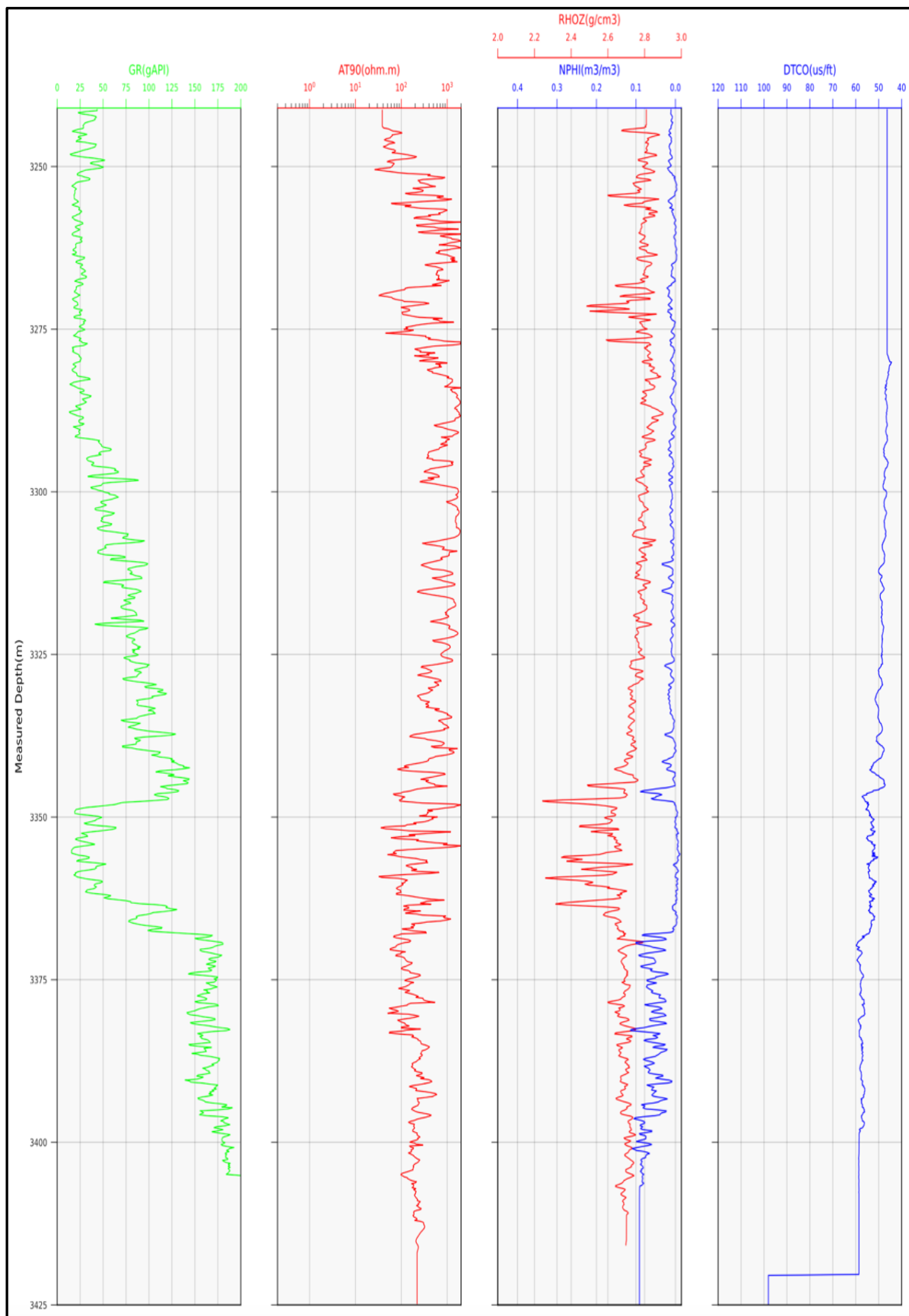


Fig. 2.3: Composite well-log of well 1-BRSA-871-MG. Track 1: Measured well depth. Track 2: Gamma-Ray (GR) log. Track 3: Induction Electric Resistivity logs. Investigation depth of 90 (AT90) inches. Track 4: Formation Density (RHOZ), and Neutron Porosity (NPHI) logs. Track 5: Compressional Wave Slowness (DTCO) log.

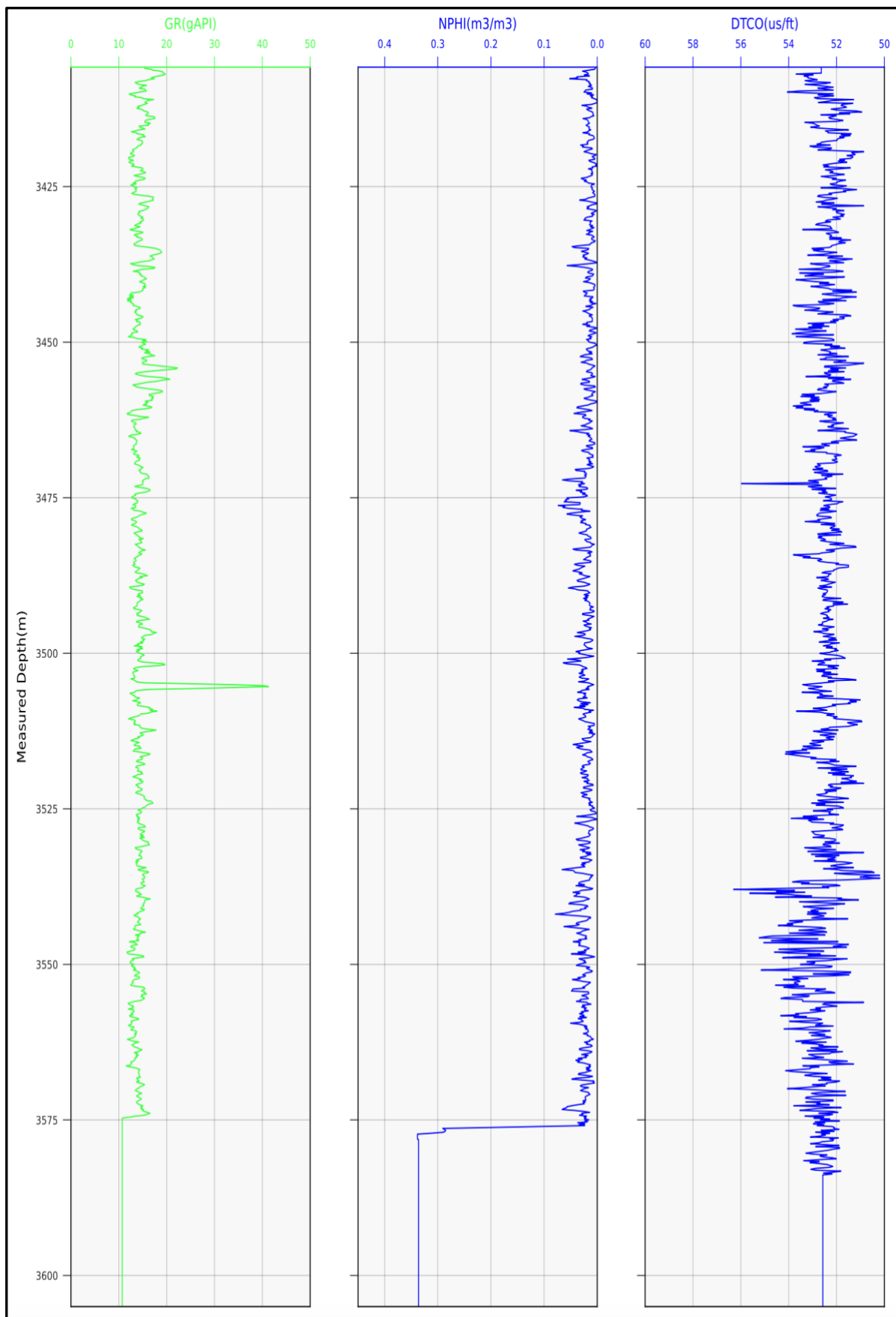


Fig. 2.4: Composite well-log of well 1-BRSA-948-MG. Track 1: Measured well depth. Track 2: Gamma-Ray (GR) log. Track 3: Neutron Porosity (NPHI) log. Track 4: Compressional Wave Slowness (DTCO) log.

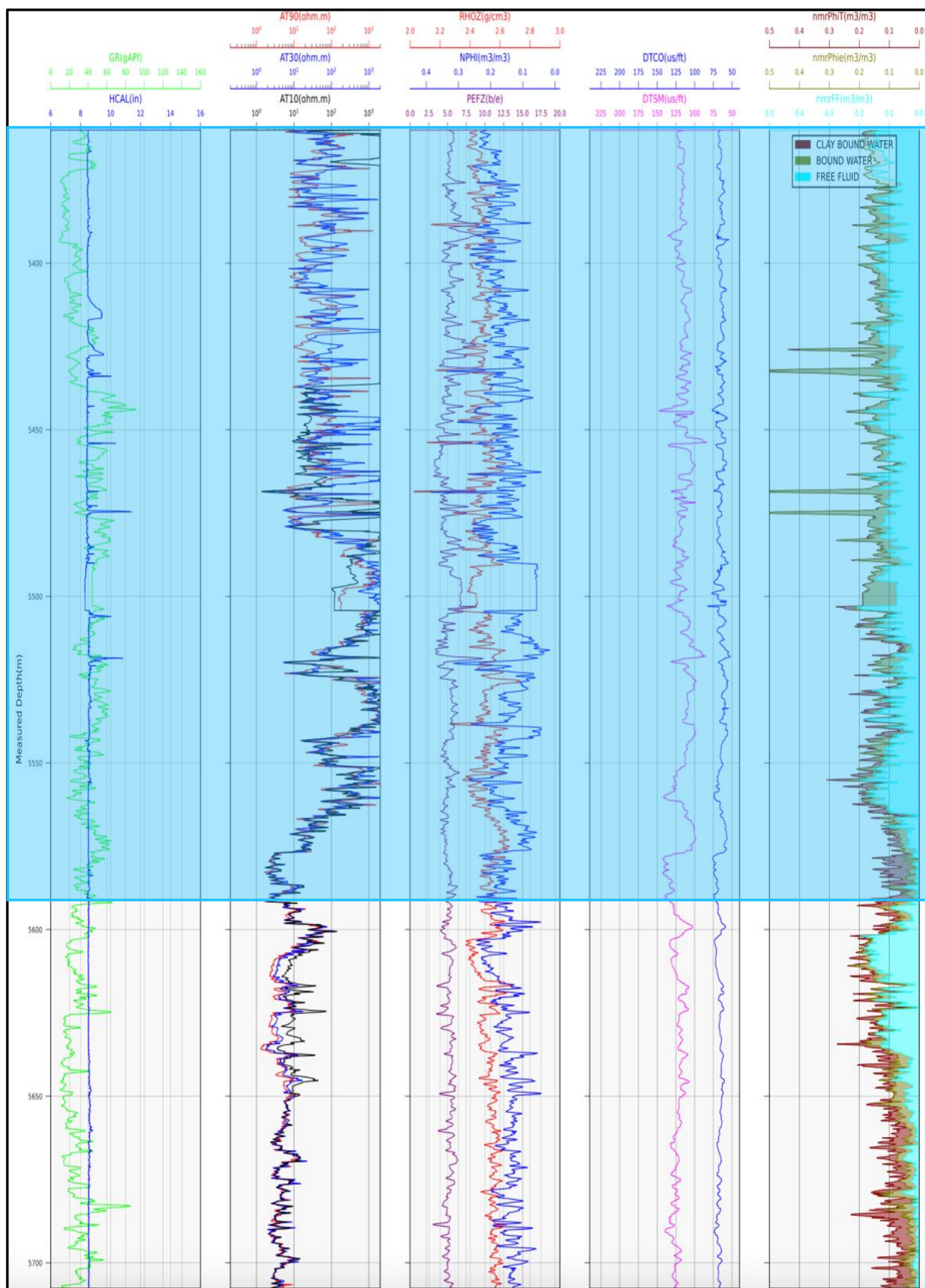


Fig. 2.5: Composite well-log of well 1-BRSA-1116-RJS. Track 1: Measured well depth. Track 2: Gamma-Ray (GR) and Caliper (HCAL) logs. Track 3: Induction Electric Resistivity logs. Investigation depths of 10 (AT10), 30 (AT30), and 90 (AT90) inches. Track 4: Formation Density (RHOZ), Neutron Porosity (NPHI), and Photoelectric Factor (PEFZ) logs. Track 5: Compressional Wave Slowness (DTCO) and Shear Wave Slowness (DTSM) logs. Track 6: Nuclear Magnetic Resonance Porosity logs. Total Porosity (nmrPHIT), Effective Porosity (nmrPHIE), and Free Fluid (nmrFF). In deepskyblue, the reservoir rock.

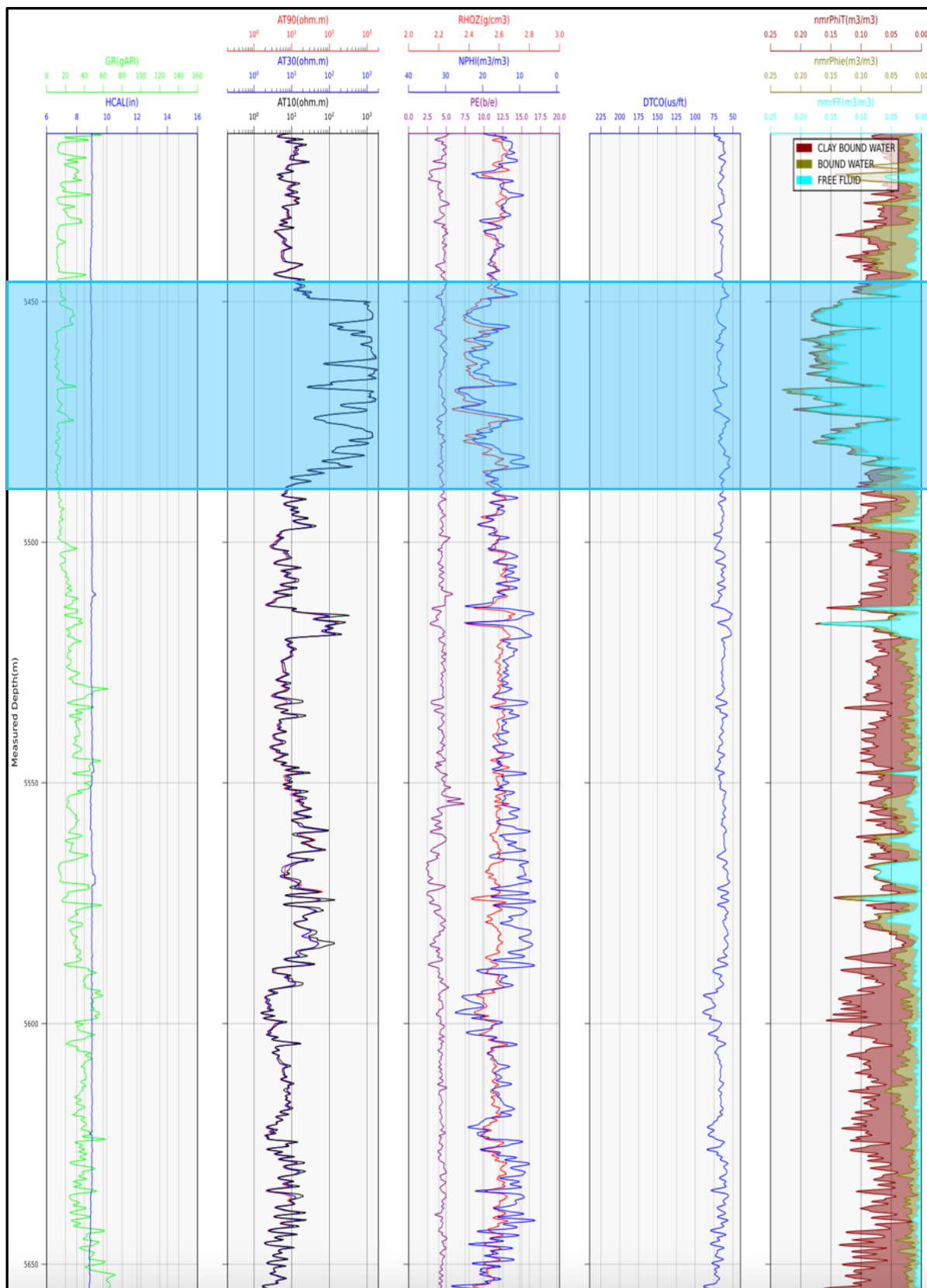


Fig. 2.6: Composite well-log of well 3-BRSA-1215-RJS. Track 1: Measured well depth. Track 2: Gamma-Ray (GR) and Caliper (HCAL) logs. Track 3: Induction Electric Resistivity logs. Investigation depths of 10 (AT10), 30 (AT30), and 90 (AT90) inches. Track 4: Formation Density (RHOZ), Neutron Porosity (NPHI), and Photoelectric Factor (PEFZ) logs. Track 5: Compressional Wave Slowness (DTCO) log. Track 6: Nuclear Magnetic Resonance Porosity logs. Total Porosity (nmrPHIT), Effective Porosity (nmrPHIE), and Free Fluid (nmrFF). In deepskyblue, the reservoir rock.

We provide a qualitative description of the plotted well-logging curves for the mentioned wells.

Well 1-BRSA-871-MG: the GR log helps identify different lithologies. From depths of 3,241 to 3,300 meters, there is dolomite; from 3,300 to 3,350 meters, there is sandstone; from 3,350 to 3,375 meters, there is sandstone with some dolomite layers, and from 3,375 to 3,425 meters, there is shale with some silt layers. The shale exhibits the highest GR values, reaching up to 200 gAPI. The NPHI and RHOZ logs indicate the presence of a gas-bearing zone within the depth interval of 3,350 to 3,375 meters, approximately.

Well 1-BRSA-948-MG: the GR log displays values ranging from 5 to 50 gAPI, which are typical for sandstone formation.

Well 1-BRSA-1116-RJS: the GR log shows values ranging from 10 to 40 gAPI, characteristic of limestone lithology. The HCAL log indicates a consistent well diameter of 8 inches, indicating a relatively uniform borehole size. The PEFZ log displays a value of 5 (barns/electron), which is typical for limestone lithology. The NMR and RES logs show a relationship, with high resistivity values corresponding to high values of free fluid (water + hydrocarbon), indicating areas with increased fluid content. Hydrocarbon-bearing zones can be identified based on these observations. The oil-bearing is located between depths of 5,360 and 5,600 meters, with the oil-water contact at approximately 5,600 meters. This implies that 90% of the well is reservoir formation, while the remaining 10% is non-reservoir formation.

Well 3-BRSA-1215-RJS: like well 1-BRSA-1116-RJS, GR, HCAL, and PEFZ logs indicate the presence of limestone lithology. The relationship between the NMR and RES logs is even more evident in this well. The oil-bearing zone is found within the depth interval of 5,450 to 5,500 meters, with only 20% of the formation representing reservoir rock, and the remaining 80% being non-reservoir rock.

2.1.3 Geological Info about São Francisco and Santos Basin

São Francisco basin belongs to the state of Minas Gerais, and it extends over an area of 350,000 square kilometers. The geological understanding of the basin is still limited, but ongoing hydrocarbon exploration activities and regional mapping initiatives are gradually improving our knowledge of the area (Reis *et al.*, 2017). The presence of surface gas seeps and gas flows observed in the drilled wells throughout the basin supports its petroleum potential (Mello, de Mio and Bruno, 2018).

According to info provided by ('ANP-TERRESTRE', 2023), the formations drilled by the wells of interest are Serra de Santa Helena, Sete Lagoas, and Jequitai. For a detailed description of lithostratigraphy, (Reis *et al.*, 2017) provide comprehensive information.

Santos basin is the largest offshore basin in Brazil, situated to the east of the states of São Paulo and Rio de Janeiro. It is currently the nation's largest oil-producing basin. With a water depth of up to 3,000 meters (deep-water basin), the basin covers an area of about 350,000 square kilometers. The formation of the Santos basin can be attributed to the separation of South America and Africa, during the Cretaceous period (145 to 66 million years ago), because of the rupture of the Supercontinent Gondwana, and the subsequent opening of the Atlantic Ocean (Lupinacci et al., 2023). Fig. 2.7 illustrates a schematic representation of the layer types found in the Santos basin.

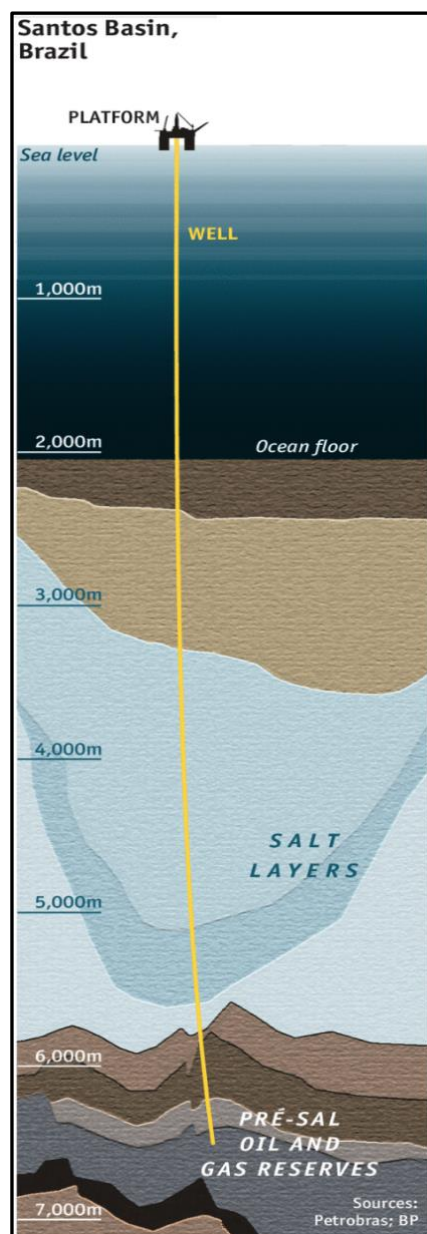


Fig. 2.7: Pre-salt oil and gas layer, salt layer, post-salt layer, and water depth for the Santos basin (PETROBRAS).

The pre-salt layers are located at a depth interval of 6,000-7,000 meters. They represent one of the most significant offshore petroleum discoveries in the last two decades. The discovery stands out due to its large volumes of resources in place, high reservoir productivity, the thick

and high-quality seal mostly composed of halite and anhydrite, and the presence of high-quality oil, with an API gravity of about 28-30° (Souza *et al.*, 2022).

To provide further insight into the lithostratigraphy, it is worth mentioning that the wells of interest have drilled the Ariri and Barra Velha formations. For a simplified stratigraphic chart of the Santos basin, (Gomes *et al.*, 2020) offers valuable information.

Moreover, Fig. 2.8 provides additional context by presenting an example of schematic lithological columns of the areas of interest. In the Santos basin, the drilled wells encountered a carbonate section consisting of limestone, which is capped by a halite layer, followed by an anhydrite layer, at the top, and by a shale layer, at the bottom.

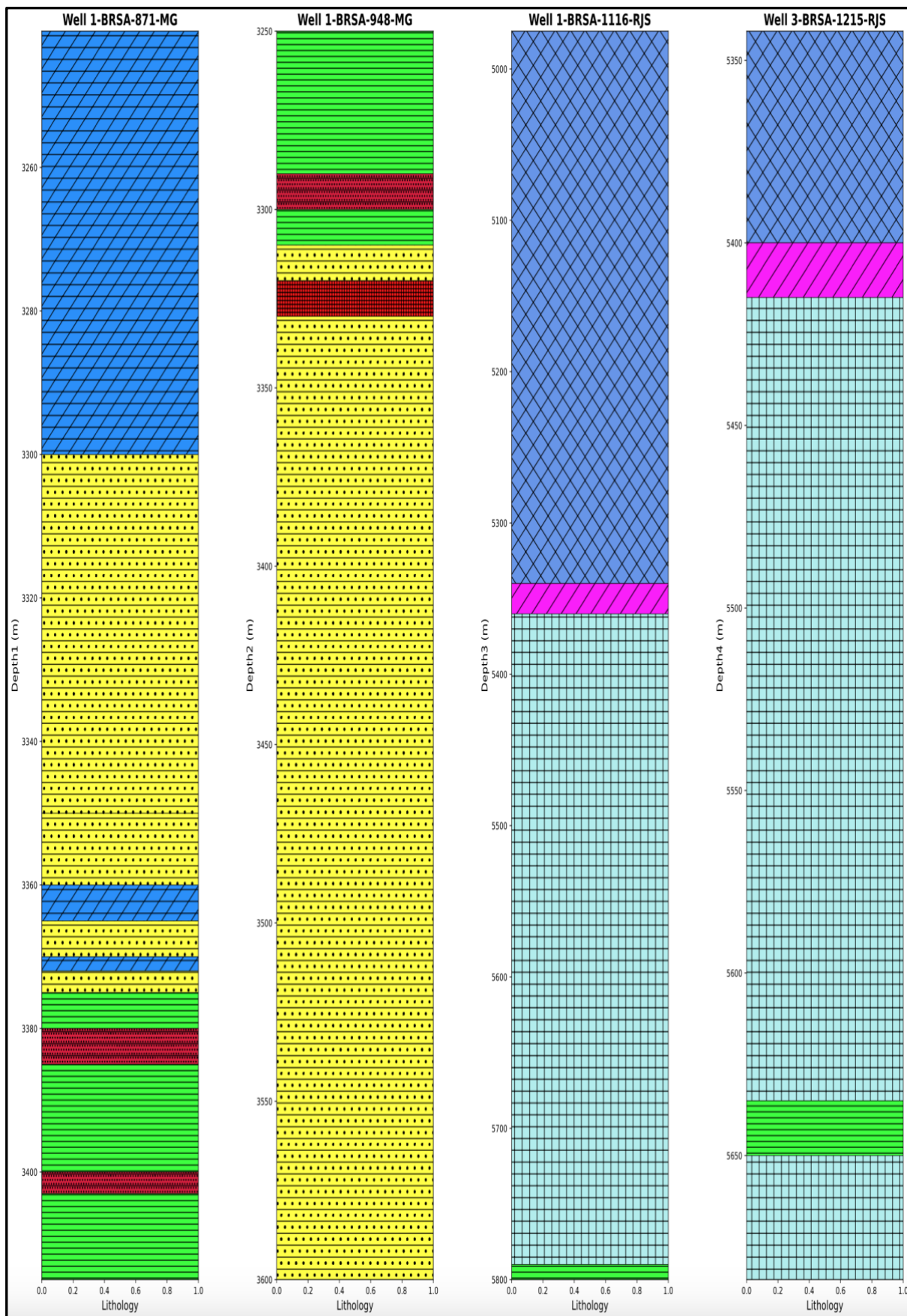


Fig. 2.8: Schematic lithological columns for well 1-BRSA-871-MG (first column), well 1-BRSA-948-MG (second column), well 1-BRSA-1116-RJS (third column), and well 3-BRSA-1215-RJS (fourth column) ('ANP-TERRESTRE', 2023).






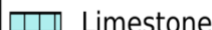

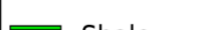


| Lithology São Francisco Basin | Lithology Santos Basin |
|---|--|
|  Dolomite |  Anhydrite |
|  Igneous |  Halite |
|  Limestone |  Limestone |
|  Sandstone |  Shale |
|  Shale | |
|  Silt | |

Table 2.3: Lithology tables for São Francisco and Santos basins.

2.1.4 General Information about the Itapu Oil Field, Santos Basin

In the Santos basin, which consists of various pre-salt fields, our study focuses on the Itapu oil field (Fig. 2.9), which is situated at a water depth of 2,000 meters and approximately 200 kilometers offshore Rio de Janeiro. Itapu is a significant oil field with an estimated in situ oil volume of 1.3 billion barrels of oil and is operated by the state-run oil and gas company, Petroleo Brasileiro, Petrobras (Offshore Technology, 2021). The discovery of Itapu was made in December 2012 through the drilling of the 1-BRSA-1116-RJS exploratory well. A formation test, conducted on the well, confirmed the excellent productivity of the reservoir (Offshore Technology, 2021).

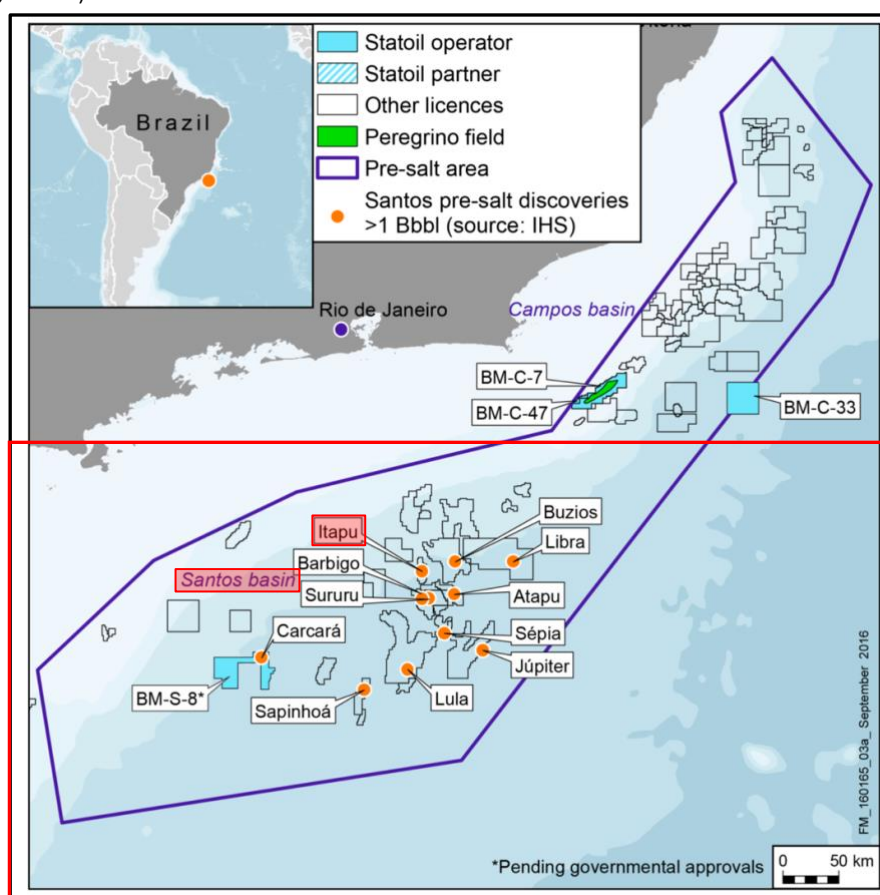


Fig. 2.9: Location map of Santos basin and the main pre-salt fields. Itapu field is highlighted in red (Equinor, 2017).

2.2 Applied Methodology

The methodology employed in this research adheres to the workflow depicted in Fig. 2.10. We will go through each phase and step.

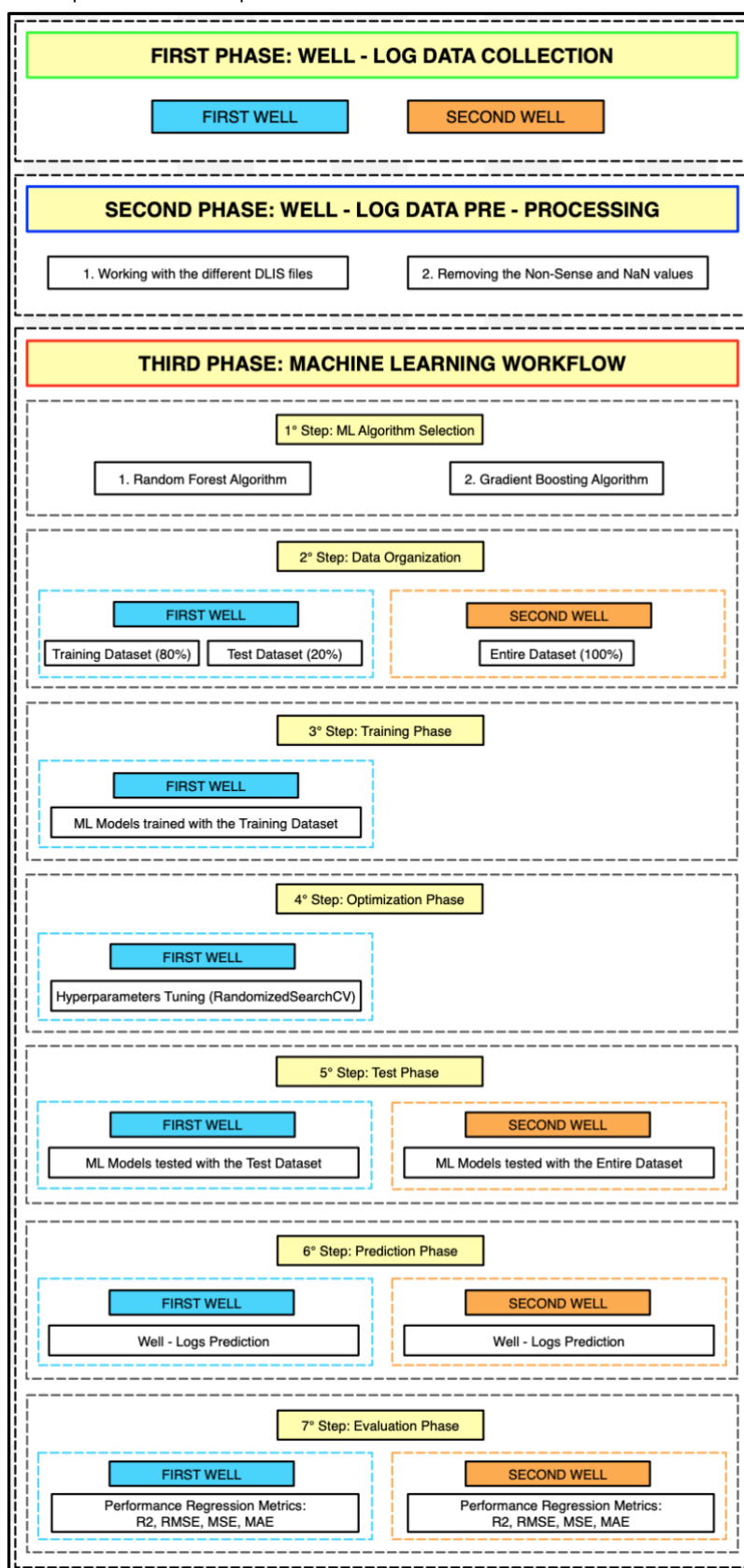


Fig. 2.10: Schematic diagram showing the research methodology applied in this study.

Moreover, we present a schematic representation (Fig. 2.11) that showcases the wells included in the study and the corresponding predicted well-logs, for each Brazilian basin.

