

From Statistical Assessment to Data-Driven Forecasting: Groundwater Level Analysis and Modelling in Southern Ontario

Amirabbas Goudarzie

A THESIS SUBMITTED TO THE FACULTY OF GRADUATE
STUDIES IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF MASTER OF APPLIED SCIENCE

GRADUATE PROGRAM IN CIVIL ENGINEERING

York University

Toronto, Ontario

April 2026

© Amirabbas Goudarzie, 2026

Abstract

Groundwater is a vital resource in Southern Ontario, supplying 90% of rural residents and around 200 municipalities with drinking water and irrigation for agriculture. Accurate groundwater level (GWL) estimation is important for sustainable water management and environmental conservation. However, direct GWL measurements are often costly and spatially limited, underscoring the need for groundwater models for decision-making. Conventional numerical models solve physics-based partial differential equations at every node of a meshed geometry. Conversely, machine learning (ML) algorithms rely on mathematical equations between inputs and outputs by iteratively adjusting the model's parameters. Each method has its strengths, but a key advantage of ML models is their ability to predict GWLs without requiring the calibration of uncertain boundary conditions, extensive hydrological parameters, and human-induced factors. In this study, Provincial Groundwater Monitoring Network (PGMN) wells, the largest publicly available dataset for GWL data in Southern Ontario, were used to detect long-term trends using the Mann-Kendall test. Results show that 92% of wells exhibit statistically significant trends, indicating the non-stationarity of the GWLs. This study also assessed the effectiveness of ensembles of Artificial Neural Networks (ANNs) – a type of ML model – and compared it to linear regression models to predict monthly GWL in PGMN wells throughout the Barrie-Oro Moraine, an important hydrological area in Southern Ontario. The models incorporated data preprocessing techniques and lagged variables such as temperature, total precipitation, hydrometric levels, and autoregressive GWL data. Results indicate that linear regression is inappropriate since its assumptions, such as linearity and homoscedasticity, are violated, with test R^2 less than 0.2 across wells. Conversely, ensembles of ANNs exhibit better performance than linear regression, particularly in single watershed modelling scenarios, capturing the non-stationarity of GWL data, with median R^2 values generally ranging between 0.5 and 0.9. In addition, single-well models generally outperformed combined and aggregated modelling configurations, and DMS-based preprocessing generally showed improved performance compared to DSM approaches. These results highlight the usefulness of data-driven models in GWL prediction studies. ML techniques may improve predictive accuracy and help water resource managers develop effective policies for groundwater management.

Acknowledgements

I would like to express my sincere gratitude to my supervisor, Dr. Usman T. Khan and Dr. Magdalena Krol, for their continuous guidance, insightful feedback, and unwavering support throughout this research. Their expertise and encouragement played a fundamental role in shaping this thesis and strengthening my analytical and research skills.

I am also thankful to the member of my supervisory committee, Dr. Matthew Perras for his constructive comments and valuable suggestions, which significantly improved the quality and clarity of this work.

I acknowledge the Provincial Groundwater Monitoring Network (PGMN), Environment and Climate Change Canada (ECCC), and the Ontario Geological Survey (OGS) for maintaining and making publicly available the datasets used in this research. The availability of high-quality open data was essential for completing this study.

Finally, I am deeply grateful to my family and friends for their constant encouragement, patience, and emotional support throughout this journey. I would especially like to thank my mother, Fatemeh, my father, Alimohammad, and my sisters, Mahsa and Mahta, whose unwavering belief in me made this achievement possible. Their love and support have been a constant source of motivation throughout my studies.

Table of Contents

Abstract.....	ii
Acknowledgements.....	iii
Table of Contents	iv
List of Tables.....	x
List of Figures.....	xiv
List of Abbreviations.....	xi
Chapter 1 Introduction.....	1
1.1 Background.....	4
1.2 Research Problem and Objectives	8
1.3 Thesis Structure	9
Chapter 2 Literature Review.....	11
2.1 Background and Importance of Groundwater Resources	11
2.2 Numerical Models for GWL Modelling	15
2.3 Machine Learning-based GWL Models	19
2.4 Multi-output and Single Output Models for GWL Prediction.....	23
2.5 Research Gaps and Contributions of the Study	28
Chapter 3 Study Area and Data.....	31
3.1 Study Area	32
3.1.1 Southern Ontario	32
3.1.2 Transition from Southern Ontario to Barrie-Oro Boundary.....	34
3.1.3 Barrie-Oro Moraine Boundary	35

3.2	Data Sources	37
3.2.1	GWL Data	37
3.2.2	Meteorological Data.....	41
3.2.3	Hydrometric Data.....	44
3.2.4	Watersheds	46
3.2.5	Aquifers.....	47
3.2.6	GWL Data Resolution	48
3.2.7	Data Inputs for GWL Modelling.....	49
Chapter 4	Methodology	54
4.1	Data Preprocessing	54
4.1.1	Handling Missing Data	54
4.1.2	Outlier Removal Algorithm Using Interquartile Range (IQR) Method	58
4.1.3	Singular Spectrum Analysis (SSA)	59
4.1.4	Data Normalization: Min-Max Scaler.....	62
4.2	Statistical Analysis Methods.....	62
4.2.1	Hypothesis Testing	63
4.2.2	Trend Analysis.....	66
4.2.3	Sen's Slope (Theil-Sen Estimator)	68
4.3	Modelling Scenarios	69
4.3.1	Single Model	69
4.3.2	Combined Model.....	70

4.3.3	Aggregated Model.....	71
4.3.4	DSM and DMS Scenarios	72
4.4	Aquifer Detection and Data-Driven Models	73
4.4.1	Aquifer Detection Model (AQM)	73
4.4.2	Correlation Analysis.....	75
4.4.3	Regression Analysis	77
4.4.4	Artificial Neural Networks (ANNs).....	80
4.4.5	Model Training and Validation Techniques.....	82
4.5	Hyperparameter Tuning.....	87
4.5.1	Tuning Method.....	87
4.5.2	Hyperparameters for Linear Regression	87
4.5.3	Hyperparameters for Neural Networks	88
4.5.4	Validation and Testing During Tuning	88
4.6	Evaluation Metrics.....	89
4.7	Sensitivity Analysis (SA) Methods.....	91
4.7.1	SHAP	91
4.7.2	One-At-a-Time (OAT).....	91
Chapter 5	Results.....	93
5.1	Introduction to Results	93
5.2	Exploratory Data Analysis Results	93
5.2.1	Meteorological Data Scarcity Analysis in Southern Ontario	94

5.2.2	Data Scarcity Analysis in Barrie-Oro Moraine	99
5.2.3	Geospatial Filtering of ERA5 Nodes	103
5.2.4	GWL Trend Analysis.....	105
5.2.5	GWL Slope Distributions in Overburden and Bedrock Aquifers.....	109
5.2.6	Spatially Averaged GWL in Southern Ontario	111
5.2.7	Seasonally Averaged GWL in Southern Ontario.....	114
5.2.8	SSA on the Daily and Monthly Data.....	118
5.2.9	Correlation Analysis Results	119
5.2.10	Aquifer Detection Model (AQM) Results	126
5.3	Inputs in the Data-Driven Models	128
5.4	Linear Regression Results	128
5.4.1	Variance Inflation Factor (VIF).....	130
5.4.2	Exogeneity	131
5.4.3	Residual Distribution	133
5.4.4	Homoscedasticity	135
5.4.5	Autocorrelation in Residuals.....	136
5.4.6	Linear Regression Results for All Single Well Scenarios	138
5.5	Artificial Neural Networks Results	140
5.5.1	Performance Plateau.....	140
5.5.2	Single, Combined, or Aggregated?	153
5.6	Model Implications.....	169

5.6.1	Prediction	170
5.6.2	Sensitivity Analysis (SA)	176
Chapter 6	Discussion	198
6.1	Interpretation of Regional GWL Trends and Non-Stationarity	198
6.2	Physical Interpretation: Aquifer Type and Hydrogeological Context.....	199
6.3	Performance of Linear Regression vs. Neural Networks	200
6.4	Single, Combined, and Aggregated Modelling Scenarios	202
6.5	Impact of SSA Preprocessing (DSM vs. DMS).....	205
6.6	Sensitivity Analysis and Model Interpretability	206
6.7	Implications for Groundwater Modelling and Contribution to the Literature	208
Chapter 7	Conclusions and Recommendations.....	210
7.1	Limitations.....	212
7.2	Recommendations	213
References.....		214
Appendices.....		236
Appendix A		236
ERA5 vs. PGMN Wells in Southern Ontario.....		236
Data Inputs for GWL Modelling.....		237
DMS vs. DSM Inputs.....		244
Linear Regression Assumptions.....		251
Exogeneity		251

Residual Distribution	254
Homoscedasticity	258
Autocorrelation in Residuals.....	261
Appendix B	265
K-Fold Cross-Validation for All Well Datasets.....	265
Performance Plateau By-Fold Plots (R^2)	272
Performance Plateau Overall Plots (R^2).....	294
Performance Plateau By-Fold Plots (MSE)	305
Performance Plateau Overall Plots (MSE).....	328
Performance Plateau By-Fold Plots (AIC).....	339
Performance Plateau Overall Plots (AIC).....	362

List of Tables

Table 2.1 -Summary of recent studies on GWL prediction using ML and clustering-based methods, highlighting model types, aquifer types, watershed scale, input variables, temporal resolution, and forecast horizons.	28
Table 3.1-Geological and administrative characteristics of selected PGMN wells within the BOM area. The table includes information on the associated conservation authority, township location, aquifer composition, and record numbers for wells equipped with pressure dataloggers.....	40
Table 3.2-Physical characteristics and geographic coordinates of selected PGMN wells in the BOM region. The table includes well depth, screened intervals, and geographic location (latitude, longitude, and elevation above mean sea level) for each well.	41
Table 3.3-Summary of the eight monitoring wells selected within the BOM area, grouped by quaternary watershed. Each well is associated with a Hydrologic Unit Code (HUC), watershed name, catchment area, watershed ID, and whether it lies within a nested or separate watershed system.	46
Table 3.4-Aquifer classification and lithological composition for selected PGMN wells within the BOM area, based on Cuddy (2013) and the OGS hydrostratigraphic model (Burt & Dodge, 2011).	48
Table 3.5-Descriptive statistics (mean, standard deviation, minimum, 25th percentile, median, 75th percentile, and maximum) of the key input variables and GWL for well 293-2 over the study period (2006-2018).....	52
Table 4.1-K-fold cross-validation (KFCV) framework illustrating the data partitioning strategy across multiple iterations. In each iteration, one fold is used for validation, the last fold is fixed for testing, and the remaining folds are used for training.	85
Table 5.1-Time lags and cross-correlation coefficients corresponding to strongest correlation between the GWL data for each well under the DSM scenario.	123
Table 5.2- Time lags and cross-correlation coefficients corresponding to strongest correlation between the GWL data for each well under the DMS scenario.	123

Table 5.3- Autocorrelation coefficients at lag 1 (rt1) and corresponding 95% confidence intervals (CI) for the eight monitored wells shown in Figure 5.24. All wells exhibit strong autocorrelation, indicating significant influence of previous GWLs on current values.....	126
Table 5.4- Comparison of aquifer assignments for the PGMN wells based on Tier 2 groundwater studies and the MATLAB-based Aquifer AQM.....	127
Table 5.5- Mean training and test performance metrics of single-well linear regression models under chronological 4-fold cross-validation. Values represent averages across folds.	130
Table 5.6-Variance Inflation Factor (VIF) results for well 293-2. High VIF values for GWL (t-1) and Hydrometric level indicate strong multicollinearity, while Temperature and Precipitation remain within acceptable linear ranges. Interpretations follow the thresholds defined by Vu et al. (2015).	131
Table 5.7-VIF values for all of the wells in the study.....	131
Table 5.8-Mean of residuals for all of the wells.....	133
Table 5.9-P-values of the K-S test for all of the wells	134
Table 5.10-P-values of the B-P test for all of the wells.....	136
Table 5.11-Maximum ACF coefficients for all of the wells.....	138
Table 5.12- Summary of linear regression assumption diagnostics for all eight wells. ✓ indicates that the assumption is satisfied based on statistical testing, while × indicates violation.	139
Table 5.13- Summary of run times and optimal number of neurons selected for each model based on R^2 performance. Fold-specific values indicate the neuron count that achieved the highest R^2 on the validation set, while the overall value reflects the point of performance plateau when combining results across all folds. Models are grouped by watershed, with the final row representing the aggregated model trained on all wells.	143
Table 5.14- Summary of run times and optimal number of neurons selected for each model based on MSE performance. Fold-specific values indicate the neuron count that resulted in the lowest MSE on the validation set, while the overall value reflects the point where adding more neurons no longer	

meaningfully reduced error. Models are grouped by watershed, with the final row showing results for the aggregated model trained on all wells.....	146
Table 5.15-The optimum neuron for each well in the DMS-FFNN scenario	151
Table 5.16- Summary statistics (Median, Q1, Q3, and IQR) of test set R^2 scores for all model configurations. The table includes results for single models, combined models, and the aggregated model.	156
Table 5.17- Summary statistics (Median, Q1, Q3, and IQR) of test set MSEs for all model configurations. The table includes results for single models, combined models, and the aggregated model.	160
Table 5.18-Summary of test set R^2 performance for DMS-FFNN models across different spatial configurations. Values represent the median, first quartile (Q1), third quartile (Q3), and interquartile range (IQR) of R^2 scores across all ensemble runs.	165
Table 5.19-Summary of test Mean Squared Error (MSE) statistics for DMS-FFNN models across different scenarios. The table presents the median, first quartile (Q1), third quartile (Q3), and IQR for each model, including single-well, combined, and aggregated configurations.	168
Table A.1-Descriptive statistics (mean, standard deviation, minimum, 25th percentile, median, 75th percentile, and maximum) of the key input variables and GWL for well 293-3 over the study period (2006–2018).	237
Table A.2-Descriptive statistics (mean, standard deviation, minimum, 25th percentile, median, 75th percentile, and maximum) of the key input variables and GWL for well 439-1 over the study period (2011–2018).	238
Table A.3-Descriptive statistics (mean, standard deviation, minimum, 25th percentile, median, 75th percentile, and maximum) of the key input variables and GWL for well 440-1 over the study period (2011–2018).	239

Table A.4-Descriptive statistics (mean, standard deviation, minimum, 25th percentile, median, 75th percentile, and maximum) of the key input variables and GWL for well 442-1 over the study period (2011–2018).....	240
--	-----

Table A.5-Descriptive statistics (mean, standard deviation, minimum, 25th percentile, median, 75th percentile, and maximum) of the key input variables and GWL for well 244-2 over the study period (2006–2020).....	241
--	-----

Table A.6-Descriptive statistics (mean, standard deviation, minimum, 25th percentile, median, 75th percentile, and maximum) of the key input variables and GWL for well 245-2 over the study period (2006–2020).....	242
--	-----

Table A.7-Descriptive statistics (mean, standard deviation, minimum, 25th percentile, median, 75th percentile, and maximum) of the key input variables and GWL for well 443-1 over the study period (2010–2018).....	243
--	-----

List of Figures

Figure 2-1-Conceptual groundwater system showing confined and unconfined aquifers and GWL as the water table (Source: Ontario Ministry of Agriculture, Food and Rural Affairs, 2021).....	14
Figure 3-1-Southern Ontario boundary and its surrounding lakes and lands.....	34
Figure 3-2-Map of Southern Ontario showing the five publicly available 3D surficial aquifer data regions provided by the Ontario Geological Survey (OGS), along with locations of PGMN wells classified by aquifer type.	35
Figure 3-3- Map of the study area showing the locations of PGMN wells, meteorological stations, ERA5 grids, and hydrometric stations. Additional features include quaternary watersheds, watercourses, waterbodies, and the boundary of the BOM study area, overlaid on a digital elevation model.	36
Figure 3-4-(a) Elevation map of Southern Ontario displaying the distribution of PGMN wells. Elevation is represented in meters above sea level. (b) Regional division of Ontario into Northern and Southern zones based on latitude (46°N).....	39
Figure 3-5-Spatial distribution of ECCC meteorological stations across Ontario, categorized into northern and southern stations based on geographic location.....	42
Figure 3-6-ERA5 grid points covering Southern Ontario, illustrating the spatial resolution and distribution of reanalysis climate data across the study region.....	44
Figure 3-7- Conceptual hydrostratigraphic cross-section of BOM area showing key aquifers and aquitards based on OGS classification, derived from Cuddy (2013).....	48
Figure 3-8-Monthly groundwater consumption trends in Ontario from the PTTW dataset, showing peak withdrawals during the late spring and summer months of 2019 and 2020.	51
Figure 3-9-Time series plots of the monthly (a) temperature, (b) precipitation, (c) hydrometric level, and (d) GWL at well 293-2, showing seasonal and long-term variations from 2006 to 2018.	53
Figure 4-1- Cross-correlation between ECCC and ERA5 temperature datasets to investigate potential time lags. The analysis shows that the maximum correlation occurs at lag 0, indicating that no	

significant temporal offset exists between the two datasets. Subsequent analyses focused on evaluating amplitude differences. 55

Figure 4-2-Scatter plots comparing ERA5 and ECCC datasets: (a) shows a strong linear relationship between ERA5 and ECCC temperature values ($R^2 = 0.92$), while (b) reveals a weak correlation between ERA5 and ECCC precipitation data ($R^2 = 0.16$). Red lines represent the fitted linear regression models..... 56

Figure 4-3-Comparison of ECCC and ERA5 temperature and precipitation datasets. Panel (a) shows observed ECCC air temperature, and (b) shows ERA5 air temperature for the same period. Panels (c) and (d) present ECCC and ERA5 precipitation data, respectively. The shaded blue region highlights the data gap in ECCC records, which was filled using ERA5 data. Temperature data were bias-corrected using a regression model, while ERA5 precipitation was used directly due to low correlation. 57

Figure 4-4-Final corrected ECCC climate data after gap-filling. Panel (a) shows the air temperature time series with the missing data section (highlighted in yellow) filled using the bias-adjusted ERA5 temperature dataset. Panel (b) presents the precipitation series, where the missing values were directly replaced with ERA5 precipitation data due to weak correlation. 58

Figure 4-5- Schematic of the single model scenario, where each feature set (F) is used independently to predict groundwater well (G) levels within the same watershed (W). For example, F1 is used separately to model G1 and G2 in W1, while F2 is used to predict G3 and G4 in W2. 70

Figure 4-6-Schematic of the combined model scenario, where multiple groundwater wells (G) within the same watershed (W) are predicted simultaneously using a shared feature set (F). For instance, F2 is used to jointly predict G3 and G4 in W2. 71

Figure 4-7-Schematic of the aggregated model scenario, where all wells (G1 to G4) across watersheds W1 and W2 are predicted using a pooled feature set (F1 and F2). This setup assesses model performance when spatial boundaries are relaxed and information is shared across regions..... 72

Figure 4-8-Aquifer identification workflow diagram. Aquifer assignments were determined by integrating various data sources, including continuous and discontinuous aquifer data from OGS, Tier 2 project document (Cuddy, 2013), and other literature. 74

Figure 4-9-Schematic architecture of a FFNN with one hidden layer. The network consists of an input layer $x_i \in R^{m1}$, a hidden layer $h_j \in R^{m2}$, and an output layer $y_k \in R^{m3}$. Each neuron in one layer is connected to all neurons in the next layer, illustrating the fully connected structure..... 81

Figure 4-10- K-fold cross-validation (KFCV) used to partition the normalized monthly GWL data from well 293-2. Each panel (a–d) illustrates a different fold where one segment is used for validation (orange) and the remaining segments for training (blue), while the most recent 20% of the data (green) is fixed as the testing set. 84

Figure 4-11-Distribution of randomly selected seed values used in the multi-start initialization process for single modelling scenarios. Each subplot corresponds to a specific monitoring well: (a) 293-2, (b) 293-3, (c) 439-1, (d) 440-1, (e) 442-1, (f) 244-2, (g) 245-2, and (h) 443-1. 86

Figure 5-1- Binary visualization of daily temperature data availability across Southern Ontario stations (2000–2023). Each vertical line represents a unique station (Climate ID), and each horizontal line corresponds to a calendar date. Black segments indicate available data, while white gaps represent missing entries. 94

Figure 5-2- Proportion of missing temperature values (NaN ratio) per station across Southern Ontario (2000–2023). Each bar represents a station’s ratio of missing daily observations, with values approaching 1 indicating nearly complete data loss and values near 0 reflecting high data completeness. 95

Figure 5-3- Distribution of climate stations in Southern Ontario based on NaN ratio categories (2000–2023). The chart groups 408 stations into five categories according to the percentage of missing temperature data..... 96

Figure 5-4- Binary visualization of daily precipitation data availability across Southern Ontario stations (2000–2023). Each vertical line represents a unique station (Climate ID), and each horizontal line

corresponds to a calendar date. Black segments indicate available data, while white gaps represent missing entries. 97

Figure 5-5- Proportion of missing precipitation values (NaN ratio) per station across Southern Ontario (2000–2023). Each bar represents a station’s ratio of missing daily observations, with values approaching 1 indicating nearly complete data loss and values near 0 reflecting high data completeness. 98

Figure 5-6- Distribution of Climate IDs by precipitation data completeness in Southern Ontario from 2000 to 2023, categorized by NaN ratio ranges. 98

Figure 5-7- Availability of daily temperature observations from ECCC stations in the BOM area (2000–2023). Black segments indicate dates with available data, while white gaps represent missing entries. 100

Figure 5-8- Availability of daily precipitation observations from ECCC stations in the BOM area (2000–2023). Black segments indicate dates with available data, while white gaps represent missing entries. 101

Figure 5-9- Availability of daily temperature data interpolated from ECCC stations in the BOM area (2000–2023). Black segments indicate dates with available data, while white gaps represent missing entries. 102

Figure 5-10- Availability of daily precipitation data interpolated from ECCC stations in the BOM area (2000–2023). Black segments indicate dates with available data, while white gaps represent missing entries. 103

Figure 5-11- Example of geospatial filtering used to select ERA5 grid nodes for gap filling. ERA5 grid nodes (blue points) identified within a 10 km search radius of a PGMN monitoring well in southern Ontario; these nodes were used to supplement missing meteorological observations from ECCC stations. The search radius is adjustable in the analysis. 104

Figure 5-12- Spatial matching of ERA5 nodes (blue crosses) and PGMN wells (black circles) in the BOM area. Dotted lines illustrate the connections used to identify nearby grid points within a 15 km radius of each well for data interpolation and analysis. 105

Figure 5-13-Results of the Mann-Kendall trend test for GWLs at nine selected monitoring wells:
 (a) Well 441-1, (b) Well 362-2, (c) Well 71-1, (d) Well 519-2, (e) Well 361-2, (f) Well 44-3, (g) Well 299-1, (h) Well 250-1, and (i) Well 427-1. 107

Figure 5-14-Distribution of statistically significant and non-significant GWL trends across different aquifer types in the PGMN network. Panel (a) shows the total number of wells with significant increasing and decreasing trends, while panel (b) breaks down the trend direction (increasing, decreasing, or no trend) by aquifer type: Overburden (OV), Bedrock (BD), and Interface (IF). Percentages reflect each category's share within the trend group. 108

Figure 5-15-Distribution of statistically significant Mann-Kendall GWL trends across shallow and deep wells in the Oak Ridges Moraine Groundwater Program (ORMGP) study area. Source: Oak Ridges Moraine Groundwater Program (2023)..... 109

Figure 5-16-Violin plot of Sen's slope estimates for wells with statistically significant Mann-Kendall trends, grouped by aquifer type. Each distribution illustrates the rate of long-term GWL change (in m/year) for bedrock, overburden, and interface aquifers. Mean (μ), standard deviation (σ), and median ($\tilde{x}$) values are reported, with overburden wells showing higher slope magnitudes on average. 110

Figure 5-17-Time series of spatially averaged GWL from overburden wells across Southern Ontario between 2000 and 2021. The black line shows the mean GWL (in meters above sea level), and the blue line represents the number of active wells contributing to each time point. The red dashed line is a fitted linear trend ($R^2 = 0.569$), indicating a long-term increase in GWL. 112

Figure 5-18- Time series of spatially averaged GWL from bedrock wells across Southern Ontario between 2000 and 2021. The black line shows the mean GWL (in meters above sea level), and the blue line represents the number of active wells contributing to each time point. The red dashed line is a fitted linear trend ($R^2 = 0.041$), indicating a long-term increase in GWL. 114

Figure 5-19-Seasonal trends of average GWL in overburden aquifers across Southern Ontario from 2005 to 2017. Each panel (a–d) represents one season: (a) Spring, (b) Summer, (c) Fall, and (d)

Winter. Black lines indicate seasonal average GWLs, while red dashed lines represent fitted linear trendlines with corresponding regression equations and R^2 values. 116

Figure 5-20- Seasonal patterns of mean GWLs in overburden aquifers across Southern Ontario from 2004 to 2017, displayed on a single graph to allow direct comparison. 117

Figure 5-21- Comparison of original monthly data (blue) and SSA-processed data (orange) for well 293-2, where SSA was applied at the daily scale prior to aggregation. The subplots represent (a) air temperature ($^{\circ}\text{C}$), (b) precipitation (mm), (c) surface water level (m), and (d) GWL (m)..... 118

Figure 5-22- Comparison of original monthly data (blue) and SSA-processed data (orange) for well 293-2, with SSA applied after aggregation to monthly frequency. The subplots show (a) air temperature ($^{\circ}\text{C}$), (b) precipitation (mm), (c) surface water level (m), and (d) GWL (m)..... 119

Figure 5-23-Cross-correlation between GWL and hydroclimatic variables for the wells: (a) 293-2, (b) 293-3, (c) 439-1, (d) 440-1, (e) 442-1, (f) 244-2, (g) 245-2, and (h) 443-1. 121

Figure 5-24-Autocorrelation functions (ACF) for the wells: (a) 293-2, (b) 293-3, (c) 439-1, (d) 440-1, (e) 442-1, (f) 244-2, (g) 245-2, and (h) 443-1..... 125

Figure 5-25-Histogram plots of test residuals for well 293-2 across four cross-validation folds. The red dashed line in each subplot indicates the zero reference (ideal unbiased residuals), while the solid blue line shows the fold-specific mean residual. The mean values are: (a) -0.1138, (b) -0.0290, (c) 0.0420, and (d) 0.0200. These results support approximate exogeneity with minor deviations across folds. 132

Figure 5-26-Histograms of test residuals for Well 293-2 across four cross-validation folds (a–d), overlaid with Kernel Density Estimate (KDE) curves. Each subplot reports the p-value from the Kolmogorov–Smirnov (K–S) test for normality. In all cases, p-values exceed the 0.05 threshold, indicating no significant deviation from a normal distribution. Despite slight asymmetries, especially in folds 2 and 3, the residuals statistically satisfy the normality assumption..... 134

Figure 5-27-Residuals versus normalized predicted GWL values across four cross-validation folds for well 293-2. The Breusch–Pagan test results are annotated on each subplot to assess homoscedasticity. Folds (a) and (b) show heteroscedastic behavior ($p < 0.05$), while folds (c) and (d) meet

the homoscedasticity assumption. This variability highlights instability in the linear model's error structure across splits. 135

Figure 5-28-Autocorrelation function (ACF) plots of residuals for each cross-validation fold for well 293-2. Each subplot (a–d) corresponds to a separate fold and shows the autocorrelation values up to lag 30. The blue shaded region indicates the 95% confidence interval. Residuals show noticeable positive autocorrelation at early lags in all folds, reflecting the autoregressive nature of the modelling framework. 137

Figure 5-29- R^2 score distributions across different neuron counts for well 293-2 in each cross-validation fold. Each subplot (a–d) represents a separate fold, with training (red), validation (blue), and test (green) results shown. The optimal number of neurons varies across folds—from as low as 3 in Fold 1 to higher values such as 11 in Fold 3—demonstrating fold-dependent behavior. While early folds show a clear plateau at lower neuron counts, Fold 3 exhibits gradual improvements over a wider range, highlighting the variability in model performance across folds..... 141

Figure 5-30- Combined R^2 score distributions for well 293-2 across training (red), validation (blue), and test (green) sets at different neuron counts. While performance begins to plateau around 4 neurons, overall improvements in R^2 and reduced variability continue up to 8 neurons, indicating it as the optimal point. 142

Figure 5-31-MSE distributions across neuron counts for well 293-2 in each cross-validation fold. Subplots (a–d) show Folds 1 to 4 with training (red), validation (blue), and test (green) results. The optimal neuron count varies by fold, with minimum MSE at lower counts (around 3–4) in Folds 1, 2, and 4. In contrast, Fold 3 shows a broader range with gradual improvement. 144

Figure 5-32-Combined MSE distributions for well 293-2 across all cross-validation folds, shown for different neuron counts. Each boxplot represents the training (red), validation (blue), and test (green) performance. While the MSE drops sharply up to around 4 neurons, smaller improvements continue beyond that point. Based on the flattening trend and reduced variability, the optimal number of neurons is identified around 8. 145

Figure 5-33-AIC distributions across neuron counts for well 293-2 in each cross-validation fold. Subplots (a–d) show Folds 1 to 4, with training (red), validation (blue), and test (green) datasets. Lower AIC values generally occur at smaller neuron counts (around 1–4), especially in Folds 1, 2, and 4. Fold 3 exhibits a broader range of low AIC scores across more neurons. 148

Figure 5-34- Combined AIC distributions for well 293-2 across different neuron counts and datasets. Boxplots show AIC values for training (red), validation (blue), and test (green) sets across all folds. The lowest AIC scores occur at smaller neuron counts, particularly around 3–5 neurons. Although AIC values rise steadily with increasing model size, the curve flattens after neuron 8, indicating diminishing gains in model fit relative to complexity..... 149

Figure 5-35- Combined R^2 distributions for the DMS-FFNN setup across different neuron counts. Each boxplot shows training (red), validation (blue), and test (green) results. The MSE decreases sharply up to around 4 neurons, with smaller gains and reduced variability continuing to 8 neurons. 151

Figure 5-36- Combined MSE distributions for the DMS-FFNN setup across different neuron counts. Each boxplot shows training (red), validation (blue), and test (green) results. The MSE decreases sharply up to around 4 neurons, with smaller gains and reduced variability continuing to 8 neurons. 152

Figure 5-37- Combined AIC distributions for the DMS-FFNN setup across different neuron counts. Boxplots show AIC values for training (red), validation (blue), and test (green) sets. The lowest AIC scores appear around 3–5 neurons, while values rise steadily with model size. 153

Figure 5-38-Boxplots of R^2 scores across training (red), validation (blue), and test (green) datasets for all model configurations, grouped by watershed. Each group includes results from individual wells and their respective combined model. The final group on the right shows the fully aggregated model trained on all wells. The gray line (right y-axis) indicates the number of available observations per model. Box annotations show the number of features (f), optimal number of neurons in the hidden layer (n), and number of output targets (o)..... 154

Figure 5-39- Boxplots of Mean Squared Error (MSE) scores across training (red), validation (blue), and test (green) datasets for all model configurations, grouped by watershed. Each group includes

results for individual wells and their corresponding combined model. The rightmost group shows the fully aggregated model trained on all wells. The gray line (right y-axis) indicates the number of available observations per model. Box annotations indicate the number of features (f), optimal number of neurons (n), and number of output variables (o). 158

Figure 5-40-Distribution of R^2 scores across training (red), validation (blue), and test (green) sets for all DMS-FFNN model configurations. Each group includes results for single-well models, combined watershed models, and the aggregated model trained on all wells. The gray line shows the number of observations per model. Box annotations indicate the number of input features (f), selected hidden neurons (n), and output targets (o). 162

Figure 5-41-Box plots of test, validation, and training MSE values for all DMS-FFNN model scenarios. Each group includes models trained on individual wells, combined wells within each watershed (W1–W3), and the aggregated model trained on all wells (far right, in red). The number of observations used for training each model is shown as a gray line, while model configuration details—number of features (f), hidden units (n), and output targets (o)—are provided in boxed annotations. 166

Figure 5-42-Comparison of actual and predicted normalized GWL over time for single models. The black solid line represents the actual GWL, while the blue dashed line indicates the median prediction from an ensemble of 4800 neural network models. The shaded cyan region represents the IQR of the predictions. Each subplot corresponds to a specific single model scenario: (a) 293-2, (b) 293-3, (c) 439-1, (d) 440-1, (e) 442-1, (f) 244-2, (g) 245-2, and (h) 443-1. The number of optimum neurons in the hidden layer (N) for each scenario is indicated in the respective subplot..... 171

Figure 5-43- Comparison of actual and predicted normalized GWL over time for combined models. The black solid line represents the actual GWL, while the blue dashed line indicates the median prediction from an ensemble of 4800 neural network models. The shaded cyan region represents the IQR of the predictions. Each subplot corresponds to a specific single model scenario: (a) 293-2, (b) 293-3, (c) 439-1, (d) 440-1, (e) 442-1, (f) 244-2, (g) 245-2, and (h) 443-1. The number of optimum neurons in the hidden layer (N) for each scenario is indicated in the respective subplot..... 173

Figure 5-44- Comparison of actual and predicted normalized GWL over time for the aggregated model. The black solid line represents the actual GWL, while the blue dashed line indicates the median prediction from an ensemble of 4800 neural network models. The shaded cyan region represents the IQR of the predictions. Each subplot corresponds to a specific single model scenario: (a) 293-2, (b) 293-3, (c) 439-1, (d) 440-1, (e) 442-1, (f) 244-2, (g) 245-2, and (h) 443-1. The number of optimum neurons in the hidden layer (N) for each scenario is indicated in the respective subplot..... 175

Figure 5-45-Mean absolute SHAP values for input features across eight wells, representing model sensitivity to key hydro-meteorological variables. Panels (a–c) show sensitivity to temperature (T), precipitation (P), and hydrometric level (H), with red stars indicating deeper wells and black squares representing shallower wells. Panels (d–e) display the distance to the nearest hydrometric and meteorological stations (d), and the elevation difference relative to these stations (e). Wells are grouped by watershed (W1–W4) as classified before..... 177

Figure 5-46- Mean GWL change (%) in response to $\pm 30\%$ changes in the autoregressive GWL feature for each well, while all other inputs are held constant. Each subplot corresponds to a different single DMS model: (a) 293-2, (b) 293-3, (c) 439-1, (d) 440-1, (e) 442-1, (f) 244-2, (g) 245-2, and (h) 443-1. A linear regression is fitted to each response data, with slope values indicating the strength of autoregressive influence. The results represent the median GWL change derived from 4800 independently trained neural network simulations. 182

Figure 5-47-Pearson correlation heatmap showing the relationships between GWL time series at different monitoring wells. Higher correlation values (closer to 1) indicate stronger linear associations between wells, while negative or near-zero values suggest weak or inverse relationships. 184

Figure 5-48- Mean GWL change (%) in response to $\pm 30\%$ changes in the autoregressive GWL feature for each well, while all other inputs are held constant. Each subplot corresponds to a different combined DMS model: (a) 293-2, (b) 293-3, (c) 439-1, (d) 440-1, (e) 442-1, (f) 244-2, (g) 245-2, and (h) 443-1. A linear regression is fitted to each response data, with slope values indicating the strength of

autoregressive influence. The results represent the median GWL change derived from 4800 independently trained neural network simulations. 187

Figure 5-49- Mean GWL change (%) as a function of $\pm 30\%$ changes in the mean values of temperature, precipitation, and hydrometric level feature. For each scenario, two features remain constant at their mean values while the third feature is varied for the purpose of sensitivity analysis. The results represent the median response of an ensemble of 200 neural network (NN) simulations. Results are shown for different single model scenarios: (a) 293-2, (b) 293-3, (c) 439-1, (d) 440-1, (e) 442-1, (f) 244-2, (g) 245-2, and (h) 443-1. 189

Figure 5-50-OAT sensitivity analysis of hydrometeorological variables for the combined models of wells 293-2 (panel a) and 293-3 (panel b). Each line represents the mean GWL (%) change in response to $\pm 30\%$ variation in a single climate input—temperature (black), precipitation (red), or hydrometric level (blue)—while all other inputs are held at their average values. Labels indicate the delay time step (e.g., $t-2$) used for each feature, corresponding to the most influential lag per well identified via cross-correlation analysis..... 192

Figure 5-51-OAT sensitivity analysis of hydrometeorological variables for the combined models of wells 439-1, 440-1, and 442-1, shown in panels (a), (b), and (c) respectively. Each line represents the mean GWL (%) change in response to $\pm 30\%$ variation in a single climate input—temperature (black), precipitation (red), or hydrometric level (blue)—while all other inputs are held constant at their mean values. Time step labels (e.g., $t-7$) indicate the most influential lag per variable, selected based on cross-correlation between each well's GWL and the corresponding input..... 195

Figure 5-52- OAT sensitivity analysis of hydrometeorological variables for the combined models of wells 244-2 and 245-2, shown in panels (a), and (b) respectively. Each line represents the mean GWL (%) change in response to $\pm 30\%$ variation in a single climate input—temperature (black), precipitation (red), or hydrometric level (blue)—while all other inputs are held constant at their mean values. Time step labels (e.g., $t-5$) indicate the most influential lag per variable, selected based on cross-correlation between each well's GWL and the corresponding input. 197

<i>Figure A.1-Spatial distribution of ERA5 grids and PGMN wells in Southern Ontario</i>	<i>236</i>
Figure A.2- Time series plots of the monthly (a) temperature, (b) precipitation, (c) hydrometric level, and (d) groundwater level (GWL) at well 293-2, showing seasonal and long-term variations from 2006 to 2017.	237
Figure A.3- Time series plots of the monthly (a) temperature, (b) precipitation, (c) hydrometric level, and (d) GWL at well 439-1, showing seasonal and long-term variations from 2011 to 2018.....	238
Figure A.4- Time series plots of the monthly (a) temperature, (b) precipitation, (c) hydrometric level, and (d) GWL at well 440-1, showing seasonal and long-term variations from 2011 to 2018.....	239
Figure A.5- Time series plots of the monthly (a) temperature, (b) precipitation, (c) hydrometric level, and (d) GWL at well 442-1, showing seasonal and long-term variations from 2011 to 2018.....	240
Figure A.6-Time series plots of the monthly (a) temperature, (b) precipitation, (c) hydrometric level, and (d) GWL at well 244-2, showing seasonal and long-term variations from 2006 to 2020.	241
Figure A.7-Time series plots of the monthly (a) temperature, (b) precipitation, (c) hydrometric level, and (d) GWL at well 245-2, showing seasonal and long-term variations from 2006 to 2020.	242
Figure A.8-Time series plots of the monthly (a) temperature, (b) precipitation, (c) hydrometric level, and (d) GWL at well 443-1, showing seasonal and long-term variations from 2010 to 2018.	243
Figure A. 9- Comparison of original monthly data (blue) and SSA-processed data (orange) for well 293-3, where SSA was applied at the daily scale prior to aggregation. The subplots represent (a) air temperature (°C), (b) precipitation (mm), (c) surface water level (m), and (d) GWL (m).....	244
Figure A.10- Comparison of original monthly data (blue) and SSA-processed data (orange) for well 293-3, with SSA applied after aggregation to monthly frequency. The subplots show (a) air temperature (°C), (b) precipitation (mm), (c) surface water level (m), and (d) GWL (m).....	244
Figure A. 11- Comparison of original monthly data (blue) and SSA-processed data (orange) for well 439-1, where SSA was applied at the daily scale prior to aggregation. The subplots represent (a) air temperature (°C), (b) precipitation (mm), (c) surface water level (m), and (d) GWL (m).....	245

Figure A.12- Comparison of original monthly data (blue) and SSA-processed data (orange) for well 439-1, with SSA applied after aggregation to monthly frequency. The subplots show (a) air temperature ($^{\circ}\text{C}$), (b) precipitation (mm), (c) surface water level (m), and (d) GWL (m)..... 245

Figure A.13- Comparison of original monthly data (blue) and SSA-processed data (orange) for well 440-1, where SSA was applied at the daily scale prior to aggregation. The subplots represent (a) air temperature ($^{\circ}\text{C}$), (b) precipitation (mm), (c) surface water level (m), and (d) GWL (m)..... 246

Figure A.14- Comparison of original monthly data (blue) and SSA-processed data (orange) for well 440-1, with SSA applied after aggregation to monthly frequency. The subplots show (a) air temperature ($^{\circ}\text{C}$), (b) precipitation (mm), (c) surface water level (m), and (d) GWL (m)..... 246

Figure A.15- Comparison of original monthly data (blue) and SSA-processed data (orange) for well 442-1, where SSA was applied at the daily scale prior to aggregation. The subplots represent (a) air temperature ($^{\circ}\text{C}$), (b) precipitation (mm), (c) surface water level (m), and (d) GWL (m)..... 247

Figure A.16- Comparison of original monthly data (blue) and SSA-processed data (orange) for well 442-1, with SSA applied after aggregation to monthly frequency. The subplots show (a) air temperature ($^{\circ}\text{C}$), (b) precipitation (mm), (c) surface water level (m), and (d) GWL (m)..... 247

Figure A.17- Comparison of original monthly data (blue) and SSA-processed data (orange) for well 244-2, where SSA was applied at the daily scale prior to aggregation. The subplots represent (a) air temperature ($^{\circ}\text{C}$), (b) precipitation (mm), (c) surface water level (m), and (d) GWL (m)..... 248

Figure A.18- Comparison of original monthly data (blue) and SSA-processed data (orange) for well 244-2, with SSA applied after aggregation to monthly frequency. The subplots show (a) air temperature ($^{\circ}\text{C}$), (b) precipitation (mm), (c) surface water level (m), and (d) GWL (m)..... 248

Figure A.19- Comparison of original monthly data (blue) and SSA-processed data (orange) for well 245-2, where SSA was applied at the daily scale prior to aggregation. The subplots represent (a) air temperature ($^{\circ}\text{C}$), (b) precipitation (mm), (c) surface water level (m), and (d) GWL (m)..... 249

Figure A.20- Comparison of original monthly data (blue) and SSA-processed data (orange) for well 245-2, with SSA applied after aggregation to monthly frequency. The subplots show (a) air temperature ($^{\circ}\text{C}$), (b) precipitation (mm), (c) surface water level (m), and (d) GWL (m)..... 249

Figure A.21- Comparison of original monthly data (blue) and SSA-processed data (orange) for well 443-1, where SSA was applied at the daily scale prior to aggregation. The subplots represent (a) air temperature ($^{\circ}\text{C}$), (b) precipitation (mm), (c) surface water level (m), and (d) GWL (m)..... 250

Figure A.22- Comparison of original monthly data (blue) and SSA-processed data (orange) for well 443-1, with SSA applied after aggregation to monthly frequency. The subplots show (a) air temperature ($^{\circ}\text{C}$), (b) precipitation (mm), (c) surface water level (m), and (d) GWL (m)..... 250

Figure A.23- Histogram plots of test residuals for well 293-3 across four cross-validation folds. The red dashed line in each subplot indicates the zero reference (ideal unbiased residuals), while the solid blue line shows the fold-specific mean residual. 251

Figure A.24- Histogram plots of test residuals for well 439-1 across four cross-validation folds. The red dashed line in each subplot indicates the zero reference (ideal unbiased residuals), while the solid blue line shows the fold-specific mean residual. 251

Figure A.25- Histogram plots of test residuals for well 440-1 across four cross-validation folds. The red dashed line in each subplot indicates the zero reference (ideal unbiased residuals), while the solid blue line shows the fold-specific mean residual. 252

Figure A.26- Histogram plots of test residuals for well 442-1 across four cross-validation folds. The red dashed line in each subplot indicates the zero reference (ideal unbiased residuals), while the solid blue line shows the fold-specific mean residual. 252

Figure A.27- Histogram plots of test residuals for well 244-2 across four cross-validation folds. The red dashed line in each subplot indicates the zero reference (ideal unbiased residuals), while the solid blue line shows the fold-specific mean residual. 253

Figure A.28- Histogram plots of test residuals for well 245-2 across four cross-validation folds. The red dashed line in each subplot indicates the zero reference (ideal unbiased residuals), while the solid blue line shows the fold-specific mean residual.	253
Figure A.29- Histogram plots of test residuals for well 443-1 across four cross-validation folds. The red dashed line in each subplot indicates the zero reference (ideal unbiased residuals), while the solid blue line shows the fold-specific mean residual.	254
Figure A.30- Histograms of test residuals for Well 293-3 across four cross-validation folds (a–d), overlaid with Kernel Density Estimate (KDE) curves. Each subplot reports the p-value from the Kolmogorov–Smirnov (K–S) test for normality.	254
Figure A.31- Histograms of test residuals for Well 439-1 across four cross-validation folds (a–d), overlaid with Kernel Density Estimate (KDE) curves. Each subplot reports the p-value from the Kolmogorov–Smirnov (K–S) test for normality.	255
Figure A.32- Histograms of test residuals for Well 440-1 across four cross-validation folds (a–d), overlaid with Kernel Density Estimate (KDE) curves. Each subplot reports the p-value from the Kolmogorov–Smirnov (K–S) test for normality.	255
Figure A.33- Histograms of test residuals for Well 442-1 across four cross-validation folds (a–d), overlaid with Kernel Density Estimate (KDE) curves. Each subplot reports the p-value from the Kolmogorov–Smirnov (K–S) test for normality.	256
Figure A.34- Histograms of test residuals for Well 244-2 across four cross-validation folds (a–d), overlaid with Kernel Density Estimate (KDE) curves. Each subplot reports the p-value from the Kolmogorov–Smirnov (K–S) test for normality.	256
Figure A.35- Histograms of test residuals for Well 245-2 across four cross-validation folds (a–d), overlaid with Kernel Density Estimate (KDE) curves. Each subplot reports the p-value from the Kolmogorov–Smirnov (K–S) test for normality.	257

Figure A.36- Histograms of test residuals for Well 443-1 across four cross-validation folds (a–d), overlaid with Kernel Density Estimate (KDE) curves. Each subplot reports the p-value from the Kolmogorov–Smirnov (K–S) test for normality.	257
Figure A.37- Residuals versus normalized predicted GWL values across four cross-validation folds for well 293-3.....	258
Figure A.38- Residuals versus normalized predicted GWL values across four cross-validation folds for well 439-1.....	258
Figure A.39- Residuals versus normalized predicted GWL values across four cross-validation folds for well 440-1.....	259
Figure A.40- Residuals versus normalized predicted GWL values across four cross-validation folds for well 442-1.....	259
Figure A.41- Residuals versus normalized predicted GWL values across four cross-validation folds for well 244-2.....	260
Figure A.42- Residuals versus normalized predicted GWL values across four cross-validation folds for well 245-2.....	260
Figure A.43- Residuals versus normalized predicted GWL values across four cross-validation folds for well 443-1.....	261
Figure A.44- Autocorrelation function (ACF) plots of residuals for each cross-validation fold for well 293-3.	261
Figure A.45- Autocorrelation function (ACF) plots of residuals for each cross-validation fold for well 439-1.	262
Figure A.46- Autocorrelation function (ACF) plots of residuals for each cross-validation fold for well 440-1.	262
Figure A.47- Autocorrelation function (ACF) plots of residuals for each cross-validation fold for well 442-1.	263

Figure A.48- Autocorrelation function (ACF) plots of residuals for each cross-validation fold for well 244-2.	263
Figure A.49- Autocorrelation function (ACF) plots of residuals for each cross-validation fold for well 245-2.	264
Figure A.50- Autocorrelation function (ACF) plots of residuals for each cross-validation fold for well 443-1.	264
Figure B.1-KFCV used to partition the normalized monthly GWL data from well 293-3, DSM scenario.	265
Figure B.2- KFCV used to partition the normalized monthly GWL data from well 293-3, DMS scenario.	265
Figure B.3- KFCV used to partition the normalized monthly GWL data from well 439-1, DSM scenario.	266
Figure B. 4- KFCV used to partition the normalized monthly GWL data from well 439-1, DMS scenario.	266
Figure B.5- KFCV used to partition the normalized monthly GWL data from well 440-1, DSM scenario.	267
Figure B.6-KFCV used to partition the normalized monthly GWL data from well 440-1, DMS scenario.	267
Figure B.7-KFCV used to partition the normalized monthly GWL data from well 442-1, DSM scenario.	268
Figure B.8-KFCV used to partition the normalized monthly GWL data from well 442-1, DMS scenario.	268
Figure B.9-KFCV used to partition the normalized monthly GWL data from well 244-2, DSM scenario.	269
Figure B.10-KFCV used to partition the normalized monthly GWL data from well 244-2, DMS scenario.	269

Figure B. 11- KFCV used to partition the normalized monthly GWL data from well 245-2, DSM scenario.	270
Figure B.12-KFCV used to partition the normalized monthly GWL data from well 245-2, DMS scenario.	270
Figure B.13-KFCV used to partition the normalized monthly GWL data from well 443-1, DSM scenario.	271
<i>Figure B.14- KFCV used to partition the normalized monthly GWL data from well 443-1, DMS scenario.</i>	<i>271</i>
Figure B.15- R^2 score distributions across different neuron counts for well 293-3 in each cross-validation fold, DSM scenario.	272
Figure B.16- R^2 score distributions across different neuron counts for well 293-3 in each cross-validation fold, DMS scenario.	273
Figure B.17- R^2 score distributions across different neuron counts for W1 in each cross-validation fold, DSM scenario.	274
Figure B.18- R^2 score distributions across different neuron counts for W1 in each cross-validation fold, DMS scenario.	275
Figure B.19- R^2 score distributions across different neuron counts for well 439-1 in each cross-validation fold, DSM scenario.	276
Figure B.20- R^2 score distributions across different neuron counts for well 439-1 in each cross-validation fold, DMS scenario.	277
Figure B.21- R^2 score distributions across different neuron counts for well 440-1 in each cross-validation fold, DSM scenario.	278
Figure B.22- R^2 score distributions across different neuron counts for well 440-1 in each cross-validation fold, DMS scenario.	279
Figure B.23- R^2 score distributions across different neuron counts for well 442-1 in each cross-validation fold, DSM scenario.	280

Figure B.24- R^2 score distributions across different neuron counts for well 442-1 in each cross-validation fold, DMS scenario.	281
Figure B.25- R^2 score distributions across different neuron counts for W2 combined in each cross-validation fold, DSM scenario.	282
Figure B.26- R^2 score distributions across different neuron counts for W2 combined in each cross-validation fold, DMS scenario.	283
Figure B.27- R^2 score distributions across different neuron counts for well 244-2 in each cross-validation fold, DSM scenario.	284
Figure B.28- R^2 score distributions across different neuron counts for well 244-2 in each cross-validation fold, DMS scenario.	285
Figure B.29- R^2 score distributions across different neuron counts for well 245-2 in each cross-validation fold, DSM scenario.	286
Figure B.30- R^2 score distributions across different neuron counts for well 245-2 in each cross-validation fold, DMS scenario.	287
Figure B.31- R^2 score distributions across different neuron counts for W3 combined in each cross-validation fold, DSM scenario.	288
Figure B.32- R^2 score distributions across different neuron counts for W3 combined in each cross-validation fold, DMS scenario.	289
Figure B.33- R^2 score distributions across different neuron counts for well 443-1 in each cross-validation fold, DSM scenario.	290
Figure B.34- R^2 score distributions across different neuron counts for well 443-1 in each cross-validation fold, DMS scenario.	291
Figure B.35- R^2 score distributions across different neuron counts for the aggregated model in each cross-validation fold, DSM scenario.	292
Figure B.36- R^2 score distributions across different neuron counts for the aggregated model in each cross-validation fold, DMS scenario.	293

Figure B.37- Combined R^2 score distributions for well 293-3, DSM scenario.....	294
Figure B.38- Combined R^2 score distributions for well 293-3, DMS scenario.....	294
Figure B.39- Combined R^2 score distributions for W1 combined, DSM scenario.	295
Figure B.40- Combined R^2 score distributions for W1 combined, DMS scenario.	295
Figure B.41- Combined R^2 score distributions for well 439-1, DSM scenario.....	296
Figure B.42- Combined R^2 score distributions for well 439-1, DMS scenario.....	296
Figure B.43- Combined R^2 score distributions for well 440-1, DSM scenario.....	297
Figure B.44- Combined R^2 score distributions for well 440-1, DMS scenario.....	297
Figure B.45- Combined R^2 score distributions for well 442-1, DSM scenario.....	298
Figure B.46- Combined R^2 score distributions for well 442-1, DMS scenario.....	298
Figure B.47- Combined R^2 score distributions for W2 combined, DSM scenario.	299
Figure B.48- Combined R^2 score distributions for W2 combined, DMS scenario.	299
Figure B.49- Combined R^2 score distributions for well 244-2, DSM scenario.....	300
Figure B.50- Combined R^2 score distributions for well 244-2, DMS scenario.....	300
Figure B.51- Combined R^2 score distributions for well 245-2, DSM scenario.....	301
Figure B.52- Combined R^2 score distributions for well 245-2, DMS scenario.....	301
Figure B.53- Combined R^2 score distributions for W3 combined, DSM scenario.	302
Figure B.54- Combined R^2 score distributions for W3 combined, DMS scenario.	302
Figure B.55- Combined R^2 score distributions for well 443-1, DSM scenario.....	303
Figure B.56- Combined R^2 score distributions for well 443-1, DMS scenario.....	303
Figure B.57- Combined R^2 score distributions for the aggregated model, DSM scenario.	304
Figure B.58- Combined R^2 score distributions for the aggregated model, DMS scenario.	304
Figure B.59- MSE distributions across different neuron counts for well 293-2 in each cross-validation fold, DMS scenario.	305
Figure B.60- MSE distributions across different neuron counts for well 293-3 in each cross-validation fold, DSM scenario.	306

Figure B.61- MSE distributions across different neuron counts for well 293-3 in each cross-validation fold, DMS scenario.	307
Figure B.62- MSE distributions across different neuron counts for W1 combined in each cross-validation fold, DSM scenario.	308
Figure B.63- MSE distributions across different neuron counts for W1 combined in each cross-validation fold, DMS scenario.	309
Figure B.64- MSE distributions across different neuron counts for well 439-1 in each cross-validation fold, DSM scenario.	310
Figure B.65- MSE distributions across different neuron counts for well 439-1 in each cross-validation fold, DMS scenario.	311
Figure B.66- MSE distributions across different neuron counts for well 440-1 in each cross-validation fold, DSM scenario.	312
Figure B.67- MSE distributions across different neuron counts for well 440-1 in each cross-validation fold, DMS scenario.	313
Figure B.68- MSE distributions across different neuron counts for well 442-1 in each cross-validation fold, DSM scenario.	314
Figure B.69- MSE distributions across different neuron counts for well 442-1 in each cross-validation fold, DMS scenario.	315
Figure B.70- MSE distributions across different neuron counts for W2 combined in each cross-validation fold, DSM scenario.	316
Figure B.71- MSE distributions across different neuron counts for W2 combined in each cross-validation fold, DMS scenario.	317
Figure B.72- MSE distributions across different neuron counts for well 244-2 in each cross-validation fold, DSM scenario.	318
Figure B.73- MSE distributions across different neuron counts for well 244-2 in each cross-validation fold, DMS scenario.	319

Figure B.74- MSE distributions across different neuron counts for well 245-2 in each cross-validation fold, DSM scenario.	320
Figure B.75- MSE distributions across different neuron counts for well 245-2 in each cross-validation fold, DMS scenario.	321
Figure B.76- MSE distributions across different neuron counts for W3 combined in each cross-validation fold, DSM scenario.	322
Figure B.77- MSE distributions across different neuron counts for W3 combined in each cross-validation fold, DMS scenario.	323
Figure B.78- MSE distributions across different neuron counts for well 443-1 in each cross-validation fold, DSM scenario.	324
Figure B.79- MSE distributions across different neuron counts for well 443-1 in each cross-validation fold, DMS scenario.	325
Figure B.80- MSE distributions across different neuron counts for the aggregated model in each cross-validation fold, DSM scenario.....	326
Figure B.81- MSE distributions across different neuron counts for the aggregated model in each cross-validation fold, DMS scenario.....	327
Figure B.82- Combined MSE distributions for well 293-3, DSM scenario.....	328
Figure B.83- Combined MSE distributions for well 293-3, DMS scenario.....	328
Figure B.84- Combined MSE distributions for W1 combined model, DSM scenario.	329
Figure B.85- Combined MSE distributions for W1 combined model, DMS scenario.	329
Figure B.86- Combined MSE distributions for well 439-1, DSM scenario.....	330
Figure B.87- Combined MSE distributions for well 439-1, DMS scenario.....	330
Figure B.88- Combined MSE distributions for well 440-1, DSM scenario.....	331
Figure B.89- Combined MSE distributions for well 440-1, DMS scenario.....	331
Figure B.90- Combined MSE distributions for well 442-1, DSM scenario.....	332
Figure B.91- Combined MSE distributions for well 442-1, DMS scenario.....	332

Figure B.92- Combined MSE distributions for W2 combined model, DSM scenario.	333
Figure B.93- Combined MSE distributions for W2 combined model, DMS scenario.	333
Figure B.94- Combined score distributions for well 244-2, DSM scenario.	334
Figure B.95- Combined MSE distributions for well 244-2, DMS scenario.....	334
Figure B.96- Combined MSE distributions for well 245-2, DSM scenario.....	335
Figure B.97- Combined MSE distributions for well 245-2, DMS scenario.....	335
Figure B.98- Combined MSE distributions for W3 combined model, DSM scenario.	336
Figure B.99- Combined MSE distributions for W3 combined, DMS scenario.	336
Figure B.100- Combined MSE distributions for well 443-1, DSM scenario.....	337
Figure B.101- Combined MSE distributions for well 443-1, DMS scenario.....	337
Figure B.102- Combined MSE distributions for the aggregated model, DSM scenario.	338
Figure B.103- Combined MSE distributions for the aggregated model, DMS scenario.	338
Figure B.104- AIC distributions across different neuron counts for well 293-2 in each cross-validation fold, DMS scenario.	339
Figure B.105- AIC distributions across different neuron counts for well 293-3 in each cross-validation fold, DSM scenario.	340
Figure B.106- AIC distributions across different neuron counts for well 293-3 in each cross-validation fold, DMS scenario.	341
Figure B.107- AIC distributions across different neuron counts for W1 combined model in each cross-validation fold, DSM scenario.....	342
Figure B.108- AIC distributions across different neuron counts for W1 combined model in each cross-validation fold, DMS scenario.....	343
Figure B.109- AIC distributions across different neuron counts for well 439-1 in each cross-validation fold, DSM scenario.	344
Figure B.110- AIC distributions across different neuron counts for well 439-1 in each cross-validation fold, DMS scenario.	345

Figure B.111- AIC distributions across different neuron counts for well 440-1 in each cross-validation fold, DSM scenario.	346
Figure B.112- AIC distributions across different neuron counts for well 440-1 in each cross-validation fold, DMS scenario.	347
Figure B.113- AIC distributions across different neuron counts for well 442-1 in each cross-validation fold, DSM scenario.	348
Figure B.114- AIC distributions across different neuron counts for well 442-1 in each cross-validation fold, DMS scenario.	349
Figure B.115- AIC distributions across different neuron counts for W2 combined in each cross-validation fold, DSM scenario.	350
Figure B.116- AIC distributions across different neuron counts for W2 combined in each cross-validation fold, DMS scenario.	351
Figure B.117- AIC distributions across different neuron counts for well 244-2 in each cross-validation fold, DSM scenario.	352
Figure B.118- AIC distributions across different neuron counts for well 244-2 in each cross-validation fold, DMS scenario.	353
Figure B.119- AIC distributions across different neuron counts for well 245-2 in each cross-validation fold, DSM scenario.	354
Figure B.120- AIC distributions across different neuron counts for well 245-2 in each cross-validation fold, DMS scenario.	355
Figure B.121- AIC distributions across different neuron counts for W3 combined model in each cross-validation fold, DSM scenario.	356
Figure B.122- AIC distributions across different neuron counts for W3 combined model in each cross-validation fold, DMS scenario.	357
Figure B.123- AIC distributions across different neuron counts for well 443-1 in each cross-validation fold, DSM scenario.	358

Figure B.124- AIC distributions across different neuron counts for well 443-1 in each cross-validation fold, DMS scenario.	359
Figure B.125- AIC distributions across different neuron counts for the aggregated model in each cross-validation fold, DSM scenario.	360
Figure B.126- AIC distributions across different neuron counts for the aggregated model in each cross-validation fold, DMS scenario.	361
Figure B.127- Combined AIC distributions for well 293-3, DSM scenario.	362
Figure B.128- Combined AIC distributions for well 293-3, DMS scenario.	362
Figure B.129- Combined AIC distributions for W1 combined model, DSM scenario.	363
Figure B.130- Combined AIC distributions for W1 combined model, DMS scenario.	363
Figure B.131- Combined AIC distributions for well 439-1, DSM scenario.	364
Figure B.132- Combined AIC distributions for well 439-1, DMS scenario.	364
Figure B.133- Combined AIC distributions for well 440-1, DSM scenario.	365
Figure B.134- Combined AIC distributions for well 440-1, DMS scenario.	365
Figure B.135- Combined AIC distributions for well 442-1, DSM scenario.	366
Figure B.136- Combined AIC distributions for well 442-1, DMS scenario.	366
Figure B.137- Combined AIC distributions for W2 combined model, DSM scenario.	367
Figure B.138- Combined AIC distributions for W2 combined model, DMS scenario.	367
Figure B.139- Combined AIC distributions for well 244-2, DSM scenario.	368
Figure B.140- Combined AIC distributions for well 244-2, DMS scenario.	368
Figure B.141- Combined AIC distributions for well 245-2, DSM scenario.	369
Figure B.142- Combined AIC distributions for well 245-2, DMS scenario.	369
Figure B.143- Combined AIC distributions for W3 combined model, DSM scenario.	370
Figure B.144- Combined AIC distributions for W3 combined model, DMS scenario.	370
Figure B.145- Combined AIC distributions for well 443-1, DSM scenario.	371
Figure B.146- Combined AIC distributions for well 443-1, DMS scenario.	371

Figure B.147- Combined AIC distributions for the aggregated model, DSM scenario. 372

Figure B.148- Combined AIC distributions for the aggregated model, DMS scenario. 372

List of Abbreviations

ACF	Auto-Correlation Function
AI	Artificial Intelligence
ANN	Artificial Neural Network
AQM	Aquifer Detection Model
BG	Breusch-Godfrey Test
BOM	Barrie-Oro Moraine
BP	Breusch-Pagan Test
CNN	Convolutional Neural Network
DMS	Daily-Monthly-SSA
DSM	Daily-SSA-Monthly
DW	Durbin-Watson Test
ECCC	Environment and Climate Change Canada
ERA5	Fifth Generation of European Centre for Medium-Range Weather Forecasts Reanalysis
FFNN	Feedforward Neural Network
GWL	Groundwater Level
HYDAT	Hydrometric Historical Data
IQR	Interquartile Range
IVS	Input Variable Selection
KFCV	K-Fold Cross-Validation
KS	Kolmogorov-Smirnov Test (two-sample)
LAMPS	Laboratory of Mathematical Parallel System
LSTM	Long Short-Term Memory
ML	Machine Learning
MLP	Multi-layer Perceptron
MLR	Multiple Linear Regression
MSE	Mean Squared Error
OAT	One-At-a-Time
OCDP	Ontario Climate Data Portal

OGS	Ontario Geological Survey
PCA	Principal Component Analysis
PGMN	Provincial Groundwater Monitoring Network
PTTW	Permit to Take Water
R^2	Coefficient of Determination
SHAP	SHapley Additive exPlanations
SLR	Simple Linear Regression
RNN	Recurrent Neural Network
SA	Sensitivity Analysis
SSA	Singular Spectrum Analysis
VIF	Variance Inflation Factor
WSC	Water Survey of Canada

Chapter 1 Introduction

Groundwater reservoirs are the largest sources of freshwater, supplying approximately 50% of the world's drinking water and 43% of the water used for irrigation (Rajeevan & Mishra, 2020). In recent years, groundwater resources have faced increasing pressure from both natural and human-induced factors, including climate change, rising agricultural demands, population growth, and the overextraction of water from illegally drilled wells (McDonough et al., 2020; Uc Castillo et al., 2022).

Changes in groundwater levels (GWLs) are governed by complex recharge-discharge interactions, aquifer storage properties, and delayed flow responses within porous media. Groundwater systems commonly exhibit delayed flow response, nonlinear dynamics, and spatial heterogeneity driven by hydrostratigraphic variability (Freeze & Cherry, 1979; Fetter, 2018). These characteristics complicate both statistical trend detection and predictive modelling. While descriptive monitoring reports in parts of Canada suggest that widespread depletion is not universally evident (Bhatti, Farooque, Li, et al., 2021), systematic well-level statistical assessment of long-term GWL behavior across multiple locations remains limited, particularly at the provincial scale in Ontario. Quantitative trend analysis is therefore necessary to distinguish natural variability from persistent directional change and to support evidence-based groundwater management.

Traditionally, groundwater prediction has relied on physics-based numerical models such as MODFLOW and FEFLOW, which simulate subsurface flow using Darcy's law and mass conservation principles (McDonald & Harbaugh, 1988; Anderson et al., 2015). Although these models provide strong physical representation, their application at regional scales requires detailed hydrogeological parameterization, boundary condition specification, and high-quality spatial datasets, which are often incomplete or uncertain (Secci et al., 2023; Zeydlinejad, 2022). As a result, data-driven approaches have increasingly been adopted to complement or substitute numerical simulations, particularly where

observational datasets are available but subsurface characterization is limited (Rajaei et al., 2019; Tao et al., 2022).

Among data-driven techniques, artificial neural networks (ANNs) and their variants – including feedforward neural networks (FFNN), convolutional neural networks (CNN), recurrent neural networks (RNN), and long short-term memory (LSTM) models – have demonstrated strong capability in capturing nonlinear groundwater dynamics (Müller et al., 2021; Patra et al., 2023; Chidepudi et al., 2025). However, despite reported improvements in predictive accuracy, several methodological challenges remain unresolved. First, many studies either focus on predictive modelling without prior statistical diagnosis of long-term trends, or conduct trend analysis without integrating the findings into modelling design. Second, spatial modelling strategies differ substantially. While single-well models often demonstrate strong local performance (Müller et al., 2021; Patra et al., 2023), multi-output or multi-station models aim to exploit shared patterns across wells (Chidepudi et al., 2025). Yet, explicit comparison of single, physically grouped (e.g., watershed-based), and fully aggregated cross-watershed modelling structures within a unified regional framework remains limited.

In addition, clustering and grouping approaches in previous studies are frequently based on statistical similarity or hydrograph behavior (Sharghi et al., 2022; Kardan Moghaddam et al., 2021), rather than on physically defined hydrogeological boundaries such as watersheds. Given that wells within the same watershed may share recharge sources, climatic forcing, and surface-groundwater interactions, evaluating watershed-based grouping provides a physically grounded alternative to purely data-driven clustering. Furthermore, preprocessing decisions – particularly the timing of noise-reduction techniques relative to temporal aggregation – are rarely evaluated systematically in groundwater studies, even though signal smoothing can influence model stability and generalization performance.

Southern Ontario provides a suitable setting to address these methodological gaps. The region includes diverse hydrostratigraphic conditions, ranging from shallow overburden aquifers to deeper bedrock systems, and supports significant municipal and agricultural groundwater use. Long-term records

from the Provincial Groundwater Monitoring Network (PGMN) enable comprehensive statistical assessment of non-stationarity, while also supporting structured evaluation of predictive modelling strategies across multiple watersheds.

Accordingly, this thesis develops an integrated framework that combines statistical trend analysis, hydrogeological classification, and data-driven modelling within a single study. The research begins with systematic evaluation of long-term GWL trends using established statistical methods, followed by predictive modelling initiated with linear regression as a baseline benchmark. Given the frequent violation of regression assumptions in temporally persistent groundwater systems, neural network models are subsequently implemented to capture nonlinear relationships. Different spatial configurations – including single, combined, and aggregated models – are compared, alongside alternative preprocessing pipelines that evaluate the timing of Singular Spectrum Analysis (SSA). Model interpretability is further examined using structured sensitivity analysis approaches to assess how input variables influence predictive behavior under different spatial configurations.

By integrating statistical rigor, hydrogeological context, modelling design, and interpretability within a unified framework, this study aims to advance both methodological understanding and practical application of GWL forecasting in regions characterized by hydrogeological heterogeneity and climatic variability.

1.1 Background

Groundwater is an invaluable resource globally, playing a critical role in multiple aspects of human life, including agriculture, municipal services, industry, and drinking water supply (Sharpe et al., 2014). Groundwater refers to water that occupies the pore spaces and fractures in soil and rock. Groundwater does not remain static beneath the surface; instead, it moves through soil pores and rock fissures under natural hydraulic gradients. It is replenished mainly by the infiltration of precipitation (rain and snow), surface runoff, and water from surface bodies, and it discharges to rivers, lakes, springs, the sea, and seepage areas (Giordano, 2009). Groundwater also serves as a dependable source of fresh water for numerous regions worldwide (Morsy et al., 2018), providing drinking water for 2 billion people, meeting 40% of irrigation needs, and supporting ecosystem health (Kundzewicz & Döll, 2009).

In Canada, around 9 million people, making up 24% of the population, use groundwater as their drinking water source. This resource is especially crucial in rural areas, where 82% of residents depend on it (Bhatti, Farooque, Li, et al., 2021a). The use of groundwater varies widely across Canada. It meets municipal needs in Ontario, Yukon, and New Brunswick; serves industrial purposes in British Columbia, Quebec, and the Northwest Territories; supplies domestic wells in Newfoundland and Nova Scotia; and provides water for livestock in Alberta, Saskatchewan, and Manitoba. The province of Prince Edward Island (PEI) entirely depends on this resource for drinking, industrial, and agricultural needs (Nowlan, 2005).

Overall groundwater consumption in Canada remains well below sustainable limits proposed in the literature. In a national-scale assessment, sustainable groundwater use was defined as withdrawals of up to 40% of the estimated annual recharge, based on simplified recharge assumptions and intended for provincial-level comparison rather than local assessment (Bhatti, Farooque, Krouglicof, et al., 2021). Using this benchmark, reported groundwater withdrawals were below 4% of renewable availability for most provinces, with New Brunswick showing the highest use at approximately 8%, followed by Ontario at 4%, while Nova Scotia exhibited the lowest consumption at about 0.31% (Bhatti, Farooque, Krouglicof, et al., 2021). This indicates a positive regional scenario, as groundwater usage in Canada is generally sustainable,

with most groundwater recharge remaining untapped by humans (Nowlan, 2005). However, significant concerns persist at the local level, where issues of groundwater depletion and quality degradation arise in various scattered locations across Canada. Additionally, GWL decreases have been shown to have an inverse relationship with population rise, posing a potential threat if the population in a specific area increases in the future (Bhatti, Farooque, Li, et al., 2021)

Climate change, on the other hand, affects all hydrological patterns and parameters and is believed to be already impacting protected areas, ecosystems, and other natural assets on the globe (Lemieux & Scott, 2011). This phenomenon leads to rising temperatures, which directly affect key components of the hydrologic cycle, including evaporation, evapotranspiration, and precipitation (Bhatti, Farooque, Krouglicof, et al., 2021).

In Canada, average temperature have increased substantially since the mid-20th century, with an overall warming of approximately 1.7-2.0°C since 1948. Globally, average temperatures have increased by approximately 1.1°C since the pre-industrial period, indicating that Canada's temperature rise is exceeding the global average (Environment and Climate Change Canada, 2023; Our World in Data, 2024). Additionally, it has caused the southern and western regions of Canada, where most of the population resides, to become warmer. This has resulted in milder winters with increased rainfall. Moreover, climate change is expected to lead to an increase in extreme weather events, such as higher intensity and longer duration of precipitation, more severe droughts, summer heat waves, thunderstorms, ice storms, wind storms, and hurricanes (Gutierrez & LePrevost, 2016).

Although the full response of groundwater to climate change is not yet thoroughly understood since the climate system is delicate and is formed by a web of complicated interactions (Andrei et al., 2021; Munday et al., 2017), climate change is expected to intensify human pressure on groundwater resources. This will occur by altering the frequency and pattern of aquifer recharge due to changes in land use, soil properties, and rainfall (Ricolfi et al., 2020). Additionally, the recharge of aquifers is closely linked to the distribution and amount of precipitation, which is directly affected by climate change and the resulting

extreme weather events. Assessing the impact of climate change not only requires accurate prediction of climatic parameters but also needs reliable estimation of groundwater recharge. This complexifies the analysis of hydrologic alterations' impact on the groundwater resources compared to the surface water bodies that directly interact with the climatic factors (Jyrkama & Sykes, 2007). Additionally, groundwater is considered more resilient to hydrological changes than surface water because it is protected by the overlying soil, which shields it from direct contact with the surrounding environment (Barbieri et al., 2023).

Approximately 3 million residents in Ontario depend on groundwater for their drinking water needs. Climate change, in this province, is expected to increase river flow due to more intense and frequent precipitation, create warmer winters that boost recharge by reducing ground frost and enhancing infiltration, and raise evapotranspiration levels due to temperature fluctuations. It remains unclear which of these factors will have the most significant impact (Bhatti, Farooque, Li, et al., 2021b). In a study conducted on a sub-catchment of Lower Whitemans Creek, Ontario, Larocque et al., (2019) found that groundwater is largely resilient to climate change. However, they observed increased recharge and GWL during winter and fall, along with a depletion of groundwater recharge during the growing season.

Reporting based on the PGMN dataset that contains GWL observations in Ontario, Bhatti, Farooque, Li, et al. (2021) explained that most of the sites did not indicate any depletion of GWL in Ontario in a long-term period. For instance, at Nottawasaga, they reported that most wells showed increasing water levels between 2003 and 2015, indicating that extraction from these groundwater resources is below the safe yield while GWLs indicated downward trends in areas close to Toronto. However, this research has not investigated the GWL trends quantitatively and only explained the overall observable trends. According to the Toronto and Region Conservation Authority (TRCA) information regarding groundwater quantity, GWLs typically fluctuate 1 to 2 meters per year with higher amounts in spring and lower amounts in summer. Among the 23 PGMN wells located within TRCA's area, only one site exhibits a declining trend which is likely due to the limited number of observations, and it is anticipated that as more data is collected, the trend will stabilize and reach a plateau (TRCA, 2024). In a similar former report, (Markle & Harris

(2019) evaluated GWL trends using both site- and field-significance testing, and only field-significant results were interpreted at the regional scale.

While these studies provide valuable descriptive insights at selected sites, a comprehensive, province-wide quantitative assessment of long-term GWL trends in Ontario remains limited, particularly one that moves beyond visual interpretation toward systematic statistical analysis. Specifically, a precise and comprehensive analysis of long-term GWL changes in Ontario and its correlation with hydrological parameters such as precipitation and temperature has yet to be conducted. This study accurately analyzes the PGMN dataset, which includes approximately 500 well observations in Ontario, to detect GWL alterations over a long-term period (2000-2020). Few prior studies have thoroughly examined long-term site-level GWL trends across Ontario. While some Canadian provinces provide province-wide information on GWL behaviour using their own monitoring networks, these efforts are generally descriptive and indicator-based rather than comprehensive statistical trend analyses. For example, British Columbia reports long-term GWL trends using a provincial observation well network; however, these results are presented as monitoring indicators, where GWL trends are summarized into predefined categories (e.g., increasing, stable, moderate decline, or large decline) based on statistical thresholds, rather than comprehensive site-level trend assessment or quantitatively linked to hydroclimatic drivers (Environmental Reporting BC, 2024). This research answers whether GWLs in Southern Ontario, where most of Ontario's population resides, have increased or decreased, using statistical tests and data analysis processes rather than merely observing trends. For some wells in this dataset, missing values complicate the analysis; these gaps will be addressed in future steps of the project using various algorithms that are able to capture temporal cycles, including machine learning (ML) methods. For the hydrological parameters, stationary data from Environment and Climate Change Canada (ECCC) were processed to ensure more reliable information. Due to the high sparsity of this dataset across Southern Ontario, mainly caused by the intermittent operation of sensors and devices, the analysis was supplemented with secondary information such as the ERA5 (Fifth Generation of European Centre for Medium-Range Weather Forecasts Reanalysis) dataset. Each dataset has

its own drawbacks, necessitating the use of multiple data sources. For example, while ERA5 provides satisfactory temporal information, it lacks high spatial resolution, requiring the use of additional downscaling techniques. To understand the impact of climate change on groundwater resources in Southern Ontario, various ML models will be utilized in future studies. Additionally, by applying different climate change scenarios that forecast hydrological parameters for the coming years, future GWL changes will be evaluated.

1.2 Research Problem and Objectives

This study begins with a statistical analysis of GWL data obtained from the PGMN in Ontario, followed by the development of a data-driven framework to predict GWL at multiple wells within a local region, as described in Section 3.1.3. The first component investigates whether GWLs across Ontario exhibit increasing or decreasing trends by examining spatially and seasonally aggregated patterns. The second component explores appropriate modelling strategies for predicting GWL when multiple wells are located within the same area.

The main objective of this thesis is to enhance the predictive performance of data-driven models for GWL prediction using historical climate, hydrological, and groundwater data from PGMN wells in Ontario. The study aims to improve the accuracy and reliability of GWL forecasting by exploring different modelling configurations and preprocessing techniques. The primary modelling considerations include the evaluation of single-output and multi-output modelling frameworks (see Section 4.3), followed by systematic model fine-tuning to enhance predictive performance (Section 4.4).

The research addresses the following main goals:

- To assess long-term non-stationarity in groundwater levels across Ontario using statistical trend analysis.

- To compare neural network models with linear regression baselines, examining whether traditional statistical approaches can match the performance of more flexible ML models.
- To assess the performance of single-output and multi-output neural network models (including single, combined, and aggregated scenarios) in predicting monthly GWL values across different watersheds.
- To evaluate the effect of time series preprocessing using SSA applied at different stages – before and after temporal aggregation – on model accuracy and trend detection.
- To investigate model interpretability by applying sensitivity analysis methods, including SHapley Additive exPlanations (SHAP) and One-At-a-Time (OAT), to determine the relative influence of input features on model predictions.

To the best of current knowledge, this thesis represents one of the few studies focused on GWL data that begins with a comprehensive statistical analysis and integrates multiple analytical tools – such as SSA, linear regression and ML approaches, single- and multi-output modelling frameworks, model tuning, and sensitivity analysis – to evaluate their impact on final GWL predictions. By following the procedures outlined in this study, hydrogeologists and engineers can obtain a structured framework for developing reliable data-driven models to predict GWL in the absence of traditional modelling approaches.

Overall, the thesis aims to contribute to the development of robust and interpretable modelling frameworks for groundwater forecasting. While Ontario serves as the primary case study, the methodological framework is intended to be applicable to other regions in support of sustainable water resource management.

1.3 Thesis Structure

This thesis is organized into seven chapters. Chapter 1 introduces the background and motivation for GWL prediction, highlighting the limitations of current practices and the need for data-driven

approaches. Chapter 2 presents a detailed literature review of GWL forecasting, covering both traditional and ML-based techniques, as well as key concepts such as single, combined, and the aggregated models.

Chapter 3 outlines the study area and the data sources used in the analysis, including PGMN well data, climate inputs, and hydrometric information. Chapter 4 describes the modelling framework, including data preprocessing methods such as SSA, model architectures, training setups, and scenario configurations (Single, Combined, Aggregated, DSM, and DMS).

Chapter 5 reports the experimental results and includes a comparison of model performances using statistical metrics, as well as sensitivity analyses using SHAP and One-At-a-Time (OAT) methods. Chapter 6 presents the discussion of the results, including interpretation of findings, methodological implications, and comparison with previous studies. Chapter 7 concludes the thesis by summarizing the key findings, discussing limitations, and proposing directions for future research.

Chapter 2 Literature Review

This literature review discusses the importance of groundwater resources and GWL prediction for water resources management using different modelling approaches. Groundwater is a critical freshwater resource globally, and Canada possesses one of the largest freshwater endowments worldwide. Effective management of groundwater resources and GWL is essential to meet the needs of households, agriculture, and industry. .

2.1 Background and Importance of Groundwater Resources

Water is essential for life on Earth, as all living beings depend on it for survival, and it has been a key natural resource for human civilization throughout history (Banihashem et al., 2024). Only 3% of the planet's total water is freshwater, and the majority of it is stored in ice (A. Ghosh et al., 2022). The remaining 97% is found as sea water which is generally not suitable for human use. Groundwater accounts for around 30% of the global freshwater (and 99% of the liquid freshwater) while lakes and rivers form less than 1% of such resource (Lall et al., 2020; Musie & Gonfa, 2023). Thus, groundwater represents the largest accessible source of freshwater for human use.

Water resource management has gained greater significance globally due to the growing population and the scarcity of water (Kumar et al., 2021). As an example, one of the main objectives of the Sustainable Development Goals (SDGs), adopted by United Nations Member States at the 2015 UN Summit, is the global need for sustainable water and sanitation management to be achieved by 2030 (United Nations, n.d.). Therefore, ensuring the sustainable management of groundwater resources is essential to meet both current and future water needs, mitigate the impacts of climate change on water resources, and support socioeconomic development.

Groundwater plays a central role in achieving sustainable water management goals. It is important for supporting domestic, agricultural, and industrial activities, while maintaining the ecological balance of

the environment (Qadir et al., 2007). Groundwater also enhances climate resilience. It is a valuable resource during droughts, as surface water sources are more vulnerable to weather conditions (Di Nunno & Granata, 2020). Furthermore, the infrastructure required to extract groundwater is generally more cost-effective to build and maintain compared to that used for surface water, which often spans large and open areas (Ifediegwu, 2021).

Given its widespread use and economic importance, monitoring and predicting GWL fluctuations is essential. Such efforts are not only key to sustainable groundwater management but also influence the stability of surface water systems, soil integrity, and stream flow (Solgi et al., 2021). In addition to quantity, the quality of groundwater differs from surface water due to its prolonged underground residence. Groundwater tends to contain more dissolved minerals as it interacts with geological formations, whereas surface water has fewer minerals because of its exposure to the atmosphere (D. Ghosh et al., 2020). While groundwater quality is an important consideration, this study focuses on GWL variability as a key metric for evaluating the groundwater system dynamics.

Groundwater is contained in regions known as aquifers. An aquifer is a geological formation capable of storing and conveying substantial quantities of water. These formations vary in dimensions and soil composition, ranging from local to extensive areas (Singer et al., 2003). Aquifers have recently been stressed by overdrawn of groundwater, wastewater pollution, and salt intrusion in coastal areas, while the potential influence of climate variability and climate change on GWL has mainly been examined through regional and scenario-based modelling studies rather than direct long-term observations (Jeiouni et al., 2019).

Based on the geological characteristics, there are two primary types of aquifers: overburden and bedrock aquifers. Overburden aquifers mostly comprise gravel, sand, and silt, whereas bedrock aquifers retain water within the fissures and cracks of the underlying rock formations (Singer et al., 2003). Overburden deposits are located above bedrock and closer to the surface, rendering them more vulnerable to climate effects and the interactions between groundwater and surface water (O'Neill, 2023). A research

by Lipka (2010) evaluated overburden and bedrock aquifers in Fredericton, New Brunswick, Canada, focusing on their reactions to variations of temperature, precipitation, and recharge rates using FEFLOW. The findings demonstrated that overburden aquifers displayed heightened vulnerability to climate change and fluctuations in surface water relative to bedrock aquifers.

In addition to aquifers, groundwater systems also include low-permeability geological units called aquitards that impact the groundwater storage and movement (Borchardt et al., 2007). An aquitard commonly exists in overburden strata characterized by clay deposits or bedrock exhibiting few fissures or apertures. Despite remaining saturated, its extremely poor permeability prevents it from yielding sufficient water for practical use (Şen, 2014).

Aquifers are typically classified as unconfined or confined based on the presence or absence of an overlying aquitard. Unconfined aquifers are directly connected to the surface and experience GWLs that fluctuate freely in response to recharge and discharge. In contrast, confined aquifers are overlain by an aquitard that restricts vertical groundwater flow, causing water to be stored under pressure and resulting in a potentiometric surface that may differ from the physical depth of the aquifer (Fetter, 2018; Freeze & Cherry, 1979).

GWL is a measure of groundwater availability and provides key insights into aquifer system dynamics (Evans et al., 2020). While monitoring, short-term prediction, and forecasting of GWL are crucial for the sustainable water management (Daliakopoulos et al., 2005), the complexity of the hydrogeology and factors such as groundwater-surface water interactions introduce uncertainties that make GWL prediction a difficult task (Patra et al., 2023). Figure 2.1 illustrates the conceptual structure of aquifers and aquitards and the location where GWLs are measured.

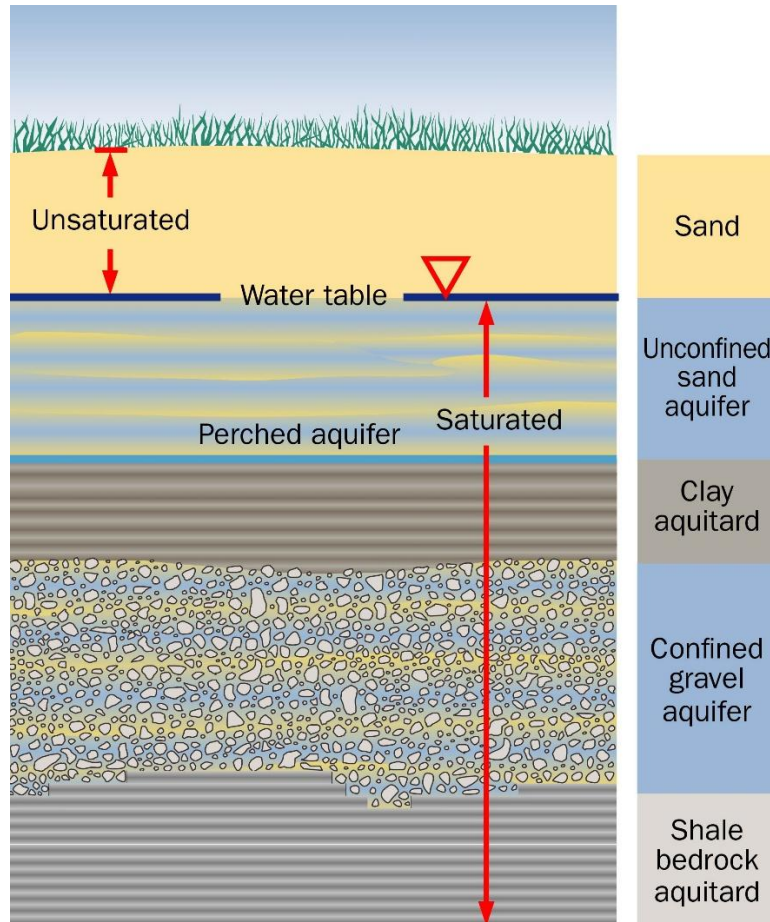


Figure 2-1-Conceptual groundwater system showing confined and unconfined aquifers and GWL as the water table (Source: Ontario Ministry of Agriculture, Food and Rural Affairs, 2021).

GWL is affected by numerous spatial and temporal factors, including groundwater extraction, climatic conditions, hydrometric parameters, precipitation, land use and land cover, terrain slope, and subsurface geology (Ifediegwu, 2021; Saroughi et al., 2023). In addition, groundwater systems are sensitive to human activities such as pumping and overextraction of water from aquifers (Chidepudi et al., 2025). High pumping rates that exceed the natural replenishment rate lead to depletion of GWL (Famiglietti, 2014). Given the impacts of climate change and the growing global population, groundwater reserves are increasingly being depleted in many countries around the world (Jia et al., 2019; Lall et al., 2020).

Given the importance of groundwater resources and the challenges associated with monitoring and managing them, especially under increasing anthropogenic and climatic pressures, accurate prediction of GWL has become a vital component of sustainable water resource management. To address this need, a

variety of numerical models have been developed over the past decades. These models, which rely on physical principles to simulate subsurface flow and recharge dynamics, are discussed in the following section.

2.2 Numerical Models for GWL Modelling

Numerical models integrate the physical principles of the problem and solve the corresponding equations within a defined domain. Mesh or grid system within the geometry are created where the equations – typically partial differential equations (PDEs) – are calculated at each discrete node.

As these models are grounded in the underlying physics of the system, they provide a comprehensive understanding of how GWLs fluctuate across various physics-based scenarios, which are inherently linked to the system's fundamental PDEs (Malekzadeh et al., 2019). To be able to solve the PDEs successfully, numerical models need a specified set of boundary and initial conditions; in the absence of these, the equations cannot be properly solved (Sahoo et al., 2017). The finite element method (FEM) and the finite difference method (FDM) are currently the most widely used numerical approaches for discretizing the domain and solving the PDEs at each node of the system in the field of GWL prediction (Anderson et al., 2015). In addition to these methods, finite volume (Jamin et al., 2020) and element-free (Pathania et al., 2019) approaches have also been applied in groundwater modelling, expanding the range of numerical techniques available for simulating groundwater systems.

Traditionally, GWL and other relevant water budget components such as soil moisture and runoff, are simulated using physics-based and mechanistic numerical software packages such as MODFLOW (Langevin et al., 2017), FEFLOW (J. Li et al., 2017), SUTRA (Mahmoodzadeh et al., 2014), ParFlow, HydroGeoSphere (Sahu et al., 2020a), and TOUGH (Steefel et al., 2015). ParFlow, HydroGeoSphere, MikeSHE, and SWAT are primarily designed to simulate large-scale land water cycles and are not specifically tailored for GWL modelling.

MODFLOW – originally called USGS Modular 3-Dimensional Finite-Difference Groundwater Flow Model – is a 3-dimensional model that uses FDM for groundwater flow simulation in heterogeneous and homogeneous aquifers (both confined and unconfined) under transient and steady-state conditions. It was first introduced by U.S. Geological Survey in 1984 – developed by Michael G. McDonald and Arlen W. Harbaugh – and modified over three years (McDonald & Harbaugh, 1988). Afterwards, it became the de facto standard in simulating groundwater conditions and groundwater-surface water interactions among hydrogeologists (McDonald et al., 2003). Shortly after, various modelling packages were processed and MODFLOW updates continued and released among users for more advanced and reliable groundwater modelling. For instance, particle-tracking programs like PATH3D, MODPATH, and MT3D were produced and assumed as one the most important modules since contaminant transport and particle advection scenarios could be incorporated with MODFLOW (McDonald et al., 2003). Using this software, several features such as waterbodies (rivers and lakes), wells, drainage areas, plants, rainfall events, and evapotranspiration rates could be added to the domain. First step in MODFLOW is to create a conceptual model (3-D geometry) of the domain. For this, one should know the aquifer's three-dimensional spatial extent (x-y-z geometry), digital elevation of the area (including bedrock, overburden, and surface layers), recharge rates, outflow and inflow boundaries, and any water extraction from operational wells in the area (Goodarzi et al., 2024).

The governing equation in MODFLOW is the combination of Darcy's Law (describing groundwater flow through a porous medium) and the continuity equation (for the conservation of mass in the system) for saturated porous media:

$$\frac{\partial}{\partial x} \left(K_x \frac{\partial h}{\partial x} \right) + \frac{\partial}{\partial y} \left(K_y \frac{\partial h}{\partial y} \right) + \frac{\partial}{\partial z} \left(K_z \frac{\partial h}{\partial z} \right) \pm W = S_s \frac{\partial h}{\partial t} \quad (1.1)$$

Where h represents the hydraulic head, K_x , K_y , and K_z are hydraulic conductivities in x , y , and z directions, W is the volumetric flux (e.g., pumping and recharge), S_s is the specific storage (how much water is released per unit change in hydraulic head per unit volume of the aquifer) and t is time.

In addition to the core flow equation, MODFLOW incorporates several supporting formulations to represent real-world boundary conditions, aquifer storage properties, and external fluxes. For example, specific storage and specific yield are used to simulate water retention and release in confined and unconfined aquifers, respectively. Boundary conditions such as General Head Boundaries (GHB), river interactions, and recharge zones are represented through package-specific equations, which allow the model to capture complex interactions between groundwater and surrounding hydrologic features. MODFLOW ensures mass conservation through a water balance equation applied at each time step. These equations, along with numerical solvers, form the computational backbone of MODFLOW and enable accurate simulation of transient and steady-state GWLs under diverse conditions (U.S. Geological Survey, 2005). With this foundation, numerous studies have applied MODFLOW to predict groundwater behavior across a wide range of geological and climatic contexts.

In their study, Radmanesh et al. (2020) modeled GWLs using MODFLOW within an unconfined alluvial aquifer located in a regional plain characterized by multiple inflow and outflow sources, applying the General Head Boundary (GHB) approach. The performance of MODFLOW was compared to two data-driven models: Gene Expression Programming (GEP) and the Adaptive Neuro-Fuzzy Inference System (ANFIS). MODFLOW exhibited superior performance, achieving a validation correlation coefficient (R) of 0.88 and an RMSE of 0.449. In comparison, GEP recorded an R of 0.86 and RMSE of 0.12, while ANFIS obtained an R of 0.80 and RMSE of 0.15. Despite MODFLOW's strength as a physically based and process-driven model, its higher RMSE during the validation phase indicates larger prediction errors compared to the data-driven GEP and ANFIS models, which showed better predictive accuracy.

Coelho et al. (2017) employed MODFLOW to evaluate three different numerical hydrological models based on varying boundary conditions (e.g., river, GHB, and stream) as field data did not provide a comprehensive understanding of the exact boundary conditions. They modelled GWL in an aquifer – located under one watershed – using Visual MODFLOW platform and calibrated the outputs using WinPEST. They mentioned that each boundary condition produced a model that was successfully calibrated

and validated with similar normalized RMSE quantities (ranging from 7.3% to 13.02%) and high correlation coefficients (94% to 97%). This means that each model was reliable in its own way, all models remained accurate across different boundary conditions, and they were all valid and reliable for simulating groundwater flow.

FEFLOW, another widely used numerical model for GWL simulations, uses finite element method (FEM) which is more versatile and can easily model irregular geometries and heterogeneous media compared to FDM implemented in MODFLOW which is simpler but less flexible for complex domains. It is particularly well-suited for modelling complex boundary conditions, irregular geometries, and heterogeneous aquifers. FEFLOW supports both saturated and unsaturated zone simulations and is widely used in environmental impact assessments, including those related to large-scale mining operations (Brown et al., 2018).

While MODFLOW and FEFLOW are specialized for simulating groundwater flow within aquifers, hydrological models such as SWAT, MIKE SHE, or HEC-HMS are broader in scope and primarily designed to simulate surface hydrological processes like rainfall-runoff generation, evapotranspiration, and river flow. Hydrological models often include simplified groundwater modules, but their main focus is the land surface and watershed-scale water balance. In contrast, MODFLOW and FEFLOW solve the governing equations of subsurface flow in detail and are used when the focus is on understanding groundwater dynamics, particularly in saturated zones. Integrated models like MIKE SHE and HydroGeoSphere attempt to couple surface and subsurface systems.

Traditional numerical groundwater models have served as a foundational tool for simulating GWLs across diverse hydrogeological contexts. However, their accuracy strongly depends on detailed input data such as aquifer properties, boundary conditions, and recharge rates, which are not always available or reliable in real-world scenarios (Zeydallinejad, 2022). With such a large number of parameters required for these models, the necessary data might be incomplete or uncertain, making it challenging to develop robust numerical models (Secci et al., 2023).

Physics-based models are still primarily applied to aquifers in small watersheds because extending them to large or regional domains is computationally demanding and typically requires coarser grid resolutions, which can reduce spatial detail and predictive reliability. While significant progress has been made in numerical modelling, developing large-scale (or regional scale) hydrological models remains challenging particularly when data are sparse or when future conditions must be projected. In such cases, data-driven tools have shown great potential as valuable models, either as standalone predictors of future GWL or as complementary approaches to physic-based models, especially in regions with limited hydrogeological data or where detailed parameterization is not feasible.

2.3 Machine Learning-based GWL Models

In recent years, the digital revolution, advancements in artificial intelligence (AI), and data analysis have significantly influenced various aspects of human life (Huang et al., 2021). AI has been employed in numerous fields such as energy (H. Li et al., 2025), manufacturing (D. Li et al., 2025), finance (Kou et al., 2024), healthcare (McCradden et al., 2025), and transportation (Gupta et al., 2025). To overcome the limitations of traditional numerical models in defining complex relationships between targets and influencing factors, AI techniques often provide more effective solutions for accurately predicting targets while reducing computational time and minimizing the need for extensive datasets (Uc-Castillo et al., 2023). Due to the accuracy and ease of implementation of AI models, numerous researchers have explored their performance across various domains, including GWL prediction studies across different regions worldwide (Rajaei et al., 2019). AI models comprise a very broad range of approaches used in different fields that are sometimes combined with each other for GWL prediction, such as Evolutionary Algorithms (EA), Fuzzy Logic, Deep Learning (DL), Reinforcement Learning (RL) and Symbolic AI.

ML models are a subset of AI algorithms that have proven to effectively capture patterns and relationships from complex data and to support prediction or decision-making based on those data (Asnicar

et al., 2024). ML was originally defined as programming computers to exhibit learning behavior similar to humans by improving performance through experience (Samuel, 1959). In the context of groundwater systems, these models can represent the complex interactions between environmental variables and GWL, enabling more accurate modelling and prediction (Tao et al., 2022). ML models can address complex problems involving numerous influencing parameters by transforming a physics-based system into a mathematical framework. The goal is to minimize model error, ensuring the results closely align with actual observations (Huang et al., 2021). ML models have also been recently used as surrogate models and as universal function approximators. In this role, they can learn complex nonlinear relationships directly from data or from outputs generated by numerical models, enabling efficient approximation of computationally intensive simulations (Fernandes et al., 2024a).

Due to non-stationary behavior of GWL data, implementing traditional time series forecasting methods such as Autoregressive Integrated Moving Average (ARIMA) and regressions is not suitable (Müller et al., 2021). Regression models typically rely on multiple assumptions which makes them sensitive to the behaviour of their residuals. Residual or the error of the regression models is the difference between the prediction and actual target quantities. It should have specific requirements otherwise the assumptions for regression are not fulfilled and the model may not generalize well on the unseen data even if its performance appears strong for a specific scenario. In the context of GWL prediction, linear regression have been used widely, but its assumptions are not capable of handling nonlinear relationships between inputs such as hydrometric level, temperature, precipitation, and human-induced factors and the GWL data (Poursaeid et al., 2022).

Artificial Neural Network (ANN) is a type of ML model which is inspired by how biological elements in nature are connected to each other. It uses advanced operators which overcome the limitations of linear regression and predicts GWL with improved accuracy (Ehteram & Banadkooki, 2023). ANNs can be utilized to predict future values of a time series by leveraging its historical data and using back propagation (BP) training techniques (Ebrahimi & Rajaei, 2017; Foroumandi et al., 2021). These models

consist of multiple layers – the input, hidden, and output layers – and are inspired by biological neural networks, where connections between elements are structured as networks with components operating in parallel (Beale et al., 2018). The effectiveness of such networks has drawn the attention of researchers, especially since the introduction of BP methods in the mid-1980s by Rumelhart et al. (1986) although implementing hidden layers in the network was understood to be powerful by diminishing their incapability in learning nonlinearly separable problems. This technique optimizes model parameters (weights and bias) by minimizing the error which results in a better performance (Hsieh, 2023; Schmidhuber, 2015).

In practical applications of GWL prediction, many studies have focused on developing different ANN models and comparing their performance using different evaluation criteria. Since these models are inherently sensitive to data patterns and influencing factors, it is not possible to claim that one model will always outperform others.

Roshni et al. (2019) utilized multiple wells within a complex alluvial aquifer system – all located in the same watershed – and determined the optimal number of weekly observations for their models using the novel Gamma Test (GT) and M-tests. They simultaneously predicted GWL fluctuations using ANNs and Wavelet-ANNs (WANNs), demonstrating the superior performance of WANNs for 1-week and 3-week ahead forecasts. However, the study lacks comprehensive details on the modelling hyperparameters, and the approach they used to find the parameters of their ANNs. In another study, Secci et al. (2023) examined the impact of climate change by predicting GWLs at 10 wells using Non-Linear Autoregressive Neural Network (NARX), LSTM, and CNN models. Wells were in four basins in Northern Italy (Magra, Coastal, Serchio, and Arno), belonging to different groundwater bodies, mainly characterized as phreatic (unconfined) aquifers, although local confining layers are present in some areas. The input data consisted of monthly historical temperature and precipitation from 18 monitoring stations, along with previous GWLs. The models were calibrated and tested using data from 2005 to 2020 before being applied for long-term projections. LSTM ($MSE_{test} = 0.3$) outperformed, NARX ($MSE_{test} = 0.82$) indicated reduced performance in validation and testing, and CNN ($MSE_{test} = 0.31$) struggled to generalize beyond the training

data. They suggest that climate change will likely cause GWL depletion, with more severe effects under high-emission scenarios (RCP8.5). LSTMs were the most reliable models to capture the long-term trend of GWL until the year 2100 due to their ability to process sequential dependencies. However, they assumed that all wells were facing the same hydrological conditions although they were drilled in different basins.

Beyond comparing individual ANN models, some studies have explored alternative modelling strategies to enhance prediction performance. ML models are sometimes stacked to assess whether their combined performance enhances results. In this case, the output from one ML model could be used as input to the next subsequent model, and this is continued until the last model gives the predicted GWL. This approach does not always yield better performance, though, since ML techniques highly rely on site-specific parameters and the nature of the dataset (Afshari Nia et al., 2023).

In addition to model structure, the selection and representation of input variables also play a critical role in GWL prediction performance. Standardized indices, such as the standardized precipitation index (SPI), standardized precipitation-evapotranspiration index (SPEI), and standardized groundwater index (SGI), are commonly used to represent the key variables in GWL studies (Secci et al., 2023).

Feedforward neural networks (FFNNs), described in detail in Section 4.4.4, are ANN architectures in which information flows unidirectionally from the input layer to the output layer. FFNNs are commonly used to predict the next time step of the target variable based on a fixed number of preceding observations (Sharghi et al., 2022). For example, Lee et al. (2019) applied a FFNN model to predict hourly GWL in eight observation wells of Yangpyeong riverside area, South Korea and subsequently conducted a sensitivity analysis to evaluate the influence of the input variables. In another study, Kouziokas et al. (2018) developed multiple FFNN configurations to forecast daily GWL in Pennsylvania, USA. Their results indicated that the model trained using the Levenberg-Marquardt (LM) algorithm achieved superior performance.

Overall, the reviewed studies demonstrate that the performance of ML-based GWL models is strongly influenced by model structure, input configuration, training strategy, and site-specific

hydrogeological conditions. No single ANN architecture consistently outperforms others across different aquifer systems. These findings highlight the importance of carefully selecting both modelling framework and spatial configuration when developing predictive models for GWL forecasting.

2.4 Multi-output and Single Output Models for GWL Prediction

Several studies have reported that incorporating all available input variables does not necessarily improve model performance and may even decrease predictive accuracy. For example, Lee et al. (2019) observed that precipitation had a limited influence on GWL fluctuations and therefore excluded the parameter from the modelling process. In another study, Chang et al. (2015) developed two FFNN models: one that incorporated temperature, precipitation, and previous GWL values as inputs, and another that used only temperature and precipitation. Their results demonstrated that the model including all three input variables achieved better predictive performance. Accordingly, a good physical grasp of the system has been emphasized in the literature to help guarantee that only pertinent features are selected (Shen, 2017). As a result, previous studies have often identified influential input variables by examining findings from similar hydrogeological settings and modelling frameworks (Rajaei et al., 2019).

In addition to the input selection, several studies have examined whether predicting multiple GWLs from different wells at the same time would enhance model performance (Chidepudi et al., 2025; Patra et al., 2023; Vu et al., 2021). Hsieh (2023), suggests that when target variables are interrelated, utilizing a single ML model to predict multiple targets simultaneously often results in improved performance. This is because the model can capture a wider range of patterns from diverse outputs while optimizing weight and bias parameters during each epoch (Zhang & Yang, 2022). This technique is known as multi-output modelling. On the other hand, if the target variables are uncorrelated, it is more effective to develop separate ML models for each target. This single-output modelling approach minimizes potential interference from unrelated data sources, enhancing model accuracy (Caruana, 1997).

In GWL modelling, the “global model” designation is typically applied to multi-output models that are calibrated with data from numerous wells or monitoring stations. Chidepudi et al. (2025) contend that this designation can be misleading, implying a broader applicability than is intended. They recommend utilizing the multi-station or multi-output approach to more clearly characterize this process. Multi-output approaches have been utilized in certain GWL prediction studies, but their application has primarily been limited to using data from nearby wells to forecast multiple GWLs simultaneously (Chidepudi et al., 2025). Furthermore, current research primarily concentrates on wells located within the same unconfined aquifer and watershed or within unconfined aquifers that extend across multiple watersheds. In contrast, deep and complex aquifer systems remain relatively underexplored, despite their crucial role in supporting various water uses. This is particularly significant in rural areas, where communities heavily rely on these deeper aquifers for drinking water and irrigation purposes.

Recent research has compared multi-station models with single-station models, but without clustering GWLs based on hydrogeological units like watersheds. Before applying advanced clustering techniques, it would be logical to explore physics-based clustering approaches, such as grouping wells within the same watershed. This would allow for the development of multi-output models for wells located within the same watershed boundary, as they are influenced by similar hydrological and environmental factors like precipitation, temperature, and hydrometric levels or discharge data. This would be particularly pertinent for unconfined aquifers.

Within this framework, researchers have compared single-output and multi-output modelling approaches to evaluate whether predicting several wells simultaneously improves performance. For example, Müller et al. (2021) implemented multi-well scenarios using various hyperparameter optimization (HPO) techniques and ML models like MLP, CNN, and LSTM within a single watershed and an unconfined aquifer. Their findings revealed that single-output ANNs outperformed multi-output models with MLPs as the best models and RNNs as the worst models as they failed to fit properly during backpropagation. The mean RMSE for single-output models was 0.12 for MLP, 0.16 for CNN, and 0.19 for LSTM. In comparison,

the multi-output models recorded slightly higher RMSE values of 0.15 for MLP, 0.18 for CNN, and 0.21 for LSTM. This indicates that single-output models yielded better predictive accuracy, likely due to their ability to focus on the unique characteristics of individual wells. Additionally, the MLP models optimized using the Radial Basis Function (RBF) – a hyperparameter fine-tuning surrogate method – achieved the best performance, consistently producing lower errors compared to other single-output MLPs. This optimization approach effectively enhanced model accuracy by refining hyperparameters, ensuring better convergence and more reliable predictions.

Similar conclusions were reported by Chidepudi et al. (2025), who found that single-well models exhibit better performance compared to global models using deep learning models such as LSTM, biLSTM, and GRU models. In addition, they stated that hybridizing multi-station models with wavelet-based spectral clustering using Boundary-Corrected Maximal Overlap Discrete Wavelet Transform (BC-MODWT) methods can significantly enhance their performance, still not better than the single model. It is important to highlight that the wells analyzed in this study are situated in different watersheds, including coastal regions. Multi-station models trained after clustering performed better specially for low-frequency dominated GWLs. It was also mentioned that multi-station models without clustering should be used cautiously as models may adopt dominant behavior which is not always a model-enhancing point.

Patra et al. (2023) further compared single-output and multi-output LSTM models in coastal regions characterized by complex groundwater dynamics. Their results indicated that single-output LSTMs consistently outperformed multi-output LSTMs. To improve multi-output performance, they introduced Transfer Learning (TL), allowing global models to initialize weights from nearby wells rather than random initialization. Although this enhanced convergence and predictive stability, multi-output models still did not surpass single-output models. They also compared two multi-output structures: a Multi-Sequence (MS) LSTM that used separate input nodes for each well, and a Single-Sequence (SS) LSTM that stacked all GWL data into one sequence. Results showed that the SS-LSTM performed significantly better than the MS model. The researchers suggested that the MS input structure introduced complexity and confusion,

which could impair the model's ability to effectively learn distinct patterns. For instance, the single-output LSTM model in coastal regions achieved an RMSE of 0.232, whereas the multi-output MS-LSTM and SS-LSTM models exhibited higher RMSE values.

Beyond direct single-versus-multi comparisons, clustering techniques are frequently applied prior to modelling. However, most clustering approaches are data-driven and rely on statistical similarity measures rather than physical or hydrogeological criteria. For instance, Sharghi et al. (2022) grouped 49 wells within the same watershed using clustering techniques like K-mean, Fuzzy C-Mean (FCM), and Growing Neural Gas (GNG). They then applied FFNN, ANFIS, LSTM, and GMDH models to predict multi-step centroid GWL values. GMDH achieved superior performance (testing RMSE = 0.42, $R^2 = 0.84$ for one-step-ahead forecasting), and performance further improved when outputs were combined using a Neural Averaging Ensemble (NAE) (testing RMSE = 0.38, $R^2 = 0.90$). However, LSTM underperformed, likely due to limited data availability.

Kardan Moghaddam et al. (2021) similarly applied K-Means clustering to select representative wells in an unconfined aquifer and evaluated FFNN, RNN, GMDH, and Bayesian Network models using different combinations of hydrological input variables. Their results showed that GMDH model outperformed the other models in the study, achieving up to 6% higher R^2 and 16% lower RMSE when predicting the GWLs at each well individually. These findings reinforce the importance of model structure and input configuration when clustering is used.

Other studies have used clustering primarily for categorization or long-term trend assessment rather than direct predictive comparison. Yin et al. (2022) clustered 104 wells within the San Joaquin River Basin, California, which includes unconfined, semi-confined (SC), and confined aquifers. They applied agglomerative hierarchical clustering using seasonal GWL data as inputs, with a complete linkage method that grouped wells based on the similarity of their GWL trends. The clustering process relied on calculating the Euclidean distance between clusters to determine similarity, resulting in five distinct well groups that reflected long-term GWL behavior. The study focused on trend analysis and drought categorization using

SGI rather than predictive modelling. Similarly, Jeihouni et al. (2019) grouped wells within the same watershed and applied LSSVM, ANFIS, NARX, and Wavelet-NARX models for long-term climate-based projections. Wavelet-NARX achieved the highest accuracy (R^2 up to 0.99; RMSE as low as 0.02), and projections suggested accelerated GWL depletion under severe climate scenarios (RCP 8.5). Wunsch et al. (2022) used Self-Organizing Maps (SOM) to cluster wells prior to CNN-based long-term projections, also reporting stronger GWL declines under high-emission scenarios.

Overall, although several studies have compared single-output and multi-output models or applied clustering prior to modelling, most grouping strategies are based on hydrograph similarity or statistical measures. Explicit comparisons using watershed-based clustering or physically defined hydrogeological units remain limited. This gap motivates the present study, which investigates whether grouping wells based on watershed boundaries leads to improved modelling performance compared to purely data-driven clustering approaches.

Table 2.1 -Summary of recent studies on GWL prediction using ML and clustering-based methods, highlighting model types, aquifer types, watershed scale, input variables, temporal resolution, and forecast horizons.

Study	Models	Aquifer	Watershed	Inputs	Resolution	Forecast
Müller et al. (2021)	MLP, CNN, LSTM	C*	M	T, P, D, G	Daily	G_{t+1}
Chidepudi et al. (2025)	GRU, LSTM, BiLSTM	C, UC	M	$T_x, P_x, EP_x, G_x - x:[t-1, t-7]$	Daily	G_{t+1}
Patra et al., (2023)	LSTM	C, UC	M	$G_{t-1}, G_{t-2}, G_{t-3}$	Daily	$G_{t+1}, G_{t+2}, G_{t+3}$
Sharghi et al., (2022)	FFNN, ANFIS, GMDH, LSTM, NAE	UC	S	R, RU, G	Monthly	$G_{t+1}, G_{t+2}, G_{t+3}$
Kardan Moghaddam et al. (2021)	FFNN, RNN, GMDH, BN	UC	N.A.	G_{t-1}, D, R, E, T, P (varied per scenario)	Monthly	G_{t+1}
Yin et al. (2022)	N.A. (Clustering Study)	UC, SC, CC	S	G	Seasonally	N.A.
Jeihouni et al. (2019)	LSSVM, ANFIS, NARX, Wavelet-NARX	UC	S	Lagged P, T, G	Monthly	G_{t+1}
Wunsch et al. (2022)	CNN	UC	M	P, T, G	Weekly	G_{t+1}

***Abbreviations:** C: Confined aquifer; UC: Unconfined aquifer; SC: Semi-confined aquifer; CC: Confined coastal aquifer; M: Multiple watersheds; S: Single watershed; N.A.: Not available; T: Temperature; P: Precipitation; D: Discharge; R: Recharge; RU: Runoff; E: Evaporation; G: GWL; T_x, P_x, EP_x : Temperature, precipitation, and evapotranspiration at time step x; G_{t-1} : GWL at previous time step; G_{t+1} : One-step-ahead forecast.

2.5 Research Gaps and Contributions of the Study

To the best of current knowledge, no previous study has systematically compared single, combined, and aggregated modelling approaches within a unified framework for GWL prediction. While prior research has explored single-output and multi-output models individually, no existing work has systematically evaluated all three scenarios – single models, combined models (with clustering based on watershed boundaries), and aggregated models (which group wells across different watersheds into a single model). In particular, the role of watershed-based spatial grouping in influencing predictive performance remains insufficiently explored.

To address this gap, wells are categorized based on watershed boundaries rather than hydrograph similarity or purely data-driven clustering techniques. This structure enables evaluation of whether physically defined grouping improves model accuracy and reliability when predicting GWL across multiple watersheds.

Conducting this comparison requires a modelling framework that can be consistently applied across all configurations. Physically based groundwater models, such as MODFLOW, require extensive hydrogeological and environmental inputs that are not always available. Therefore, a data-driven modelling framework is adopted, relying on mathematical relationships between input variables and GWL responses to ensure comparability across scenarios.

Before developing predictive models, a statistical analysis of GWL trends across Southern Ontario was conducted. Specifically, the Mann-Kendall trend test was applied to all available GWL time series from the PGMN dataset for the period 2000-2023. This allowed us to identify the presence and strength of trends at different monitoring wells throughout the region.

Following this initial analysis, modelling process was initiated using simple regression models to predict GWLs within specific regional boundaries (further detailed in Section 3.1.2). These models were selected for their straightforward implementation, low computational requirements, and high interpretability. However, their performance is strongly influenced by underlying assumptions, which, if violated, can compromise the model's generalizability on unseen data.

To overcome these limitations, neural network models were introduced, which impose fewer assumptions and are capable of learning complex, nonlinear relationships. Despite being more computationally demanding and susceptible to variability due to their stochastic nature, this issue was addressed using the method described in Section 4.4.5.5 to stabilize model outputs. Additionally, a high-performance local server with 144 GB of RAM at York University was used to support parallel model training and reduce runtime. Although similar data-driven approaches have been applied in previous

studies, many researchers adopt specific models without explaining why simpler alternatives may fall short. In this study, we provided a clear justification for moving beyond basic models.

During the neural network-based predictive modelling process, we applied several preprocessing techniques to address issues related to noisy inputs and outliers, aiming to enhance data interpretability and improve model reliability and accuracy. One of the key methods used was (SSA). We evaluated the effectiveness of SSA at two different stages: before aggregating daily data into monthly averages (referred to as the Daily-SSA-Monthly or DSM models) and after the data had already been aggregated (referred to as the Daily-Monthly-SSA or DMS models). In both the DSM and DMS setups, the performance of single, combined, and aggregated modelling approaches was further compared to examine the consistency of the findings across different preprocessing strategies. This is the first known comparison of its kind, not only in groundwater modelling but also across the broader field of hydrological modelling.

To clearly highlight the novelty and scope of this work, the following bullet points summarize the identified research gaps and the corresponding contributions made by this thesis:

- The study conducts a comprehensive Mann-Kendall trend analysis of GWL time series across Southern Ontario (2000-2023), addressing a major gap in regional-scale statistical evaluations of GWL behavior.
- This is the first study to systematically compare single, combined, and aggregated modelling approaches for GWL prediction within a unified data-driven framework.
- It introduces a novel distinction between combined models (grouping wells within the same watershed) and aggregated models (grouping wells across different watersheds), a separation rarely addressed in the literature.
- The thesis proposes watershed-based well grouping as a physically meaningful clustering method, offering an alternative to purely data-driven clustering.
- It provides a clear rationale for progressing from simple regression models to neural networks, highlighting when and why simpler models may fail in GWL forecasting.

- It evaluates the impact of SSA applied at different stages of preprocessing (before vs. after data aggregation), a comparison not found in prior GWL studies.
- It offers the first known comparison of DSM and DMS pipelines across all three modelling scenarios (single, combined, aggregated), filling a major methodological gap in hydrological modelling research.

Chapter 3 Study Area and Data

This chapter provides a detailed overview of the study area, and the datasets used in this research. Section 3.1 describes the geographical scope of the study, beginning with Southern Ontario and continuing through key transition zones and boundaries relevant to Barrie-Oro Moraine. Section 3.2 outlines the

various data sources, including GWLs, meteorologic data, hydrometric records, and geographic features such as watersheds and aquifers. Each data source is introduced along with its spatial and temporal resolution, ensuring a clear understanding of the information underpinning the analysis.

3.1 Study Area

Southern Ontario was selected as the initial focus of the statistical analyses to gain a broader understanding of regional GWL patterns across a densely populated and hydrologically significant area. The statistical component involved applying the Mann–Kendall test to GWL time series across the entire Southern Ontario region to evaluate long-term trends.

Following this regional assessment, the study transitioned to a smaller subregion – the Barrie-Oro Moraine (BOM) – for the development of data-driven ML models for GWL prediction. This BOM was chosen due to its high reliance on groundwater for municipal and agricultural use, making it particularly relevant for groundwater resource assessment. While the subsequent data-driven models cover the GWL prediction using ML approaches in addition, the availability of consistent meteorological and hydrometric data in this region supported the development of more reliable models. While the BOM served as a case study, the same modelling framework can be applied to other subregions, provided that sufficient data are available.

3.1.1 Southern Ontario

Southern Ontario is the southernmost part of the province of Ontario, Canada. It is bounded by the Great Lakes to the south and west (Lake Huron, Lake Erie, and Lake Ontario), northern Ontario to the north, and the province of Quebec to the east spanning from 74.3°W to 82.9°W longitude and 41.6°N to 45.0°N latitude as shown on Figure 3.1 (Statistics Canada, 2016). According to Government of Canada

(2021), Ontario is the most populous province in Canada, housing 14.22 million people, which accounts for 38.5% of the national population, with a density of 15.9 people per square kilometer. FedDev Ontario Pocketbook (2022) reports that Southern Ontario is particularly dense, accommodating 13.5 million residents, resulting in a population density of 118 people per square kilometer. This region contains several important cities such as Toronto (the capital of Ontario and the largest city in Canada), Ottawa (the capital of Canada), Hamilton, London, and Windsor with the core region being situated south of Algonquin Park. Southern Ontario's area is about 114, 217 km² which approximately covers 13% of the area of the province depending on the inclusion of Parry Sound and Muskoka districts which lie in the area between northern and southern forest regions. Southern Ontario forms a distinct hydrogeological region characterized by Paleozoic sedimentary bedrock overlain by thick unconsolidated glacial deposits, which locally reach up to 200 m in thickness. The bedrock consists mainly of Cambrian to Devonian marine sedimentary rocks preserved within regional sedimentary basins, and variations in bedrock topography and sediment thickness strongly influence groundwater flow patterns and aquifer distribution across the region. Important aquifer systems occur within both sediment and bedrock units, including thick sediment aquifers in stratified moraine deposits and contact zone aquifers located at the interface between bedrock and overlying sediments. As shown in Figure 3.1, the region is bordered by the Great Lakes and contains abundant surface water resources; however, groundwater remains a critical resource for both agricultural and drinking purposes. According to Sharpe et al. (2014), approximately 25% of residents rely on groundwater for drinking water, with dependence increasing to about 90% in rural areas, and around 200 municipalities depend primarily on groundwater as their main water source. Sharpe et al. (2014) also report that long-term mean groundwater recharge ranges between approximately 83 and 285 mm per year depending on soil and slope conditions. Although groundwater quality is generally good across much of the region, shallow aquifers remain vulnerable to contamination from agricultural activities, waste disposal, and road salt application.

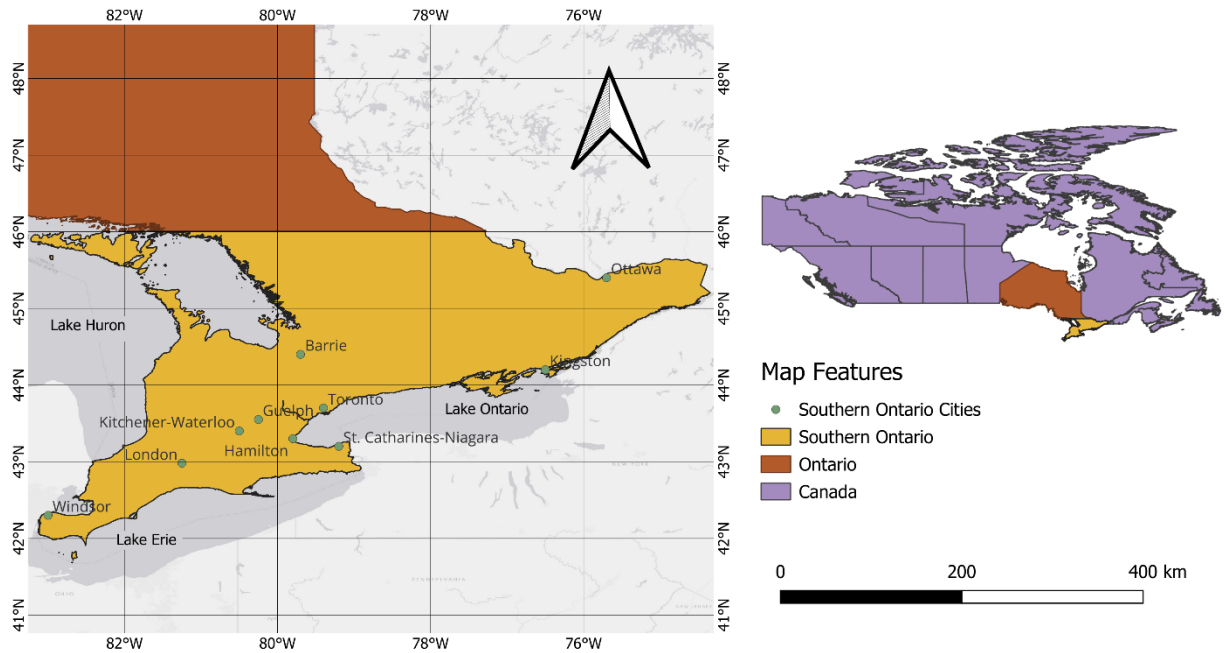


Figure 3-1-Southern Ontario boundary and its surrounding lakes and lands

3.1.2 Transition from Southern Ontario to Barrie-Oro Boundary

The Southern Ontario region was examined as part of the initial data analysis, focusing on GWL time series across the broader region. The study then narrowed its scope to a smaller area within Southern Ontario, known as the Barrie-Oro Moraine (BOM) region, to develop and apply a data-driven GWL modelling framework. This area was selected due to the availability of high-resolution, three-dimensional surficial aquifer data provided by the Ontario Geological Survey (OGS), which offers publicly accessible aquifer distribution data for five key regions in Southern Ontario.

Among these five OGS three-dimensional modelling areas (Figure 3.2), the northernmost region – corresponding to the BOM – was chosen for detailed modelling. The three-dimensional hydrostratigraphic framework allows for improved representation of aquifer geometry, layer continuity, and subsurface heterogeneity, which are essential for understanding groundwater flow systems and supporting reliable groundwater modelling (Burt & Dodge, 2011; Sharpe et al., 2014).

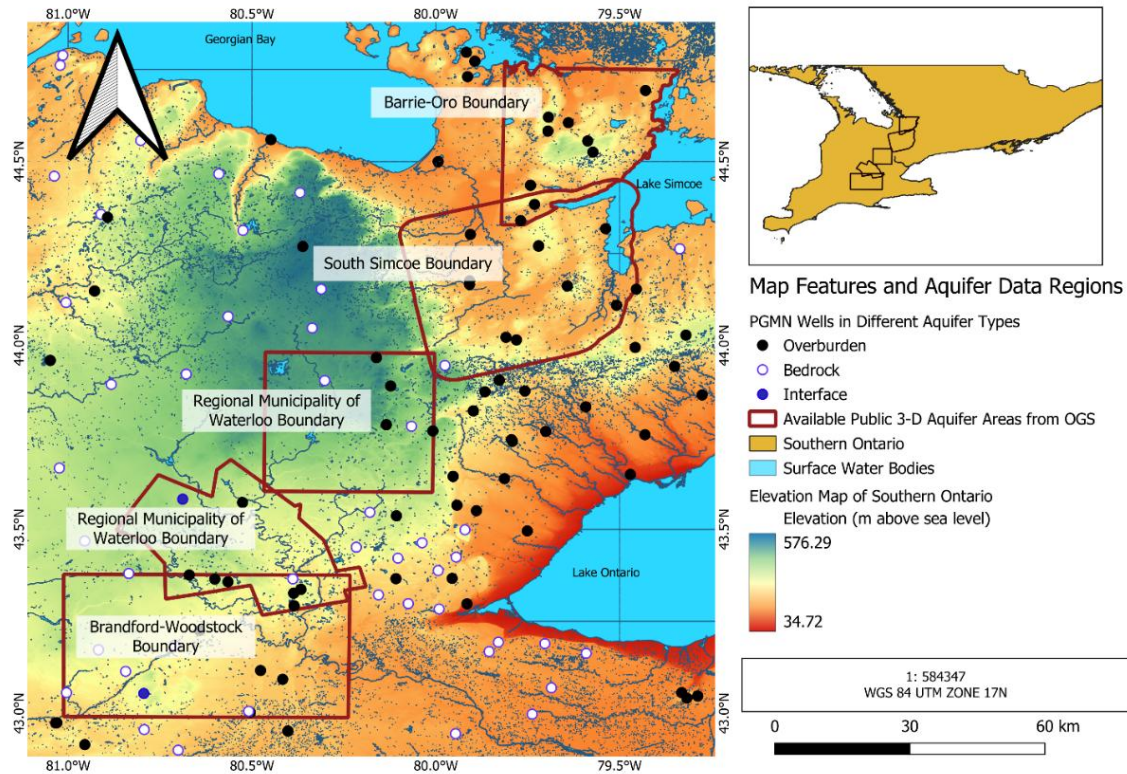


Figure 3-2-Map of Southern Ontario showing the five publicly available 3D surficial aquifer data regions provided by the Ontario Geological Survey (OGS), along with locations of PGMN wells classified by aquifer type.

3.1.3 Barrie-Oro Moraine Boundary

The area selected for GWL modelling covers approximately 1200 km² within Simcoe County, Ontario, Canada, as shown in Figure 3.3. It is located between Georgian Bay and Lake Simcoe, with the Oro Moraine at its center (Burt & Dodge, 2011). This area was formed during the last glacial period and is mainly made up of sand and gravel that serve as aquifers, along with clay and silt that act as aquitards. It is considered a kame moraine, where meltwater deposited highly permeable materials like sand and gravel in ice-walled basins (Chapman & Putnam, 1984). The BOM consists of multiple stacked and laterally extensive aquifers varying from unconfined to semi-confined and confined layers where the groundwater flows radially outward from the crest of moraine toward Lake Simcoe and Lake Couchiching as the discharge areas (Burt & Dodge, 2011).

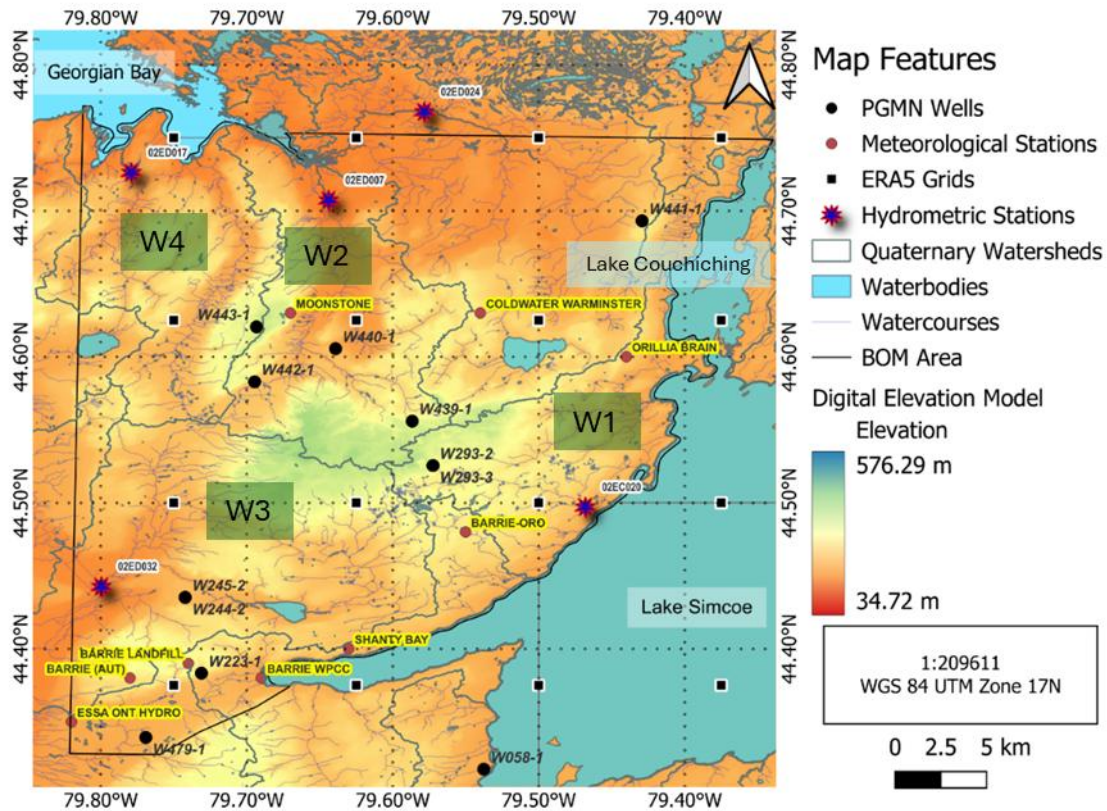


Figure 3-3- Map of the study area showing the locations of PGMN wells, meteorological stations, ERA5 grids, and hydrometric stations. Additional features include quaternary watersheds, watercourses, waterbodies, and the boundary of the BOM study area, overlaid on a digital elevation model.

The study area boundaries follow the delineation provided by Burt & Dodge (2011) and Cuddy (2013), which are defined by a combination of natural and mapped geological limits. To the east and southeast, Lakes Couchiching and Simcoe form clear natural borders. The western edge, however, is defined arbitrarily along the margin of a large clay basin, marked by a straight vertical line. As one moves north, the soil cover becomes thinner and gives way to exposed bedrock, indicating a major shift in the area's geology (Burt & Dodge, 2011). The area extends from the City of Barrie in the south, to Orillia in the north, and is covered by several meteorological and hydrometric stations as well as GWL monitoring wells distributed across its various watersheds (W1 through W4), as illustrated in Figure 3.3. The BOM

region is notable for its dependence on groundwater, which supports both municipal supply and agricultural activities, particularly in the municipalities of Barrie, Oro-Medonte, and Springwater (Burt & Dodge, 2011).

In Barrie, a significant portion of the water comes from groundwater with 14 municipal wells drilled in overburden aquifers. As of 2003, the groundwater accounted for 44% of the city's total water supply capacity with 10.2 million m³/year. Considering the population growth and urban expansion, the total water demand was projected to rise by 65% by 2031, potentially increasing the stress on these aquifers unless managed sustainably (City of Barrie, 2008). The Township of Oro-Medonte relies entirely on groundwater to supply its 13 municipal residential drinking water systems – including the Hillsdale, Horseshoe, and Craighurst systems – all of which draw water from overburden aquifers. The total average annual water demand across all systems is approximately 1.08 million m³, with Hillsdale and Horseshoe systems being the top consumers, accounting for 30.3% and 23.9% of the total demand, respectively (Township of Oro-Medonte, 2008). These figures highlight the critical role groundwater plays in maintaining water security in the BOM region.

The OGS selected the BOM as one of the five regions for which three-dimensional surficial aquifer data were developed to support land-use planning and groundwater management, indicating the importance of this area for groundwater flow and availability modelling (Burt & Dodge, 2011). As a result, this area was selected for the development of data-driven GWL models, allowing for detailed analysis of groundwater dynamics in a region where aquifer structure and availability are well characterized.

3.2 Data Sources

3.2.1 GWL Data

Various datasets offer groundwater-related information in Ontario. The Oil, Gas, and Salt Resources (OGSR) dataset, for instance, provides paid access to data on Paleozoic geology, natural gas storage reservoirs, subsurface hydrocarbon storage, solution mining of salt, and oil and gas exploration data (Oil, Gas and Salt Resources Library, 2020). However, most sources in Ontario recommend using the PGMN

dataset for GWL trend analysis. This recommendation is due to its comprehensive regional coverage, detailed metadata, consistent recording and updating practices with minimal missing values, and credibility, as it is managed by the Ontario Ministry of Environment. As a result, the GWL was downloaded and accessed from the PGMN program. This program, initiated in 2000, aims to monitor ambient GWLs and chemistry throughout Ontario. It operates through a partnership between the Ministry of Environment, Conservation, and Parks, all 36 Conservation Authorities in Ontario, and the Severn Environmental Association. According to the PGMN dataset, there are 522 monitoring wells located in Ontario where 480 of them are actively operating and 391 of them reporting the data publicly within the southern Ontario region on an hourly basis. The aquifer types monitored by a PGMN well are categorized as overburden, bedrock, or the interface between overburden and bedrock (Government of Ontario, 2019). As depicted in Figure 3.4, this study analyzed 391 wells. Of these, 219 wells (56%) were drilled into overburden aquifers, 156 wells (~40%) into bedrock, and 16 wells (~4%) at the interface between these layers. Additionally, well observations were converted to daily data to ensure compatibility with other datasets used in the analysis.

The PGMN dataset also includes vital information such as the ground elevation at the well site, well depth, and the lithology of the surrounding aquifer. Furthermore, 66 of the wells are equipped with rain gauges to record precipitation every 15 minutes from early spring to late fall. Although these precipitation data are available within the PGMN program, they were not utilized in this study, as the analysis focused exclusively on GWL time series.

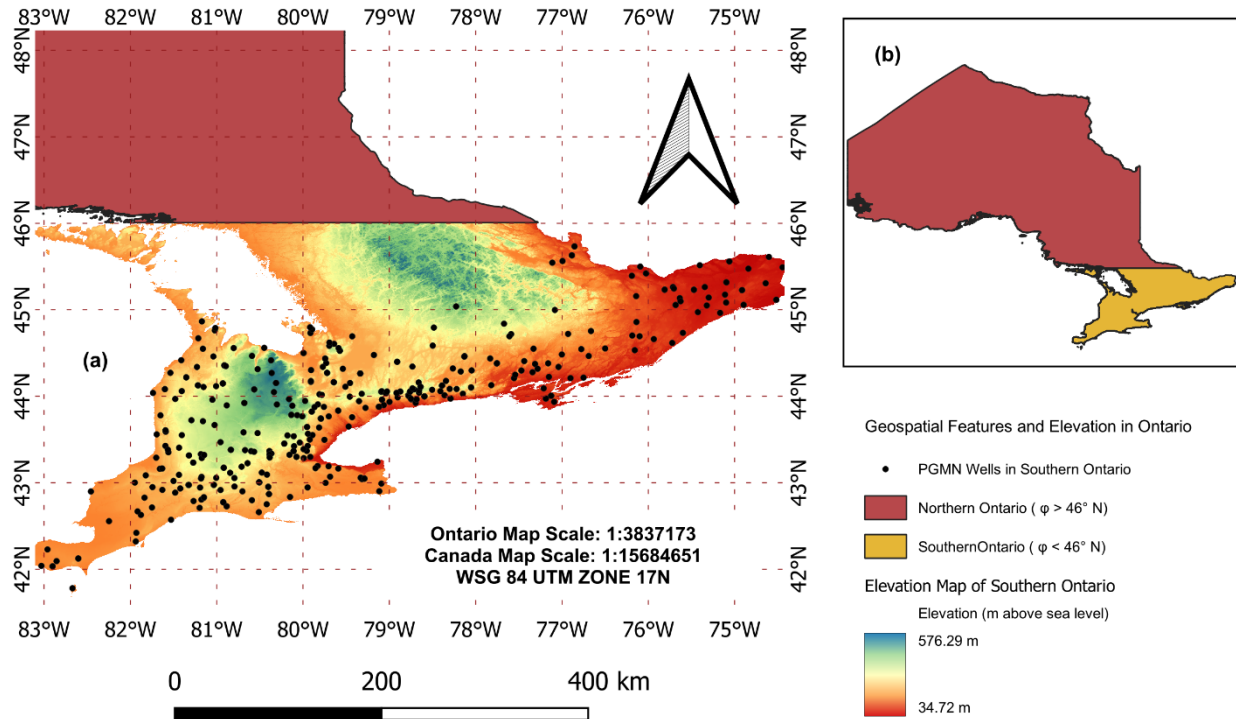


Figure 3-4-(a) Elevation map of Southern Ontario displaying the distribution of PGMN wells. Elevation is represented in meters above sea level. (b) Regional division of Ontario into Northern and Southern zones based on latitude (46°N).

In the BOM area, eight wells were selected (as referenced in Section 3.2.5), despite there being a total of eleven wells within the BOM boundary. The selection was primarily based on data availability and hydrostratigraphic characterization. Specifically, these eight wells were previously incorporated into the Tier 2 Water Budget Analysis conducted for the Oro Moraine (Cuddy, 2013), where they were used for groundwater flow model calibration and comparison between simulated and observed heads. Their inclusion in that study indicates that sufficient hydrostratigraphic information, reliable GWL records, and clear aquifer assignments were available for these wells. In contrast, the remaining three wells lacked either detailed aquifer characterization, consistent long-term monitoring records, or compatibility with the established hydrostratigraphic model framework, and were therefore excluded from the present modelling

analysis. For clarity, the wells were grouped into four watershed units defined in this study. Wells 293-2 and 293-3 are located in Watershed 1, Wells 439-1, 440-1, and 442-1 in Watershed 2, Wells 244-2 and 245-2 in Watershed 3, and Well 443-1 in Watershed 4.

Wells 293-2 and 293-3 are situated within the same aquifer and share identical geographic coordinates, classifying them as clustered wells. Well 293-2 monitors a shallower depth, while well 293-3 is screened at a greater depth. Due to the greater overburden weight at well 293-3, its GWLs are typically higher than those at well 293-2. This is because GWLs are recorded using pressure dataloggers, which measure pressure and convert it into water level relative to mean sea level. Table 3.1 summarizes the geological and administrative characteristics of the selected PGMN wells located within the BOM. It includes each well's conservation authority, township, aquifer composition, and the record number where applicable.

Table 3.1-Geological and administrative characteristics of selected PGMN wells within the BOM area. The table includes information on the associated conservation authority, township location, aquifer composition, and record numbers for wells equipped with pressure dataloggers.