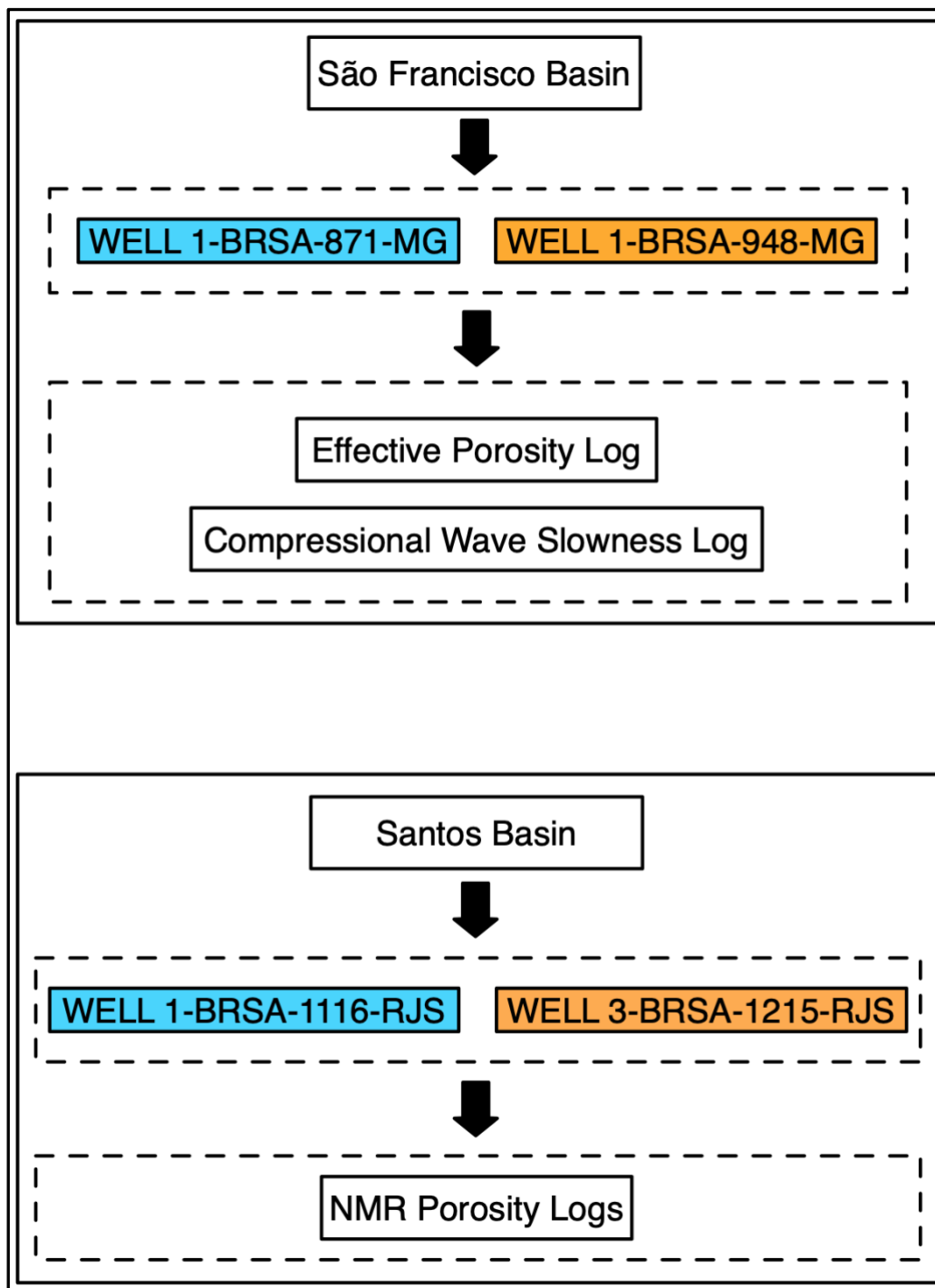


Fig. 2.11 Scheme of the available wells and predicted well-logs for each basin.

Each phase and step, summarized in Fig. 2.10, is entirely implemented using the Python Programming Language. The following Python libraries are employed in this study:

1. DLISIO library is used to read DLIS files, which contain well-log data provided by Brazil's National Agency for Petroleum, Natural Gas and Biofuels ('ANP-TERRESTRE', 2023).
2. Folium library is employed for mapping purposes, specifically to visualize the locations of the wells.
3. JSON library is used to read JSON files, which assist in mapping the coordinates of different regions and sub-regions in Brazil ('Brazilian States Coordinates', 2023)('Brazilian Sub-Regions Coordinates', 2023).
4. Matplotlib library, known as the plotting library, is utilized for visualizing and plotting well-log data.
5. LASIO library is employed to read LAS files. In some cases, the DLIS files may be converted to LAS format, for convenience.
6. NumPy library is used to provide support to multi-dimensional arrays. Throughout the implementation of the code, different arrays are manipulated.
7. Pandas library is employed to import data from different formats and to create different data frames.
8. Pickle library is used to read pickle files, which are used to save the ML Models in this research.
9. QB Styles library, short for "Quantum Black Styles", contains light and black Matplotlib styles. In this study, the light style is chosen for the plots.
10. Scikit-Learn library, known as a comprehensive ML library, offers efficient tools for various data analysis tasks, including classification, regression, clustering, dimensionality reduction, model selection, and pre-processing. In this research, the focus is on regression tasks ('Scikit-learn, Machine learning in Python', 2023).

In Appendix 7 we report some lines of the Python code that we implemented to predict NMR porosity logs. The code follows the structure outlined in Fig. 2.10. Furthermore, the complete work is available on GitHub at the following link:

[\[https://github.com/VittoDePe98/Well-Logs_Predictive_Models.git\]](https://github.com/VittoDePe98/Well-Logs_Predictive_Models.git).

We created a repository named **Well-Logs_Predictive_Models**, that is organized as follows:

- Introduction – Project Objective.
- Well-Log Datasets Used.
- 3 Folders Containing Notebooks.
- Software Used.
- Citation.

Of particular interest is the "3 Folders Containing Notebooks" section, which is organized as follows (Fig. 2.12):




1.  **Lithology_Visualization:**
 - `Lithology_Columns_SãoFrancisco_Santos_Basins.ipynb`
2.  **LogData_Collection_and_Visualization:**
 - `RequiredDataset_CompressionalWaveSlownessLogPrediction_São FranciscoBasin.ipynb`
 - `RequiredDataset_EffectivePorosityLogPrediction_São FranciscoBasin.ipynb`
 - `RequiredDataset_nmrPorosityLogsPrediction_SantosBasin.ipynb`
3.  **Logs_Prediction:**
 - `CompressionalWaveSlownessLog_Prediction.ipynb`
 - `EffectivePorosityLog_Prediction.ipynb`
 - `nmrPorosityLogs_Prediction_FirstAttempt.ipynb`
 - `nmrPorosityLogs_Prediction_SecondAttempt.ipynb`

Fig. 2.12: Notebooks present within the GitHub repository.

In the **Lithology_Visualization** folder, a tool has been developed to display lithological columns. This tool allows for the visualization of lithological information in a graphical format. In the **LogData_Collection_and_Visualization** folder, the well-log data obtained from different DLIS files are saved into data frames. This step involves extracting the required well-log data from the available DLIS files and organizing them in a structured format. Additionally, a tool has been implemented to visualize the well-log data, enabling the analysis of the log data through plots and visual representations. Finally, the **Logs_Prediction** folder is of particular importance as it contains the implementation of the ML predictive models required for the prediction task. These three folders collectively represent the key components of the research implementation.

2.3 Well-Log Data Collection

In the first phase of the workflow (Fig. 2.10), the focus is on collecting the well-log data from the available DLIS files for the selected wells. For case study 1, wells 1-BRSA-871-MG and 1-BRSA-948-MG are chosen for data collection. Similarly, for case study 2, wells 1-BRSA-1116-RJS and 3-BRSA-1215-RJS are selected.

The **LogData_Collection_and_Visualization** folder contains the necessary information and tools to collect and visualize the well-log data for these wells.

2.4 Well-Log Data Pre-Processing

The “Well-Log Data Cleaning” phase, as part of the **LogData_Collection_and_Visualization** folder, focuses on preparing the collected well-log data for modeling. This phase consists of two steps:

- Splicing data into a single data frame: since the well-log data are sourced from different DLIS files, the first step involves performing a splice operation to merge the data into a single data frame, for each well. This operation ensures that all the relevant data, for a particular well, are consolidated in one place.
- Removal of non-sense and NaN values: the focus is on removing non-sense and NaN (not a number) values from the well-log data. These NaN values can interfere with the accuracy and performance of the ML models. Therefore, it is crucial to identify and remove these values to ensure the quality of the dataset.

2.5 Machine Learning Workflow

The ML workflow consists of different steps. As follows, we will provide only a brief and qualitative explanation of each step. The detailed implementation of each step can be found in Appendix 7, or in the **Logs_Prediction** folder.

2.5.1 Machine Learning Algorithm Selection

As already mentioned, we decide to work with two supervised ML algorithms, RF and GB.

2.5.1.1 Random Forest Algorithm Implementation

To implement this algorithm, we import the `RandomForestRegressor` function from the Scikit-Learn library. This function uses different hyper-parameters to fit the model to the input data; they will be specified during the Optimization phase. Moreover, this function can predict more than one variable at a time.

Additionally, the `"tree.plot_tree"` function is employed to extract and visualize individual decision trees. In Fig. 2.13, the first decision tree, with a limited depth (`max_depth=2`), built for the prediction of NMR porosity logs when the RF model is trained on well 3-BRSA-1215-RJS, is shown.

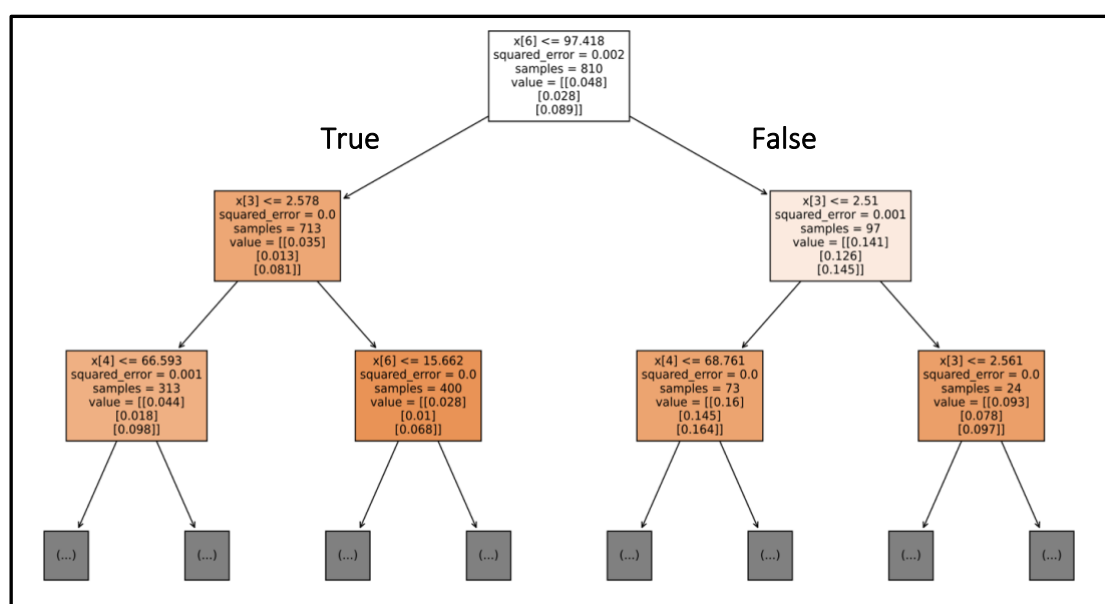


Fig. 2.13: Schematic representation of the decision process occurring within the RF algorithm.

The first node (white square) represents the Root Node, whereas all the other colored nodes are the Decision Nodes. They contain the following elements:

1. The input feature and its value to split the node on, which are chosen randomly. They are indicated with the index value; for instance, 6 corresponds to AT90.
2. The mean squared error (MSE) of the node.
3. The number of data points in the node.
4. The prediction target; in this case, three prediction outcomes.

The darker the node, the lower the error for that prediction.

2.5.1.2 Gradient Boosting Algorithm Implementation

The GB algorithm is implemented using the GradientBoostingRegressor function that is imported from the Scikit-Learn library. It uses similar hyper-parameters as those defined for the RandomForestRegressor, in the Optimization phase. This function can predict only one variable at a time; thus, we need to import the MultiOutputRegressor function.

2.5.2 Well-Log Data Organization

We select the training and the validation/test wells. The training well is used to build the model and it helps to capture the relationships between the input and the target variables. On the other hand, the test well is an “unseen” well, meaning it is not used during the ML model-building process. It allows for assessing the performance and generalization ability of the trained ML model on new and unseen data.

In case study 1, the training well is 1-BRSA-871-MG, and the validation well is 1-BRSA-948-MG. In case study 2, we distinguish two attempts. In the first attempt, the training well is 3-BRSA-1215-RJS, and the test well is 1-BRSA-1116-RJS. In the second attempt, the training well is 1-BRSA-1116-RJS, and the validation well is 3-BRSA-1215-RJS.

Moreover, we specify the input variables to the ML models, which in this case are limited to conventional well-log data, and the output variables that the ML models aim to predict.

In case study 1, we have two different predictions. For the calculated PHIE_HILT log, the input parameters used are GR, NPHI, and RHGX_HILT. For the measured DTCO log, only two input variables are used, which are GR and NPHI.

In case study 2, the focus is on predicting NMR porosity logs. The input parameters for this prediction include AT90, DTCO, GR, HCAL, NPHI, PEFZ, and RHOZ.

The selection of the input features is not random but based on the knowledge of how the target variables are calculated or related to specific well-log measurements.

For the prediction of the calculated PHIE_HILT log, we know that GR, NPHI, and RHOZ logs provide information about the total porosity, which is then used in equation (Eq. 2.1):

$$\phi_E = \phi_T - \phi_{WB} \quad \text{Eq. 2.1}$$

where ϕ_E is the effective porosity, ϕ_T is the total porosity, and ϕ_{WB} is the porosity occupied by the clay-bound water. According to Eq. 2.1, we can understand that PHIE_HILT is indirectly related to GR, NPHI, and RHOZ logs. In addition, instead of the RHOZ log, we use the RHGX_HILT log, which is calculated from it.

The DTCO log provides information about the total porosity, and the GR and NPHI logs contribute to estimating this total porosity.

In the case of predicting NMR porosity logs, RHOZ, DTCO, GR, HCAL, NPHI, PEFZ, and AT90 logs, collectively provide information about the rock matrix, mineral composition, fluid change, and rock porosity; these are relevant factors in predicting NMR porosities.

We summarize the well-log predictors and outputs in Table 2.4, Table 2.5, and Table 2.6:

| WELL NUMBER | LITHOLOGY | WELL LOG PREDICTORS | WELL LOG OUTPUTS |
|---------------------|---|-------------------------|------------------|
| 1 – BRSA – 871 – MG | Dolomite Diamictite Limestone Loam Sandstone Shale Silt | GR NPHI RHGX_HILT | PHIE_HILT |
| 1 – BRSA – 948 – MG | Grainstone Limestone Shale | GR NPHI RHGX_HILT | PHIE_HILT |

Table 2.4: Predictors and Outputs for each well, with the corresponding lithology (São Francisco basin)

| WELL NUMBER | LITHOLOGY | WELL LOG PREDICTORS | WELL LOG OUTPUTS |
|---------------------|---|---------------------|------------------|
| 1 – BRSA – 871 – MG | Dolomite Limestone Sandstone Shale Silt | GR NPHI | DTCO |
| 1 – BRSA – 948 – MG | Sandstone | GR NPHI | DTCO |

Table 2.5: Predictors and Outputs for each well, with the corresponding lithology (São Francisco basin)

| WELL NUMBER | LITHOLOGY | WELL LOG PREDICTORS | WELL LOG OUTPUTS |
|-----------------------|-----------|--|-----------------------------|
| 1 – BRSA – 1116 – RJS | Limestone | AT90 DTCO GR HCAL NPHI PEFZ RHOZ | NMRFF NMRPHIE NMRPHIT |
| 3 – BRSA – 1215 – RJS | Limestone | AT90 DTCO GR HCAL NPHI PEFZ RHOZ | NMRFF NMRPHIE NMRPHIT |

Table 2.6: Predictors and Outputs for each well, with the corresponding lithology (Santos basin)

We develop a total of eight supervised models: two models for predicting the PHIE_HILT log (single-output models), two models for predicting the DTCO log (single-output models), and four models for predicting NMR porosity logs (multi-output models).

Just in case the mentioned abbreviations are not clear, Table 2.7 provides a reference table that associated each abbreviation with its corresponding extended log name.

| ABBREVIATION | EXTENDED LOG NAME |
|--------------|--|
| AT90 | Array induction resistivity log, investigation 90 in |
| DTCO | Delta T compressional wave, compressional slowness log |
| GR | Gamma-ray log |
| HCAL | Caliper log |
| NMR | Nuclear magnetic resonance log |
| NPHI | Thermal neutron porosity log |
| PEFZ | Photo-electric log |
| PHIE_HILT | HILT effective porosity log |
| RHGX_HILT | HILT grain density log |
| RHOZ | Formation density log |

Table 2.7: Abbreviations and Extended Log Names.

2.5.3 Training Phase

We consider the original dataset of the training wells, and by means of the “train_test_split” function, it is randomly split into two parts: the training and the test datasets. The training dataset comprises 80% of the data and is used to train the model, while the remaining 20% is used to test the model. The training dataset allows the algorithm to understand the patterns that exist between the input and the output data, whereas the test dataset, which is the unseen dataset by the algorithm, serves to evaluate the predictive power of the trained model. The relationship that is established between the input and output variables is unknown, thus this function can be seen as a “black box”. The random shuffling process employed in the RF and GB algorithms is a key feature. In Fig. 2.14, we provide a schematic representation of the random shuffling process.

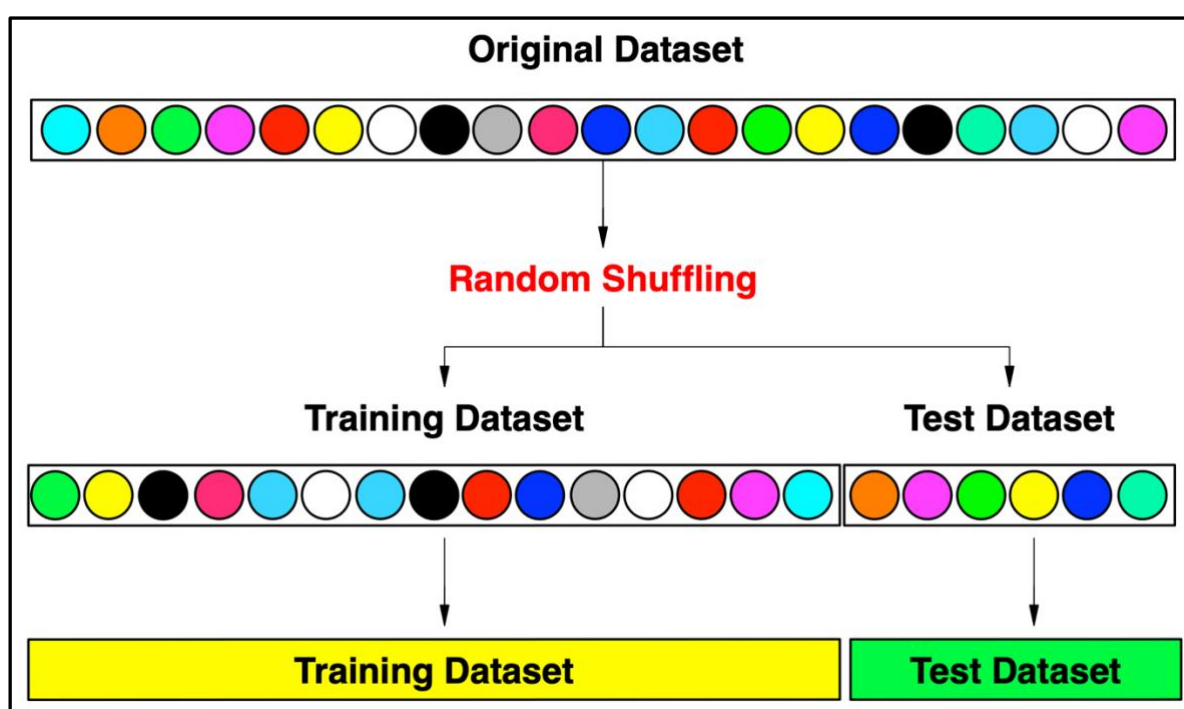


Fig. 2.14: Random Shuffling of the original dataset performed during Training and Test Split

Now we consider only the training dataset, and we proceed to train our models using it. However, to ensure more accurate training, it is recommended to apply one of the Cross-Validation techniques. We employ the K-Fold Cross-Validation (KFCV) technique. The data points of the training dataset are randomly assigned to K folds or called groups. In this case, we set K equal to 10, resulting in 10-Fold Cross-Validation. In an iterative process consisting of 10 iterations, each time one of the K folds is selected as the validation dataset, while the remaining (K-1) folds form the new training dataset. This procedure is repeated K times, ensuring that each group is selected as the validation dataset once. During each iteration, an accuracy value is calculated and an average accuracy value, across the 10 iterations, is obtained (Maleki *et al.*, 2020). Fig. 2.15 presents a schematic representation of the KFCV technique.

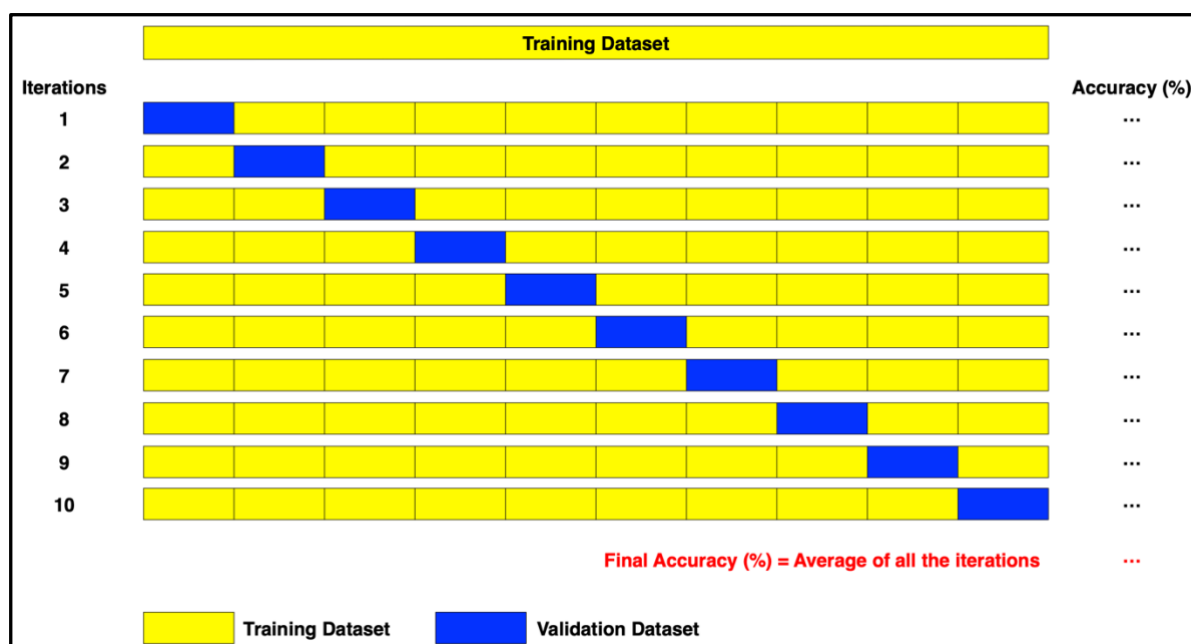


Fig. 2.15: *k*-Fold Cross-Validation, where *k*=10.

The result of the training phase is a trained model that has learned the relationship between the input and output features. However, since ensemble-based decision tree approaches involve randomness, the results may vary each time the model is trained. To ensure reproducibility, we fixed the value of the “random_state” parameter, for instance, 42. The next step is to tune/optimize the trained models.

2.5.4 Optimization Phase

The `RandomForestRegressor`, and the `GradientBoostingRegressor` functions have a set of hyperparameters that can be tuned to optimize the model’s performance. The complete list and meanings of all the hyperparameters are available on the official Scikit-Learn webpage (‘Scikit-learn, Machine learning in Python’, 2023). For simplicity, we consider only the following hyper-parameters: `n_estimators`, and `max_depth` for the RF Regressor; `n_estimators`, `max_depth`, and `learning_rate` for the GB Regressor. We summarize the meaning of each parameter (Farmanov *et al.*, 2023):

- `n_estimators` is the number of independent trees in the ensemble. The RF and GB models perform better when it is increased, but doing so increases the computational time.
- `max_depth` is the maximum depth of each tree in the ensemble. Larger depths sometimes result in better model performance but exponentially longer computational time.
- `learning_rate` represents how fast the model learns. It controls the change of the weights applied to the input data.

Additional hyper-parameters that we do not consider can be the `min_samples_split`, which is the lowest number of samples required to split a node, and the `min_samples_leaf`, which is the minimum number of samples required to be at a leaf node.

To choose the best hyperparameter values we need to perform “Hyperparameter Optimization or Hyperparameter Tuning”, which is extremely important since our target is to optimize the performance of RF and GB regressors. In this process, we rely on the Randomized Search Cross-Validation technique, using the “`RandomizedSearchCV`” function which allows determining the optimal hyperparameter values for RF and GB, while the models are cross-validated on these random combinations. The best hyperparameter couple is the one that maximizes the accuracy (R^2) of the model and minimizes its error (MSE, RMSE, and MAE). The “`RandomizedSearchCV`” function explores multiple random combinations of hyperparameters. In Fig. 2.16 a schematic representation of the random search grid is provided.

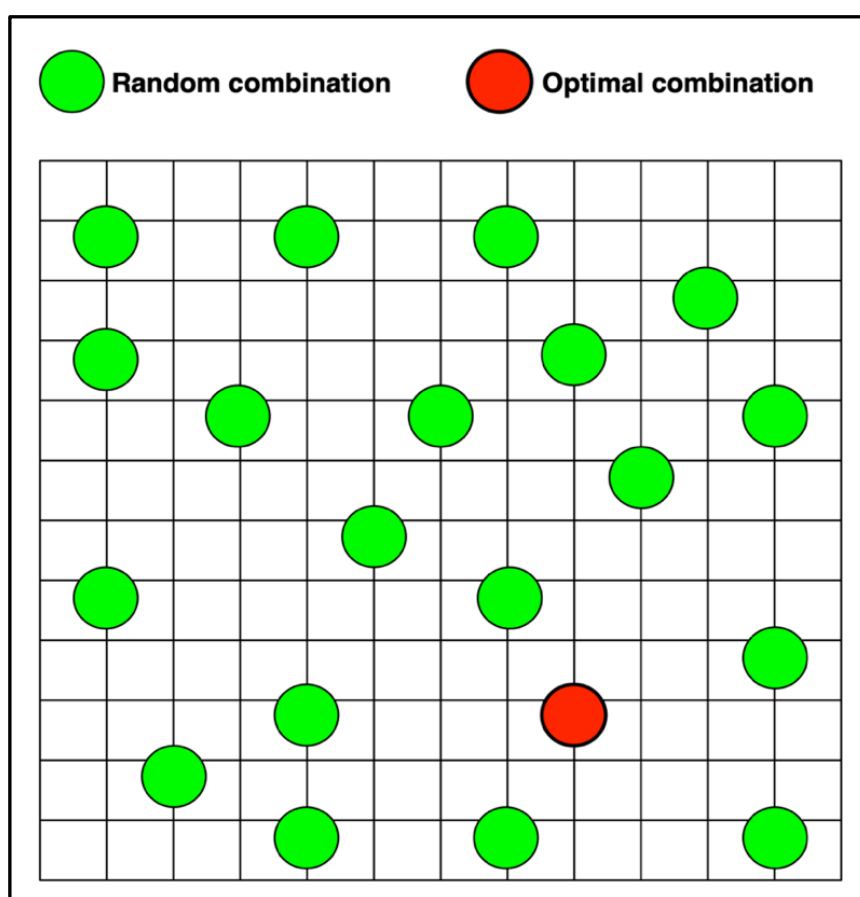


Fig. 2.16: Random Search Grid

The point of strength of the ensemble-based decision tree approaches is that intensive parameter tuning is not required compared to other models (Babasafari *et al.*, 2022).

The optimized hyperparameters found for each algorithm are summarized in Table 2.8, Table 2.9, and Table 2.10 as follows:

| Effective Porosity Prediction – Optimal Couple | | | | |
|--|-----------|-------------------|-----------|---------------|
| Random Forest | | Gradient Boosting | | |
| n_estimators | max_depth | n_estimators | max_depth | learning_rate |
| 250 | 20 | 250 | 5 | 0.3 |

Table 2.8: Optimal couple for Effective Porosity Prediction for both RF and GB (São Francisco basin)

| Compressional Wave Slowness Prediction – Optimal Couple | | | | |
|---|-----------|-------------------|-----------|---------------|
| Random Forest | | Gradient Boosting | | |
| n_estimators | max_depth | n_estimators | max_depth | learning_rate |
| 150 | 10 | 250 | 5 | 0.3 |

Table 2.9: Optimal couple for Compressional Wave Slowness Prediction for both RF and GB (São Francisco basin)

| NMR Porosities Prediction – Optimal Couple | | | | |
|--|-----------|-------------------|-----------|---------------|
| Random Forest | | Gradient Boosting | | |
| n_estimators | max_depth | n_estimators | max_depth | learning_rate |
| 400 | 20 | 250 | 5 | 0.3 |

Table 2.10: Optimal couple for NMR Porosities Prediction for both RF and GB (Santos basin)

2.5.5 Test Phase

Once our models are trained and tuned, we need to test them. The test phase consists of two steps. In the first step, we apply the models to the test dataset (20%) of the same training well. In the second step, the most important and critical one, we apply the models to the entire dataset of the test well. This last step is crucial because it tests the ability of the models to generalize and make accurate predictions on completely unseen data.

2.5.6 Prediction Phase

The predicted well-log results are compared with the measured well-log data. The synthetic well-logs, generated by the models, are plotted alongside the measured ones. This allows us to have a first assessment of the quality of the predictions by visually inspecting how well the synthetic well-logs align with the measured well-logs.

2.5.7 Evaluation Phase

We use regression/evaluation metrics such as R^2 , MSE, RMSE, and MAE to quantitatively assess the performance of the ML models. These metrics allow for comparing the predicted well-log data with the measured well-log data. Having high accuracy values and low errors does not necessarily mean that a model is capable of generalizing for any dataset.

3 Methodology Application & Outcomes

In this chapter, a comprehensive presentation of the results obtained from the applied methodology is provided to observe the differences in the performance of the two models. We compare the measured well-log data with their corresponding synthetic counterparts, presenting the findings in the form of scatter plots and as a function of the dataset index and measured depth.

The results encompass the test dataset, the total dataset for the training well, and the total dataset for the test well. By including both the test and the total datasets, we aim to visualize the predicted curves not only in relation to the data sample index but also in terms of the measured depth.

To emphasize the most significant outcomes, we highlight the relevant plots in red.

3.1 Case Study 1: Effective Porosity Log

In this section, we aim to assess the ability of RF and GB models to predict a calculated well-log parameter, the effective porosity, thereby evaluating their understanding of the unknown deterministic function that relates the input well-log data and the target well-log variable. Our expectation is that the models perform well on training well 1-BRSA-871-MG, indicating their ability to learn the underlying function. Moreover, we anticipate that the models can accurately predict the data of the test well 1-BRSA-948-MG, as we assume the same underlying function applies to both wells.

It is important to note that the training dataset consists of 7596 data points (80% of 9496), while the test dataset comprises only 1288 data points. Therefore, we anticipate that the predictions for the second well should be highly accurate given the ample training data available.

The following results (Fig. 3.1, Fig. 3.2, Fig. 3.3, Fig. 3.4, Fig. 3.5, Fig. 3.6, Fig. 3.7, Fig. 3.8, Fig. 3.9) illustrate the outcomes of our analysis and support our evaluation of the model's prediction capabilities.

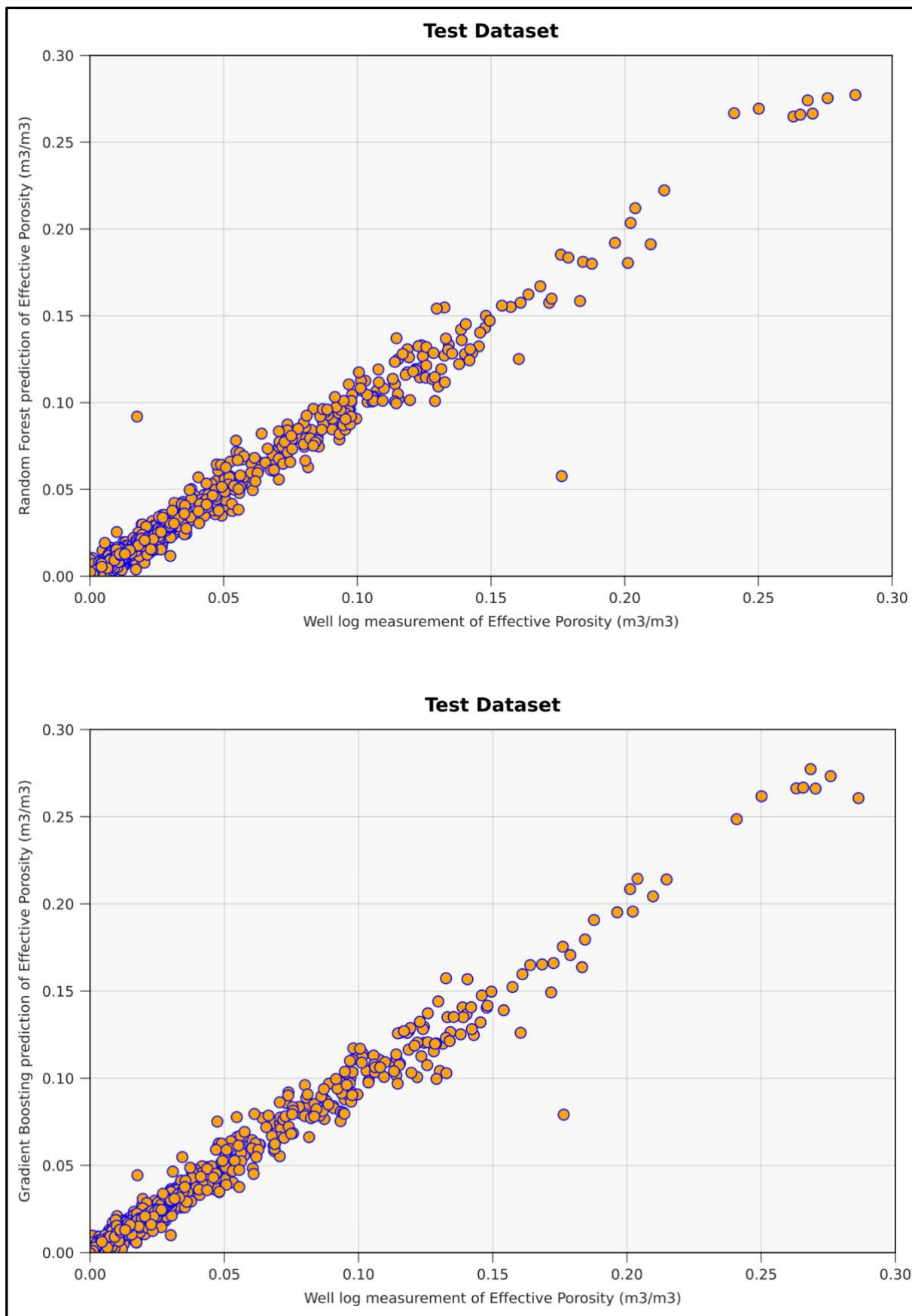


Fig. 3.1: Scatter plots of predicted versus measured effective porosity (m^3/m^3), for the RF and GB Models, applied to the Test Dataset of well 1-BRSA-871-MG.