PGMN Well	Conservation Authority	Township	Composition	Record Number
W223-1	Nottawasaga Valley C.A.	Vespra	Sand	N.A.
W244-2	Nottawasaga Valley C.A.	Vespra	Sandy silt	N.A.
W245-2	Nottawasaga Valley C.A.	Vespra	Sand	N.A.
W293-2	Lake Simcoe Region C.A.	Oro	Silt, silty sand	5737770
W293-3	Lake Simcoe Region C.A.	Oro	Sandy silt, sand	5737769
W439-1	Severn Sound Environmental Association	Medonte	Sand and Gravel	N.A.
W440-1	Severn Sound Environmental Association	Medonte	Sandstone	N.A.
W441-1	Severn Sound Environmental Association	Severn	Sand and Gravel	N.A.
W442-1	Severn Sound Environmental Association	Medonte	Medium to coarse light brown to yellow Sand	N.A.
W443-1	Severn Sound Environmental Association	Medonte	Clay, Sand and Gravel	N.A.
W479-1	Nottawasaga Valley C.A.	Essa	N.A.	5740549

To complement this information, Table 3.2 presents the physical attributes and geographic coordinates of the same wells, including well depth, screened intervals, and elevation above mean sea level.

Table 3.2-Physical characteristics and geographic coordinates of selected PGMN wells in the BOM region. The table includes well depth, screened intervals, and geographic location (latitude, longitude, and elevation above mean sea level) for each well.

PGMN Well	Well Depth (m)	Screen (m)	Latitude	Longitude	Elevation (m.a.s.l)
W223-1	77.72	73.8 - 76.8	44.38	-79.73	284.9
W244-2	20.27	18.3 - 19.8	44.44	-79.74	239.9
W245-2	76.2	61.6 - 67.4	44.44	-79.74	240.59
W293-2	17.68	13.72 - 16.76	44.53	-79.57	322.92
W293-3	25.91	22.9 - 25.9	44.53	-79.57	322.92
W439-1	58.57	54.04 - 59.4	44.56	-79.59	326.06
W440-1	66.41	58 - 66.4	44.61	-79.64	238.26
W441-1	37.81	33.00 - 37.81	44.69	-79.43	261.68
W442-1	19.83	18.29 - 19.81	44.58	-79.69	256.72
W443-1	118.14	68.5 - 71.63	44.62	-79.69	344.39
W479-1	117.2	83.5 - 90.2	44.34	-79.77	230.01

3.2.2 Meteorological Data

Temperature and total precipitation were selected as the primary exogenous meteorological features for GWL prediction, as these variables are among the most used inputs in groundwater prediction models according to the literature (Ahmadi et al., 2022; Boo et al., 2024; Rajaei et al., 2019). Station data were prioritized over modeled or simulated data whenever possible; however, in some instances, modelled data was used due to insufficient station observations, and in one case, station data was adjusted with nearby modelled data to fill the gaps in the station records. A detailed description of the data sources, preprocessing steps, and gap-filling procedure is provided in Section 4.1.1.

3.2.2.1 Environment and Climate Change Canada (ECCC) Datasets

Hourly historical data from all stations operating in Southern Ontario between 2000 and 2023 were collected using the “env_canada” Python library from the ECCC platform. A total of 314 unique stations were filtered both temporally and spatially to meet the specific needs of this research in Southern Ontario. From the various historical weather parameters provided by the library, such as gust direction, gust speed, and snow on the ground, only mean temperature and total precipitation were selected for this stage of the study. Analysis of the hydrological data revealed numerous instances where data reporting was not continuous, resulting in a heterogeneous dataset.

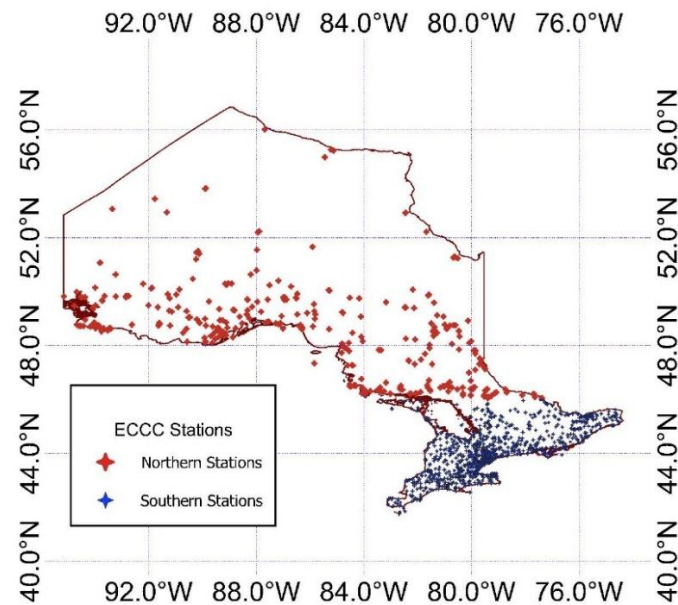


Figure 3-5-Spatial distribution of ECCC meteorological stations across Ontario, categorized into northern and southern stations based on geographic location.

Figure 3.5 illustrates a higher number of ECCC stations in Southern Ontario compared to the northern region, with around 1173 stations (75%) located in the south and 399 stations (25%) situated in the north.

3.2.2.2 Fifth Generation of European Centre for Medium-Range Weather Forecasts Reanalysis Data (ERA5)

The fifth generation of atmospheric reanalysis data produced by the European Centre for Medium-Range Weather Forecasts (ECMWF), or more commonly, ERA5 dataset was downloaded and processed to extract the historical temperature and total precipitation data across Southern Ontario. This dataset offers a global detailed package of records such as temperature, total precipitation, surface pressure, and surface solar radiation from 1950 to the present by employing data assimilation techniques and integrating stationary data, remote sensing data, numerical models, and buoys information (Hersbach, H. et al., 2020). This dataset was downloaded from the Ontario Climate Data Portal (OCDP) which is an initiative led by the Laboratory of Mathematical Parallel Systems (LAMPS) at York University providing high-resolution climate change projections and historical climate data specific to Ontario (Zhu et al., 2020). ERA5 provides meteorological data at a spatial resolution of approximately 0.25 degree. This region, spanning longitudes from 74.3°W to 82.9°W (resulting in about 44 grids) and latitudes from 41.6°N to 46.0°N (about 24 grids), requires higher spatial resolution (Figure 3.6).

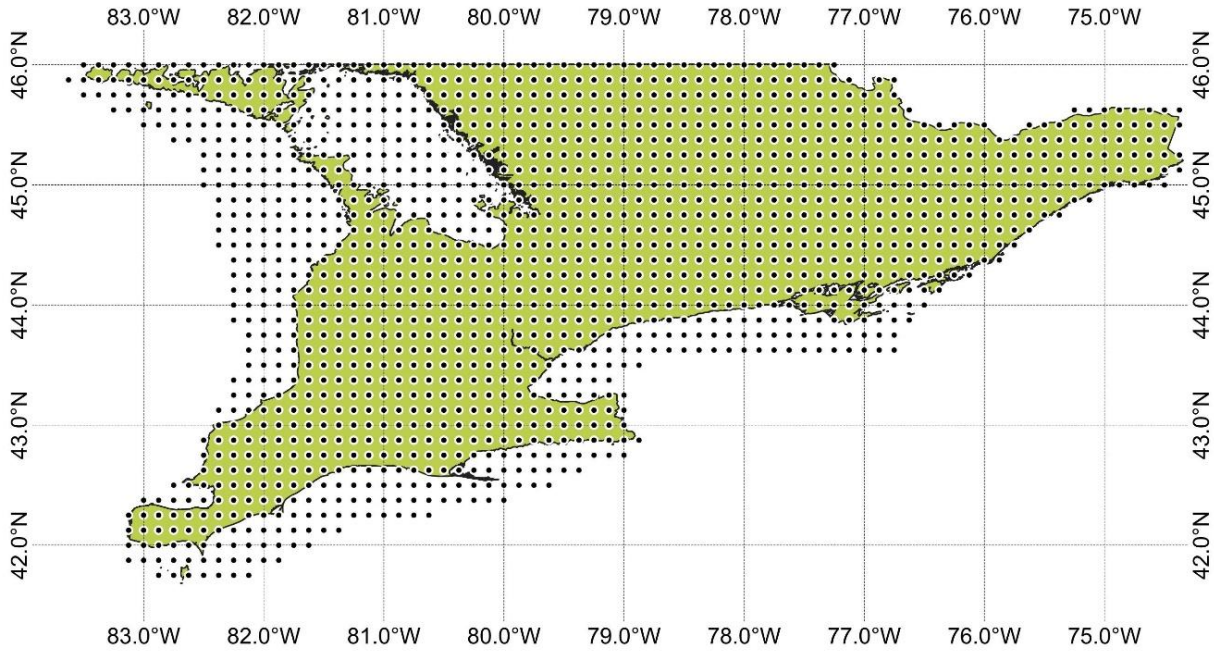


Figure 3-6-ERA5 grid points covering Southern Ontario, illustrating the spatial resolution and distribution of reanalysis climate data across the study region.

ERA5 reanalysis data can provide high-quality inputs for ML models, but it may not capture real-world conditions or observational errors from stations, which can lead to decreased ML performance when dealing with actual noisy data (Camps-Valls et al., 2025). This is particularly relevant for the present study, as GWL prediction relies on meteorological inputs that influence recharge processes at the local scale. Differences between reanalysis data and station observations may therefore affect model accuracy and generalizability when applied to individual monitoring wells.

3.2.3 Hydrometric Data

Hydrometric data refers to measurements of water levels and flow rates in rivers, lakes, and other surface water bodies. In this study, daily hydrometric level measurements were obtained from the Hydrometric Historical Data (HYDAT) database, maintained by the Water Survey of Canada (WSC). These

data represent the water surface elevation (in meters) at designated gauging stations near the selected groundwater wells.

For each well used in the modelling analysis, the nearest active hydrometric station with continuous records during the study period (approximately 2000–2020, depending on the availability of well data) was selected. Stations were chosen based on their proximity, location within the same watershed, and the completeness of their records to ensure that surface water conditions could reasonably represent potential recharge influences on GWLs.

Specifically, hydrometric station 02EC020 – Hawkestone Creek at Hawkestone was used for wells 293-2 and 293-3; 02ED032 – Willow Creek near Minesing for wells 245-2 and 244-2; 02ED007 – Coldwater River at Coldwater for wells 439-1, 440-1, and 442-1; and 02ED017 – Hogg Creek near Victoria Harbour for well 443-1. Only the water level data (not streamflow) from these stations were used in the analysis, as GWL prediction studies commonly rely on hydraulic and meteorological variables rather than streamflow values (Ahmadi et al., 2022; Tao et al., 2022; Uc-Castillo et al., 2023). The locations of these hydrometric stations in relation to the monitoring wells are shown in Figure 3.3 in Section 3.1.3.

In the subsequent statistical (Section 4.2.2) and ML (Section 4.4.4) modelling analyses conducted in this thesis, GWLs were used within different modelling configurations. The raw daily measurements were processed into monthly averages to match the resolution of other input features in the modelling framework. These monthly values were then included as one of the main predictors in the models developed for GWL forecasting. The hydrometric level variable helps account for the effect of nearby rivers or streams, particularly in areas with shallow, unconfined aquifers where direct surface-groundwater interactions are more likely (Sophocleous, 2002).

3.2.4 Watersheds

Eight monitoring wells within the BOM area were selected within four quaternary watersheds. A watershed (also referred to as a drainage basin, river basin, or catchment) is the topographically defined area that contributes all water to a specified cross-section of a stream, known as the outlet (Dingman, 2015). These wells are distributed among the Severn River-Lake Simcoe and Nottawasaga River watersheds, spanning catchment areas from 189 to 327 km². As illustrated in Figure 3.3, wells 293-2 and 293-3 are nested within the Severn River-Lake Simcoe watershed (catchment: 234 km², Watershed W1), while Wells 439-1, 440-1, and 442-1 lie in separate zones of the Nottawasaga River watershed (217 km², Watershed W2). Additionally, Wells 244-2 and 245-2 are nested within another part of the Nottawasaga River watershed (327 km², Watershed W3), and Well 443-1 is located in a separate segment with a smaller catchment area (189 km², Watershed W4).

Table 3.3-Summary of the eight monitoring wells selected within the BOM area, grouped by quaternary watershed. Each well is associated with a Hydrologic Unit Code (HUC), watershed name, catchment area, watershed ID, and whether it lies within a nested or separate watershed system.

Well	*HUC	Quaternary Watershed	Catchment Area (km ²)	Watershed ID	Type
293-3	02EC-09	Severn River – Lake Simcoe	234	W1	Nested
293-2	02EC-09	Severn River – Lake Simcoe	234	W1	Nested
439-1	02ED-05	Nottawasaga River	217	W2	Separate
440-1	02ED-05	Nottawasaga River	217	W2	Separate
442-1	02ED-05	Nottawasaga River	217	W2	Separate
245-2	02ED-14	Nottawasaga River	327	W3	Nested
244-2	02ED-14	Nottawasaga River	327	W3	Nested
443-1	02ED-04	Nottawasaga River	189	W4	Separate

3.2.5 Aquifers

The PGMN dataset categorizes aquifers into overburden, bedrock, and interface aquifers – where groundwater flow occurs at the contact zone between overburden and bedrock. However, the PGMN database does not specify the exact lithological characterization or geological formation data of these aquifers, posing challenges for developing more precise modelling scenarios (Rogojin & Geo, 2024).

Within the BOM region, 11 PGMN wells have been installed, but this study only focuses on a subset of eight wells. These were selected based on the availability of detailed aquifer descriptions in the report by Cuddy (2013) and Burt & Dodge (2011), which compiled stratigraphic interpretations for wells across the Lake Simcoe watershed. Using this information, along with the Aquifer Detection Model (AQM) developed in this study (Section 3.2.6), the likely aquifer types associated with each well were identified. The AQM used both continuous and discontinuous three-dimensional geological datasets from the Ontario Geological Survey (OGS), combined with k-d tree spatial analysis, to assign aquifer units based on proximity to screened intervals. All selected wells are screened in overburden or bedrock aquifers, but they span a range of stratigraphic units (as shown in Figure 3.7), from surficial Algonquin aquifers (GLAF) to deeper glaciofluvial systems such as upper valley-fill aquifer (CAF1), lower valley fill aquifer (CAF3), and upper regional aquifer (AF1)-regional aquifer (AF4). The aquifer assignments also reflect variations in lithology, including sand, silty sand, and gravel-rich deposits. Table 3.4 summarizes the selected wells, their associated watersheds, assigned aquifer units (based on Cuddy, 2013), and the corresponding aquifer materials derived from the OGS hydrostratigraphic framework. To support the interpretation, Figure 3.7 presents a conceptual cross-section of the study area's hydrostratigraphy, showing the spatial relationships among aquifers and aquitards identified in this region.

Table 3.4-Aquifer classification and lithological composition for selected PGMN wells within the BOM area, based on Cuddy (2013) and the OGS hydrostratigraphic model (Burt & Dodge, 2011).

Well ID	Sub-Watershed	Aquifer Name (Cuddy, 2013)	Aquifer Material (Burt & Dodge, 2011)
293-2	W1	AF1-GLAF	Gravel, sand and silty sand
293-3	W1	AF1-GLAF	Gravel, sand and silty sand
439-1	W2	AF2	Gravel, sand and silty sand
440-1	W2	Bedrock	Bedrock
442-1	W2	CAF1-AF4	Sand and silty sand
244-2	W3	AF1-GLAF	Gravel, sand and silty sand
245-2	W3	CAF3-LAF	Gravel, sand and silty sand
443-1	W4	AF1-GLAF	Gravel, sand and silty sand

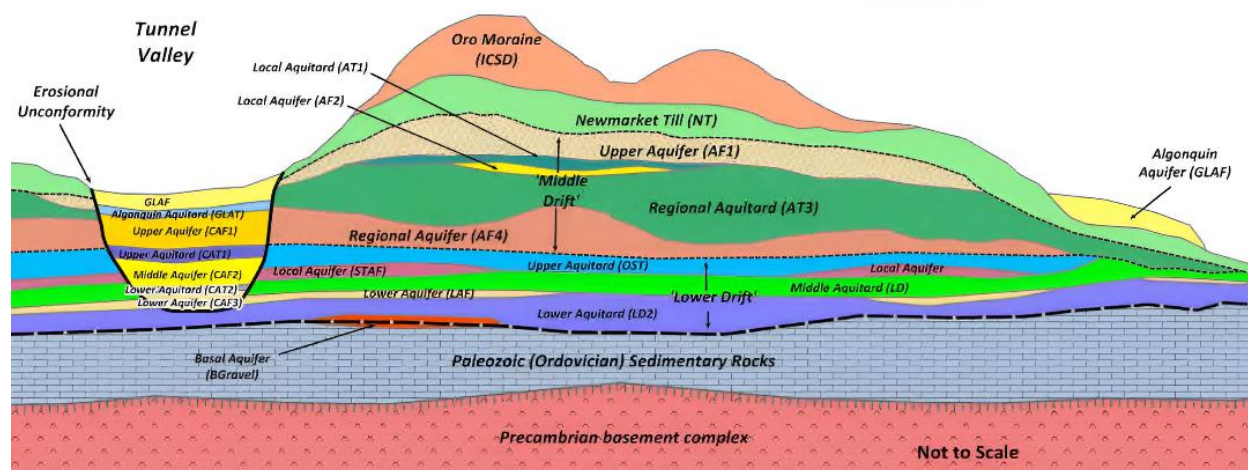


Figure 3-7- Conceptual hydrostratigraphic cross-section of BOM area showing key aquifers and aquitards based on OGS classification, derived from Cuddy (2013).

3.2.6 GWL Data Resolution

The frequency of GWL data transmission to the Ministry of the Environment, Conservation and Parks (MECP) varies among conservation authorities. This data is then stored in the Provincial Monitoring

Groundwater Information System (PMGIS) and is accessible to program partners. Under the PGMN framework, monitoring wells within the Nottawasaga Valley Conservation Authority and the Lake Simcoe Region Conservation Authority are equipped with automated data loggers that record GWL at an hourly resolution. In the Nottawasaga Valley, documentation indicates that although GWL is recorded hourly, water level data are formally downloaded twice per year. For the Lake Simcoe Region, publicly available documentation confirms continuous monitoring under the PGMN framework; however, the exact frequency of data transmission to MECP is not explicitly stated.

While retrieval and transmission schedules differ administratively among conservation authorities, the underlying measurement resolution remains hourly. For the data-driven modelling component of this study, hourly GWL observations were aggregated to a monthly resolution to ensure temporal consistency across wells and to reduce short-term variability. This aggregation approach is supported by the fact that all wells operate with high-frequency (hourly) measurements, regardless of the administrative data download schedule.

3.2.7 Data Inputs for GWL Modelling

In all the data-driven models developed in this study, temperature, precipitation, and hydrometric level data were used as the main exogenous input features to predict GWL, while the previous time step of GWL was included as an autoregressive input. These variables are commonly used in data-driven GWL modelling studies, as they capture key hydrological and climatic drivers of GWL fluctuations (Rajaei et al., 2019; Tao et al., 2022; Uc-Castillo et al., 2023). In contrast, physically based groundwater models may additionally incorporate pumping rates, hydraulic conductivity, boundary conditions, and aquifer properties; however, such detailed hydrogeological and abstraction data were not consistently available for all wells in the study area. The GWL data used in this study were obtained from the PGMN wells, which provide long-term, continuous monitoring across Ontario. These records capture natural and climate-related

variations in groundwater systems (Lake Simcoe Region Conservation Authority, 2024). However, the reliability of PGMN data may decline in areas near heavy pumping or quarry operations, where local disturbances can influence the readings (Cuddy, 2013).

Although groundwater use is a known factor influencing GWL changes, it was not directly included as a model input. This is because variations in precipitation, runoff, and GWL itself can implicitly reflect water consumption patterns, such as irrigation or pumping, even without direct usage records (Sharghi et al., 2022). A similar approach has been adopted in studies facing limited access to reliable pumping records. For example, Sahu et al. (2020) noted that higher pumping rates often coincide with periods of low precipitation and elevated temperatures, indicating that abstraction can be indirectly associated with climatic conditions. Due to the absence of consistent, long-term pumping records, they excluded explicit withdrawal data and relied on hydroclimatic variables as practical proxy drivers. In another modelling study by Janardhanan et al. (2023), groundwater consumptive use was incorporated indirectly in MODFLOW using evapotranspiration and well packages and calibrated against observed GWLs, instead of prescribing fixed pumping rates, due to the lack of reliable spatial and seasonal pumping data. Another practical limitation relates to the restricted temporal coverage of publicly available abstraction records in Southern Ontario. While the Permit to Take Water (PTTW) dataset offers daily withdrawal data for users extracting more than 50,000 liters per day, it only covers 2019 and 2020 (Xu & Lofranco, 2023). As shown in Figure 3.8, usage tends to peak in late spring to early summer; however, the short time span of the data reduces its value for long-term modelling efforts.

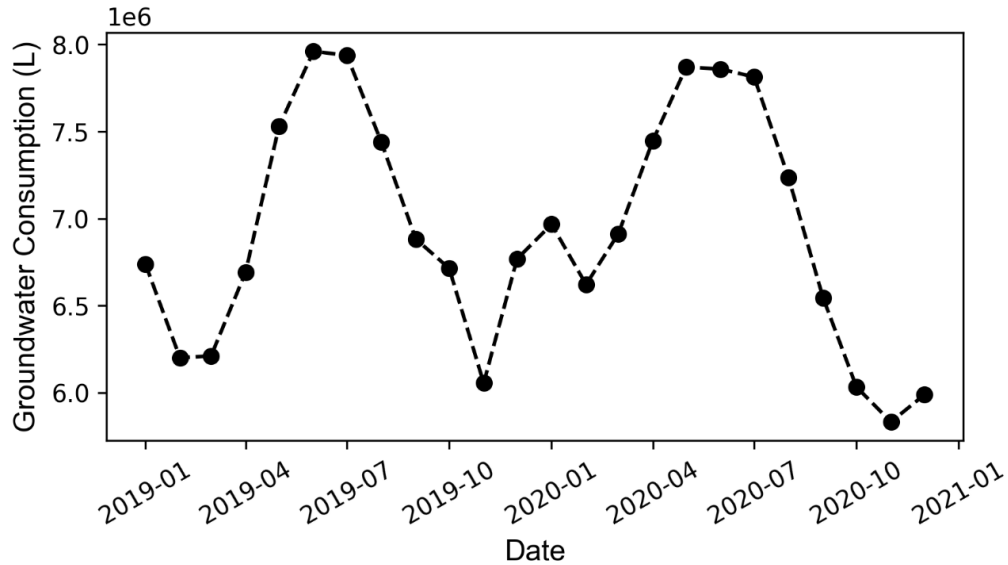


Figure 3-8-Monthly groundwater consumption trends in Ontario from the PTTW dataset, showing peak withdrawals during the late spring and summer months of 2019 and 2020.

These factors highlight the challenges of relying on consumption rates as explicit model inputs. To ensure broader applicability and usability of the trained models, especially in future scenarios where consumption data may not be readily available, this study focused exclusively on hydrometeorological inputs and autoregressive GWL data. This approach enables the models to infer the influence of groundwater use through the observed behavior of the groundwater system itself.

To summarize the data used in model training, Table 3.5 presents the descriptive statistics of the key input variables and the GWL for well 293-2. The average monthly temperature was 6.85°C, with a standard deviation of 10.13°C, and ranged from -14.89°C to 23.07°C, showing strong seasonal variation. Precipitation had a wide range from 0 to 277.44 mm, with a mean of 78.72 mm and a high standard deviation of 55.66 mm, indicating notable month-to-month variability. The hydrometric level showed relatively stable values, with a mean of 2.27 m and a small standard deviation of 0.12 m, ranging between 2.11 m and 2.87 m. The GWL values were also relatively consistent, ranging from 306.64 m to 308.48 m, with a mean of 307.69 m and a standard deviation of 0.41 m.

Table 3.5-Descriptive statistics (mean, standard deviation, minimum, 25th percentile, median, 75th percentile, and maximum) of the key input variables and GWL for well 293-2 over the study period (2006-2018).

	Temperature (°C)	Precipitation (mm)	Hydrometric Level (m)	GWL
count	150	150	150	150
mean	6.85	78.72	2.27	307.69
std	10.13	55.66	0.12	0.41
min	-14.89	0	2.11	306.64
25%	-1.64	40.28	2.18	307.39
50%	6.7	75.9	2.24	307.63
75%	16.07	111.28	2.33	307.99
max	23.07	277.44	2.87	308.48

Figure 3.9 displays the full time series of monthly temperature, precipitation, hydrometric level, and GWL values at this well over the study period (2006-2018), revealing clear seasonal cycles and long-term fluctuations across all variables. Summary statistics and graphs for the remaining wells of BOM are provided in Appendix 7.2.

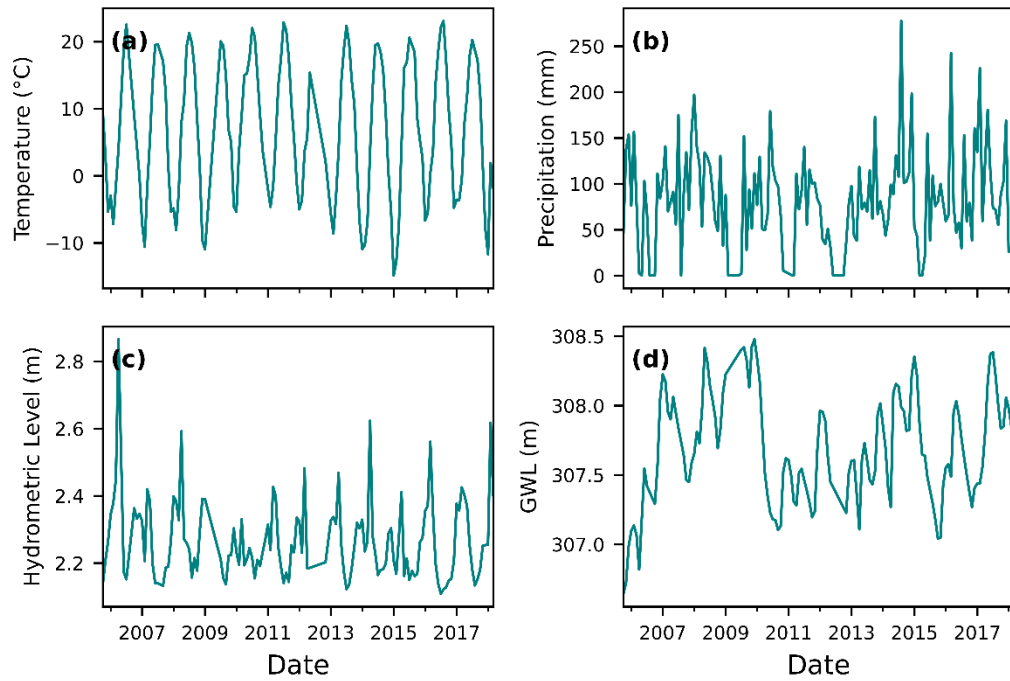


Figure 3-9-Time series plots of the monthly (a) temperature, (b) precipitation, (c) hydrometric level, and (d) GWL at well 293-2, showing seasonal and long-term variations from 2006 to 2018.

Chapter 4 Methodology

This chapter explains the main steps used to analyze GWLs in the study area. It starts with data preprocessing to handle missing values and clean inputs from PGMN, ECCC, and HYDAT sources. Then, it covers the statistical methods used to detect trends and test assumptions. Different modelling scenarios are introduced, including single, combined, and neural network models. The chapter also presents the Aquifer Detection Model (AQM), which identifies the likely aquifer for each well using spatial and geological data. Finally, model training, tuning, and evaluation steps are described.

4.1 Data Preprocessing

Since this study prioritizes data from stations (PGMN, ECCC, and HYDAT), this information should be preprocessed in advance to reduce the uncertainties incorporated within the data. ERA5, a modelled atmospheric reanalysis dataset, is a complete dataset offering climatic information from 1940 to present; however, this dataset may not indicate the real conditions of temperature and precipitation in different sub-watersheds on the study area.

4.1.1 Handling Missing Data

The study prioritized using data from ECCC stations due to their higher reliability and closer alignment with real-world observations than simulated or modeled data, such as ERA5 (Fatolahzadeh Gheysari et al., 2024; Goswami et al., 2024). There were instances where a significant gap existed in the station data records. In such cases, ERA5 data was used to fill the gaps using a correlation-based adjustment approach. Specifically, the ERA5 dataset was first compared to the corresponding ECCC station observations to assess the strength of their relationship (R^2). When the correlation between the two datasets was high, a linear regression (trendline) was developed between ERA5 and ECCC values, and the fitted relationship was used to estimate missing observations. Conversely, when the correlation was low, ERA5

values were used directly without adjustment to avoid introducing additional uncertainty. Additionally, for specific watersheds with no active ECCC stations, the ERA5 grid points within the watershed served as the primary source of meteorological data. However, reanalysis data products, such as ERA5, often exhibit biases compared to actual observations (Goswami et al., 2024; Niazkar et al., 2023), necessitating the prioritization of adjustments to align with real-world measurements.

A cross-correlation analysis was conducted between the ECCC and ERA5 datasets to examine if a temporal lag existed in the ERA5 data, as illustrated in Figure 4.1. This diagnostic step was used to assess whether the regression relationship should be applied to concurrent values or lagged data. The results revealed that the highest correlation occurred at zero lag, prompting the next steps of the modification to focus on identifying the amplitude relationships between the two datasets.

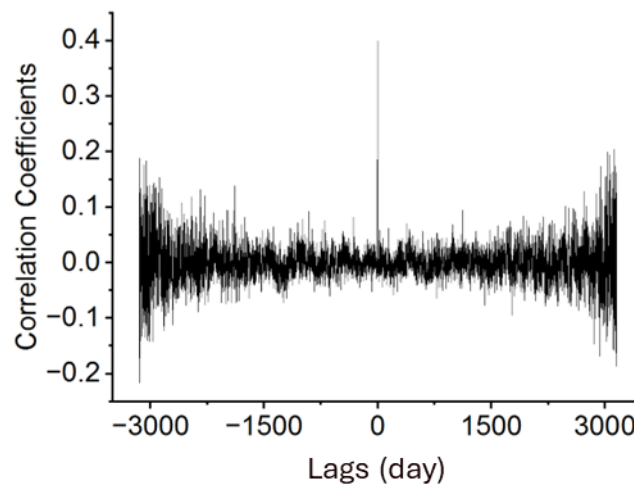


Figure 4-1- Cross-correlation between ECCC and ERA5 temperature datasets to investigate potential time lags. The analysis shows that the maximum correlation occurs at lag 0, indicating that no significant temporal offset exists between the two datasets. Subsequent analyses focused on evaluating amplitude differences.

To achieve this, a linear regression was established between the observed and reanalysis data to determine the relationship between the two datasets. The resulting trendline shown in Figure 4.2 represents the best-fit linear regression between ERA5 and ECCC values. Gheysari et al. (2024) suggested using linear

regression between ERA5 and actual mean air temperature and precipitation data to potentially address temporal and spatial biases in the Canadian Prairies, including Manitoba.

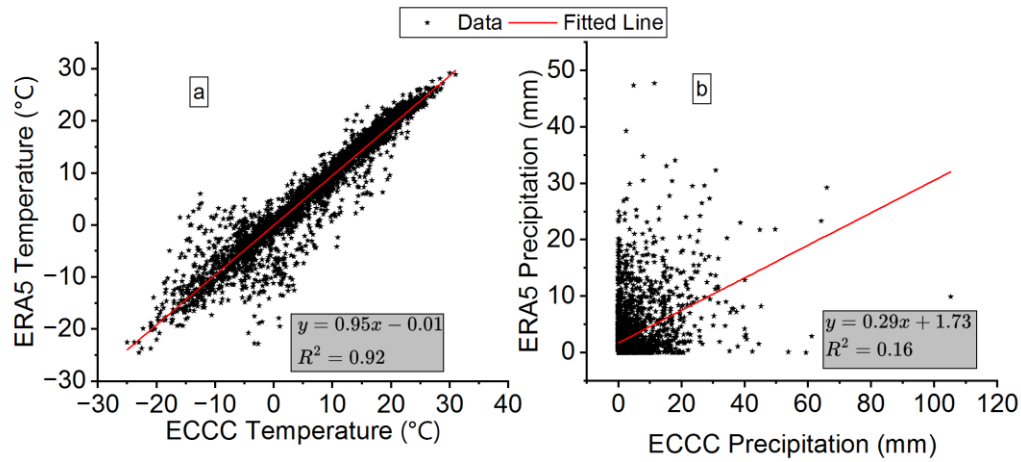


Figure 4-2-Scatter plots comparing ERA5 and ECCC datasets: (a) shows a strong linear relationship between ERA5 and ECCC temperature values ($R^2 = 0.92$), while (b) reveals a weak correlation between ERA5 and ECCC precipitation data ($R^2 = 0.16$). Red lines represent the fitted linear regression models.

This approach was similarly applied in the current study to fill in missing data. A strong correlation was found between ERA5 and ECCC temperature records ($R^2 = 0.92$), and the following equation was used to align the datasets:

$$ERA5 = 0.95(ECCC) - 0.01 \quad (4.1)$$

This is illustrated in Figure 4.3-a. For precipitation, however, the correlation between ERA5 and ECCC was weak, and the resulting regression equation showed a low R^2 value. Therefore, applying the regression adjustment would have increased uncertainty rather than reduced it. In this case, ERA5 data were used directly to fill the missing precipitation records, as this represented the most practical approach given the limited data availability. The process is represented in Figure 4.3-b.

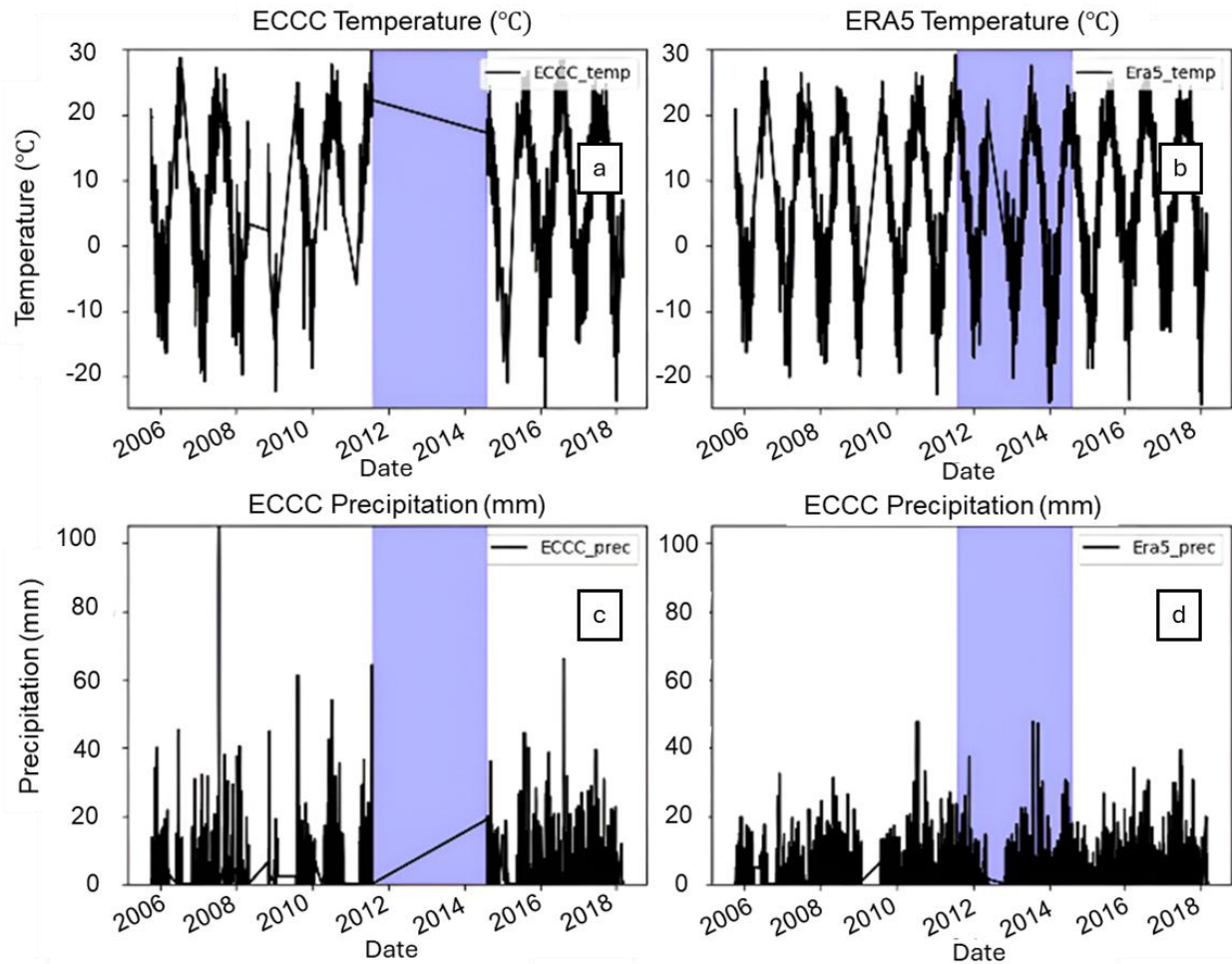


Figure 4-3-Comparison of ECCC and ERA5 temperature and precipitation datasets. Panel (a) shows observed ECCC air temperature, and (b) shows ERA5 air temperature for the same period. Panels (c) and (d) present ECCC and ERA5 precipitation data, respectively. The shaded blue region highlights the data gap in ECCC records, which was filled using ERA5 data. Temperature data were bias-corrected using a regression model, while ERA5 precipitation was used directly due to low correlation.

The modified version of the ECCC dataset for this specific weather station is depicted below in Figure 4.4. Finally, for those watersheds (W2, W3, and W4), where no ECCC weather station is available, ERA5 grid boxes were used as the only available data in that watershed.

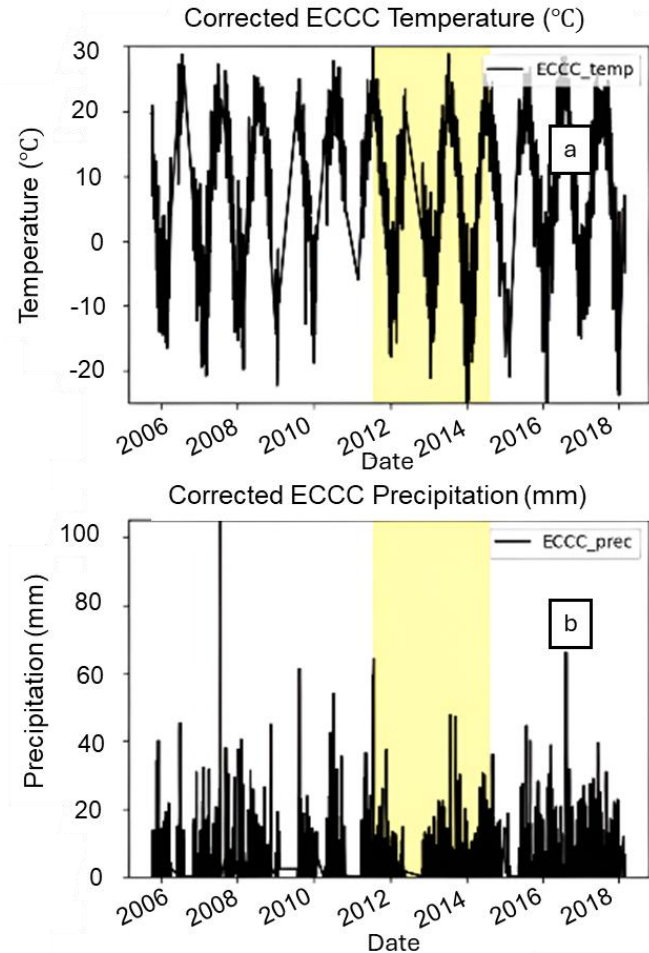


Figure 4-4-Final corrected ECCC climate data after gap-filling. Panel (a) shows the air temperature time series with the missing data section (highlighted in yellow) filled using the bias-adjusted ERA5 temperature dataset. Panel (b) presents the precipitation series, where the missing values were directly replaced with ERA5 precipitation data due to weak correlation.

4.1.2 Outlier Removal Algorithm Using Interquartile Range (IQR) Method

Efficiently identifying outliers is crucial in statistical analysis, as their presence can lead to inaccurate or misleading results. Conventional outlier detection methods used in hydrology typically include the three-sigma rule (3σ), median absolute deviation (MAD), and interquartile range (IQR) techniques (Filzmoser et al., 2008). The 3σ method, being a parametric approach, is suitable for normally distributed data, whereas MAD and IQR serve as non-parametric techniques for identifying and removing

outliers (Jeong et al., 2017). Jeong et al. (2017) found that the IQR method generally surpasses the MAD method in identifying outliers in both real-world and simulated GWL data. The study highlights that IQR is more accurate and robust, particularly for handling non-Gaussian datasets with higher densities of outliers. The formula for this method is provided below:

$$IQR = Q_3 - Q_1 \quad (4.2)$$

Where Q_3 represents the upper quartile (75th percentile) and Q_1 denotes the lower quartile (25th percentile) of the dataset. The threshold for identifying an observation as an outlier is determined by a range extending below the first quartile and above the third quartile:

$$Data_{min}: Q_1 - \alpha \times IQR; Data_{max}: Q_3 + \alpha \times IQR \quad (4.3)$$

The parameter α is typically set to 1.5 but can be adjusted depending on the specific analysis and the desired level of certainty (Song et al., 2022).

4.1.3 Singular Spectrum Analysis (SSA)

Time series data, such as GWL and precipitation, often require smoothing and denoising to enhance stationarity and prepare the data for modelling (Cao & Zheng, 2016). SSA is a commonly used form of Principal Component Analysis (PCA), applied in the lag-time domain. It enhances the signal-to-noise ratio in a data-adaptive manner, enabling the detection of periodic patterns in short and noisy time series (Vautard et al., 1992). Compared to traditional spectral analysis tools such as Fourier Analysis, SSA offers higher resolution in both the time and frequency domains. While SSA does not directly produce spectral estimates, it is used to break down the time series into its trend, cyclic components, and noise. The reconstructed periodic signals (RCs) are more distinct, making them easier to analyze spectrally (Cao & Zheng, 2016). The SSA can be used for denoising algorithm by decomposing and reconstructing the time series data as follows (Hassani & Zhigljavsky, 2009).

The SSA algorithm was applied to the ECCC and ERA5 timeseries datasets (GWL, temperature, and precipitation) using Python 3.11 with the NumPy, Pandas, and scikit-learn libraries. The process consisted of four stages - embedding, singular value decomposition, grouping, and reconstruction – presented below as Steps 1 to 4.

Step 1: Embedding The embedding stage, that is the first stage, transforms a one-dimensional time series data $Y_T = \{y_1, y_2, \dots, y_T\}$ into a multi-dimensional series by constructing the initial trajectory matrix X , that is a Hankel matrix:

$$X = [X_1, X_2, \dots, X_K] = \begin{pmatrix} y_1 & y_2 & \dots & y_K \\ y_2 & y_3 & \dots & y_{K+1} \\ \vdots & \vdots & \ddots & \vdots \\ y_L & y_{L+1} & \dots & y_T \end{pmatrix}; X_i = (y_i, y_{i+1}, \dots, y_{i+L-1})^T \quad (4.4)$$

Where L represents the window length, $K = T - L + 1$ denotes the number of columns, and each X_i is a lagged vector with L dimensions.

Step 2: Singular Value Decomposition (SVD) The initial trajectory matrix X is subjected to Singular Value Decomposition (SVD), which can be expressed as:

$$X = \sum_{i=1}^{i=d} \sqrt{\lambda_i} U_i V_i^T \quad (4.5)$$

Where λ_i are the eigenvalues of XX^T (ordered as $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_d > 0$), U_i are the eigenvectors (columns of matrix U) corresponding to λ_i , $V_i = \frac{X^T U_i}{\sqrt{\lambda_i}}$ is the factor vector (right singular vectors), and d is the rank of trajectory matrix X . Each term $\sqrt{\lambda_i} U_i V_i^T$ is a rank-1 matrix called an elementary matrix, and the collection is referred to as eigentriples $\{(\sqrt{\lambda_i}, U_i, V_i)\}$.

Step 3: Grouping At this stage, the SVD components, $\sqrt{\lambda_i} U_i V_i^T$, are categorized according to their contribution to the signal. If $I = \{i_1, i_2, \dots, i_p\}$ is a list of group indices (up to p eigentriples), then the grouped matrix X_I is formed as:

$$X_I = X_{i_1} + X_{i_2} + \dots + X_{i_p} = \sum_{i \in I} \sqrt{\lambda_i} U_i V_i^T \quad (4.6)$$

Grouping the data transforms the trajectory matrix into distinct components, including trends, oscillations, and noise.

Step 4: Reconstruction (Diagonal Averaging) The grouped matrices X_I are then transformed back into time series using diagonal averaging (Hankelization):

$$y_K = \frac{\sum_{(i,j): i,j=K+1} x_{ij}}{\text{number of elements in diagonal } K} \quad (4.7)$$

The final SSA formula sums the reconstructed components (Y_T^m) and the residuals (R_T) for each group (M) of trends, oscillation, and noise (Hassani & Zhigljavsky, 2009):

$$Y_T = \{y_1, y_2, \dots, y_T\} = \sum_{m=1}^M Y_T^m + R_T \quad (4.8)$$

4.1.3.1 Window Size (l) in SSA

The window size (l) is a crucial parameter in the SSA algorithm as it directly influences how the time series is decomposed. A small l value primarily captures short-term fluctuations, whereas a larger l shifts the focus toward long-term trends (Sun & Li, 2017). Given this, selecting the optimal l is both challenging and essential, as it significantly affects the patterns extracted from the input data. The choice of l is typically based on domain knowledge (Marques et al., 2006), trial and error until non-noisy trends are achieved (Chau & Wu, 2010), or theoretical approaches designed for optimal selection (Golyandina, 2011; Hassani et al., 2011; Sun & Li, 2017). In this study, the Singular SSA algorithm was used differently in the DSM and DMS setups. In the DSM configuration, SSA was applied to the daily data using a 30-day window to capture monthly patterns before aggregating. Then, temperature, hydrometric level, and GWL values were averaged by month, while precipitation was summed using the *resample* function in Pandas. In the DMS setup (detailed in 4.3.4), however, SSA was applied after the data were aggregated to monthly

resolution, using a shorter window length ($l = 3$) to smooth the already averaged data. A separate scenario also examined the effect of applying SSA directly on monthly values.

4.1.4 Data Normalization: Min-Max Scaler

Both for regression analysis and neural network modelling, scaling the variables with different magnitudes to a fixed range can enhance the stability of the model and avoid overfitting (Belsley et al., 2005; Hsieh, 2023).

Normalization is a type of data scaling technique and adjusts data to a specific range, ensuring that no single feature disproportionately influences the analysis due to its scale. In this study, the Min-Max scaling method is applied to transform all data, including lagged and auto-regressive features as well as the targets, to a range of $[0, 1]$ using the formula (Pedregosa et al., 2011):

$$\tilde{x} = \frac{x - \min(x)}{\max(x) - \min(x)} \quad (4.9)$$

4.2 Statistical Analysis Methods

This section outlines the statistical techniques used to support the interpretation of long-term GWL time series data and to evaluate the assumptions underlying the linear regression models. These methods were selected to ensure the robustness and validity of the modelling process, particularly when dealing with time-dependent data.

To assess the statistical assumptions of normality, linearity, homoscedasticity, and independence of residuals, a range of hypothesis tests were conducted, including the Kolmogorov-Smirnov test, t-test, and tests for autocorrelation and variance. Additionally, trend analysis methods such as the Mann-Kendall test and Sen's slope estimator were employed to detect and quantify monotonic trends in the data.

Since hydrometeorological variables like precipitation and temperature may influence GWL with time lags, correlation analysis methods—including cross-correlation and autocorrelation—were also applied. These methods helped identify temporal dependencies and lag structures, supporting a more accurate understanding of the underlying physical processes. The following subsections describe each of these statistical methods in detail.

4.2.1 Hypothesis Testing

To evaluate the assumptions of the model and residual behavior in the linear regression section, several statistical tests were performed. The one-sample Kolmogorov-Smirnov (K-S) test was used to verify whether the residual distributions matched the expected theoretical (normal) distribution. Homoscedasticity was assessed using the Breusch-Pagan test to check for constant variance in the residuals. To detect autocorrelation, the Durbin-Watson statistic was used as a preliminary test, followed by the Breusch-Godfrey test with a lag of one to formally evaluate whether the residuals were correlated. These tests collectively ensured that the linear regression assumptions were satisfied, and the residuals exhibited the desired properties.

4.2.1.1 Two-sample Kolmogorov-Smirnov Test

The two-sample Kolmogorov-Smirnov (K-S) test is a non-parametric statistical test used to determine if two independent samples come from the same distribution. It compares the cumulative distribution functions (CDFs) of the two samples and identifies the largest absolute difference between them. The K-S test does not assume any specific data distribution and is widely applicable for testing differences between empirical distributions (Massey Jr., 1951).

Mathematically, the test statistic D is defined as:

$$D = \sup_x |F_1(x) - F_2(x)| \quad (4.10)$$

Where $F_1(x)$ and $F_2(x)$ are the empirical cumulative distribution functions (ECDFs) of the two samples and $\sup_x g(x)$ is the supremum (largest difference) over all observed values x .

The null hypothesis (H_0) states that the two samples are drawn from the same distribution. If the p-value is less than the chosen significance level (e.g., 0.05), the null hypothesis is rejected, indicating a significant difference between the distributions.

4.2.1.2 Breusch-Pagan Test

The Breusch-Pagan (BP) test is used to detect heteroscedasticity in a regression model, i.e., whether the variance of the residuals is constant across all levels of the independent variable(s). It tests the null hypothesis that the variance of the errors is constant (H_0 , homoscedastic) against the alternative that it varies (H_1 , heteroscedastic). The test is performed by regressing the squared residuals from the original regression on the original independent variables. The test statistic is calculated as:

$$BP = nR^2 \quad (4.11)$$

Where n is the number of observations and R^2 is the coefficient of determination from the auxiliary regression.

The BP statistic follows a chi-squared (χ^2) distribution with degrees of freedom equal to the number of independent variables. If the p-value is less than the chosen significance level ($\alpha = 0.05$), the null hypothesis of constant variance is rejected (T. S. Breusch & Pagan, 1979).

4.2.1.3 Durbin-Watson Test

The Durbin-Watson (DW) test checks for first-order autocorrelation in the residuals of a regression model. It tests whether the residuals from a linear regression are independent over time. The test statistic is defined as:

$$DW = \frac{\sum_{t=2}^n (e_t - e_{t-1})^2}{\sum_{t=1}^n e_t^2} \quad (4.12)$$

Where e_t is the residual of time series at time t and n is the number of observations.

If the test statistic is equal to 2, it indicates no autocorrelation in the residuals. A value less than 2 suggests positive autocorrelation, while a value greater than 2 indicates negative autocorrelation. However, the Durbin-Watson test does not formally test a null hypothesis, and it only provides an approximate indication of the presence and direction of first-order autocorrelation (Durbin & Watson, 1950).

4.2.1.4 Breusch-Godfrey Test

The Breusch-Godfrey (BG) test is a general test for higher-order autocorrelation in regression residuals. Unlike the Durbin-Watson test, the BG test can test for lags greater than one and allows for lagged dependent variables as regressors. The BG test procedure involves: (a) estimating the regression model and obtaining residuals (b) regressing these residuals on the original regressors and lagged values of the residuals up to the desired lag order. Finally, the test statistics could be calculated as:

$$BG = nR^2 \quad (4.13)$$

Where n is the number of observations in the dataset and R^2 comes from the auxiliary regression of residuals on lagged residuals and independent variables.

The BG statistic follows a chi-squared (χ^2) distribution with degrees of freedom equal to the number of lags. A low p-value (lower than $\alpha = 0.05$) indicates the presence of autocorrelation in the residuals (T. s. Breusch, 1978; Godfrey, 1978).

4.2.2 Trend Analysis

All groundwater time series data were analyzed using trend analysis algorithms to examine their long-term behavior within the study area. The Mann-Kendall test is among the most robust hypothesis tests for this purpose, which is explained in more detail in the following subsection. This test was applied exclusively to overburden aquifers due to their susceptibility to climatic conditions.

4.2.2.1 Mann-Kendall Test

The Mann-Kendall test is a non-parametric statistical test utilized to detect trends in a timeseries data. This test compares the rank of each data point with all the subsequent observations and calculates their relative difference according to their rank by a sign function. This results in a distribution-free test as no data was transformed to any distribution and only depends on the order of the timeseries data rather than their quantities. For a timeseries data of $X = \{x_i, x_{i+1}, \dots, x_{n-1}, x_n\}$ with orders of $R = \{r_i, r_{i+1}, \dots, r_{n-1}, r_n\}$, the test statistic parameter (S) is introduced as (Kendall, 1948; Mann, 1945):

$$S = \sum_{i=1}^{i=n-1} \sum_{j=i+1}^{j=n} \text{sgn}(r_j - r_i) \quad (4.14)$$

Where:

$$\text{sgn}(r_j - r_i) = \begin{cases} 1 & r_j > r_i \\ 0 & r_j = r_i \\ -1 & r_j < r_i \end{cases} \quad (4.15)$$

If there are no tied values and all the data points are different from each other, the mean and variance of the parameter S is as follows:

$$E(S) = 0 \quad (4.16)$$

$$Var(S) = \frac{n(n-1)(2n+5)}{18} \quad (4.17)$$

Where n is the number of data point in the dataset X . However, if there are instances where some points take the same value creating tied groups, a reduction in variance will happen based on the following equation with the same value for the mean:

$$Var(S) = \frac{n(n-1)(2n+5) - \sum_{i=1}^{i=q} t_i(t_i-1)(2t_i+5)}{18} \quad (4.18)$$

Where q is the number of tied groups and t_i is the number of equal observations at each tied group.

Kendall (1975) proves that when the number of observations becomes vast, the distribution of S tends to be normal. The significance of the trend, as a result, could be tested by comparing the standardized test statistic (Z) with the standard normal random variable at the desired user-identified significance level α :

$$Z = \begin{cases} \frac{S-1}{\sigma} & S > 0 \\ 0 & S = 0 \\ \frac{S+1}{\sigma} & S < 0 \end{cases} \quad (4.19)$$

If $|Z| \leq \alpha$, the null hypothesis (H_0) cannot be rejected resulting in no trend within the timeseries data. However, if $|Z| > \alpha$, there is satisfying evidence to reject the null hypothesis and the alternative hypothesis (H_a) will be chosen indicating that there is a significant trend in the timeseries data.

4.2.3 Sen's Slope (Theil-Sen Estimator)

Sen's Slope is a non-parametric method used to estimate the rate of change in a time series by computing the median of all pairwise slopes between data points. It is robust to outliers and provides a reliable measure of trend magnitude. Developed by Sen, this approach calculates the slope for each pair of observations in a dataset with N values and uses the median of these slopes as the overall trend estimate (Frimpong et al., 2022; Sen, 1968):

$$Q_i = \frac{x_j + x_k}{j - k} \text{ for } i = 1, \dots, N \text{ and } j > k \quad (4.20)$$

Where, x_j and x_k are the data values at times j and k . If there is only one observation per each time in the dataset, the number of pairs, N , could be defined as:

$$N = \frac{n(n-1)}{2} \quad (4.21)$$

Where n is the number of observations in the dataset. The N values of the Q_i are then ranked from the minimum to the maximum and the median of Sen's slope is calculated using the following rule:

$$Q_{med} = \begin{cases} Q_{[\frac{N+1}{2}]} \rightarrow \text{if } Q \text{ is odd} \\ \frac{Q_{[\frac{N}{2}]} + Q_{[\frac{(N+2)}{2}]}}{2} \rightarrow \text{if } Q \text{ is even} \end{cases} \quad (4.22)$$

The sign of Q_{med} indicates the sign of trend (positive for upward and negative for downward) and its value reflects the steepness of the trend. Using this robust estimator, we can determine both the direction and magnitude of trends in the GWL time series and gain insight into how these trends are statistically distributed.

In this study, the Mann–Kendall test and the associated Sen's slope estimator were implemented using the *pymannkendall* library in Python. This library provides a standardized implementation of both

methods, including handling of tied values and calculation of Sen's slope, consistent with the formulations described above.

4.3 Modelling Scenarios

At the core of this research lies the development and evaluation of three modelling scenarios: .single, combined, and aggregated – to predict GWLs across complex aquifer systems. These scenarios represent different ways of structuring input-output relationships between features and groundwater responses, each reflecting a unique spatial configuration. By applying both neural networks and linear regression models under these scenarios, this study aims to determine the most reliable and informative approach for GWL prediction across multiple sub-watersheds. The design and comparison of these scenarios form the foundation for understanding how modelling choices influence prediction accuracy in hydrological applications. The following subsections provide a detailed explanation of each scenario.

4.3.1 Single Model

The single model scenario establishes individual models for each well using only the feature set derived from its corresponding watershed. In this configuration, the model learns localized groundwater dynamics by linking one feature set to one target well within the same region. This approach is particularly useful for capturing fine-scale hydrological behavior that may be lost in broader models. For instance, feature set F1 is used to train separate models to predict G1 and G2, both located in watershed W1, but the predictions are made independently (see Figure 4.5). A similar structure is applied in watershed W2, where

F2 is used to individually predict G3 and G4. The single model setup serves as a baseline for evaluating whether broader models (combined or aggregated) offer performance advantages.

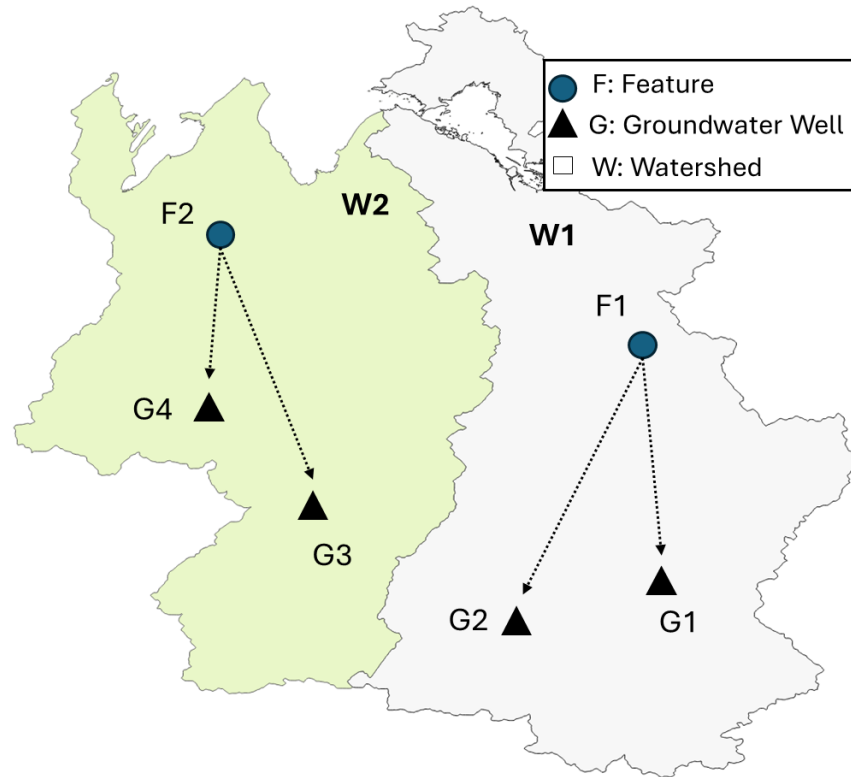


Figure 4-5- Schematic of the single model scenario, where each feature set (F) is used independently to predict groundwater well (G) levels within the same watershed (W). For example, F1 is used separately to model G1 and G2 in W1, while F2 is used to predict G3 and G4 in W2.

4.3.2 Combined Model

The combined model scenario predicts GWLs at multiple wells within the same watershed using a shared feature set. Unlike the single model setup, which develops separate models per well, the combined model treats all wells in a watershed as joint outputs of a single model. This setup evaluates whether learning from multiple wells in the same hydrological context improves generalization and reduces overfitting. For example, feature set F2 is used to simultaneously predict GWLs at G3 and G4 within watershed W2. By incorporating more than one target variable, the model leverages interdependencies among wells while still preserving the spatial relevance of inputs (see Figure 4.6).

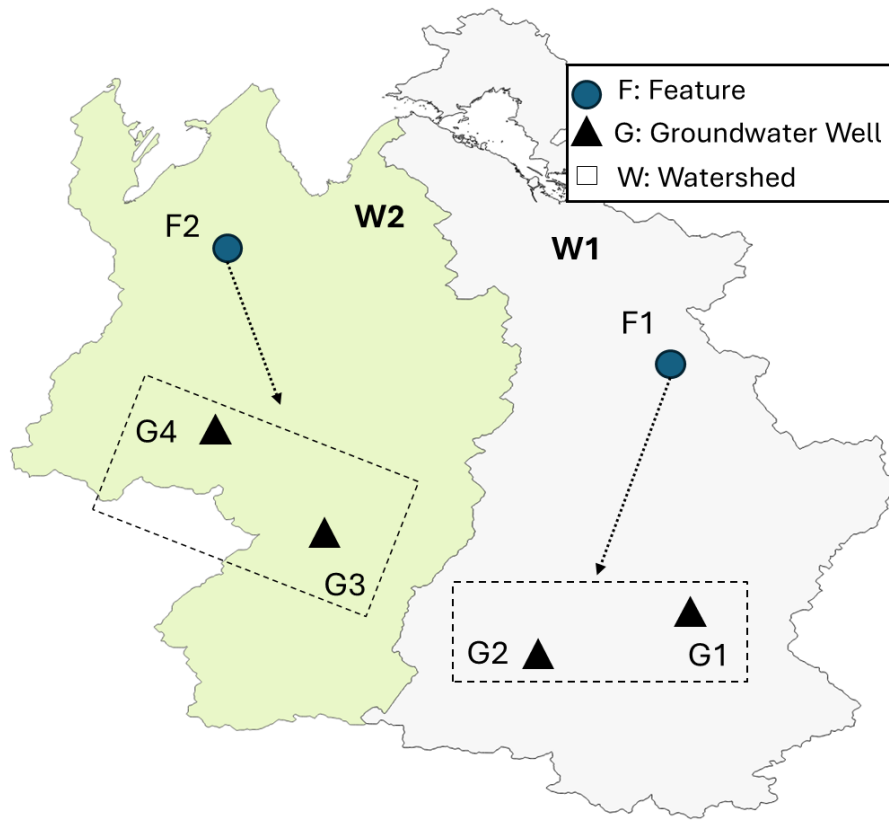


Figure 4-6-Schematic of the combined model scenario, where multiple groundwater wells (G) within the same watershed (W) are predicted simultaneously using a shared feature set (F). For instance, F2 is used to jointly predict G3 and G4 in W2.

4.3.3 Aggregated Model

The aggregated model integrates feature-well data from multiple watersheds into a single generalized model. In this approach, inputs from different watersheds (e.g., F1 from W1 and F2 from W2) are merged to predict all corresponding wells (e.g., G1 to G4) across the regions. This scenario tests the model's ability to generalize GWLs predictions across diverse hydrogeological conditions. By training on a broader dataset, the aggregated model captures shared patterns in groundwater dynamics, allowing it to perform robustly even in data-sparse settings. However, it may sacrifice some location-specific accuracy compared to single or combined models due to the increased variability in input-output relationships (see Figure 4.7).

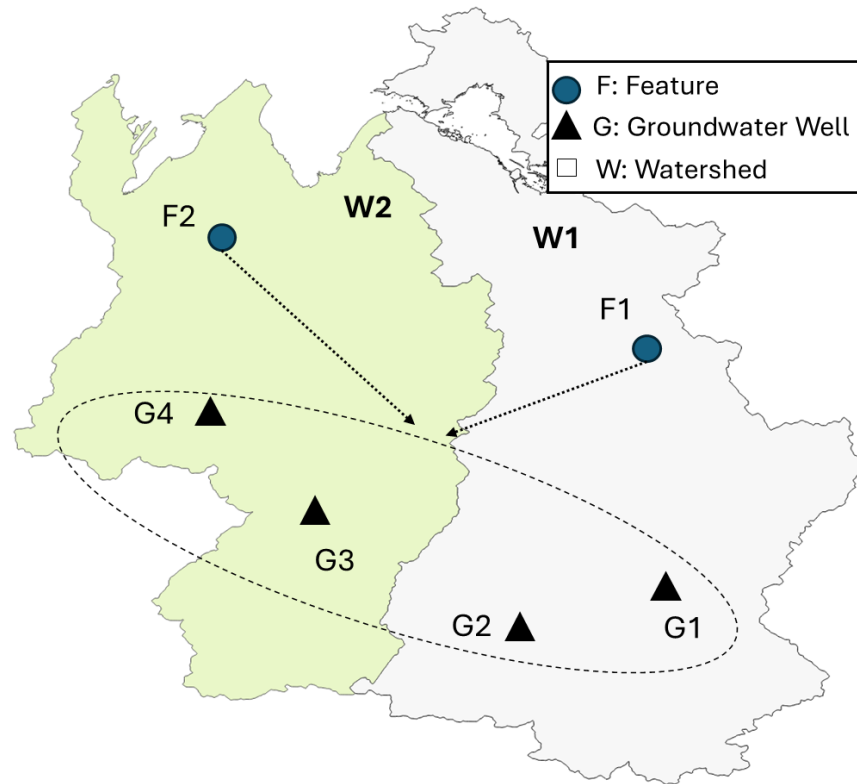


Figure 4-7-Schematic of the aggregated model scenario, where all wells (G1 to G4) across watersheds W1 and W2 are predicted using a pooled feature set (F1 and F2). This setup assesses model performance when spatial boundaries are relaxed and information is shared across regions.

4.3.4 DSM and DMS Scenarios

Two preprocessing configurations were tested to assess the effect of noise reduction on model performance: the Daily-SSA-Monthly (DSM) and the Daily-Monthly-SSA (DMS) setups. In the DSM configuration, the SSA algorithm was applied to daily time series data before monthly aggregation, helping to capture temporal trends and reduce short-term fluctuations. In the DMS configuration, the same SSA method was applied after the data were aggregated to monthly resolution. Both versions were evaluated under the single, combined, and aggregated model scenarios to investigate whether the timing of noise reduction influences prediction accuracy. The impact of SSA-based preprocessing timing was assessed only in the neural network models (Single, Combined, and Aggregated), as this approach was not applicable to the linear regression models due to assumption violations (see Section 5.4 for details).

4.4 Aquifer Detection and Data-Driven Models

This section outlines the modelling framework developed for GWL prediction, consisting of two main approaches: linear regression and neural networks. Prior to implementing these models, an Aquifer Detection Model (AQM) was developed to identify the most likely aquifer associated with each monitoring well. This step was conducted to improve geological understanding and provide supporting context for the modelling analysis. The subsections that follow first introduce the AQM process, then detail the development and evaluation of the linear and neural network models.

4.4.1 Aquifer Detection Model (AQM)

To better understand the geological setting and hydrogeological behavior of the study area, the AQM was developed to identify the most likely aquifer associated with each monitoring well. This classification provides valuable context for interpreting GWL dynamics by linking each well to its corresponding subsurface unit. By incorporating geological data into the analysis, the study ensures a more informed and spatially grounded understanding of groundwater trends across the region.

To achieve this, the AQM used three-dimensional spatial data of stratigraphic units provided by the OGS and well locations from the PGMN program to identify the nearest stratigraphic unit to each well based on spatial proximity. It incorporated both continuous and discontinuous aquifer data in comma-separated-value (.csv) for this purpose (Burt & Dodge, 2011).

The continuous and discontinuous datasets are both constructed on a 100-meter grid. The continuous data depict all 23 hydrostratigraphic units as uninterrupted surfaces across the region, ensuring full coverage of the study area. In contrast, the discontinuous data reflect the actual distributions and boundaries of these layers, providing elevation values only where the layers are confirmed to exist in borehole cores.

In this study, both continuous and discontinuous datasets were employed to improve the reliability and accuracy of the aquifer detection model. Given the uncertainty associated with borehole pick point qualities (categorized as definitive: clearly identified stratigraphic boundaries, high: reasonably clear boundaries, medium: interpreted but not directly observed boundaries, and low: ambiguous boundaries), the model identified the 10 nearest aquifer points from both datasets based on the average screening depth of each well. If a single aquifer was repeated in more than 5 out of the 10 nearest points, it was selected as the corresponding aquifer. Additionally, findings from other reports were integrated to validate the results produced by the AQM model (Cuddy, 2013; LSRCA, 2013) as shown in Figure 4.8.

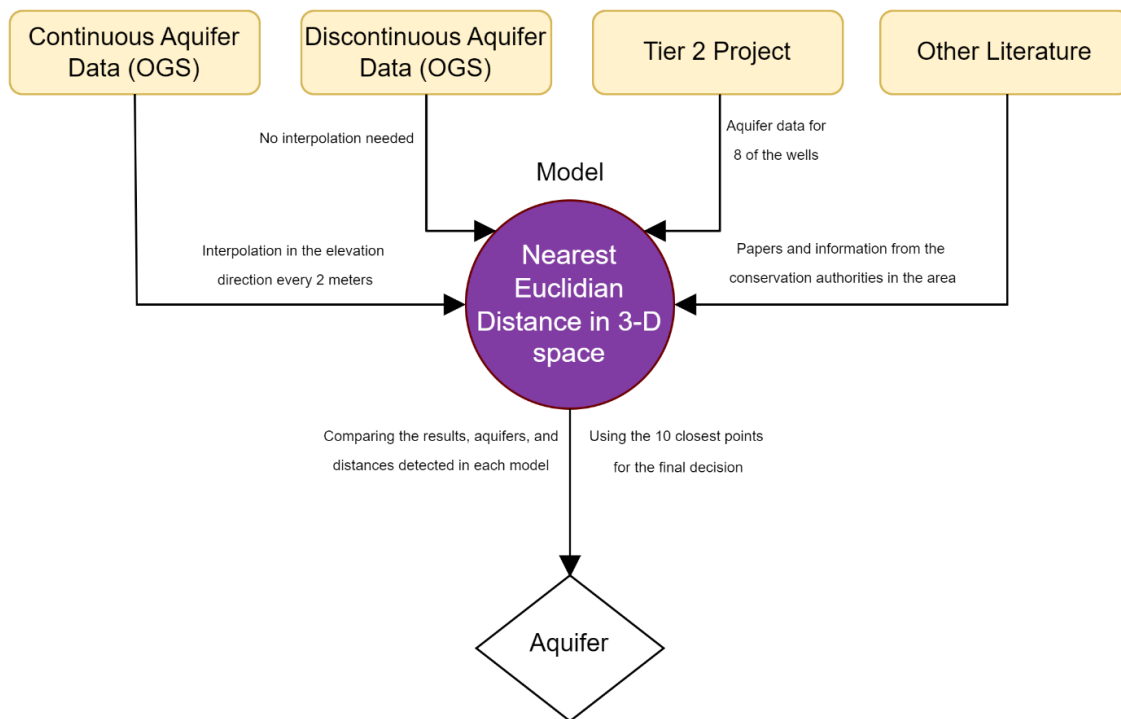


Figure 4-8-Aquifer identification workflow diagram. Aquifer assignments were determined by integrating various data sources, including continuous and discontinuous aquifer data from OGS, Tier 2 project document (Cuddy, 2013), and other literature.

The model computes the three-dimensional Euclidean distance from the aquifer points derived from the OGS continuous and discontinuous data using the following equation:

$$d = \sqrt{(X_w - X_a)^2 + (Y_w - Y_a)^2 + (Z_s - Z_a)^2} \quad (4.23)$$

Where X_w, Y_w, Z_w are the UTM coordinates (easting, northing, and adjusted screening elevation above the sea level) of the well, and X_a, Y_a, Z_a are the UTM coordinates of the aquifer points. To account for the well's interaction depth with the aquifer, the adjusted screening elevation for each well was calculated by subtracting the average screening depth from the ground elevation:

$$Z_s = Z_g - Z_{avg_w} \quad (4.24)$$

The AQM model utilizes the k-d tree algorithm to efficiently identify the 10 nearest aquifer points by partitioning the 3D space of aquifer points into hierarchical and axis-aligned bounding boxes (Bentley, 1975), which reduces the computational time to search for neighbors by deleting the irrelevant regions (Friedman et al., 1977).

4.4.2 Correlation Analysis

Detecting which lags quantities are more important in the model is important specifically in deep aquifer systems as GWL data responds slowly to the hydrometeorological features compared to shallower aquifers. Similar approach was taken by Chidepudi et al. (2025) to detect the influencing lags using auto-correlation function (ACF) and cross-correlation concepts.

4.4.2.1 Cross-Correlation Analysis

Lagged cross-correlation function (r_{xy}) was employed to analyze the temporal relationships between GWLs at each well and hydrometeorological variables, including temperature, precipitation, and hydrometric levels, as outlined by Box et al. (2015). GWL data and independent features were normalized for consistent scaling, then all feature time series were shifted forward (positive lag, $k > 0$, delaying the features) and backward (negative lag, $k < 0$, advancing the features) relative to the GWL data (following

Javadzadeh et al., 2020). The following equations describe the process of calculating cross-correlation by first determining the cross-covariance, denoted as $cov(k)$, between two data series, x and y :

$$cov(k) = \begin{cases} \text{Forward: } \frac{1}{n} \sum_{t=1}^{t=n-k} (x_t - \bar{x})(y_{t+k} - \bar{y}), & k = 0, 1, 2, \dots \\ \text{Backward: } \frac{1}{n} \sum_{t=1}^{t=n+k} (y_t - \bar{y})(x_{t+k} - \bar{x}), & k = 0, -1, -2, \dots \end{cases} \quad (4.25)$$

Where $\bar{x}$ is the mean of x and $\bar{y}$ represents the mean of y . Here, k denotes the number of time lags and n refers to the length of both time series data. Then, the cross-correlation for each time lag k is calculated based on the cross-covariance mentioned in Eq. 12:

$$r_{xy} = \frac{cov(k)}{\sigma_x \sigma_y} \quad (4.26)$$

Where σ_x and σ_y are the standard deviations of time series x and y , respectively.

4.4.2.2 Autocorrelation Analysis

In hydrology, the autocorrelation function (ACF) is used to evaluate how much a current value in a timeseries is correlated with its past values. It helps determine the strength of the linear relationship between observations over time (Machiwal & Jha, 2006). For a timeseries containing N observations, the autocorrelation coefficient at time lag k is defined as (Box et al., 2015b):

$$r_k = \frac{c_k}{c_0} \quad (4.27)$$

Where $c_k = \frac{1}{N} \sum_{t=1}^{N-k} (x_t - \bar{x})(x_{t+k} - \bar{x})$ and $c_0 = \frac{1}{N} \sum_{t=1}^N (x_t - \bar{x})^2$. As a result, the ACF coefficient could be reformulated as:

$$r_k = \frac{\sum_{t=1}^{N-k} (x_t - \bar{x})(x_{t+k} - \bar{x})}{\sum_{t=1}^N (x_t - \bar{x})^2} \quad (4.28)$$

4.4.3 Regression Analysis

Regression analysis is a widely used statistical method that reliably explores how changes in independent variables (x) influence the expected values of a dependent variable (y). It is commonly applied for prediction, forecasting, identifying causal relationships, and estimating missing data (interpolation and extrapolation techniques). The widespread use of this model stems from its ease of application and well-known techniques. The specific regression types employed in this study are described in detail in the subsequent sections (Araghinejad, 2014).

4.4.3.1 Simple Linear Regression (SLR)

The concept of regression was first introduced by (Galton, 1886), where he developed the regression line to illustrate the relationship between the heights of parents and their offspring. Linear regression is a method used to model the relationship between independent (input) variables and a dependent (output) variable in a linear manner. The simplest form of linear regression, known as simple linear regression (SLR), involves only one independent variable and one dependent variable. The equation for this approach, applied to a dataset containing N pairs of independent (x) and dependent (y) values, is given as follows (Hsieh, 2023):

$$y_i = a_0 + a_1x_i + \epsilon_i = \hat{y}_i + \epsilon_i, \quad i = 1, \dots, N \quad (4.29)$$

Where a_0 and a_1 represent the regression coefficients, $\hat{y}_i$ is the predicted value of y_i , and ϵ_i denotes the residual (error) of the prediction. The parameters in the linear regression were estimated using the Ordinary Least Squares (OLS) method. This approach determines the coefficient values by minimizing the sum of squared errors (SSE), as shown in the following equation:

$$SSE = \sum_{i=1}^{i=N} \epsilon_i^2 = \sum_{i=1}^{i=N} (y_i - \hat{y}_i)^2 \quad (4.30)$$

By substituting Equation (4.30) into Equation (4.29) for $\hat{y}_i$, and setting the partial derivatives of Equation (4.29) with respect to both parameters a_0 and a_1 to zero, the optimal values of these coefficients ($\hat{a}_0$ and $\hat{a}_1$) can be derived as follows:

$$\hat{a}_1 = \frac{\sum_{i=1}^N x_i y_i - N \bar{x} \bar{y}}{\sum_{i=1}^N x_i^2 - N \bar{x}^2} \quad (4.31)$$

$$\hat{a}_0 = \bar{y} - \hat{a}_1 \bar{x} \quad (4.32)$$

4.4.3.2 Multiple Linear Regression (MLR)

When the model includes more than one independent variable (referred to as single models in this study) to predict a single dependent variable, this form of linear regression is known as multiple linear regression (MLR). For a dataset with N measurements, j independent variables (x_j), and one dependent variable (y), the MLR formula is given as (Hsieh, 2023):

$$y_i = a_0 + \sum_{j=1}^{j=m} x_{ij} a_j + \epsilon_j = \hat{y}_i + \epsilon_i, \quad i = 1, \dots, N \quad (4.33)$$

In the vector form, the Equation (4.33) could be written as:

$$\tilde{y} = \tilde{X} \tilde{a} + \tilde{\epsilon} \quad (4.34)$$

Where $\tilde{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_N \end{pmatrix}$, $\tilde{a} = \begin{pmatrix} a_1 \\ \vdots \\ a_m \end{pmatrix}$, $\tilde{\epsilon} = \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_N \end{pmatrix}$, and matrix $\tilde{X}$ that is called the design matrix could be written as:

$$\tilde{X} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1m} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{N1} & \cdots & x_{Nm} \end{pmatrix} \quad (4.35)$$

Additionally, the OLS method minimizes the SSE of the model to obtain the optimal values for the model's coefficients:

$$SSE = \tilde{\epsilon}^T \tilde{\epsilon} = (\tilde{y} - \tilde{X}\tilde{a})^T (\tilde{y} - \tilde{X}\tilde{a}) \quad (4.36)$$

By minimizing Equation (4.36) with respect to the model coefficients, the optimal values for the coefficients (vector $\tilde{a}$) can be determined using the Equation (4.34), which can be solved through QR decomposition or the singular value decomposition (SVD) method (Hansen et al., 2013).

4.4.3.3 Multivariate Linear Regression

This model represents a generalized form of the MLR model, accommodating multiple dependent variables (combined and aggregated modelling frameworks in this study are conceptually based on this formulation). For a dataset with a sample size of N , l dependent variables, and j independent variables, the multivariate linear regression can be expressed using the following equation (Hsieh, 2023):

$$y_{ik} = a_{0k} + \sum_{j=1}^{j=m} x_{ij}a_{jk} + \epsilon_{ik} = \hat{y}_{ik} + \epsilon_{ik}, \quad i = 1, \dots, N, \quad k = 1, \dots, l \quad (4.37)$$

Using the vector format, Equation (4.37) could be expressed as:

$$\tilde{Y} = \tilde{X}\tilde{A} + \tilde{E} \quad (4.38)$$

Where $\tilde{Y} = \begin{pmatrix} y_{11} & \cdots & y_{1l} \\ \vdots & \ddots & \vdots \\ y_{N1} & \cdots & y_{Nl} \end{pmatrix}$, $\tilde{X} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1m} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{N1} & \cdots & x_{Nm} \end{pmatrix}$, $\tilde{A} = \begin{pmatrix} a_{01} & \cdots & a_{0l} \\ \vdots & \ddots & \vdots \\ a_{m1} & \cdots & a_{ml} \end{pmatrix}$, and $\tilde{E} = \begin{pmatrix} \epsilon_{11} & \cdots & \epsilon_{1l} \\ \vdots & \ddots & \vdots \\ \epsilon_{N1} & \cdots & \epsilon_{Nl} \end{pmatrix}$; and are called response matrix, design matrix, matrix of regression parameters, and residual matrix respectively.

Using the OLS method, the parameters can be determined with the following solution:

$$\tilde{A} = \tilde{X}^+ \tilde{Y} \quad (4.39)$$

Where the $\tilde{X}^+$ is the Moore-Penrose inverse of $\tilde{X}$:

$$\tilde{X}^+ = (\tilde{X}^T \tilde{X})^{-1} \tilde{X}^T \quad (4.40)$$

4.4.4 Artificial Neural Networks (ANNs)

ANNs have demonstrated strong performance in modelling complex, nonlinear relationships in environmental and hydrological datasets, often outperforming traditional statistical methods (Rajaei et al., 2019). Due to their flexibility and learning capability, they are well-suited for predicting GWL dynamics. In the following subsections, we describe the architecture of Feed-forward Neural Networks (FFNNs), explain how these models function, and outline the training and validation techniques used in this study, including k-fold cross-validation, random seed initialization, and hyperparameter tuning techniques implemented.

4.4.4.1 Feed-forward Neural Networks (FFNNs)

FFNN, a type of multi-layer perceptron (MLP) network, is the simplest and most utilized type of ANNs for predicting GWL (Ahmadi et al., 2022; Uc-Castillo et al., 2023). In this architecture, the data flows in a single direction from the input layer to the output layer, passing through hidden layers if they are present. An MLP network containing one input, hidden, and output layer is called a two-layer network as shown in Figure 4.9.

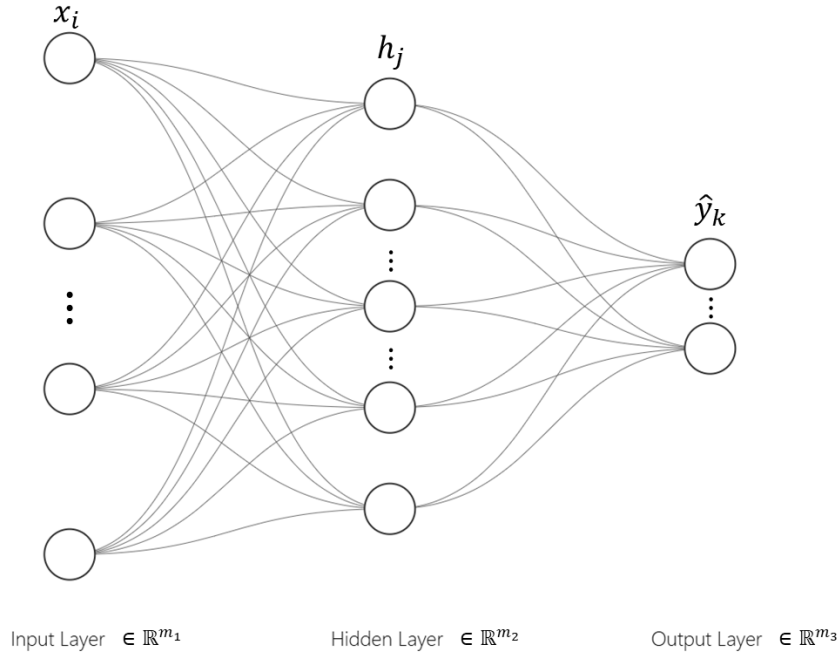


Figure 4-9-Schematic architecture of a FFNN with one hidden layer. The network consists of an input layer $x_i \in \mathbb{R}^{m_1}$, a hidden layer $h_j \in \mathbb{R}^{m_2}$, and an output layer $\hat{y}_k \in \mathbb{R}^{m_3}$. Each neuron in one layer is connected to all neurons in the next layer, illustrating the fully connected structure

According to Figure 4.9, For m_1 independent variables ($x_i: i \in \{1, \dots, m_1\}$), m_2 neurons in the hidden layer ($h_j: j \in \{1, \dots, m_2\}$), and m_3 outputs or dependent variables in the output layer ($\hat{y}_k: k \in \{1, \dots, m_3\}$), the following equations define the relationships between the hidden and input layers:

$$h_j = f \left(\sum_{i=1}^{m_1} w_{ji} x_i + b_j \right) \quad (4.40)$$

Where w_{ji} is the weight connecting the j th neuron in the hidden layer to the i th independent variable (x_i) in the input layer, b_j denotes the bias parameter for the j th neuron in the hidden layer, and f represents the activation function applied to the weighted sum of the inputs and biases.

Additionally, the relationship between the output and hidden layer is expressed as:

$$\hat{y}_k = g \left(\sum_{j=1}^{m_2} \tilde{w}_{kj} h_j + \tilde{b}_k \right) \quad (4.41)$$

Here, $\tilde{w}_{kj}$ is the weight connecting the k th neuron in the output layer to the j th node in the hidden layer, $\tilde{b}_k$ shows the offset parameter for the k th node in the output layer, and g is the activation function applied to the weighted sum of nodes in the hidden layer and biases.

Using activation functions (f and g functions) can introduce non-linearity in the system so that the model is more powerful in capturing complex relationships in the data (Fernandes et al., 2024). The ReLU activation function has been extensively utilized in GWL models (Müller et al., 2021). The ReLU function outputs the input directly if it is positive and zero otherwise, as defined by:

$$f(x) = \max(0, x) \quad (4.42)$$

In this study, the ReLU activation was applied to the hidden layer neurons to improve convergence speed and minimize vanishing gradient issues during training. This means that any negative input values to the hidden layer are replaced with zero, introducing sparsity while maintaining non-linear behavior in the network.

4.4.5 Model Training and Validation Techniques

To ensure that the ANN models developed in this study can generalize well to unseen data, careful training and validation procedures were followed. This section describes the key steps used to optimize model performance and avoid overfitting, including the choice of loss function, the use of back-propagation for parameter updates, and evaluation metrics. Furthermore, we employed cross-validation, fixed random seed initialization, and systematic hyperparameter tuning to enhance model reliability and reproducibility.

4.4.5.1 Model Loss Function and Back-Propagation

Model's performance is assessed using the mean squared error (MSE) during training, which, in academic terms, serves as the loss function for the system. As the data reaches the output nodes and creates

the predicted target ($\hat{y}_i$), it is compared with the actual target (y_i), and the loss function is calculated as follows:

$$L = \frac{\sum_{i=1}^{N_d} (\hat{y}_i - y_i)^2}{N_d} \quad (4.43)$$

Where N_d accounts for the number observations in the training and validation dataset.

The aim is to minimize this loss function by choosing appropriate weights and bias parameters. To do so, Adam back-propagation method was chosen in the ANN architecture to optimize the weights and bias parameters in each modelling scenario. This optimization process was implemented in Python using the TensorFlow/Keras library, which automates the back-propagation and adaptive learning rate adjustments.

4.4.5.2 Data Partitioning

The dataset was partitioned into 60% for training, 20% for validation, and 20% for testing, maintaining this split across all scenarios to ensure a fair comparison between models, as the same proportional data split was applied consistently to each scenario despite differences in the total number of observations. This partitioning scheme was selected because it is among the most commonly used data splits in GWL prediction studies, allowing sufficient data for model training while preserving independent validation and testing sets (Ahmadi et al., 2022; Rajaei et al., 2019; Uc-Castillo et al., 2023). The final 20% of the data, representing the most recent observations of inputs and outputs, was consistently used for testing. The remaining data was allocated for training and validation. In addition, data splitting performed chronologically to preserve the sequential nature of the data following Nourani et al. (2022), where they used LSTM models without shuffling the observations.

4.4.5.3 K-Fold Cross-Validation (KFCV)

K-fold Cross-Validation (KFCV), a widely used cross-validation technique, was selected to address the uncertainty associated with splitting calibration data and reduce biases that might occur when the model has limited observations, especially when choosing which data portion to use for validation (Arabameri et al., 2021; Bennett et al., 2013; Hastie et al., 2009b). As shown in Figure 4.10, which presents the monthly normalized GWL time series in well 293-2, all observations are divided into K partitions (folds), where the final fixed fold (comprising the most recent 20% of the data) is reserved for testing the model, and the remaining $K - 1$ folds are used for model calibration.

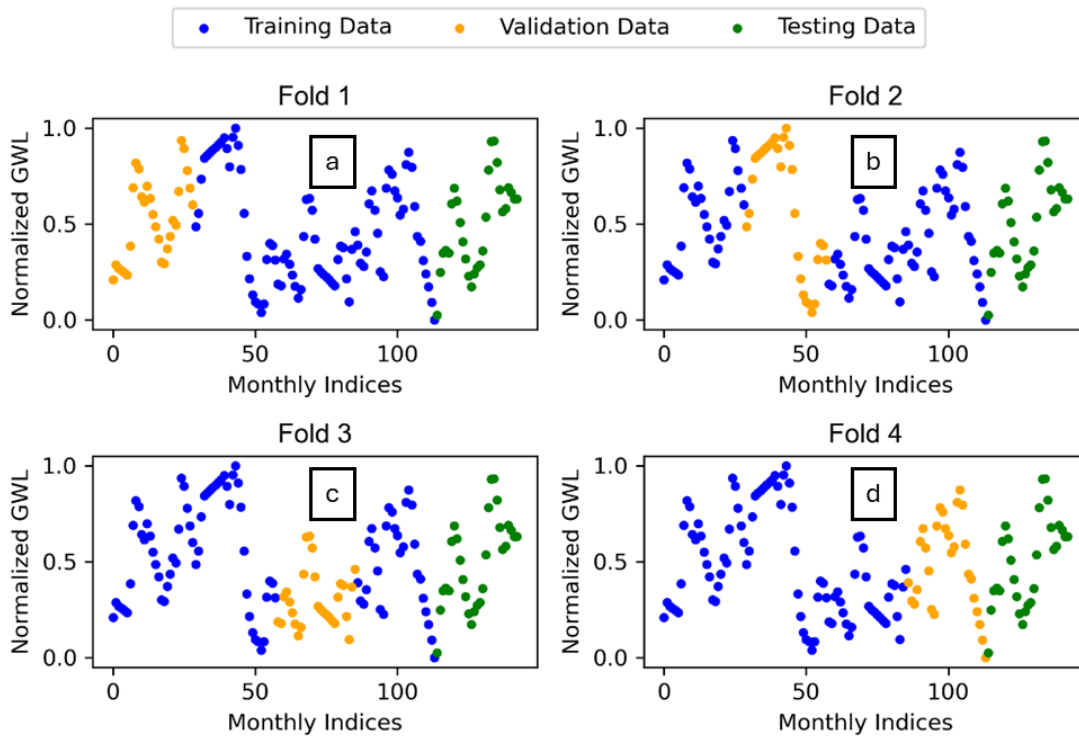


Figure 4-10- K-fold cross-validation (KFCV) used to partition the normalized monthly GWL data from well 293-2. Each panel (a–d) illustrates a different fold where one segment is used for validation (orange) and the remaining segments for training (blue), while the most recent 20% of the data (green) is fixed as the testing set.

Within the calibration dataset, the validation fold shifts sequentially from the first partition to the last (also shown in Table 4.1), ensuring that the model is validated on the entirety of the calibration data during the training process (Hsieh, 2023; Snieder et al., 2020a). A similar approach is implemented for all other scenarios and plots are provided in Appendix B.

Table 4.1-K-fold cross-validation (KFCV) framework illustrating the data partitioning strategy across multiple iterations. In each iteration, one fold is used for validation, the last fold is fixed for testing, and the remaining folds are used for training.

Iteration	Fold				
	1	2	3	4	5
1	Validation	Training	Training	Training	Testing
2	Training	Validation	Training	Training	Testing
3	Training	Training	Validation	Training	Testing
4	Training	Training	Training	Validation	Testing
⋮	⋮	⋮	⋮	⋮	⋮
n	Training	Training	Training	Validation	Testing

4.4.5.4 Random Seed for Multi-Start Initialization Approach

The Adam optimizer updates the weights and biases of the ANN structure iteratively and adjusts parameters (weights and biases), which minimizes the model's loss function. Variability in the outputs and predictions could be due to convergence to different local minima because of uncertainty in the original parameter initialization. To reduce this intrinsic variability, the model is run fifty times using fifty different seed values randomly sampled from the interval $[0, 1000]$. To be able to assign different values to seeds in advance of running the NNs, random and Tensorflow libraries in Python programming language were used. This process, named as multi-start initialization, will provide a quantitative estimate of the uncertainty in

the model. Snieder et al. (2020b) also followed a similar approach while developing neural networks for flood forecasting using various input variable selection (IVS) methods.

Figure 4.11 illustrates the distribution of seed values used in the multi-start initialization approach in the single modelling scenarios. The graph represents the frequency of randomly selected seed values.

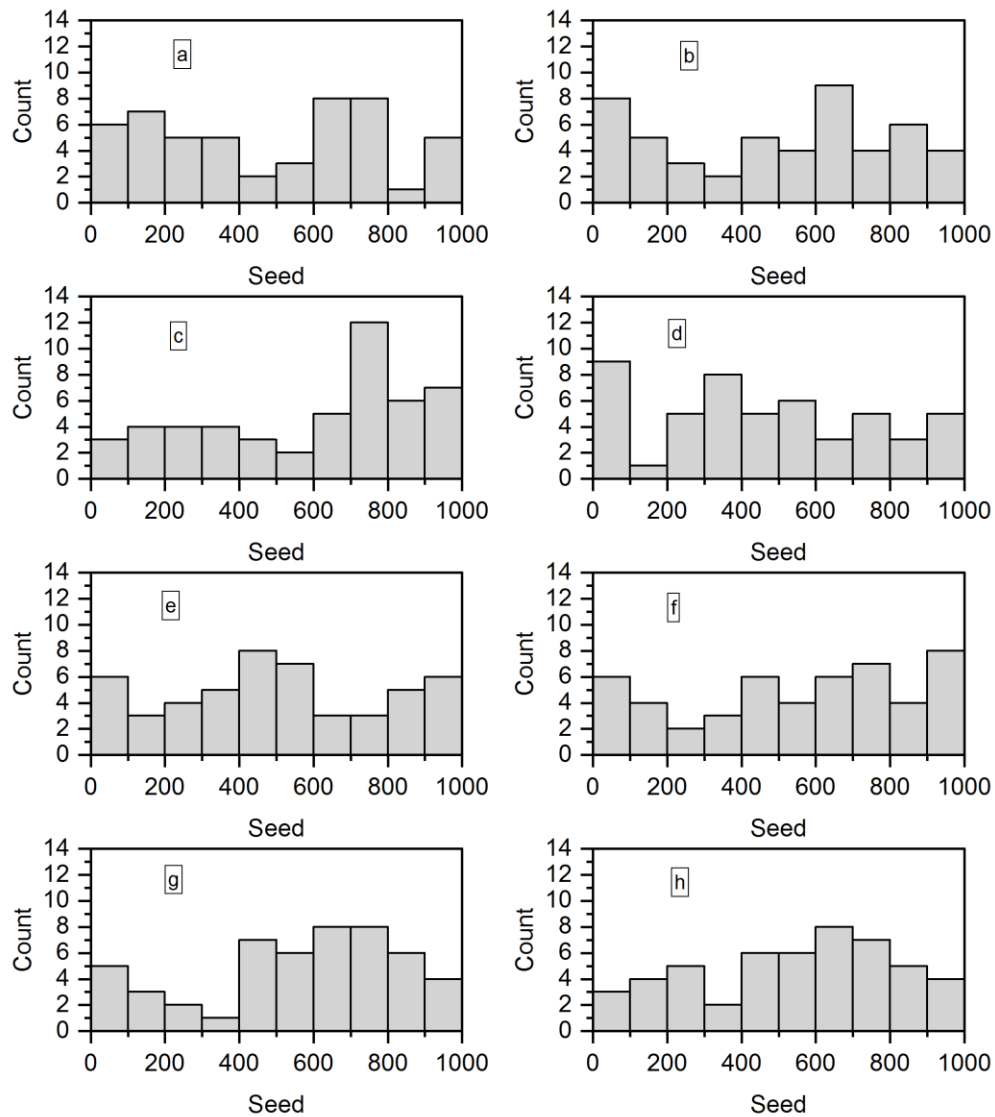


Figure 4-11-Distribution of randomly selected seed values used in the multi-start initialization process for single modelling scenarios. Each subplot corresponds to a specific monitoring well: (a) 293-2, (b) 293-3, (c) 439-1, (d) 440-1, (e) 442-1, (f) 244-2, (g) 245-2, and (h) 443-1.

4.5 Hyperparameter Tuning

Selecting appropriate hyperparameters in ANNs is critical, as they significantly influence model behavior and performance. To ensure reliable modelling, it is essential to either adopt values supported by previous research or evaluate the model under various hyperparameter settings. This process, known as hyperparameter tuning, allows for comparison across different configurations to identify the most effective setup. Given the high flexibility of neural networks, careful tuning helps achieve optimal results.

4.5.1 Tuning Method

Hyperparameter tuning involves systematically exploring combinations of model settings to find the configuration that yields the best performance. Two commonly used methods are:

- Grid Search: This brute-force method tests all possible combinations of specified hyperparameters. Though exhaustive, it is computationally expensive.
- Random Search: This method randomly samples from the hyperparameter space and is more efficient for large search spaces (Bergstra & Bengio, 2012).

In this study, random search was selected due to its computational efficiency and capability to identify promising hyperparameter regions without exhaustively testing all combinations.

4.5.2 Hyperparameters for Linear Regression

Linear regression typically does not require extensive hyperparameter tuning. However, in regularized forms like Ridge and Lasso regression, the regularization parameter (λ) plays a critical role in controlling overfitting by penalizing large coefficients. In this study, the ordinary least squares (OLS) version was used, and no additional hyperparameter was tuned beyond ensuring assumptions were met (refer to the residual diagnostics Section 5.4).

4.5.3 Hyperparameters for Neural Networks

Varying the number of hidden layers and the neurons within them—typically determined through trial and error—can impact model performance. Several hyperparameters were explored for the ANN models:

- **Number of Hidden Layers and Neurons:** Set through trial and error, with one hidden layer typically being sufficient for hydrological problems. A single hidden layer is understood to be sufficient for most hydrological applications (Maier & Dandy, 2000).
- **Learning Rate:** Controls the step size during gradient descent. A fixed learning rate of 0.001 was used across all models, which is commonly considered a stable default for neural networks.
- **Batch Size:** Number of samples processed before the model is updated. Affects training stability and speed.
- **Epochs and Patience:** The number of training epochs was predefined, with early stopping implemented to avoid overfitting. The training process was terminated if the validation loss did not improve over 10 consecutive epochs.
- **Activation Function:** ReLU was selected for hidden layers based on its widespread use and performance in GWL modelling (Müller et al., 2021).
- **Weight Initialization & Optimizer:** Weights were initialized randomly using a random initialization method, while the Adam optimizer was employed to ensure stable and efficient convergence during training.

4.5.4 Validation and Testing During Tuning

During the training process, the model's performance must be assessed using an independent dataset, known as validation data, which is not utilized during training (Maier & Dandy, 2000). The

evaluation metrics are computed continuously for both the training and validation datasets, and the training process is halted under two conditions:

- **Early stopping:** if the model's performance (MSE) does not improve for a specified number of consecutive steps – referred to as patience – the training is terminated. In this study, the patience value is set to 10. If the validation MSE does not improve over 10 consecutive steps, training stops, and the best weight and bias parameters are stored.
- **Maximum Epochs Reached:** If the model continues to improve but reaches the predefined maximum number of epochs, training is stopped, regardless of ongoing performance gains on the validation set.

4.6 Evaluation Metrics

Since various metrics capture different model properties, best practices in hydrological modelling recommend utilizing multiple performance indicators to evaluate model's performance (Ewen, 2011; Maier et al., 2010; Snieder et al., 2020). In this study, the coefficient of determination (R^2) and mean squared error (MSE) were used as statistical indicators to evaluate the model's performance. Additionally, the MSE served as the objective function during the calibration process. The R^2 measures the proportion of variance in the observed GWLs explained by the model, while the MSE quantifies the average squared difference between the observed and predicted values, indicating the model's error magnitude. These indicators are widely used in GWLforecasting to assess model accuracy and reliability (Ahmadi et al., 2022; Boo et al., 2024; Uc-Castillo et al., 2023).

The MSE, used to assess the performance of the models during the training, validation, and testing stages, can be expressed as:

$$MSE = \frac{\sum_{i=d}^{N_d} (\hat{y}_i - y_i)^2}{N_d} \quad (4.44)$$

Where $\hat{y}_i$ represents the GWL value predicted by the MLP model, y_i is the actual observed GWL, and N_d denotes the total number of observations across the training, validation, testing, or the entire dataset. The Mean Squared Error (MSE) can range from 0 to infinity, where values closer to 0 indicate that the model's predictions are highly accurate, as the errors are minimal. On the other hand, higher MSE values suggest poorer model performance, with larger discrepancies between the predicted and actual values (Hastie et al., 2009a).

The coefficient of determination (R^2) is another key metric used to evaluate the model's performance. It is expressed as:

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}} = 1 - \frac{\sum_{i=1}^{N_d} (\hat{y}_i - y_i)^2}{\sum_{i=1}^{N_d} (\bar{y} - y_i)^2} \quad (4.45)$$

where the SS_{res} represents the residual sum of squares which quantifies the error between the observed and predicted value and SS_{tot} is the total sum of squares that measures the total variance in the observed data relative to its mean ($\bar{y}$). The R^2 value ranges between $-\infty$ and 1. A value of $R^2 = 1$ indicates a perfect fit where the model explains 100% of the variance in the data. An R^2 value of 0 suggests the model does no better than simply predicting the mean of the observed values. Negative R^2 values occur when the model performs worse than the mean prediction (Draper & Smith, 1998).

In addition to R^2 and MSE, the Akaike Information Criterion (AIC) was used as a complementary metric to evaluate model fit while accounting for model complexity. Unlike R^2 and MSE, which focus solely on predictive accuracy, AIC introduces a penalty term proportional to the number of estimated parameters, thereby discouraging overly complex models. The AIC is defined as:

$$AIC = 2k - 2 \ln(L) \quad (4.46)$$

where k denotes the number of estimated model parameters and L represents the maximum likelihood of the model. Lower AIC values indicate a better trade-off between goodness-of-fit and parsimony. In the context of neural network models, k corresponds to the total number of trainable weights

and biases. Therefore, AIC enables comparison of architectures with different neuron counts by balancing accuracy improvements against increases in model complexity (Akaike, 1974).

4.7 Sensitivity Analysis (SA) Methods

Sensitivity Analysis (SA) is a critical step in model interpretation that quantifies how changes in input features affect the model output. In this study, two different SA methods were applied: SHAP (SHapley Additive exPlanations) and One-At-a-Time (OAT).

4.7.1 SHAP

SHAP is a game-theoretic approach to explain the output of ML models. It is based on Shapley values from cooperative game theory, which attribute a fair contribution of each feature to the final model prediction. For a model f and a set of features $x = \{x_1, x_2, \dots, x_n\}$, the SHAP value ϕ_i for feature x_i is computed as:

$$\phi_i = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|! (n - |S| - 1)!}{n!} [f(S \cup \{i\}) - f(S)] \quad (4.47)$$

where:

- N is the set of all features,
- S is a subset of features not containing i ,
- $f(S)$ is the model prediction using only features in S .

SHAP values satisfy local accuracy, missingness, and consistency, making them reliable for interpreting both individual predictions and global feature importance (Lundberg & Lee, 2017).

4.7.2 One-At-a-Time (OAT)

The One-At-a-Time (OAT) method is a simple but effective global sensitivity analysis technique. It involves perturbing one input variable at a time while keeping all other inputs constant and observing the

resulting change in model output. This helps identify the relative impact of each input variable on the prediction. Mathematically, the sensitivity index S_i for feature x_i is calculated as:

$$S_i = \frac{|f(x_1, \dots, x_i + \Delta x_i, \dots, x_n) - f(x_1, \dots, x_i, \dots, x_n)|}{\Delta x_i} \quad (4.48)$$

where:

- Δx_i is a small change in the value of feature x_i ,
- f is the trained model.

This approach is intuitive and easy to implement, but it does not account for interaction effects between variables (Saltelli et al., 2004).

Chapter 5 Results

5.1 Introduction to Results

This chapter presents the results of the analysis conducted on the GWL data from Southern Ontario. It begins with an exploratory analysis to understand the characteristics and distribution of the GWL data in Southern Ontario. Subsequently, the following sections present the main findings derived from data-driven modelling of GWL data in the BOM area.

5.2 Exploratory Data Analysis Results

Within this section, the explanatory analysis results are presented to better understand the behavior of GWL data in Southern Ontario. To achieve a comprehensive understanding of GWL variations, it is important to examine the GWL datasets and also to explore related hydrological and meteorological factors that may influence GWL trends. Therefore, the analyses include assessments of precipitation, temperature, and hydrological levels in both independent and comparative contexts.

This section begins by evaluating data availability, covering meteorological data scarcity across Southern Ontario and in specific areas such as the BOM area. It then proceeds to explore the spatial filtering of ERA5 climate nodes and trends observed in GWL measurements. Subsequent subsections investigate slope distributions in both overburden and bedrock units and analyze spatial and seasonal patterns in the GWL data. SSA is conducted on both daily and monthly scales to detect periodic signals. Additional attention is given to the relationship between GWL and meteorological variables, followed by the calculation of Pearson correlation coefficients for individual GWL time series.

Finally, the section concludes with the results of a correlation analysis and the implementation of the AQM, which is designed to classify groundwater behavior and contribute to a more informed interpretation of subsurface hydrological processes.

5.2.1 Meteorological Data Scarcity Analysis in Southern Ontario

Before performing any statistical analysis or modelling, it is essential to assess the completeness of the raw meteorological data as the inputs in the models. Temperature and precipitation records from ECCC stations located across Southern Ontario were examined to evaluate the extent of missing data over time. Given that missing values can introduce bias, reduce model accuracy, or limit trend detection, understanding the spatial and temporal distribution of data gaps is a critical first step.

To visually assess the availability of temperature data, Figure 5.1 presents a binary representation of daily observations across all available stations in Southern Ontario from 2000 to 2023. Each vertical strip corresponds to a unique climate ID (station), and each row indicates a calendar date. To maintain visual clarity and prevent overcrowding, exact date labels and full seven-digit Climate IDs (Environment and Climate Change Canada, n.d.) were omitted from the graph. Black segments indicate dates with available data, while white segments show missing entries. The figure reveals a wide variation in data completeness across stations – some exhibit nearly full coverage, while others are dominated by long or frequent gaps.

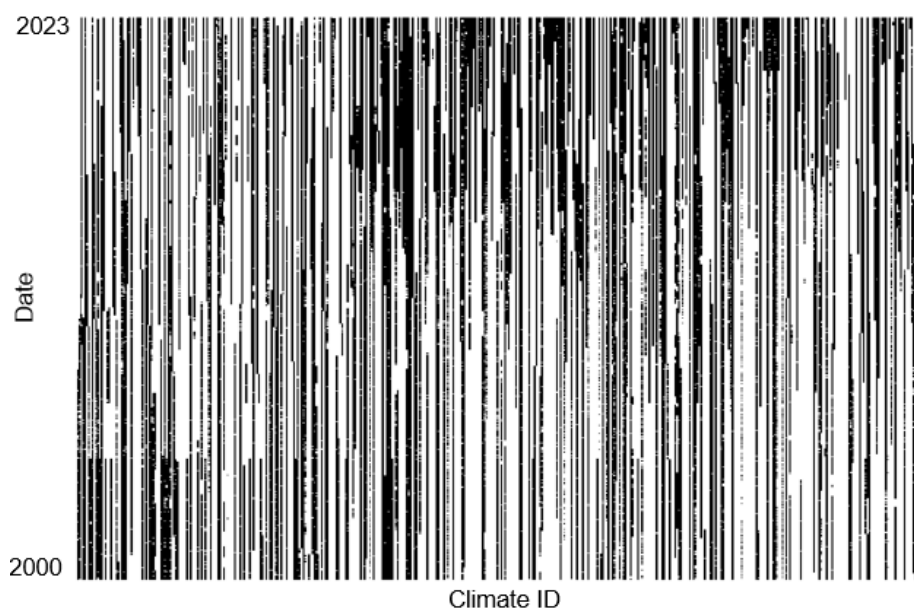


Figure 5-1- Binary visualization of daily temperature data availability across Southern Ontario stations (2000–2023). Each vertical line represents a unique station (Climate ID), and each horizontal line corresponds to a calendar date. Black segments indicate available data, while white gaps represent missing entries.

To gain a clearer understanding of data incompleteness in the temperature dataset, the proportion of missing values (NaN ratio) was calculated for each station. As shown in Figure 5.2, a significant number of stations exhibit extremely high NaN ratios, with bars approaching a value of 1. A NaN ratio of 1 indicates that all data points for a station are missing during the study period. This indicates that many stations were nearly or entirely out of service during the study period, rendering their records unsuitable for statistical analysis or predictive modelling.

In addition to identifying fully inactive stations, this figure also highlights the broad variability in data availability across the network. While some stations maintain relatively low NaN ratios, suggesting more reliable records, others demonstrate inconsistent data presence that may still require interpolation or exclusion depending on the intended use.

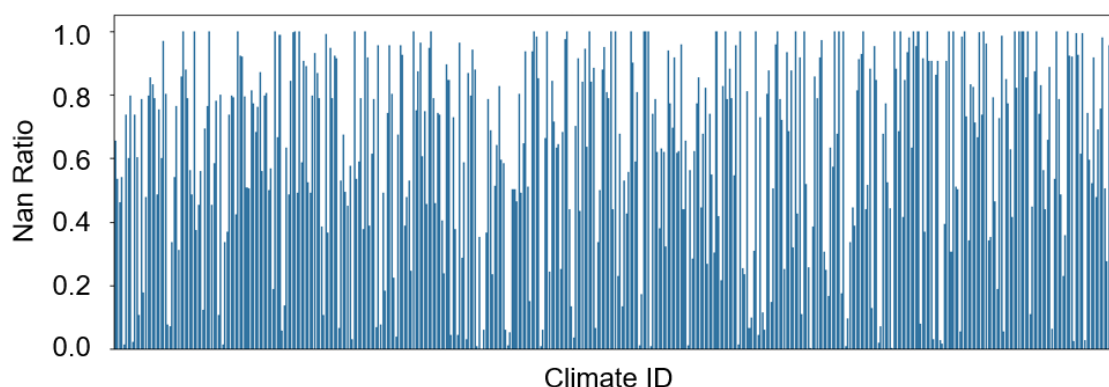


Figure 5-2- Proportion of missing temperature values (NaN ratio) per station across Southern Ontario (2000–2023). Each bar represents a station's ratio of missing daily observations, with values approaching 1 indicating nearly complete data loss and values near 0 reflecting high data completeness.

To numerically summarize and interpret the distribution of data availability, Figure 5.3 classifies the stations into five categories based on their NaN ratios. Stations with no recorded data were excluded from this classification, as they lack the completeness required for any meaningful analysis. The pie chart in Figure 5.3 shows that 36.5% of climate IDs had missing data for more than 75% of the time period under study, while another 25.7% fell between 50% and 75% missing. In other words, over 60% of the 408 climate

IDs suffered from severe data scarcity – IDs with more than 50% incompleteness between 2000 and 2023. On the other hand, only 11.8% of stations had less than 10% missing values, indicating a relatively small portion of highly reliable IDs. The overall average NaN ratios across all stations is 57%, reflecting a general trend of limited data completeness in the raw temperature dataset across Southern Ontario.

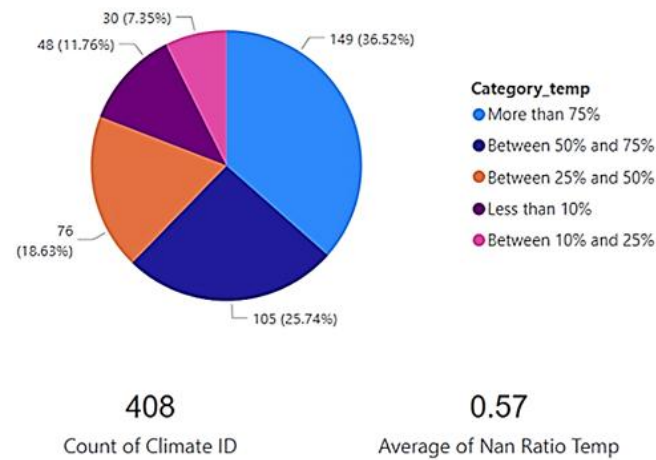


Figure 5-3- Distribution of climate stations in Southern Ontario based on NaN ratio categories (2000–2023). The chart groups 408 stations into five categories according to the percentage of missing temperature data.

A similar approach was applied to assess the completeness of daily precipitation records across Southern Ontario. Figure 5.4 presents a binary visualization of missing and available precipitation data for each station from 2000 to 2023. As with the temperature graph, each vertical bar corresponds to a Climate ID, while each horizontal row represents a specific date. Black segments indicate days with recorded precipitation values, and white gaps denote missing entries. The visual pattern reveals a comparable level of inconsistency in precipitation data coverage, with some stations maintaining dense and continuous records, while others experience frequent and prolonged data gaps.

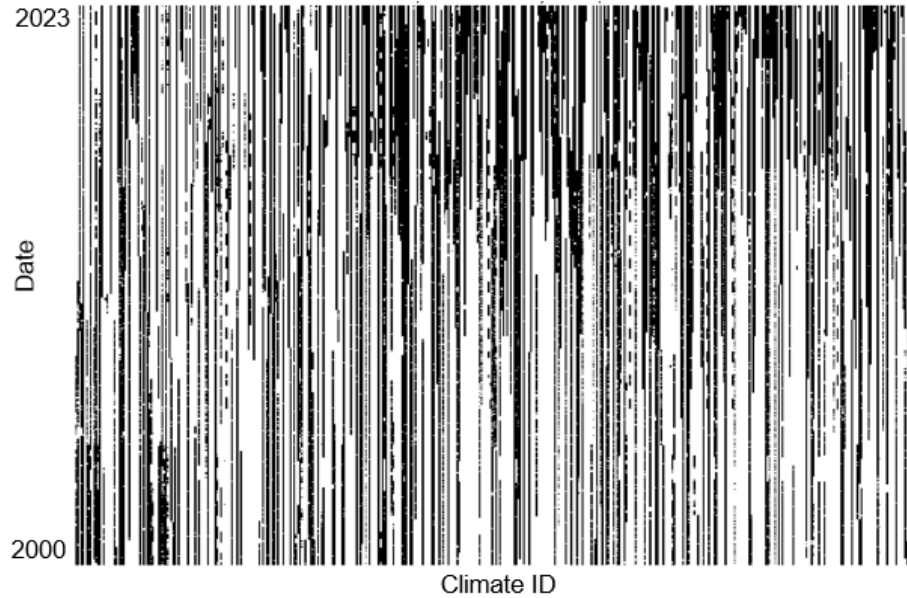


Figure 5-4- Binary visualization of daily precipitation data availability across Southern Ontario stations (2000–2023). Each vertical line represents a unique station (Climate ID), and each horizontal line corresponds to a calendar date. Black segments indicate available data, while white gaps represent missing entries.

Building on the temperature analysis, Figure 5.5 presents the NaN ratios for each station in the precipitation dataset. Similar patterns emerge, with a large portion of stations exhibiting high levels of missing data – many with NaN ratios close to 1. While some stations offer more reliable data with lower NaN ratios, the overall trend indicates the need for careful station selection before incorporating precipitation data into modelling workflows.

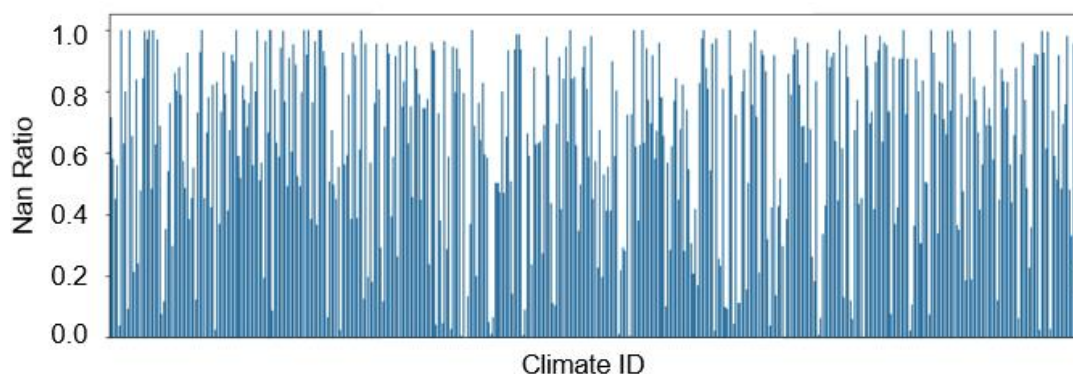


Figure 5-5- Proportion of missing precipitation values (NaN ratio) per station across Southern Ontario (2000–2023). Each bar represents a station's ratio of missing daily observations, with values approaching 1 indicating nearly complete data loss and values near 0 reflecting high data completeness.

To better interpret the distribution of missing precipitation data, Figure 5.6 categorizes the Climate IDs based on their NaN ratios in numbers. Before categorizing the ratios, stations with complete missing observations were excluded similar to temperature analysis. In this case, 438 stations had usable precipitation data with NaN ratios below 1, which is slightly more than the 408 stations available for the temperature analysis. Among these 438 stations, a substantial portion, 35.9%, had more than 75% missing data, while another 28.0% fell within the 50-75% range. In contrast, only 8.9% of the stations had less than 10% missing values, indicating a relatively small subset of highly reliable records.

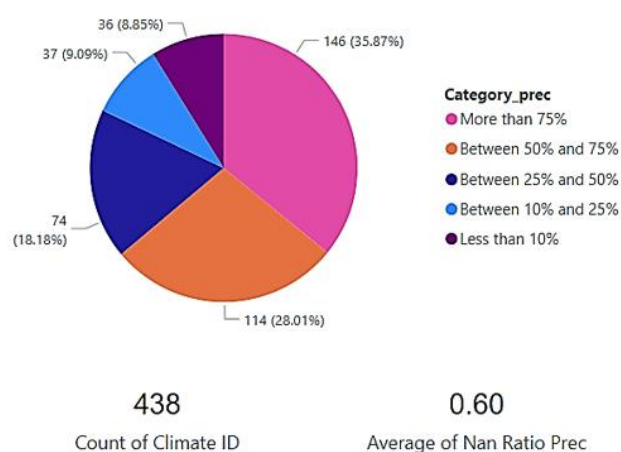


Figure 5-6-Distribution of Climate IDs by precipitation data completeness in Southern Ontario from 2000 to 2023, categorized by NaN ratio ranges.

The analysis of meteorological data from ECCC stations across Southern Ontario shows considerable variability in data availability. A significant number of Climate IDs suffer from high levels of missing values, with many exceeding 75% incompleteness – as shown in the pie charts. This inconsistency underscores the need for further preprocessing and interpolation, as well reducing the area size from Southern Ontario to BOM for prediction, which are addressed in the following sections.

5.2.2 Data Scarcity Analysis in Barrie-Oro Moraine

This section examines the availability and completeness of temperature and precipitation data from ECCC weather stations located within and around the BOM region. While not all stations fall strictly within the BOM boundary, they were selected based on their proximity to the area.

Figure 5.7 represents a visual summary of the daily data completeness from 2000 to 2023. It provides a binary view of the dataset's temporal gaps (similar to the previous section), with black indicating available records and white indicating missing values. In addition, the spatial locations of the stations within the BOM area are shown in Figure 3.3, located in Section 3.1.3.

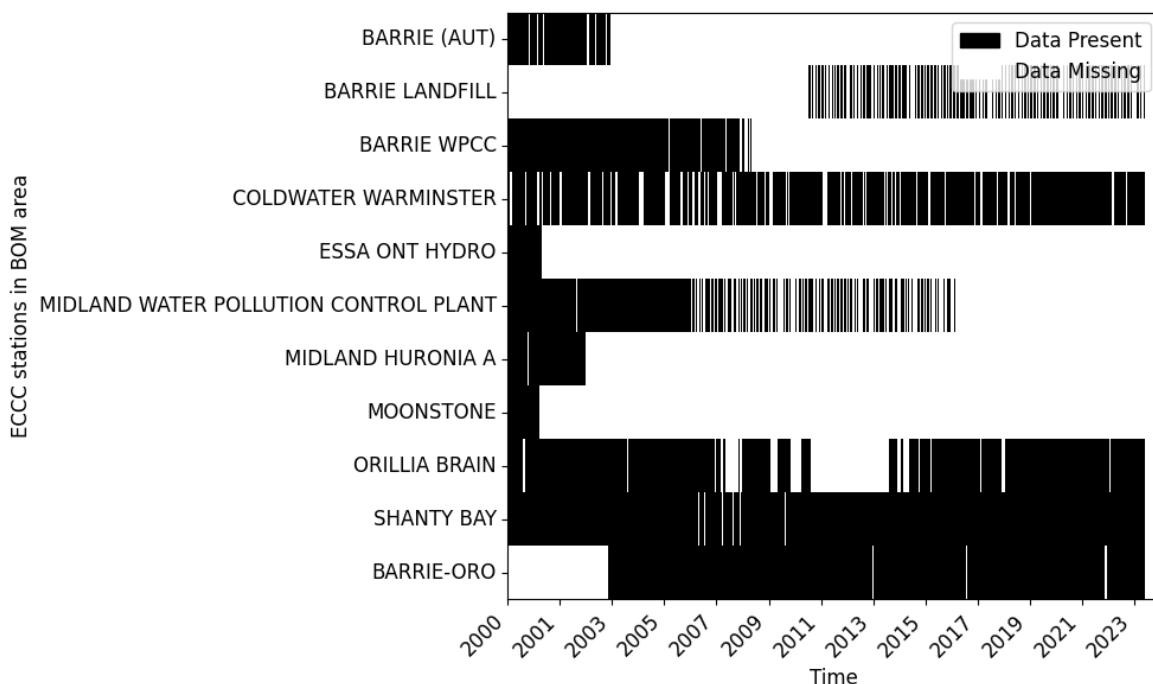


Figure 5-7-Availability of daily temperature observations from ECCC stations in the BOM area (2000–2023). Black segments indicate dates with available data, while white gaps represent missing entries.

The extent of data incompleteness varies significantly across the ECCC stations. Some stations experienced extensive data loss throughout the 20-year study period. For instance, Moonstone and Essa Ont Hydro had missing temperature data for 95.1% and 94.7% of the days between 2000–2023, respectively. Similarly, Midland Huronia A and Barrie (AUT) also exhibited high rates of data absence, with over 85% of their records missing. In contrast, several stations had more complete datasets. Notably, Shanty Bay and Coldwater Warminster had relatively low data incompleteness, with only 3.5% and 14.45% of days missing, respectively.

The precipitation data also shows substantial variation in completeness. As visualized in Figure 5.8, several stations experienced high levels of missing data across the study period. In particular, Moonstone and Essa Ont Hydro had over 94% of their daily precipitation records missing, while Barrie (AUT) and Midland Huronia A exhibited similarly poor coverage, each missing more than 88% of the data.

On the other hand of the spectrum, Shanty Bay and Coldwater Warminster retained relatively complete datasets, with only 3.5% and 13.1% of their entries missing, respectively. The remaining stations, including Barrie WPCC, Barrie Landfill, and Midland Water Pollution Control Plant, displayed moderate gaps, with missing values ranging from 51.9% to 64.5%. These findings further emphasize the inconsistent data availability in the region, which could hinder reliable GWL prediction.

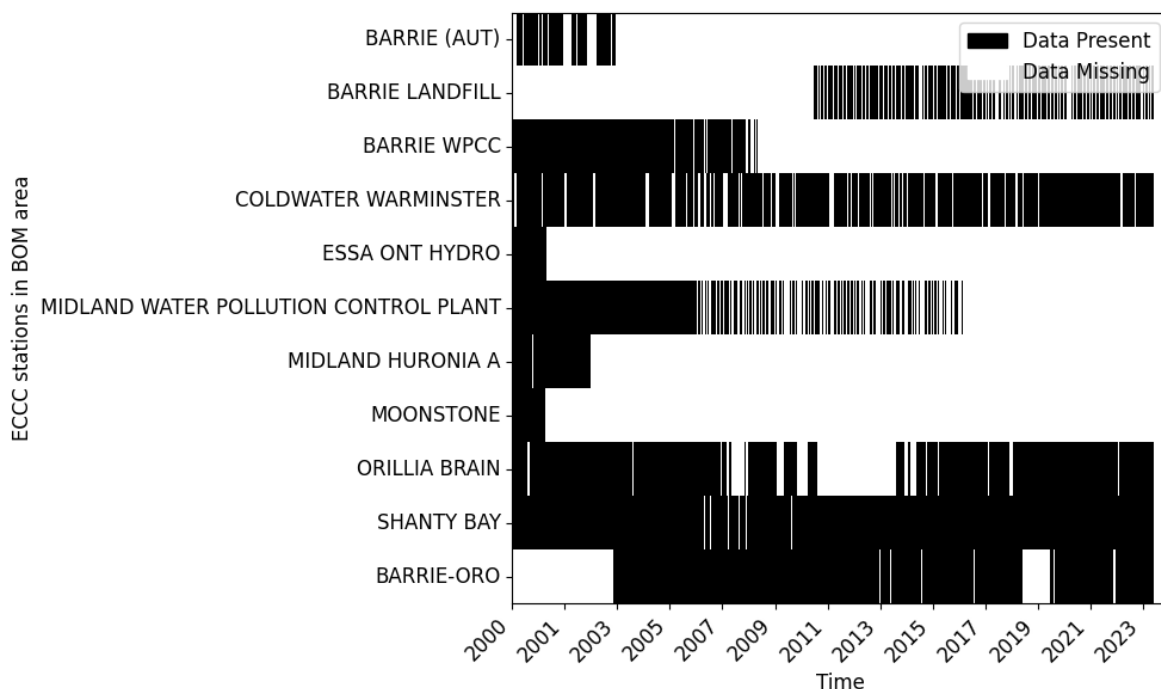


Figure 5-8- Availability of daily precipitation observations from ECCC stations in the BOM area (2000–2023). Black segments indicate dates with available data, while white gaps represent missing entries.

To address this issue, linear interpolation was applied to fill the identified gaps and enhance the continuity of the dataset for subsequent analysis. As shown in Figure 5.9, this approach significantly enhanced the completeness of most station records, especially those with moderate to high missingness. Stations like Coldwater Warminster, Orillia Brain, and Shanty Bay now exhibit near-complete time series, enabling more robust downstream analysis. While interpolation cannot fully recover long continuous gaps

– such as those observed in Moonstone and Essa Ont Hydro – it provides a practical solution for maintaining temporal consistency.

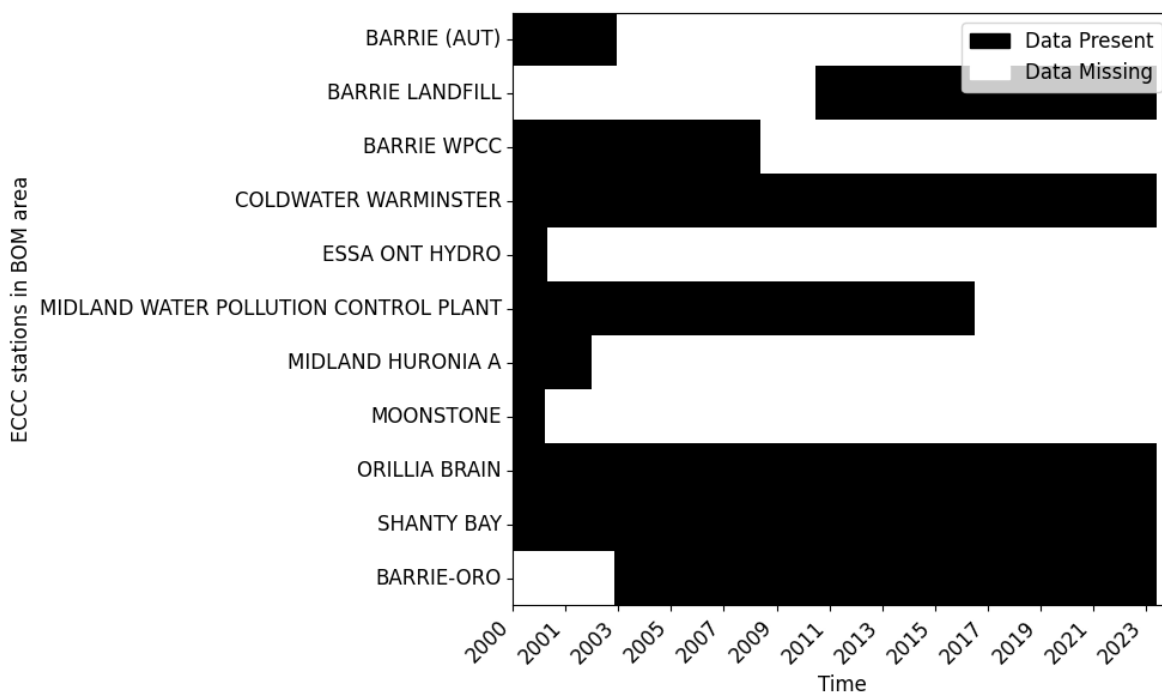


Figure 5-9- Availability of daily temperature data interpolated from ECCC stations in the BOM area (2000–2023). Black segments indicate dates with available data, while white gaps represent missing entries.

A similar interpolation process was applied to the precipitation dataset to address the gaps identified in the earlier analysis. As shown in Figure 5.10, linear interpolation substantially improved the continuity of daily precipitation records across most stations. Stations that originally had low to moderate data gaps – such as Shanty Bay, Orillia Brain, and Coldwater Warminster – now show more continuous time series. Interpolation also substantially improved data completeness in stations like Barrie WPCC, Barrie Landfill, and Midland Water Pollution Control Plant, which had originally lost between 51.9% and 64.5% of their observations.

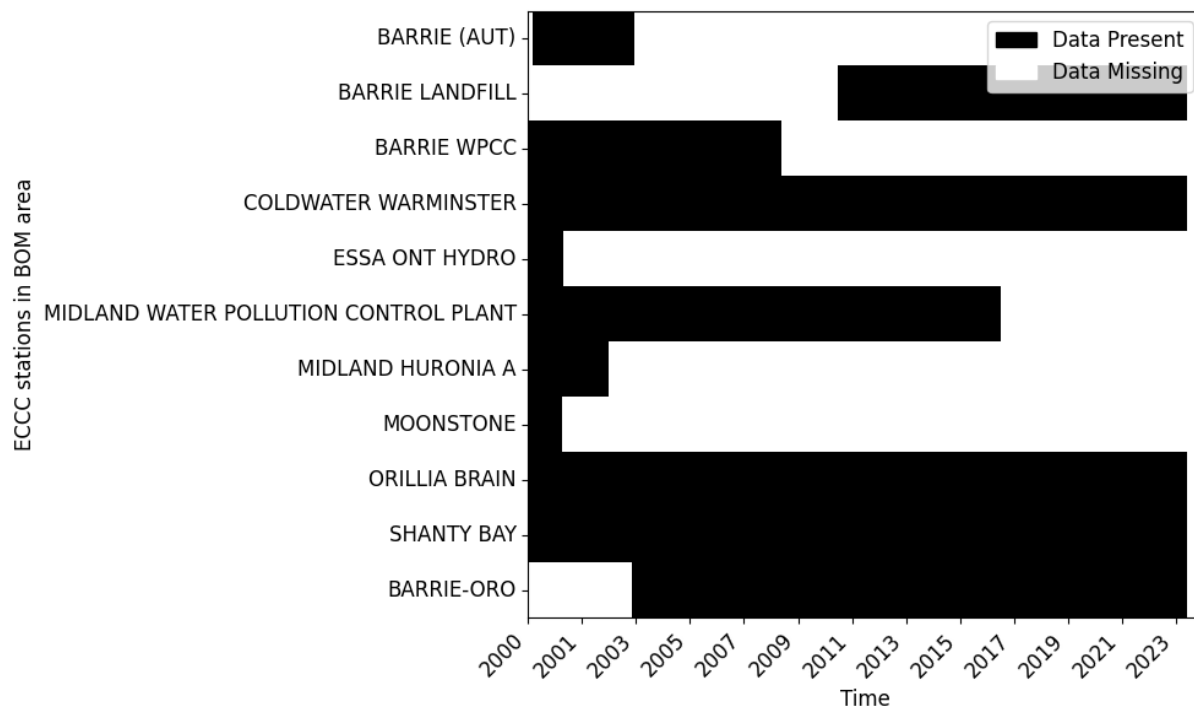


Figure 5-10- Availability of daily precipitation data interpolated from ECCC stations in the BOM area (2000–2023). Black segments indicate dates with available data, while white gaps represent missing entries.

However, similar to the temperature dataset, stations with extensive missing periods – such as Moonstone (95.0%) and ESSA ONT HYDRO (94.7%) – could not be fully recovered, as interpolation cannot reliably reconstruct very long gaps.

5.2.3 Geospatial Filtering of ERA5 Nodes

Due to the extensive data gaps identified in ECCC weather station records, it was necessary to supplement missing observations using external reanalysis products. In this study, ERA5 gridded climate data was utilized to fill such gaps where high-quality local observations were unavailable. However, because ERA5 provides data over a regular spatial grid, it is essential to extract only those nodes that are geographically relevant to the site of interest.

As a result, a geospatial filtering algorithm was used to identify ERA5 nodes near each selected well. For this reason, a defined 15 km radius was set around each well, and all ERA5 data points falling within that radius were selected. Figure 5.11 illustrates the geospatial filtering approach used to identify ERA5 grid nodes within a 15 km radius of a selected PGMN monitoring well in Southern Ontario for the purpose of gap filling. The radius is adjustable and can be increased or decreased depending on the search requirements. This buffer ensures that only spatially representative ERA5 data is included in the analysis.

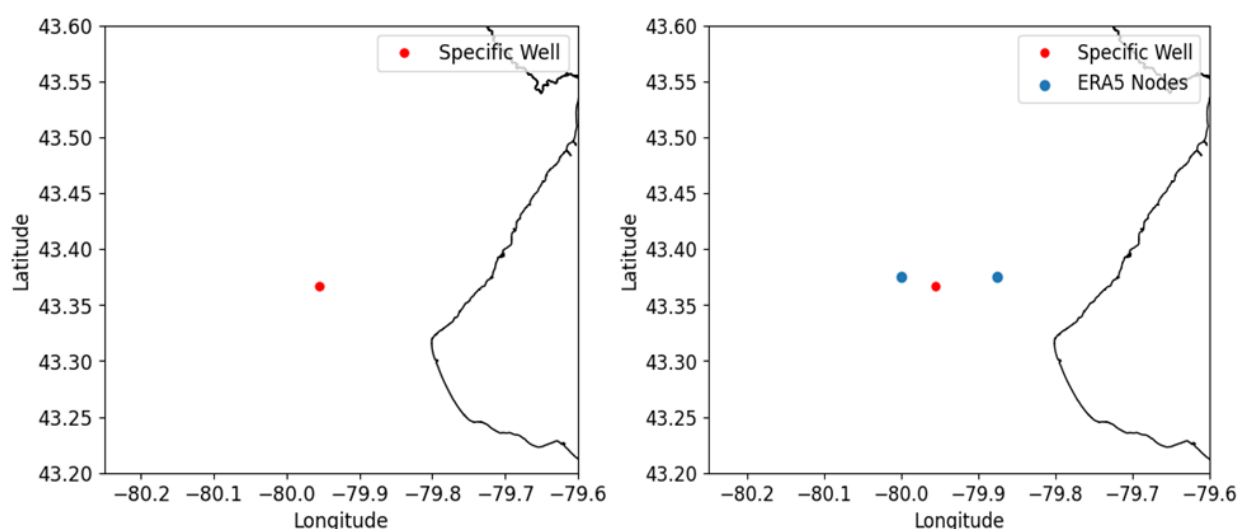


Figure 5-11-Example of geospatial filtering used to select ERA5 grid nodes for gap filling. ERA5 grid nodes (blue points) identified within a 10 km search radius of a PGMN monitoring well in southern Ontario; these nodes were used to supplement missing meteorological observations from ECCC stations. The search radius is adjustable in the analysis

The code used to implement this filtering process begins by reading the location of a selected PGMN well and extracting its geographic coordinates. It then iterates over all ERA5 temperature and precipitation stations, each containing multiple grid points with geographic coordinates. Using the GeoPandas library (Jordahl et al., 2020), the script calculates the geodesic distance between the well and each ERA5 grid point.

This filtering approach helps ensure that only ERA5 data from nearby locations are used for filling data gaps and analyzing groundwater–climate interactions. Figure 5.12 illustrates the identification of

ERA5 nodes located within a 15 km radius of a selected PGMN well in the BOM area. In some cases, the same ERA5 nodes may be associated with multiple wells due to overlapping buffer zones created by their close proximity.

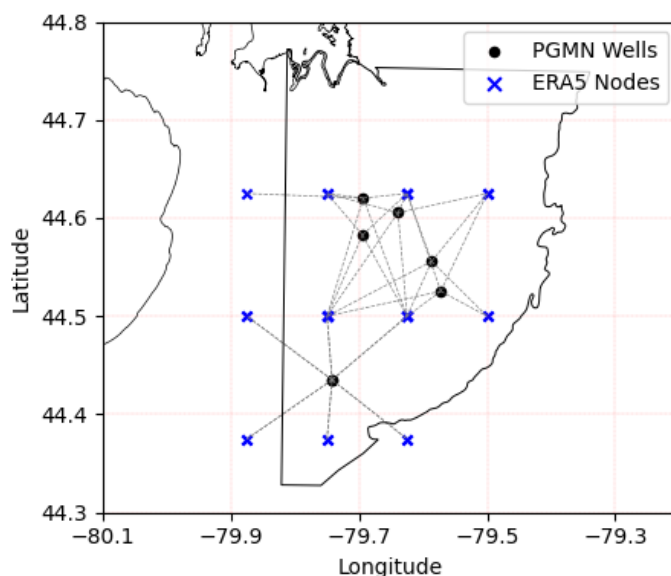


Figure 5-12-Spatial matching of ERA5 nodes (blue crosses) and PGMN wells (black circles) in the BOM area. Dotted lines illustrate the connections used to identify nearby grid points within a 15 km radius of each well for data interpolation and analysis.

Watershed boundaries also need to be taken into account when linking stations to wells. After selecting nearby nodes, only those falling within the same watershed as the corresponding well are retained for further analysis. As noted earlier, ECCC observational data were given priority over ERA5 reanalysis data, since they are generally more reflective of actual local climate conditions. As shown in Figure 3.3, Section 3.1.3, watershed boundaries are overlaid with ERA5 nodes to illustrate the selected nodes corresponding to each watershed in the BOM study area.