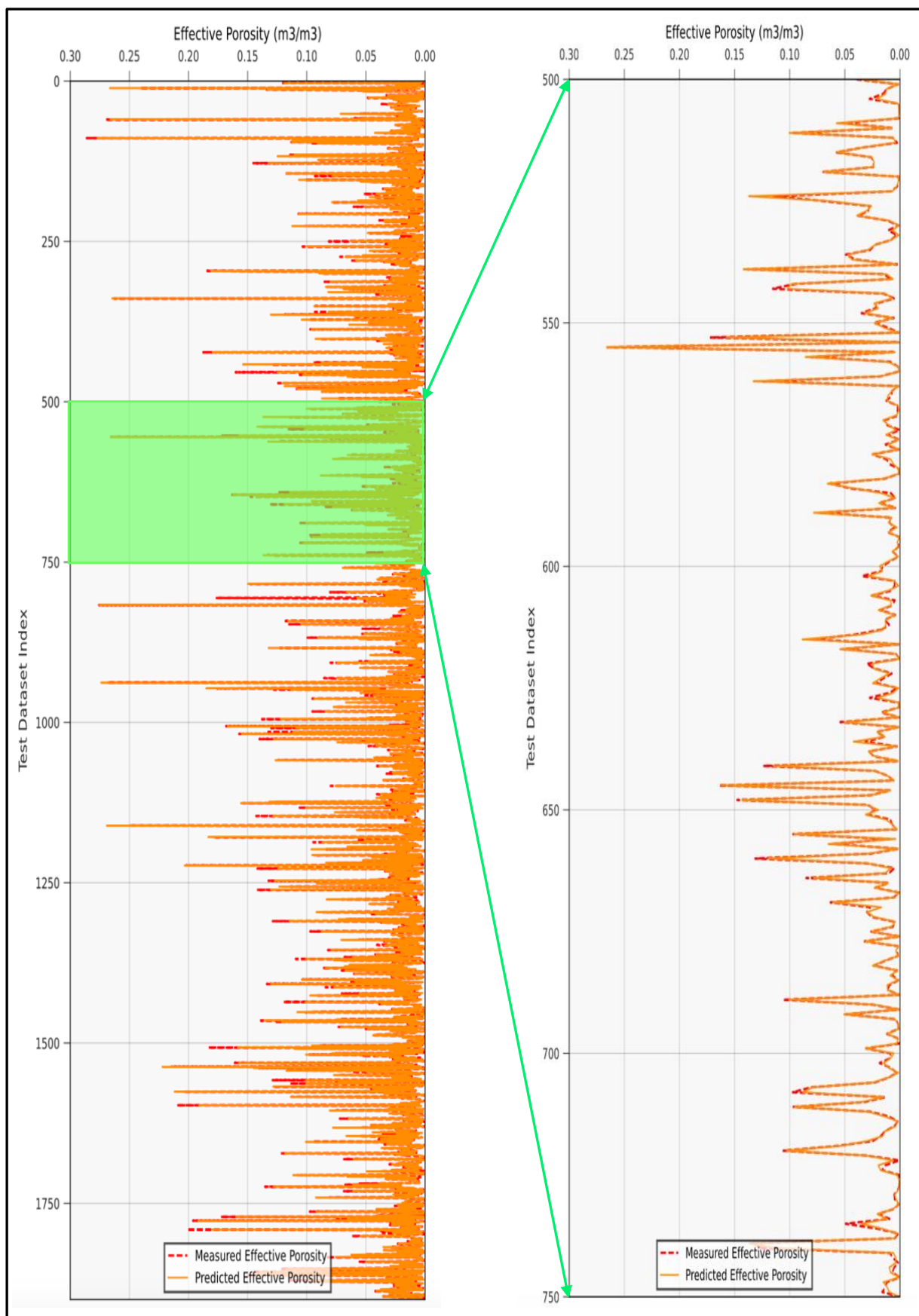


Fig. 3.2: Comparing the match between the predicted and measured effective porosity, for the RF Model, across the Test Dataset Index of well 1-BRSA-871-MG. Plot 1: Predicted and Measured Effective Porosity (m³/m³) for the range 0 - 1900. Plot 2: Predicted and Measured Effective Porosity (m³/m³) for the range 500 - 750.

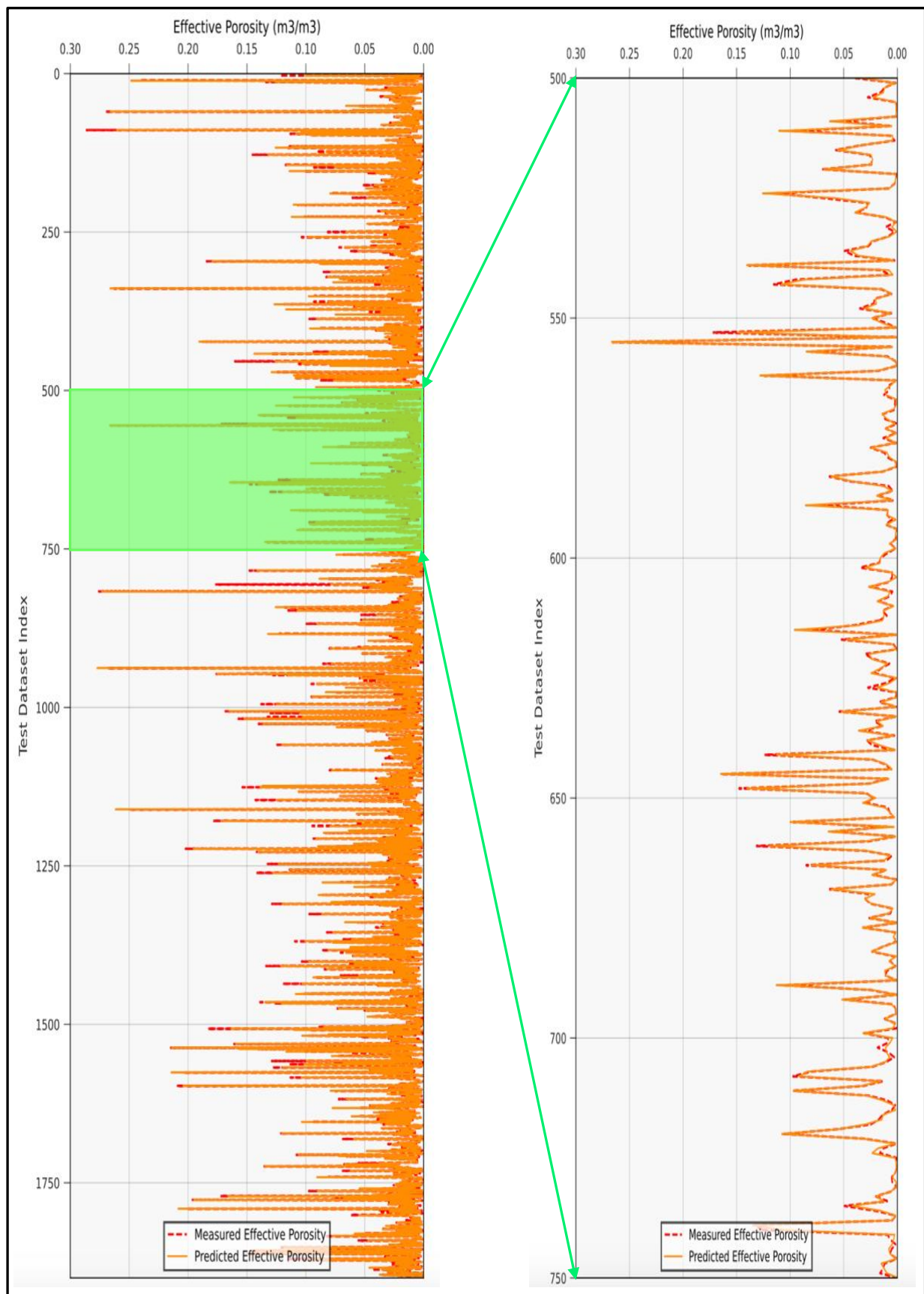


Fig. 3.3: Comparing the match between the predicted and measured effective porosity, for the GB Model, across the Test Dataset Index of well 1-BRSA-871-MG. Plot 1: Predicted and Measured Effective Porosity (m³/m³) for the range 0 - 1900. Plot 2: Predicted and Measured Effective Porosity (m³/m³) for the range 500 - 750.

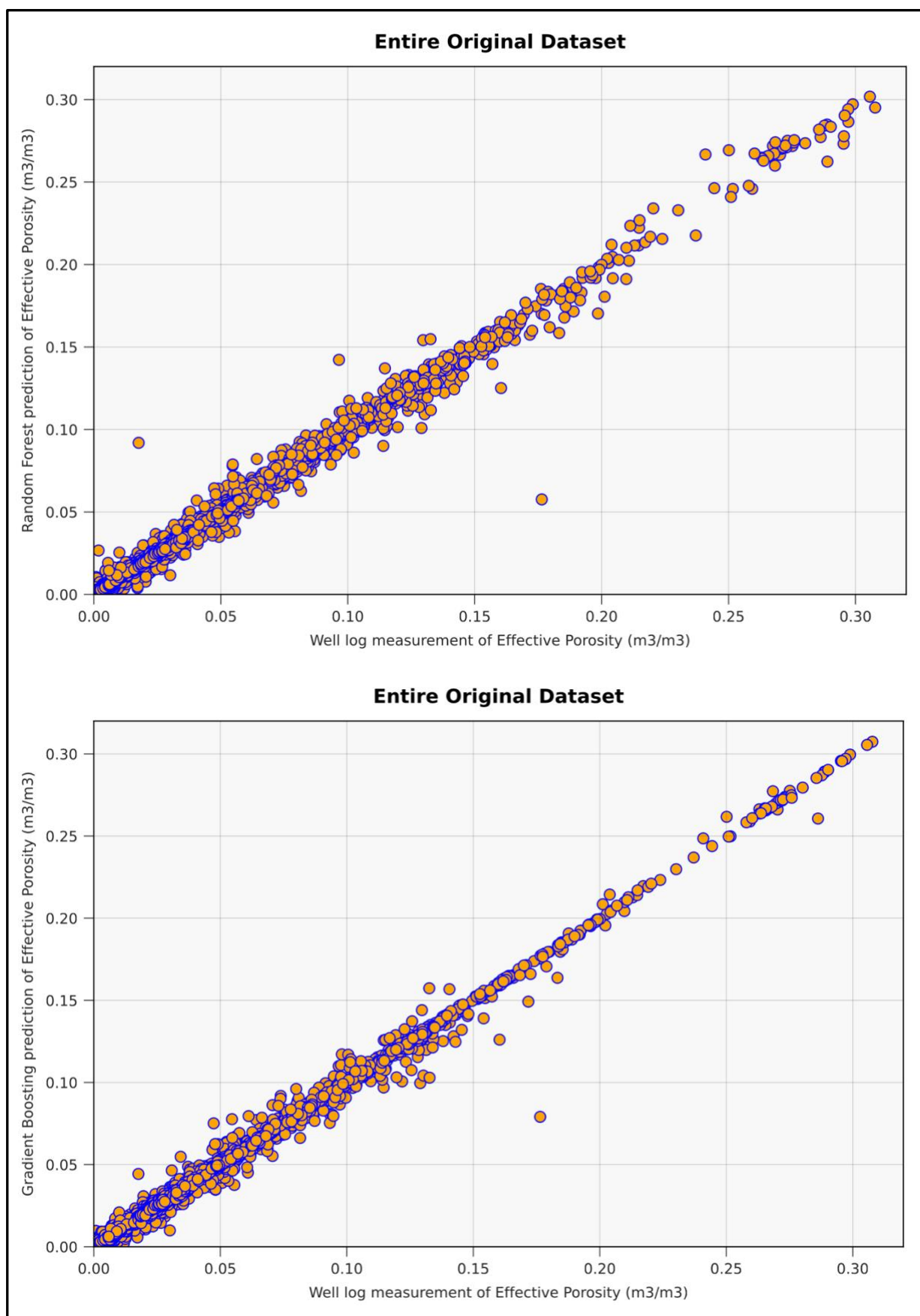


Fig. 3.4: Scatter plots of predicted versus measured effective porosity (m^3/m^3), for the RF and GB Models, applied to the Entire Original Dataset of well 1-BRSA-871-MG.

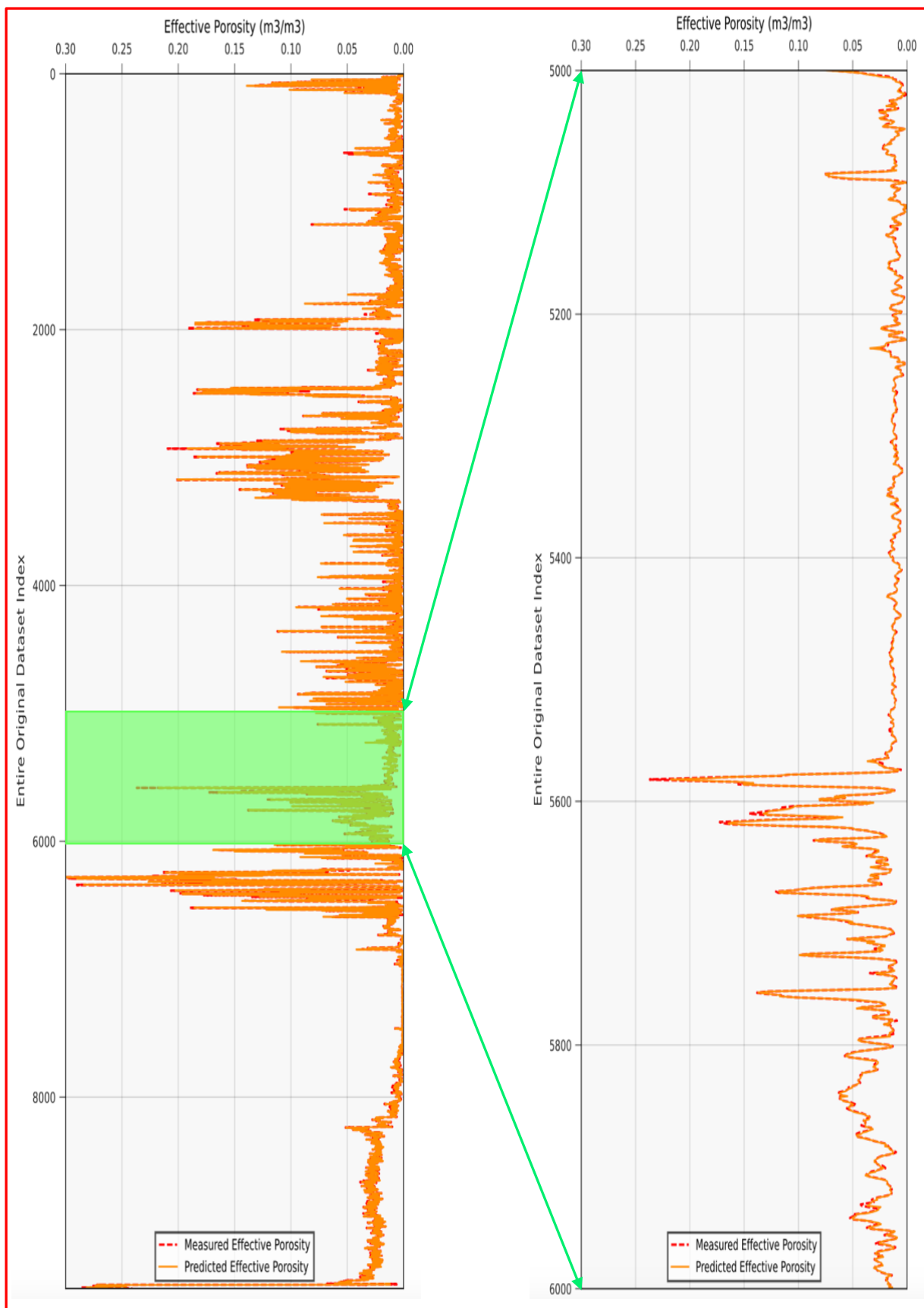


Fig. 3.5: Comparing the match between the predicted and measured effective porosity, for the RF Model, across the Entire Original Dataset Index of well 1-BRSA-871-MG, which corresponds to the Measured Well Depth. Plot 1: Predicted and Measured Effective Porosity (m^3/m^3) for the range 0 - 9496. Plot 2: Predicted and Measured Effective Porosity (m^3/m^3) for the range 5000 - 6000.

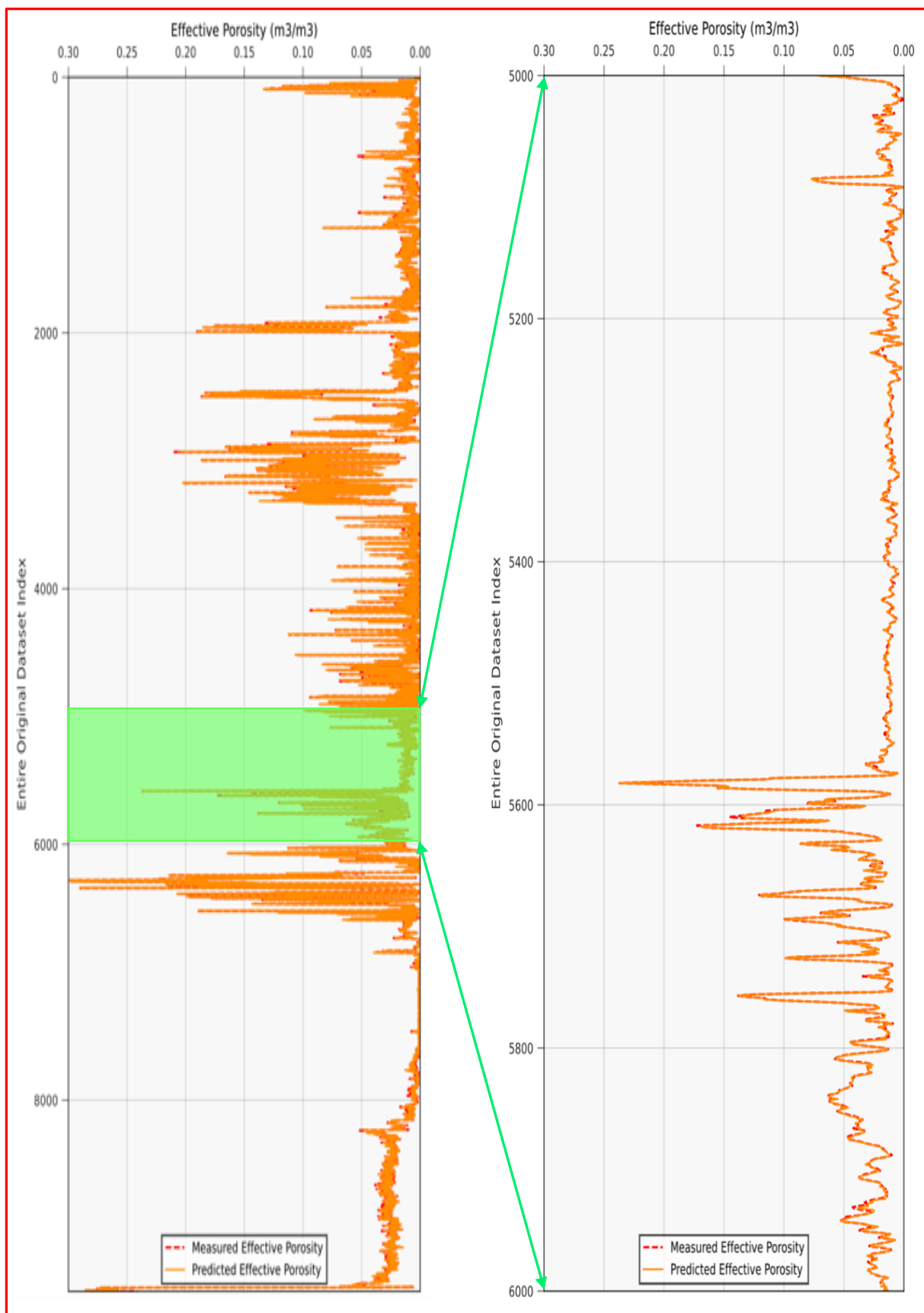


Fig. 3.6: Comparing the match between the predicted and measured effective porosity, for the GB Model, across the Entire Original Dataset Index of well 1-BRSA-871-MG, which corresponds to the Measured Well Depth. Plot 1: Predicted and Measured Effective Porosity (m^3/m^3) for the range 0 - 9496. Plot 2: Predicted and Measured Effective Porosity (m^3/m^3) for the range 5000 - 6000.

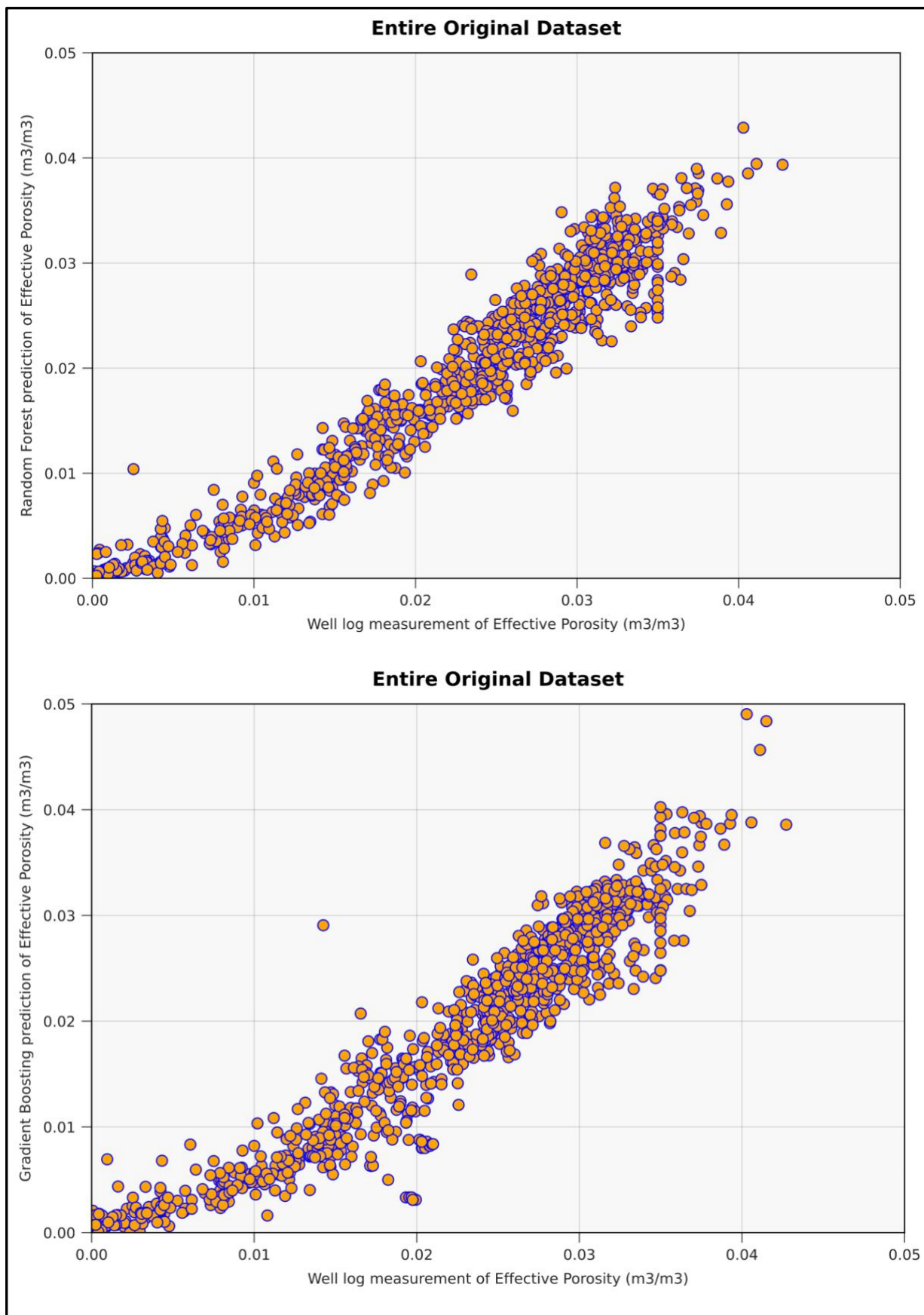


Fig. 3.7: Scatter plots of predicted versus measured effective porosity (m³/m³), for the RF and GB Models, applied to the Entire Original Dataset of well 1-BRSA-948-MG.

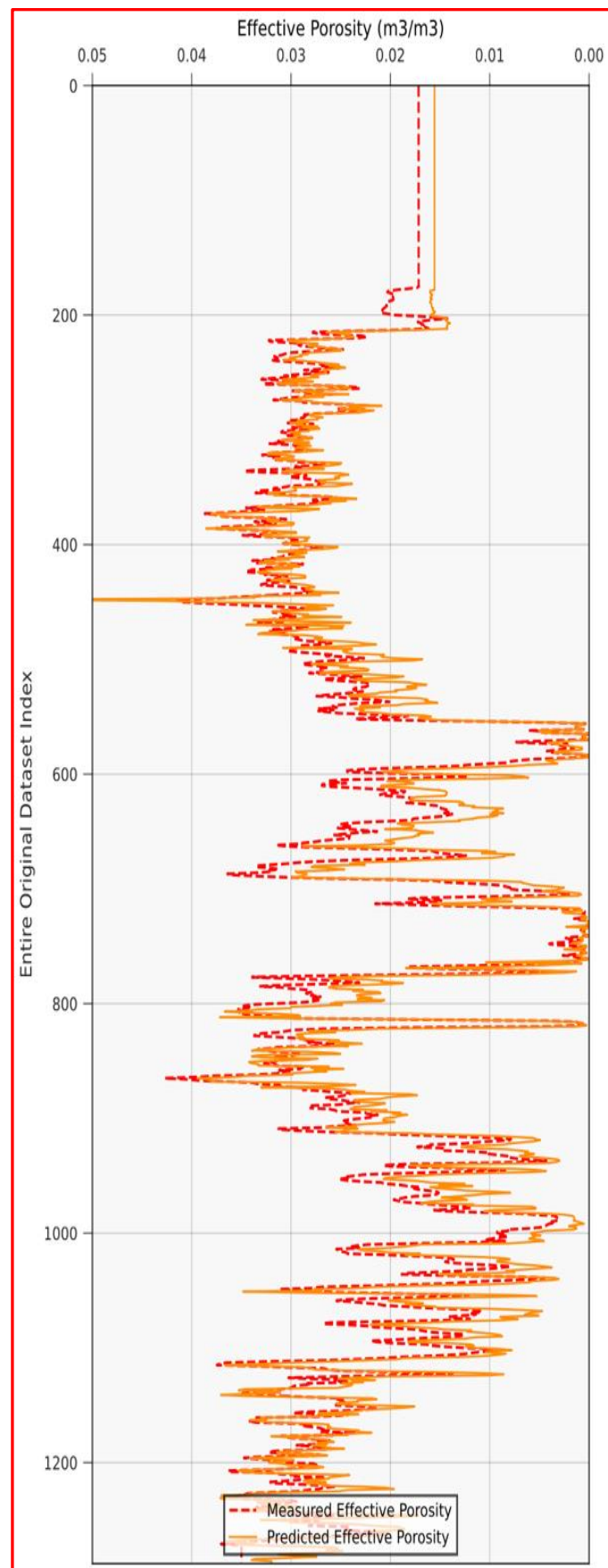


Fig. 3.8: Comparing the match between the predicted and measured effective porosity, for the RF Model, across the Entire Original Dataset Index of well 1-BRSA-948-MG, which corresponds to the Measured Well Depth.

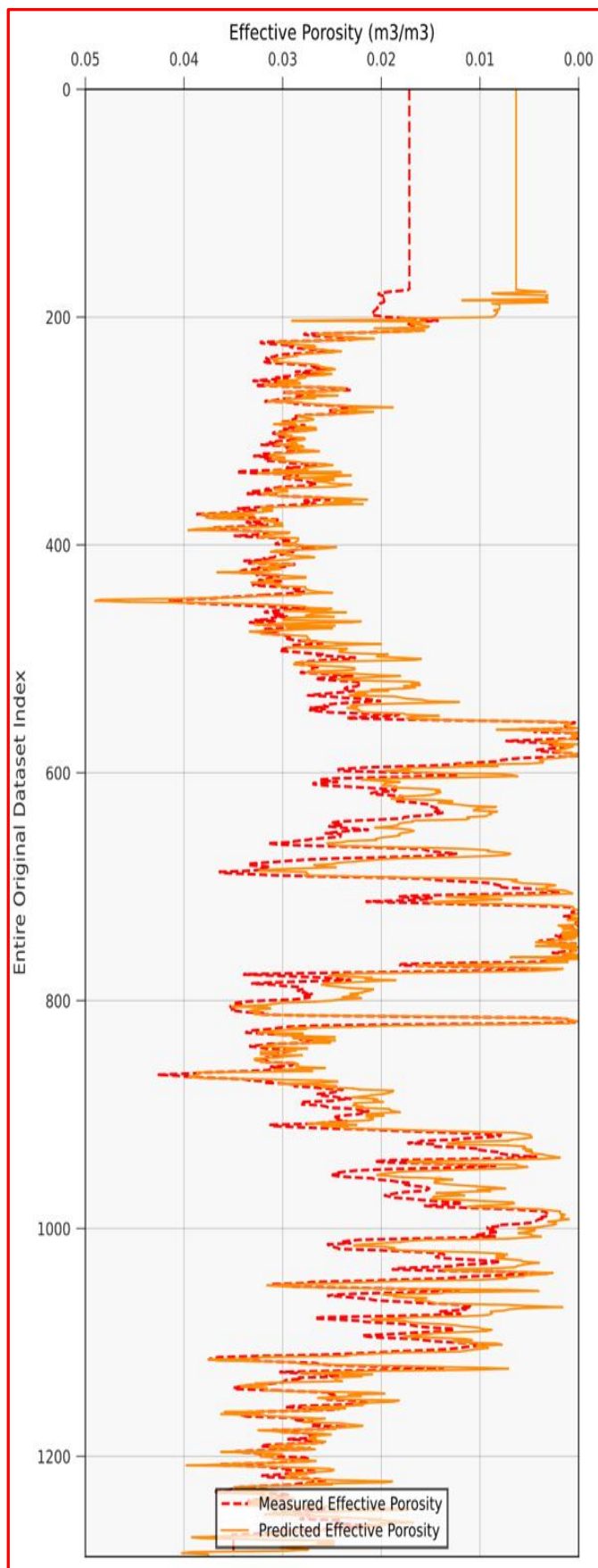


Fig. 3.9: Comparing the match between the predicted and measured effective porosity, for the GB Model, across the Entire Original Dataset Index of well 1-BRSA-948-MG, which corresponds to the Measured Well Depth.

A qualitative interpretation of the plotted results confirms our expectations, as both ML models demonstrate excellent performance on both training and test wells. The visual analysis of the results indicates that the developed ML models effectively capture the underlying patterns and relationships within the data, leading to accurate predictions. The satisfactory performance of the models on unseen test data validates their robustness and generalizability.

3.2 Case Study 1: Compressional Wave Slowness Log

In this section, we focus on evaluating the performance of the RF and GB models in predicting compressional wave slowness, a measured well-log parameter. Considering the limitations of our training dataset, which comprises only 916 data points (80% of 1146), we anticipate that the models will perform well on the training well 1-BRSA-871-MG but may struggle on the test well 1-BRSA-948-MG. This is further compounded by the fact that the range of the compressional wave slowness log we aim to predict differs from the one used in the training phase.

The results presented in Fig. 3.10, Fig. 3.11, Fig. 3.12, Fig. 3.13, Fig. 3.14, Fig. 3.15, Fig. 3.16, Fig. 3.17, and Fig. 3.18 showcase the outcomes of our analysis, confirm our expectations, and provide insights into the initial limitations of our models concerning the size of the training dataset and the range of prediction.

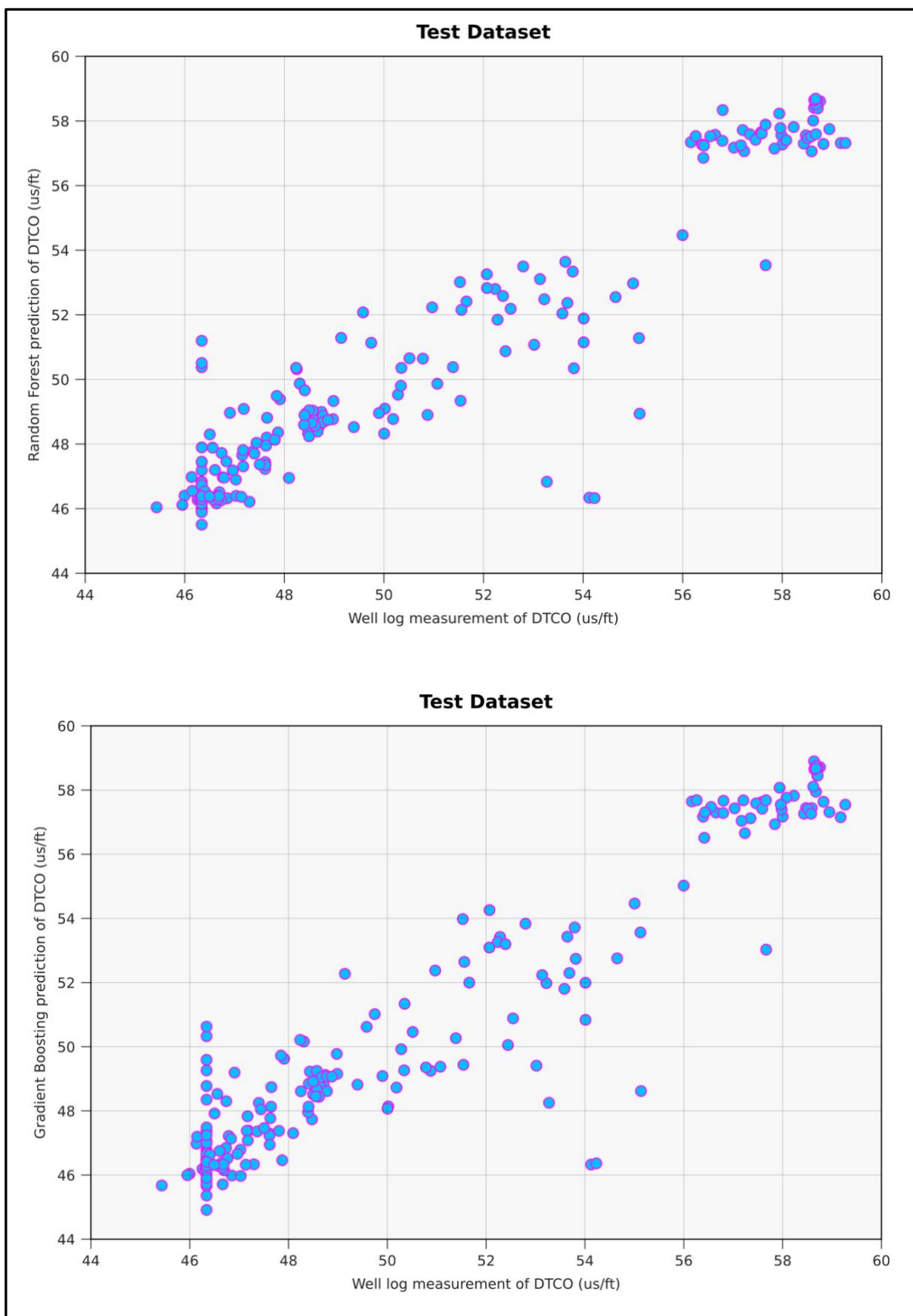


Fig. 3.10: Scatter plots of predicted versus measured compressional wave slowness ($\mu\text{s}/\text{ft}$), for the RF and GB Models, applied to the Test Dataset of well 1-BRSA-871-MG.

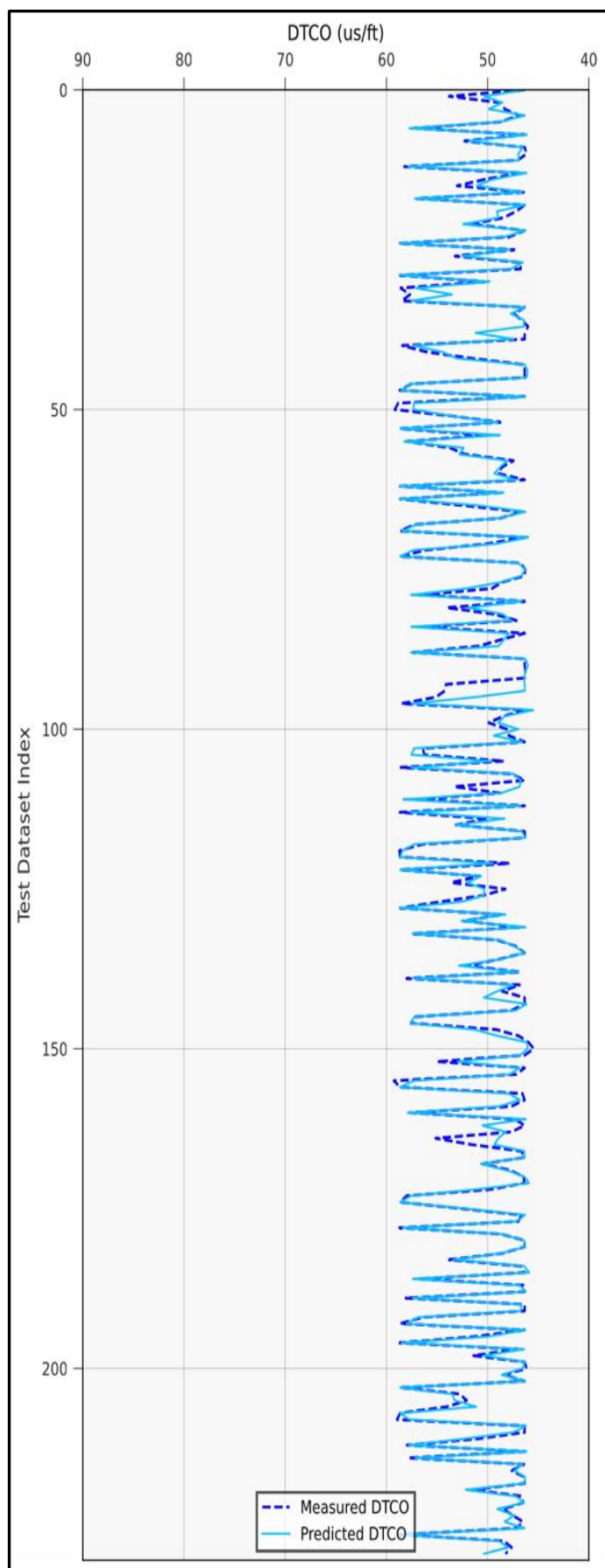


Fig. 3.11: Comparing the match between the predicted and measured compressional wave slowness ($\mu\text{s}/\text{ft}$), for the RF Model, across the Test Dataset Index of well 1-BRSA-871-MG.