5.2.4 GWL Trend Analysis

The Mann-Kendall test was applied to all wells in Southern Ontario to assess whether the long-term GWL trends show statistically significant changes or remain stable. As shown in Figure 5.13, the

Mann-Kendall test results are presented for nine randomly selected monitoring wells to evaluate whether their GWL trends are statistically significant. Each subplot displays a time series of daily GWL observations (black dots) and a fitted linear trendline (red). The label $H = true$ indicates a statistically significant monotonic trend ($p < 0.05$), while $H = false$ suggests no significant trend.

Several wells (e.g., 71-1, 519-2, 299-1, 250-1) show statistically significant increasing trends, which may point to rising GWLs potentially linked to changes in recharge, reduced withdrawals, or regional hydrological shifts. For instance, well 71-1 shows a clear upward trend over more than a decade, while 519-2, despite shorter records, also shows a consistent seasonal pattern with a positive slope.

In contrast, wells such as 44-3 and 427-1 have $H = false$, indicating no significant trend despite visible fluctuations. These could be due to seasonal variability, short time windows, or inconsistent data that obscure long-term changes. Interestingly, some wells (e.g., 361-2) show a significant *declining* trend ($H = true$), reflecting potential localized depletion or increased pumping stress.

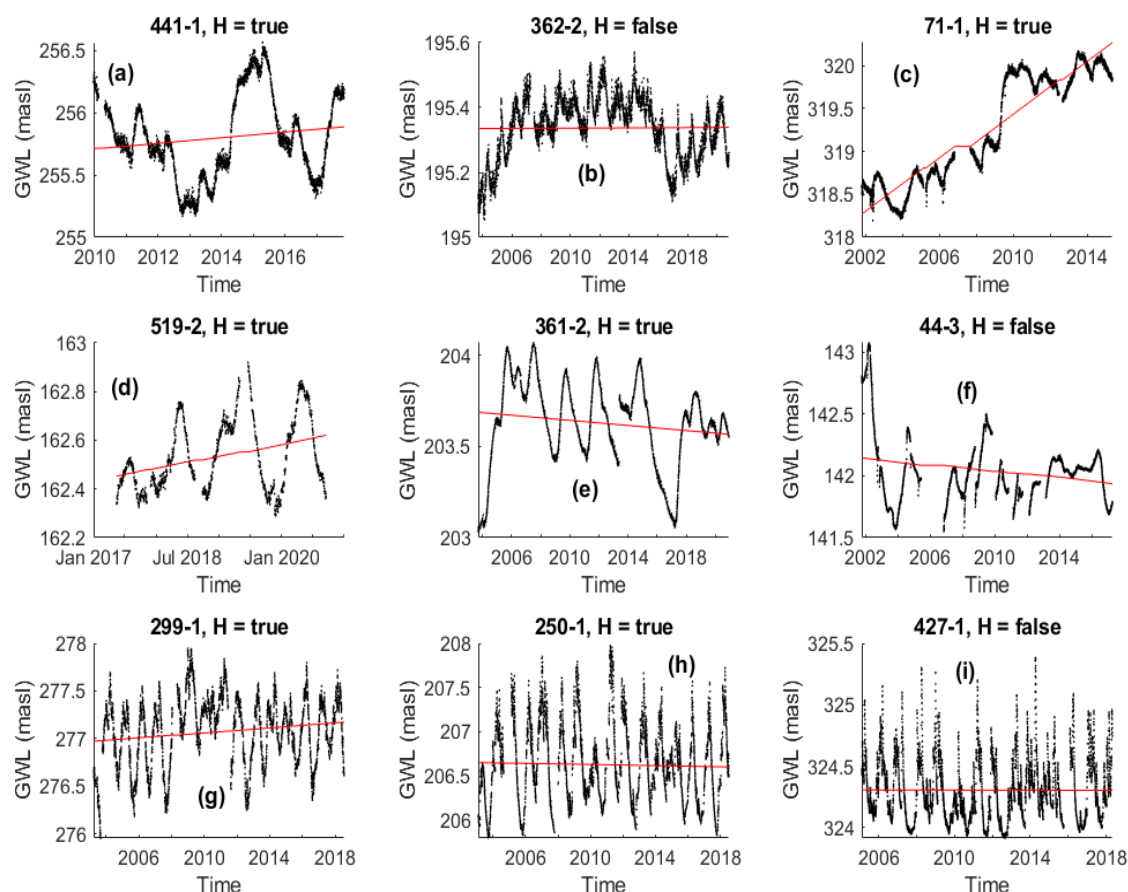


Figure 5-13-Results of the Mann-Kendall trend test for GWLs at nine selected monitoring wells: (a) Well 441-1, (b) Well 362-2, (c) Well 71-1, (d) Well 519-2, (e) Well 361-2, (f) Well 44-3, (g) Well 299-1, (h) Well 250-1, and (i) Well 427-1.

The same Mann-Kendall trend analysis was applied to all wells in the PGMN dataset across Southern Ontario to evaluate the distribution of increasing, decreasing, and non-significant trends across different aquifer types. Among the 389 GWL wells analyzed, 359 wells (92.29% of the total) exhibited statistically significant long-term trends, as the null hypothesis was rejected during the hypothesis testing. These trends are illustrated in Figure 5.14 (a). Of these wells, 239 (66.57%) displayed an increasing trend with positive Sen's slope, while 120 (33.43%) showed a decreasing trend with negative Sen's slope, indicating that increasing trends are more prevalent across the dataset. Among the wells with increasing

trends, 139 (58.16%) were drilled in overburden aquifers, and 92 (38.49%) were in bedrock aquifers. Similarly, among the wells with decreasing trends, 120 (50.83%) were in overburden aquifers, and 53 (44.17%) were in bedrock aquifers.

Only 30 wells (7.71%) showed no statistically significant trends, indicating a strong regional pattern in GWL changes. Of these, 17 were in overburden aquifers, 11 in bedrock aquifers, and 2 at the interface. The higher occurrence of increasing trends, particularly in overburden wells, may be linked to shifts in climate conditions, land use practices, or decreased groundwater withdrawal in some regions.

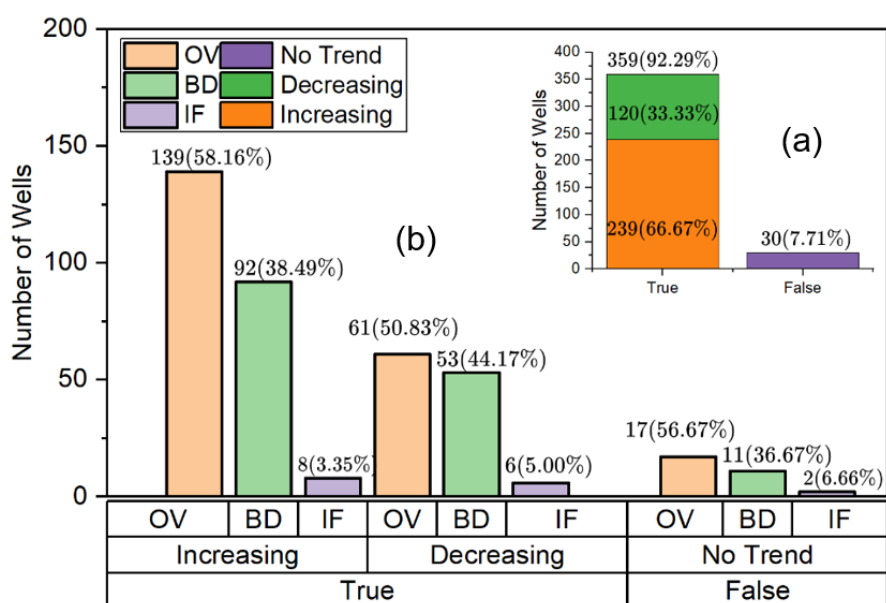


Figure 5-14-Distribution of statistically significant and non-significant GWL trends across different aquifer types in the PGMN network. Panel (a) shows the total number of wells with significant increasing and decreasing trends, while panel (b) breaks down the trend direction (increasing, decreasing, or no trend) by aquifer type: Overburden (OV), Bedrock (BD), and Interface (IF). Percentages reflect each category's share within the trend group.

The Oak Ridges Moraine Groundwater Program similarly used the Mann-Kendall test to detect trends in GWLs, but there are key differences compared to this study. Their analysis is limited to the most recent 10 years, while this study spans a longer 20-year period, providing a more extended perspective on long-term changes. In addition, they classify wells based on depth – shallow wells being less than 20 meters

and deep wells with more than 20 meters depth from the ground surface – whereas the classification in this section is based primarily on aquifer type, distinguishing between overburden and bedrock systems as included in the PGMN dataset. According to Figure 5.15, both shallow and deep wells show a higher count of decreasing trends compared to increasing ones. Additionally, the “none” category – which includes wells without statistically significant trends – accounts for nearly half of the total. This suggests that using a different classification criterion, such as well depth instead of aquifer type, or adjusting the analysis period, can noticeably impact the trend detection results.

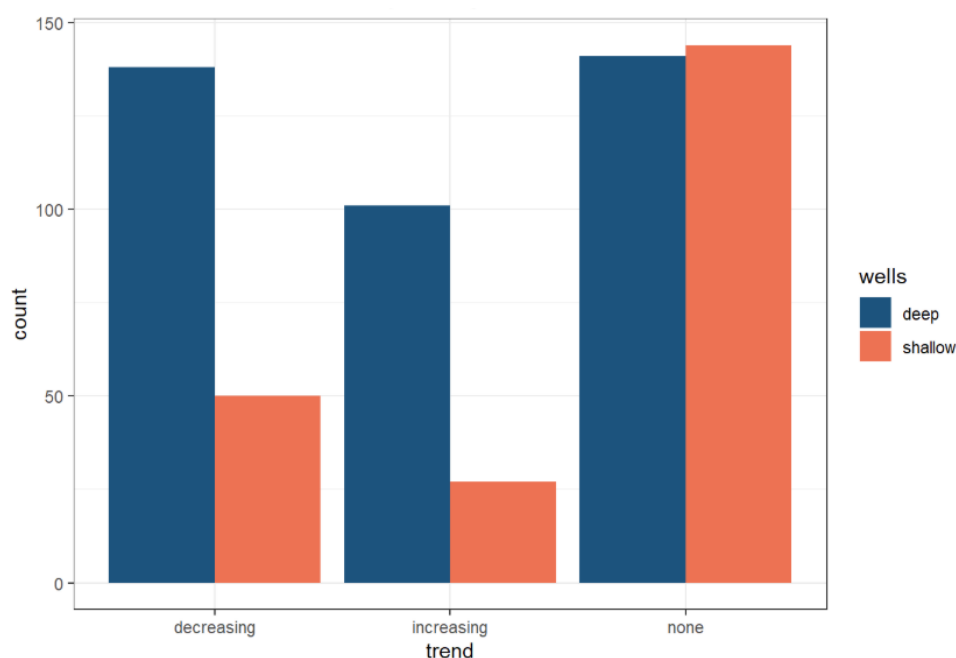


Figure 5-15-Distribution of statistically significant Mann-Kendall GWL trends across shallow and deep wells in the Oak Ridges Moraine Groundwater Program (ORMGP) study area. Source: Oak Ridges Moraine Groundwater Program (2023)

5.2.5 GWL Slope Distributions in Overburden and Bedrock Aquifers

For all positive Mann-Kendall test results, the Sen’s slope distribution is plotted to compare the two main categories -overburden and bedrock- within this dataset, wells constructed in bedrock and overburden aquifers. The Sen’s slope is less sensitive to outliers in the slope distribution and is a direct output.

As illustrated in Figure 5.16, the mean (μ) and median ($\tilde{x}$) slopes in the distributions for the overburden category are $4.49 \times 10^{-5} \frac{m}{day}$ and $2.90 \times 10^{-5} \frac{m}{day}$, respectively, compared to $2.56 \times 10^{-5} \frac{m}{day}$ and $1.77 \times 10^{-5} \frac{m}{day}$ for bedrock wells. The standard deviation (σ) values are also comparable between the two aquifer types – both around $2 \times 10^{-4} \frac{m}{day}$ – suggesting similar variability in GWL change rates across the dataset. Overall, the mean and median slopes for overburden wells are approximately 1.75 and 1.64 times higher than those for bedrock wells. This finding highlights the more pronounced long-term changes in GWL within overburden aquifers, probably due to their greater exposure to climatic variations and groundwater extraction. This aligns with findings from Singer et al. (2003) suggesting that GWL changes are more dynamic in overburden aquifers than bedrock ones due to their composition and hydraulic properties in Southern Ontario.

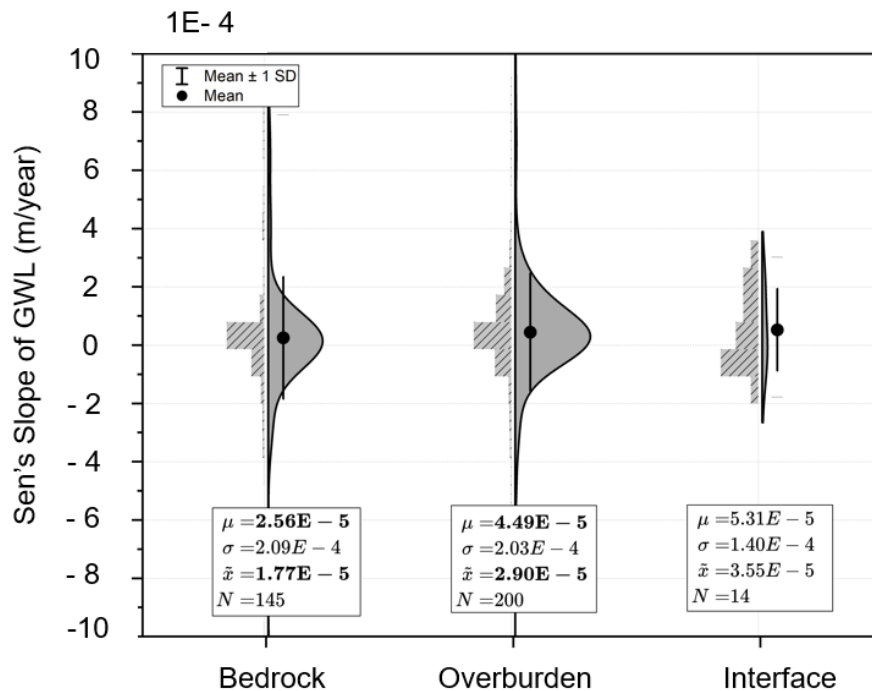


Figure 5-16-Violin plot of Sen's slope estimates for wells with statistically significant Mann-Kendall trends, grouped by aquifer type. Each distribution illustrates the rate of long-term GWL change (in m/year) for bedrock, overburden, and interface aquifers. Mean (μ), standard deviation (σ), and median ($\tilde{x}$) values are reported, with overburden wells showing higher slope magnitudes on average.

5.2.6 Spatially Averaged GWL in Southern Ontario

Only 14 wells were drilled at the interface of the bedrock and overburden aquifers, so they were excluded from further analysis. To estimate the spatially averaged GWLs across Southern Ontario, all other available observations were used. Since not all wells were active at every time point, the number of observations contributing to the average varied over time. As in earlier steps, bedrock and overburden wells were analyzed separately to compare their trends.

For the overburden wells, Figure 5.17 shows the temporal average GWLs from 2000 to 2021, alongside the number of observation wells contributing data over time. The black line indicates the average GWL, while the blue line represents how many stations were active at each point.

Between 2003 and 2017, when stations coverage was highest (reaching over 150 wells), the GWLs exhibit a general upward trend, with fluctuations possibly driven by seasonal or climatic factors. Prior to 2003, the number of contributing wells was steadily growing but still insufficient to produce a stable, consistent average, resulting in noticeable short-term variability in the GWL data. After 2017, both the number of stations and the GWL begin to decline, with the drop in well coverage likely affecting the stability and representativeness of the calculated average.

The red dashed line represents a fitted linear trend with an R^2 of 0.569, indicating a moderate correlation and a long-term increase in GWL. The equation $GWL = 2.430 \times 10^{-3}Time + 235.905$ shows a positive slope, confirming the overall upward trend in GWL across the study area. The time variable in the fitted equation is measured in days, with day zero corresponding to January 1, 2003. As a result, the slope represents the rate of GWL change per day. To express this rate in terms of annual change between 2003 and 2017, the slope must be multiplied by 365.25 days per year, yielding an average yearly GWL increase of approximately $0.89 \frac{m}{year}$. This indicates a commonly high annual GWL variation even if it is between the years 2003 and 2017. If the entire study period is considered – instead of focusing only on the peak station coverage between 2003 and 2017 – the estimated slope decreases to approximately $0.44 \frac{m}{year}$.

This suggests that including the early and late years of the dataset, when station coverage was lower, affects the trend. However, this estimate is less reliable due to the reduced number of observation wells during those periods, making the mid-period (2003–2017) results more dependable. From a hydrogeological perspective, both the mid-period and full-period slope values appear relatively high. Referring to Figure 5.16 – which shows the Sen’s Slope Estimator rather than a linear regression fit – the average GWL change for overburden wells was $4.49 \times 10^{-5} \frac{m}{day}$, or about $0.016 \frac{m}{year}$. This estimate is more reliable than the previous slope values, as the Sen’s method is less influenced by outliers and better reflects typical long-term groundwater changes across Southern Ontario.

In addition, the decrease seen in the GWL after 2017 is likely caused by the drop in the number of active wells used in the average, rather than a real change in the aquifer. Overall, while the long-term trend appears positive, reduced data coverage in recent years introduces uncertainty and highlights the importance of consistent monitoring for accurate groundwater assessment.

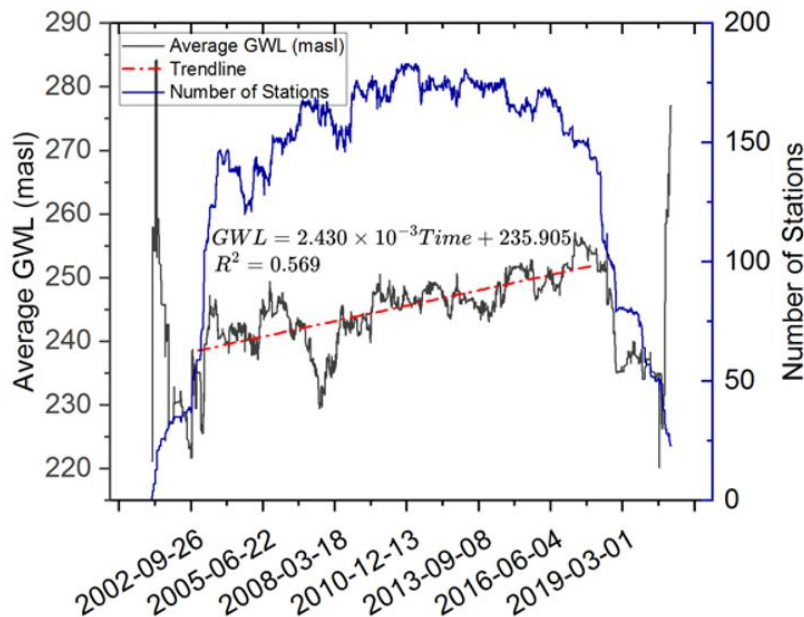


Figure 5-17-Time series of spatially averaged GWL from overburden wells across Southern Ontario between 2000 and 2021. The black line shows the mean GWL (in meters above sea level), and the blue line represents the number of active wells contributing to each time point. The red dashed line is a fitted linear trend ($R^2 = 0.569$), indicating a long-term increase in GWL.

In the case of bedrock wells, Figure 5.18 illustrates the trend of average GWLs for these wells from 2000 to 2021, along with changes in the number of active observation wells over time. As previously described, the overall pattern reflects both GWL variation and shifts in monitoring coverage.

Between 2003 and 2017, when station coverage for bedrock wells peaked (exceeding 120 wells), the average GWL showed a generally increasing trend, though with a gentler slope compared to the overburden wells. To support this observation, a linear regression was applied to this period, resulting in an equation of $GWL = 8.637 \times 10^{-4}Time + 220.299$ with an R^2 value of 0.041, indicating a weak correlation but a slight upward trend. The slope of this trendline is approximately 35.5% of the one derived from the average GWL of the overburden wells, indicating a significantly stronger upward trend in the overburden aquifers of the PGMN compared to the bedrock aquifers.

Following the same approach used for the overburden aquifers, the daily slope was converted to an annual GWL change, resulting in an estimated rate of approximately $0.32 \frac{m}{year}$. When the entire time period with available bedrock observations is considered, the average slope drops to a negative value of $-0.097 \frac{m}{year}$. This decline is likely influenced by the low number of active wells during the early and late parts of the record, which reduces the reliability of the slope as a true indicator of long-term GWL trends from 2000 to 2021. In contrast, the Sen's Slope Estimator gives a more robust mean value of approximately $9.34 \times 10^{-3} \frac{m}{year}$ for bedrock wells, which is more consistent with hydrogeological expectations.

Similar to the overburden averages, a more stable trend was observed during the period of peak data availability. However, before 2003 and after 2017, sharp fluctuations appeared in the average GWLs, highlighting the impact of limited observations on data-driven analyses. Relying on these edge periods may not accurately reflect the true behavior of average GWL trends and can introduce uncertainty into the results.

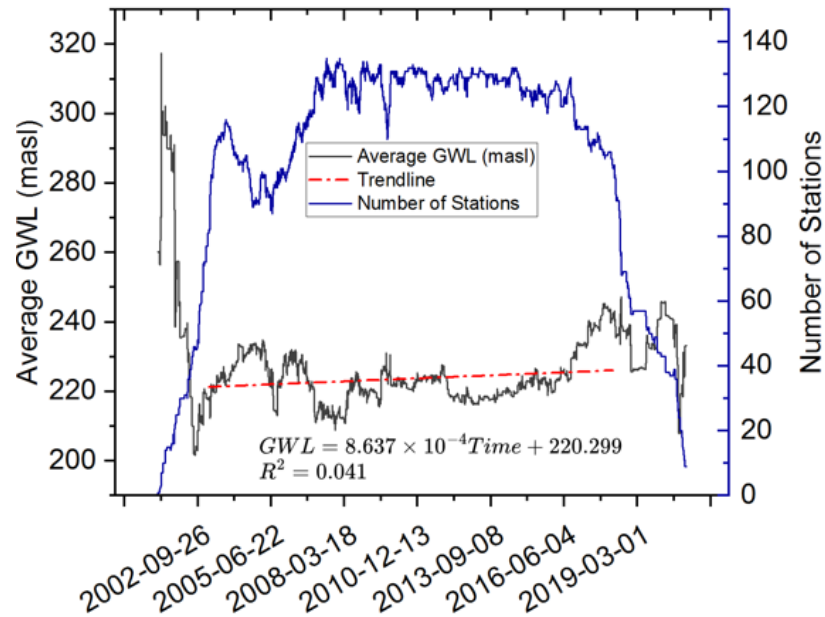


Figure 5-18- Time series of spatially averaged GWL from bedrock wells across Southern Ontario between 2000 and 2021. The black line shows the mean GWL (in meters above sea level), and the blue line represents the number of active wells contributing to each time point. The red dashed line is a fitted linear trend ($R^2 = 0.041$), indicating a long-term increase in GWL.

This spatial averaging process suggests that long-term average GWL values may not accurately capture the true behavior of GWL changes, as they can be influenced by outliers in certain locations. Instead, using alternative slope estimators can better represent the trends in a way that aligns more closely with established hydrogeological understanding, especially since the overall slopes showed fluctuations that don't fully reflect typical groundwater patterns.

5.2.7 Seasonally Averaged GWL in Southern Ontario

To gain deeper insight into temporal variations of GWL data, the spatially averaged data were further averaged into seasonal time series.. This seasonal breakdown helps reveal whether the observed long-term trends are consistent across different times of the year or driven by seasonal patterns. Furthermore, given that both the mean and median Sen's slope values were higher for overburden wells compared to bedrock wells – indicating more pronounced long-term changes – the analysis in the following

steps focused only on overburden wells. This decision was made to prioritize the aquifer type showing stronger and more consistent responses to climate and land-use variability.

Figure 5.19 displays yearly average GWLs for spring, summer, fall, and winter from 2003 to 2017 – a period with the highest station coverage for both overburden and bedrock wells – showing how GWLs have evolved over time within each season. For this reason, all daily GWL observations within each season were averaged, and the resulting yearly trends were plotted to track seasonal variations over time.

In all four seasons, the GWLs show a clear upward trend, indicating a consistent rise over the study period. The linear trendlines fitted to each seasonal dataset are strong, with R^2 values above 0.89 in all cases, suggesting that the seasonal GWL changes are well explained by a linear model. Among the seasons, fall exhibits the steepest slope (0.74), pointing to the most pronounced increase in GWLs during that time. Winter, on the other hand, shows the highest R^2 value (0.945), indicating a very stable and consistent rise with minimal year-to-year variation. Spring and summer also show positive trends with slopes of 0.66 and 0.70, respectively, and strong linear fits. Overall, the results suggest that GWLs in overburden wells have been steadily increasing in all seasons, with some seasonal differences in the strength and consistency of the trend.

The slight differences between the seasonal slopes and the overall annual slope ($0.89 \frac{m}{year}$ from Section 5.2.6) arise because, in the seasonal analysis, each data point represents the average GWL for a specific season in a given year. This method reduces short-term fluctuations and results in trendlines that reflect seasonal patterns more clearly, rather than the combined effect of all daily values throughout the year.

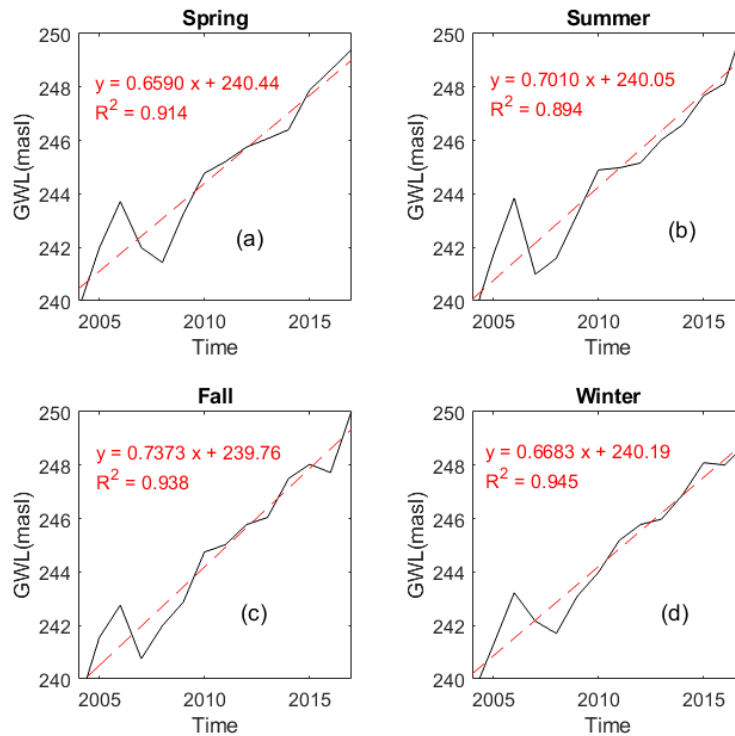


Figure 5-19-Seasonal trends of average GWL in overburden aquifers across Southern Ontario from 2005 to 2017. Each panel (a–d) represents one season: (a) Spring, (b) Summer, (c) Fall, and (d) Winter. Black lines indicate seasonal average GWLs, while red dashed lines represent fitted linear trendlines with corresponding regression equations and R^2 values.

To compare seasonal variations more effectively, the trends for all four seasons were plotted on a single graph, allowing for direct comparison at each point in time. Figure 5.20 presents the seasonal GWL patterns throughout the study period on the same axis. By analyzing the maximum GWL per year, it was found that spring had the highest levels in 9 out of 13 years, confirming its dominant role in seasonal recharge. Winter followed with 6 peaks years, likely due to minimal water extraction and low evapotranspiration losses. Summer was the highest in 5 years, often later in the record, suggesting potential shifts in seasonal recharge timing. Fall had the fewest peak years (2), indicating it is less commonly associated with annual GWL maxima.

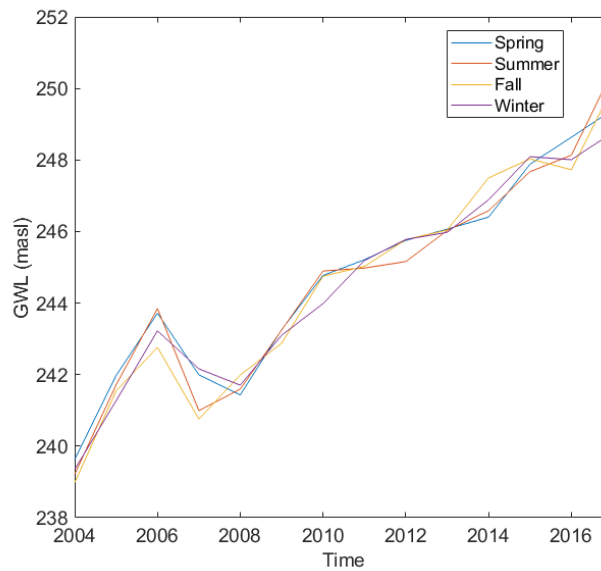


Figure 5-20- Seasonal patterns of mean GWLs in overburden aquifers across Southern Ontario from 2004 to 2017, displayed on a single graph to allow direct comparison.

The results are generally aligned with the analysis conducted by Li & Wang (2022), which examined the long-term GWL trend across the Canadian landmass using the Gravity Recovery and Climate Experiment (GRACE) satellite observations and the Canadian Groundwater Observation Well Network (CGOWN) data. During the 20-year period from 2002 to 2020, they found that GWLs generally peak in spring-particularly in shallow overburden aquifers-due to snowmelt and spring precipitation recharge. They also noted that winter peaks may occur in some cases, driven by reduced evapotranspiration and minimal groundwater extraction. According to their national-scale analysis, GWLs typically reach their lowest levels in late summer or early fall. In contrast, the regional analysis occasionally showed summer as the peak season, which may reflect localized precipitation patterns or differences in data availability in Southern Ontario.

5.2.8 SSA on the Daily and Monthly Data

As outlined in Section 4.1.3.4, SSA ($l = 30$ days) was first applied to the daily time series to reduce short-term fluctuations prior to converting the data to a monthly frequency. This preprocessing step aimed to minimize the impact of high-frequency noise before calculating monthly averages or totals. As shown in Figure 5.21, which presents four variables for well 293-2 – (a) temperature, (b) precipitation, (c) hydrometric level, and (d) GWL – the SSA-processed curves (orange) closely follow the original monthly data (blue). This indicates that applying SSA at the daily scale effectively preserves the overall trend while reducing daily-level noise.

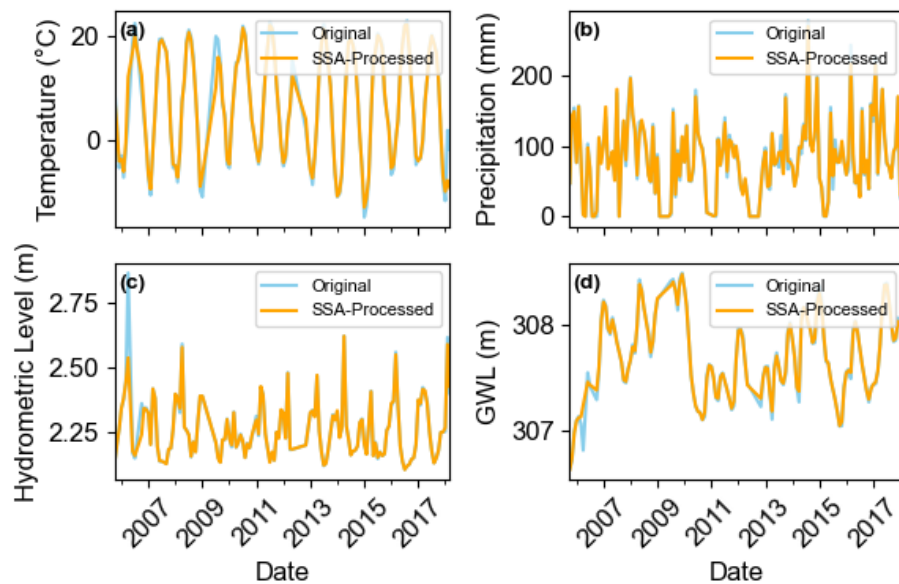


Figure 5-21- Comparison of original monthly data (blue) and SSA-processed data (orange) for well 293-2, where SSA was applied at the daily scale prior to aggregation. The subplots represent (a) air temperature (°C), (b) precipitation (mm), (c) surface water level (m), and (d) GWL (m).

To compare, SSA ($l = 3$ months) was also applied after the data had been aggregated to a monthly resolution. As shown in Figure 5.22, this approach produces smoother curves compared to the original monthly data, especially for groundwater and hydrometric levels. While this method enhances overall smoothness, it may also suppress subtle seasonal or short-term signals that could be important for capturing hydrological variability. The differences between the two approaches highlight the need for further

evaluation to determine which preprocessing method better supports GWL modelling in terms of preserving meaningful temporal patterns and producing more accurate results.

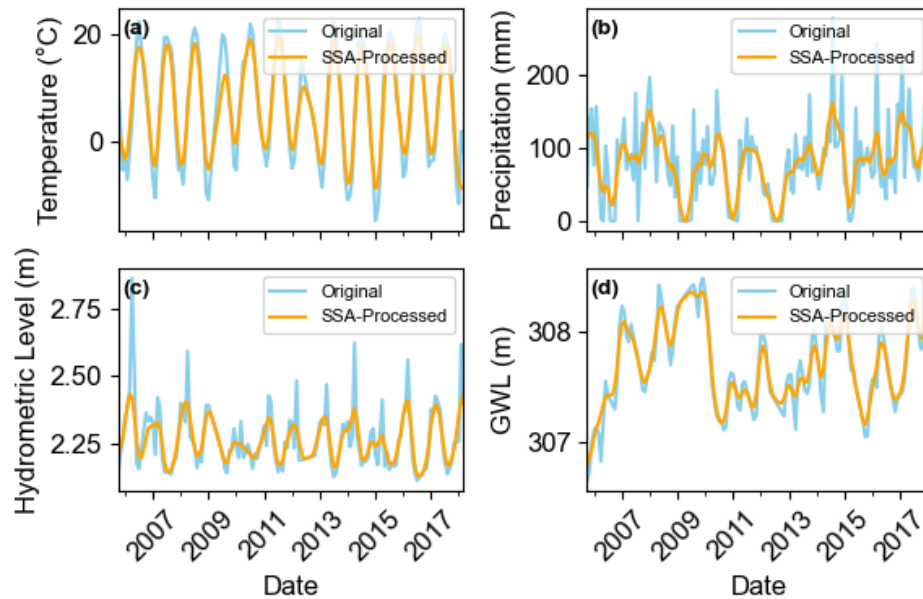


Figure 5-22- Comparison of original monthly data (blue) and SSA-processed data (orange) for well 293-2, with SSA applied after aggregation to monthly frequency. The subplots show (a) air temperature (°C), (b) precipitation (mm), (c) surface water level (m), and (d) GWL (m).

5.2.9 Correlation Analysis Results

As outlined in Section 4.2.4, a more reliable modelling approach considers the time lag between hydroclimatic events and GWL responses. To capture this effect, cross-correlation was used to calculate monthly time lags between GWL and variables such as hydrometric levels, precipitation, and temperature. Additionally, auto-correlation was applied to assess how past GWL values influence future ones, verifying the usefulness of previous GWL observations across all time series. The detailed outcomes of these analyses are discussed further in this section.

5.2.9.1 Cross-Correlation Analysis Results

The cross-correlation coefficient was calculated between GWL data and non-autoregressive features, such as temperature, precipitation, and hydrometric level. This analysis enabled the identification of optimal lags by selecting those with the highest correlation coefficients. As illustrated in Figure 5.23, cross-correlation values with positive time lags were selected, as GWL predictions are expected to depend on preceding months' hydroclimatic conditions. Additionally, cross-correlation trends display cyclic variations across lag times, reflecting the seasonal influence of temperature, precipitation, and hydrometric levels on groundwater fluctuations. The optimal lag is identified as the point of maximum correlation for each cyclic trend shown in Figure 5.23. These values can vary from one well to another based on factors such as geological structure, soil characteristics, permeability, porosity, and other key hydrogeological conditions.

Each subplot shows the correlation coefficients between GWL and three key hydroclimatic variables-temperature (black squares), precipitation (red circles), and hydrometric level (blue triangles)-plotted against monthly time lags ranging from 0 to 12 months.

In general, these plots reveal how GWL responds to each hydrometeorological factors over time. For instance, in wells (a), (b), and (f), hydrometric levels tend to show a consistent positive correlation across multiple months, indicating that surface water levels exert a relatively stable influence on GWL. In wells such as (e) and (g), precipitation displays strong negative correlations at shorter lags and positive correlations at longer ones, suggesting delayed recharge or seasonal effects.

Temperature, on the other hand, tends to show either weak or negative correlations in many wells-particularly visible in wells (d), (f), and (h)-possibly due to the inverse relationship between evapotranspiration and groundwater recharge. Wells (c) and (e) demonstrate a strong positive correlation between GWL and both precipitation and hydrometric levels at different lag times, highlighting the complex and site-specific nature of groundwater responses.

These lag-dependent patterns emphasize the importance of accounting for temporal delays when modelling groundwater dynamics. They also reflect the varying influence of climatic and surface water variables depending on well characteristics and hydrogeological settings.

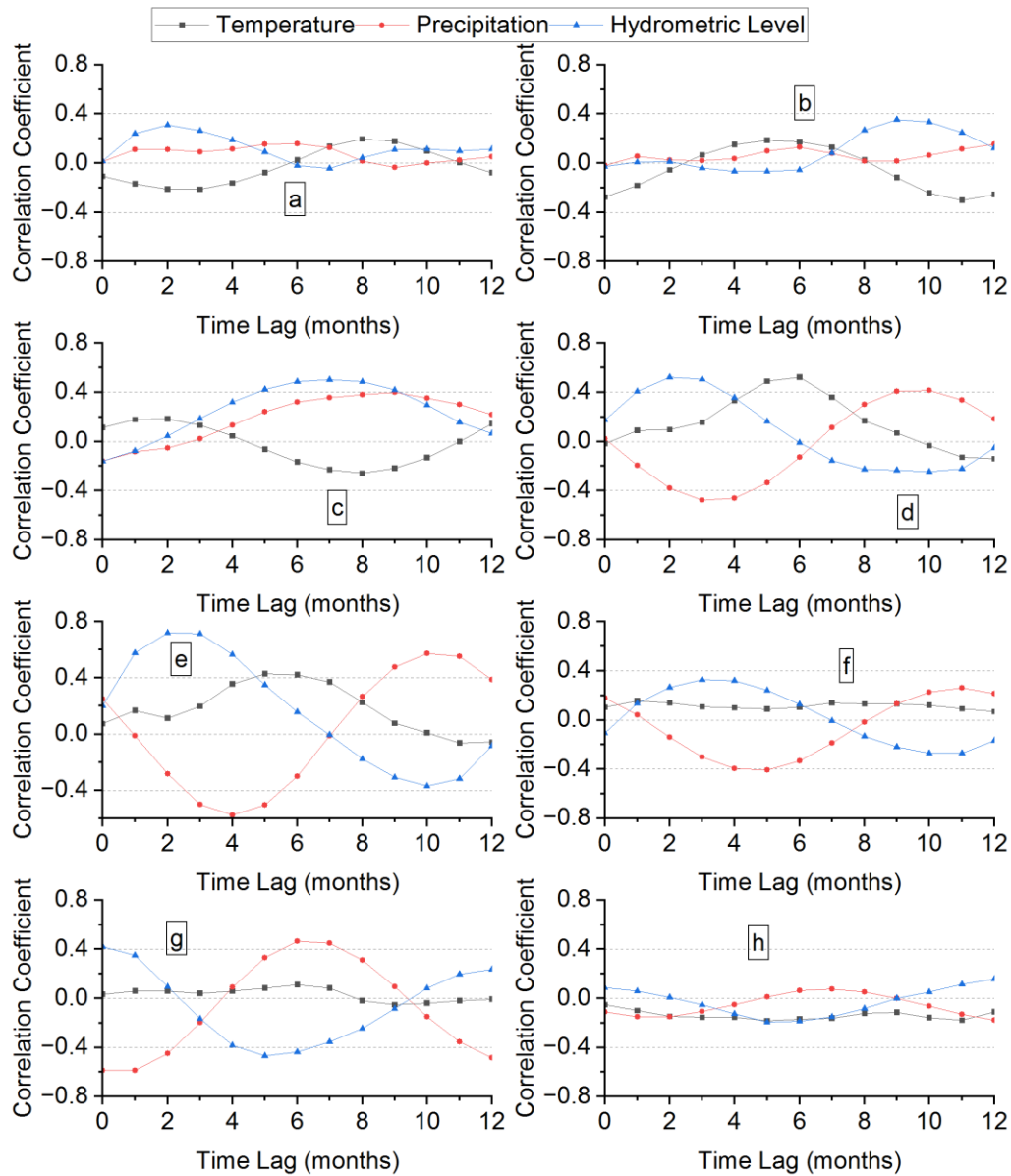


Figure 5-23-Cross-correlation between GWL and hydroclimatic variables for the wells: (a) 293-2, (b) 293-3, (c) 439-1, (d) 440-1, (e) 442-1, (f) 244-2, (g) 245-2, and (h) 443-1.

Tables 5.1 and 5.2 summarize the optimal time lags and corresponding cross-correlation coefficients between GWL and the hydrometeorological variables, temperature, precipitation, and hydrometric level for each well shown in Figure 5.23 under the DSM and DMS scenarios, respectively. Each table contains wells ranging from well 293-2 in watershed W1 to well 443-1 in watershed W4 as the first column, followed by columns indicating the monthly lag at which each hydrometeorological variable shows the strongest correlation with GWL. Then, the corresponding cross-correlation coefficient between GWL and each variable is reported. In general, most wells exhibit moderate to strong correlations, with hydrometeorological variables showing delayed effects.

For the DSM scenario, cross-correlation results indicate variable-dependent relationships between GWL and the hydrometeorological drivers (Table 5.1). Temperature shows both positive and negative correlations with GWL across wells, with correlation coefficients ranging from -0.59 (well 245-2, 1-month lag) to 0.60 (well 442-1, 10-month lag), indicating substantial spatial and temporal variability in the groundwater response to thermal conditions. For the DSM, the strongest positive correlation is found between GWL and hydrometric level at well 442-1 ($r = 0.72$ at a 3-month lag), which perhaps indicates a relatively stable and consistent influence of surface water on GWLs in this deeper aquifer. Conversely, the weakest correlation is observed for well 443-1, where hydrometric level correlates negatively with GWL ($r = -0.19$ at a 5-month lag), possibly due to more complex or indirect hydrological interactions in that region. The strongest correlation between the temperature and GWL data takes place at well 442-1 with a 10-month delay while Across all wells, the shortest effective lag is 1 month (e.g., hydrometric level at wells 244-2, 440-1, and 443-1), while the longest reaches up to 12 months (e.g., precipitation at well 293-3).

Table 5.1-Time lags and cross-correlation coefficients corresponding to strongest correlation between the GWL data for each well under the DSM scenario.

Well	T	T values	P	P values	HL	HL values
293-2	2	-0.21	6	0.15	2	0.30
293-3	11	-0.28	12	0.16	9	0.35
W1	2, 11	-0.21, -0.28	6, 12	0.16, 0.15	2, 9	0.30, 0.35
439-1	7	-0.24	8	0.29	7	0.48
440-1	3	-0.49	1	0.35	3	0.50
442-1	10	0.60	1	0.47	3	0.72
W2	7, 3, 10	-0.24, -0.49, 0.60	8, 1, 1	0.29, 0.35, 0.47	7, 3, 3	0.48, 0.50, 0.72
244-2	5	-0.40	7	0.15	3	0.29
245-2	1	-0.59	6	0.11	5	-0.47
W3	5, 1	-0.40, -0.51	7, 6	0.15, 0.11	3, 5	0.29, -0.47
443-1	12	-0.18	5	-0.19	5	-0.19

Table 5.2- Time lags and cross-correlation coefficients corresponding to strongest correlation between the GWL data for each well under the DMS scenario.

Well	T	T values	P	P values	HL	HL values
293-2	2	-0.21	5	0.19	3	0.30
293-3	11	-0.25	12	0.23	10	0.40
W1	2, 11	-0.21, -0.25	5, 12	0.19, 0.23	3, 10	0.30, 0.40
439-1	7	-0.23	12	0.43	7	0.56
440-1	4	-0.48	1	0.41	3	0.53
442-1	10	0.62	1	0.57	3	0.81
W2	7, 4, 10	-0.23, -0.48, 0.61	12, 1, 1	0.43, 0.41, 0.57	7, 3, 3	0.56, 0.53, 0.81
244-2	5	-0.38	8	0.19	10	0.29
245-2	1	-0.57	5	0.10	6	-0.52
W3	5, 1	-0.38, -0.57	8, 5	0.19, 0.10	10, 6	-0.29, -0.52
443-1	12	-0.18	5	-0.27	6	-0.19

These results highlight the delayed and variable nature of GWL responses in deeper aquifers, influenced by subsurface properties such as permeability, porosity, and geological layering. Therefore,

accounting for appropriate lag structures is essential when modelling groundwater dynamics in such systems.

5.2.9.2 Autocorrelation Analysis Results

The auto-correlation analysis was performed on the GWL time series to evaluate the degree to which current GWLs are influenced by their past values. This method helps identify the persistence or memory effect within the GWL data, which is important for timeseries modelling and prediction. Positive auto-correlation values at specific time lags indicate that past GWLs at those intervals are strongly correlated with the current levels. By identifying the lag at which the autocorrelation is highest, we can determine how far back previous GWL observations remain relevant for predicting future values.

Figure 5.24 shows the results of autocorrelation analysis – commonly named as autocorrelation function (ACF) – conducted on the GWL data. Similar to the cross-correlation analysis, only positive monthly time lags were considered in the autocorrelation process to ensure that the relationship between past GWLs and their subsequent values is captured. In other words, the ACF plots depict how strongly GWL values at a given time are correlated with their values in previous months, across time lags ranging from 0 to 12 months. The solid lines represent autocorrelation coefficients for each lag, while the shaded region marks the 95% confidence interval (CI). Coefficients that fall outside this region are statistically significant.

Across all wells, a clear pattern of gradually decreasing correlation with increasing lag is evident, indicating that past GWL values exert a strong influence on future values – especially at shorter lags. For example, Wells (c) 439-1 and (h) 443-1 exhibit very high lag-1 autocorrelation values of 0.967 and 0.958, respectively, suggesting a strong temporal persistence in GWL behavior. This pattern is consistent with the slow-moving nature of groundwater systems and reinforces the usefulness of prior GWL data – GWL (t-1) - in time series forecasting.

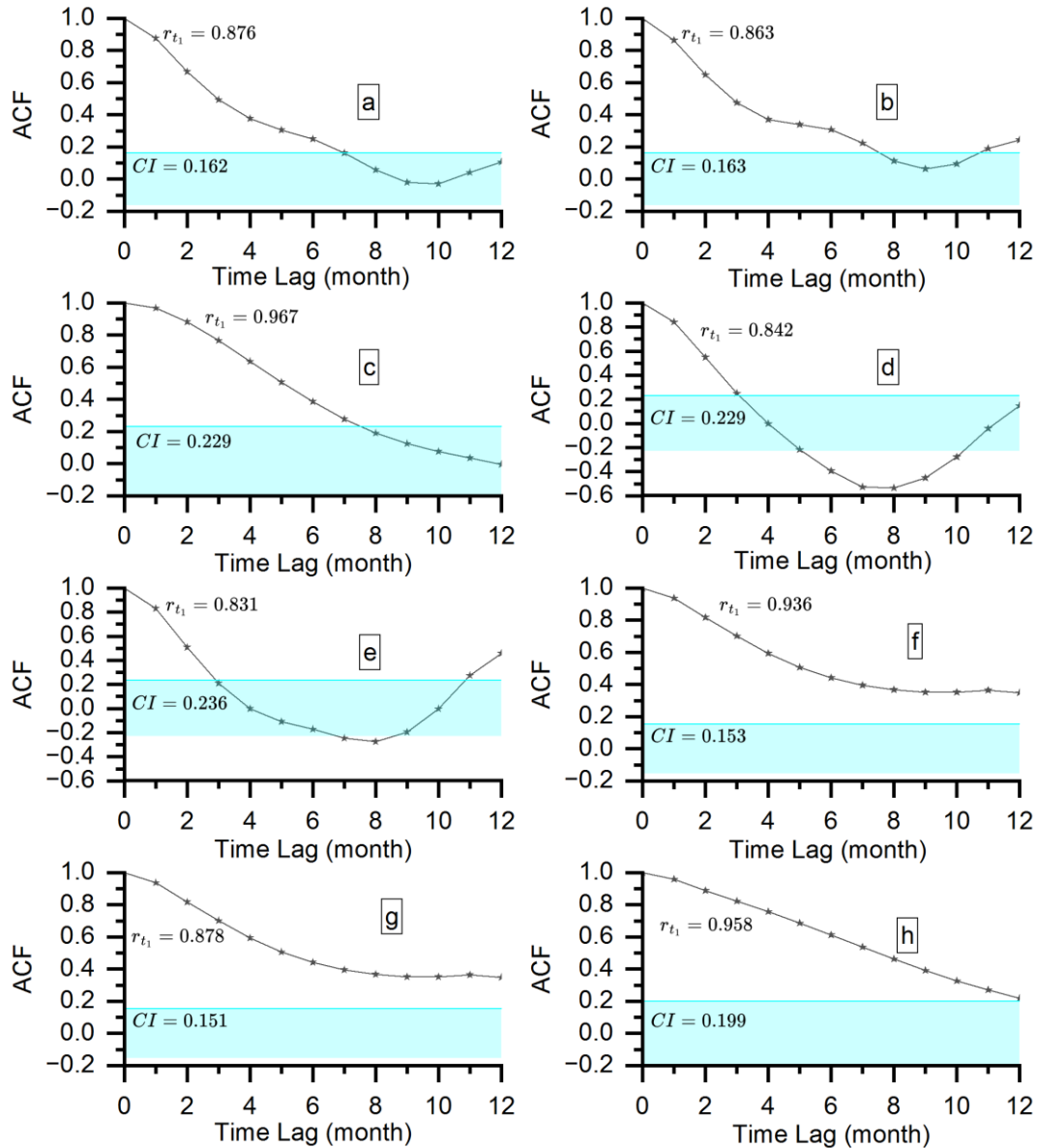


Figure 5-24-Autocorrelation functions (ACF) for the wells: (a) 293-2, (b) 293-3, (c) 439-1, (d) 440-1, (e) 442-1, (f) 244-2, (g) 245-2, and (h) 443-1.

Table 5.3 summarizes the maximum autocorrelation coefficient at lag 1 (r_{t_1}) and the corresponding confidence intervals for each well. As previously discussed, and depicted in Figure 5.24, the strongest autocorrelation consistently occurs at the first-time lag, indicating that the most recent GWL measurement holds the greatest predictive value. Consequently, the first lag and its corresponding CI for each well are

presented in this table. All lag-1 autocorrelation coefficients exceed 0.8, ranging from 0.831 for well 442-1 to a peak of 0.967 for well 439-1. Importantly, none of the maximum ACF values fall within the non-significant region of the CI. This strong temporal persistence further validates the chosen modelling strategy, which relies on autoregressive prediction of GWL using past GWL observations.

Table 5.3- Autocorrelation coefficients at lag 1 (r_{t_1}) and corresponding 95% confidence intervals (CI) for the eight monitored wells shown in Figure 5.24. All wells exhibit strong autocorrelation, indicating significant influence of previous GWLs on current values.

Subplot	Well ID	ACF at Lag 1 (r_{t_1})	95% CI
(a)	293-2	0.876	0.162
(b)	293-3	0.863	0.163
(c)	439-1	0.967	0.229
(d)	440-1	0.842	0.229
(e)	442-1	0.831	0.236
(f)	244-2	0.936	0.153
(g)	245-2	0.878	0.151
(h)	443-1	0.958	0.199

5.2.10 Aquifer Detection Model (AQM) Results

Understanding which aquifer each monitoring well is screened in is a crucial step for interpreting GWL dynamics and identifying hydrogeological patterns in the study area. Due to the complexity and limited availability of subsurface geological information, an automated Aquifer Detection Model (AQM) was developed to classify wells based on spatial relationships with mapped aquifer layers as discussed in Section 4.4.1. The model's outputs were subsequently compared with existing classification records from previous studies. This step enhances the geological understanding of the system and supports the validation of the data used in the analysis.

Table 5.4 compares aquifer assignments for the eight representative PGMN wells as identified by the Tier 2 groundwater modelling projects and those determined using a MATLAB-based AQM developed in this study. In most cases, the results align closely with those from the Tier 2 assessments (see Cuddy, 2013; LSRCA, 2013), supporting the validity of the approach developed in this research. For instance, wells 293-3 and 439-1 are consistently matched with AF1 and AF2 aquifers respectively by both sources, reflecting strong agreement.

However, discrepancies arise in a few cases. For example, well 293-2 is identified as ICSD (Interbedded Coarse and Sandy Deposits) in the analysis, whereas the Tier 2 project assigns it to a combined AF1–GLAF1 system. Similarly, for well 244, the AQM algorithm assigns it to GLAF, while the Tier 2 designation spans both AF1 and GLAF1. These mismatches are likely attributable to inherent uncertainties in geological interpretation – particularly in areas with complex stratigraphy or sparse high-confidence borehole data.

The differences are not unexpected, given that the Tier 2 classifications may have incorporated site-specific hydrogeological expertise and historical context, whereas the AQM model strictly applies spatial proximity and elevation-based logic using continuous and discontinuous datasets. Moreover, geological boundary ambiguity (e.g., medium or low confidence pick points) can influence aquifer identification, especially for wells situated near transitions between multiple hydrostratigraphic units.

Table 5.4- Comparison of aquifer assignments for the PGMN wells based on Tier 2 groundwater studies and the MATLAB-based Aquifer AQM.

PGMN Well	Tier 2	MATLAB Algorithm
293-2	AF1-GLAF1	ICSD
293-3	AF1-GLAF1	AF1
244	AF1-GLAF1	GLAF
245	CAF3-LAF	LAF
439-1	AF2	AF2
440-1	Bedrock	LAF
442-1	CAF1-AF4	AF4
443-1	AF1-GLAF	AF2

In summary, while the AQM model shows a high degree of consistency with expert-derived aquifer classifications, minor discrepancies are understandable and highlight the need for combining spatial modelling with expert review, particularly when preparing inputs for groundwater prediction models.

5.3 Inputs in the Data-Driven Models

The predictive model results presented in this section were developed using the hydrometeorological and autoregressive inputs described in Section 3.2. Monthly temperature, precipitation, hydrometric level, and lagged GWL were used in the modelling framework. The statistical characteristics of these variables for well 293-2 are summarized in Table 3.5, while their temporal behaviour over the study period (2006-2018) is illustrated in Figure 3.9.

As shown in Table 3.5, temperature exhibits strong seasonal variability (mean = 6.85 °C; standard deviation = 10.13 °C), and precipitation displays substantial month-to-month dispersion (mean = 78.72 mm; standard deviation = 55.66 mm). In contrast, hydrometric level (mean = 2.27 m; standard deviation = 0.12 m) and GWL (mean = 307.69 m; standard deviation = 0.41 m) show relatively smaller variability, indicating more stable water level dynamics compared to climatic drivers. Equivalent descriptive summaries for the remaining wells are provided in Appendix 7.2.

5.4 Linear Regression Results

The predictive performance of the linear regression model was first evaluated under the single configuration using chronological 4-fold cross-validation. Table 5.5 summarizes the training, validation, and test performance metrics for Well 293-2.

For Well 293-2, training R^2 values ranged from 0.07 to 0.20 across folds, indicating limited explanatory power even within the training period. Validation R^2 varied between -0.26 and 0.33, reflecting

instability in generalization performance. Test R^2 values were negative in three out of four folds, ranging from -0.32 to 0.25. Negative test R^2 values indicate that, in several temporal splits, the model performed worse than a naïve mean-based predictor.

The corresponding test MSE values varied between 0.044 and 0.103, further highlighting performance inconsistency across time periods. These results suggest that although lagged GWL was included as a predictor, the autoregressive relationship is not temporally stable, and regression coefficients vary across folds.

The summary statistics reveal substantial variability between wells. Two wells (244-2 and 442-1) achieved positive mean test R^2 values (0.19 and 0.18, respectively), suggesting moderate predictive capability under certain hydrogeological conditions. However, several wells demonstrated strong instability. For example, Well 439-1 yielded a mean test R^2 of -4.22, while Well 443-1 exhibited severe performance degradation with a mean test R^2 of -61.68 and negative training R^2 values. In these cases, the linear model failed even within the training folds, indicating a fundamental mismatch between the linear assumption and the underlying groundwater dynamics.

A similar pattern was observed across the remaining seven wells (Table 5.5). While some wells exhibited modest positive R^2 values in isolated folds, the majority showed low or negative test R^2 under chronological validation. This indicates weak temporal generalization of the linear regression framework across the study area.

Table 5.5- Mean training and test performance metrics of single-well linear regression models under chronological 4-fold cross-validation. Values represent averages across folds.

Well	Mean Train R²	Mean Test R²	Mean Train MSE	Mean Test MSE
293-2	0.13	-0.05	0.061	0.067
293-3	0.24	-0.60	0.034	0.051
439-1	0.39	-4.22	0.047	0.096
440-1	0.38	-0.99	0.021	0.058
442-1	0.63	0.18	0.023	0.043
244-2	0.27	0.19	0.031	0.032
245-2	0.39	-0.53	0.028	0.049
443-1	-16.12	-61.68	0.041	0.094

To better understand the causes of this instability, the following sub-chapters examine the core assumptions of linear regression in detail, including linearity, multicollinearity, exogeneity, normality of residuals, homoscedasticity, and autocorrelation. These diagnostic analyses provide further insight into whether the observed predictive limitations arise from structural violations of regression assumptions or from inherent nonlinearity in groundwater system dynamics.

5.4.1 Variance Inflation Factor (VIF)

To assess multicollinearity among the independent variables, the Variance Inflation Factor (VIF) was calculated for each feature. According to the interpretation criteria from Vu et al. (2015), a VIF above 10 typically indicates problematic multicollinearity. The results show that GWL (t-1) and hydrometric level have extremely high VIF values of 1046.26 and 1011.32, respectively, suggesting severe collinearity and potential redundancy in the predictors. In contrast, temperature (VIF = 2.90) and precipitation (VIF = 3.19) fall within acceptable ranges, indicating they contribute unique, non-redundant information to the model.

In this case, linearity is violated in this case, as the extremely high VIF values for GWL (t-1) and hydrometric level indicate that these variables are not providing independent explanatory power.

Table 5.6-Variance Inflation Factor (VIF) results for well 293-2. High VIF values for GWL (t-1) and Hydrometric level indicate strong multicollinearity, while Temperature and Precipitation remain within acceptable linear ranges. Interpretations follow the thresholds defined by Vu et al. (2015).

Feature	VIF	Interpretation (Vu et al., 2015)
GWL (t-1)	1046.26	High collinearity
Temperature	2.90	Acceptable
Precipitation	3.19	Acceptable
Hydrometric level	1011.32	High collinearity

The VIF analysis was extended to all eight wells using individual linear regression models for each case. Table 5.7 presents the VIF values for each well under these single-model scenarios. As the results indicate, every model exhibits severe multicollinearity, with at least one predictor showing an excessively high VIF.

Table 5.7-VIF values for all of the wells in the study

Well	GWL(t-1)	T	P	HL	Well	GWL(t-1)	T	P	HL
293-2	1046.26	2.90	3.19	1011.32	442-1	398.83	3.05	8.49	434.02
293-3	869.38	2.36	3.17	851.20	244-2	681.85	2.65	4.11	656.57
439-1	654.06	4.44	9.02	655.30	245-2	361.35	1.49	4.15	365.35
440-1	442.07	2.85	7.66	442.85	443-1	89.11	4.29	10.39	126.63

5.4.2 Exogeneity

Figure 5.25 shows the distribution of test residuals for well 293-2 across four folds of cross-validation. Each subplot includes a red dashed line at zero, which serves as the ideal reference where residuals would be perfectly unbiased, and a solid blue line that marks the mean residual for that fold. In

Fold 1 (Figure 5.25-a), the mean residual is -0.11, positioned left of the zero line, indicating a moderate underestimation. Fold 2 (5.25-b) shows a smaller negative mean of -0.03, suggesting a mild bias toward underestimation. In contrast, Fold 3 (5.25-c) and Fold 4 (5.25-d) display positive mean residuals of 0.042 and 0.020, respectively, indicating slight overestimation.

While the histograms in some folds appear skewed, the mean residuals remain close to zero, which supports the assumption of exogeneity for this well.

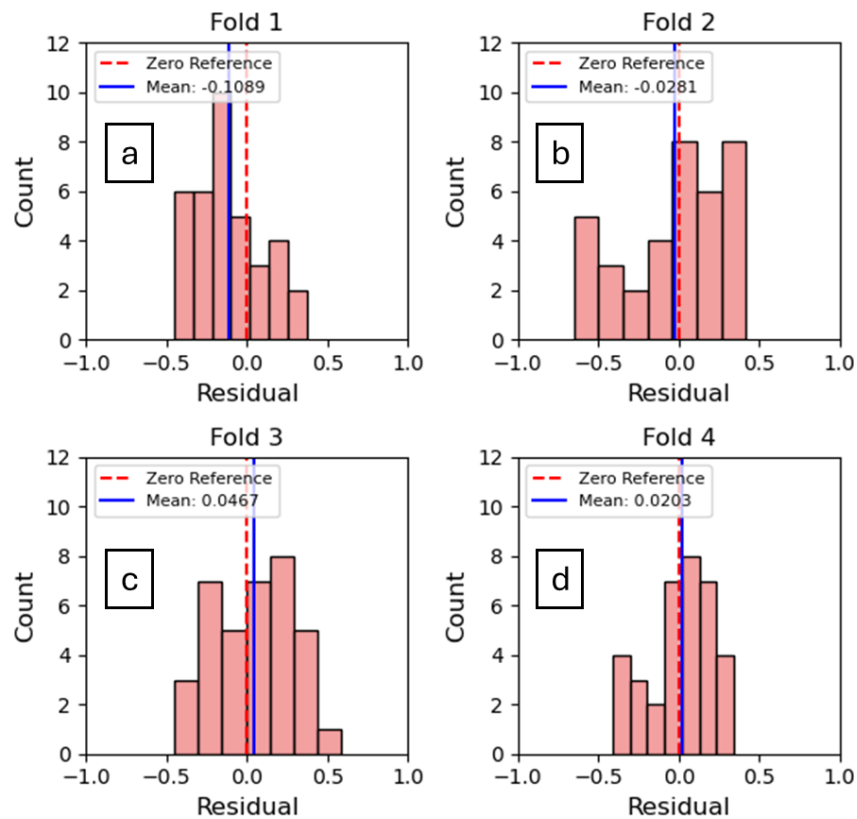


Figure 5-25-Histogram plots of test residuals for well 293-2 across four cross-validation folds. The red dashed line in each subplot indicates the zero reference (ideal unbiased residuals), while the solid blue line shows the fold-specific mean residual. The mean values are: (a) -0.1138, (b) -0.0290, (c) 0.0420, and (d) 0.0200. These results support approximate exogeneity with minor deviations across folds.

A full summary of mean residuals for all eight wells is presented in Table 5.8.

Table 5.8-Mean of residuals for all of the wells

Well	Fold 1	Fold 2	Fold 3	Fold 4	Well	Fold 1	Fold 2	Fold 3	Fold 4
293-2	-0.11	-0.03	0.04	0.02	442-1	0.11	-0.20	-0.04	-0.04
293-3	-0.18	0.06	0.06	0.05	244-2	0.05	-0.03	-0.01	-0.08
439-1	0.34	-0.05	-0.18	-0.17	245-2	0.14	0.10	-0.16	-0.25
440-1	-0.13	-0.12	0.03	0.003	443-1	0.21	0.20	-0.36	0.20

5.4.3 Residual Distribution

To evaluate whether the residuals of the linear regression model follow a normal distribution – a key assumption in linear regression – we plotted histograms of the test residuals along with fitted KDE (Kernel Density Estimate) curves for one representative well (Well 293-2) across four cross-validation folds (Figure 5.26). Each subplot includes the p-value from the Kolmogorov-Smirnov (K-S) test, which was used to statistically assess the normality of residuals by comparing their empirical distribution against a standard normal distribution.

In all four folds (a–d), the visual shape of the residuals suggests mild skewness in some cases, especially in Folds 2 and 3. However, the K-S p-values – ranging from 0.160 to 0.802 – are all above the standard $\alpha = 0.05$ significance threshold. This indicates that the residuals do not significantly deviate from normality in any fold, and the normality assumption holds statistically for this well. It is worth noting that while the statistical tests confirm normality, visual inspections still reveal slight asymmetries in the distribution, which could influence model behavior in sensitive applications.

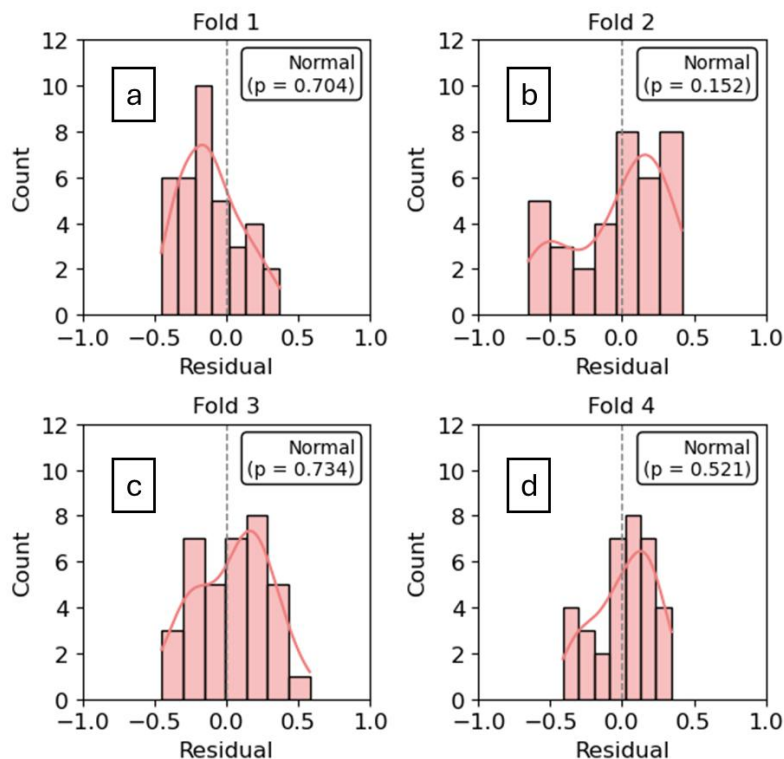


Figure 5-26-Histograms of test residuals for Well 293-2 across four cross-validation folds (a–d), overlaid with Kernel Density Estimate (KDE) curves. Each subplot reports the p -value from the Kolmogorov–Smirnov (K–S) test for normality. In all cases, p -values exceed the 0.05 threshold, indicating no significant deviation from a normal distribution. Despite slight asymmetries, especially in folds 2 and 3, the residuals statistically satisfy the normality assumption.

For all other wells in the area, the results are provide in Table 5.9 where the p -values of K-S test are provided.

Table 5.9- P -values of the K-S test for all of the wells

Well	Fold 1	Fold 2	Fold 3	Fold 4	Well	Fold 1	Fold 2	Fold 3	Fold 4
293-2	0.70	0.15	0.73	0.52	442-1	0.86	0.79	0.38	0.51
293-3	0.99	0.23	0.76	0.58	244-2	0.44	0.56	0.69	0.50
439-1	0.49	0.48	0.87	0.80	245-2	0.82	0.25	0.97	0.66
440-1	0.70	0.53	0.53	0.74	443-1	0.20	0.79	0.20	0.34

5.4.4 Homoscedasticity

To assess whether the linear regression model satisfies the assumption of homoscedasticity, the Breusch-Pagan test was applied to test residuals across four cross-validation folds. Figure 5.27 (subplots a-d) shows the test residuals plotted against the normalized predicted GWLs for each fold. Based on the p-values shown in the legend of each subplot, Folds 1 (a) and 2 (b) show significant heteroscedasticity ($p = 0.013$), violating the constant variance assumption. Folds 3 (c) and 4 (d) show no evidence of heteroscedasticity ($p = 0.182$ and $p = 0.964$, respectively), indicating that the assumption holds in these cases. The inconsistent performance across folds reflects poor reliability of the linear model in this well and supports the need for more robust models.