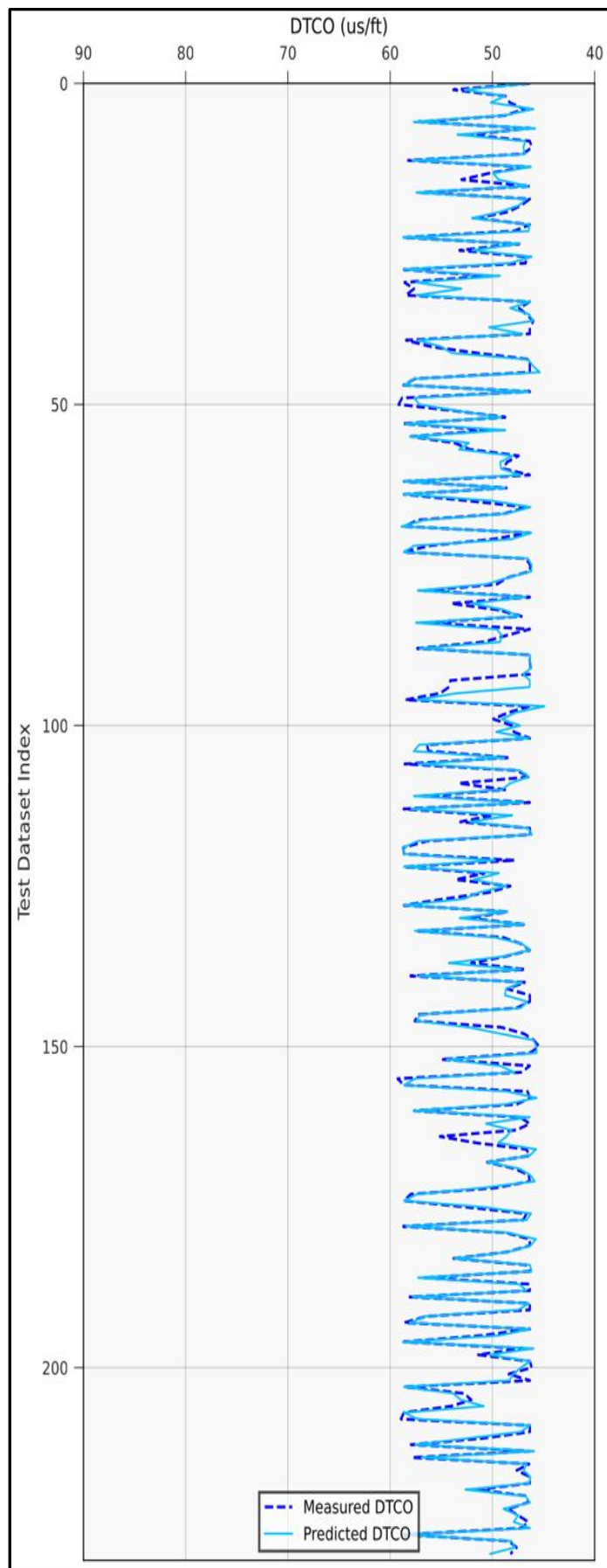


Fig. 3.12: Comparing the match between the predicted and measured compressional wave slowness ($\mu\text{s}/\text{ft}$), for the GB Model, across the Test Dataset Index of well 1-BRSA-871-MG.

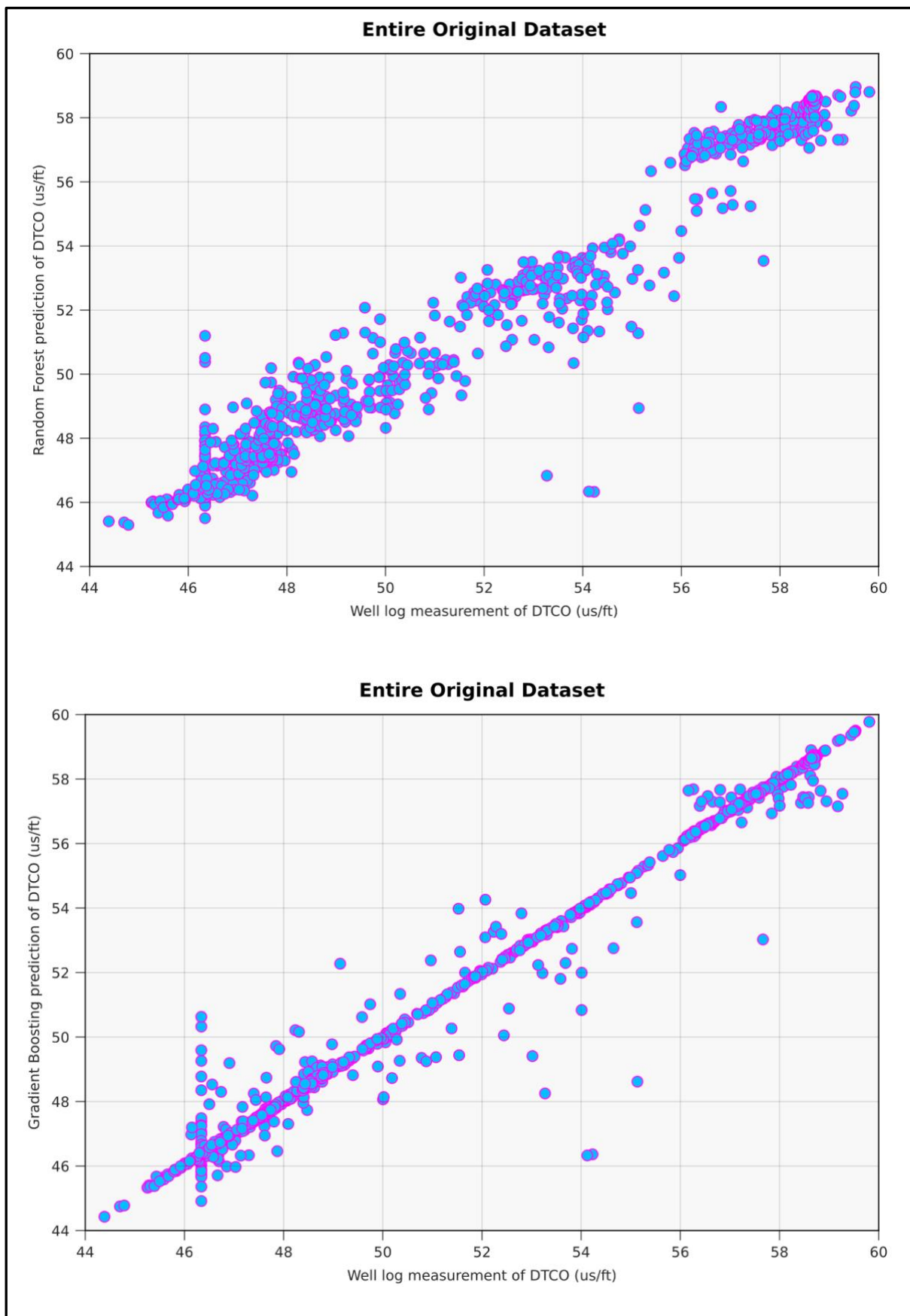


Fig. 3.13: Scatter plots of predicted versus measured compressional wave slowness ($\mu\text{s}/\text{ft}$), for the RF and GB Models, applied to the Entire Original Dataset of well 1-BRSA-871-MG.

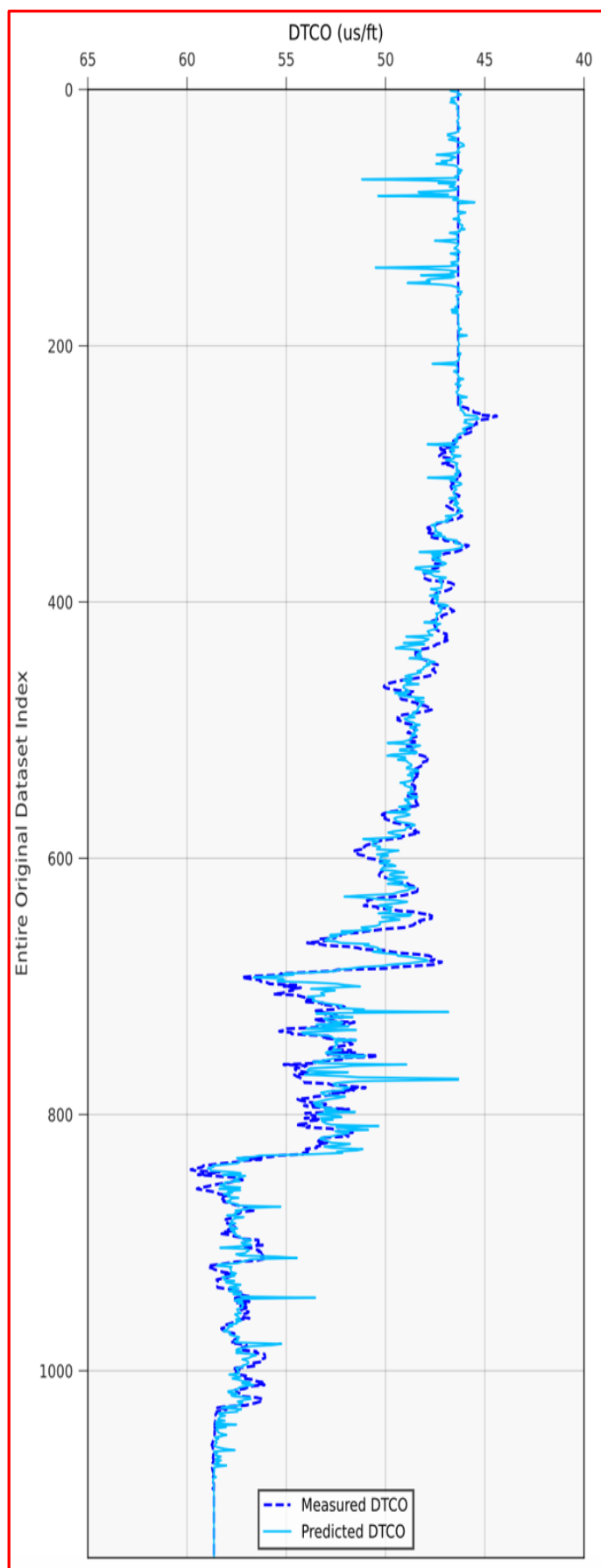


Fig. 3.14: Comparing the match between the predicted and measured compressional wave slowness ($\mu\text{s}/\text{ft}$), for the RF Model, across the Entire Original Dataset Index of well 1-BRSA-871-MG.

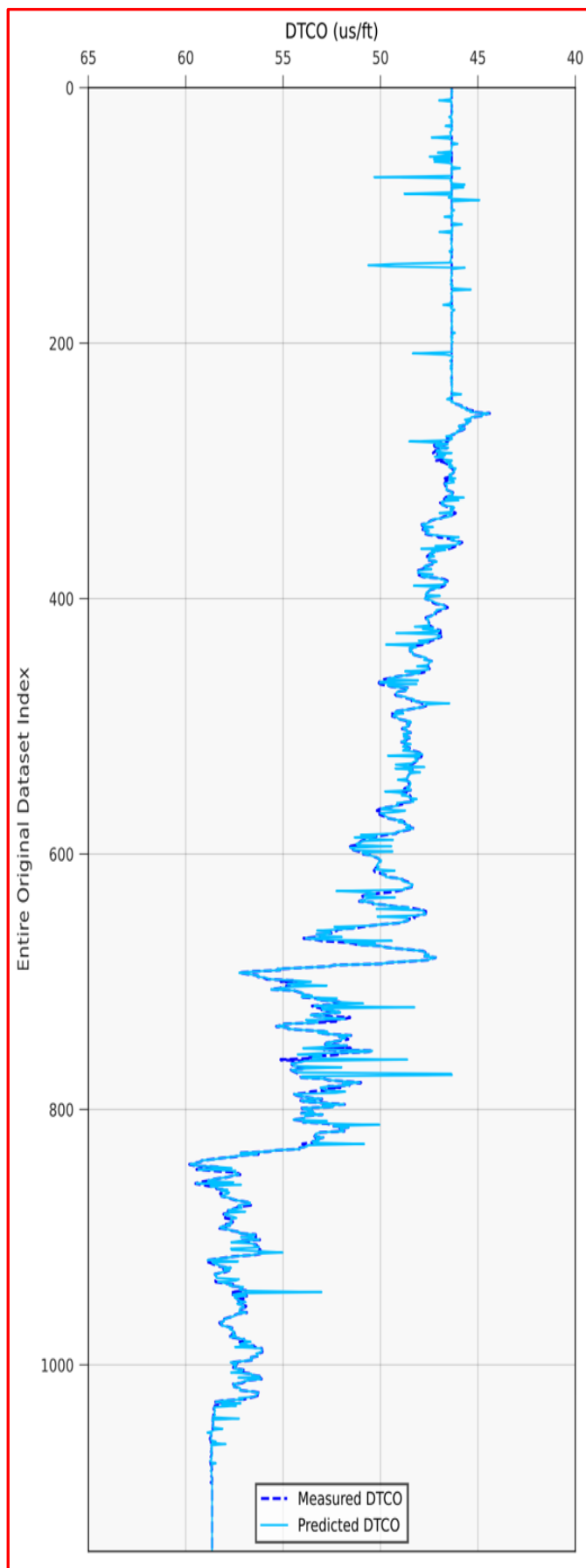


Fig. 3.15: Comparing the match between the predicted and measured compressional wave slowness ($\mu\text{s}/\text{ft}$), for the RF Model, across the Entire Original Dataset Index of well 1-BRSA-871-MG.

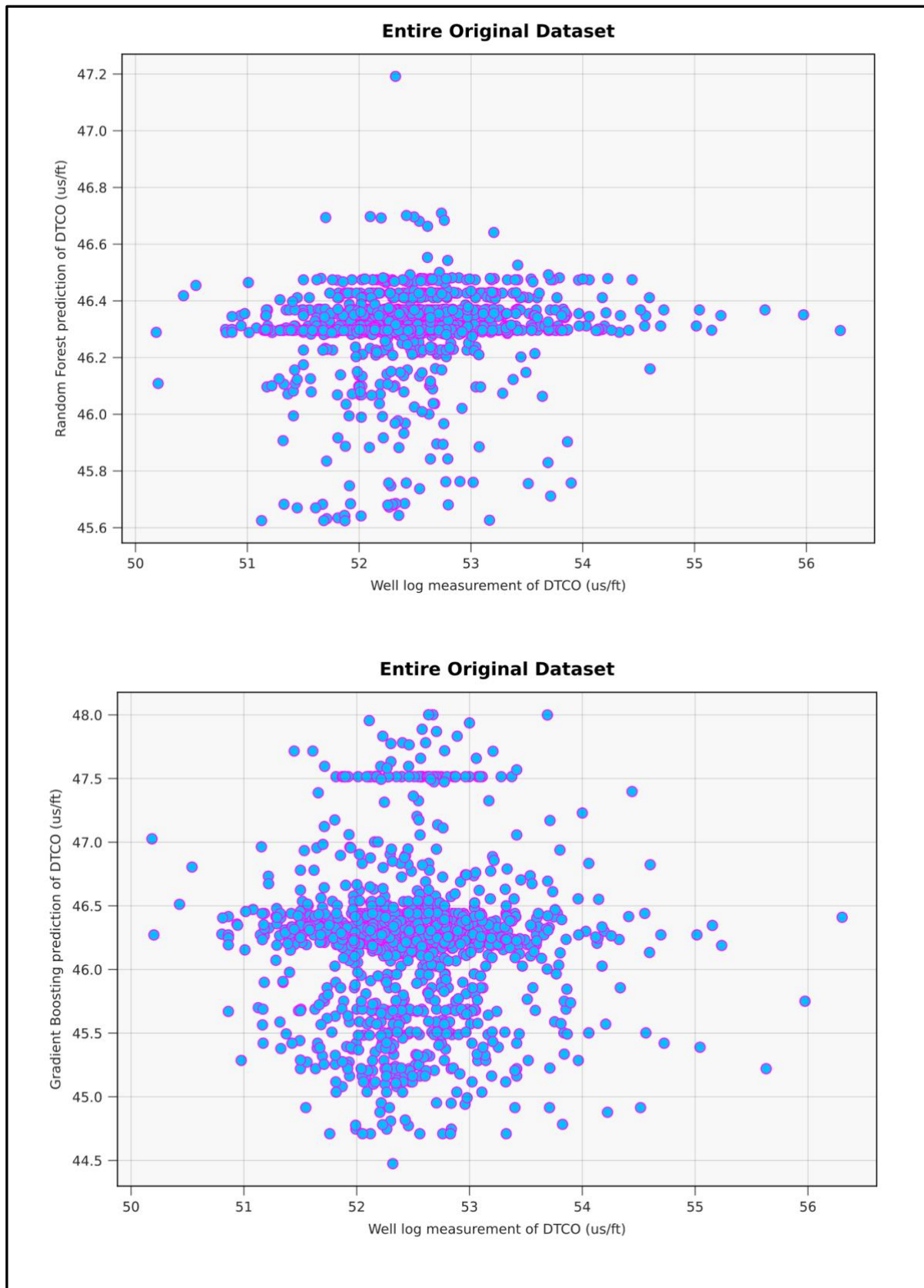


Fig. 3.16: Scatter plots of predicted versus measured compressional wave slowness ($\mu\text{s}/\text{ft}$), for the RF and GB Models, applied to the Entire Original Dataset of well 1-BRSA-948-MG.

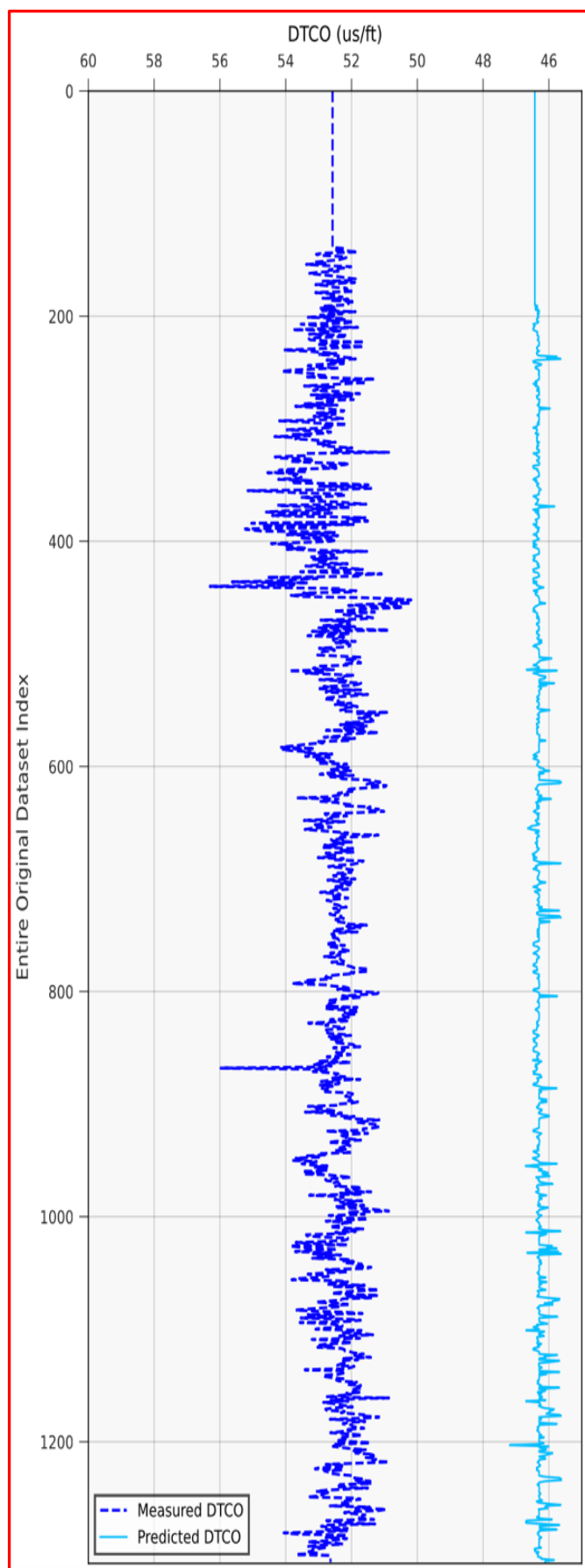


Fig. 3.17: Comparing the match between the predicted and measured compressional wave slowness ($\mu\text{s}/\text{ft}$), for the RF Model, across the Entire Original Dataset Index of well 1-BRSA-948-MG.

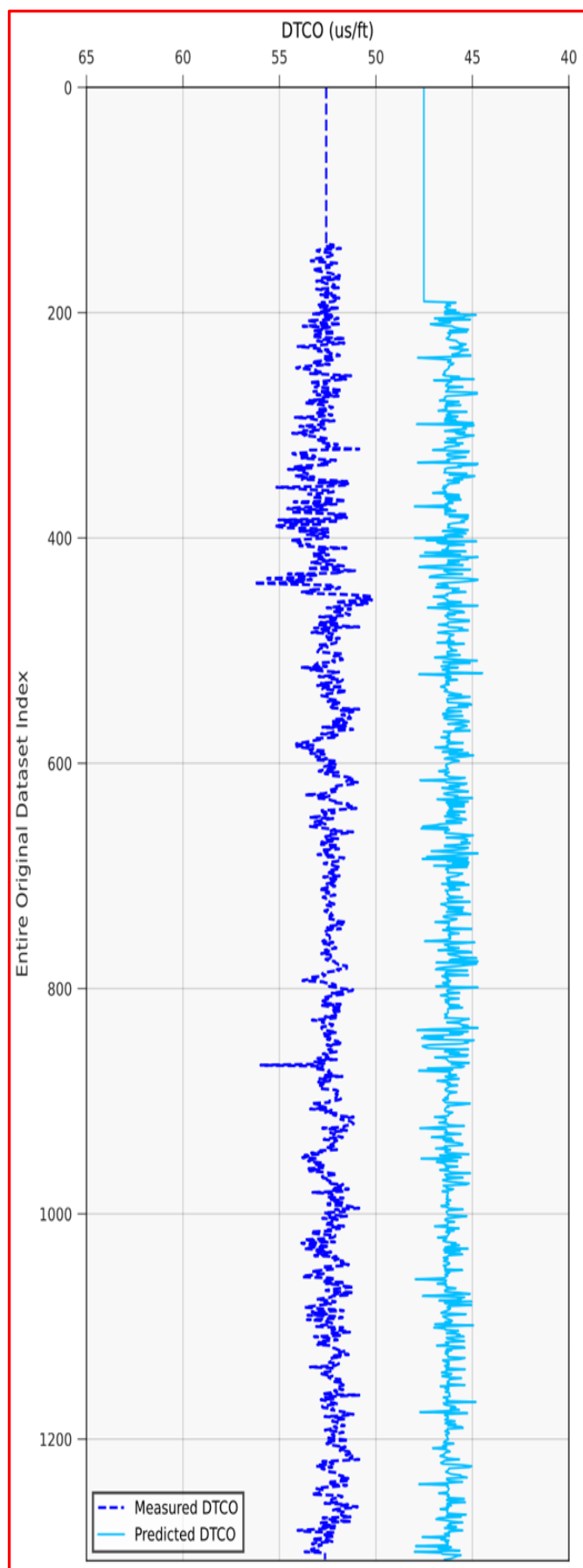


Fig. 3.18: Comparing the match between the predicted and measured compressional wave slowness ($\mu\text{s}/\text{ft}$), for the RF Model, across the Entire Original Dataset Index of well 1-BRSA-948-MG.

3.3 Case Study 2: NMR Porosity Logs

In this section, our focus is on evaluating the performance of the RF and GB models in predicting NMR porosities, as measured well-log parameters. We want to conduct a more accurate analysis, and for this reason, we undertake two attempts to assess the model's predictive capabilities. By conducting two attempts, we can gather more insights and validate the robustness of the models in different scenarios.

In the first attempt, we use well 3-BRSA-1215-RJS as the training well, and well 1-BRSA-1116-RJS as the test well. Conversely, in the second attempt, we reverse the roles, using well 1-BRSA-1116-RJS as the training well and well 3-BRSA-1215-RJS as the validation well. Our supposition is that, in both cases, the models will exhibit excellent predictive performance on the training well. However, we expect to achieve improved predictions on the test well on the second attempt, and there are several reasons to support this expectation.

Firstly, well 1-BRSA-1116-RJS is predominantly composed of good reservoir rock, accounting for 90% of its composition, with the remaining 10% being a non-reservoir rock. This indicates that it exhibits favorable properties for hydrocarbon reservoirs. Conversely, well 3-BRSA-1116-RJS consists of only 20% of reservoir rock, with the remaining 80% being non-reservoir rock. This distinction implies that the model trained on well 1-BRSA-1116-RJS may have a higher capability to predict the reservoir portion of the test well, which represents 20% of the total formation. And that the model trained on well 3-BRSA-1215-RJS may not be able to predict the reservoir portion of the test well, which is the predominant portion.

Secondly, well 1-BRSA-1116-RJS has a larger starting dataset of 2282 datapoints, providing a substantial training dataset compared to well 3-BRSA-1215-RJS, which has 1579 data points and a smaller training dataset. This difference in dataset size suggests that the model trained on well 1-BRSA-1116-RJS may have more information to learn from, potentially leading to better predictive performance.

Consequently, considering these factors, we anticipate that the second attempt will result in improved predictions on the test well compared to the first attempt. For this specific case, the larger training dataset, and the higher concentration of reservoir rock, in the training well, provide an advantage for the model to learn, enabling it to make more accurate predictions on unseen data.

However, it is worth noting that the predictions on the test well may still be weak, even in the second attempt, due to significant differences in the NMR porosity distributions between the training and the test well. The dissimilarity in the distributions, despite the similarity in lithology between the two wells, suggests potential challenges in accurately predicting NMR porosities. This difference is highlighted in a histogram overview, specifically shown in Fig. 3.19, illustrating the disparities in the porosity distributions of the training and test wells.

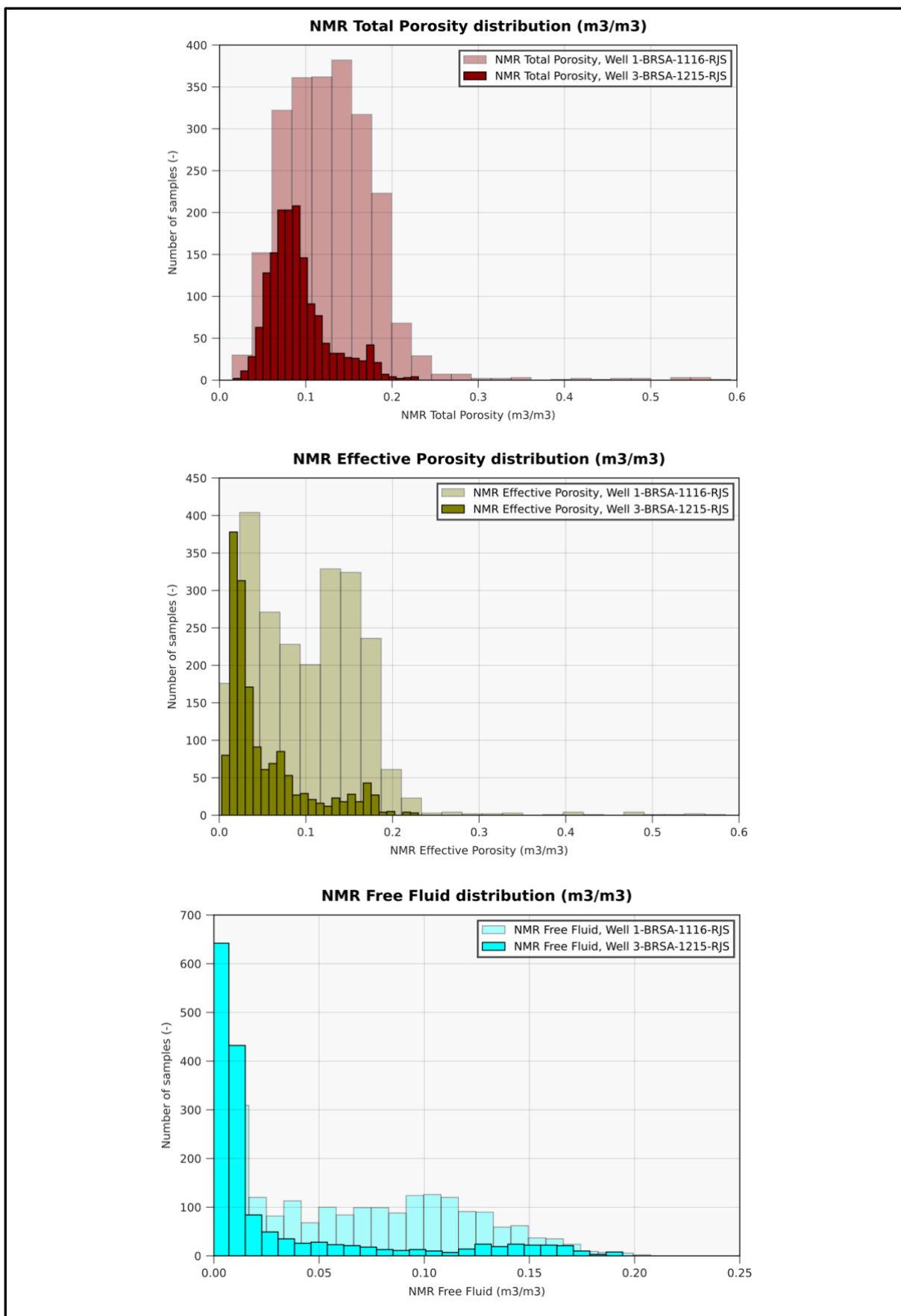


Fig. 3.19: NMR Porosity distribution for well 1-BRSA-1116-RJS, and well 3-BRSA-1215-RJS. Histogram 1: NMR Total Porosity distribution (m³/m³). Histogram 2: NMR Effective Porosity distribution (m³/m³). Histogram 3: NMR Free Fluid distribution (m³/m³).

3.3.1 First Attempt: Training on Well 3-BRSA-1215-RJS and Validation on Well 1-BRSA-1116-RJS

The results depicted in Fig. 3.20, Fig. 3.21, Fig. 3.22, Fig. 3.23, Fig. 3.24, Fig. 3.25, Fig. 3.26, Fig. 3.27, Fig. 3.28, Fig. 3.29, Fig. 3.30, Fig. 3.31 provide further evidence to support our previous statements. These figures illustrate the outcomes of our analysis and reinforce the observations made regarding the predictive capabilities of the RF and GB models on test well. The visual comparison between the predicted and the measured well-logs confirms that our expectations are indeed met. However, for a more accurate analysis, we need to rely on regression metrics, which will be presented at a later stage, to provide a quantitative assessment of the model's performance.

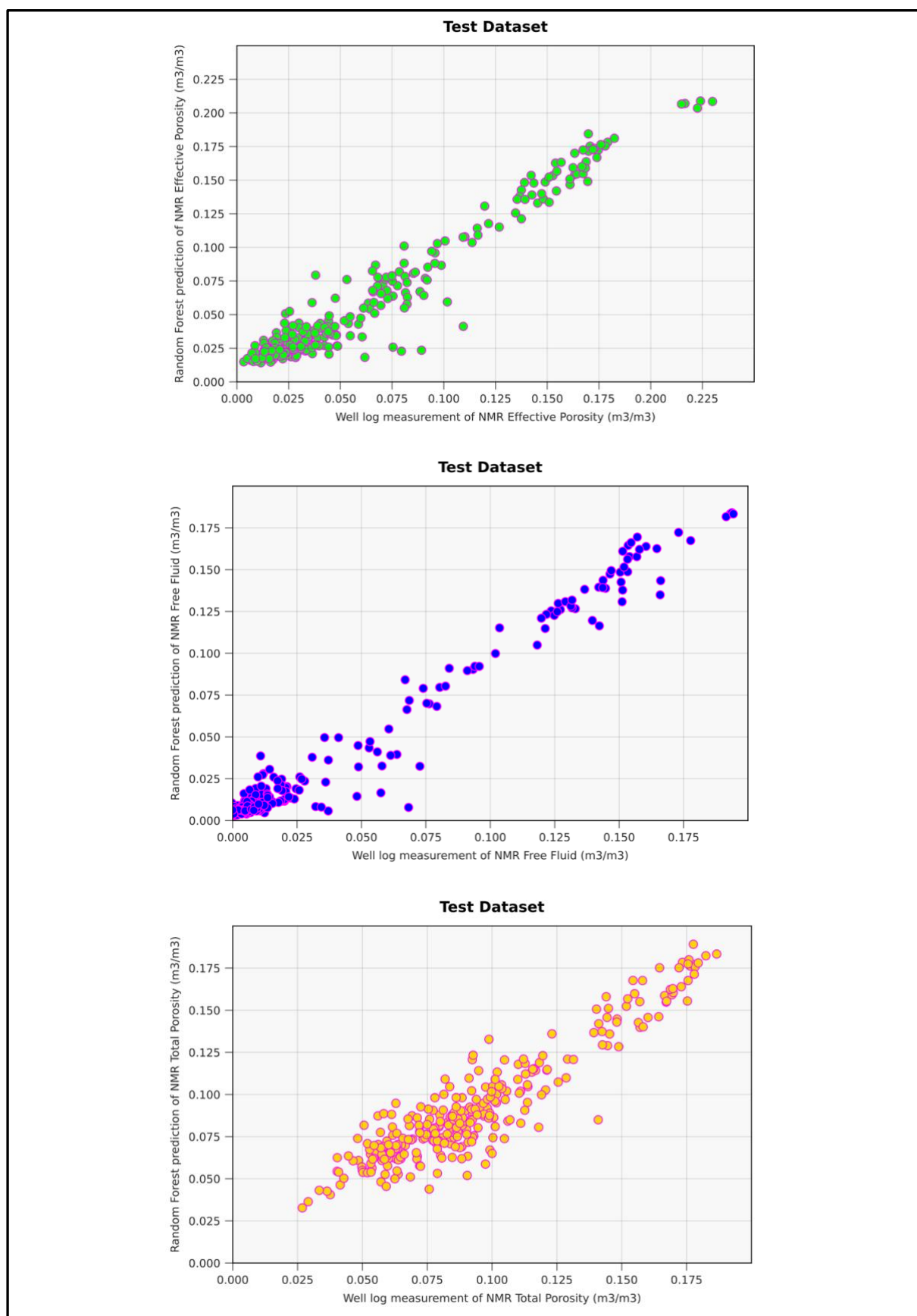


Fig. 3.20: Scatter plots of predicted versus measured NMR porosity, for the RF Model, applied to the Test Dataset of well 3-BRSA-1215-RJS. Plot 1: NMR Effective Porosity (m^3/m^3). Plot 2: NMR Free Fluid (m^3/m^3). Plot 3: NMR Total Porosity (m^3/m^3).