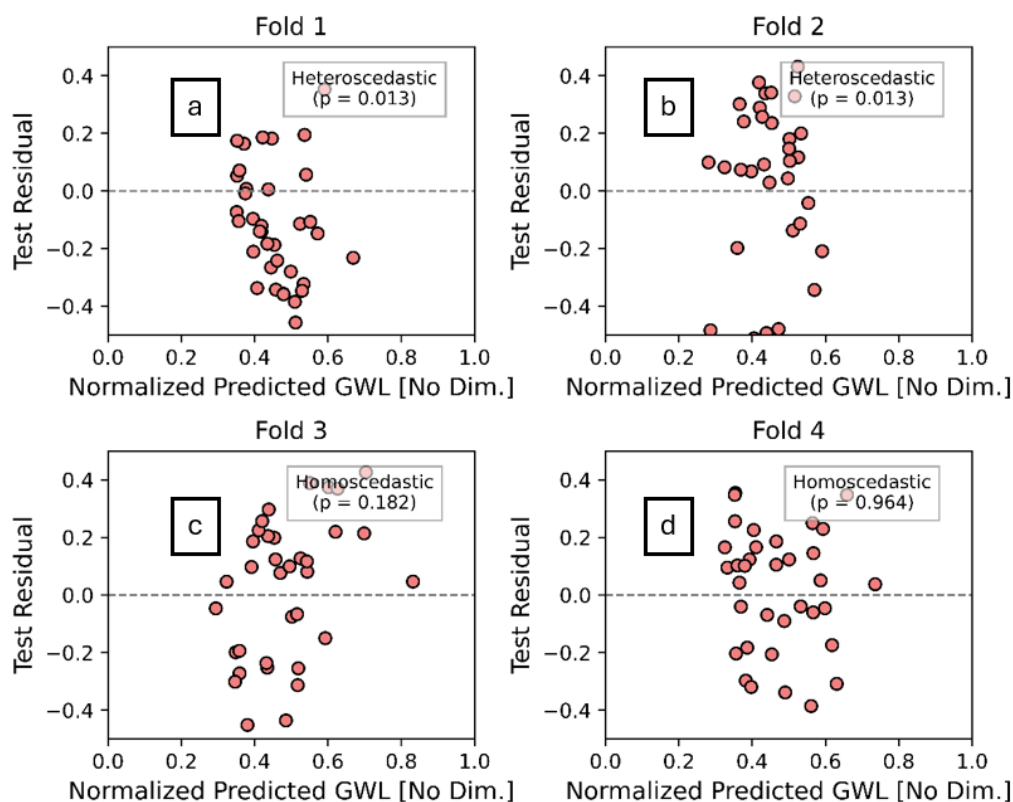


Figure 5-27-Residuals versus normalized predicted GWL values across four cross-validation folds for well 293-2. The Breusch-Pagan test results are annotated on each subplot to assess homoscedasticity. Folds (a) and (b) show heteroscedastic behavior ($p < 0.05$), while folds (c) and (d) meet the homoscedasticity assumption. This variability highlights instability in the linear model's error structure across splits.

To assess the linear regression model's performance across the study area, the Breusch–Pagan test was applied to all eight wells using 4-fold cross-validation. Table 5.10 shows the resulting p-values. Low p-values ($\alpha < 0.05$) indicate heteroscedasticity, meaning the model fails to maintain constant residual variance. Some wells consistently exhibit heteroscedasticity across folds, while others show mixed or acceptable results. Overall, the variability across wells highlights the model's limited reliability and supports the need for more flexible approaches like neural networks.

Table 5.10-P-values of the B-P test for all of the wells

Well	Fold 1	Fold 2	Fold 3	Fold 4	Well	Fold 1	Fold 2	Fold 3	Fold 4
293-2	0.01	0.01	0.18	0.96	442-1	0.11	0.07	0.72	0.22
293-3	0.004	0.03	0.56	0.67	244-2	0.86	0.55	0.48	0.65
439-1	0.005	0.56	0.95	0.65	245-2	0.16	0.44	0.04	0.85
440-1	0.93	0.89	0.60	0.31	443-1	0.12	0.001	0.33	0.46

5.4.5 Autocorrelation in Residuals

In Figure 5.28, the autocorrelation of residuals for each of the four cross-validation folds is shown for well 293-2. These plots visualize the autocorrelation function (ACF) up to lag 30, allowing assessment of the temporal dependency within the residuals of the linear regression model. In an ideal linear regression model, residuals should be uncorrelated, meaning the ACF values would hover around zero and fall within the 95% confidence band (shaded region). However, all four plots exhibit a clear pattern of positive autocorrelation at early lags that gradually declines, with some values extending beyond the confidence bands. This indicates that residuals are not entirely independent across time.

In classical regression analysis, such autocorrelation violates the Gauss-Markov assumption of uncorrelated errors, potentially leading to underestimated standard errors and overconfident predictions. However, in this study, the model is designed to be autoregressive in nature, meaning it intentionally incorporates lagged values of the target variable as predictors. Therefore, some degree of autocorrelation in

the residuals is expected and not necessarily problematic. The model structure inherently acknowledges temporal dependencies, and as a result, strict residual independence is not a primary assumption.

In the first fold for well 293-2, the ACF shows a strong second-highest peak at lag 1 with a coefficient of 0.718, just after the expected value of 1 at lag 0. The ACF then steadily declines until approximately lag 9, indicating a gradual loss of temporal dependence. However, beyond lag 9, the ACF values begin to rise again, forming a periodic pattern with a local maximum around lag 17 and a local minimum near lag 23. Similar patterns are observed in the remaining three folds. Each displays a strong second-highest autocorrelation at lag 1, with coefficients of 0.916, 0.720, and 0.841 for folds 2, 3, and 4, respectively. These consistently high lag-1 values across folds highlight the autoregressive nature of the signal. All folds exhibit periodic behavior beyond the initial lags, indicating that GWL variations are seasonally driven and influenced by recurring factors such as rainfall, snowmelt, agricultural activities, and groundwater consumption which introduce predictable fluctuations into the data.

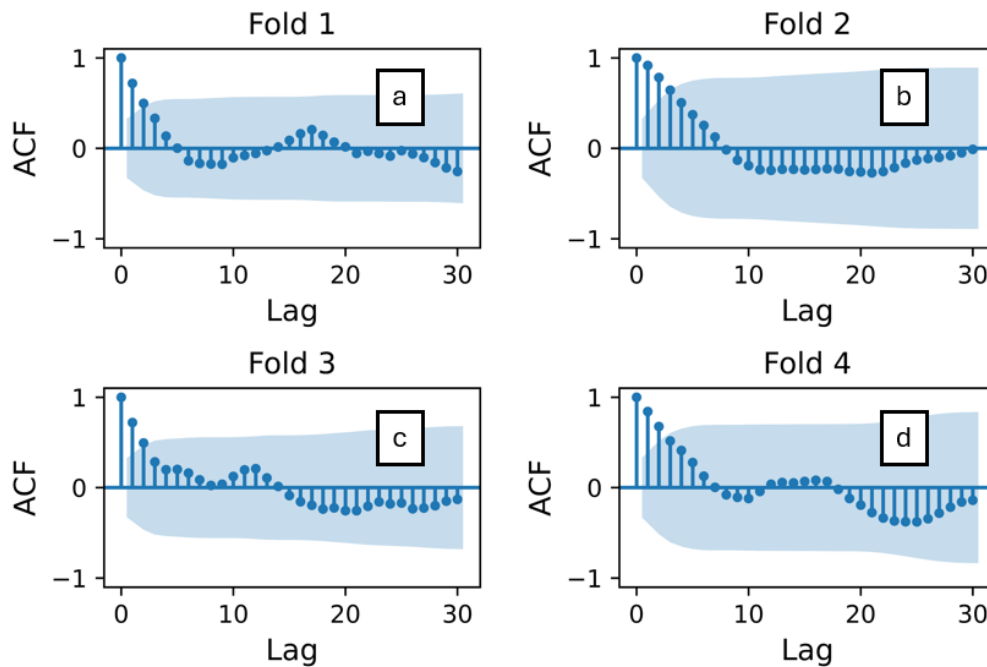


Figure 5-28-Autocorrelation function (ACF) plots of residuals for each cross-validation fold for well 293-2. Each subplot (a–d) corresponds to a separate fold and shows the autocorrelation values up to lag 30. The blue shaded region indicates the 95% confidence interval. Residuals show noticeable positive autocorrelation at early lags in all folds, reflecting the autoregressive nature of the modelling framework.

The ACF coefficients for all 8 wells are summarized in Table 5.11. For each fold, the second highest ACF value up to lag 30 is reported. In every case, this second peak occurs at lag 1, indicating strong month-to-month autocorrelation in the residuals across all wells and folds.

Table 5.11-Maximum ACF coefficients for all of the wells

Well	Fold 1	Fold 2	Fold 3	Fold 4	Well	Fold 1	Fold 2	Fold 3	Fold 4
293-2	0.72	0.92	0.72	0.84	442-1	0.35	0.60	0.46	0.75
293-3	0.79	0.80	0.60	0.77	244-2	0.93	0.47	0.75	0.78
439-1	0.57	0.68	-0.45	0.40	245-2	0.77	0.74	0.54	0.52
440-1	0.57	0.81	0.50	0.89	443-1	0.82	0.39	0.91	0.48

5.4.6 Linear Regression Results for All Single Well Scenarios

The results of the linear regression assumption checks for all eight wells in the study area are summarized in Table 5.12. Overall, the linearity assumption was not satisfied for any of the wells upon implementing the VIF. Normality of residuals, evaluated using the Kolmogorov-Smirnov test, was satisfied in all cases, suggesting that deviations from linearity did not severely distort the residual distribution. The residual mean was close to zero in most wells, confirming that the models were unbiased on average. Homoscedasticity, tested with the Breusch–Pagan test, was satisfied in half of the wells, indicating that in the remaining cases, the variance of residuals changed across predicted values. Finally, autocorrelation was detected in most wells, except for well 440-1, where the Durbin-Watson statistic and the Breusch-Godfrey test indicated no significant correlation between residuals over time. The presence of autocorrelation in the wells, especially at lag 1, reflects the temporal dependencies in the data and aligns with the autoregressive nature of GWL modelling.

These findings highlight the limitations of linear regression in this context and motivate the need for more flexible models that can better capture the temporal and nonlinear characteristics of GWL fluctuations.

Table 5.12- Summary of linear regression assumption diagnostics for all eight wells. ✓ indicates that the assumption is satisfied based on statistical testing, while × indicates violation.

Well	Linearity	Normality	Mean	Homoscedasticity	Autocorrelation
293-2	×	✓	✓	✓	×
293-3	×	✓	✓	×	×
244-2	×	✓	✓	✓	×
245-2	×	✓	✓	✓	×
443-1	×	✓	×	×	×
439-1	×	✓	×	×	×
440-1	×	✓	✓	✓	✓
442-1	×	✓	✓	×	×

5.5 Artificial Neural Networks Results

Given that key regression assumptions – such as linearity, homoscedasticity, and exogeneity – were violated in several cases, a more flexible and robust modelling approach was required. FFNNs, being less sensitive to these assumptions, were adopted to model the GWL changes under various input configurations.

5.5.1 Performance Plateau

In this study, an artificial FFNN with a single hidden layer was employed. The number of neurons in this layer was varied systematically to identify the optimal configuration for each model. Initially, models were tested with up to 12 neurons. If no noticeable improvement in performance was observed – based on evaluation metrics like R^2 or MSE – the configuration was considered to have reached a performance plateau. In scenarios where model performance continued to improve with increasing neurons, the network size was extended to 24 neurons to ensure sufficient model capacity. Regardless of early convergence, all models were trained up to 24 neurons to maintain consistency across wells, and final decisions about optimal complexity were made based on the evaluation results.

To explore how SSA timing influences model complexity and neuron selection, we performed two sets of experiments: one using inputs smoothed before monthly aggregation (DSM-FFNN) and another using inputs smoothed after aggregation (DMS-FFNN). For each case, we analyzed how model performance evolved as the number of hidden neurons increased, identifying the point of diminishing returns – i.e., the performance plateau. The following sections present these results separately for DSM-FFNN and DMS-FFNN.

5.5.1.1 DSM-FFNN

This section presents the DSM-FFNN results, where SSA was applied to the daily GWL series prior to aggregating the data to a monthly resolution. By smoothing earlier in the pipeline, the goal was to reduce short-term fluctuations while preserving longer-term patterns that might influence model learning. The number of hidden neurons was varied across folds to evaluate how performance metrics such as R^2 , MSE,

and AIC responded to increasing model complexity. The results help identify the performance plateau and guide the selection of optimal model size under this preprocessing setup.

Figure 5.29 illustrates this process for well 293-2, showing the distribution of R^2 scores across four cross-validation folds. Each subplot (a-d) corresponds to one fold, where training (red), validation (blue), and test (green) performances are shown. The number of neurons associated with the best performance varies between folds – for instance, 3, 5, 13, and 4 neurons were optimal for folds 1 to 4, respectively. This variation reflects the influence of different training-validation splits on the optimal model configuration.

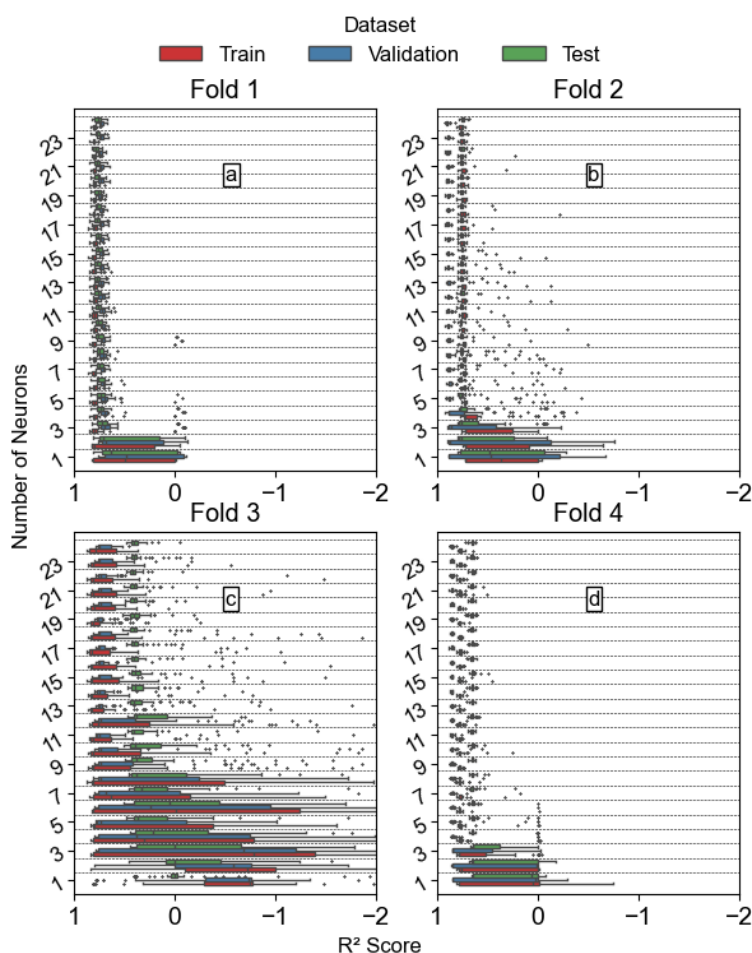


Figure 5-29- R^2 score distributions across different neuron counts for well 293-2 in each cross-validation fold. Each subplot (a–d) represents a separate fold, with training (red), validation (blue), and test (green) results shown. The optimal number of neurons varies across folds—from as low as 3 in Fold 1 to higher values such as 11 in Fold 3—demonstrating fold-dependent behavior. While early folds show a clear plateau at lower neuron counts, Fold 3 exhibits gradual improvements over a wider range, highlighting the variability in model performance across folds.

To identify the most robust overall configuration, results from all folds were combined. This approach avoids over-reliance on fold-specific behaviors and focuses on generalizable model performance. As shown in Figure 5.30, the combined R^2 distributions for well 293-2 narrow and peak around 4 neurons. Although some improvements continue up to 17 neurons, the combined R^2 performance and its consistency already stabilize by neuron 8. Therefore, the model with 4 neurons is selected as the optimal point to avoid unnecessary model complexity and limit the number of parameters, which would increase with additional neurons without providing meaningful performance gains.

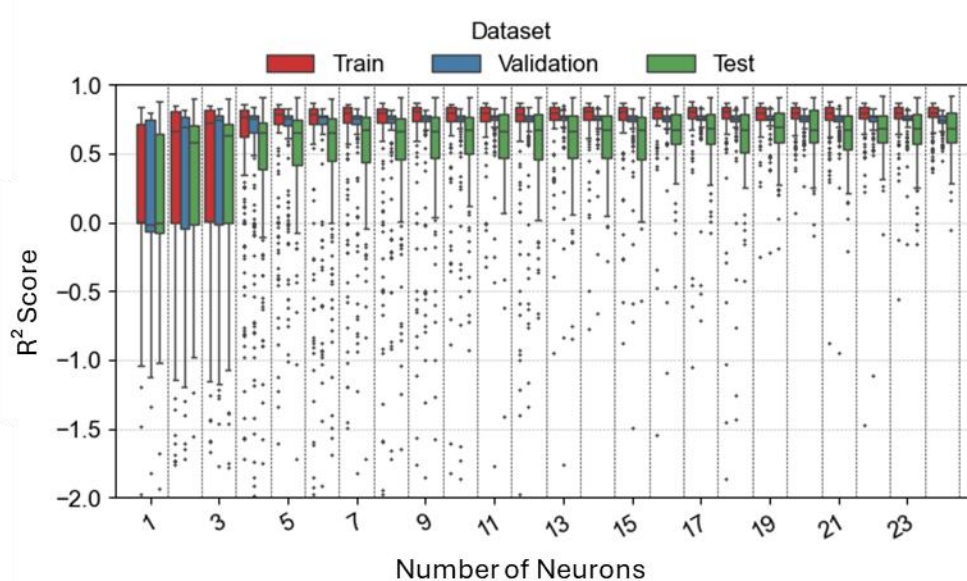


Figure 5-30- Combined R^2 score distributions for well 293-2 across training (red), validation (blue), and test (green) sets at different neuron counts. While performance begins to plateau around 4 neurons, overall improvements in R^2 and reduced variability continue up to 8 neurons, indicating it as the optimal point.

Table 5.13 summarizes the results for all wells, showing the optimal number of neurons for each fold as well as the overall best configuration based on combined R^2 distributions. Considerable variability is evident between folds. For example, well W293-2 shows optimal values ranging from 3 to 15 neurons across folds, while the final model used 8 neurons. In contrast, wells like W439-1 and W442-1 consistently favor higher neuron counts (21 and 23 neurons respectively), suggesting that more complex patterns in

these wells require deeper network capacity. Additionally, runtime durations increase with larger neuron counts, as seen in wells W244-2 and W245-2.

Table 5.13- Summary of run times and optimal number of neurons selected for each model based on R^2 performance. Fold-specific values indicate the neuron count that achieved the highest R^2 on the validation set, while the overall value reflects the point of performance plateau when combining results across all folds. Models are grouped by watershed, with the final row representing the aggregated model trained on all wells.

Model	Run time (min)	Number of Neurons				
		Fold 1	Fold 2	Fold 3	Fold 4	Overall
W293-2	1072	3	5	13	4	5
W293-3	1367	8	8	15	4	8
Watershed 1 (W1)	1265	3	4	11	4	17
W439-1	638	14	13	12	16	22
W440-1	817	11	8	14	9	24
W442-1	679	12	7	14	9	23
Watershed 2 (W2)	891	22	23	23	24	24
W244-2	1610	19	9	6	3	18
W245-2	1739	11	5	3	7	23
Watershed 3 (W3)	1768	24	12	13	14	22
W443-1	728	20	21	16	19	23
Aggregated	1761	24	24	24	24	24

While R^2 captures the proportion of explained variance, MSE provides insight into absolute error. Figure 5.31 displays MSE distributions for well 293-2 across the four folds. Again, optimal neuron counts differ per fold. Folds 1, 2, and 4 show minimum MSE at lower neuron counts (around 3-5), whereas Fold 3 exhibits a broader improvement curve, suggesting that a higher number of neurons contributes to error reduction.

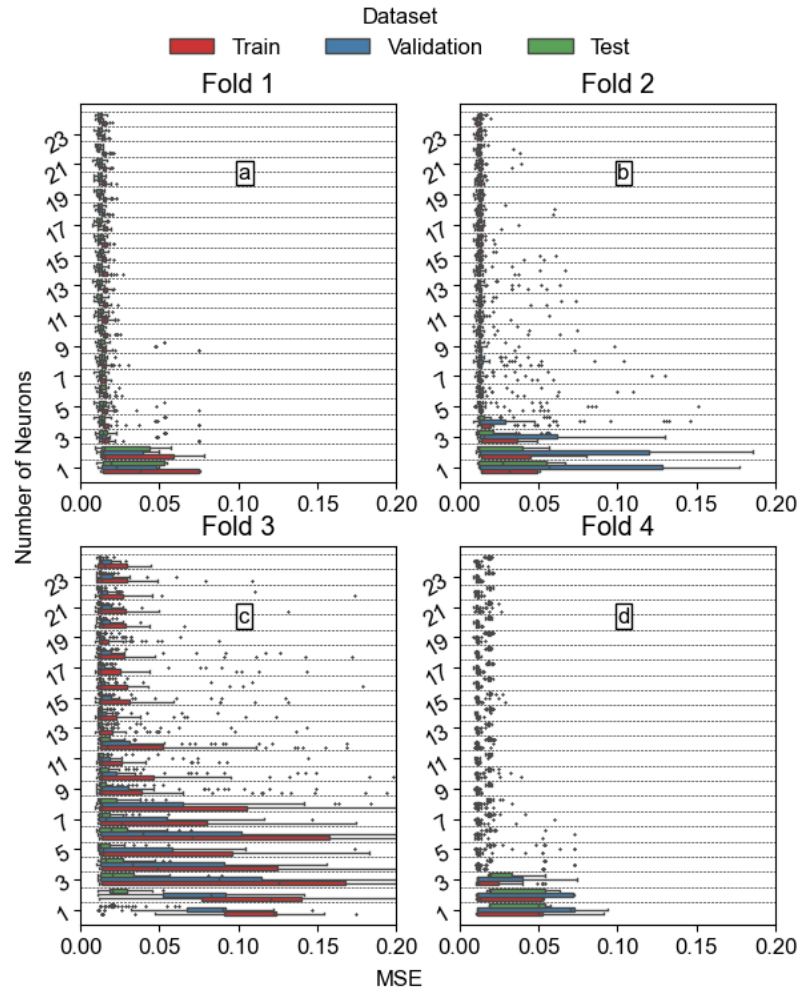


Figure 5-31-MSE distributions across neuron counts for well 293-2 in each cross-validation fold. Subplots (a–d) show Folds 1 to 4 with training (red), validation (blue), and test (green) results. The optimal neuron count varies by fold, with minimum MSE at lower counts (around 3–4) in Folds 1, 2, and 4. In contrast, Fold 3 shows a broader range with gradual improvement.

The combined MSE outcomes (Figure 5.32) support this pattern, showing a steady decrease up to about 4 neurons. Beyond this point, improvements become marginal, and the curve noticeably levels off. The reduced variability around 4 neurons further supports selecting it as the optimal configuration for well 293-2.

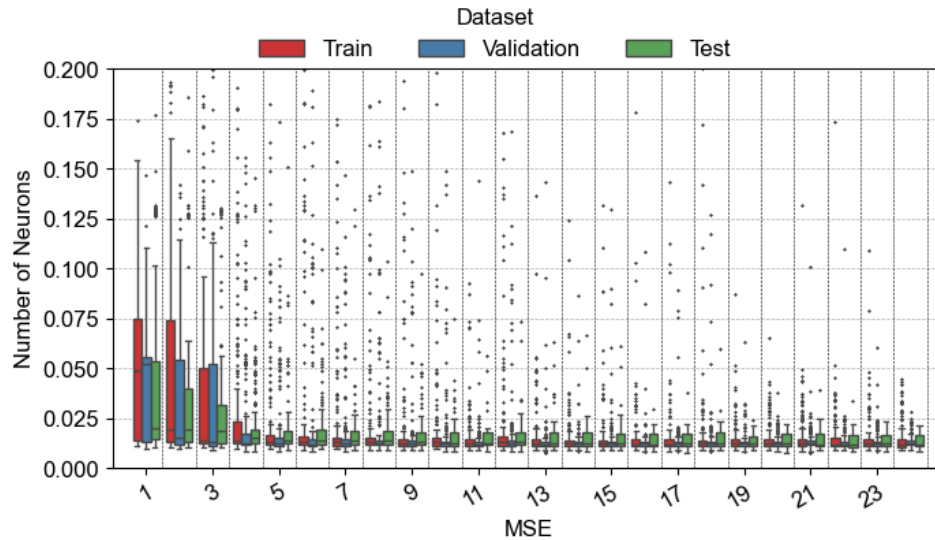


Figure 5-32-Combined MSE distributions for well 293-2 across all cross-validation folds, shown for different neuron counts. Each boxplot represents the training (red), validation (blue), and test (green) performance. While the MSE drops sharply up to around 4 neurons, smaller improvements continue beyond that point. Based on the flattening trend and reduced variability, the optimal number of neurons is identified around 8.

Table 5.14 presents the optimal number of neurons determined for each fold, along with the overall selection based on the combined MSE results. While individual fold minima differ, the final choice reflects the point where the MSE curve levels off and shows reduced variability. Moreover, the selected neuron count should also include the optimum found using the R^2 metric, ensuring that the final configuration maintains strong overall performance without reverting to a suboptimal model.

Table 5.14- Summary of run times and optimal number of neurons selected for each model based on MSE performance. Fold-specific values indicate the neuron count that resulted in the lowest MSE on the validation set, while the overall value reflects the point where adding more neurons no longer meaningfully reduced error. Models are grouped by watershed, with the final row showing results for the aggregated model trained on all wells.

Model	Run time (min)	Number of Neurons				
		Fold 1	Fold 2	Fold 3	Fold 4	Overall
W293-2	1072	3	5	13	4	5
W293-3	1367	6	8	15	6	8
Watershed 1 (W1)	1265	3	5	7	5	17
W439-1	638	13	13	13	13	22
W440-1	817	10	8	9	20	24
W442-1	679	18	9	13	6	24
Watershed 2 (W2)	891	21	24	23	24	24
W244-2	1610	13	11	8	3	8
W245-2	1739	9	5	3	3	12
Watershed 3 (W3)	1768	16	12	7	7	22
W443-1	728	14	15	9	15	23
Aggregated	1761	14	22	12	13	24

As described in Section 4.6, the AIC was used as a supporting metric. Figure 5.33 shows AIC values across different neuron counts and cross-validation folds for well 293-2. As expected, the lowest AIC values generally occur at smaller neuron counts (around 1–4), especially in Folds 1, 2, and 4. Fold 3 shows a broader range of competitive values, but still favors smaller architectures. This trend reinforces a core principle: while more neurons may slightly improve accuracy metrics like R^2 or MSE, they also increase the number of parameters and may lead to overfitting.

The model selected for well 293-2 uses 8 neurons, based primarily on performance metrics. Although the AIC does not reach its minimum at this exact size, its value remains within an acceptable range, indicating that this configuration does not introduce unnecessary complexity. In this context, AIC serves as a diagnostic tool – not to select the model – but to ensure that the chosen configuration achieves a reasonable trade-off between fit and simplicity. In some wells, AIC values may increase notably with larger neuron counts; however, when R^2 and MSE scores clearly favor those higher-capacity models, selecting a larger architecture is necessary despite the AIC penalty. This reflects a practical compromise where predictive performance takes priority, while still using AIC to avoid excessive overfitting.

Figure 5.34 displays the combined AIC results across folds. Again, the lowest AIC values tend to appear at smaller neuron counts (typically 3–5), and the curve flattens beyond 8 neurons. This plateau suggests diminishing returns with added complexity, making it clear that increasing neuron counts beyond this point is unlikely to produce meaningful improvements.

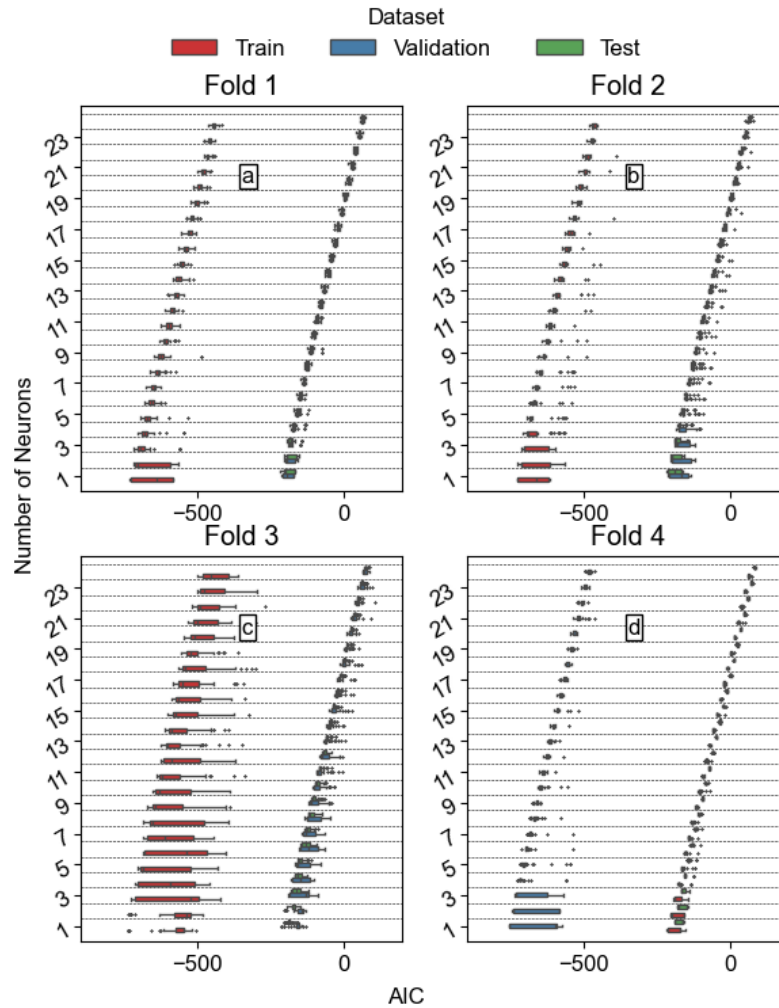


Figure 5-33-AIC distributions across neuron counts for well 293-2 in each cross-validation fold. Subplots (a–d) show Folds 1 to 4, with training (red), validation (blue), and test (green) datasets. Lower AIC values generally occur at smaller neuron counts (around 1–4), especially in Folds 1, 2, and 4. Fold 3 exhibits a broader range of low AIC scores across more neurons.

The combined AIC results (Figure 5.34) align with earlier findings. The lowest AIC scores are found between 3–5 neurons. While AIC increases with model size, the rate of increase slows after 8 neurons – marking this as a reasonable trade-off point between accuracy and complexity.

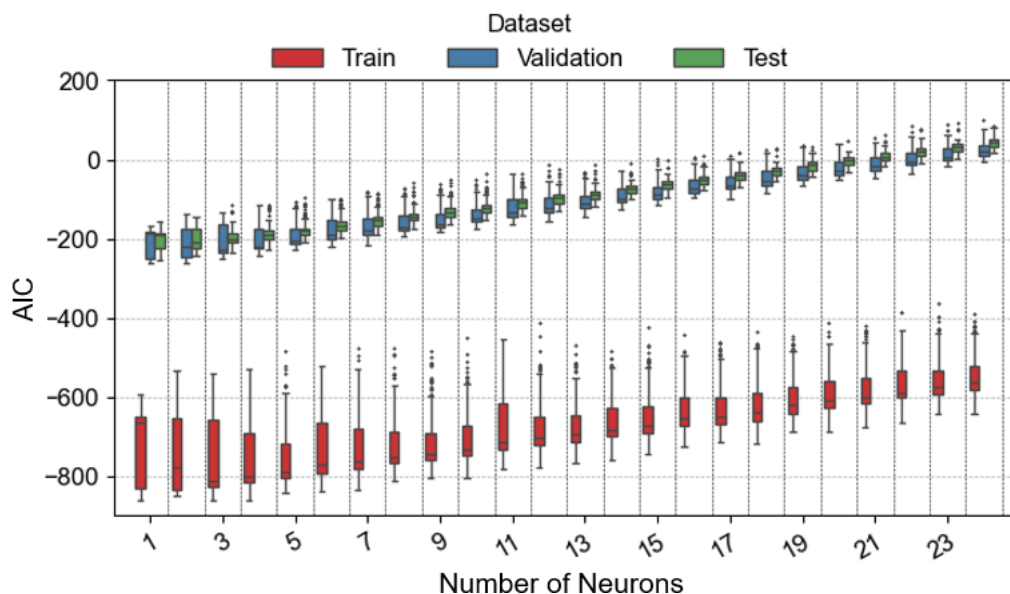


Figure 5-34- Combined AIC distributions for well 293-2 across different neuron counts and datasets. Boxplots show AIC values for training (red), validation (blue), and test (green) sets across all folds. The lowest AIC scores occur at smaller neuron counts, particularly around 3–5 neurons. Although AIC values rise steadily with increasing model size, the curve flattens after neuron 8, indicating diminishing gains in model fit relative to complexity.

Overall, AIC is a useful diagnostic tool for assessing model parsimony by penalizing unnecessary complexity. It often favors smaller architectures, with the lowest values typically observed around 3–5 neurons. However, model selection should not rely solely on AIC. In cases where R^2 and MSE continue to improve with increased model size, it may be appropriate to accept a higher AIC value in exchange for better predictive performance. This perspective is supported by hydrological modelling literature, including Pérez-Sánchez et al. (2019), who emphasize using AIC as a guiding rather than deciding criterion. Similarly, Boo et al. (2024) note that most GWL modelling studies rely on multiple evaluation metrics – typically two to four – which supports the multi-metric approach in this study. Uc-Castillo et al. (2023) further recommend combining different metrics to mitigate individual limitations and enhance overall performance assessment. In line with these findings, we use AIC as a secondary metric to help detect unnecessary complexity without overriding performance-driven decisions.

5.5.1.2 DMS-FFNN

To further investigate the effect of SSA timing on model performance, the same modelling process was repeated using inputs smoothed after the daily data had been aggregated to monthly resolution. In this setup, SSA was applied using a three-month window to reduce short-term variability in the monthly series. This approach generally produced smoother input curves than the daily-SSA scenario. To keep the presentation clear and focused, only the combined distributions of evaluation metrics (R^2 , MSE, and AIC) are shown for this case, as they more directly reflect the final model selection criteria. Since the purpose here is to compare the overall effectiveness of this SSA strategy rather than evaluate fold-specific variations, detailed fold-wise plots were omitted to avoid redundancy and simplify interpretation.

To evaluate the effect of applying SSA after monthly aggregation (DMS setup), the model was retrained using inputs smoothed with a 3-month SSA window. As seen in Figure 5.35, the R^2 scores for train, validation, and test sets gradually improve with neuron count. The gains level off around 7-9 neurons, suggesting this as the range of optimal complexity. Compared to the DSM setup (Figure 5.30), where SSA was applied to daily data before aggregation, the DMS results are slightly more stable, with fewer extreme outliers, particularly in the test set. This indicates that post-aggregation smoothing may help reduce variability and improve generalization, although the differences in median R^2 are small.

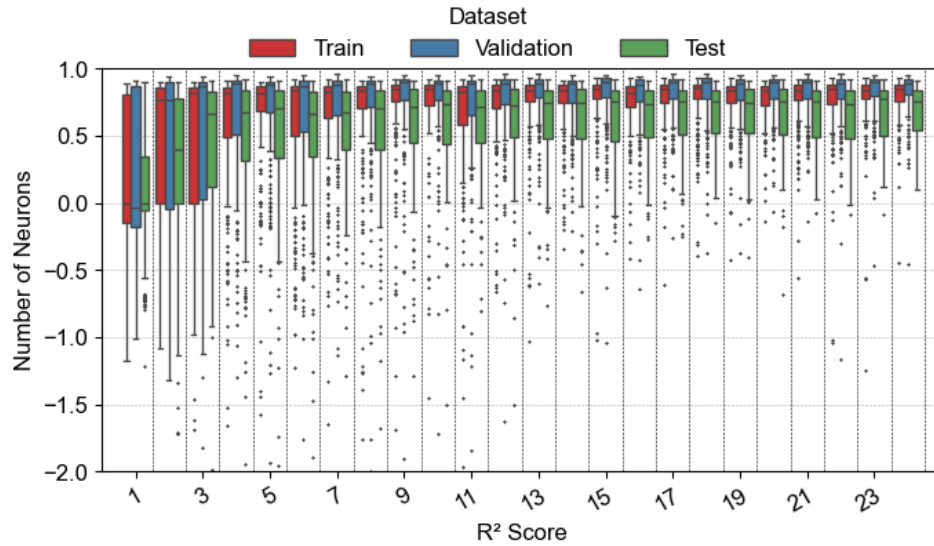


Figure 5-35- Combined R^2 distributions for the DMS-FFNN setup across different neuron counts. Each boxplot shows training (red), validation (blue), and test (green) results. The MSE decreases sharply up to around 4 neurons, with smaller gains and reduced variability continuing to 8 neurons.

Table 5.15 presents the overall optimal neuron count based on the combined R^2 results for each well. Fold-wise details for DMS are not shown (see Appendix 7.5 to 7.11 for the fold-wise and combined plots), and the combined values alone are used for comparison.

Table 5.15-The optimum neuron for each well in the DMS-FFNN scenario

Model	Optimum Neuron	Model	Optimum Neuron
293-2	24	Watershed 2 (W2)	24
293-2	13	244-2	24
Watershed 1 (W1)	11	245-2	24
439-1	24	Watershed 3 (W3)	22
440-1	24	443-1	22
442-1	23	Aggregated	23

To compare the impact of SSA timing on model accuracy, the MSE distributions were examined under both preprocessing setups. Building on the earlier results, Figure 5.36 presents the MSE distributions for the DMS scenario. As the number of neurons increases, MSE generally decreases and the spread narrows, particularly between 3 and 7 neurons. Compared to the earlier DSM setup, DMS shows slightly lower variability and more stable performance across the train, validation, and test sets. These patterns suggest that applying SSA after monthly aggregation not only maintains accuracy but may also improve consistency across different folds.

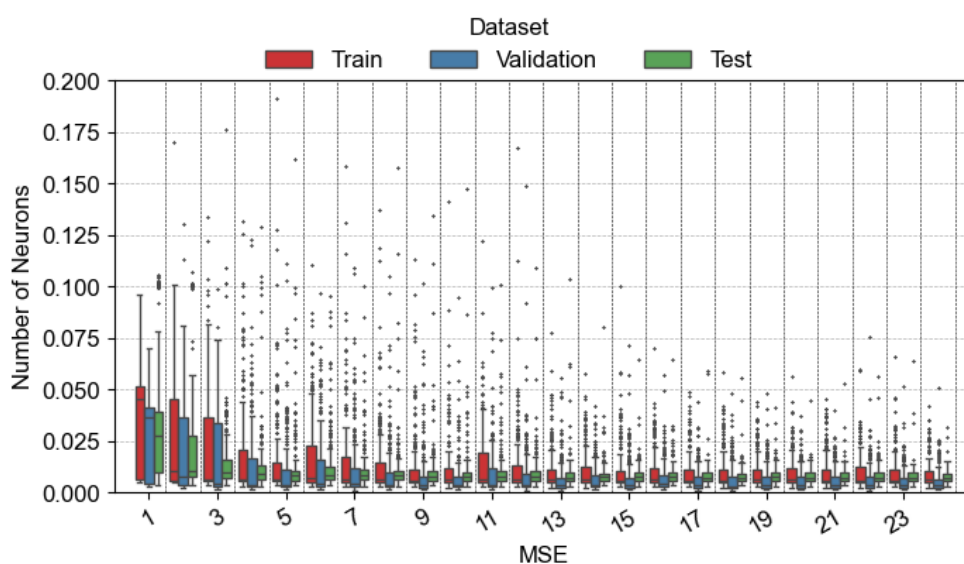


Figure 5-36- Combined MSE distributions for the DMS-FFNN setup across different neuron counts. Each boxplot shows training (red), validation (blue), and test (green) results. The MSE decreases sharply up to around 4 neurons, with smaller gains and reduced variability continuing to 8 neurons.

To complete the comparison between the two SSA setups, the AIC distributions were also examined for the DMS scenario. Figure 5.37 presents the AIC distributions for the DMS setup. Similar to the DSM case shown earlier, lower AIC values appear at smaller neuron counts, especially between 1 and 5. However, in DMS, the AIC scores show more consistent separation between training and testing sets, indicating improved generalization. Although AIC tends to increase with neuron count, the selected model with 8

neurons remains within a reasonable range, confirming that the chosen architecture balances model fit and complexity, without excessive overfitting.

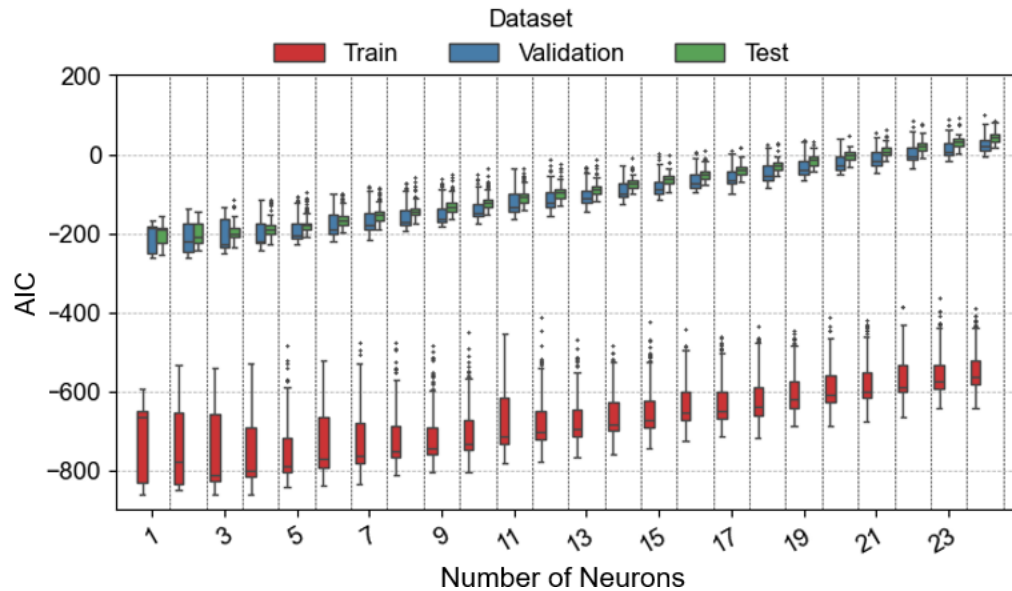


Figure 5-37- Combined AIC distributions for the DMS-FFNN setup across different neuron counts. Boxplots show AIC values for training (red), validation (blue), and test (green) sets. The lowest AIC scores appear around 3–5 neurons, while values rise steadily with model size.

Overall, these findings support the use of AIC as a secondary criterion – helpful for detecting unnecessary complexity but not overriding improvements in predictive performance measured by R^2 and MSE. As with the DSM setup, the final DMS model selection reflects a balance between accuracy, stability, and parsimony, guided by multiple evaluation metrics.

5.5.2 Single, Combined, or Aggregated?

5.5.2.1 DSM-FFNN

To assess how model performance is influenced by the spatial structure of the training data, we evaluated three modelling strategies: single models, combined models, and the aggregated model trained on all available wells. These strategies allow us to explore the trade-offs between increasing sample size

and preserving spatial specificity, particularly in terms of generalization performance on validation and test data.

Figure 5.38 displays the distribution of R^2 scores for training, validation, and test sets across all model scenarios. Each watershed (W1 to W3) includes results from individual wells and a combined model trained on all wells within the watershed. Watershed 4 (W4) contains only one well (443-1) and is therefore represented solely by its single-well model. The rightmost group shows the aggregated model trained on all wells across all watersheds. The number of available observations for each model is shown as a gray line, while box annotations denote the number of input features (f), optimal hidden neurons (n), and output targets (o).

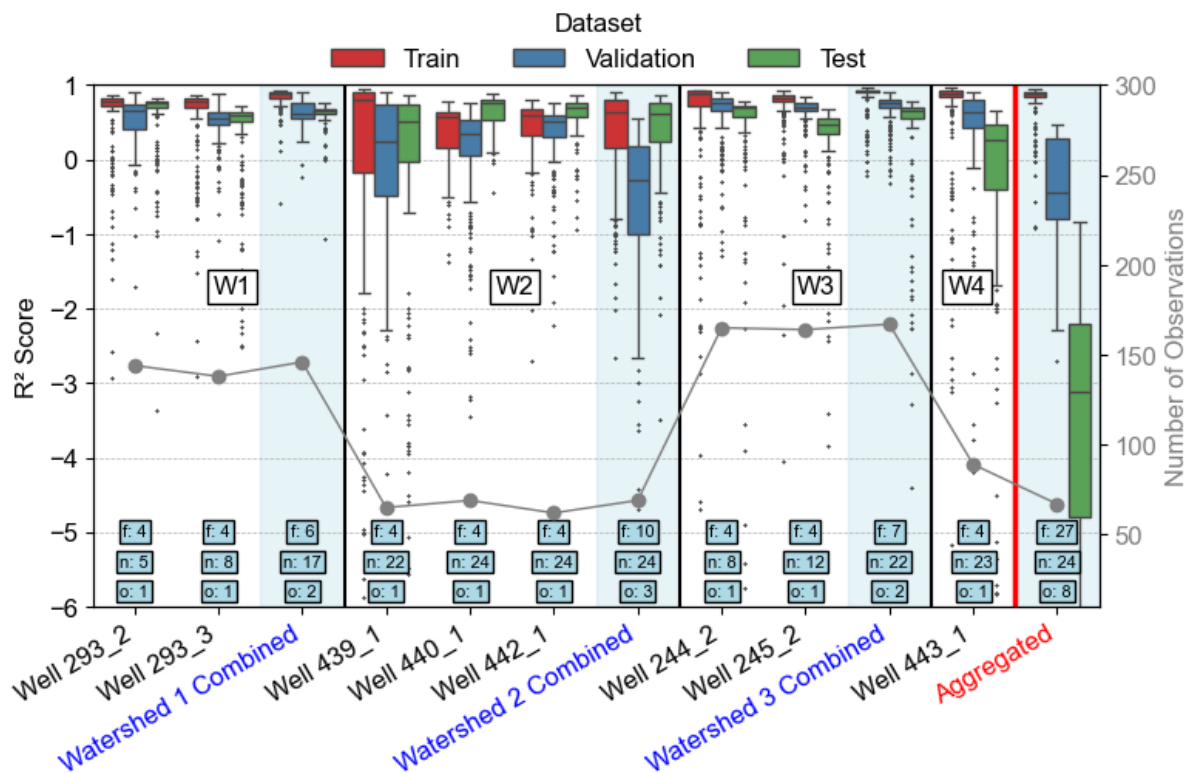


Figure 5-38-Boxplots of R^2 scores across training (red), validation (blue), and test (green) datasets for all model configurations, grouped by watershed. Each group includes results from individual wells and their respective combined model. The final group on the right shows the fully aggregated model trained on all wells. The gray line (right y-axis) indicates the number of available observations per model. Box annotations show the number of features (f), optimal number of neurons in the hidden layer (n), and number of output targets (o).

As shown in Table 5.16, the single models consistently achieved high and stable performance on the test set. Well 293-2 achieved the best overall results, with a median test R^2 of 0.74 and a very narrow IQR of 0.15, suggesting stable and generalizable performance. Well 293-3 followed with a median of 0.62 and an IQR of 0.26. These values reflect good balance between model fit and generalization. The W1 combined model, aggregating data from both wells, preserved much of this strength, with a median of 0.69 and a moderate IQR of 0.26.

In Watershed 2, performance was more variable. Well 439-1 yielded a median R^2 of 0.57 with a large IQR of 1.08, reflecting high inconsistency across training runs. Wells 440-1 and 442-1 performed slightly similar in terms of stability, with medians of 0.54 and 0.59, and IQRs of 0.47 and 0.29, respectively. However, the W2 combined model suffered from the poorest performance among all combined models, with a low median R^2 of 0.29 and a wide IQR of 1.01. This suggests that combining these wells introduced conflicting hydrological patterns that the model failed to reconcile, leading to instability and weak generalization.

By contrast, Watershed 3 showed strong and consistent performance. Wells 244-2 and 245-2 achieved medians of 0.73 and 0.68, with IQRs of 0.27 and 0.32, respectively. Their combined model (W3) also performed well, with a median R^2 of 0.75 and a relatively low IQR of 0.25, indicating that this watershed benefited from internal similarity between wells. Well 443-1 in W4, a single model, showed moderate performance with a median of 0.59 and a higher IQR of 0.65, which likely reflects data scarcity and local variability.

The aggregated model, which includes input from all wells across all watersheds, performed the worst. Despite having access to the largest number of wells, its median test R^2 was negative (-0.45), and its IQR was extremely wide (3.07). This poor performance stems from a key limitation: although the dataset includes all available wells, the training period is constrained by wells with fewer observations – particularly those in Watershed 2. To ensure temporal consistency, the model can only be trained on timestamps where all wells simultaneously report data, drastically reducing the effective sample size and

leading to overfitting. This issue is clearly visible in the Figure 5.38, where Watershed 2 wells show much lower observation counts (gray line). As a result, the model is forced to generalize from fewer, less representative data points, which amplifies noise and spatial heterogeneity.

These findings are summarized in Table 5.16, which presents the test R^2 median, first quartile (Q1), third quartile (Q3), and interquartile range (IQR) for each model scenario.

Table 5.16- Summary statistics (Median, Q1, Q3, and IQR) of test set R^2 scores for all model configurations. The table includes results for single models, combined models, and the aggregated model.

Scenario	Median	Q1	Q3	IQR
293-2	0.74	0.64	0.79	0.15
293-3	0.62	0.50	0.76	0.26
Watershed 1 (W1)	0.69	0.61	0.86	0.26
439-1	0.57	-0.29	0.79	1.08
440-1	0.54	0.25	0.71	0.47
442-1	0.59	0.43	0.70	0.27
Watershed 2 (W2)	0.30	-0.29	0.72	1.01
244-2	0.73	0.61	0.89	0.27
245-2	0.68	0.49	0.81	0.32
Watershed 3 (W3)	0.75	0.65	0.90	0.25
443-1	0.59	0.21	0.86	0.65
Aggregated	-0.45	-2.24	0.84	3.07

In conclusion, model performance tends to deteriorate as the level of spatial aggregation increases. Single models consistently demonstrate the highest stability and accuracy, likely because they can capture well-specific hydrological patterns. Combined models can also perform effectively when the contributing wells exhibit similar behaviors, as seen in Watersheds 1 and 3, where the wells are nested and more uniform.

However, in cases like Watershed 2, where well dynamics differ, combining data leads to reduced performance. Although the aggregated model utilizes the most features and wells, its ability to generalize is limited by inconsistencies in both spatial characteristics and the timing of available observations. These findings emphasize a key point: increasing data volume does not guarantee better performance if spatial and temporal consistency is compromised.

To complement the analysis of model performance using R^2 scores, we also evaluated the MSE across the same modelling configurations: single models, combined models, and the aggregated model. MSE provides a more direct measure of the average prediction error magnitude, making it especially useful when assessing absolute model accuracy and error stability.

Figure 5.39 presents boxplots of the MSE values for training, validation, and test sets across all scenarios. The layout is consistent with the R^2 plot: wells are grouped by watershed (W1 to W4), with combined models placed at the end of each group and the aggregated model shown on the far right. As with the R^2 Figure 5.38, the gray line tracks the number of observations available for each model, and the inset boxes indicate the number of features (f), optimal hidden neurons (n), and output targets (o).

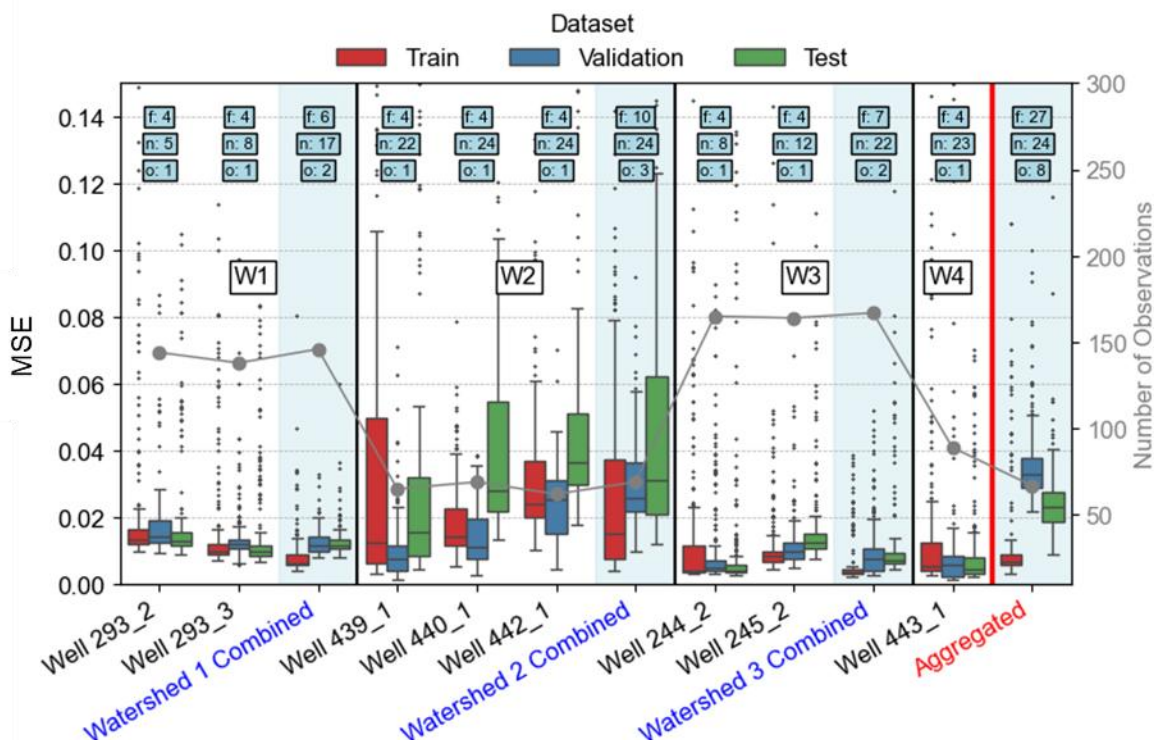


Figure 5-39- Boxplots of Mean Squared Error (MSE) scores across training (red), validation (blue), and test (green) datasets for all model configurations, grouped by watershed. Each group includes results for individual wells and their corresponding combined model. The rightmost group shows the fully aggregated model trained on all wells. The gray line (right y-axis) indicates the number of available observations per model. Box annotations indicate the number of features (f), optimal number of neurons (n), and number of output variables (o).

In general, the trends observed in R^2 analysis are echoed in the MSE results. The single models in Watershed 1 (Well 293-2 and Well 293-3) show low median MSE values (0.0132 and 0.0104, respectively) and tight interquartile ranges (IQRs of 0.0060 and 0.0041), reflecting consistent and low-error predictions. The combined model for Watershed 1 maintains comparable performance, with MSE in the middle (median = 0.0107, IQR = 0.0046), reinforcing the earlier conclusion that combining data from wells with similar dynamics (i.e., nested wells) can still result in a reliable model.

Watershed 2 shows more variability. While Well 439-1 has a moderate median MSE (0.0117), its wide IQR (0.0216) indicates instability across different data splits. Other wells, like 440-1 and 442-1, show higher median errors (0.0190 and 0.0297) and wide IQRs (0.0162 and 0.0169), reflecting both lower

accuracy and poor stability. The combined model for Watershed 2 has the highest overall MSE (median = 0.0321, IQR = 0.0298), confirming that combining wells with dissimilar temporal or spatial behavior reduces model performance and generalizability.

Watershed 3 and 4 show performance similar to Watershed 1. Both individual wells (244-2 and 245-2) achieve low median MSEs (0.0042 and 0.0101) and relatively narrow IQRs, suggesting high accuracy and robustness. The Watershed 3 combined model remains competitive (median = 0.0064, IQR = 0.0052), and Well 443-1 from Watershed 4 also maintains strong performance (median = 0.0052, IQR = 0.0065), despite being the only well in its group.

The aggregated model, trained on all wells, again shows poor performance. With a median MSE of 0.0240 and a wide IQR of 0.0244, it displays significant variability and reduced predictive reliability. This is likely due to inconsistencies in data availability across wells – especially those with sparse records, like in Watershed 2 – since the aggregated model must be trained on overlapping time steps. These misalignments limit the model’s ability to generalize, highlighting the challenge of combining heterogeneous sources without sufficient temporal coverage.

Table 5.17 summarizes the detailed statistics for each scenario, including the median, first quartile (Q1), third quartile (Q3), and IQR of the test MSE values.

Table 5.17- Summary statistics (Median, Q1, Q3, and IQR) of test set MSEs for all model configurations. The table includes results for single models, combined models, and the aggregated model.

Scenario	Median	Q1	Q3	IQR
293-2	0.0132	0.0119	0.0179	0.0060
293-3	0.0104	0.0089	0.0130	0.0041
Watershed 1 (W1)	0.0108	0.0084	0.013	0.0044
439-1	0.0103	0.0062	0.0266	0.0203
440-1	0.0190	0.0114	0.0277	0.0162
442-1	0.0297	0.0218	0.0387	0.0169
Watershed 2 (W2)	0.0259	0.0177	0.0431	0.0254
244-2	0.0040	0.0035	0.0062	0.0027
245-2	0.0101	0.0081	0.0131	0.0050
Watershed 3 (W3)	0.0064	0.0038	0.0090	0.0052
443-1	0.0055	0.0031	0.0099	0.0068
Aggregated	0.0240	0.0089	0.0327	0.0239

In summary, both R^2 and MSE analyses point to consistent patterns: single-well models tend to deliver more stable and accurate predictions, while the aggregated model underperforms due to inconsistencies in spatial and temporal data. However, some discrepancies emerge between the two metrics in certain models. For example, the R^2 for Well 439-1 indicates moderate median performance but high instability, whereas its MSE appears more stable and less severe. Similarly, the W2 combined model shows a slightly better R^2 than the aggregated model, yet its MSE remains among the highest. These differences reflect the fact that R^2 captures relative fit while MSE quantifies absolute error. Therefore, interpreting both metrics together provides a more balanced evaluation of model generalization and robustness. This dual perspective reinforces the earlier conclusion: combining data from wells with similar hydrological behavior (as in W1 and W3) can enhance model performance but forced aggregation across heterogeneous or

temporally misaligned sources (as in W2 and the aggregated model) degrades both accuracy and consistency. Similar findings were reported by Müller et al. (2021).

5.5.2.2 DMS-FFNN

To evaluate the effect of applying SSA after aggregating the daily data to monthly resolution (DMS setup), the models were retrained using inputs smoothed with a 3-month SSA window. This approach, which reduces short-term fluctuations while preserving broader trends, offers a clearer signal for the model to learn from the inputs.

Figure 5.40 shows the distribution of R^2 scores for training, validation, and test sets across all scenarios. Watersheds W1 to W3 include both individual wells and a combined model, while W4 (with a single well) appears only as a single-well model. The rightmost group represents the aggregated model. The gray line tracks the number of observations, and the box annotations show the number of input features (f), hidden neurons (n), and output targets (o) per model.

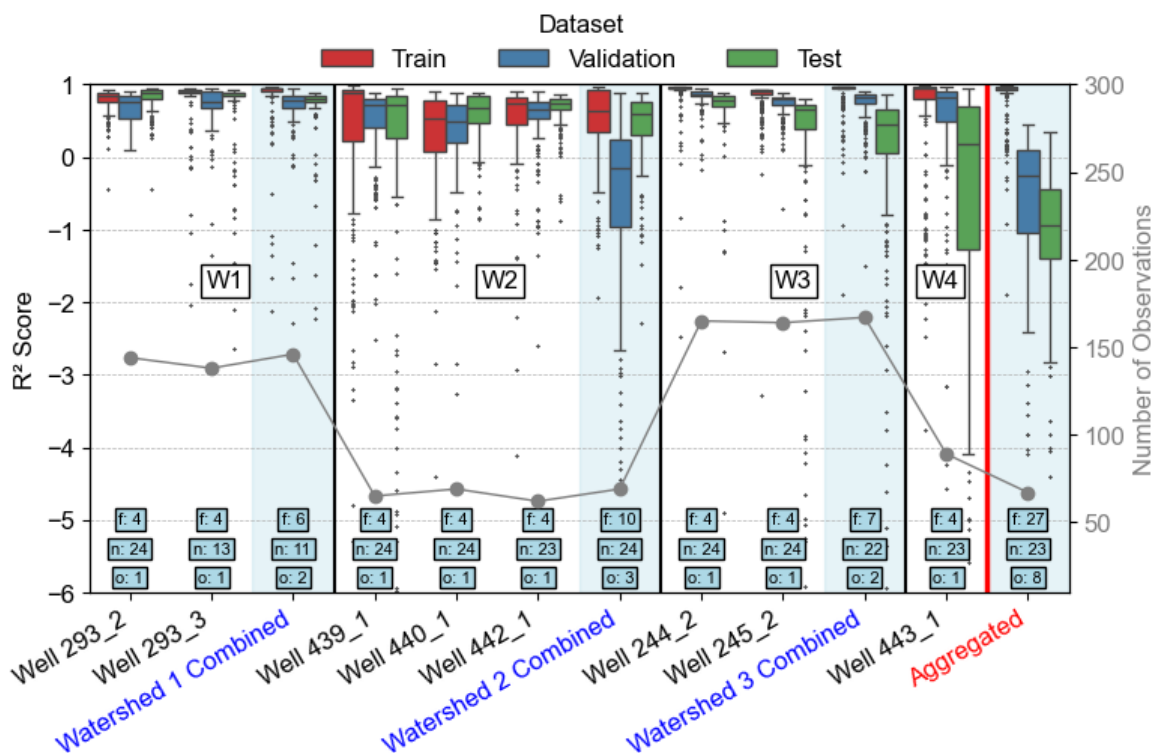


Figure 5-40-Distribution of R^2 scores across training (red), validation (blue), and test (green) sets for all DMS-FFNN model configurations. Each group includes results for single-well models, combined watershed models, and the aggregated model trained on all wells. The gray line shows the number of observations per model. Box annotations indicate the number of input features (f), selected hidden neurons (n), and output targets (o).

Following a similar modelling approach to DSM-FFNN, the DMS-based single models also demonstrated consistently strong and reliable performance on the test set, as summarized in Table 5.18. Models trained on W1 wells generally performed well in the DMS setting. Wells 293-2 and 293-3 achieved high median R^2 scores of 0.84 and 0.88, with IQRs of 0.16 and 0.12, respectively. These represent clear improvements over the DSM models, where the corresponding medians were 0.74 and 0.62, with wider IQRs of 0.15 and 0.26. This suggests that DMS inputs enhanced both accuracy and consistency for these nested wells. The W1 combined model under DMS also showed a higher median R^2 of 0.83 (IQR: 0.17) compared to 0.69 (IQR: 0.26) in DSM, reinforcing the idea that spatial proximity and data aggregation within a nested watershed were effectively leveraged under DMS.

In W2, the DMS models showed noticeable improvements over their DSM counterparts. For instance, the single model for well 439-1 achieved a median R^2 of 0.75 under DMS, significantly higher than the 0.57 reported for DSM. Despite the improvement, the IQR remained wide (0.52 in DMS vs. 1.08 in DSM), indicating variability in model stability across folds. Similarly, well 440-1 showed a slightly improved median R^2 (0.58 in DMS vs. 0.54 in DSM), with the IQR increasing slightly from 0.47 to 0.53, suggesting moderate improvement in central performance but continued variation. Well 442-1 also benefited from the DMS input structure, increasing its median R^2 from 0.59 (DSM) to 0.72 (DMS), with a small decrease in IQR (0.27 to 0.25), reflecting both improved and more stable predictions. At the watershed level, the combined W2 model improved from 0.30 (DSM) to 0.37 (DMS) in median R^2 , with a reduction in IQR from 1.01 to 0.82. These results highlight that for W2, which includes wells with more complex or variable dynamics, the DMS approach provided more reliable and accurate modelling outcomes compared to the DSM inputs.

In W3, the DMS models also outperformed their DSM equivalents, both in terms of accuracy and consistency. For well 244-2, the median R^2 increased from 0.73 (DSM) to 0.87 (DMS), while the IQR dropped from 0.27 to 0.13, indicating a substantial gain in performance and model stability across the test folds. Well 245-2 showed a similar trend, with its median R^2 rising from 0.68 to 0.77 and IQR decreasing from 0.32 to 0.24 under the DMS setup. These improvements demonstrate that the DMS approach helped capture the temporal and spatial patterns in W3 more effectively than DSM. The combined model for W3 also saw an increase in median R^2 from 0.75 (DSM) to 0.80 (DMS). However, in contrast to single models, the IQR widened from 0.25 to 0.47, suggesting increased variability across folds when wells were combined in the DMS configuration. Despite this rise in IQR, the overall accuracy remained strong, and the consistent gains across all three W3 models.

For Watershed 4 (W4), which includes only well 443-1, the DMS-based model also outperformed its DSM counterpart. The median R^2 improved from 0.59 (DSM) to 0.78 (DMS), showing a notable gain in predictive accuracy. However, this came with a trade-off in consistency: the IQR increased from 0.65 to

0.87, indicating greater variation in performance across folds. This suggests that while the DMS approach enhanced the model's ability to capture the general behavior of the GWL for this well, the performance remained sensitive to data splits, possibly due to limited data availability or distinct characteristics of the well.

For the aggregated model, which was trained on all wells combined into a single dataset, performance remained poor across both the DMS and DSM settings. While the DMS model showed a slight improvement, with the median R^2 increasing from -0.45 (DSM) to -0.12 (DMS), the result was still negative, indicating that the model generally failed to make accurate predictions. The IQR also remained high in both cases, though reduced from 3.07 (DSM) to 1.90 (DMS), suggesting some gain in consistency but still substantial variability across folds. This reflects the challenges of applying a single generalized model to wells with different hydrogeological behaviors.

Despite having the same number of input samples for each well (due to time alignment), the model struggled to capture the well-specific dynamics, especially in wells with fewer observations or more complex temporal patterns. These results reinforce the limitations of spatial aggregation when local variability is important and emphasize the value of using spatially distinct models, particularly under the DMS structure. Table 5.18 summarizes these results by presenting the median, first quartile (Q1), third quartile (Q3), and IQR of test R^2 scores for each modelling scenario.

Table 5.18-Summary of test set R^2 performance for DMS-FFNN models across different spatial configurations. Values represent the median, first quartile (Q1), third quartile (Q3), and interquartile range (IQR) of R^2 scores across all ensemble runs.

Scenario	Median	Q1	Q3	IQR
293-2	0.84	0.73	0.89	0.16
293-3	0.88	0.80	0.91	0.12
Watershed 1 (W1)	0.83	0.76	0.93	0.17
439-1	0.75	0.35	0.87	0.52
440-1	0.58	0.26	0.79	0.53
442-1	0.72	0.55	0.80	0.25
Watershed 2 (W2)	0.37	-0.08	0.74	0.82
244-2	0.87	0.81	0.94	0.13
245-2	0.77	0.64	0.87	0.24
Watershed 3 (W3)	0.80	0.48	0.96	0.47
443-1	0.78	0.06	0.93	0.87
Aggregated	-0.12	-0.97	0.94	1.90

Overall, the R^2 analysis of DMS-FFNN models demonstrated improved and more stable predictive performance compared to the DSM-FFNN counterparts across most scenarios. Single models under DMS achieved higher median R^2 scores and narrower IQRs, particularly in nested wells and regions with sufficient training data, suggesting enhanced model generalization due to better input structuring. Combined models also showed moderate gains in performance and stability. In contrast, the aggregated model remained the least effective under both input strategies, likely due to its inability to capture localized variations. These comparisons reinforce the advantage of using DMS inputs, particularly for preserving site-specific temporal patterns and improving the reliability of test set predictions. Having established this,

we now turn to a complementary evaluation using MSE to assess model performance from an error-focused perspective.

To further evaluate model accuracy and stability, the distribution of MSE values were examined across the test sets for all DMS-FFNN model scenarios (Figure 5.41). While R^2 offers insight into the proportion of explained variance, MSE provides a direct measure of the average squared difference between predicted and actual GWLs, making it sensitive to large errors. Lower MSE values indicate more precise predictions. Similar to the R^2 analysis, MSE results were summarized for individual wells, combined models within each watershed, and the aggregated model.