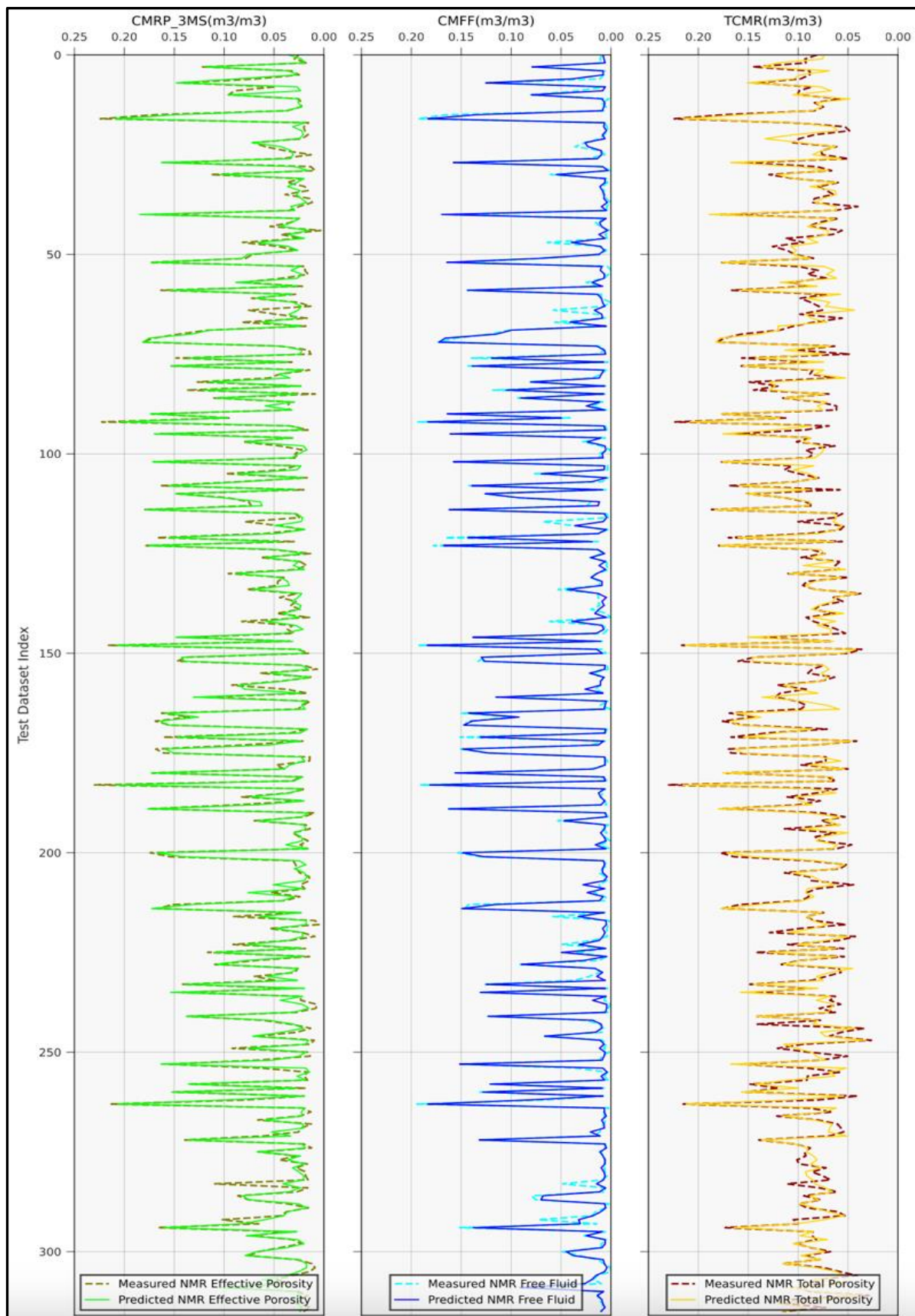


Fig. 3.21: Comparing the match between the predicted and measured NMR porosity, for the RF Model, across the Test Dataset Index of well 3-BRSA-1215-RJS. Track 1: Predicted and Measured NMR Effective Porosity (m^3/m^3). Track 2: Predicted and Measured NMR Free Fluid (m^3/m^3). Track 3: Predicted and Measured NMR Total Porosity (m^3/m^3).

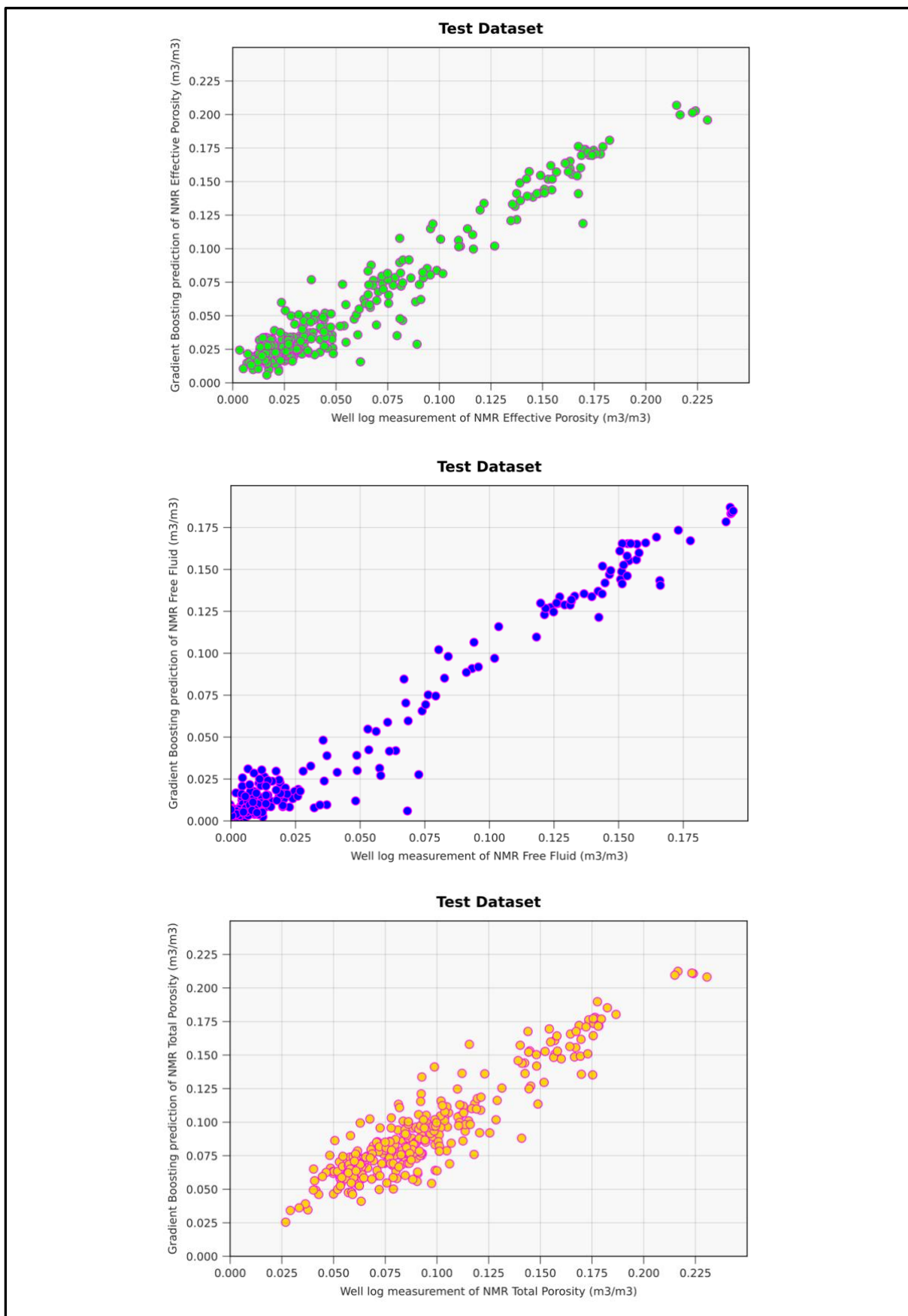


Fig. 3.22: Scatter plots of predicted versus measured NMR porosity, for the GB Model, applied to the Test Dataset of well 3-BRSA-1215-RJS. Plot 1: NMR Effective Porosity (m³/m³). Plot 2: NMR Free Fluid (m³/m³). Plot 3: NMR Total Porosity (m³/m³).

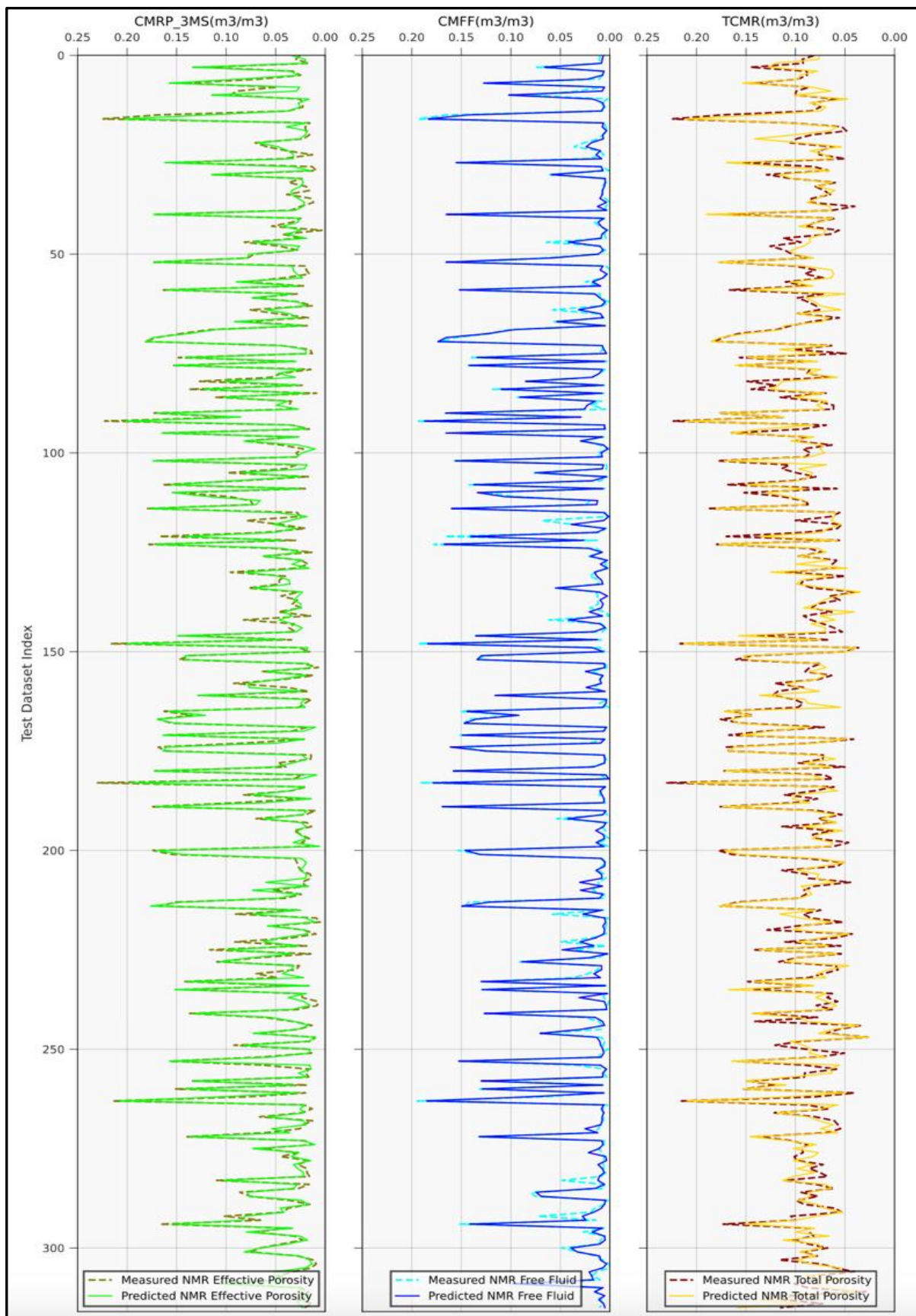


Fig. 3.23: Comparing the match between the predicted and measured NMR porosity, for the GB Model, across the Test Dataset Index of well 3-BRSA-1215-RJS. Track 1: Predicted and Measured NMR Effective Porosity (m^3/m^3). Track 2: Predicted and Measured NMR Free Fluid (m^3/m^3). Track 3: Predicted and Measured NMR Total Porosity (m^3/m^3).

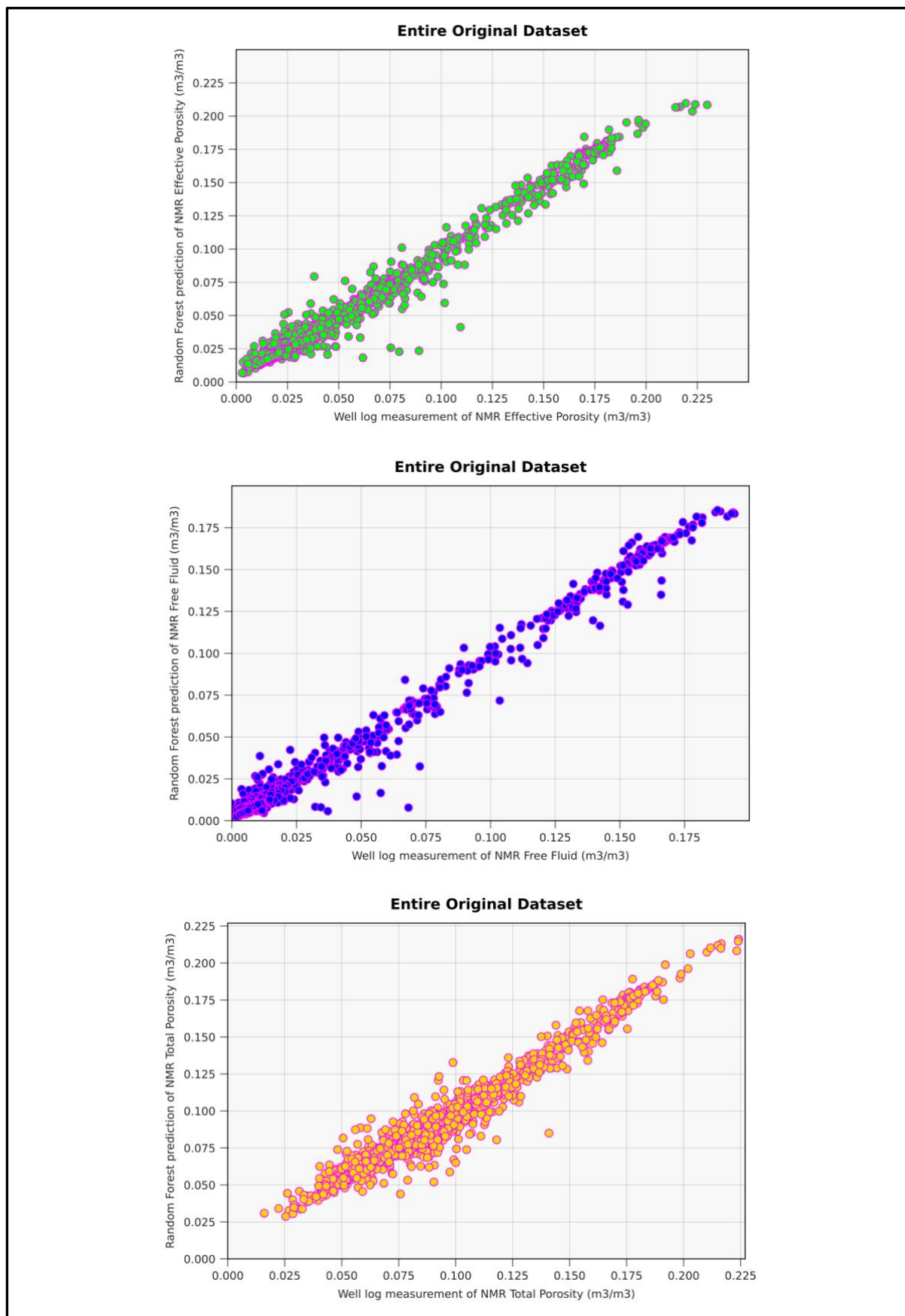


Fig. 3.24: Scatter plots of predicted versus measured NMR porosity, for the RF Model, applied to the Entire Original Dataset of well 3-BRSA-1215-RJS. Plot 1: NMR Effective Porosity (m³/m³). Plot 2: NMR Free Fluid (m³/m³). Plot 3: NMR Total Porosity (m³/m³).

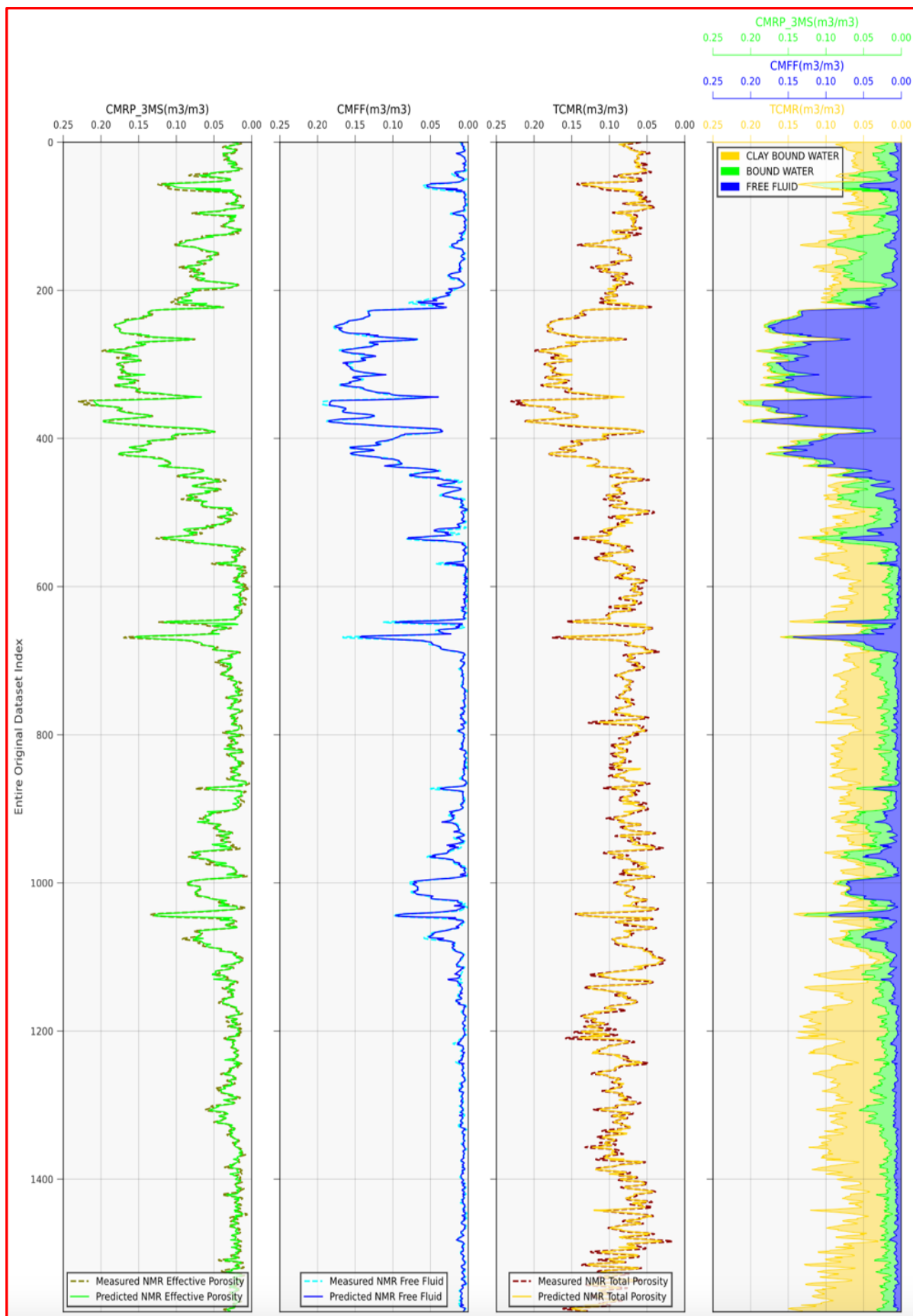


Fig. 3.25: Comparing the match between the predicted and measured NMR porosity, for the RF Model, across the Entire Original Dataset Index of well 3-BRSA-1215-RJS, which corresponds to the Measured Well Depth. Track 1: Predicted and Measured NMR Effective Porosity (m^3/m^3). Track 2: Predicted and Measured NMR Free Fluid (m^3/m^3). Track 3: Predicted and Measured NMR Total Porosity (m^3/m^3).

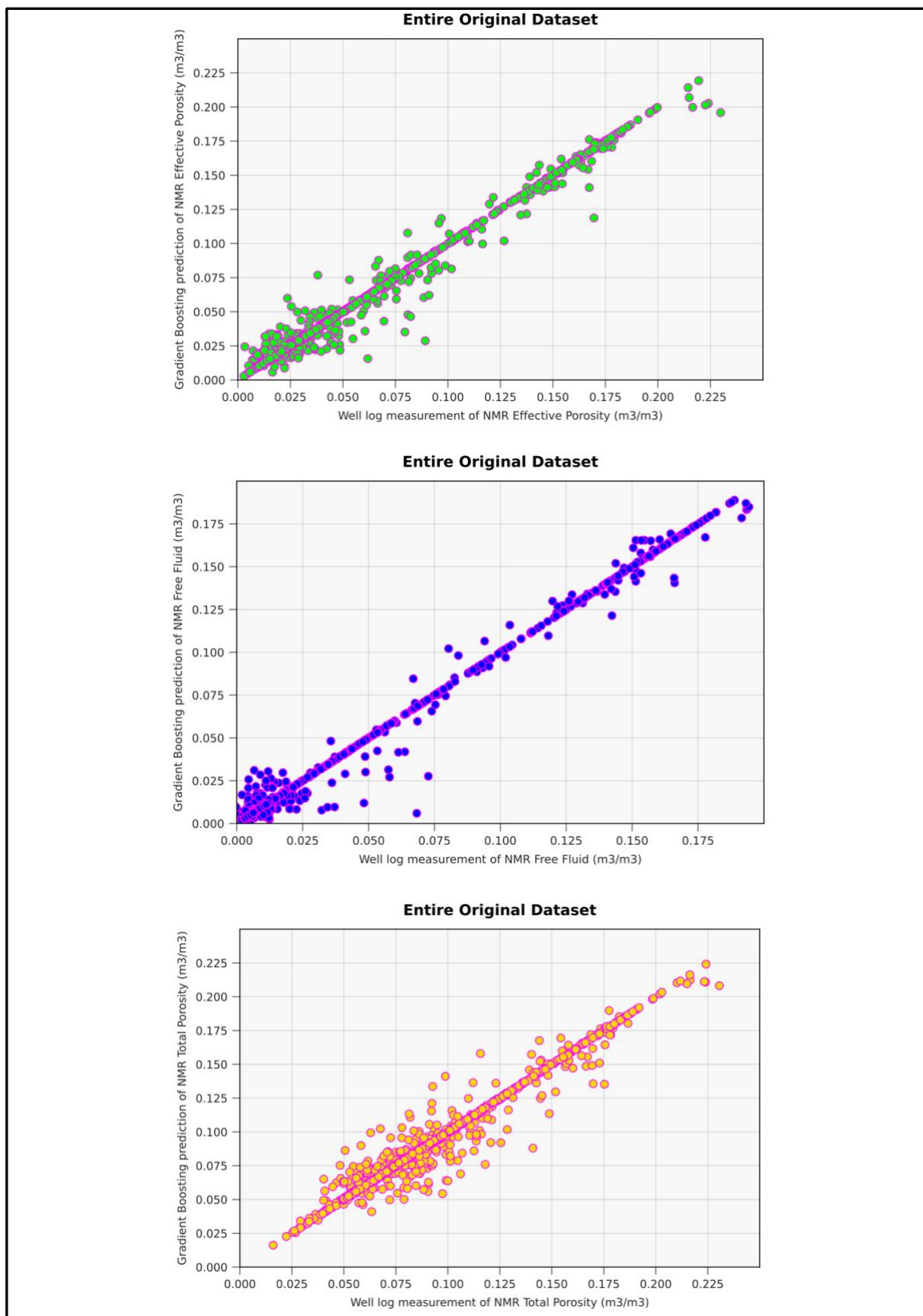


Fig. 3.26: Scatter plots of predicted versus measured NMR porosity, for the GB Model, applied to the Entire Original Dataset of well 3-BRSA-1215-RJS. Plot 1: NMR Effective Porosity (m³/m³). Plot 2: NMR Free Fluid (m³/m³). Plot 3: NMR Total Porosity (m³/m³).

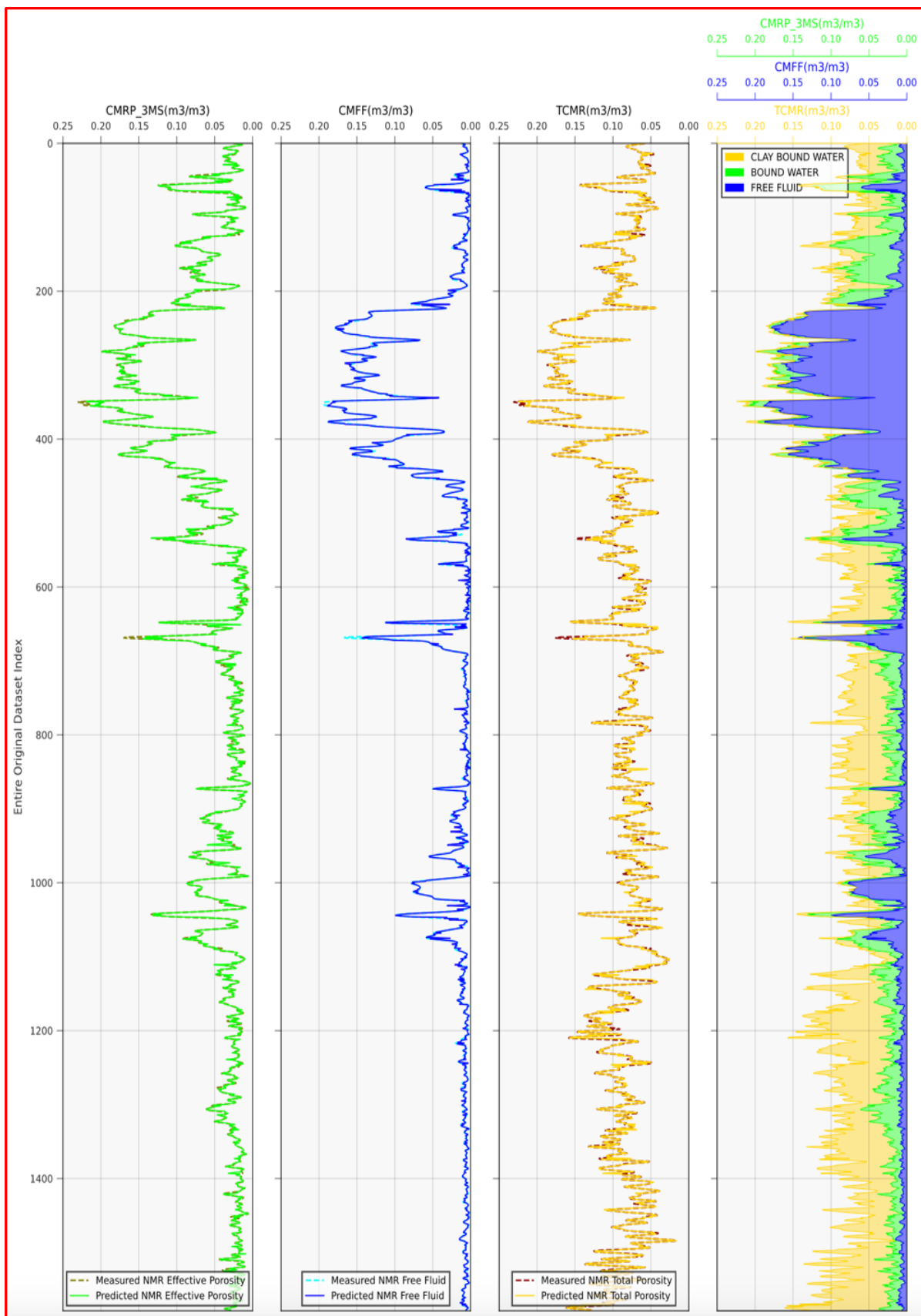


Fig. 3.27: Comparing the match between the predicted and measured NMR porosity, for the GB Model, across the Entire Original Dataset Index of well 3-BRSA-1215-RJS, which corresponds to the Measured Well Depth. Track 1: Predicted and Measured NMR Effective Porosity (m^3/m^3). Track 2: Predicted and Measured NMR Free Fluid (m^3/m^3). Track 3: Predicted and Measured NMR Total Porosity (m^3/m^3).



Fig. 3.28: Scatter plots of predicted versus measured NMR porosity, for the RF Model, applied to the Entire Original Dataset of well 1-BRSA-1116-RJS. Plot 1: NMR Effective Porosity (m³/m³). Plot 2: NMR Free Fluid (m³/m³). Plot 3: NMR Total Porosity (m³/m³).

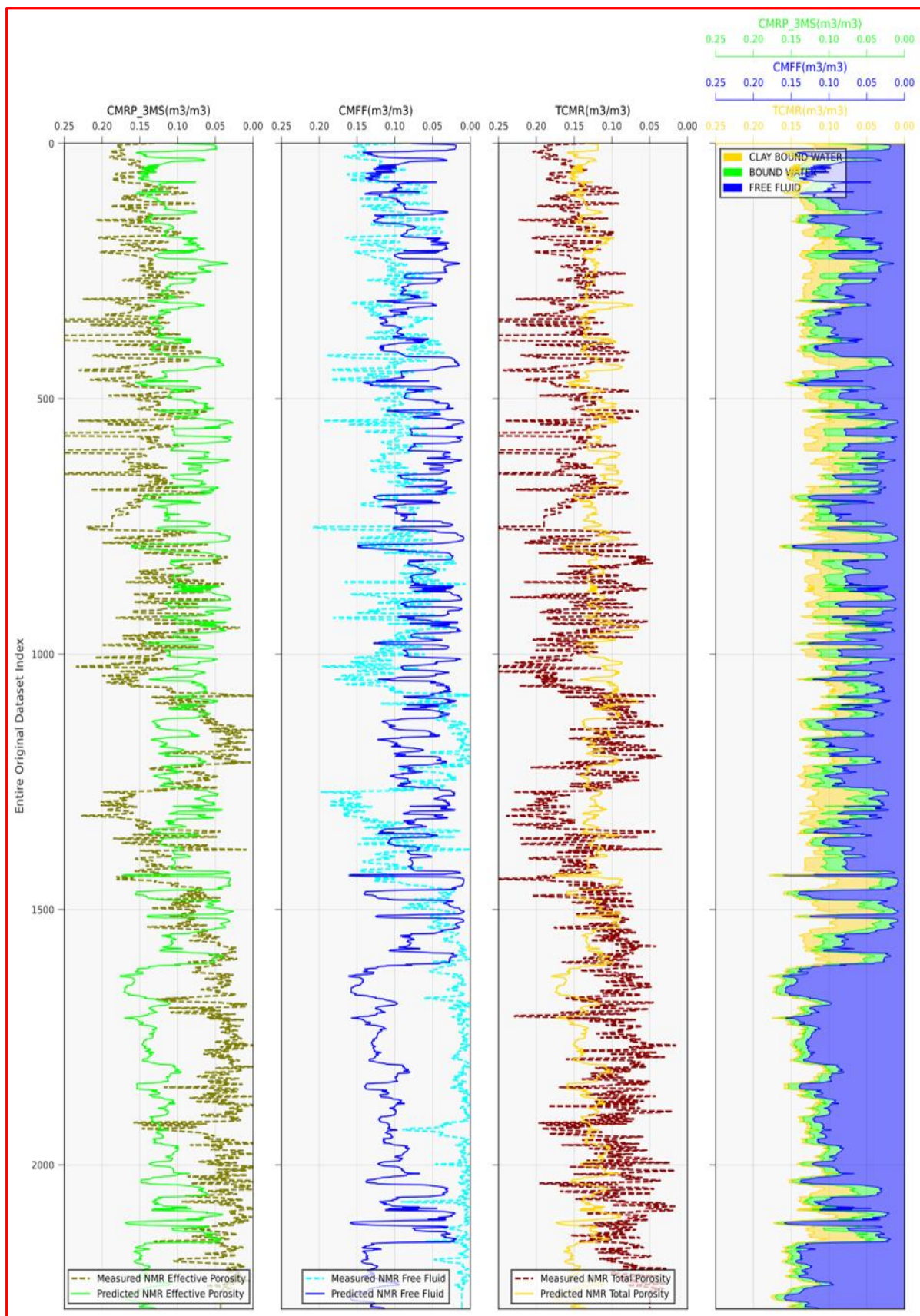


Fig. 3.29: Comparing the match between the predicted and measured NMR porosity, for the RF Model, across the Entire Original Dataset Index of well 1-BRSA-1116-RJS, which corresponds to the Measured Well Depth. Track 1: Predicted and Measured NMR Effective Porosity (m^3/m^3). Track 2: Predicted and Measured NMR Free Fluid (m^3/m^3). Track 3: Predicted and Measured NMR Total Porosity (m^3/m^3).



Fig. 3.30: Scatter plots of predicted versus measured NMR porosity, for the GB Model, applied to the Entire Original Dataset of well 1-BRSA-1116-RJS. Plot 1: NMR Effective Porosity (m³/m³). Plot 2: NMR Free Fluid (m³/m³). Plot 3: NMR Total Porosity (m³/m³).

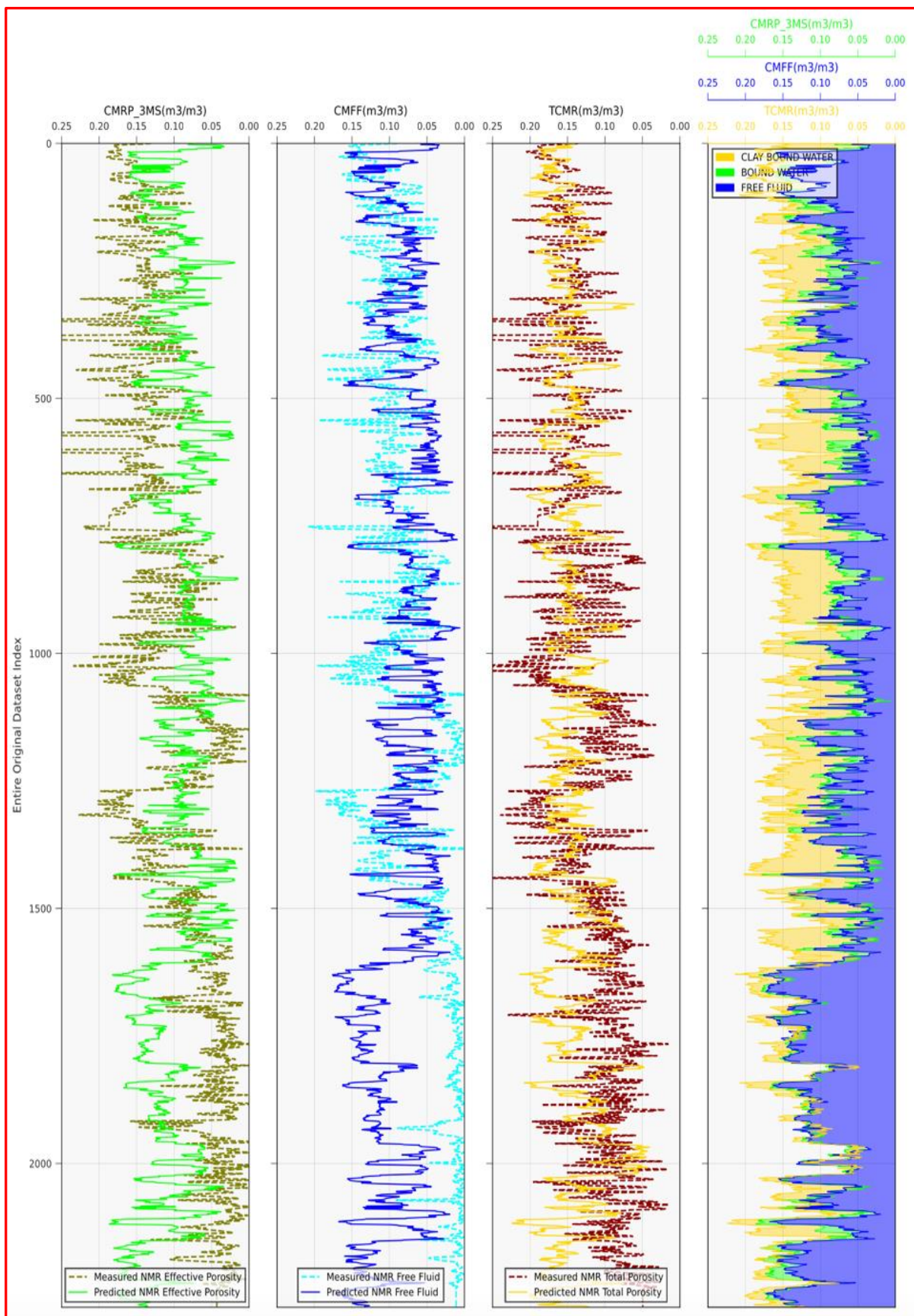


Fig. 3.31: Comparing the match between the predicted and measured NMR porosity, for the GB Model, across the Entire Original Dataset Index of well 1-BRSA-1116-RJS, which corresponds to the Measured Well Depth. Track 1: Predicted and Measured NMR Effective Porosity (m^3/m^3). Track 2: Predicted and Measured NMR Free Fluid (m^3/m^3). Track 3: Predicted and Measured NMR Total Porosity (m^3/m^3).

3.3.2 Second Attempt: Training on Well 1-BRSA-1116-RJS and Validation on Well 3-BRSA-1215-RJS

Also in this case, results shown in Fig. 3.32, Fig. 3.33, Fig. 3.34, Fig. 3.35, Fig. 3.36, Fig. 3.37, Fig. 3.38, Fig. 3.39, Fig. 3.40, Fig. 3.41, Fig. 3.42, Fig. 3.43 support our previous statements and reinforce the observations regarding the predictive capabilities of RF and GB models on the test well. These figures visually confirm that our expectations have been met. A more detailed analysis, including regression metrics, will be presented later to provide a comprehensive and quantitative assessment of the RF and GB models performance.

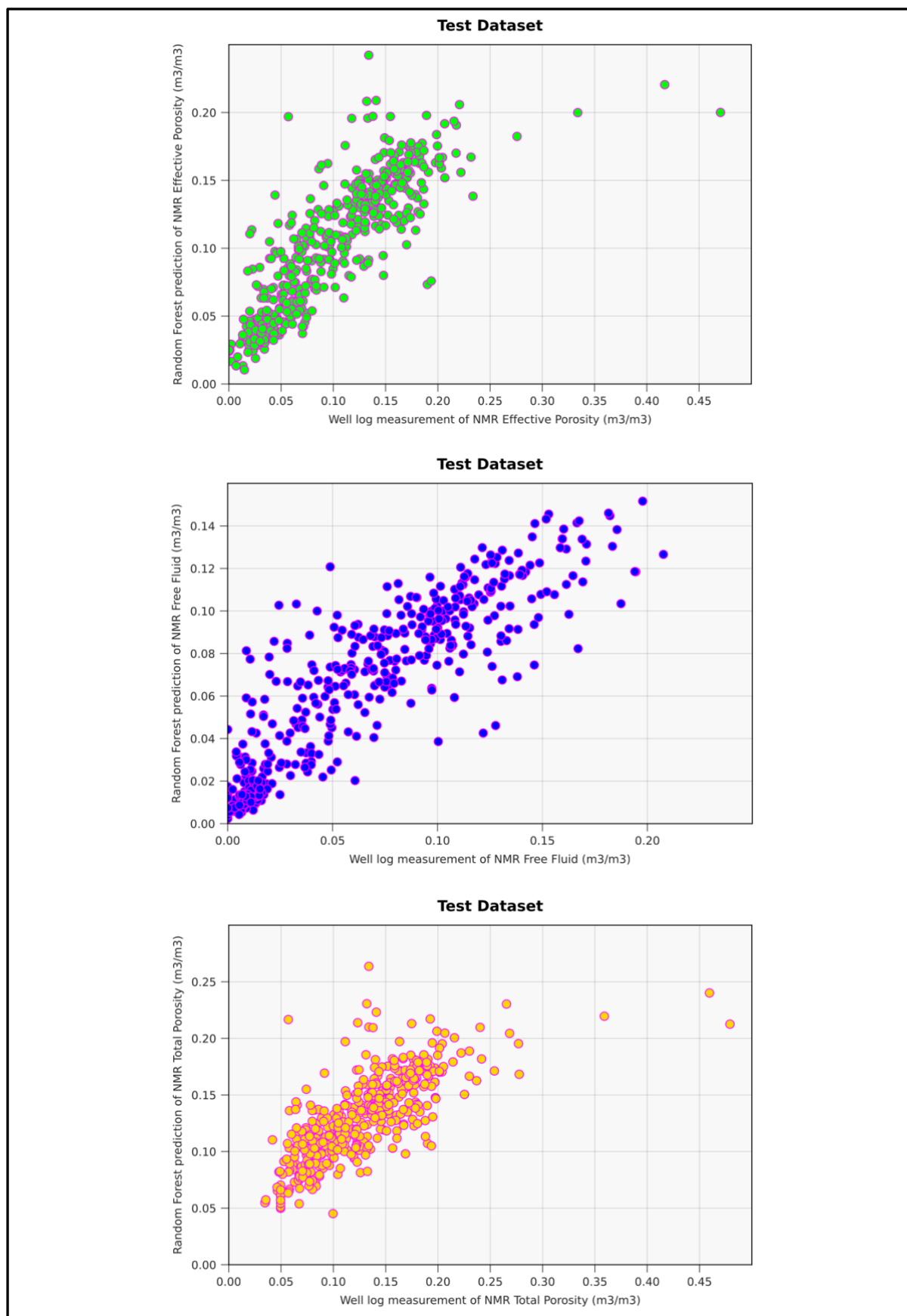


Fig. 3.32: Scatter plots of predicted versus measured NMR porosity, for the RF Model, applied to the Test Dataset of well 1-BRSA-1116-RJS. Plot 1: NMR Effective Porosity (m^3/m^3). Plot 2: NMR Free Fluid (m^3/m^3). Plot 3: NMR Total Porosity (m^3/m^3).

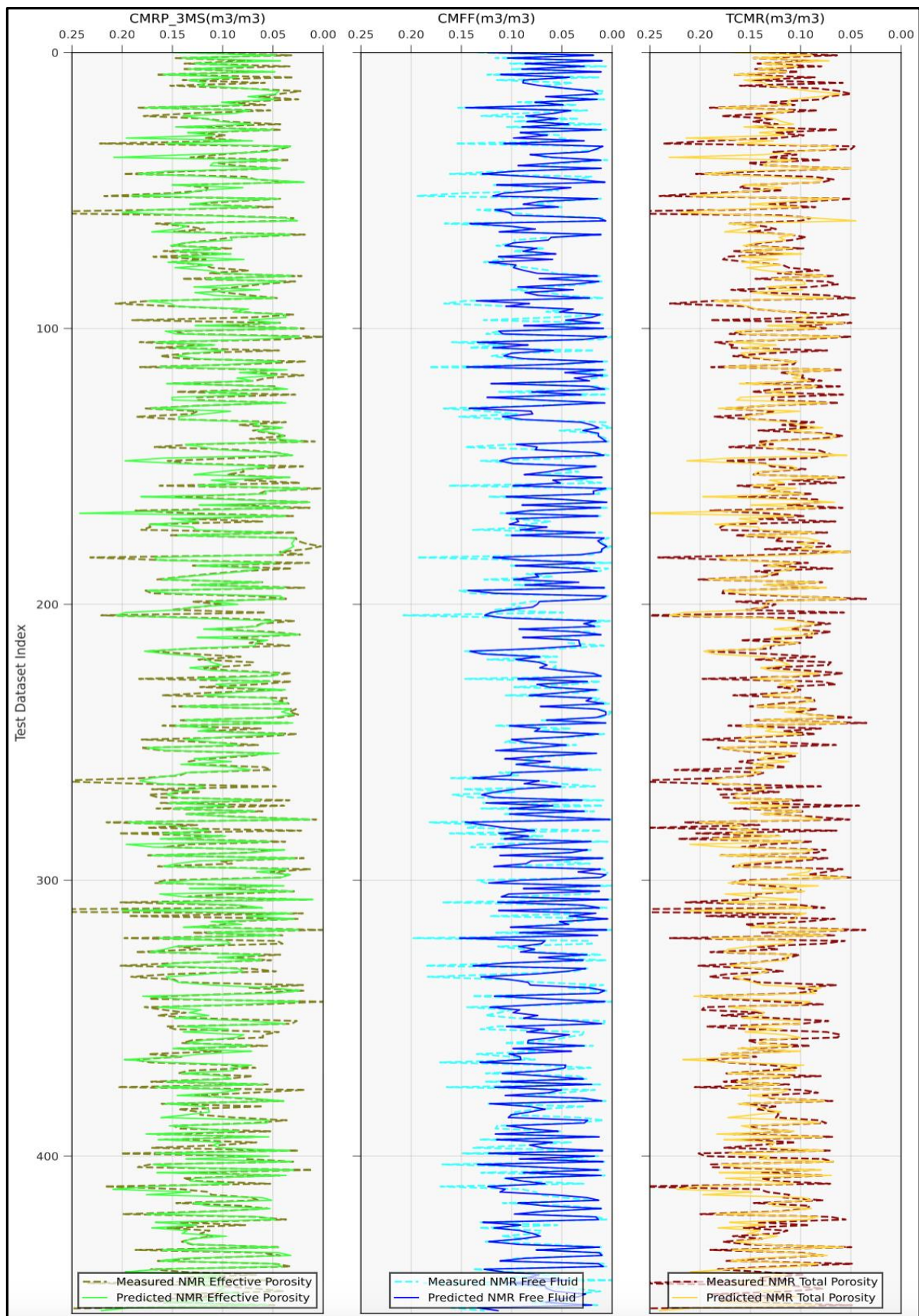


Fig. 3.33: Comparing the match between the predicted and measured NMR porosity, for the RF Model, across the Test Dataset Index of well 1-BRSA-1116-RJS. Track 1: Predicted and Measured NMR Effective Porosity (m^3/m^3). Track 2: Predicted and Measured NMR Free Fluid (m^3/m^3). Track 3: Predicted and Measured NMR Total Porosity (m^3/m^3).

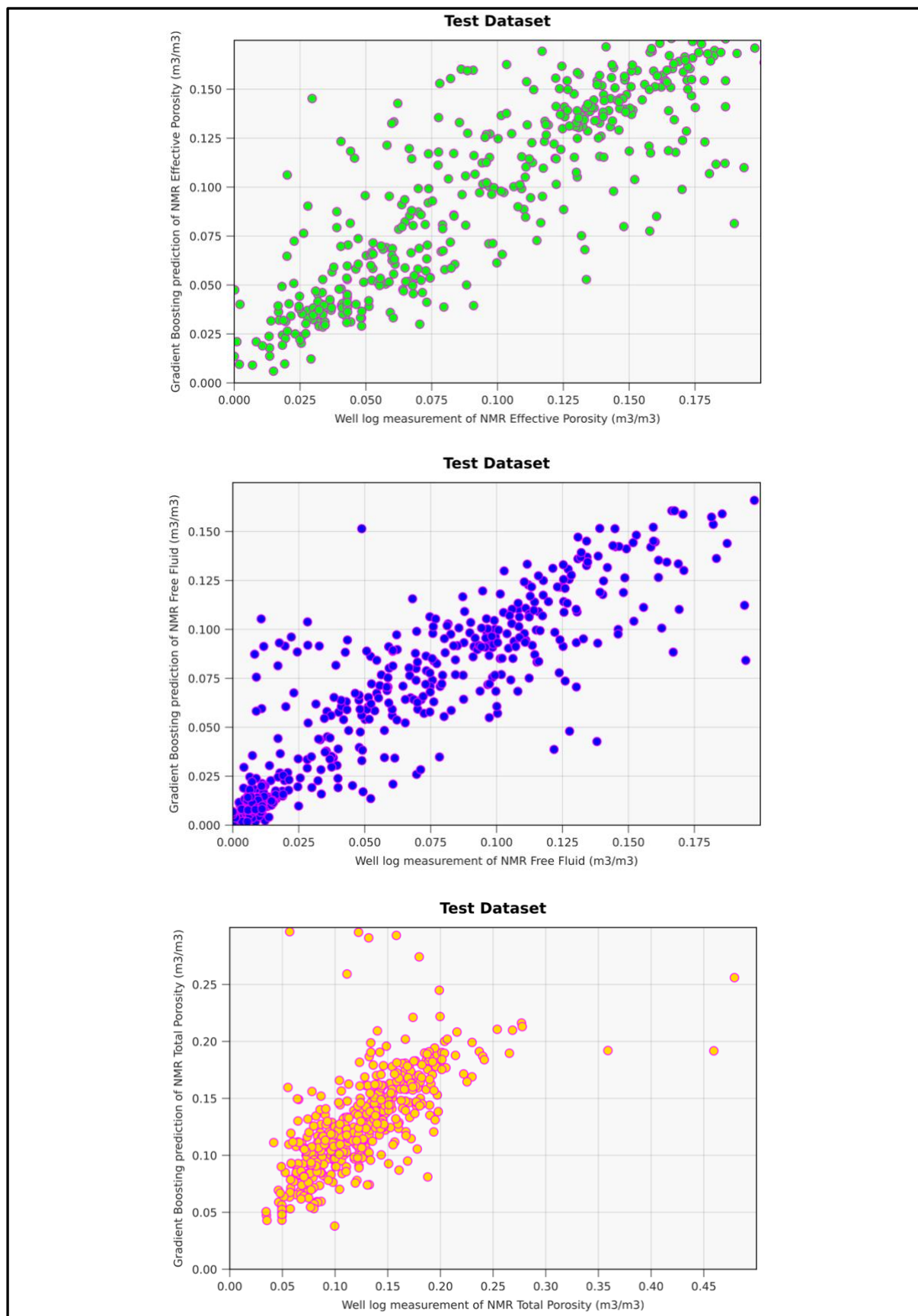


Fig. 3.34: Scatter plots of predicted versus measured NMR porosity, for the GB Model, applied to the Test Dataset of well 1-BRSA-1116-RJS. Plot 1: NMR Effective Porosity (m³/m³). Plot 2: NMR Free Fluid (m³/m³). Plot 3: NMR Total Porosity (m³/m³).

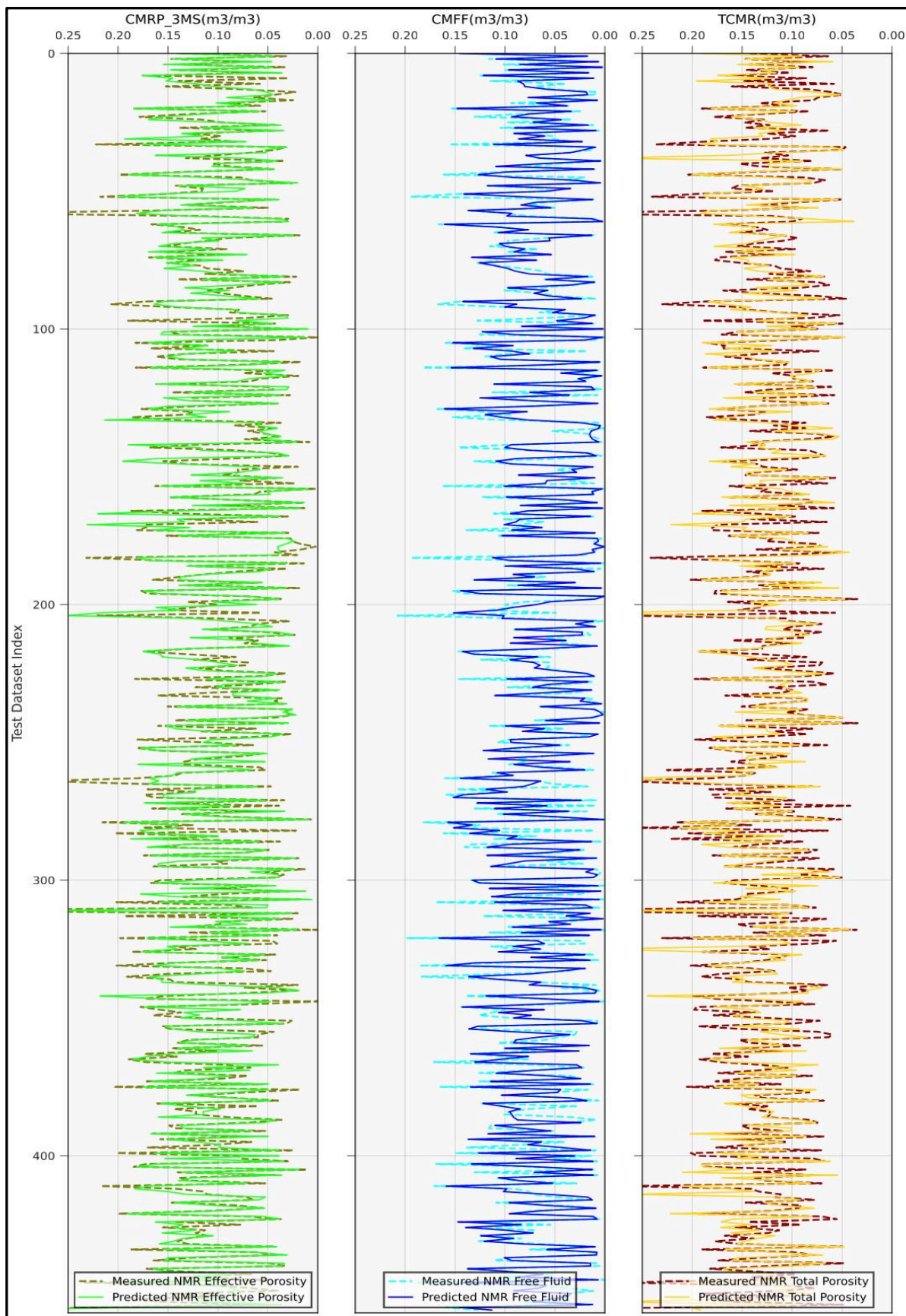


Fig. 3.35: Comparing the match between the predicted and measured NMR porosity, for the RF Model, across the Test Dataset Index of well 1-BRSA-1116-RJS. Track 1: Predicted and Measured NMR Effective Porosity (m³/m³). Track 2: Predicted and Measured NMR Free Fluid (m³/m³). Track 3: Predicted and Measured NMR Total Porosity (m³/m³).

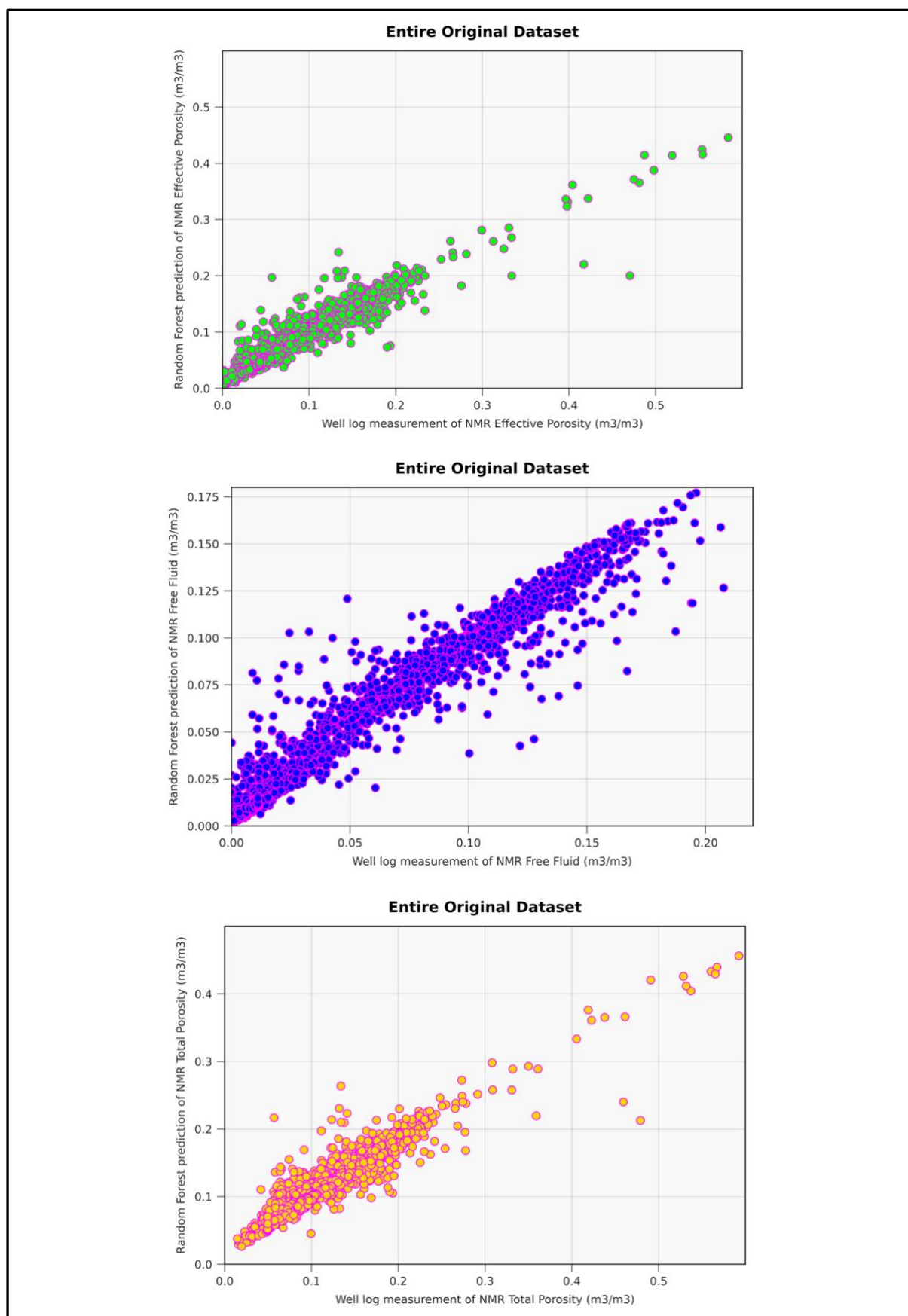


Fig. 3.36: Scatter plots of predicted versus measured NMR porosity, for the RF Model, applied to the Entire Original Dataset of well 1-BRSA-1116-RJS. Plot 1: NMR Effective Porosity (m³/m³). Plot 2: NMR Free Fluid (m³/m³). Plot 3: NMR Total Porosity (m³/m³).

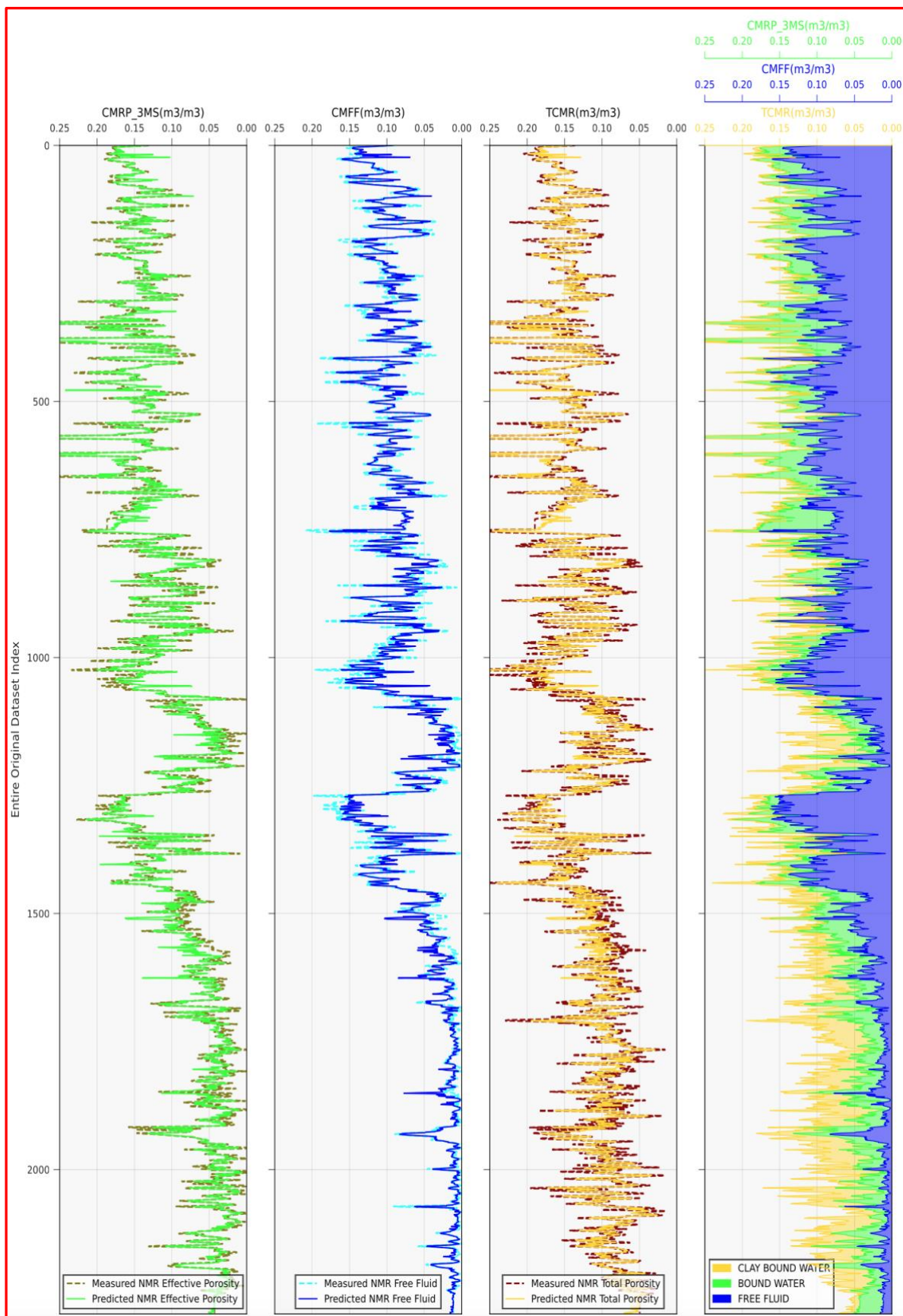


Fig. 3.37: Comparing the match between the predicted and measured NMR porosity, for the RF Model, across the Entire Original Dataset Index of well 1-BRSA-1116-RJS, which corresponds to the Measured Well Depth. Track 1: Predicted and Measured NMR Effective Porosity (m^3/m^3). Track 2: Predicted and Measured NMR Free Fluid (m^3/m^3). Track 3: Predicted and Measured NMR Total Porosity (m^3/m^3).

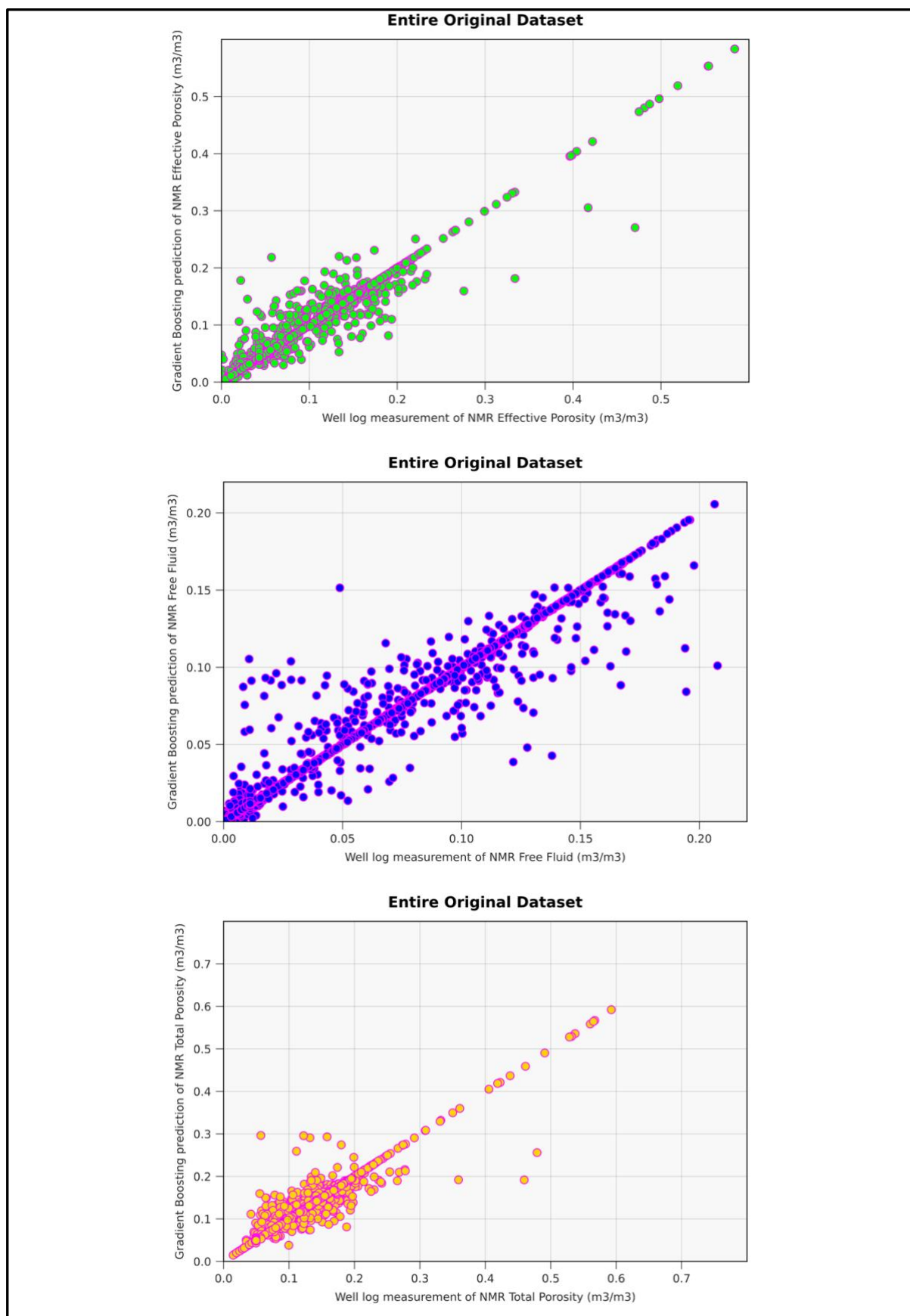


Fig. 3.38: Scatter plots of predicted versus measured NMR porosity, for the GB Model, applied to the Entire Original Dataset of well 1-BRSA-1116-RJS. Plot 1: NMR Effective Porosity (m³/m³). Plot 2: NMR Free Fluid (m³/m³). Plot 3: NMR Total Porosity (m³/m³).

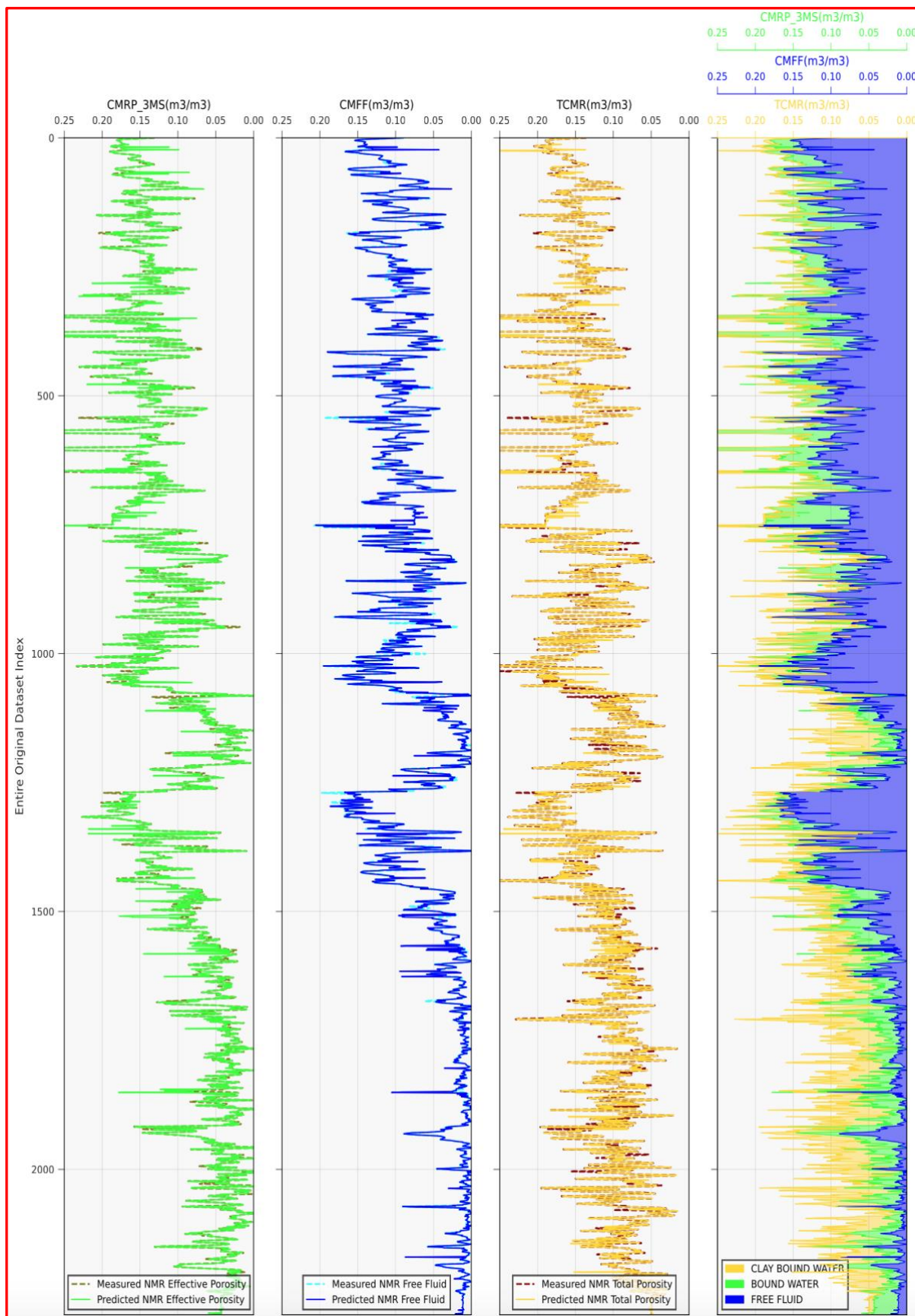


Fig. 3.39: Comparing the match between the predicted and measured NMR porosity, for the GB Model, across the Entire Original Dataset Index of well 1-BRSA-1116-RJS, which corresponds to the Measured Well Depth. Track 1: Predicted and Measured NMR Effective Porosity (m^3/m^3). Track 2: Predicted and Measured NMR Free Fluid (m^3/m^3). Track 3: Predicted and Measured NMR Total Porosity (m^3/m^3).