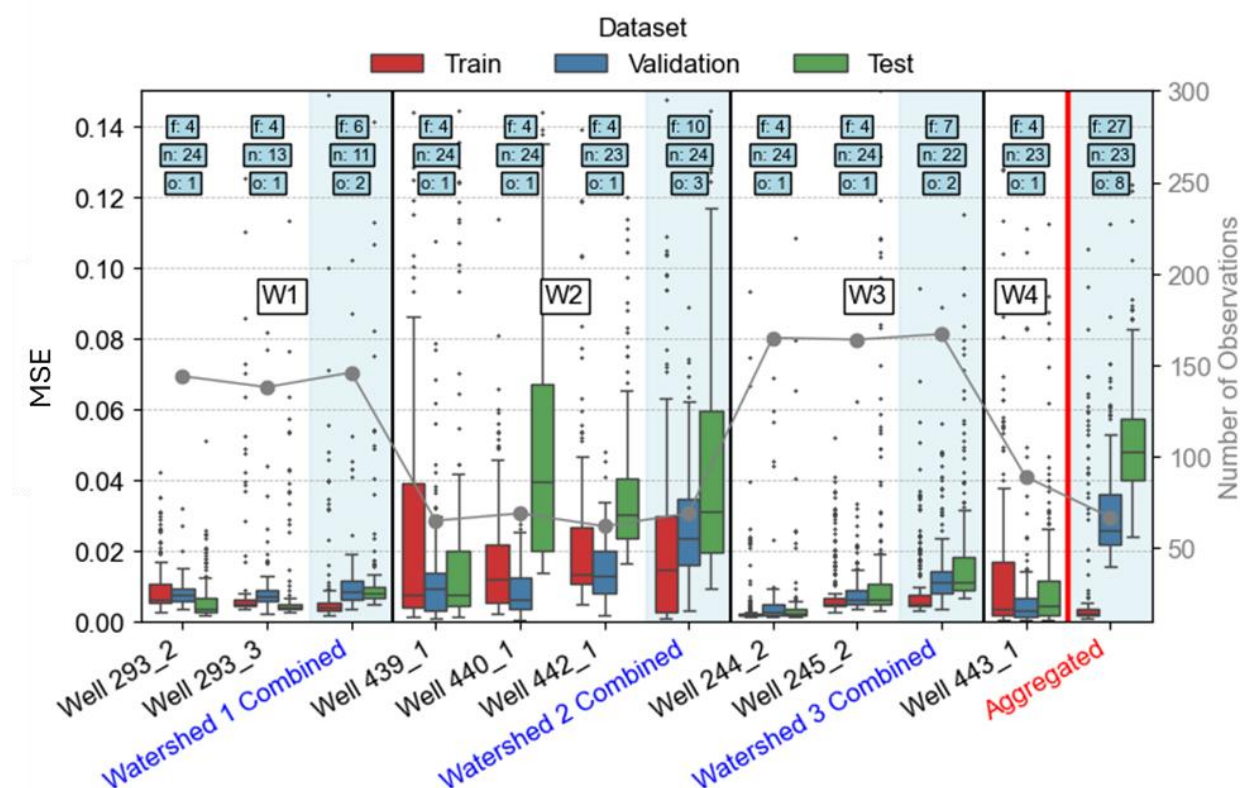


Figure 5-41-Box plots of test, validation, and training MSE values for all DMS-FFNN model scenarios. Each group includes models trained on individual wells, combined wells within each watershed (W1–W3), and the aggregated model trained on all wells (far right, in red). The number of observations used for training each model is shown as a gray line, while model configuration details—number of features (f), hidden units (n), and output targets (o)—are provided in boxed annotations.

For the DMS-FFNN, wells 293-2 and 293-3 achieved low MSE medians of 0.0061 and 0.0050, with narrow IQRs of 0.0047 and 0.0032, respectively. The W1 combined model had a slightly higher median of 0.0070 and an IQR of 0.0050, still indicating stable and accurate performance. Compared to the DSM-FFNN, where the IQRs were 0.0060 (293-2), 0.0041 (293-3), and 0.0044 (W1 combined), the DMS-FFNN offered a marginal improvement in central accuracy and model stability for 293-3, while remaining comparable for the other two scenarios.

The single wells showed a range of MSE medians in the second watershed: 0.0080 (439-1), 0.0155 (440-1), and 0.0205 (442-1). Notably, the IQRs were large for 439-1 (0.0154) and especially for 440-1 (0.0255), indicating more variability in model performance. The W2 combined model had an even higher median of 0.0236 and a wide IQR of 0.0258. Compared to the DSM-FFNN, the W2 combined model showed similar variability (DSM IQR = 0.0254), but DMS-FFNN slightly reduced the median error from 0.0259 to 0.0236. Still, both approaches struggled to generalize well in W2.

For the DMS-FFNN of Watershed 3, well 244-2 performed best with a very low median MSE of 0.0020 and an IQR of 0.0019. Well 245-2 had a median of 0.0055 and an IQR of 0.0045. The combined W3 model had a median of 0.0096 and an IQR of 0.0097. Compared to DSM-FFNN values (244-2: IQR = 0.0027, 245-2: 0.0050, W3 combined: 0.0052), the DMS model reduced the median MSE for well 244-2 while slightly increasing variability in the combined model. Overall, DMS-FFNN improved performance for the single wells in W3. The single well 443-1 also had a low DMS-FFNN MSE median of 0.0037 with a moderate IQR of 0.0079. Compared to the DSM-FFNN (median = 0.0055, IQR = 0.0068), the DMS model slightly improved accuracy but with slightly higher variability.

The DMS-FFNN aggregated model yielded the highest test MSE (median = 0.0291) and the widest IQR (0.0431), indicating poor and inconsistent performance across folds. Although this model was trained using data from all wells, the time alignment requirement across datasets limited the number of usable observations, so it did not actually benefit from the largest sample size. Compared to the DSM-FFNN (median = 0.0240, IQR = 0.0239), the DMS version showed both higher error and greater variability. This

result suggests that simply combining all available wells under a uniform DMS structure did not enhance performance and may have introduced additional temporal noise or incompatibility across sites.

Table 5.19-Summary of test Mean Squared Error (MSE) statistics for DMS-FFNN models across different scenarios. The table presents the median, first quartile (Q1), third quartile (Q3), and IQR for each model, including single-well, combined, and aggregated configurations.

Scenario	Median	Q1	Q3	IQR
293-2	0.0061	0.0043	0.0090	0.0047
293-3	0.0050	0.0039	0.0071	0.0032
Watershed 1 (W1)	0.0070	0.0047	0.0097	0.0050
439-1	0.0080	0.0039	0.0193	0.0154
440-1	0.0155	0.0059	0.0314	0.0255
442-1	0.0205	0.0112	0.0303	0.0191
Watershed 2 (W2)	0.0236	0.0143	0.0401	0.0258
244-2	0.0020	0.0017	0.0036	0.0019
245-2	0.0055	0.0046	0.0091	0.0045
Watershed 3 (W3)	0.0096	0.0056	0.0153	0.0097
443-1	0.0037	0.0016	0.0095	0.0079
Aggregated	0.0291	0.0033	0.0464	0.0431

In summary, the comparison of single, combined, and aggregated model configurations across DSM-FFNN and DMS-FFNN pipelines provides critical insights into how spatial and data aggregation strategies impact predictive performance. Single models often achieved the most stable and interpretable results, especially in watersheds with nested wells (e.g., W1 and W3), where their performance remained consistent across folds and both R^2 and MSE metrics demonstrated low variability. Combined models occasionally improved performance by increasing the training sample size, but this gain was not uniform

across all watersheds. In fact, in more heterogeneous settings such as W2, combining wells sometimes introduced additional variability, likely due to differences in local hydrogeologic characteristics.

The aggregated model – while trained on the largest spatial extent – did not consistently outperform more localized models. Notably, its performance suffered despite access to a higher number of features, which may reflect overfitting, increased data heterogeneity, or the challenge of generalizing across distinct watershed conditions. Importantly, the alignment of time frames across all wells constrained the actual sample size available for training the aggregated model, limiting the expected benefit of data quantity.

Furthermore, while DMS-based models consistently outperformed their DSM-based counterparts in both accuracy and stability, the trends observed across model configurations remained similar between the two setups. This consistency reinforces the generalizability of the modelling framework and emphasizes that tailoring model scope to the spatial characteristics of the data – rather than merely increasing size – is more beneficial for predictive reliability. Overall, these findings highlight the value of site-specific modelling and careful aggregation decisions, which together define the trade-off between local precision and generalizability in data-driven environmental modelling.

5.6 Model Implications

After training and calibrating the models under both SSA timing scenarios (DSM-FFNN and DMS-FFNN), the single models generally demonstrated better performance compared to combined and aggregated setups across most wells. As no clear superiority was observed between DSM and DMS in all cases – each performed better in different wells – we selected the DSM scenario as a representative example for downstream analysis. Consequently, the prediction results presented in the next subsection are based on the DSM-FFNN models, although the same procedure could be applied to DMS-FFNN with similar steps. Similarly, for the sensitivity analysis (SA), we focused on comparing single and combined models only, as

aggregated models consistently underperformed and offered limited added value for understanding feature impacts.

5.6.1 Prediction

To assess predictive performance across different spatial modelling strategies, we generated ensembles of DMS-FFNN models for each scenario (single, combined, and aggregated models). Each ensemble consisted of 4,800 models, trained using 50 random seeds, up to 24 neuron counts, and 4-fold cross-validation. All models were saved in .h5 format, enabling consistent inference using the optimal neuron configuration determined for each well or scenario. At each time point, the final prediction was calculated as the median of ensemble outputs to ensure robustness against outliers and enhance stability:

$$\tilde{y} = \text{median}(\hat{y}_1, \hat{y}_2, \dots, \hat{y}_n) \quad (5.1)$$

While both DSM-FFNN and DMS-FFNN models showed strong performance in different cases, DMS-FFNN generally outperformed DSM-FFNN across watersheds and configurations. As such, the prediction results presented here are based on DMS-FFNN models to provide a representative and fair comparison between the single, combined, and aggregated modelling strategies.

Figure 5.42 illustrates representative prediction results for each scenario. The black solid line shows the observed GWL, the dashed blue line indicates the ensemble median prediction, and the shaded cyan region marks the IQR across ensemble outputs. We begin by presenting the results for single models, which generally tracked observed GWLs more closely and showed narrower IQRs, reflecting lower uncertainty. Prediction results for combined and aggregated models are also shown and discussed, offering insight into how spatial generalization and data merging influence accuracy.

The single models produced highly accurate and stable predictions, closely following the observed GWL patterns across different wells. Seasonal rises and recessions were well captured by the ensemble

medians, and the uncertainty – reflected in the narrow IQRs – remained consistently low. This stability highlights the strength of localized models in learning site-specific temporal dynamics. Particularly in nested well configurations, such as those in Watershed 1 and Watershed 3, the predictions showed minimal deviation across ensemble members. Even in wells with more irregular fluctuations or limited data, the models maintained reasonable accuracy, further demonstrating the robustness of the single-model approach for well-specific forecasting.

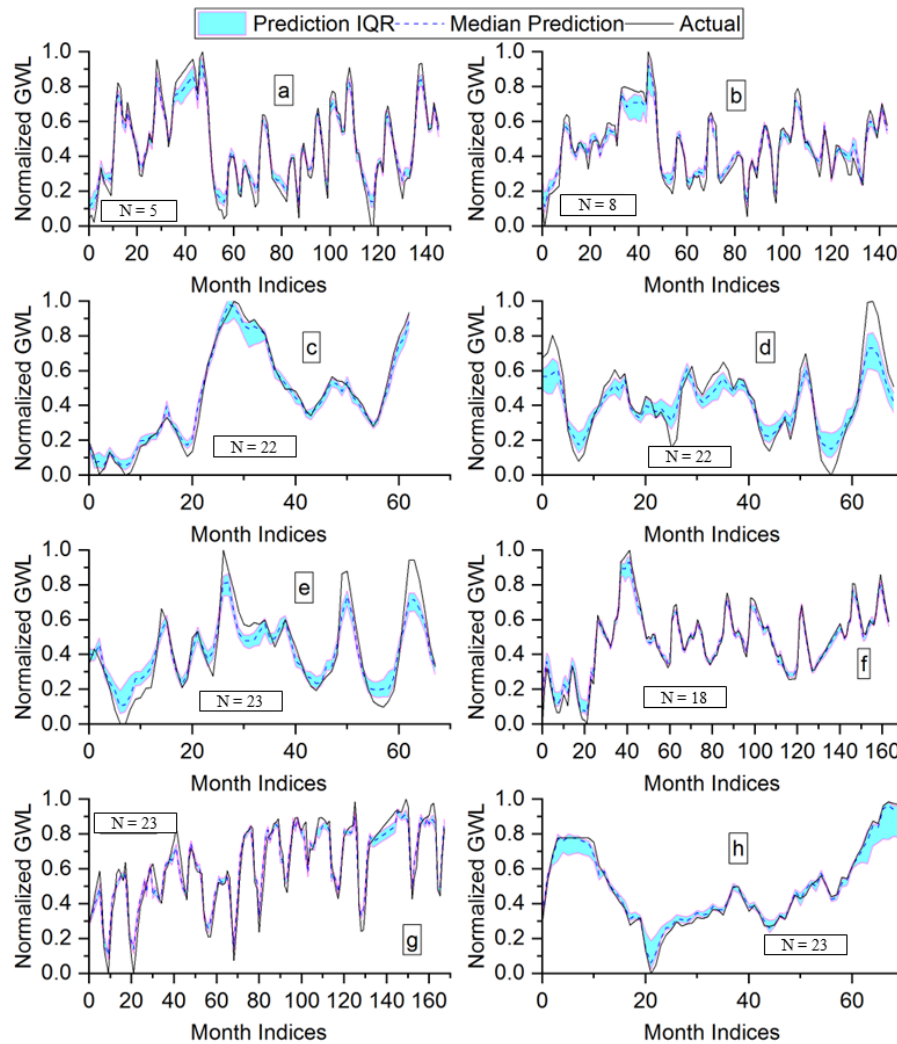


Figure 5-42-Comparison of actual and predicted normalized GWL over time for single models. The black solid line represents the actual GWL, while the blue dashed line indicates the median prediction from an ensemble of 4800 neural network models. The shaded cyan region represents the IQR of the predictions. Each subplot corresponds to a specific single model scenario: (a) 293-2, (b) 293-3, (c) 439-1, (d) 440-1, (e) 442-1, (f) 244-2, (g) 245-2, and (h) 443-1. The number of optimum neurons in the hidden layer (N) for each scenario is indicated in the respective subplot.

Similar to the single case, ensembles were generated for combined models by pooling data from all wells within each watershed. This setup maintained model diversity while increasing training data volume, again yielding 4,800 models per scenario. Predictions were made using the best-performing architecture for each combined model. Figure 5.43 displays the results, showing that the combined approach retains reasonable accuracy but exhibits wider IQRs in some wells. While it smooths extreme fluctuations better in some cases, the added data blending may obscure local patterns, reducing precision for highly variable wells.

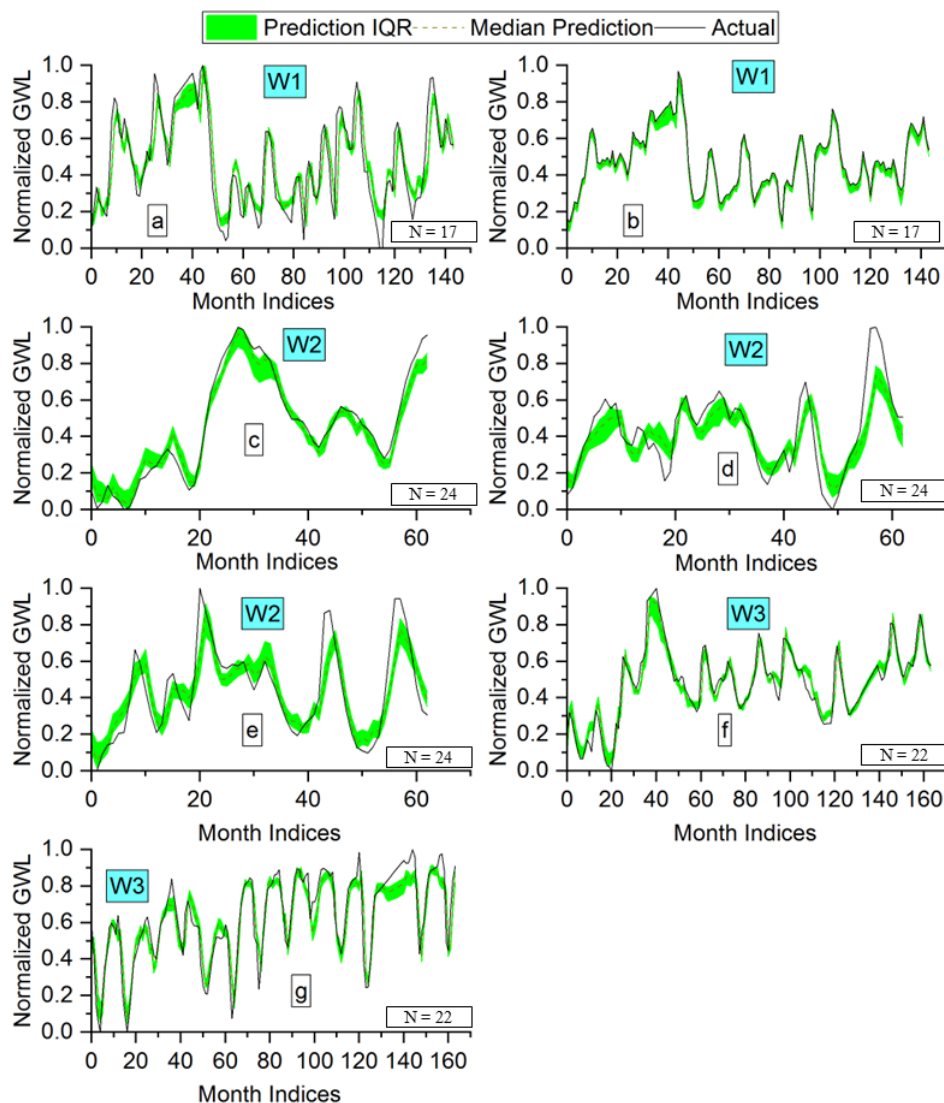


Figure 5-43- Comparison of actual and predicted normalized GWL over time for combined models. The black solid line represents the actual GWL, while the blue dashed line indicates the median prediction from an ensemble of 4800 neural network models. The shaded cyan region represents the IQR of the predictions. Each subplot corresponds to a specific single model scenario: (a) 293-2, (b) 293-3, (c) 439-1, (d) 440-1, (e) 442-1, (f) 244-2, (g) 245-2, and (h) 443-1. The number of optimum neurons in the hidden layer (N) for each scenario is indicated in the respective subplot.

A comparison between the single and combined model predictions (Figures 5.42 and 5.43) reveals several important distinctions. First, the prediction intervals represented by the IQR are generally narrower in the single model plots. This indicates that single models provide more stable and confident predictions, likely due to their exclusive training on site-specific data. In contrast, the combined models, trained across multiple wells, show wider IQRs, reflecting greater uncertainty introduced by cross-well variability.

Notably, in wells 440-1 and 442-1, the combined model displays slight temporal shifts in its median predictions compared to the actual GWL trends. These shifts are not evident in the single model results, suggesting that the combined model may struggle to precisely align predictions when site-specific dynamics differ across wells. In addition, in wells such as 440-1 and 443-1, the combined model slightly underestimates peak GWL values, whereas the single models capture these peaks more accurately. This underestimation further illustrates the limitations of a generalized model in handling localized extremes.

Moreover, the combined model predictions tend to appear smoother and less reactive to abrupt changes. While this smoothing may offer stability, it sometimes fails to reflect short-term fluctuations that are well captured by the single models. Even in wells with strong seasonal signals and longer time series, such as 245-2 and 443-1, the single model predictions remain more tightly aligned with the observed data and show less uncertainty. These visual patterns confirm the earlier findings based on quantitative metrics – R^2 , MSE, and AIC – and further support the conclusion that single models provide superior predictive accuracy and confidence when the goal is to preserve site-level temporal and statistical fidelity.

For the aggregated model, all wells were merged into a single dataset to train one generalized model across the entire region. Although this configuration increases the number of wells, the overall time frame must remain consistent across all sites. Because wells from W2 and W3 (e.g., 439, 440, 442, 443) have fewer observations – approximately 60 months – the aggregated dataset is truncated to match this shorter duration. As a result, each well contributes fewer data points, potentially limiting the model's ability to learn long-term trends. Figure 5.44 shows the ensemble-based predictions. While the median predictions follow general patterns, site-specific dynamics are less accurately captured. Wider IQRs in many plots suggest increased uncertainty, likely due to the reduced input length and diluted local variability.

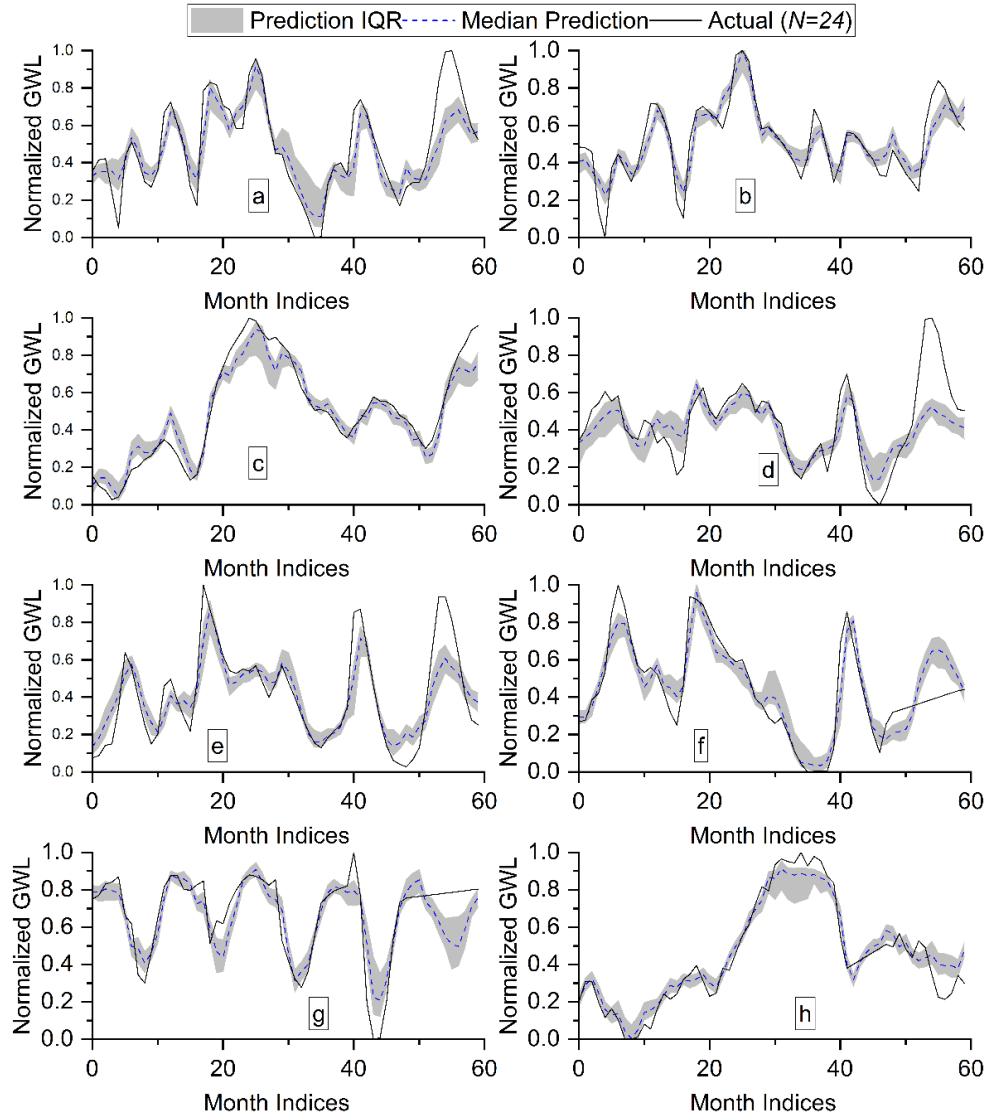


Figure 5-44- Comparison of actual and predicted normalized GWL over time for the aggregated model. The black solid line represents the actual GWL, while the blue dashed line indicates the median prediction from an ensemble of 4800 neural network models. The shaded cyan region represents the IQR of the predictions. Each subplot corresponds to a specific single model scenario: (a) 293-2, (b) 293-3, (c) 439-1, (d) 440-1, (e) 442-1, (f) 244-2, (g) 245-2, and (h) 443-1. The number of optimum neurons in the hidden layer (N) for each scenario is indicated in the respective subplot.

A comparison across the three modelling strategies shows that the aggregated model consistently underperforms relative to the combined and single-well approaches. This is most clearly reflected in its wider prediction IQRs, indicating higher uncertainty. Although the number of entries per well was equalized in the aggregated model, its generalized structure leads to a loss of local specificity. While the model can

often follow general trends, it fails to capture finer fluctuations seen in several wells. For instance, in well 440, the prediction diverges toward the end of the series. In wells 244 and 245, although the observed values are missing in the final months and were interpolated, the aggregated model predicts marked fluctuations. This behavior may suggest that, had actual observations been available, they might have confirmed similar patterns – potentially improving the model’s reliability. Overall, the decline in accuracy stems from the model’s attempt to generalize across heterogeneous wells, sacrificing precision in site-specific patterns despite using a uniform input structure.

5.6.2 Sensitivity Analysis (SA)

Beyond assessing predictive accuracy, it is also essential to understand which input features most influence model behavior. To achieve this, SA was performed on all trained DMS-FFNN models using SHAP-based attribution methods. For each scenario, the models with the optimized neuron count were reloaded from their saved .h5 files, and their parameter sets were used to evaluate feature impacts. By computing the mean absolute SHAP values across ensemble members, we identified which predictors had the greatest influence on GWL forecasts.

5.6.2.1 SHAP Results

Mean absolute SHAP values can be utilized to rank features based on their level of influence, from the most to the least significant. A similar method was employed by Chidepudi et al. (2025) and Wunsch et al., (2022) to identify the key factors impacting GWL predictions in their ANN models.

Figure 5.45 summarizes the mean absolute SHAP values for key input features across all wells, highlighting the relative influence of temperature (T), precipitation (P), hydrometric levels (H), and spatial attributes including distance and elevation differences from hydrometric and meteorological stations.

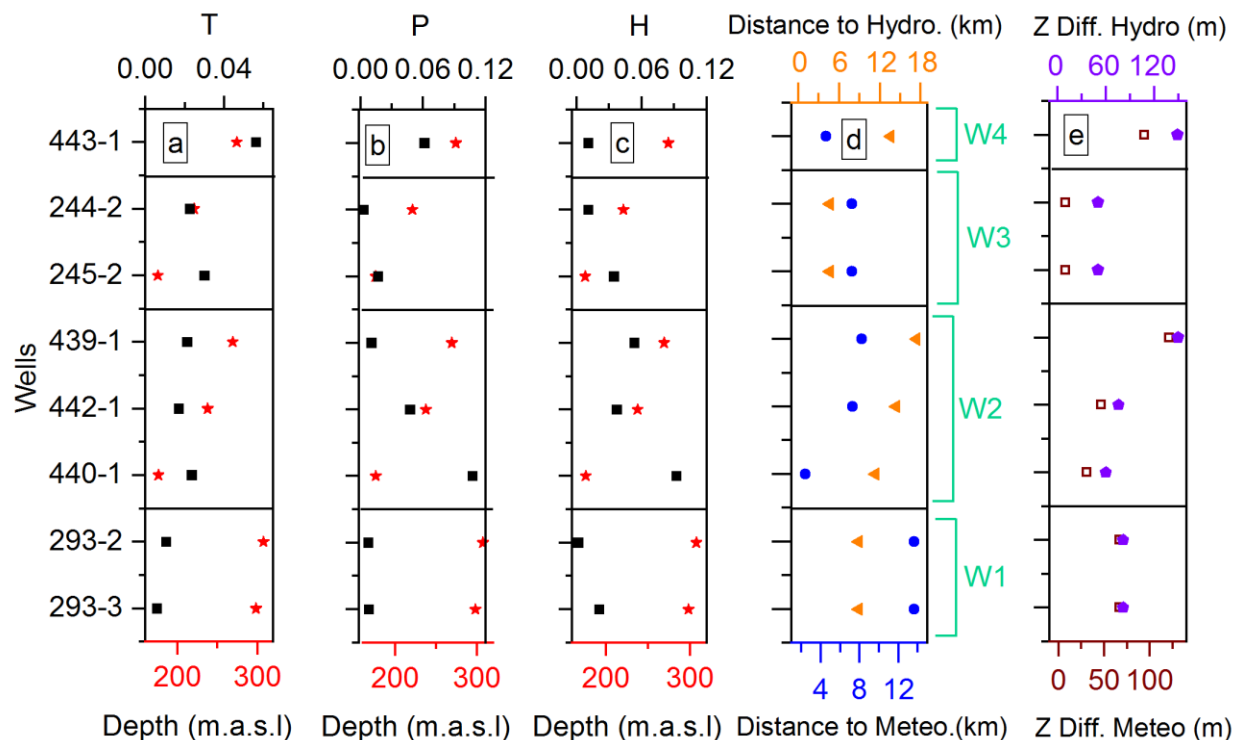


Figure 5-45-Mean absolute SHAP values for input features across eight wells, representing model sensitivity to key hydro-meteorological variables. Panels (a–c) show sensitivity to temperature (T), precipitation (P), and hydrometric level (H), with red stars indicating deeper wells and black squares representing shallower wells. Panels (d–e) display the distance to the nearest hydrometric and meteorological stations (d), and the elevation difference relative to these stations (e). Wells are grouped by watershed (W1–W4) as classified before.

Wells 293-2 and 293-3, situated in the eastern part of the study area, are both screened within the Upper Newmarket Till, as confirmed by ORMGP. While the Tier 2 assessment places both wells in the AF1-GLAF aquifer, the AQM model assigns well 293-2 to the overlying ICSD aquifer, which is geologically adjacent. Despite this minor classification difference, both wells share a similar stratigraphic setting. The SHAP analysis confirms this similarity in behavior but also highlights key differences: temperature has a stronger impact on well 293-2 (mean absolute SHAP = 0.01059) compared to 293-3 (0.0059), while precipitation influences 293-3 (0.00811) slightly more than 293-2 (0.00754). Notably, hydrometric levels have a much stronger influence on 293-3 (0.02092) compared to 293-2 (0.00151), with more than a 13-fold difference. While this may appear related to the deeper placement of 293-3, the greater sensitivity is more likely due to local hydrogeological conditions – such as stronger hydraulic connectivity between certain

confined layers and nearby lakes or streams. In contrast, 293-2 shows a higher sensitivity to temperature (0.01059 vs. 0.0059), which may counterbalance or mask the effect of hydrometric levels in that well. These results suggest that even wells screened within the same formation can display distinct sensitivity patterns due to complex vertical and lateral interactions with surface water systems and aquifer stratification. It also reinforces the ongoing observation that hydrometric influences may not act uniformly across depths, but instead interact with different layers in a non-linear and spatially variable manner.

In Watershed 2 (W2), Well 440-1 exhibits the strongest sensitivity to all three dynamic variables – T, P, and H – despite being the deepest well in the group. This heightened responsiveness in Well 440-1 can be explained by its close proximity to both the meteorological ERA5 grid node and the hydrometric station (02ED007), as shown in Figure 5.45 (panels d and e) and supported by the spatial distribution map (Figure 3.3). Being located near these stations, 440-1 is likely more directly affected by short-term stream responses and climatic changes, despite being the deepest well in Watershed 2. Its SHAP scores (temperature: 0.02355, precipitation: 0.10759, hydrometric level: 0.09208) confirm its strong sensitivity to all three dynamic variables. In contrast, wells 439-1 and 442-1, although located within the same watershed, are farther from the hydrometeorological stations and situated at different elevations. According to the map (Figure 3.3), both are somewhat equally distant from the ERA5 grid node and 02ED007 hydrometric stations, but 439-1 lies at a higher elevation (Z diff. to hydrometric station: 149.69 m) compared to 442-1 (75.6 m). This is mirrored in the SHAP values: 439-1 shows higher sensitivity to temperature (0.0212) and hydrometric levels (0.05329), while 442-1 is more sensitive to precipitation (0.04754). These distinctions suggest that not only horizontal distance but also vertical separation from surface monitoring infrastructure shapes GWL sensitivity patterns. Given the spatial scatter of wells in W2 – unlike the clustered, nested setup in W1 – it was particularly important to account for both horizontal and vertical distances to hydrometeorological stations when interpreting model behavior. The spatial map reinforces that 440-1's location offers more immediate exposure to surface forcing, whereas 439-1 and 442-1 experience more diluted or modulated responses due to their spatial positioning within the watershed.

In Watershed 3 (W3), the nested well pair 244-2 and 245-2 provides additional insight into vertical sensitivity differences and subsurface dynamics. Well 245-2, constructed deeper than 244-2, consistently shows higher SHAP values across all three variables – temperature (0.00871), precipitation (0.00453), and hydrometric level (0.00715) – compared to its shallower counterpart 244-2, which reports SHAP values of 0.00522, 0.00351, and 0.00173, respectively. These results align with the greater seasonal variation previously observed in 245-2 by Azimuth Environmental Consulting, Inc. (2019), which described it as screened in a confined aquifer, in contrast to 244-2 being located in a shallow overburden unit. Interestingly, aquifer classifications vary across sources. According to Cuddy (2013) and the AQM model, well 245-2 is located within the CAF3-LAF aquifer, consistent with its deeper and more confined nature. Meanwhile, well 244-2 is classified as AF1-GLAF in both the AQM model and Cuddy (2013), supporting its identification as part of a shallower, unconfined system. The convergence of SHAP-based sensitivity, depth profiles, and aquifer assignment across independent sources reinforces the conclusion that 245-2 is more dynamically responsive due to both its hydrogeological setting and potential interactions with deeper regional flow systems. It also highlights how vertical position, aquifer type, and possibly unmonitored factors such as groundwater extraction contribute to the observed differences in model behavior.

Overall, the SHAP-based sensitivity patterns align with hydrogeological expectations and emphasize the role of depth, proximity, elevation differences, and aquifer type in shaping model behavior. They also highlight the importance of using detailed local data when interpreting GWL fluctuations, especially in nested or topographically complex environments.

5.6.2.2 One-At-a-Time (OAT) Results

While SHAP values provided robust, localized insights into feature importance across time steps, we also sought to understand how changes in input means affect the average predicted GWL. To achieve this, OAT sensitivity analysis was applied, which isolates each input feature to observe its direct impact on

the mean output when perturbed. A thorough explanation of this method was also provided in the Methodology Section 4.7.2. In this method, all features were fixed at their average values, except for one, which was varied $\pm 30\%$ around its mean. This allowed us to quantify how shifts in a single input influence the mean GWL, complementing the time-resolved and feature-ranking perspective offered by SHAP.

5.6.2.2.1 Trend of the Autoregressive Features

Autoregressive features represent the influence of prior GWLs on future values (in this case, a past monthly GWL as the autoregressive feature), capturing the persistence and memory of the system. Analyzing their trends helps to quantify the extent to which past groundwater conditions shape upcoming levels, independent of external hydrometeorological drivers. This section examines how variations in these features around their mean influence the predicted GWL and the direction of that impact, using the OAT framework.

Beginning with the single DMS models, Figure 5.46 shows that all autoregressive features have a clear linear relationship with GWL at time $(t + 1)$. Each subplot corresponds to a specific well, and the high R^2 values (around 0.99) confirm a strong linear fit. Since each single model uses just one autoregressive feature in this scenario, only one regression line is shown per subplot. By comparison, the combined and aggregated models display multiple lines, representing autoregressive inputs from several wells.

A linear fit equation was added to each subplot to quantify the relationship between changes in the autoregressive input and the resulting changes in mean predicted GWL. The slope of each regression line represents the sensitivity of GWL to a $\pm 30\%$ change in its own lagged value. High slope values suggest strong persistence in the groundwater system – i.e., a strong influence of past GWLs on future levels – while lower slopes indicate greater susceptibility to external drivers.

For instance, wells 443-1, 245-2, 244-2, 439-1, and 293-2 exhibit steep slopes (0.93, 0.90, 0.89, 0.88, and 0.85, respectively), reinforcing the idea that their future GWLs are largely governed by prior

values. These wells demonstrate a clear autoregressive dominance, meaning they are likely located in hydrogeologically stable areas or formations where external fluctuations have a smaller impact.

For instance, wells 443-1, 245-2, 244-2, 439-1, and 293-2 exhibit steep slopes (0.93, 0.90, 0.89, 0.88, and 0.85, respectively), reinforcing the idea that their future GWLs are largely governed by prior values. These wells demonstrate a clear autoregressive dominance, meaning they are likely located in hydrogeologically stable areas where external fluctuations have a smaller impact. A 30% increase in the autoregressive input led to mean GWL rises of +22.6% (443-1 and 245-2), +27.1% (244-2), +25.6% (439-1), and +22.7% (293-2), whereas a 30% decrease led to drops of -32.1%, -32.1%, -26.2%, -27.4%, and -28.2%, respectively. The asymmetry in the response also reflects the compound effects of slope and intercept.

In contrast, wells 293-3, 440-1, and 442-1 show lower slopes (0.79, 0.63, and 0.59), indicating weaker autoregressive control. These wells may be more dynamically influenced by hydrometeorological conditions such as recharge or stream-aquifer interactions. For these wells, a +30% change in the autoregressive input led to +19.8% (293-3), +13.4% (440-1), and +16.7% (442-1) in predicted GWL, while -30% changes resulted in -28.0%, -23.3%, and -18.3% drops, respectively. These comparatively smaller positive shifts and weaker negative responses point to reduced sensitivity to past GWL values.

Furthermore, the R^2 values across all subplots are close to 0.99 (not shown on the plots), confirming excellent linear fits and validating the use of a linear model to approximate these relationships. This consistency indicates that, despite differences in slope magnitudes, the relationship between autoregressive changes and predicted GWL remains linear across all wells.

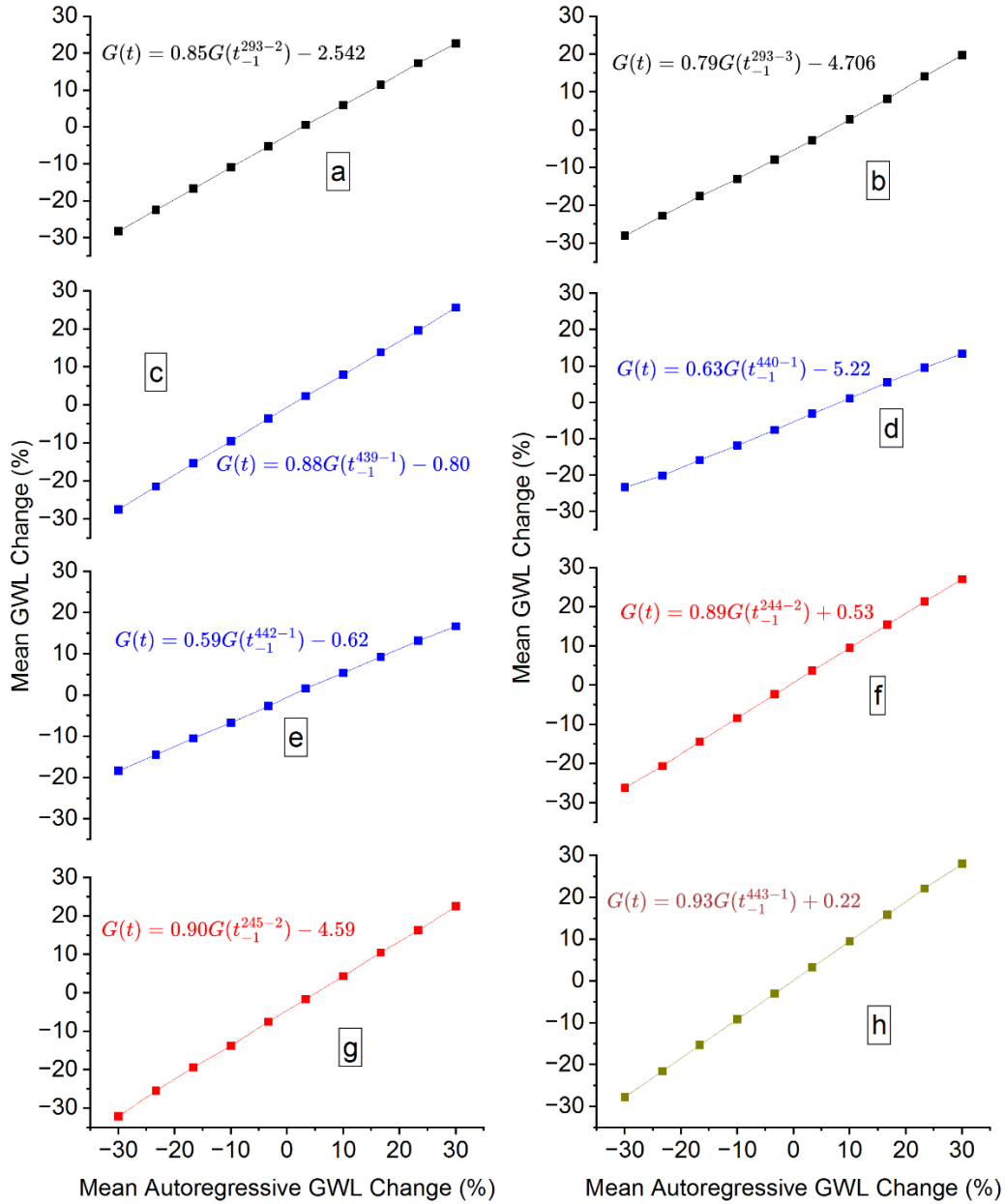


Figure 5-46- Mean GWL change (%) in response to $\pm 30\%$ changes in the autoregressive GWL feature for each well, while all other inputs are held constant. Each subplot corresponds to a different single DMS model: (a) 293-2, (b) 293-3, (c) 439-1, (d) 440-1, (e) 442-1, (f) 244-2, (g) 245-2, and (h) 443-1. A linear regression is fitted to each response data, with slope values indicating the strength of autoregressive influence. The results represent the median GWL change derived from 4800 independently trained neural network simulations.

The analysis of autoregressive features using the OAT method was extended to the combined models, where each model includes one autoregressive feature per well within the respective watershed. As

shown in Figure 5.48, subplots (a) and (b) correspond to Watershed 1 (W1) which includes two wells (293-2 and 293-3); hence, each panel contains two lines representing the individual impacts of the autoregressive features from both wells on the predicted GWL ($t + 1$). Watershed 2 (W2) contains three wells (439-1, 440-1, and 442-1), leading to three lines in panels (c), (d), and (e). Similarly, Watershed 3 (W3), which includes wells 244-2 and 245-2, is represented in panels (f) and (g), each with two lines. Since well 443-1 does not belong to a combined model, it is not included in any of these panels.

Across all panels, the results align with those from the individual OAT analysis: the relationships between the autoregressive features and the predicted GWL remain linear, as indicated by the consistent straight-line patterns. This confirms that altering the mean of each autoregressive input by $\pm 30\%$ while keeping all hydrometeorological variables fixed at their average values leads to a proportional response in the predicted GWL. The slopes in Figure 5.48 reflect how sensitive each well's predicted GWL is to changes in its own and neighboring wells' historical GWL values.

In subplot (a), a 30% increase in the mean autoregressive feature of well 293-2 results in an 11.00% rise in its predicted GWL (slope = 0.60), while a 30% decrease leads to a 25.28% drop, showing a strong and roughly linear effect. In contrast, changes in well 293-3 cause smaller shifts – 1.83% increase and -17.14% decrease (slope = 0.32) – indicating a weaker influence. This suggests that a well's own past values have a greater effect on its GWL prediction than other wells in the same watershed, as seen by the steeper slope for 293-2 compared to 293-3.

Similarly, in panel (b), where the effects of the autoregressive features of wells 293-2 and 293-3 on the GWL of well 293-3 are shown, the slopes are quite close -0.36 for 293-3 and 0.39 for 293-2. This small difference may be explained by the high similarity between their time series (Pearson correlation ≈ 0.8), as illustrated in Figure 5.47.

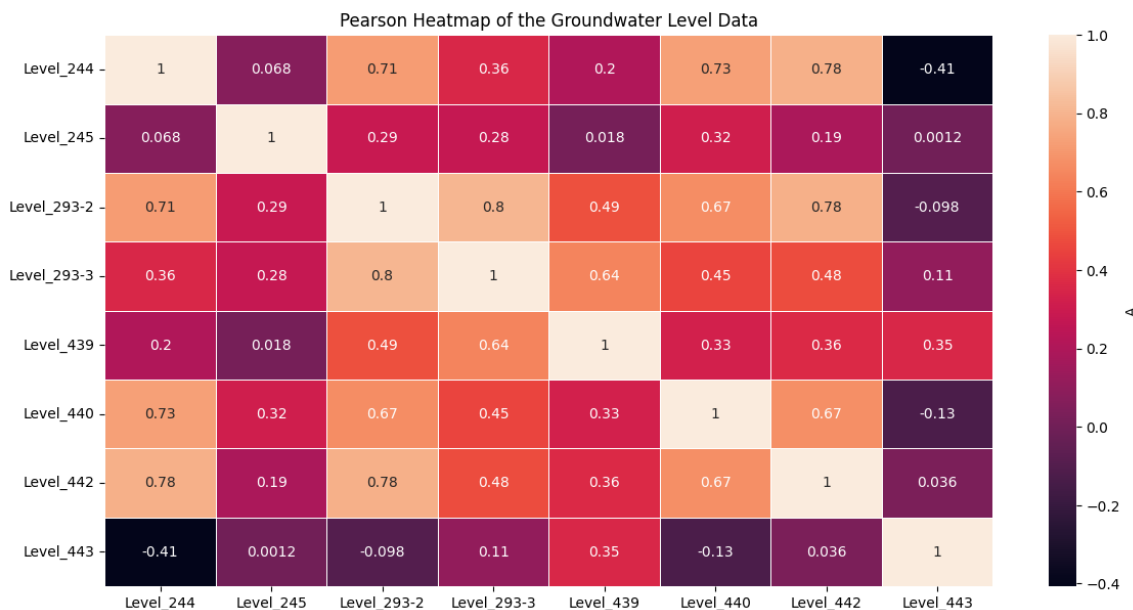


Figure 5-47-Pearson correlation heatmap showing the relationships between GWL time series at different monitoring wells. Higher correlation values (closer to 1) indicate stronger linear associations between wells, while negative or near-zero values suggest weak or inverse relationships.

More importantly, hydrometeorological variables were found to have a stronger influence on the combined model for well 293-3 (to be detailed in Section 5.6.2.2.2), which may reduce the relative contribution of autoregressive features. In this case, the model may struggle to distinguish which input – 293-2 or 293-3 – has more influence on the predicted GWL of 293-3. This confusion does not occur with well 293-2, where the impact of hydrometeorological parameters is lower overall, allowing the model to more clearly differentiate between its autoregressive features.

In panel (c), well 439-1 shows strong self-influence: a +30% increase in its autoregressive feature causes a +13.99% change in predicted GWL, and a -30% change leads to a -21.84% drop (slope = 0.61). Meanwhile, well 440-1 has a weak and slightly negative influence: +30% results in -3.34% and -30% gives -3.78% (slope = 0.01). Well 442-1 shows a mild effect: -7.96% for -30% and +1.17% for +30% (slope = 0.15).

Panel (d) illustrates how the predicted GWL of well 440-1 responds to changes in the autoregressive features of all three wells in Watershed 2. The influence of its own autoregressive input is moderate, with a slope of 0.19: increasing its mean value by 30% results in a surprising -2.30% drop in predicted GWL, while decreasing it by 30% leads to a much larger decline of -13.65%. This unexpected negative response to a positive change suggests some confusion in the model's interpretation of inputs.

Well 439-1 shows a weak inverse relationship with a slope of -0.05; its +30% change reduces 440-1's predicted GWL by -8.66%, and a -30% change causes a -6.07% drop. Similarly, well 442-1 exerts a mild positive influence, with a slope of 0.14, but still produces a negative response: -12.71% for a -30% change and -4.28% even when increased by +30%. These results indicate a non-symmetric and possibly nonlinear model behavior for this well.

Similarly, panel (e) presents the effect of autoregressive features from all three wells in Watershed 2 on the predicted GWL of well 442-1. The strongest influence comes from 442-1's own autoregressive input, with a slope of 0.25: increasing its mean by 30% leads to a +3.95% rise in predicted GWL, while a 30% decrease results in an -11.10% drop. This indicates a moderate and somewhat asymmetric response, with a stronger negative reaction when the input is reduced.

Well 439-1 exerts a weak inverse effect, with a slope of 0.05. A +30% increase in its autoregressive feature decreases the predicted GWL at 442-1 by -2.30%, and a -30% change leads to a -5.58% drop. Meanwhile, the influence of well 440-1 is slightly stronger, with a slope of 0.11: a -30% reduction in its input results in a -7.39% decrease in GWL at 442-1, while a +30% increase still causes a slight negative impact of -0.80%.

In subplot (f), the pattern becomes more pronounced: well 244-2 shows the strongest influence on its own predicted GWL, with a steep slope of 0.81. A 30% decrease in its autoregressive feature leads to a -24.03% drop in GWL, while a 30% increase results in a +24.33% rise. In contrast, well 245-2 has a much

smaller effect, with a slope of 0.15, producing -4.60% and +4.44% changes for -30% and +30% input shifts, respectively.

Finally, subplot (g) highlights the dominant influence of well 245-2 on its own GWL prediction. A 30% decrease in its autoregressive feature leads to a substantial -26.91% drop, while a 30% increase results in a +16.10% rise, corresponding to a steep slope of 0.72. In contrast, the effect of well 244-2 is minimal, with a slope of just 0.05: a 30% decrease in its input causes a -7.67% change, and even a 30% increase unexpectedly leads to a further decline of -4.71%, indicating a weak and possibly confusing influence on the prediction.

Compared to Watersheds 1 and 3, all slope values in Watershed 2 are relatively small, suggesting that the autoregressive features play a less dominant role here. This is likely due to the higher importance of hydrometeorological variables in this watershed, which may cause the model to rely less on past GWL values and struggle to clearly differentiate between the wells' autoregressive effects. In the case of well 440-1, this confusion is especially pronounced, as its own past values unexpectedly reduce its future GWL prediction when increased, highlighting the complex interactions within the combined model.

Overall, these results confirm that, even in the presence of multiple inputs from nearby wells, the dominant predictor for GWL at time $t+1$ remains the well's own past GWL values. The clear and linear relationship captured in the OAT analysis highlights the autoregressive memory of the system and reinforces the importance of including temporal self-dependence in GWL modelling.

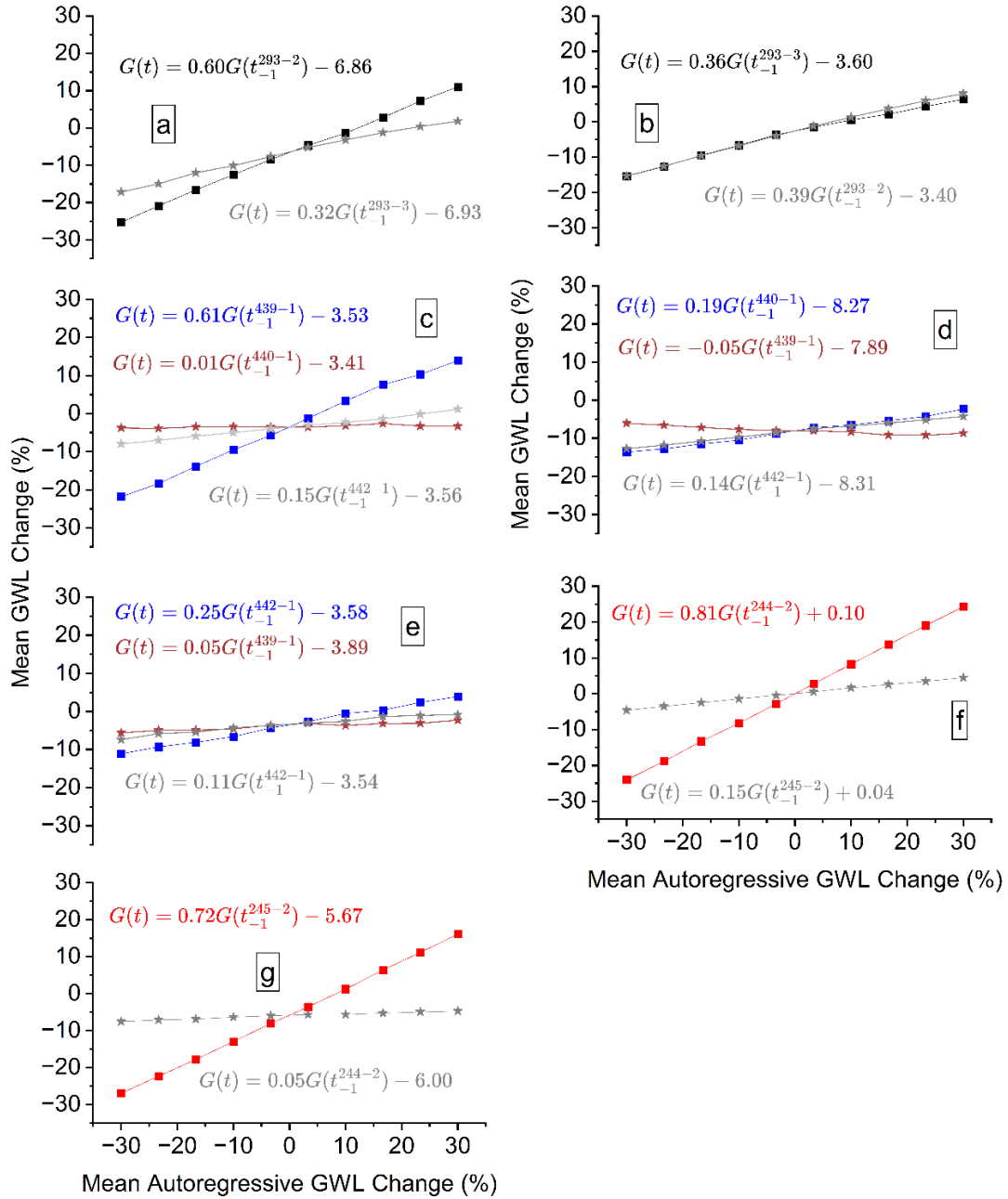


Figure 5-48- Mean GWL change (%) in response to $\pm 30\%$ changes in the autoregressive GWL feature for each well, while all other inputs are held constant. Each subplot corresponds to a different combined DMS model: (a) 293-2, (b) 293-3, (c) 439-1, (d) 440-1, (e) 442-1, (f) 244-2, (g) 245-2, and (h) 443-1. A linear regression is fitted to each response data, with slope values indicating the strength of autoregressive influence. The results represent the median GWL change derived from 4800 independently trained neural network simulations.

5.6.2.2.2 Trend of Hydrometeorological Features

To assess the influence of individual hydrometeorological variables on GWL predictions, OAT sensitivity analysis was conducted on such parameters without changing the autoregressive feature anymore. In this method, the mean of each feature – temperature, precipitation, or hydrometric level – was independently varied between $\pm 30\%$ while keeping the other parameters fixed at their mean values. The resulting GWL changes were obtained from the median response of an ensemble of 4800 neural network simulations. Figure 5.49 displays the results for each of the eight single DMS models, labeled (a) through (h), corresponding to well IDs 293-2, 293-3, 439-1, 440-1, 442-1, 244-2, 245-2, and 443-1, respectively.

In panel (a), model 293-2 showed a negative relationship between hydrometric level and GWL, ranging from about -1.2% to -4.3%, while temperature had a milder effect from -3.9% to -1.4%. Precipitation had minimal impact. In panel (b), for model 293-3, both temperature and hydrometric level exhibited negative effects on GWL, reaching values as low as -6.9% and -4.1%, respectively, while precipitation remained nearly constant. In panel (c), model 439-1 was highly sensitive to hydrometric level, with GWL dropping from +5.4% to -7.4%, and showed a moderate positive response to temperature; precipitation influence was minor. Panel (d) for model 440-1 confirmed strong sensitivity to hydrometric level, with GWL changes from -8.7% to -2.3%, while temperature and precipitation effects remained relatively stable, around -5.5%. Panel (e), representing model 442-1, revealed a wide response range for hydrometric level from -9.4% to +7.6%, while temperature showed a small positive influence and precipitation had a nonlinear effect, switching from positive to negative.

Model 244-2, shown in panel (f), responded positively to all three features. Hydrometric level changes led to GWL increases from -1.9% to +3.1%, while temperature and precipitation also caused moderate increases. Panel (g), for model 245-2, showed a strong negative response to both temperature and precipitation, with GWL decreasing to -6.6% and -9.4%, respectively, while hydrometric level influence was positive but less pronounced. Finally, in panel (h), model 443-1 mirrored the results of 245-2, with

sharp GWL declines in response to increased temperature and precipitation, and a gradual negative trend for hydrometric level as well.

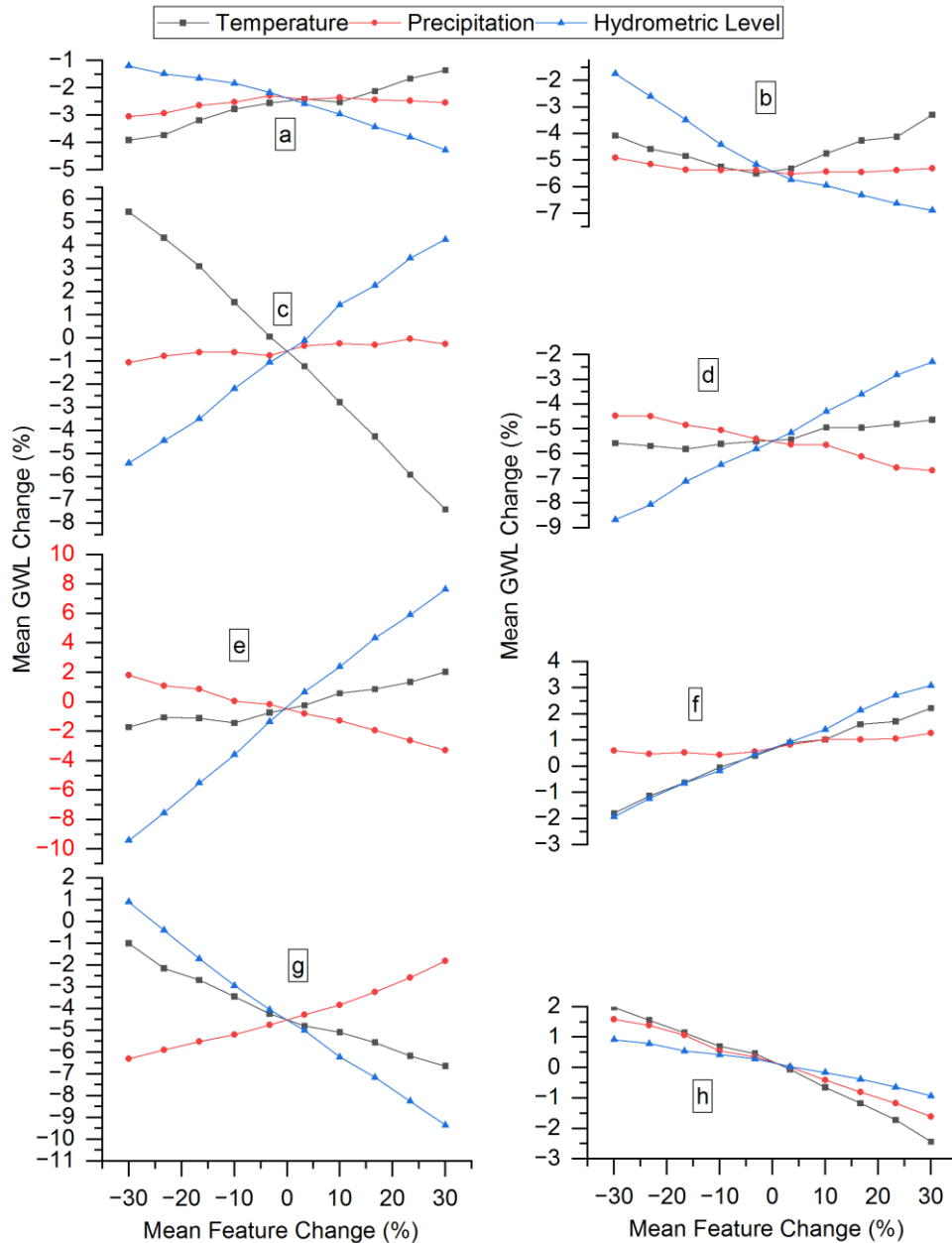


Figure 5-49- Mean GWL change (%) as a function of $\pm 30\%$ changes in the mean values of temperature, precipitation, and hydrometric level feature. For each scenario, two features remain constant at their mean values while the third feature is varied for the purpose of sensitivity analysis. The results represent the median response of an ensemble of 200 neural network (NN) simulations. Results are shown for different single model scenarios: (a) 293-2, (b) 293-3, (c) 439-1, (d) 440-1, (e) 442-1, (f) 244-2, (g) 245-2, and (h) 443-1.

These results indicate that hydrometric level is generally the most influential hydrometeorological variable across most scenarios, especially in models like 439-1, 440-1, and 442-1 (panels c, d, and e), where strong stream-aquifer interactions are likely present. Temperature generally had a negative effect, likely due to increased evapotranspiration, while precipitation often showed limited or delayed impacts. Some wells, such as 244-2, responded more evenly and positively to all three variables, suggesting better-connected or shallower groundwater systems.

Notably, the influence of hydrometeorological variables in Watershed 2 is not only stronger but also more nonlinear and asymmetric compared to other watersheds. This is clearly seen in panels (c), (d), and (e), where climate drivers cause steep and irregular changes in GWL predictions. In contrast, the effects of autoregressive features across all models were more linear, symmetric, and stable. The high influence and nonlinearity of climate variables in Watershed 2 may partly explain why the impact of autoregressive inputs appears weaker in that region – climate factors dominate the model's attention, making it harder to isolate the contribution of past GWL values. These findings underscore the complex, site-specific nature of groundwater–climate interactions and the importance of accounting for nonlinear dynamics when modelling climate-driven groundwater variability.

To further explore these patterns, OAT sensitivity analysis was also applied to the combined models. Here, each hydrometeorological variable – temperature, precipitation, or hydrometric level – was independently varied by $\pm 30\%$, while all other features, including autoregressive inputs, were held constant at their mean values. The predicted GWL responses were drawn from the median of 4800 (50 seeds, 24 neurons, and 4 folds) neural network simulations. The results are presented across three figures: Figure 5.50 includes panels (a) and (b) for wells 293-2 and 293-3; Figure 5.51 includes panels (c), (d), and (e) for wells 439-1, 440-1, and 442-1; and Figure 5.52 includes panels (f) and (g) for wells 244-2 and 245-2. The analysis for well 443-1 is not shown, as it was not included in a combined model.

Panel (a) shows the sensitivity of predicted GWL in the combined model for well 293-2 to changes in hydrometeorological variables. For temperature, two delayed inputs are present: $t-11$ and $t-2$,

corresponding to the two wells in the watershed. The response is more pronounced for $t-11$, where a -30% change leads to a -4.56% drop in GWL, and a +30% change results in a larger decline of -7.15%. For $t-3$, the response is slightly more symmetric, with -5.81% and -6.07% changes for -30% and +30%, respectively. These trends show that temperature has a consistently negative impact, with moderate sensitivity across both delays.

For precipitation, the delayed inputs $t-5$ and $t-12$ show relatively small and flat responses. A -30% change in $t-5$ results in -6.36% GWL change, while a +30% change leads to -6.83%. For $t-12$, the model predicts -7.17% with -30% input and -5.87% with +30%, again indicating a negative but nonlinear effect, especially for $t-12$.

The hydrometric level delays, $t-3$ and $t-10$, show a more stable and symmetric response. For $t-3$, GWL changes from -5.66% to -6.86% as the feature varies from -30% to +30%. For $t-10$, the response remains nearly flat, ranging from -6.14% to -6.20%. These results confirm that hydrometric level, though consistently negative, has a more linear and balanced influence compared to the more variable responses from temperature and precipitation.

Panel (b) shows the sensitivity of predicted GWL in the combined model for well 293-3 to changes in hydrometeorological variables. For temperature, two delayed inputs are used: $t-11$ and $t-2$, corresponding to the two wells in the watershed. The response is stronger for $t-2$, where a -30% change leads to a -7.10% decrease in GWL, while a +30% change results in a +0.91% increase – indicating a wider, asymmetric range. For $t-11$, the model predicts a milder drop from -1.16% to -5.43% across the same range, showing a consistently negative but less steep influence.

For precipitation, $t-5$ and $t-12$ again show relatively moderate effects. A -30% change in $t-5$ results in a -3.44% drop in GWL, while a +30% change leads to -2.62%. For $t-12$, the impact ranges from -4.31% to -1.59%, reflecting a stronger nonlinear pattern with a flatter upper end.

The hydrometric level inputs, $t-3$ and $t-10$, show smaller and more balanced effects compared to the temperature inputs. For $t-3$, GWL changes from -1.41% to -4.69% as the feature shifts from -30% to +30%, suggesting an inverted response. $t-10$ produces a narrower range, from -3.54% to -2.84%, indicating a stable but weaker influence. These results suggest that in well 293-3, temperature and precipitation have more varied and asymmetric effects, while hydrometric level contributes a smaller and more consistent impact.

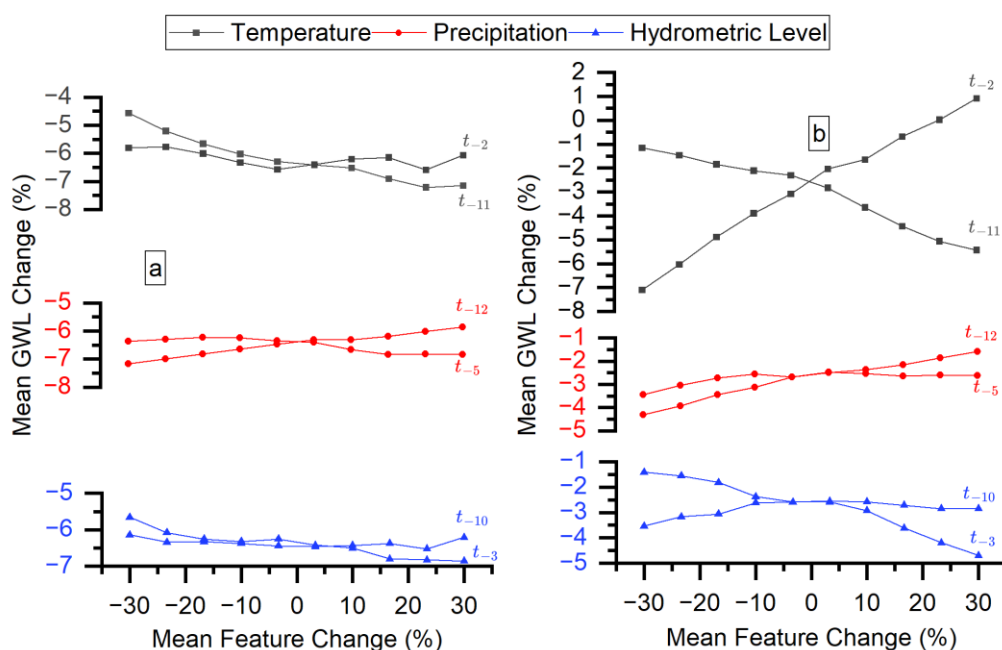


Figure 5-50-OAT sensitivity analysis of hydrometeorological variables for the combined models of wells 293-2 (panel a) and 293-3 (panel b). Each line represents the mean GWL (%) change in response to $\pm 30\%$ variation in a single climate input—temperature (black), precipitation (red), or hydrometric level (blue)—while all other inputs are held at their average values. Labels indicate the delay time step (e.g., $t-2$) used for each feature, corresponding to the most influential lag per well identified via cross-correlation analysis.

For the combined model of Watershed 2, panel (a) illustrates the sensitivity of predicted GWL in well 439-1 to changes in hydrometeorological inputs. For temperature, three delayed features are included: $t-7$, $t-4$, and $t-10$. The strongest impact is seen for $t-7$, where a -30% change leads to a -2.76% drop in GWL, and a +30% increase results in a -4.43% change – suggesting a counterintuitive effect with increased temperature worsening conditions. $t-4$ produces a relatively symmetric response, from -4.36% to -3.09%,

while $t-10$ yields a sharper decline from -4.87% to -2.27%, showing a wider but consistently negative influence.

For precipitation, delays $t-1$ and $t-12$ show negative and moderately nonlinear trends. For $t-1$, GWL decreases from -5.38% to -2.06% across the $\pm 30\%$ range, while $t-12$ shifts from -2.88% to -3.74%, showing a smaller but still negative effect overall.

The hydrometric level inputs, $t-3$ and $t-7$, reveal stronger and more distinct responses. A -30% change in $t-3$ causes a -4.56% decrease in GWL, while a +30% change leads to -2.62%, maintaining a mostly linear relationship. In contrast, $t-7$ shows a broader range from -7.01% to -0.39%, indicating a steeper and more nonlinear impact. These findings reinforce that hydrometric level is the dominant and most sensitive climate driver in this model, with more complex patterns than the mostly symmetric autoregressive responses observed earlier.

Panel (b) shows the sensitivity of predicted GWL for well 440-1 to changes in hydrometeorological variables. Among the temperature delays ($t-7$, $t-4$, and $t-10$), the responses are strongly negative and asymmetric. For example, $t-7$ ranges from -12.60% to -4.35%, $t-4$ from -11.23% to -5.86%, and $t-10$ from -6.33% to -9.70%.

Precipitation inputs ($t-1$ and $t-12$) also show substantial influence, with GWL changes between -11.20% and -4.63% for $t-1$, and -8.40% to -7.62% for $t-12$, again showing nonlinear patterns. The hydrometric level delays ($t-3$ and $t-7$) contribute slightly less, with changes from -9.56% to -7.14% and -8.44% to -7.16%, respectively. Compared to well 439-1, temperature and precipitation exert greater influence in this model, likely reducing the predictive contribution of autoregressive features.

Panel (c) finally illustrates the sensitivity of predicted GWL for well 442-1 to changes in hydrometeorological inputs. The model includes three temperature delays: $t-7$, $t-4$, and $t-10$. The responses are asymmetric and varied. For $t-7$, GWL drops by -7.94% under a -30% change, while a +30% increase leads to a slight rise of +0.71%. $t-4$ shows a more consistent negative effect, ranging from -4.69% to -

3.46%. In contrast, $t-10$ reverses this pattern, with a steeper decline for a positive shift (-7.97%) and a minimal drop for a negative one (-0.24%). Precipitation delays $t-1$ and $t-12$ have more moderate effects. A -30% change in $t-1$ leads to a -6.22% drop, while a +30% shift gives -1.58%. For $t-12$, the GWL ranges from -3.73% to -3.39%, suggesting a smaller, more symmetric impact.