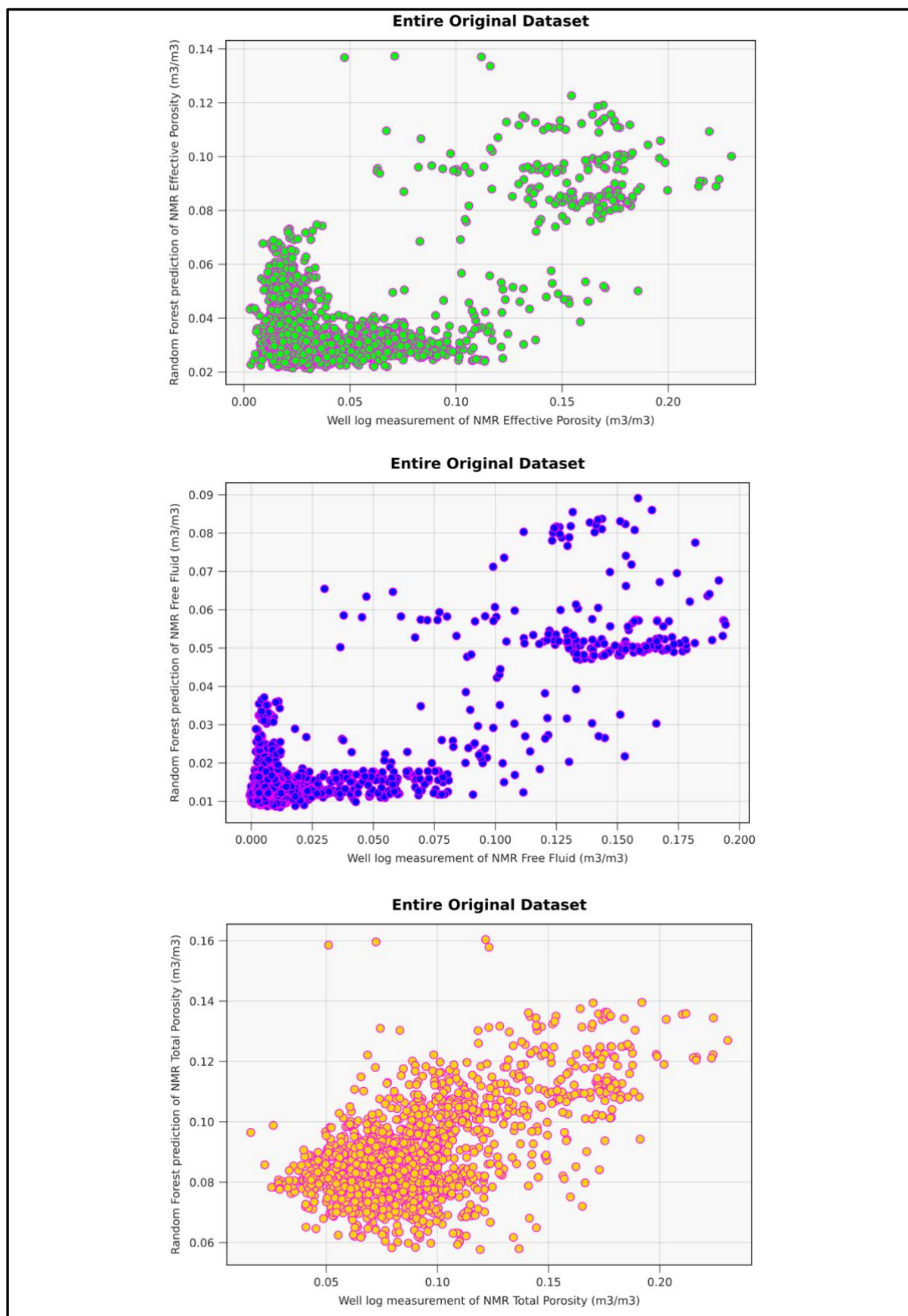


Fig. 3.40: Scatter plots of predicted versus measured NMR porosity, for the RF Model, applied to the Entire Original Dataset of well 3-BRSA-1215-RJS. Plot 1: NMR Effective Porosity (m³/m³). Plot 2: NMR Free Fluid (m³/m³). Plot 3: NMR Total Porosity (m³/m³).

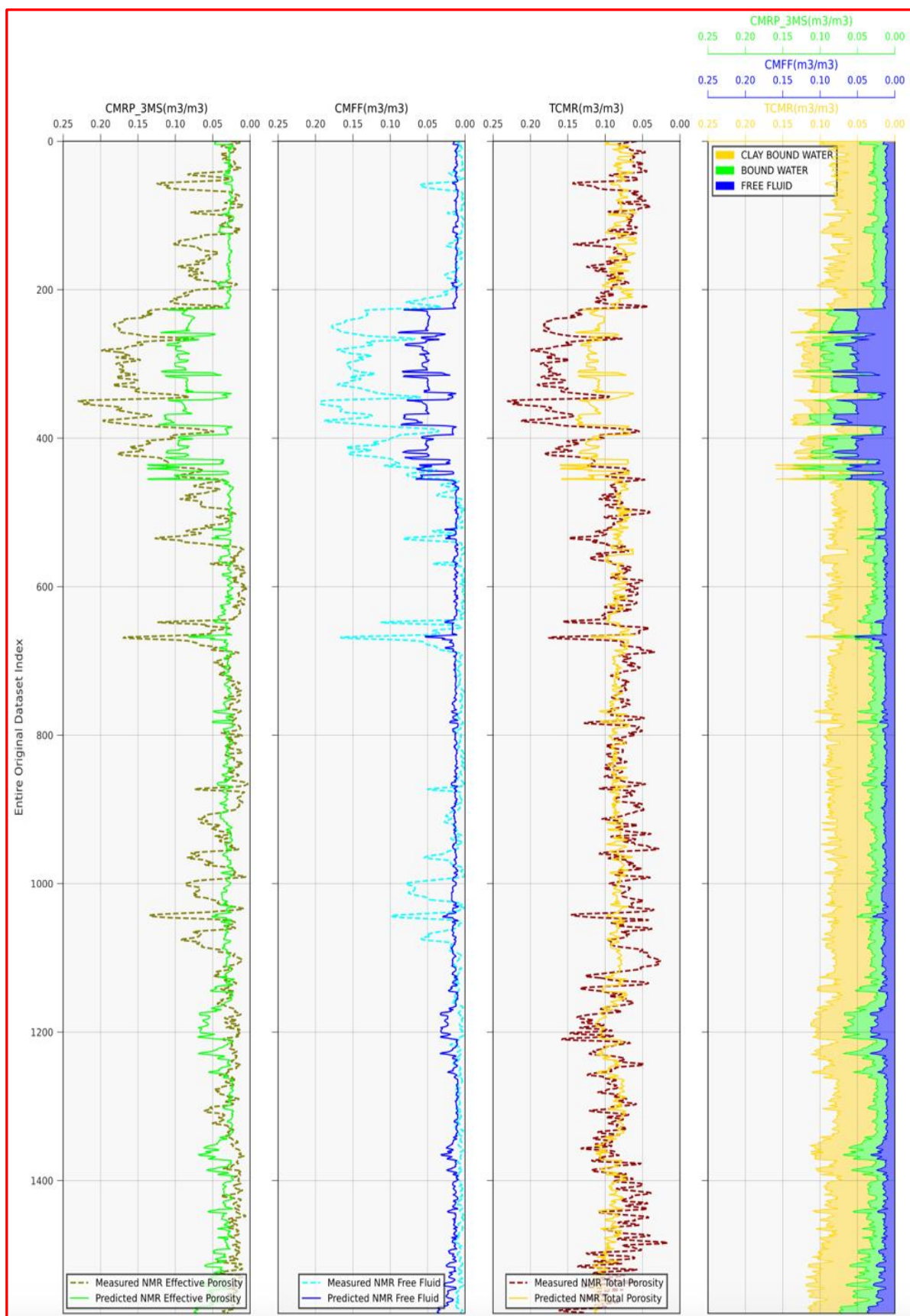


Fig. 3.41: Comparing the match between the predicted and measured NMR porosity, for the RF Model, across the Entire Original Dataset Index of well 3-BRSA-1215-RJS, which corresponds to the Measured Well Depth. Track 1: Predicted and Measured NMR Effective Porosity (m^3/m^3). Track 2: Predicted and Measured NMR Free Fluid (m^3/m^3). Track 3: Predicted and Measured NMR Total Porosity (m^3/m^3).

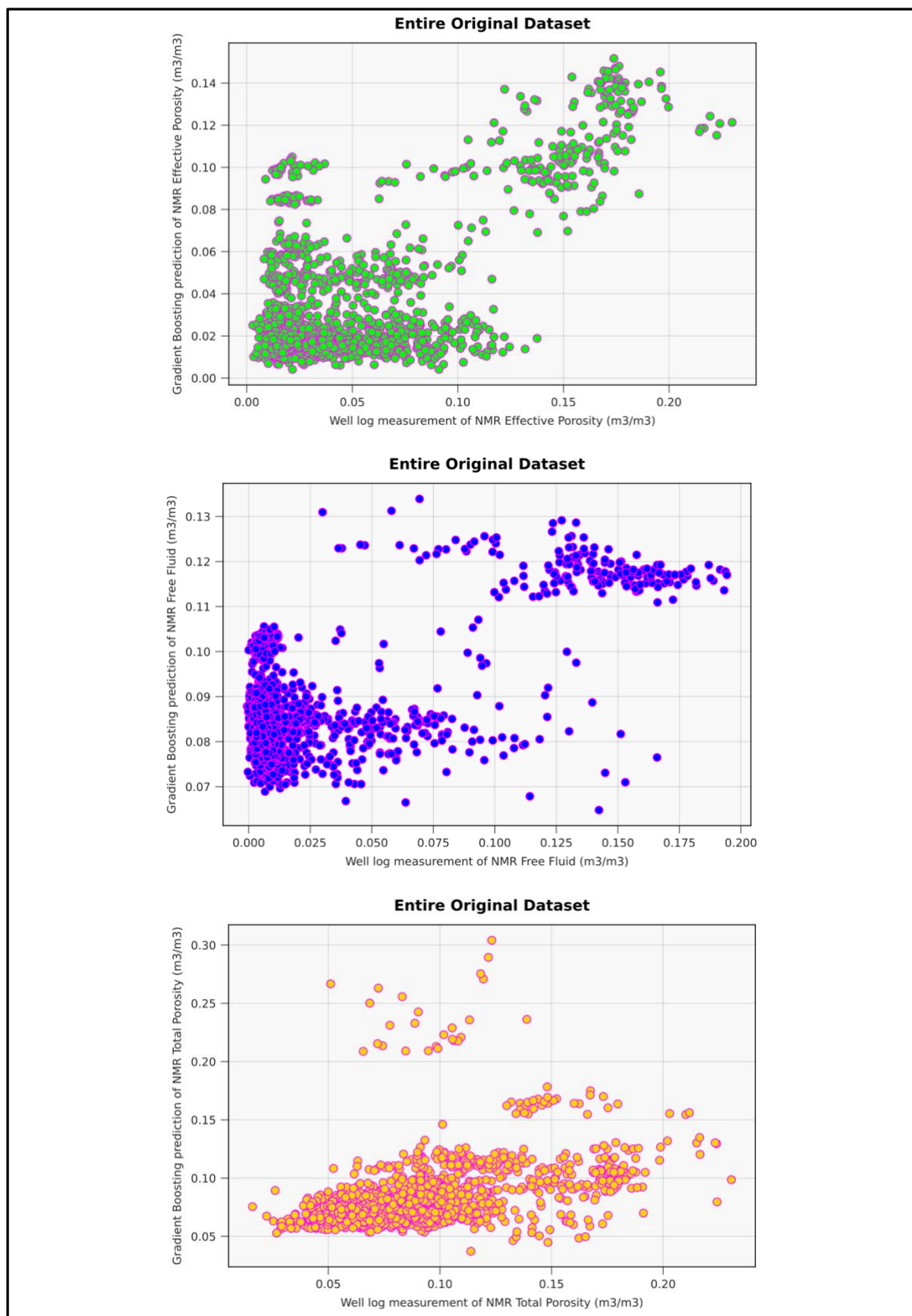


Fig. 3.42: Scatter plots of predicted versus measured NMR porosity, for the GB Model, applied to the Entire Original Dataset of well 3-BRSA-1215-RJS. Plot 1: NMR Effective Porosity (m³/m³). Plot 2: NMR Free Fluid (m³/m³). Plot 3: NMR Total Porosity (m³/m³).

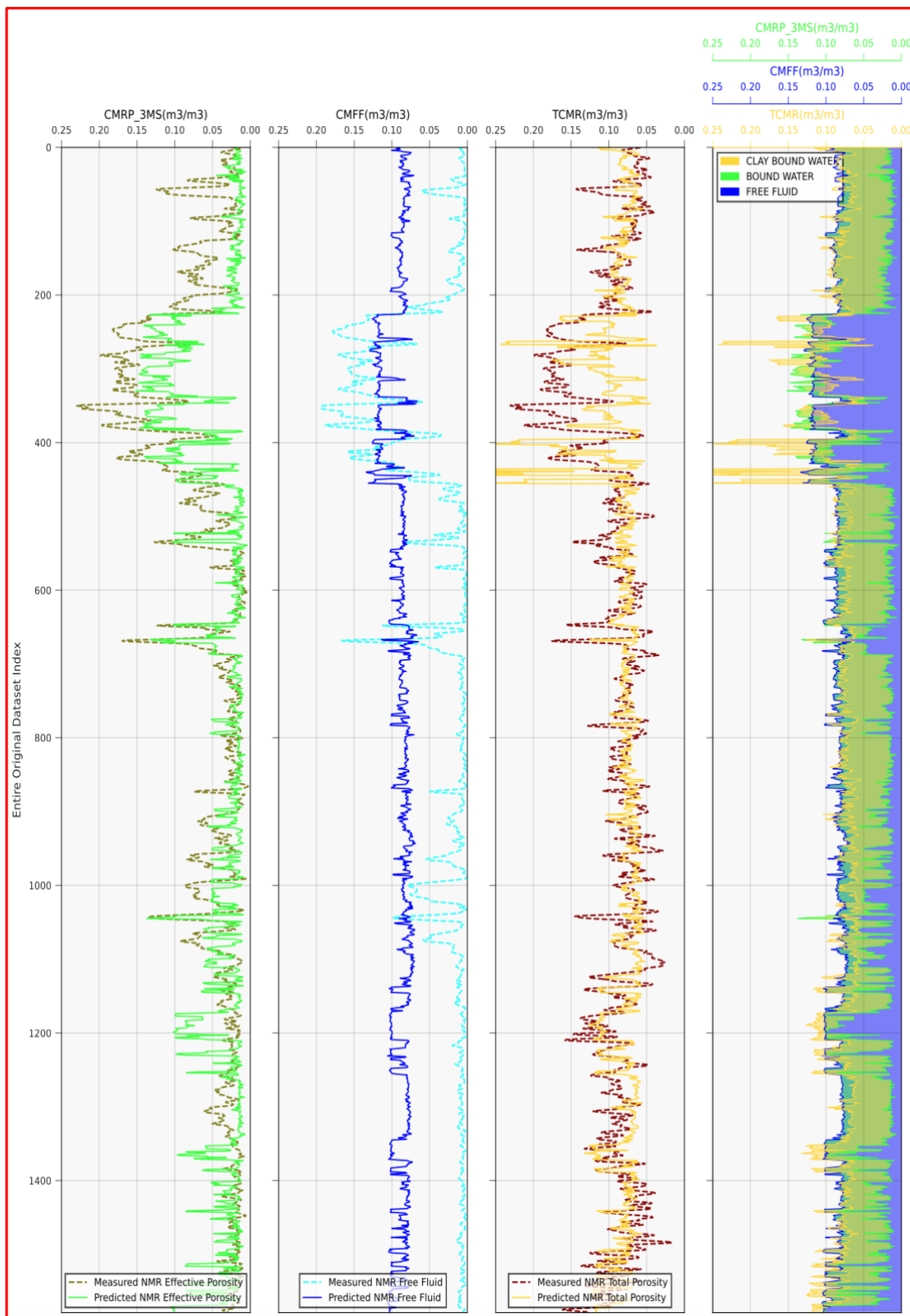


Fig. 3.43: Comparing the match between the predicted and measured NMR porosity, for the GB Model, across the Entire Original Dataset Index of well 3-BRSA-1215-RJS, which corresponds to the Measured Well Depth. Track 1: Predicted and Measured NMR Effective Porosity (m³/m³). Track 2: Predicted and Measured NMR Free Fluid (m³/m³). Track 3: Predicted and Measured NMR Total Porosity (m³/m³).

4 Discussion of the Results

To quantitatively evaluate the performance of the Tree-Based Ensemble Methods, we use the following Regression Metrics: Coefficient of Determination (R^2), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and Mean Absolute Error (MAE) (Farmanov *et al.*, 2023).

The Coefficient of Determination (R^2) represents the predictive power of a model as a value between -inf and 1.00. The closer to 1.00, the better the prediction. A negative R^2 value indicates that the fit does not follow the trend data and can occur with non-linear regression models. It is expressed in percentage. As follows, its formulation (Eq. 4.1):

$$R^2 = 1 - \frac{\sum_i^N (y_{predicted} - y_{measured})^2}{\sum_i^N (y_{measured} - y_{mean})^2} \quad \text{Eq. 4.1}$$

where $y_{predicted}$ and $y_{measured}$ represent the predicted and measured values, respectively; y_{mean} indicates the mean value of the measured values, and N is the number of samples.

The Mean Squared Error (MSE) indicates the average squared difference between the predicted and measured values. It takes numbers in the range of 0.00 and +inf. The closer to 0.00, the better. It makes use of the same scale of the measured data but to the power of two. The equation for MSE is shown below (Eq. 4.2):

$$MSE = \frac{\sum_i^N (y_{predicted} - y_{measured})^2}{N} \quad \text{Eq. 4.2}$$

The Root Mean Squared Error (RMSE) is the squared root of the Mean Squared Error, and it ranges between 0.00 and +inf. The closer to 0.00, the better. It uses the same scale as the measured data. As follows, its equation (Eq. 4.3):

$$RMSE = \sqrt{\frac{\sum_i^N (y_{predicted} - y_{measured})^2}{N}} \quad \text{Eq. 4.3}$$

The Mean Absolute Error (MAE) determines the average absolute difference between the predicted and the measured values. It can assume a value between 0.00 and +inf. The closer to 0.00, the better. It has the same scale as the measured data. The structure of MAE is illustrated by (Eq. 4.4), as follows:

$$MAE = \frac{\sum_i^N |y_{predicted} - y_{measured}|}{N} \quad \text{Eq. 4.4}$$

4.1 Case Study 1: Effective Porosity Log

The summarized results for R^2 and MSE, RMSE, and MAE scores using the best hyperparameter configuration, for RF and GB models, are presented in Table 4.1, and Table 4.2. The scores specifically refer to the predictions made on the test dataset of well 1-BRSA-871-MG (training well) and the total dataset of well 1-BRSA-948-MG (test well). These datasets were chosen as they represent unseen data for the models and allow for a reliable evaluation of their performance.

| ML Algorithm | R^2 (test) (%) | RMSE (m^3/m^3) | MSE (m^3/m^3) ² | MAE (m^3/m^3) |
|-------------------|------------------|--------------------|--------------------------------|-------------------|
| Random Forest | 97.85 | 0.0053 | 0.0 | 0.0024 |
| Gradient Boosting | 98.15 | 0.0049 | 0.0 | 0.0025 |

Table 4.1: Regression Metrics of RF and GB Models for estimation of Effective Porosity on the Test Dataset of well 1-BRSA-871-MG.

| ML Algorithm | R^2 (%) | RMSE (m^3/m^3) | MSE (m^3/m^3) ² | MAE (m^3/m^3) |
|-------------------|-----------|--------------------|--------------------------------|-------------------|
| Random Forest | 84.78 | 0.0036 | 0.0 | 0.0029 |
| Gradient Boosting | 62.32 | 0.0057 | 0.0 | 0.0044 |

Table 4.2: Regression Metrics of RF and GB Models for estimation of Effective Porosity on the Entire Original Dataset of well 1-BRSA-948-MG.

Overall, both models demonstrate reasonably good performance in predicting the calculated effective porosity. The GB model tends to outperform the RF model in terms of accuracy (R^2) and regression errors (MSE, RMSE, and MAE) on the training well. On the other hand, the RF model performs better on the test well. These results suggest that the GB model is more effective in capturing the patterns and variations in the training well data, while the RF model generalizes better to unseen data and performs well on the test well. We can conclude that the prediction is good in both wells since the two wells have similar lithology, the size of the training dataset is big, and the range of prediction is similar.

4.2 Case Study 1: Compressional Wave Slowness Log

Also in this case we present a summary of the R^2 and MSE, RMSE, and MAE scores using the best hyperparameter configuration, for RF and GB models, in Table 4.3, Table 4.4:

| ML Algorithm | R^2 (test) (%) | RMSE ($\mu s/ft$) | MSE ($\mu s/ft$) ² | MAE ($\mu s/ft$) |
|-------------------|------------------|---------------------|---------------------------------|--------------------|
| Random Forest | 90.94 | 1.4186 | 2.0124 | 0.7955 |
| Gradient Boosting | 90.78 | 1.4311 | 2.0482 | 0.8528 |

Table 4.3: Regression Metrics of RF and GB Models for estimation of Compressional Wave Slowness on the Test Dataset of well 1-BRSA-871-MG.

| ML Algorithm | R^2 (%) | RMSE ($\mu s/ft$) | MSE ($\mu s/ft$) ² | MAE ($\mu s/ft$) |
|-------------------|-----------|---------------------|---------------------------------|--------------------|
| Random Forest | -8670.76 | 6.2050 | 38.5024 | 6.1696 |
| Gradient Boosting | -8737.31 | 6.2285 | 38.7945 | 6.1530 |

Table 4.4: Regression Metrics of RF and GB Models for estimation of Compressional Wave Slowness on the Entire Original Dataset of well 1-BRSA-948-MG.

Both models perform well on the training well, but they perform poorly on the test well. The negative R^2 scores and high regression errors (MSE, RMSE, and MAE) indicate that the models have not been able to capture the patterns and relationships in the test well data effectively. As explained before, the reason is associated with the small training dataset and the different DTCO log data distribution between the two wells.

4.3 Case Study 2: NMR Porosity Logs

In this section, we present the results of the R^2 and MSE, RMSE, and MAE scores using the best hyperparameter configuration for both attempts (Table 4.5, Table 4.6, Table 4.7, Table 4.8). These scores provide a measure of the predictive performance of the models on the test dataset of the training well and on the entire dataset of the test well. Additionally, to enhance the visualization of the error scores, we use bar plots to display the results, only for the first attempt (Fig. 4.1). Furthermore, we include bar plots that showcase the importance of each input feature for predicting the output parameters (Fig. 4.2, Fig. 4.3, Fig. 4.4, Fig. 4.5). These plots help identify the most influential features in the model's predictions.

4.3.1 Regression Metrics for the First Attempt

| Variable | ML Algorithm | R^2 (test) (%) | RMSE (m^3/m^3) | MSE (m^3/m^3) ² | MAE (m^3/m^3) |
|--------------------|-------------------|------------------|--------------------|--------------------------------|-------------------|
| Total Porosity | Random Forest | 88.56 | 0.0131 | 0.0002 | 0.0099 |
| | Gradient Boosting | 85.27 | 0.0148 | 0.0002 | 0.0112 |
| Effective Porosity | Random Forest | 94.63 | 0.0123 | 0.0002 | 0.0083 |
| | Gradient Boosting | 94.54 | 0.0124 | 0.0002 | 0.0089 |
| Free Fluid | Random Forest | 97.24 | 0.0087 | 0.0001 | 0.0051 |
| | Gradient Boosting | 97.11 | 0.0089 | 0.0001 | 0.0055 |

Table 4.5: Regression Metrics of RF and GB Models for estimation of (NMR) Total Porosity, Effective Porosity, and Free Fluid on the Test Dataset of well 3-BRSA-1215-RJS.

| Variable | ML Algorithm | R^2 (%) | RMSE (m^3/m^3) | MSE (m^3/m^3) ² | MAE (m^3/m^3) |
|--------------------|-------------------|-----------|--------------------|--------------------------------|-------------------|
| Total Porosity | Random Forest | -21.25 | 0.0623 | 0.0039 | 0.0458 |
| | Gradient Boosting | -40.32 | 0.0670 | 0.0045 | 0.0503 |
| Effective Porosity | Random Forest | -53.32 | 0.0802 | 0.0064 | 0.0653 |
| | Gradient Boosting | -57.10 | 0.0812 | 0.0066 | 0.0652 |
| Free Fluid | Random Forest | -122.94 | 0.0752 | 0.0057 | 0.0632 |
| | Gradient Boosting | -118.24 | 0.0744 | 0.0055 | 0.0617 |

Table 4.6: Regression Metrics of RF and GB Models for estimation of (NMR) Total Porosity, Effective Porosity, and Free Fluid on the Entire Original Dataset of well 1-BRSA-1116-RJS.

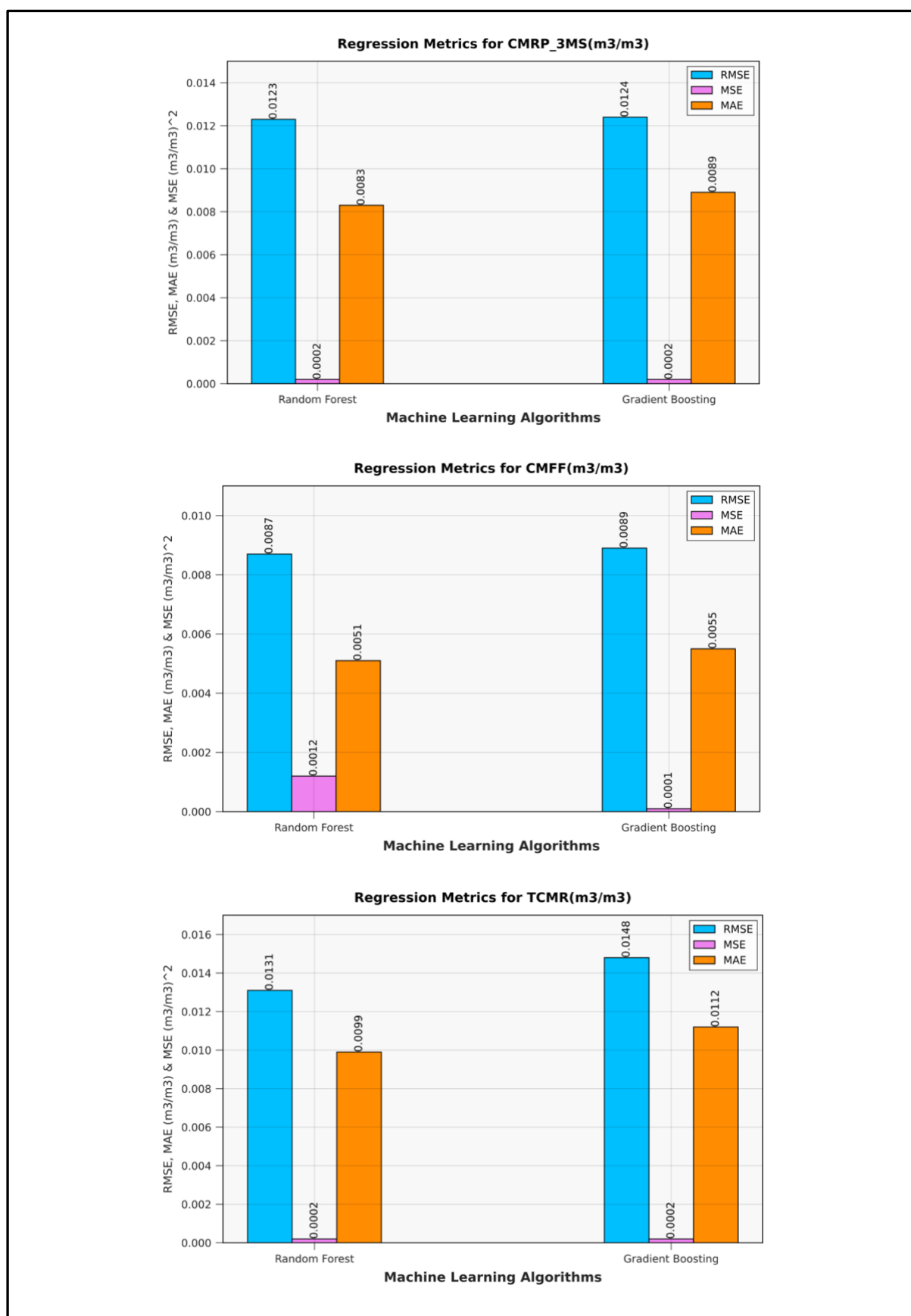


Fig. 4.1: Regression Metrics for RF and GB Models. Root Mean Squared Error (RMSE), Mean Squared Error (MSE), and Mean Absolute Error (MAE). Bar Plot 1: NMR Effective Porosity (m^3/m^3). Bar Plot 2: NMR Free Fluid (m^3/m^3). Bar Plot 3: NMR Total Porosity (m^3/m^3).

The results indicate that the RF model performs slightly better than the GB model in predicting the three output parameters, as it exhibits higher R^2 values and lower regression errors on the training well. Overall, both models can predict the output parameters reasonably well on the training well 3-BRSA-1215-RJS, as evidenced by the close resemblance between the synthetic well-logs generated by the models and the real logged data.

However, when it comes to the validation on the test well, both models show inconsistent accuracy, with negative R^2 scores. This indicates that the models struggle to accurately predict the output parameters on the test well. There are several reasons for this lack of accuracy in the validation phase.

Firstly, the training well 3-BRSA-1215-RJS predominantly consists of non-reservoir rock (80%), making it challenging for the models to generalize to the test well 1-BRSA-1116-RJS, which is almost entirely composed of reservoir rock (90%), even if they are made of the same lithology.

Secondly, the training dataset itself is limited, comprising less than 1579 data points. The small size of the training dataset restricts the model's ability to capture the underlying relationships between the input features and the output parameters, accurately.

Additionally, the porosity range used for training differs significantly from the target porosity range for prediction. This discrepancy is evident in Fig. 3.19, which depicts the porosity distributions of the training and test wells. The difference in porosity range further contributes to the model's inability to effectively predict the porosity values on the test well.

As follows the importance of each feature for predicting the output parameters (Fig. 4.2, and Fig. 4.3):

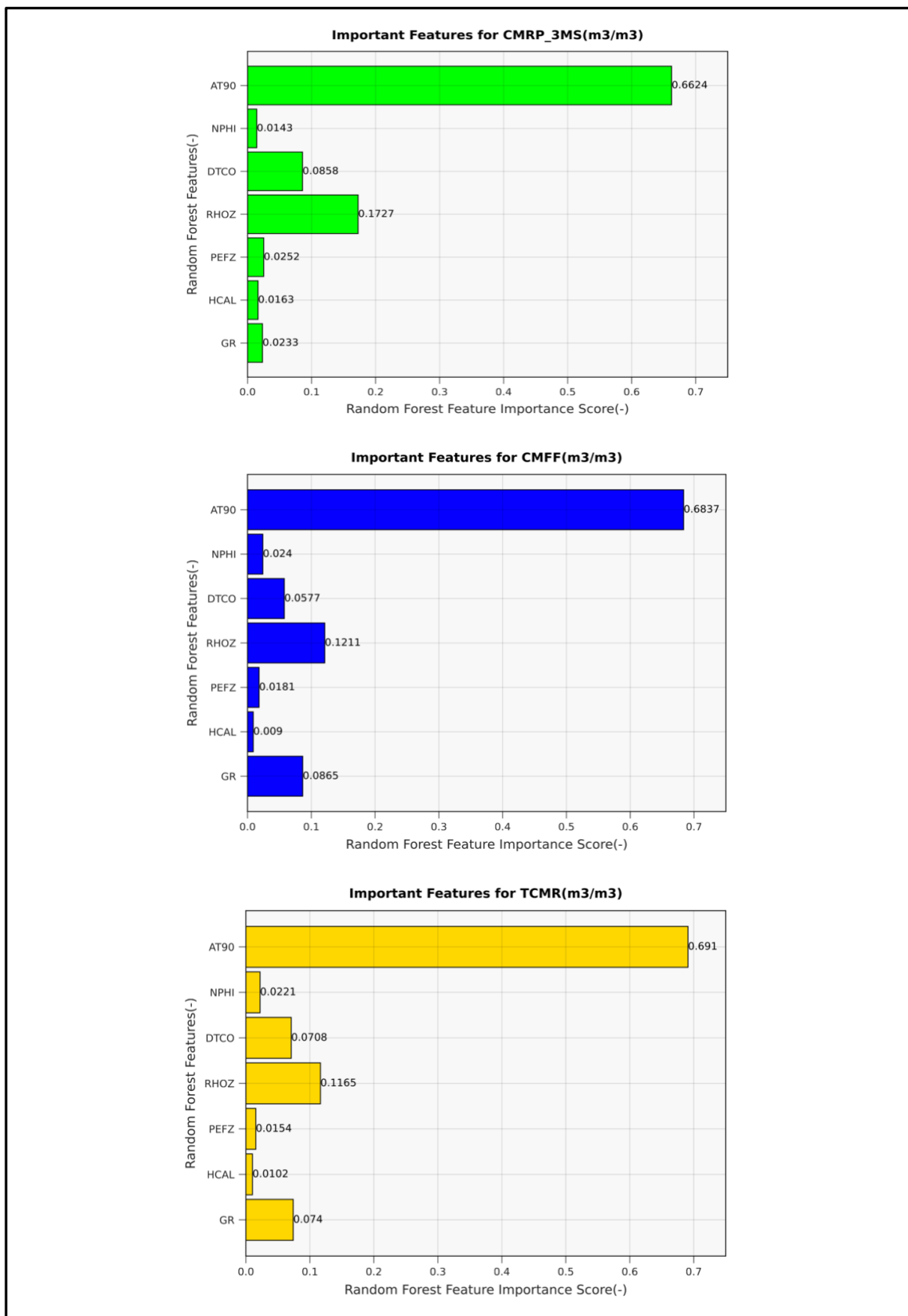


Fig. 4.2: Feature Importance for RF Model (first attempt). Bar plot 1: feature importance scores for NMR Effective Porosity (m³/m³). Bar plot 2: feature importance scores for NMR Free Fluid (m³/m³). Bar plot 3: feature importance scores for NMR Total Porosity (m³/m³).

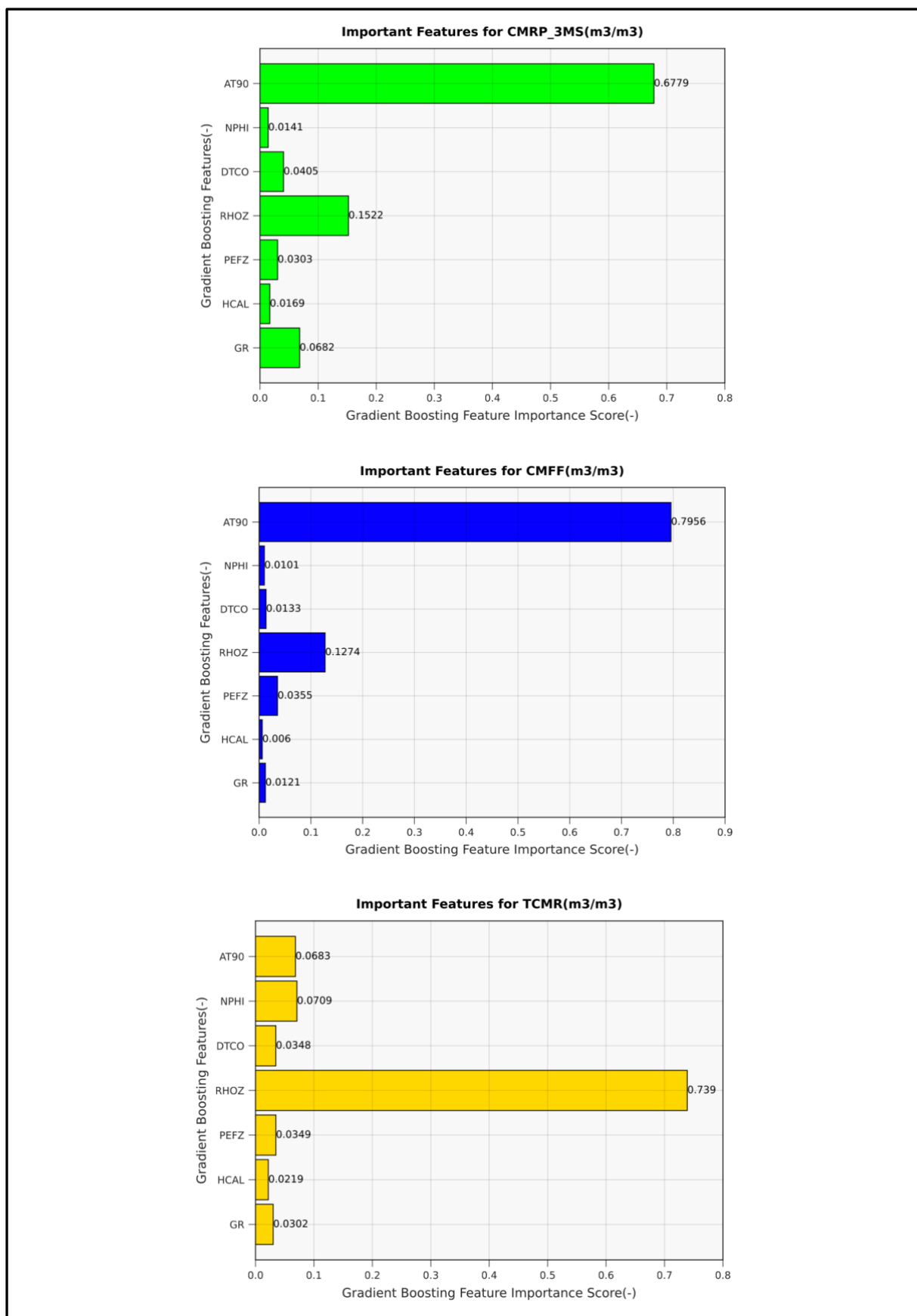


Fig. 4.3: Feature Importance for GB Model (first attempt). Bar plot 1: feature importance scores for NMR Effective Porosity (m^3/m^3). Bar plot 2: feature importance scores for NMR Free Fluid (m^3/m^3). Bar plot 3: feature importance scores for NMR Total Porosity (m^3/m^3).

The conclusions drawn from Fig. 4.2, and Fig. 4.3 suggest that the resistivity log plays a more significant role in the learning process compared to other logs. This finding aligns with expectations, as there is a strong correspondence between resistivity and NMR logs, as shown in Fig. 2.5. The resistivity log can effectively capture the fluid change that we have in well 3-BRSA-1215-RJS, while logs such as NPHI, and DTCO logs may be less affected by these changes, leading to reduced importance in the prediction process.

Furthermore, the results indicate that in the GB model, the formation density log is the primary parameter for predicting the total porosity. This suggests that the GB model can successfully identify the strong relationship between total porosity and formation density logs, which seems to not be significantly impacted by changes in fluid properties.

4.3.2 Regression Metrics for the Second Attempt

| Variable | ML Algorithm | R ² (test) (%) | RMSE (m ³ /m ³) | MSE (m ³ /m ³) ² | MAE (m ³ /m ³) |
|--------------------|-------------------|---------------------------|--|--|---------------------------------------|
| Total Porosity | Random Forest | 54.45 | 0.0356 | 0.0013 | 0.0239 |
| | Gradient Boosting | 46.93 | 0.0384 | 0.0015 | 0.0246 |
| Effective Porosity | Random Forest | 69.26 | 0.0340 | 0.0012 | 0.0221 |
| | Gradient Boosting | 69.94 | 0.0337 | 0.0011 | 0.0220 |
| Free Fluid | Random Forest | 77.17 | 0.0244 | 0.0006 | 0.0169 |
| | Gradient Boosting | 76.81 | 0.0246 | 0.0006 | 0.0157 |

Table 4.7: Regression Metrics of RF and GB Models for estimation of (NMR) Total Porosity, Effective Porosity, and Free Fluid on the Test Dataset of well 1-BRSA-1116-RJS.

| Variable | ML Algorithm | R ² (%) | RMSE (m ³ /m ³) | MSE (m ³ /m ³) ² | MAE (m ³ /m ³) |
|--------------------|-------------------|--------------------|--|--|---------------------------------------|
| Total Porosity | Random Forest | 32.86 | 0.0289 | 0.0008 | 0.0222 |
| | Gradient Boosting | -3.24 | 0.0358 | 0.0013 | 0.0242 |
| Effective Porosity | Random Forest | 36.26 | 0.0376 | 0.0014 | 0.0278 |
| | Gradient Boosting | 41.86 | 0.0359 | 0.0013 | 0.0265 |
| Free Fluid | Random Forest | 37.62 | 0.0366 | 0.0013 | 0.0214 |
| | Gradient Boosting | -134.05 | 0.0709 | 0.0050 | 0.0668 |

Table 4.8: Regression Metrics of RF and GB Models for estimation of (NMR) Total Porosity, Effective Porosity, and Free Fluid on the Entire Original Dataset of well 3-BRSA-1215-RJS.

Table 4.7, and Table 4.8 indicate that both models perform well in predicting the training well 1-BRSA-1116-RJS. However, the accuracy values are relatively lower compared to the first attempt. This can be attributed to the absence of a strong relationship between individual input parameters and the output variables in this well. Instead, all the input parameters contribute significantly to the prediction, indicating their collective importance in capturing the complex behavior of the well. The findings in Fig. 4.4, and Fig. 4.5 further support this observation.

Moreover, it is notable that almost all the accuracy values for the NMR porosities, on the test well, are positive. This suggests that the models are beginning to capture the trends in the test well, even though there is still space for improvement. Overall, these results indicate that the models are making progress in predicting the NMR porosity values on the test well, based on the considerations done previously.

As follows the importance of each feature for predicting the output parameters (Fig. 4.4, Fig. 4.5):

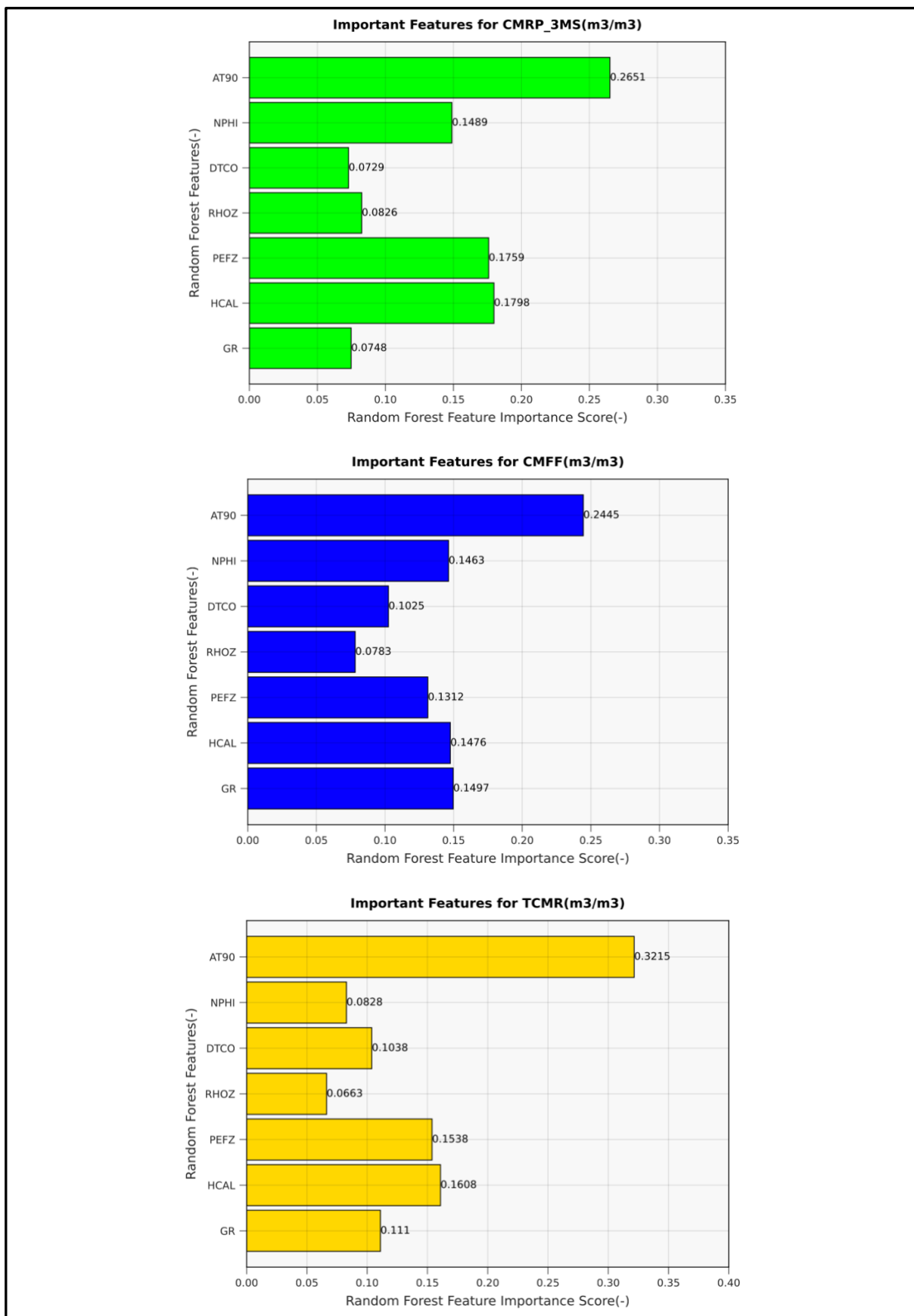


Fig. 4.4: Feature Importance for RF Model (second attempt). Bar plot 1: feature importance scores for NMR Effective Porosity (m^3/m^3). Bar plot 2: feature importance scores for NMR Free Fluid (m^3/m^3). Bar plot 3: feature importance scores for NMR Total Porosity (m^3/m^3).

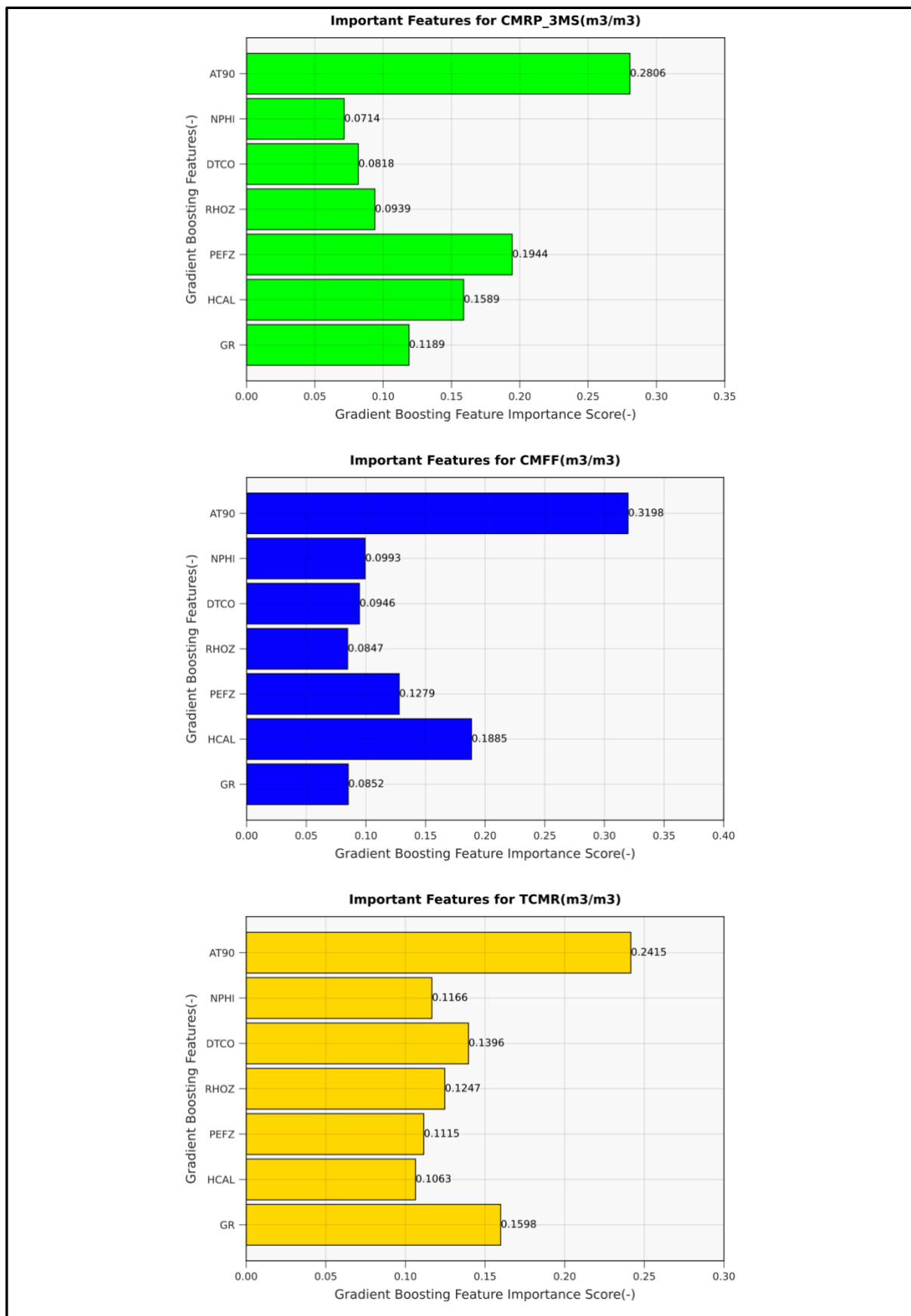


Fig. 4.5: Feature Importance for GB Model (second attempt). Bar plot 1: feature importance scores for NMR Effective Porosity (m^3/m^3). Bar plot 2: feature importance scores for NMR Free Fluid (m^3/m^3). Bar plot 3: feature importance scores for NMR Total Porosity (m^3/m^3).

The conclusions drawn from Fig. 4.4, and Fig. 4.5 support the expectation that the prediction of NMR porosities is influenced by the contribution of different input parameters. However, certain logs have a more pronounced impact on the predictions. Specifically, the DTCO, RHOZ, GR, NPHI, and AT90 logs are identified as having a predominant role in predicting NMR porosities. These logs likely contain valuable information about the formation characteristics and the fluid content that directly relate to the NMR porosity. In this case, the AT90 log is not the dominant predictor for NMR porosity logs because there is not a clear correlation between the trend of the resistivity logs and the NMR porosity logs, as shown in Fig. 2.6.

5 Conclusion

In this research, the main objective was to develop a Python code, from scratch, to implement two supervised learning algorithms, Random Forest (RF) and Gradient Boosting (GB), for the prediction of various types of well-logs using conventional well-logs as input data. We conducted two distinct case studies: one for the São Francisco onshore Brazilian basin, aiming to predict the calculated effective porosity (PHIE_HILT) and the measured compressional wave slowness (DTCO) logs; and another for the Santos offshore Brazilian basin, focusing on the prediction of nuclear magnetic resonance (NMR) porosity logs, which is of primary interest. Our implemented code successfully generated PHIE_HILT, DTCO, and NMR synthetic well-logs for the training wells. However, we obtained poor results for the prediction of DTCO, and NMR logs on the test wells. This issue can be attributed to the size and range of the training dataset initially used to train the Machine Learning (ML) models. For this reason, these results provided valuable insights into the characteristics that an ideal training dataset should possess, for this specific application, with particular attention given to the results obtained from the NMR logs prediction.

Based on our findings, we concluded that a carefully selected training dataset should contain a sufficient number of data points, preferably a minimum of 10,000. The training well should include a mix of reservoir and non-reservoir formations, and ideally, it should encompass all the lithologies that need to be predicted, ensuring a robust representation of the data, and improving the model's generalization ability. By incorporating a wide range of geological variations in the training data, the model becomes more capable of capturing the complexities and patterns present in different formations. Considering the insights gained from this research, the two models can be trained again on an improved training dataset of a larger size, which would lead to more accurate prediction results. This recommendation is further supported by a research conducted by (Tamoto, Gioria and Carneiro, 2023), in the Santos basin, where their supervised ML models, trained on a large dataset, achieved impressive results on the test wells.

Furthermore, the synthetic NMR porosity well-log curves generated can be used to calculate the irreducible water saturation of the formation. This, in turn, can be utilized to estimate the permeability of the formation using an equation that is directly proportional to the total porosity and inversely proportional to the irreducible water saturation, of the formation. These results can be further validated using core analysis data.

In conclusion, this research contributes to the understanding of using supervised ML algorithms for well-logs prediction and highlights the importance of selecting an appropriate training dataset to improve the accuracy of the model's predictions.

6 References

'ANP-TERRESTRE' (2023). Available at: <https://reate.cprm.gov.br/anp/TERRESTREen>.

Babasafari, A.A. *et al.* (2022) 'Ensemble-based machine learning application for lithofacies classification in a pre-salt carbonate reservoir, Santos Basin, Brazil', *Petroleum Science and Technology*, pp. 1–17. Available at: <https://doi.org/10.1080/10916466.2022.2143813>.

'Brazilian States Coordinates' (2023). Available at: <https://www.kaggle.com/datasets/thiagobodruk/brazil-geojson?resource=download>.

'Brazilian Sub-Regions Coordinates' (2023). Available at: <https://portal.grdc.bafg.de/applications/public.html?publicuser=PublicUser#dataDownload/Subregions>.

Crain, E.R. and Ganz, C.I. (1986) *The log analysis handbook*. Tulsa, Okla., USA: PennWell.

Equinor (2017) 'Statoil strengthens position in Brazilian licence BM-S-8'. Available at: <https://www.equinor.com/news/archive/strengthens-position-bm-s-8>.

Farmanov, R. *et al.* (2023) 'Application of Machine Learning for Estimating Petrophysical Properties of Carbonate Rocks Using NMR Core Measurements', in *Day 1 Tue, January 24, 2023. SPE Reservoir Characterisation and Simulation Conference and Exhibition*, Abu Dhabi, UAE: SPE, p. D011S001R003. Available at: <https://doi.org/10.2118/212625-MS>.

Gomes, J.P. *et al.* (2020) 'Facies classification and patterns of lacustrine carbonate deposition of the Barra Velha Formation, Santos Basin, Brazilian Pre-salt', *Marine and Petroleum Geology*, 113, p. 104176. Available at: <https://doi.org/10.1016/j.marpetgeo.2019.104176>.

González Carrasquilla, A.A. and Tapia Briones, V.H. (2019) 'SIMULATING POROSITY AND PERMEABILITY OF THE NUCLEAR MAGNETIC RESONANCE (NMR) LOG IN CARBONATE RESERVOIRS OF CAMPOS BASIN, SOUTHEASTERN BRAZIL, USING CONVENTIONAL LOGS AND ARTIFICIAL INTELLIGENCE APPROACHES', *Brazilian Journal of Geophysics*, 37(2), p. 221. Available at: <https://doi.org/10.22564/rbgf.v37i2.173>.

Lupinacci, W.M. *et al.* (2023) 'Controls of fracturing on porosity in pre-salt carbonate reservoirs', *Energy Geoscience*, 4(2), p. 100146. Available at: <https://doi.org/10.1016/j.engeos.2022.100146>.

Mahesh, B. (2020) 'Machine Learning Algorithms - A Review'. Available at: <https://www.ijsr.net/archive/v9i1/ART20203995.pdf>.

Maleki, F. *et al.* (2020) 'Machine Learning Algorithm Validation', *Neuroimaging Clinics of North America*, 30(4), pp. 433–445. Available at: <https://doi.org/10.1016/j.nic.2020.08.004>.

Mello, M., de Mio, E. and Bruno, P. (2018) *Petroleum System Overview of São Francisco Basin: A New Gas Province Onshore Brazil*. Available at: <https://www.researchgate.net/publication/329655355>.

Mustafa, A. *et al.* (2023) 'Machine learning accelerated approach to infer nuclear magnetic resonance porosity for a middle eastern carbonate reservoir', *Scientific Reports*, 13(1), p. 3956. Available at: <https://doi.org/10.1038/s41598-023-30708-7>.

Offshore Technology (2021) 'Itapu Oil and Gas Field, Santos Basin, Brazil'. Available at: <https://www.offshore-technology.com/projects/itapu-oil-and-gas-field-santos-basin-brazil/>.

Pinheiro Junior, C.R. *et al.* (2021) 'Genesis and classification of carbonate soils in the State of Rio de Janeiro, Brazil', *Journal of South American Earth Sciences*, 108, p. 103183. Available at: <https://doi.org/10.1016/j.jsames.2021.103183>.

Reis, H.L.S. *et al.* (2017) 'The São Francisco Basin', in M. Heilbron, U.G. Cordani, and F.F. Alkmim (eds) *São Francisco Craton, Eastern Brazil*. Cham: Springer International Publishing (Regional Geology Reviews), pp. 117–143. Available at: https://doi.org/10.1007/978-3-319-01715-0_7.

Rocha, H.O.D. *et al.* (2019) 'Petrophysical characterization using well log resistivity and rock grain specific surface area in a fractured carbonate pre-salt reservoir in the Santos Basin, Brazil', *Journal of Petroleum Science and Engineering*, 183, p. 106372. Available at: <https://doi.org/10.1016/j.petrol.2019.106372>.

Sarker, I.H. (2021) 'Machine Learning: Algorithms, Real-World Applications and Research Directions', *SN Computer Science*, 2(3), p. 160. Available at: <https://doi.org/10.1007/s42979-021-00592-x>.

Schlumberger (2023) 'Energy Glossary'. Available at: <https://glossary.slb.com>.

'Scikit-learn, Machine learning in Python' (2023). Available at: <https://scikit-learn.org/>.

Souza, I.V.A.F. *et al.* (2022) 'Geochemical characterization of natural gases in the pre-salt section of the Santos Basin (Brazil) focused on hydrocarbons and volatile organic sulfur compounds', *Marine and Petroleum Geology*, 144, p. 105763. Available at: <https://doi.org/10.1016/j.marpetgeo.2022.105763>.

Tamoto, H., Gioria, R.D.S. and Carneiro, C.D.C. (2023) 'Prediction of nuclear magnetic

resonance porosity well-logs in a carbonate reservoir using supervised machine learning models', *Journal of Petroleum Science and Engineering*, 220, p. 111169. Available at: <https://doi.org/10.1016/j.petrol.2022.111169>.

Westphal, H. *et al.* (2005) 'NMR Measurements in Carbonate Rocks: Problems and an Approach to a Solution', *pure and applied geophysics PAGEOPH*, 162(3), pp. 549–570. Available at: <https://doi.org/10.1007/s00024-004-2621-3>.

7 Appendixes

7.1 Building Machine Learning Models – Complete Workflow

We have included specific lines of Python code that were written for predicting NMR porosity logs. The complete Python code can be accessed through the link provided in the Introduction Chapter.

IMPORTING MODELS AND REQUIRED DEPENDENCIES

```
%pip install --upgrade scikit-learn==1.2.2
%pip install qbstyles

# Importing the models
from sklearn.ensemble import RandomForestRegressor
from sklearn.ensemble import GradientBoostingRegressor
from sklearn.multioutput import MultiOutputRegressor

# Importing the dependencies
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import matplotlib.patches as mpatches
import pickle

from qbstyles import mpl_style
mpl_style(dark=False)

from sklearn.model_selection import train_test_split
from sklearn.model_selection import cross_val_score, KFold
from sklearn.metrics import r2_score, mean_squared_error, mean_squared_error, mean_absolute_error

from matplotlib_inline.backend_inline import set_matplotlib_formats
set_matplotlib_formats('svg')
```

LOADING THE WELL LOG DATA (WELL 3-BRSA-1215-RJS)

```
# Load the csv well log data to Pandas DataFrame
df = pd.read_csv("df_1_ML.csv")
df
```

SPECIFYING PREDICTORS & TARGETS

```
predictors = ["GR", "HCAL", "PEFZ", "RHOZ", "DTCO", "NPHI", "AT90"]
outputs = ["CMRP_3MS", "CMFF", "TCMR"] # They are predicted at the same time

X = df[predictors]
y = df[outputs]
```

TRAINING & TEST WELL-LOG DATASETS

```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

# Shape of the training and test datasets
print(X.shape, y.shape, X_train.shape, X_test.shape, y_train.shape, y_test.shape)
```

TRAINING & TUNING OF THE RANDOM FOREST MODEL

```
# Import the dependencies
from sklearn.model_selection import RandomizedSearchCV

# RANDOM FOREST Hyperparameters

# Number of trees to be used
rf_n_estimators = [100, 150, 200, 250, 300, 350, 400]

# Maximum number of levels in tree
rf_max_depth = [5, 10, 15, 20, 25]

# Criterion to split on
rf_criterion = ['squared_error']

# Create the grid
rf_grid = {'n_estimators': rf_n_estimators,
          'max_depth': rf_max_depth,
          'criterion': rf_criterion}

# Model to be tuned
rf_model = RandomForestRegressor(random_state=42) # Shuffle=True by default

# Create the random search Random Forest
```



```
rf_random = RandomizedSearchCV(rf_model, rf_grid, n_iter=20, cv=10, random_state=42)

# Fit the random search model
rf_random.fit(X_train, y_train)
```

TRAINING & TUNING OF THE GRADIENT BOOSTING MODEL

```
# GRADIENT BOOSTING Hyperparameters

# Number of trees to be used
gb_n_estimators = [100, 150, 200, 250, 300, 350, 400]

# Maximum number of levels in tree
gb_max_depth = [5, 10, 15, 20, 25]

# Learning rate
gb_rate = [0.1, 0.2, 0.3, 0.4, 0.5, 0.6]

# Criterion to split on
gb_criterion = ['squared_error']          # It is optional

# Create the grid
gb_grid = {'estimator__n_estimators': gb_n_estimators,
           'estimator__max_depth': gb_max_depth,
           'estimator__learning_rate': gb_rate,
           'estimator__criterion': gb_criterion}

# Model to be tuned
gb_model = GradientBoostingRegressor(random_state=42) # Shuffle=True by default

# Create the random search Gradient Boosting
gb_random = RandomizedSearchCV(MultiOutputRegressor(gb_model), gb_grid, n_iter=20, cv=10,
                               random_state=42)

# Fit the random search model
gb_random.fit(X_train, y_train)
```

PREDICTION WITH THE TEST DATASET

```
# Create the tuned Random Forest
```

```
rf_final_model=RandomForestRegressor(n_estimators=400, max_depth=20, random_state=42, criterion =
'squared_error')

# Create the tuned Gradient Boosting
gb_final_model=MultiOutputRegressor(GradientBoostingRegressor(n_estimators=250, learning_rate=0.3,
max_depth=5, random_state=42))

# Train the tuned Random Forest
rf_final_model.fit(X_train, y_train)

# Train the tuned Gradient Boosting
gb_final_model.fit(X_train, y_train)

# Prediction on Test data (RF)
y_pred_rf = rf_final_model.predict(X_test)

# Prediction on Test data (GB)
y_pred_gb = gb_final_model.predict(X_test)

# Prediction on Test data (GB)
y_pred_gb = gb_final_model.predict(X_test)

# Extract output 1, output 2 and output 3
y_test_out1 = y_test.drop(columns=["CMFF","TCMR"]) # CMRP_3MS
y_test_out2 = y_test.drop(columns=["CMRP_3MS","TCMR"]) # CMFF
y_test_out3 = y_test.drop(columns=["CMRP_3MS","CMFF"]) # TCMR

# Output 1, Output 2 and Output 3 should be converted into arrays
y_test_out1_ = np.array(y_test_out1)
y_test_out2_ = np.array(y_test_out2)
y_test_out3_ = np.array(y_test_out3)

# Extract the predicted values for output 1, output 2 and output 3 (RF)
y_pred_out1_rf = y_pred_rf[:, 0]
y_pred_out2_rf = y_pred_rf[:, 1]
y_pred_out3_rf = y_pred_rf[:, 2]

# Extract the predicted values for output 1, output 2 and output 3
y_pred_out1_gb = y_pred_gb[:, 0]
y_pred_out2_gb = y_pred_gb[:, 1]
y_pred_out3_gb = y_pred_gb[:, 2]
```

EVALUATING THE PERFORMANCE OF THE MODELS WITH REGRESSION METRICS

```
# RANDOM FOREST MODEL
```

```
# List of variables
```

```
test_data = [y_test_out1_, y_test_out2_, y_test_out3_]
```

```
predicted_data = [y_pred_out1_rf, y_pred_out2_rf, y_pred_out3_rf]
```

```
# Create "for loop" that calculates the Regression Metrics for each variable, separately
```

```
def regression_metrics_rf():
```

```
    for i in range(len(test_data)):
```

```
        rmse = mean_squared_error(test_data[i], predicted_data[i], squared = False)
```

```
        rmse = round(rmse,4)
```

```
        mse = mean_squared_error(test_data[i], predicted_data[i], squared = True)
```

```
        mse = round(mse,4)
```

```
        mae = mean_absolute_error(test_data[i], predicted_data[i])
```

```
        mae = round(mae,4)
```

```
        print("Regression metrics for Variable", i+1)
```

```
        print("Root Mean Squared Error:", rmse)
```

```
        print("Mean Squared Error:", mse)
```

```
        print("Mean Absolute Error:", mae)
```

```
        print("="*90)
```

```
# Call function
```

```
regression_metrics_rf()
```

```
# GRADIENT BOOSTING MODEL
```

```
# List of variables
```

```
test_data = [y_test_out1_, y_test_out2_, y_test_out3_]
```

```
predicted_data = [y_pred_out1_gb, y_pred_out2_gb, y_pred_out3_gb]
```

```
# Create "for loop" that calculates the Regression Metrics for each variable, separately
```

```
def regression_metrics_gb():
```

```
    for i in range(len(test_data)):
```

```
rmse = mean_squared_error(test_data[i], predicted_data[i], squared = False)
rmse = round(rmse,4)

mse = mean_squared_error(test_data[i], predicted_data[i], squared = True)
mse = round(mse,4)

mae = mean_absolute_error(test_data[i], predicted_data[i])
mae = round(mae,4)

print("Regression metrics for Variable", i+1)
print("Root Mean Squared Error:", rmse)
print("Mean Squared Error:", mse)
print("Mean Absolute Error:", mae)
print("="*90)

# Call function
regression_metrics_gb()
```

FEATURE IMPORTANCE

RANDOM FOREST MODEL

```
features_rf_1 = np.round(rf_final_model.estimators_[0].feature_importances_,4)
features_rf_2 = np.round(rf_final_model.estimators_[1].feature_importances_,4)
features_rf_3 = np.round(rf_final_model.estimators_[2].feature_importances_,4)
```

Print Features

```
features_rf_1, features_rf_2, features_rf_3
```

GRADIENT BOOSTING MODEL

```
features_gb_1 = np.round(gb_final_model.estimators_[0].feature_importances_,4)
features_gb_2 = np.round(gb_final_model.estimators_[1].feature_importances_,4)
features_gb_3 = np.round(gb_final_model.estimators_[2].feature_importances_,4)
```

Print Features

```
features_gb_1, features_gb_2, features_gb_3
```

SAVE THE RANDOM FOREST AND GRADIENT BOOSTING MODELS AS PICKLE FILES

```
import pickle # Library for save and load scikit-learn models

# RANDOM FOREST MODEL

# Define file name. ".pickle" as file extension. A pickle file is a binary file.
filename = "random_forest.pickle"

# Save Random Forest Model by means of "pickle.dump" function to store the object data to the file.
# This function takes 2 arguments:
# Object that you want to store.
# File object you get by opening the desired file in write-binary (wb) mode.
pickle.dump(rf_final_model, open(filename, "wb"))

# Load Random Forest Model by means of the "pickle.load" function.
# The primary argument of the function is the file object you get by opening the desired file in read-binary (rb)
mode.
random_forest_model_loaded = pickle.load(open(filename, "rb"))

# To print the trained and tuned random forest model
print(random_forest_model_loaded)

# GRADIENT BOOSTING MODEL

# Define file name. ".pickle" as file extension.
filename = "gradient_boosting.pickle"

# Save Gradient Boosting Model by means of "pickle.dump" function to store the object data to the file.
pickle.dump(gb_final_model, open(filename, "wb"))

# Load Gradient Boosting Model by means of the "pickle.load" function.
gradient_boosting_model_loaded = pickle.load(open(filename, "rb"))

# Print the trained and tuned random forest model
print(gradient_boosting_model_loaded)
```

LOAD SAVED MODELS & PERFORM NEW PREDICTION ON WELL 1-BRSA-1116-RJS

```
# Load the csv well log data to Pandas DataFrame
df1 = pd.read_csv("df_0_ML.csv")
df1

# RANDOM FOREST MODEL

# Load the trained and tuned Random Forest Model
RandomForestModel = pd.read_pickle('random_forest.pickle')

# Print the Model
RandomForestModel

# Compute new predictions by using the "unseen data" of WELL 1-BRSA-1116-RJS
predictors = ["GR", "HCAL", "PEFZ", "RHOZ", "DTCO", "NPHI", "AT90"]
outputs = ["CMRP_3MS", "CMFF", "TCMR"] # They are predicted at the same time

X = df1[predictors]
y = df1[outputs]

# New prediction on the entire dataset
y_predicted_rf1 = RandomForestModel.predict(X)

# Extract the predicted values for output 1, output 2 and output 3
y_predicted_out1_rf1 = y_predicted_rf1[:, 0]
y_predicted_out2_rf1 = y_predicted_rf1[:, 1]
y_predicted_out3_rf1 = y_predicted_rf1[:, 2]

# Extract output 1, output 2 and output 3
y_out1_1 = y.drop(columns=["CMFF", "TCMR"]) # CMRP_3MS
y_out2_1 = y.drop(columns=["CMRP_3MS", "TCMR"]) # CMFF
y_out3_1 = y.drop(columns=["CMRP_3MS", "CMFF"]) # TCMR

# Output 1, Output 2 and Output 3 should be converted into arrays
y_out1_1_ = np.array(y_out1_1)
y_out2_1_ = np.array(y_out2_1)
y_out3_1_ = np.array(y_out3_1)
```



```
# GRADIENT BOOSTING MODEL

# Load the trained and tuned Gradient Boosting Model
GradientBoostingModel = pd.read_pickle('gradient_boosting.pickle')

# Print the Model
GradientBoostingModel

# Compute new predictions by using the "unseen data" of WELL 1-BRSA-1116-RJS
predictors = ["GR","HCAL","PEFZ","RHOZ","DTCO","NPHI","AT90"]
outputs = ["CMRP_3MS","CMFF","TCMR"] # They are predicted at the same time

X = df1[predictors]
y = df1[outputs]

# New prediction on the entire dataset
y_predicted_gb1 = GradientBoostingModel.predict(X)

# Extract the predicted values for output 1, output 2 and output 3
y_predicted_out1_gb1 = y_predicted_gb1[:, 0]
y_predicted_out2_gb1 = y_predicted_gb1[:, 1]
y_predicted_out3_gb1 = y_predicted_gb1[:, 2]
```

EVALUATING THE PERFORMANCE OF THE MODELS WITH REGRESSION METRICS

```
# RANDOM FOREST MODEL

# List of variables
data = [y_out1_1_, y_out2_1_, y_out3_1_]
predicted_data = [y_predicted_out1_rf1, y_predicted_out2_rf1, y_predicted_out3_rf1]

# Create "for loop" that calculates the Regression Metrics for each variable, separately
def regression_metrics_rf1():
    for i in range(len(data)):

        r2 = r2_score(data[i], predicted_data[i])
        r2 = round(r2,4)
```

```
r2 = r2*100

rmse = mean_squared_error(data[i], predicted_data[i], squared = False)
rmse = round(rmse,4)

mse = mean_squared_error(data[i], predicted_data[i], squared = True)
mse = round(mse,4)

mae = mean_absolute_error(data[i], predicted_data[i])
mae = round(mae,4)

print("Regression metrics for Variable", i+1)
print("Coefficient of Determination:", r2)
print("Root Mean Squared Error:", rmse)
print("Mean Squared Error:", mse)
print("Mean Absolute Error:", mae)
print("="*90)

# Call the function
regression_metrics_rf1()

# GRADIENT BOOSTING MODEL

# List of variables
data = [y_out1_1_, y_out2_1_, y_out3_1_]
predicted_data = [y_predicted_out1_gb1, y_predicted_out2_gb1, y_predicted_out3_gb1]

# Create "for loop" that calculates the Regression Metrics for each variable, separately
def regression_metrics_gb1():
    for i in range(len(data)):

        r2 = r2_score(data[i], predicted_data[i])
        r2 = round(r2,4)
        r2 = r2*100

        rmse = mean_squared_error(data[i], predicted_data[i], squared = False)
        rmse = round(rmse,4)
```

```
mse = mean_squared_error(data[i], predicted_data[i], squared = True)
mse = round(mse,4)

mae = mean_absolute_error(data[i], predicted_data[i])
mae = round(mae,4)

print("Regression metrics for Variable", i+1)
print("Coefficient of Determination:", r2)
print("Root Mean Squared Error:", rmse)
print("Mean Squared Error:", mse)
print("Mean Absolute Error:", mae)
print("="*90)

# Call the function
regression_metrics_gb1()
```