The hydrometric level delays $t-3$ and $t-7$ again contribute meaningful effects, though less strongly than temperature. For $t-3$, GWL changes from -6.22% to -1.40%, while for $t-7$, the response is steeper on the negative side, from -6.64% to -0.34%. These results point to more complex, nonlinear interactions in the model for well 442-1, especially with temperature inputs showing high variability, which may limit the clarity of autoregressive feature effects in this case.

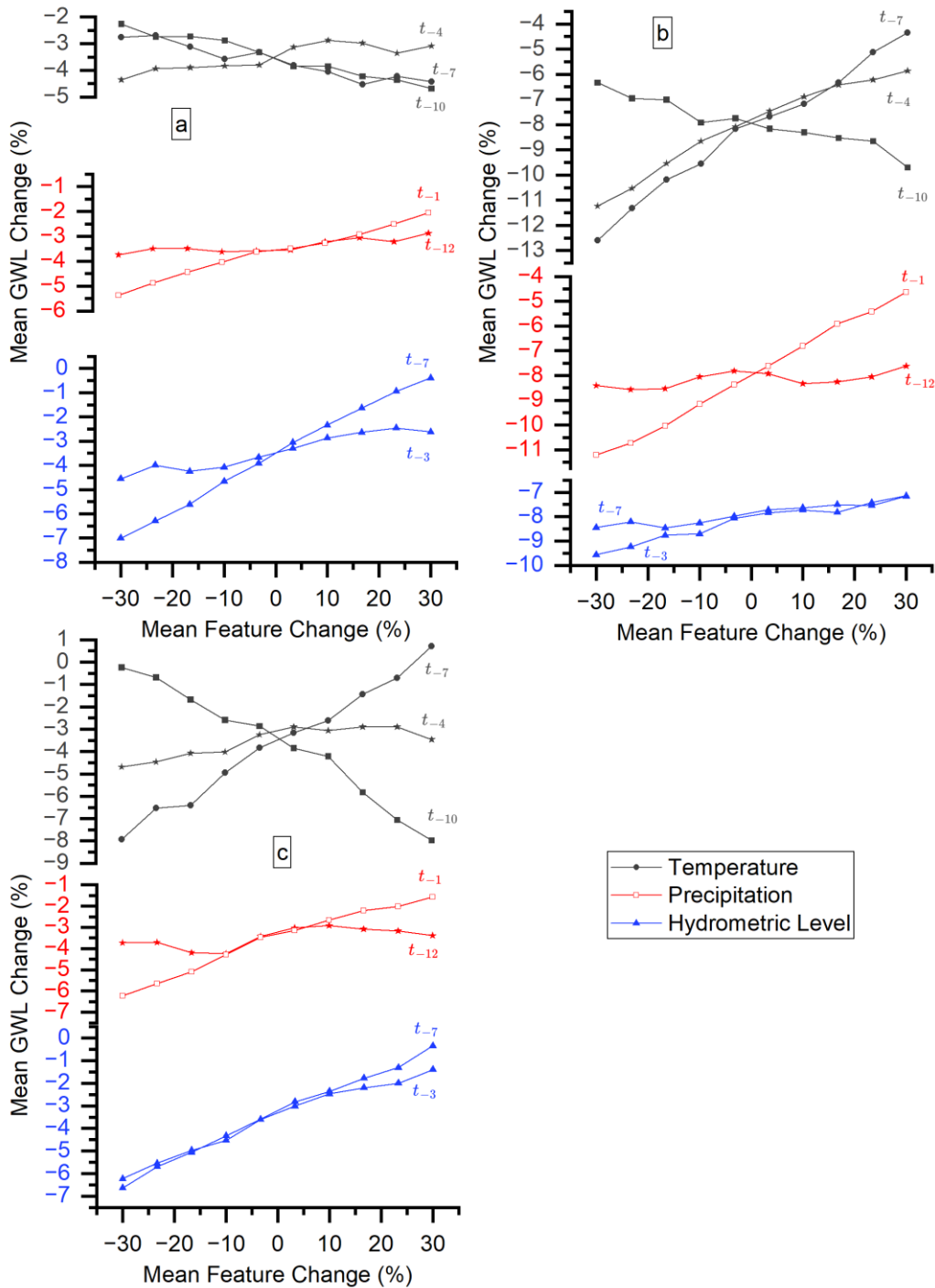


Figure 5-51-OAT sensitivity analysis of hydrometeorological variables for the combined models of wells 439-1, 440-1, and 442-1, shown in panels (a), (b), and (c) respectively. Each line represents the mean GWL (%) change in response to $\pm 30\%$ variation in a single climate input—temperature (black), precipitation (red), or hydrometric level (blue)—while all other inputs are held constant at their mean values. Time step labels (e.g., t_{-7}) indicate the most influential lag per variable, selected based on cross-correlation between each well's GWL and the corresponding input.

The analysis then proceeds to Watershed 3, where panel (a) represents the combined model for well 244-2. The sensitivity of predicted GWL to hydrometeorological variables shows more stable and symmetric behavior compared to the wells in Watershed 2 and Watershed 1. For temperature, the model includes delays $t-5$ and $t-1$, both showing nonlinear and asymmetric effects. For $t-5$, a +30% change results in a -2.39% drop in GWL, while a -30% shift leads to a +2.20% increase. Similarly, $t-1$ ranges from -3.83% to +3.68%, indicating strong but opposing effects. Precipitation delays $t-1$ and $t-6$ have weaker but still asymmetric influences: GWL changes from -0.48% to +0.44% for $t-1$, and from -0.55% to +0.49% for $t-6$.

The hydrometric level delays $t-3$ and $t-5$ contribute minimally, with changes of -0.58% to +0.59% and -0.07% to -0.18%, respectively. Overall, the model for well 244-2 responds most clearly to temperature, while precipitation and hydrometric level have smaller, more balanced effects. This pattern, along with the linearity of the responses, may reflect a shallower or more stable groundwater system with weaker surface interactions than observed in Watershed 2.

Panel (b) continues the analysis in Watershed 3, focusing on the combined model for well 245-2. The temperature inputs ($t-5$ and $t-1$) have the strongest influence on predicted GWL. Both delays show steep and symmetric responses: GWL drops from around -9.9% to -6.0% as temperature increases by 30%, and rises similarly with a -30% reduction. These patterns indicate a consistent and linear sensitivity to temperature variations. Precipitation delays $t-1$ and $t-6$ also show clear negative impacts, though less pronounced than temperature. GWL changes from -6.22% to +0.44% for $t-1$, and from -3.73% to +0.49% for $t-6$, suggesting mild nonlinearity and slightly stronger effects at lower values.

For the hydrometric level, delays $t-3$ and $t-5$ display more distinct behavior. $t-3$ shows a steep nonlinear response: a -30% change leads to about -6.6%, while a +30% change results in a positive +0.59%. $t-5$, by contrast, varies from around -0.07% to -0.18%, indicating a much smaller role.

Overall, the combined model for well 245-2 shows stronger climate sensitivity than 244-2, especially to temperature and, to a lesser extent, hydrometric level. The nonlinear responses – particularly in hydrometric inputs – suggest that local hydrogeological conditions may amplify or dampen specific climate signals in this part of the watershed.

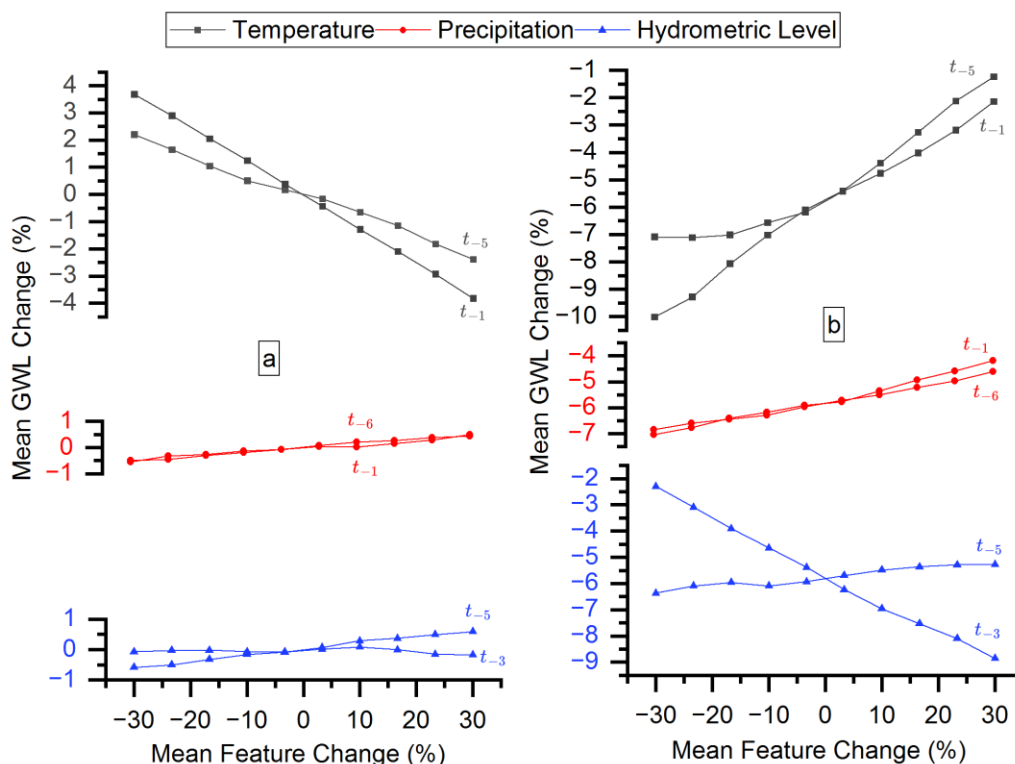


Figure 5-52- OAT sensitivity analysis of hydrometeorological variables for the combined models of wells 244-2 and 245-2, shown in panels (a), and (b) respectively. Each line represents the mean GWL (%) change in response to $\pm 30\%$ variation in a single climate input—temperature (black), precipitation (red), or hydrometric level (blue)—while all other inputs are held constant at their mean values. Time step labels (e.g., $t-5$) indicate the most influential lag per variable, selected based on cross-correlation between each well's GWL and the corresponding input.

Overall, hydrometeorological variables had a weaker influence on mean GWL compared to the autoregressive feature, indicating that GWLs are more strongly driven by their own past values. Unlike the consistently positive effect seen with autoregressive inputs, changes in temperature, precipitation, and hydrometric level did not always lead to a predictable rise or fall in GWL. In several cases, these parameters had an inverse effect, where increasing the input led to a decrease in GWL or vice versa. Among them,

hydrometric level had the strongest influence, though its overall effect remained modest relative to the dominant autoregressive control. These findings highlight the limited predictive power of external variables in comparison to the internal memory of the groundwater system. Since the Aggregated model performed poorly, even under the DMS modelling framework, it was excluded from SHAP and OAT analyses to maintain focus on more reliable and informative results.

Chapter 6 Discussion

6.1 Interpretation of Regional GWL Trends and Non-Stationarity

The first objective of this thesis was to evaluate long-term GWL trends across Southern Ontario using a systematic statistical framework. The Mann-Kendall results demonstrate that the majority of wells exhibit statistically significant trends over the study period. This confirms that GWL dynamics in the region are largely non-stationary (showing a trend upwards or downwards) rather than fluctuating around a stable long-term mean.

This finding strengthens and extends earlier descriptive and regional assessments conducted in Ontario. For example, Bhatti et al. (2021) reported observable increases or localized declines in selected PGMN wells but did not perform a systematic site-level statistical trend analysis across the entire province. Similarly, the Toronto and Region Conservation Authority (TRCA, 2024) provided monitoring-based summaries indicating generally stable conditions within its jurisdiction, yet these assessments relied largely on indicator-based reporting rather than formal statistical hypothesis testing.

In the Oak Ridges Moraine Groundwater Program (ORMGP) assessment, the Mann-Kendall test was applied to daily groundwater records; however, the analysis was limited to approximately a 10-year observation window. Shorter temporal windows reduce sensitivity to long-term monotonic changes and may underestimate gradual but persistent trends. As a result, the proportion and magnitude of significant trends reported in ORMGP studies appear smaller compared to the findings of this thesis.

Similarly, Markle and Harris (2019) conducted a province-wide Mann-Kendall trend analysis of Ontario wells but relied primarily on annually aggregated data. Annual aggregation substantially reduces the number of observations and may decrease statistical power relative to analyses conducted on higher temporal resolution datasets. In contrast, this thesis applies consistent site-level testing across a longer period (2000–2023) and preserves temporal variability before aggregation for modelling purposes. This approach enhances sensitivity to subtle long-term trends and provides a more comprehensive regional-scale statistical evaluation.

Therefore, while earlier Ontario-based studies identified localized or descriptive patterns, this thesis extends prior work by implementing a unified, province-wide, site-level statistical framework. This directly addresses the research gap identified in Chapter 2 regarding the lack of comprehensive regional statistical evaluation of groundwater behaviour in Southern Ontario.

6.2 Physical Interpretation: Aquifer Type and Hydrogeological Context

The slope distribution analysis comparing overburden and bedrock aquifers reveals clear and systematic differences in groundwater responsiveness and variability. Overburden aquifers generally exhibit stronger seasonal signals and larger slope magnitudes, indicating greater sensitivity to climatic drivers such as precipitation and temperature. In contrast, bedrock aquifers show more dampened responses, with smaller slope magnitudes and reduced short-term variability. These differences suggest that aquifer type significantly influences both the magnitude and temporal behaviour of groundwater trends across Southern Ontario. This pattern is consistent with the known hydrogeological behaviour of these aquifer types, where overburden aquifers typically respond more strongly to climatic drivers than bedrock aquifers.

Importantly, aquifer type in this thesis was determined using three-dimensional hydrostratigraphic data from the Ontario Geological Survey, supported by regional hydrogeological reports (Burt & Dodge, 2011; Cuddy, 2013; LSRCA, 2013). By linking each well to its most likely subsurface unit, the AQM framework provided geological context for interpreting groundwater trends. This approach moves beyond

surface-based or purely statistical grouping methods and ensures that trend interpretation is grounded in subsurface structure.

The literature review emphasizes that groundwater systems are inherently heterogeneous and that spatial configuration plays a critical role in controlling groundwater dynamics. It also notes that deeper and more complex aquifer systems remain relatively underexplored in regional assessments. Without accounting for aquifer type, regional-scale analyses could mask important differences in responsiveness. For example, stronger seasonal variability in overburden aquifers likely reflects rapid recharge transmission, whereas gradual long-term behaviour in bedrock systems may indicate slower hydraulic adjustment within deeper flow systems.

Overall, the observed contrast between overburden and bedrock aquifers confirms that groundwater trend behaviour in Southern Ontario is controlled by hydrostratigraphic conditions. By embedding statistical findings within established hydrogeological theory and validated geological classification, this thesis ensures that regional groundwater analysis remains physically grounded rather than treated as a purely numerical exercise.

6.3 Performance of Linear Regression vs. Neural Networks

A core objective of this study was to evaluate whether traditional regression approaches can match the performance of neural networks for monthly GWL prediction. Rather than assuming the superiority of machine learning methods, this thesis first implemented linear models as baseline approaches and evaluated their statistical validity through diagnostic testing.

The regression diagnostics clearly reveal violations of key classical assumptions, including heteroscedasticity, autocorrelation of residuals, and nonlinear relationships between predictors and GWLs. These violations are not unexpected in hydrological systems, where groundwater dynamics are influenced by delayed recharge, threshold behaviour, seasonal variability, and interaction effects among climatic and

hydrometric variables. Linear regression assumes linearity, exogeneity, and homoscedastic errors, assumptions that are often unrealistic for hydrogeological time series. While regression models provide transparency and computational simplicity, their structural rigidity limits their ability to represent complex subsurface–climate interactions.

The limitations of purely linear approaches are consistent with findings in the groundwater modelling literature. Rajaei et al. (2019) reviewed AI applications in groundwater prediction and reported that ANNs frequently outperform traditional statistical models when modelling nonlinear hydrological relationships. Similarly, Ahmadi et al. (2022) and Boo et al. (2024) emphasized the ability of neural networks to capture nonlinear dependencies between climatic inputs and groundwater responses. These studies highlight that groundwater systems rarely behave as simple linear systems, particularly when lagged inputs and multivariate climatic interactions are involved.

Empirical comparisons within the reviewed literature further support this pattern. Sharghi et al. (2022) compared FFNN, ANFIS, GMDH, LSTM, and NAE models and reported superior predictive performance for nonlinear machine learning models compared to simpler approaches. Kardan Moghaddam et al. (2021) similarly found that nonlinear GMDH models outperformed other modelling strategies when predicting GWLs. These findings suggest that groundwater systems benefit from flexible model structures capable of representing nonlinear and multivariate interactions.

Consistent with these studies, neural networks in this thesis demonstrate improved predictive performance across most wells and modelling configurations. Their architecture allows them to learn nonlinear relationships between lagged GWL, precipitation, temperature, and hydrometric inputs without requiring strict distributional assumptions. Unlike regression, neural networks do not impose linear functional forms or independence constraints, enabling them to adapt to complex seasonal and interannual variability.

Importantly, this progression from regression to neural networks was intentionally structured within the thesis design. Many previous groundwater studies directly implement neural networks without explicitly demonstrating the failure of simpler models. In contrast, this thesis first establishes the empirical limitations of regression within the Ontario dataset before introducing more flexible modelling frameworks. The diagnostic evidence shows that regression shortcomings are not theoretical but observable in the data. This step strengthens the methodological transparency of the study and provides a justified rationale for adopting neural network approaches.

Overall, the comparison confirms that while regression models offer interpretability and computational efficiency, their structural assumptions restrict their applicability in complex hydrogeological systems. Neural networks provide improved predictive capability by accommodating nonlinear, lagged, and multivariate relationships that characterize monthly groundwater dynamics. This finding aligns with the broader groundwater modelling literature and supports the decision to employ advanced machine learning techniques within a physically informed framework.

6.4 Single, Combined, and Aggregated Modelling Scenarios

A central methodological contribution of this thesis is the structured comparison between three modelling configurations: (1) single models, (2) combined models, and (3) aggregated models. This directly addresses the objective stated in Chapter 1 of evaluating single-output and multi-output frameworks under different spatial configurations. Unlike many previous studies that restrict comparisons to single versus multi-output settings within one aquifer or watershed, this thesis introduces a physically defined intermediate grouping based on watershed boundaries.

The results of this study largely align with prior findings that single-output models often outperform multi-output or global models. Müller et al. (2021) reported that single-output MLP, CNN, and LSTM models consistently achieved lower RMSE values than their multi-output counterparts within the same

watershed. Similarly, Chidepudi et al. (2025) found that single-well deep learning models (LSTM, GRU, BiLSTM) generally outperformed multi-station models unless additional clustering strategies were applied. Patra et al. (2023) also demonstrated superior performance of single-output LSTM models compared to multi-output configurations in coastal aquifer systems. These findings are consistent with the results of this thesis, where single-well models frequently achieved the highest predictive accuracy by focusing on localized hydrogeological dynamics.

From a theoretical perspective, this behaviour is expected when target variables are not strongly interrelated. Caruana (1997) argued that multi-task or multi-output learning can lead to negative transfer when outputs are weakly correlated, as shared representations introduce interference between unrelated signals. Hsieh (2023) similarly notes that multi-output modelling is beneficial primarily when target variables are interdependent and share common structure. In heterogeneous groundwater systems spanning multiple aquifers and watersheds, such strong correlation cannot be assumed. The dominance of single models in this thesis therefore aligns with both empirical and theoretical expectations.

However, most prior comparisons have been limited to wells located within the same aquifer or watershed. As discussed in Section 2.4, multi-output approaches in groundwater studies are typically applied to nearby wells within a shared hydrogeological setting (Chidepudi et al., 2025). Explicit evaluation of physically defined grouping strategies – such as watershed boundaries – remains limited. Many clustering-based studies rely on statistical similarity rather than hydrogeological reasoning. For example, Sharghi et al. (2022) grouped wells using K-means, FCM, and GNG clustering before applying predictive models, while Kardan Moghaddam et al. (2021) used clustering to select representative wells prior to modelling. Although these approaches improved performance under certain conditions, they were primarily data-driven and not explicitly based on watershed-scale physical boundaries.

The watershed-based combined models introduced in this thesis provide a physically meaningful compromise between purely local and fully aggregated modelling. Wells within the same watershed often share recharge sources, climatic forcing patterns, and surface-groundwater interactions, although local

hydrogeological differences – such as variations in aquifer material, degree of confinement, or well depth – can still produce distinct groundwater responses. Grouping wells according to watershed boundaries introduces hydrogeological logic into the modelling structure rather than relying solely on statistical similarity. The results show that combined models often achieve performance close to single-well models while benefiting from increased training data and shared information. This suggests that physically informed grouping may mitigate some of the negative transfer effects observed in unconstrained multi-output models.

In contrast, aggregated models that combine wells across multiple watersheds demonstrate stronger generalization potential but reduced local precision. While these models benefit from exposure to a broader range of hydrogeological conditions, their performance is influenced by two important structural factors observed in this study. First, the inclusion of wells with inherently weaker predictive behaviour can negatively affect the shared parameter space of the aggregated model. When one or more wells exhibit high noise, complex dynamics, or limited predictive signal, the model attempts to fit these patterns alongside better-behaved wells. This can lead to negative transfer, where the learning of one target degrades the performance of others (Caruana, 1997). As a result, strong wells may lose some predictive accuracy because the model must accommodate weaker signals within a unified structure.

Second, the aggregated configuration reduces the effective number of usable observations. Because not all wells share identical monitoring periods or record lengths, the combined dataset becomes constrained by overlapping temporal windows. Wells with shorter time series reduce the number of training samples available across the shared modelling period. This imbalance limits the model's ability to fully leverage the theoretical benefit of larger data volume. Instead of gaining purely from scale, the aggregated model must operate under uneven and sometimes reduced data coverage.

As spatial heterogeneity increases – due to differences in aquifer type, recharge mechanisms, and climatic forcing – the model must balance diverse groundwater behaviours within a single parameter space. This often results in smoother, regionally averaged predictions that sacrifice well-specific accuracy.

Therefore, the trade-off observed in this thesis reflects not only a bias-variance balance but also the impact of heterogeneity and unequal data representation. Single models preserve localized dynamics, watershed-combined models retain physical coherence while sharing information, and aggregated models prioritize generalization at the expense of local specificity and stability.

6.5 Impact of SSA Preprocessing (DSM vs. DMS)

Another major methodological contribution of this research is the systematic evaluation of SSA timing within the modelling pipeline. While SSA has been widely used for time-series decomposition in hydrological and environmental studies (Golyandina, 2011; Hassani et al., 2011), its placement within predictive modelling workflows has received limited attention in groundwater forecasting applications.

The comparison between DSM and DMS configurations demonstrates that preprocessing timing can influence both model accuracy and stability. Across most watersheds, the DMS configuration produced higher median R^2 values and lower MSE compared to DSM, particularly for single-well and combined models. The improvements observed in Watersheds 1 and 3 suggest that smoothing after aggregation better preserves dominant seasonal and interannual patterns relevant to monthly prediction. These findings indicate that input structuring plays a critical role in determining model generalization performance. However, the benefits of DMS were not uniform across all spatial configurations. In heterogeneous settings and especially within the aggregated model, improvements were limited. This reinforces a key insight from the broader groundwater modelling literature: model performance depends not only on architecture but also on how input signals are structured and aligned (Rajaei et al., 2019; Shen, 2017). Decomposition alone cannot compensate for spatial heterogeneity or temporal misalignment.

Previous studies have incorporated signal decomposition prior to predictive modelling. For example, Jeihouni et al. (2019) used wavelet-based preprocessing in long-term climate projections, and Chidepudi et al. (2025) applied spectral techniques in multi-station modelling frameworks. Sharghi et al. (2022) and Kardan Moghaddam et al. (2021) similarly integrated preprocessing within their modelling

pipelines. However, these studies generally treated decomposition as a fixed preparatory step rather than evaluating its placement within the modelling sequence. A direct comparison of preprocessing timing relative to temporal aggregation has not been systematically addressed.

By comparing DSM and DMS structures under identical modelling conditions, this thesis demonstrates that preprocessing timing is a design parameter rather than a neutral preprocessing choice. The results suggest that applying smoothing after aggregation can enhance predictive stability for wells with coherent seasonal structure, while the overall spatial configuration of the model remains the dominant determinant of performance.

Overall, this analysis extends the groundwater forecasting literature by explicitly treating signal decomposition as an adjustable modelling component. It shows that preprocessing decisions should be informed by temporal scale and spatial coherence, rather than applied uniformly across modelling scenarios.

6.6 Sensitivity Analysis and Model Interpretability

Sensitivity analysis using SHAP and OAT approaches provides insight into the relative influence of input variables across spatial modelling configurations. While neural networks are often criticized for limited transparency, interpretability tools allow the identification of dominant drivers and assessment of whether learned relationships are physically plausible.

Across all spatial scenarios and preprocessing configurations, SHAP results consistently show that lagged GWL variables exert the strongest influence on model predictions. This is partly expected, as the adopted framework is autoregressive and explicitly incorporates previous GWLs as predictors. However, the strong influence of lagged terms is also physically consistent with groundwater system behaviour. Hydraulic head dynamics are governed by aquifer storage properties and delayed flow responses, meaning that GWLs adjust gradually to recharge-discharge processes (Freeze & Cherry, 1979; Fetter, 2018). Similar dominance of previous GWL observations has been reported in ML-based forecasting studies (Müller et

al., 2021; Patra et al., 2023; Chidepudi et al., 2025). Therefore, although structurally anticipated, the sensitivity results confirm that the neural networks capture realistic system memory effects.

Beyond lagged groundwater inputs, SHAP and OAT analyses indicate that precipitation and hydrometric variables contribute meaningfully, particularly in overburden aquifers and watershed-based combined models. Groundwater recharge is primarily driven by precipitation infiltration and surface–groundwater interactions (Giordano, 2009; Jyrkama & Sykes, 2007). Overburden aquifers, being closer to the surface and often unconfined, are more responsive to climatic forcing than deeper bedrock systems (Singer et al., 2003; Lipka, 2010). Although wells within the same watershed do not necessarily have to share the same recharge zone, the observed sensitivity patterns show that recharge-related variables have greater influence in shallower systems and within hydrologically coherent watershed groupings.

Differences in SHAP and OAT patterns across spatial configurations further highlight the effect of model structure. In single models, lagged GWL dominates, reflecting localized storage-controlled behaviour. In combined models, precipitation and hydrometric inputs gain relatively greater influence, suggesting that shared climatic forcing becomes more detectable when wells are grouped within the same hydrological boundary. In contrast, aggregated models that combine wells across multiple watersheds exhibit less stable sensitivity patterns, likely due to spatial heterogeneity and competing aquifer responses within a shared parameter space. This behaviour is consistent with discussions in multi-output learning, where shared models may obscure localized dynamics when target variables are not strongly correlated (Caruana, 1997; Hsieh, 2023).

Overall, integrating SHAP and OAT into the modelling framework strengthens interpretability. The sensitivity results demonstrate that neural network predictions are driven by physically plausible processes – aquifer storage, recharge inputs, and surface-groundwater interactions – rather than arbitrary statistical relationships. By linking model behaviour to established hydrogeological principles, this study addresses concerns regarding the opacity of neural networks and enhances confidence in their application for groundwater forecasting.

6.7 Implications for Groundwater Modelling and Contribution to the Literature

The results of this thesis have important implications for groundwater modelling practice and regional water management. The statistical analysis confirms that GWLs in Southern Ontario are largely non-stationary, emphasizing that trend testing should precede predictive modelling. Many modelling approaches assume stable behaviour, yet groundwater systems are influenced by climatic variability, recharge dynamics, and human pressures (Müller et al., 2021; Patra et al., 2023). Embedding Mann-Kendall testing within the modelling workflow ensures that predictive strategies are aligned with observed system behaviour.

The structured comparison between single, combined, and aggregated models demonstrates that spatial configuration significantly affects predictive performance. While previous studies have compared single-output and multi-output models (Chidepudi et al., 2025; Patra et al., 2023), physically defined watershed grouping remains underexplored. By organizing wells according to hydrological boundaries rather than purely statistical similarity (Sharghi et al., 2022), this study shows that model reliability depends on hydrogeological coherence and temporal data consistency, not simply on increasing data volume.

The evaluation of SSA timing further highlights that preprocessing decisions influence model stability. Although denoising techniques are commonly applied in ML-based hydrological studies (Ahmadi et al., 2022; Boo et al., 2024), their placement within the modelling pipeline is rarely examined. The comparison of DSM and DMS configurations demonstrates that input structuring can materially affect predictive robustness.

From a modelling perspective, the progression from linear regression to neural networks provides empirical justification for adopting flexible ML approaches when regression assumptions are violated (Huang et al., 2021). FFNNs improved predictive performance under nonlinear and temporally persistent conditions, while interpretability analyses confirmed that predictions remained physically plausible. From a management perspective, the presence of widespread significant trends reinforces the need for continued

long-term monitoring. Single models offer the highest local precision, combined models support sub-regional planning, and aggregated models may inform broader policy assessments.

Overall, this thesis contributes to the literature by integrating regional trend analysis, aquifer-type classification (bedrock versus overburden), spatially structured modelling, preprocessing evaluation, and interpretability within a unified framework. By directly linking statistical diagnosis to predictive model design, the study advances a structured approach to groundwater forecasting that is transferable beyond Southern Ontario to other regions facing hydroclimatic variability and data limitations.

Chapter 7 Conclusions and Recommendations

This thesis evaluated long-term GWL dynamics across Southern Ontario and examined how statistical diagnosis, hydrogeological context, and modelling design influence predictive performance. The study addressed several main objectives: (1) assessing regional non-stationarity using a systematic statistical framework, and interpreting trends within hydrogeological context, (2) examining the suitability of linear regression models and, upon identifying violations of key assumptions, adopting FFNN as a more flexible alternative, (3) evaluating spatial modelling structures including single, combined, and aggregated configurations, (4) assessing preprocessing strategies through DSM and DMS input arrangements, and (5) systematically optimizing FFNN architecture by identifying the optimal number of hidden neurons under multiple random initializations to ensure stable and reliable performance.

The Mann-Kendall analysis demonstrated that the majority of monitoring wells exhibit statistically significant long-term trends between 2000 and 2023. This indicates that GWLs across Southern Ontario tend to show long-term trends over time rather than simply fluctuating around a constant mean. The widespread presence of monotonic trends highlights the importance of explicitly testing for non-stationarity prior to predictive modelling.

Following the hydrogeological interpretation of groundwater trends, the study proceeded to develop a data-driven framework for predicting GWL. The modelling process began with linear regression as a simple and interpretable benchmark model. This initial step allowed evaluation of fundamental assumptions and provided a baseline for predictive performance. However, diagnostic analysis revealed violations of key regression assumptions, including heteroscedasticity, residual autocorrelation, and nonlinear predictor relationships.

Given these empirical limitations, the modelling framework transitioned to FFNN, which do not require strict distributional or linearity assumptions. The FFNN demonstrated improved predictive performance, particularly under nonlinear and temporally persistent conditions characteristic of

groundwater systems. This progression from regression to flexible machine learning models was therefore methodologically justified by observed data behaviour rather than assumed a priori.

To ensure fair and stable model comparison, neural network architecture was systematically optimized for each modelling scenario. The optimal number of hidden neurons was identified through structured experimentation, and multiple random initializations were applied to reduce sensitivity to weight initialization and local minima. This approach ensured that reported performance metrics reflected robust model behaviour rather than artefacts of a single training run.

Spatial modelling structure proved to be a critical design decision. Single models consistently achieved the highest stability and predictive accuracy. Combined models performed well when wells shared hydrogeological similarity. In contrast, the fully aggregated model performed poorly due to spatial heterogeneity and reduced effective sample size caused by temporal alignment constraints. These results demonstrate that increasing spatial scope does not automatically improve generalization.

The comparison of DSM and DMS preprocessing pipelines showed that the timing of SSA smoothing influences predictive stability. In most cases, DMS inputs improved model performance and reduced variability across folds, although improvements were not uniform across all wells. This confirms that data structuring decisions materially affect neural network behaviour.

Sensitivity analysis using SHAP and OAT approaches confirmed that lagged GWL variables dominate model behaviour, reflecting the strong temporal persistence of groundwater systems. Precipitation and hydrometric variables contributed more strongly in shallow and watershed-based configurations. These results demonstrate that the models learned physically coherent relationships.

7.1 Limitations

Despite the comprehensive analytical framework adopted in this thesis, several limitations should be acknowledged.

First, meteorological data availability posed a significant constraint. Many Environment and ECCC stations exhibited substantial missing records, necessitating interpolation and supplementation with ERA5 reanalysis data. Although this approach improved spatial coverage and temporal continuity, reanalysis datasets may introduce systematic bias compared to direct observations, potentially affecting correlation strength and model calibration.

Second, temporal alignment across wells limited the effective sample size in combined and aggregated modelling scenarios. Because models required synchronized timestamps across predictors and target wells, wells with shorter or incomplete records constrained the overall training window. This reduction in usable observations partially explains the poor performance of the aggregated model and limits the generalizability of region-wide modelling conclusions.

Third, the predictive framework focused on FFNNs. While FFNNs demonstrated strong performance, alternative architectures such as RNNs or LSTM models may better capture temporal dependencies in groundwater systems. The present study did not evaluate these architectures.

Fourth, anthropogenic influences such as groundwater extraction, land-use change, and engineered drainage systems were not explicitly incorporated into the modelling framework. Although hydrometric and meteorological variables capture part of the system response, omission of direct pumping and management data may limit the explanatory power of the models in heavily influenced regions.

Finally, the study focused on Southern Ontario, with detailed modelling centered on the BOM region. While the modelling framework is transferable, hydrogeological characteristics vary across regions, and results may not directly generalize to fundamentally different aquifer systems.

7.2 Recommendations

Building upon the findings of this thesis, several avenues for future research are recommended.

First, future studies should incorporate anthropogenic variables such as pumping records, land-use dynamics, and groundwater management data. Including these drivers may improve predictive accuracy and provide clearer interpretation of long-term trends.

Second, advanced deep learning architectures designed for sequential data, including LSTM and gated recurrent units (GRU), should be evaluated and compared with FFNNs. These architectures may better represent temporal memory effects observed in groundwater systems.

Third, transfer learning and cross-regional validation approaches should be explored. Evaluating whether models trained in one watershed can be adapted to another would provide insight into model portability and regional scalability.

Fourth, hybrid modelling frameworks that integrate physically based groundwater models with data-driven approaches may enhance interpretability while preserving predictive flexibility. Such integration could reduce reliance on purely empirical relationships.

Fifth, climate change scenario analysis using projected precipitation and temperature inputs would extend the modelling framework toward forward-looking water resource planning.

Finally, future work should continue evaluating preprocessing strategies and spatial aggregation structures in different hydrogeological settings to determine whether the findings observed in this study hold under varying climatic and geological conditions.

References

- Afshari Nia, M., Panahi, F., & Ehteram, M. (2023). Convolutional Neural Network- ANN- E (Tanh): A New Deep Learning Model for Predicting Rainfall. *Water Resources Management*, 37(4), 1785–1810. <https://doi.org/10.1007/s11269-023-03454-8>
- Ahmadi, A., Olyaei, M., Heydari, Z., Emami, M., Zeynolabedin, A., Ghomlaghi, A., Daccache, A., Fogg, G. E., & Sadegh, M. (2022). Groundwater Level Modeling with Machine Learning: A Systematic Review and Meta-Analysis. *Water*, 14(6), Article 6. <https://doi.org/10.3390/w14060949>
- Akaike, H. (1974). A New Look at the Statistical Model Identification. *IEEE Transactions on Automatic Control*, 19, 716–723. <https://doi.org/10.1109/TAC.1974.1100705>
- Anderson, M. P., Woessner, W. W., & Hunt, R. J. (2015). *Applied Groundwater Modeling: Simulation of Flow and Advective Transport*. Academic Press.
- Andrei, F., Barbieri, M., Muteto, P. V., Ricolfi, L., Sappa, G., & Vitale, S. (2021). Water Resources Management Under Climate Change Pressure in Limpopo National Park Buffer Zone. In M. Abrunhosa, A. Chambel, S. Peppoloni, & H. I. Chaminé (Eds.), *Advances in Geoethics and Groundwater Management: Theory and Practice for a Sustainable Development* (pp. 129–132). Springer International Publishing. https://doi.org/10.1007/978-3-030-59320-9_28
- Arabameri, A., Arora, A., Pal, S. C., Mitra, S., Saha, A., Nalivan, O. A., Panahi, S., & Moayedi, H. (2021). K-Fold and State-of-the-Art Metaheuristic Machine Learning Approaches for Groundwater Potential Modelling. *Water Resources Management*, 35(6), 1837–1869. <https://doi.org/10.1007/s11269-021-02815-5>
- Araghinejad, S. (2014). *Data-Driven Modeling: Using MATLAB® in Water Resources and Environmental Engineering* (Vol. 67). Springer Netherlands. <https://doi.org/10.1007/978-94-007-7506-0>
- Asnicar, F., Thomas, A. M., Passerini, A., Waldron, L., & Segata, N. (2024). Machine learning for microbiologists. *Nature Reviews Microbiology*, 22(4), 191–205. <https://doi.org/10.1038/s41579-023-00984-1>

- Azimuth Environmental Consulting, Inc. (2019). *Phase Two Environmental Site Assessment: 1132 Snow Valley Road, Township of Springwater, Ontario*. Azimuth Environmental Consulting, Inc.
- Banihashem, S. M., Moradi, A., Evazzadeh, B., Namvar, F., & Fang, Z. N. (2024). Biogenically synthesized nanoparticles in wastewater treatment; a greener approach: A review. *Clean Technologies and Environmental Policy*, 26(6), 1731–1754. <https://doi.org/10.1007/s10098-023-02720-y>
- Barbieri, M., Barberio, M. D., Banzato, F., Billi, A., Boschetti, T., Franchini, S., Gori, F., & Petitta, M. (2023). Climate change and its effect on groundwater quality. *Environmental Geochemistry and Health*, 45(4), 1133–1144. <https://doi.org/10.1007/s10653-021-01140-5>
- Beale, M. H., Hagan, M. T., & Demuth, H. B. (2018). *Neural Network Toolbox™ Reference*. The MathWorks, Inc. <https://www.mathworks.com>
- Belsley, D. A., Kuh, E., & Welsch, R. E. (2005). *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*. John Wiley & Sons.
- Bennett, N. D., Croke, B. F. W., Guariso, G., Guillaume, J. H. A., Hamilton, S. H., Jakeman, A. J., Marsili-Libelli, S., Newham, L. T. H., Norton, J. P., Perrin, C., Pierce, S. A., Robson, B., Seppelt, R., Voinov, A. A., Fath, B. D., & Andreassian, V. (2013). Characterising performance of environmental models. *Environmental Modelling & Software*, 40, 1–20. <https://doi.org/10.1016/j.envsoft.2012.09.011>
- Bentley, J. L. (1975). Multidimensional binary search trees used for associative searching. *Commun. ACM*, 18(9), 509–517. <https://doi.org/10.1145/361002.361007>
- Bergstra, J., & Bengio, Y. (2012). Random search for hyper-parameter optimization. *J. Mach. Learn. Res.*, 13(null), 281–305.
- Bhatti, A. Z., Farooque, A. A., Krouglicof, N., Li, Q., Peters, W., Abbas, F., & Acharya, B. (2021). An Overview of Climate Change Induced Hydrological Variations in Canada for Irrigation Strategies. *Sustainability*, 13(9), Article 9. <https://doi.org/10.3390/su13094833>

- Bhatti, A. Z., Farooque, A. A., Li, Q., Abbas, F., & Acharya, B. (2021a). Spatial Distribution and Sustainability Implications of the Canadian Groundwater Resources under Changing Climate. *Sustainability*, 13(17), Article 17. <https://doi.org/10.3390/su13179778>
- Bhatti, A. Z., Farooque, A. A., Li, Q., Abbas, F., & Acharya, B. (2021b). Spatial Distribution and Sustainability Implications of the Canadian Groundwater Resources under Changing Climate. *Sustainability*, 13(17), 9778. <https://doi.org/10.3390/su13179778>
- Boo, K. B. W., El-Shafie, A., Othman, F., Khan, Md. M. H., Birima, A. H., & Ahmed, A. N. (2024). Groundwater level forecasting with machine learning models: A review. *Water Research*, 252, 121249. <https://doi.org/10.1016/j.watres.2024.121249>
- Borchardt, M. A., Bradbury, K. R., Gotkowitz, M. B., Cherry, J. A., & Parker, B. L. (2007). Human Enteric Viruses in Groundwater from a Confined Bedrock Aquifer. *Environmental Science & Technology*, 41(18), 6606–6612. <https://doi.org/10.1021/es071110+>
- Box, G. E. P., Jenkins, G. M., Reinsel, G. C., & Ljung, G. M. (2015a). *Time Series Analysis: Forecasting and Control*. John Wiley & Sons.
- Box, G. E. P., Jenkins, G. M., Reinsel, G. C., & Ljung, G. M. (2015b). *Time Series Analysis: Forecasting and Control*. John Wiley & Sons.
- Breusch, T. s. (1978). Testing for Autocorrelation in Dynamic Linear Models. *Australian Economic Papers*, 17(31), 334. <https://doi.org/10.1111/j.1467-8454.1978.tb00635.x>
- Breusch, T. S., & Pagan, A. R. (1979). A Simple Test for Heteroscedasticity and Random Coefficient Variation. *Econometrica*, 47(5), 1287–1294. <https://doi.org/10.2307/1911963>
- Brown, J., Ward, J., & Hutchinson, M. F. (2018). *Groundwater modelling for large scale mining impact assessments* [Technical Report]. Commonwealth Scientific and Industrial Research Organisation (CSIRO). <https://publications.csiro.au/rpr/download?pid=csiro:EP183502&dsid=DS1>
- Burt, A. K., & Dodge, J. E. P. (2011). *Three-dimensional modelling of surficial deposits in the Barrie-Oro Moraine area of southern Ontario* (Groundwater Resources Study 11). Ontario Geological Survey.

- Camps-Valls, G., Fernández-Torres, M.-Á., Cohrs, K.-H., Höhl, A., Castelletti, A., Pacal, A., Robin, C., Martinuzzi, F., Papoutsis, I., Prapas, I., Pérez-Aracil, J., Weigel, K., Gonzalez-Calabuig, M., Reichstein, M., Rabel, M., Giuliani, M., Mahecha, M. D., Popescu, O.-I., Pellicer-Valero, O. J., ... Williams, T. (2025). Artificial intelligence for modeling and understanding extreme weather and climate events. *Nature Communications*, 16(1), 1919. <https://doi.org/10.1038/s41467-025-56573-8>
- Cao, G., & Zheng, C. (2016). Signals of short-term climatic periodicities detected in the groundwater of North China Plain. *Hydrological Processes*, 30(4), 515–533. <https://doi.org/10.1002/hyp.10631>
- Caruana, R. (1997). Multitask Learning. *Machine Learning*, 28(1), 41–75. <https://doi.org/10.1023/A:1007379606734>
- Chang, J., Wang, G., & Mao, T. (2015). Simulation and prediction of suprapermafrost groundwater level variation in response to climate change using a neural network model. *Journal of Hydrology*, 529, 1211–1220. <https://doi.org/10.1016/j.jhydrol.2015.09.038>
- Chapman, L. J., & Putnam, D. F. (1984). *The Physiography of Southern Ontario* (Special Volume 2) [Topographic/Physiographic Map]. Ontario Geological Survey.
- Chau, K. W., & Wu, C. L. (2010). A Hybrid Model Coupled with Singular Spectrum Analysis for Daily Rainfall Prediction. *Journal of Hydroinformatics*, 12(4), 458–473. <https://doi.org/10.2166/hydro.2010.032>
- Chidepudi, S. K. R., Massei, N., Jardani, A., Dieppois, B., Henriot, A., & Fournier, M. (2025). Training deep learning models with a multi-station approach and static aquifer attributes for groundwater level simulation: What is the best way to leverage regionalised information? *Hydrology and Earth System Sciences*, 29(4), 841–861. <https://doi.org/10.5194/hess-29-841-2025>
- City of Barrie. (2008). *Lake Simcoe Phosphorus Reduction Strategy: Chapter 8 – City of Barrie*. https://www.simcoe.com/news/new-wave-of-water-in-barrie/article_274bc0cb-c855-5ae2-b705-6249ead38723.html

- Coelho, C. D., Faria, A. C. S., & Marques, E. A. G. (2017). Comparative analysis of different boundary conditions and their influence on numerical hydrogeological modeling of Palmital watershed, southeast Brazil. *Journal of Hydrology: Regional Studies*, 12, 210–219.
<https://doi.org/10.1016/j.ejrh.2017.05.006>
- Cuddy, D. (2013). *Tier 2 Water Budget Analysis and Water Quantity Stress Assessment for the Oro North, Oro South, and Hawkestone Creeks Subwatersheds*.
- Daliakopoulos, I. N., Coulibaly, P., & Tsanis, I. K. (2005). Groundwater level forecasting using artificial neural networks. *Journal of Hydrology*, 309(1), 229–240.
<https://doi.org/10.1016/j.jhydrol.2004.12.001>
- Di Nunno, F., & Granata, F. (2020). Groundwater level prediction in Apulia region (Southern Italy) using NARX neural network. *Environmental Research*, 190, 110062.
<https://doi.org/10.1016/j.envres.2020.110062>
- Dingman, S. L. (2015). *Physical Hydrology: Third Edition*. Waveland Press.
- Draper, N. R., & Smith, H. (1998). *Applied Regression Analysis* (1st ed.). Wiley.
<https://doi.org/10.1002/9781118625590>
- Durbin, J., & Watson, G. S. (1950). Testing for serial correlation in least squares regression. I. *Biometrika*, 37(3–4), 409–428. <https://doi.org/10.1093/biomet/37.3-4.409>
- Ebrahimi, H., & Rajaei, T. (2017). Simulation of groundwater level variations using wavelet combined with neural network, linear regression and support vector machine. *Global and Planetary Change*, 148, 181–191. <https://doi.org/10.1016/j.gloplacha.2016.11.014>
- Ehteram, M., & Banadkooki, F. B. (2023). A Developed Multiple Linear Regression (MLR) Model for Monthly Groundwater Level Prediction. *Water*, 15(22), Article 22.
<https://doi.org/10.3390/w15223940>
- Environment and Climate Change Canada. (n.d.). *Glossary*. Government of Canada - Environment and Climate Change Canada. Retrieved May 24, 2025, from https://climate.weather.gc.ca/glossary_e.html

- Environment and Climate Change Canada. (2023). *Adjusted and homogenized Canadian climate data* [Datasets]. <https://www.canada.ca/en/environment-climate-change/services/climate-change/canadian-climate-data-homogenization.html>
- Environmental Reporting BC. (2024). *Long-term Trends in Groundwater Levels in B.C.* Ministry of Environment and Climate Change Strategy, Government of British Columbia. <https://www2.gov.bc.ca/gov/content/environment/air-land-water/water/water-quality/groundwater/groundwater-levels>
- Evans, S. W., Jones, N. L., Williams, G. P., Ames, D. P., & Nelson, E. J. (2020). Groundwater Level Mapping Tool: An open source web application for assessing groundwater sustainability. *Environmental Modelling & Software*, 131, 104782. <https://doi.org/10.1016/j.envsoft.2020.104782>
- Ewen, J. (2011). Hydrograph matching method for measuring model performance. *Journal of Hydrology*, 408(1), 178–187. <https://doi.org/10.1016/j.jhydrol.2011.07.038>
- Famiglietti, J. S. (2014). The global groundwater crisis. *Nature Climate Change*, 4(11), 945–948. <https://doi.org/10.1038/nclimate2425>
- Fatolahzadeh Gheysari, A., Maghoul, P., Ojo, E. R., & Shalaby, A. (2024). Reliability of ERA5 and ERA5-Land reanalysis data in the Canadian Prairies. *Theoretical and Applied Climatology*, 155(4), 3087–3098. <https://doi.org/10.1007/s00704-023-04771-z>
- FedDev Ontario Pocketbook*. (2022). Government of Canada. <https://feddev-ontario.canada.ca/en/transparency/briefing-documents/transition-materials-prepared-minister-2022/briefing-material-honourable-filomena-tassi-feddev-ontario-agency-overview-2022-pocketbook>
- Fernandes, V. J., de Louw, P. G. B., Bartholomeus, R. P., & Ritsema, C. J. (2024a). Machine learning for faster estimates of groundwater response to artificial aquifer recharge. *Journal of Hydrology*, 637, 131418. <https://doi.org/10.1016/j.jhydrol.2024.131418>

- Fernandes, V. J., de Louw, P. G. B., Bartholomeus, R. P., & Ritsema, C. J. (2024b). Machine learning for faster estimates of groundwater response to artificial aquifer recharge. *Journal of Hydrology*, 637, 131418. <https://doi.org/10.1016/j.jhydrol.2024.131418>
- Fetter, C. W. (2018). *Applied Hydrogeology: Fourth Edition*. Waveland Press.
- Filzmoser, P., Maronna, R., & Werner, M. (2008). Outlier identification in high dimensions. *Computational Statistics & Data Analysis*, 52(3), 1694–1711. <https://doi.org/10.1016/j.csda.2007.05.018>
- Foroumandi, E., Nourani, V., & Sharghi, E. (2021). Climate change or regional human impacts? Remote sensing tools, artificial neural networks, and wavelet approaches aim to solve the problem. *Hydrology Research*, 52(1), 176–195. <https://doi.org/10.2166/nh.2020.112>
- Freeze, R. A., & Cherry, J. A. (1979). *Groundwater*. Prentice-Hall.
- Friedman, J. H., Bentley, J. L., & Finkel, R. A. (1977). An Algorithm for Finding Best Matches in Logarithmic Expected Time. *ACM Transactions on Mathematical Software*, 3(3), 209–226. <https://doi.org/10.1145/355744.355745>
- Frimpong, B. F., Koranteng, A., & Molkenthin, F. (2022). Analysis of temperature variability utilising Mann–Kendall and Sen’s slope estimator tests in the Accra and Kumasi Metropolises in Ghana. *Environmental Systems Research*, 11(1), 24. <https://doi.org/10.1186/s40068-022-00269-1>
- Galton, F. (1886). Regression Towards Mediocrity in Hereditary Stature. *The Journal of the Anthropological Institute of Great Britain and Ireland*, 15, 246–263. <https://doi.org/10.2307/2841583>
- Ghosh, A., Adhikary, P. P., Bera, B., Bhunia, G. S., & Shit, P. K. (2022). Assessment of groundwater potential zone using MCDA and AHP techniques: Case study from a tropical river basin of India. *Applied Water Science*, 12(3), 37. <https://doi.org/10.1007/s13201-021-01548-5>
- Ghosh, D., Mandal, M., Karmakar, M., Banerjee, M., & Mandal, D. (2020). Application of geospatial technology for delineating groundwater potential zones in the Gandheswari watershed, West

- Bengal. *Sustainable Water Resources Management*, 6(1), 14. <https://doi.org/10.1007/s40899-020-00372-0>
- Giordano, M. (2009). Global Groundwater? Issues and Solutions. *Annual Review of Environment and Resources*, 34(Volume 34, 2009), 153–178.
<https://doi.org/10.1146/annurev.enviro.030308.100251>
- Godfrey, L. G. (1978). Testing for Higher Order Serial Correlation in Regression Equations when the Regressors Include Lagged Dependent Variables. *Econometrica*, 46(6), 1303–1310.
<https://doi.org/10.2307/1913830>
- Golyandina, N. (2011). On the choice of parameters in Singular Spectrum Analysis and related subspace-based methods. *arXiv Preprint*. <http://arxiv.org/abs/1005.4374v2>
- Goodarzi, M. R., Bafrouei, H. B., & Vazirian, M. (2024). Insight into groundwater level prediction with feature effectiveness: Comparison of machine learning and numerical methods. *Hydrology Research*, 56(1), 74–92. <https://doi.org/10.2166/nh.2024.135>
- Goswami, U. P., Déry, S. J., & Fortin, V. (2024). Performance Evaluation of High-resolution Reanalysis Datasets Over North-central British Columbia. *Atmosphere-Ocean*, 62(3), 222–242.
<https://doi.org/10.1080/07055900.2024.2308878>
- Government of Canada, S. C. (2022, February 9). *Profile table, Census Profile, 2021 Census of Population—Ontario [Province]*. <https://www12.statcan.gc.ca/census-recensement/2021/dp-pd/prof/index.cfm?Lang=E>
- Government of Ontario. (2019). *Provincial Groundwater Monitoring Network*.
<https://data.ontario.ca/dataset/provincial-groundwater-monitoring-network/resource/2f97f6cf-9b2d-4dd7-8c41-64f51f1773b9>
- Gupta, M. S., Srivastava, A., & Kumar, K. (2025). RORA: Reinforcement learning based optimal distributed resource allocation strategies in vehicular cognitive radio networks for 6G. *Vehicular Communications*, 52, 100882. <https://doi.org/10.1016/j.vehcom.2025.100882>

- Gutierrez, K. S., & LePrevost, C. E. (2016). Climate Justice in Rural Southeastern United States: A Review of Climate Change Impacts and Effects on Human Health. *International Journal of Environmental Research and Public Health*, 13(2), Article 2.
<https://doi.org/10.3390/ijerph13020189>
- Hansen, P. C., Pereyra, V., & Scherer, G. (2013). *Least Squares Data Fitting with Applications*. JHU Press.
- Hassani, H., Mahmoudvand, R., & Zokaei, M. (2011). Separability and window length in singular spectrum analysis. *Comptes Rendus Mathematique*, 349(17), 987–990.
<https://doi.org/10.1016/j.crma.2011.07.012>
- Hassani, H., & Zhigljavsky, A. (2009). Singular spectrum analysis: Methodology and application to economics data. *Journal of Systems Science and Complexity*, 22(3), 372–394.
<https://doi.org/10.1007/s11424-009-9171-9>
- Hastie, T., Tibshirani, R., & Friedman, J. (2009a). Linear Methods for Regression. In *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (pp. 43–99). Springer.
https://doi.org/10.1007/978-0-387-84858-7_3
- Hastie, T., Tibshirani, R., & Friedman, J. (2009b). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (2nd Edition). Springer.
- Hersbach, H., Bell, B., & Berrisford, P. (2020). The ERA5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, 146(730), 1999–2049. <https://doi.org/10.1002/qj.3803>
- Hsieh, W. W. (2023a). *Introduction to Environmental Data Science*. Cambridge University Press.
<https://doi.org/10.1017/9781107588493>
- Hsieh, W. W. (2023b). *Introduction to Environmental Data Science* (1st ed.). Cambridge University Press.
<https://doi.org/10.1017/9781107588493>
- Huang, R., Ma, C., Ma, J., Huangfu, X., & He, Q. (2021). Machine learning in natural and engineered water systems. *Water Research*, 205, 117666. <https://doi.org/10.1016/j.watres.2021.117666>

- Ifediegwu, S. I. (2021). Assessment of groundwater potential zones using GIS and AHP techniques: A case study of the Lafia district, Nasarawa State, Nigeria. *Applied Water Science*, 12(1), 10. <https://doi.org/10.1007/s13201-021-01556-5>
- Jamin, P., Cochand, M., Dagenais, S., Lemieux, J.-M., Fortier, R., Molson, J., & Brouyère, S. (2020). Direct measurement of groundwater flux in aquifers within the discontinuous permafrost zone: An application of the finite volume point dilution method near Umiujaq (Nunavik, Canada). *Hydrogeology Journal*. <https://doi.org/10.1007/s10040-020-02108-y>
- Janardhanan, S., Nair, A. S., Indu, J., Pagendam, D., & Kaushika, G. S. (2023). Estimation of groundwater storage loss for the Indian Ganga Basin using multiple lines of evidence. *Scientific Reports*, 13, 1797. <https://doi.org/10.1038/s41598-023-28615-y>
- Javadzadeh, H., Ataie-Ashtiani, B., Hosseini, S. M., & Simmons, C. T. (2020). Interaction of lake-groundwater levels using cross-correlation analysis: A case study of Lake Urmia Basin, Iran. *Science of The Total Environment*, 729, 138822. <https://doi.org/10.1016/j.scitotenv.2020.138822>
- Jeihouni, E., Mohammadi, M., Eslamian, S., & Zareian, M. J. (2019). Potential impacts of climate change on groundwater level through hybrid soft-computing methods: A case study—Shabestar Plain, Iran. *Environmental Monitoring and Assessment*, 191(10), 620. <https://doi.org/10.1007/s10661-019-7784-6>
- Jeong, J., Park, E., Han, W. S., Kim, K., Choung, S., & Chung, I. M. (2017). Identifying outliers of non-Gaussian groundwater state data based on ensemble estimation for long-term trends. *Journal of Hydrology*, 548, 135–144. <https://doi.org/10.1016/j.jhydrol.2017.02.058>
- Jia, X., O'Connor, D., Hou, D., Jin, Y., Li, G., Zheng, C., Ok, Y. S., Tsang, D. C. W., & Luo, J. (2019). Groundwater depletion and contamination: Spatial distribution of groundwater resources sustainability in China. *Science of The Total Environment*, 672, 551–562. <https://doi.org/10.1016/j.scitotenv.2019.03.457>
- Jiang, Y., Somers, G., Mutch, J., & Island, P. E. (1999). *Application of numerical modeling to groundwater assessment and management in prince edward island*. 57th Canadian Geotechnical

- Conference (57ième Congrès Canadien de Géotechnique), Charlottetown, Prince Edward Island, Canada.
- Jordahl, K., Van den Bossche, J., Fleischmann, M., Wasserman, J., McBride, J., Gerard, J., & Bilogur, A. (2020). *geopandas/geopandas: V0.8.1*. Zenodo. <https://doi.org/10.5281/zenodo.3946761>
- Jyrkama, M. I., & Sykes, J. F. (2007). The impact of climate change on spatially varying groundwater recharge in the grand river watershed (Ontario). *Journal of Hydrology*, 338(3), 237–250. <https://doi.org/10.1016/j.jhydrol.2007.02.036>
- Kardan Moghaddam, H., Ghordoyee Milan, S., Kayhomayoon, Z., Rahimzadeh kivi, Z., & Arya Azar, N. (2021). The prediction of aquifer groundwater level based on spatial clustering approach using machine learning. *Environmental Monitoring and Assessment*, 193(4), 173. <https://doi.org/10.1007/s10661-021-08961-y>
- Kendall, M. G. (1948). *Rank correlation methods*. Griffin.
- Kou, G., Eti, S., Yüksel, S., Dinçer, H., Ergün, E., & Gökalp, Y. (2024). Innovative solution suggestions for financing electric vehicle charging infrastructure investments with a novel artificial intelligence-based fuzzy decision-making modelling. *Artificial Intelligence Review*, 58(1), 16. <https://doi.org/10.1007/s10462-024-11012-w>
- Kouziokas, G. N., Chatzigeorgiou, A., & Perakis, K. (2018). Multilayer Feed Forward Models in Groundwater Level Forecasting Using Meteorological Data in Public Management. *Water Resources Management*, 32(15), 5041–5052. <https://doi.org/10.1007/s11269-018-2126-y>
- Kumar, R., Singh, M. P., Roy, B., & Shahid, A. H. (2021). A Comparative Assessment of Metaheuristic Optimized Extreme Learning Machine and Deep Neural Network in Multi-Step-Ahead Long-term Rainfall Prediction for All-Indian Regions. *Water Resources Management*, 35(6), 1927–1960. <https://doi.org/10.1007/s11269-021-02822-6>
- Kundzewicz, Z. w., & Döll, P. (2009). Will groundwater ease freshwater stress under climate change? *Hydrological Sciences Journal*, 54(4), 665–675. <https://doi.org/10.1623/hysj.54.4.665>

- Lachassagne, P. (2021). What Is Groundwater? How to Manage and Protect Groundwater Resources. *Annals of Nutrition and Metabolism*, 76(Suppl. 1), 17–24. <https://doi.org/10.1159/000515024>
- Lake Simcoe Region Conservation Authority. (2024). *Tier 2 Water Budget Analysis—Oro*. Lake Simcoe Region Conservation Authority.
- Lake Simcoe Region Conservation Authority (LSRCA). (2013). *Oro and Hawkestone Creeks Subwatershed Plan*. Lake Simcoe Region Conservation Authority.
- Lall, U., Josset, L., & Russo, T. (2020). A Snapshot of the World's Groundwater Challenges. *Annual Review of Environment and Resources*, 45(Volume 45, 2020), 171–194. <https://doi.org/10.1146/annurev-environ-102017-025800>
- Langevin, C. D., Hughes, J. D., Banta, E. R., Niswonger, R. G., Panday, S., & Provost, A. M. (2017). Documentation for the MODFLOW 6 Groundwater Flow Model. In *Techniques and Methods* (6-A55). U.S. Geological Survey. <https://doi.org/10.3133/tm6A55>
- Larocque, M., Levison, J., Gagné, S., & Saleem, S. (2019). Groundwater use for agricultural production—Current water budget and expected trends under climate change. *Final Report*.
- Lee, S., Lee, K.-K., & Yoon, H. (2019). Using artificial neural network models for groundwater level forecasting and assessment of the relative impacts of influencing factors. *Hydrogeology Journal*, 27(2), 567–579. <https://doi.org/10.1007/s10040-018-1866-3>
- Lemieux, C. J., & Scott, D. J. (2011). Changing Climate, Challenging Choices: Identifying and Evaluating Climate Change Adaptation Options for Protected Areas Management in Ontario, Canada. *Environmental Management*, 48(4), 675–690. <https://doi.org/10.1007/s00267-011-9700-x>
- Li, D., Liu, S., Wang, B., Yu, C., Zheng, P., & Li, W. (2025). Trustworthy AI for human-centric smart manufacturing: A survey. *Journal of Manufacturing Systems*, 78, 308–327. <https://doi.org/10.1016/j.jmsy.2024.11.020>

- Li, H., Lu, Z., Zhang, Z., & Tanasescu, C. (2025). How does artificial intelligence affect manufacturing firms' energy intensity? *Energy Economics*, 141, 108109.
<https://doi.org/10.1016/j.eneco.2024.108109>
- Li, J., Mao, X., & Li, M. (2017). Modeling hydrological processes in oasis of Heihe River Basin by landscape unit-based conceptual models integrated with FEFLOW and GIS. *Agricultural Water Management, Special Issue on Improving Agricultural Water Productivity to Ensure Food Security under Changing Environments Overseen by: Brent Clothier*, 179, 338–351.
<https://doi.org/10.1016/j.agwat.2016.09.007>
- Li, J., & Wang, S. (2022). Seasonal variations and long-term trends of groundwater over the Canadian landmass. *Hydrogeology Journal*, 30(2), 401–415. <https://doi.org/10.1007/s10040-022-02460-1>
- Lippa, N. A. (2010). *Assessing Groundwater Flux from Underlying Fractured Bedrock to the Overburden Aquifer System* [Master's Thesis, University of New Brunswick]. <https://unbscholar.lib.unb.ca/>
- Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 30.
- Machiwal, D., & Jha, M. K. (2006). *Time series analysis of hydrologic data for water resources planning and management: A review*. 54, 237–257.
- Mahmoodzadeh, D., Ketabchi, H., Ataie-Ashtiani, B., & Simmons, C. T. (2014). Conceptualization of a fresh groundwater lens influenced by climate change: A modeling study of an arid-region island in the Persian Gulf, Iran. *Journal of Hydrology*, 519, 399–413.
<https://doi.org/10.1016/j.jhydrol.2014.07.010>
- Maier, H. R., & Dandy, G. C. (2000). Neural networks for the prediction and forecasting of water resources variables: A review of modelling issues and applications. *Environmental Modelling & Software*, 15(1), 101–124. [https://doi.org/10.1016/S1364-8152\(99\)00007-9](https://doi.org/10.1016/S1364-8152(99)00007-9)
- Maier, H. R., Jain, A., Dandy, G. C., & Sudheer, K. P. (2010). Methods used for the development of neural networks for the prediction of water resource variables in river systems: Current status and

- future directions. *Environmental Modelling & Software*, 25(8), 891–909.
<https://doi.org/10.1016/j.envsoft.2010.02.003>
- Malekzadeh, M., Kardar, S., Saeb, K., Shabanlou, S., & Taghavi, L. (2019). A Novel Approach for Prediction of Monthly Ground Water Level Using a Hybrid Wavelet and Non-Tuned Self-Adaptive Machine Learning Model. *Water Resources Management*, 33(4), 1609–1628.
<https://doi.org/10.1007/s11269-019-2193-8>
- Mann, H. B. (1945). Nonparametric Tests Against Trend. *Econometrica*, 13(3), 245–259.
<https://doi.org/10.2307/1907187>
- Markle, J., & Harris, M. (2019). *Groundwater Levels in Ontario: A Trend Analysis using the Provincial Groundwater Monitoring Network*.
- Marques, C. A. F., Ferreira, J. A., Rocha, A., Castanheira, J. M., Melo-Gonçalves, P., Vaz, N., & Dias, J. M. (2006). Singular spectrum analysis and forecasting of hydrological time series. *Physics and Chemistry of the Earth, Parts A/B/C, Time Series Analysis in Hydrology*, 31(18), 1172–1179.
<https://doi.org/10.1016/j.pce.2006.02.061>
- Massey Jr., F. J. (1951). The Kolmogorov-Smirnov Test for Goodness of Fit. *Journal of the American Statistical Association*, 46(253), 68–78. <https://doi.org/10.1080/01621459.1951.10500769>
- McCradden, M. D., London, A. J., Gichoya, J. W., Sendak, M., Erdman, L., Stedman, I., Oakden-Rayner, L., Akrou, I., Anderson, J. A., Farmer, L.-A., Greer, R., Goldenberg, A., Ho, Y., Joshi, S., Louise, J., Mamdani, M., Mazwi, M. L., Mohamud, A., Palmer, L. J., ... Liu, X. (2025). CANAIRI: The Collaboration for Translational Artificial Intelligence Trials in healthcare. *Nature Medicine*, 31(1), 9–11. <https://doi.org/10.1038/s41591-024-03364-1>
- McDonald, M. G., & Harbaugh, A. W. (1988). *A Modular Three-dimensional Finite-difference Groundwater Flow Model*. U.S. Geological Survey.
- McDonald, M. G., Harbaugh, A. W., & Modflow, T. O. A. O. (2003). The History of MODFLOW. *Groundwater*, 41(2), 280–283. <https://doi.org/10.1111/j.1745-6584.2003.tb02591.x>

- McDonough, L. K., Santos, I. R., Andersen, M. S., O'Carroll, D. M., Rutledge, H., Meredith, K., Oudone, P., Bridgeman, J., Gooddy, D. C., Sorensen, J. P. R., Lapworth, D. J., MacDonald, A. M., Ward, J., & Baker, A. (2020). Changes in global groundwater organic carbon driven by climate change and urbanization. *Nature Communications*, *11*(1), 1279. <https://doi.org/10.1038/s41467-020-14946-1>
- Morsy, K. M., Morsy, A. M., & Hassan, A. E. (2018). Groundwater sustainability: Opportunity out of threat. *Groundwater for Sustainable Development*, *7*, 277–285. <https://doi.org/10.1016/j.gsd.2018.06.010>
- Müller, J., Park, J., Sahu, R., Varadharajan, C., Arora, B., Faybishenko, B., & Agarwal, D. (2021). Surrogate optimization of deep neural networks for groundwater predictions. *Journal of Global Optimization*, *81*(1), 203–231. <https://doi.org/10.1007/s10898-020-00912-0>
- Munday, P. L., Donelson, J. M., & Domingos, J. A. (2017). Potential for adaptation to climate change in a coral reef fish. *Global Change Biology*, *23*(1), 307–317. <https://doi.org/10.1111/gcb.13419>
- Musie, W., & Gonfa, G. (2023). Fresh water resource, scarcity, water salinity challenges and possible remedies: A review. *Heliyon*, *9*(8), e18685. <https://doi.org/10.1016/j.heliyon.2023.e18685>
- Niazkar, M., Piraei, R., Menapace, A., Dhawan, P., Torre, D. D., Larcher, M., & Righetti, M. (2023). Bias correction of ERA5-Land temperature data using standalone and ensemble machine learning models: A case of northern Italy. *Journal of Water and Climate Change*, *15*(1), 271–283. <https://doi.org/10.2166/wcc.2023.669>
- Nourani, V., Khodkar, K., Paknezhad, N. J., & Laux, P. (2022). Deep learning-based uncertainty quantification of groundwater level predictions. *Stochastic Environmental Research and Risk Assessment*, *36*(10), 3081–3107. <https://doi.org/10.1007/s00477-022-02181-7>
- Nowlan, L. (2005). *Buried treasure: Groundwater permitting and pricing in Canada* | *Policy Commons*. <https://policycommons.net/artifacts/1191797/buried-treasure/1744917/>
- Oil, Gas and Salt Resources Library. (2020). *OGSR Library*. <https://www.ogsrlibrary.com/>

- O'Neill, C. (2023). *A Comparison of Methods to Estimate Vertical Groundwater Flux Through Overburden Materials in Prince Edward Island* [Master's Thesis, University of New Brunswick].
Not available
- Ontario Ministry of Agriculture, Food and Rural Affairs. (2021, June). *Understanding groundwater*.
Government of Ontario. <https://www.ontario.ca/page/understanding-groundwater>
- Our World in Data. (2024). *Average temperature anomaly*. Our World in Data.
<https://ourworldindata.org/grapher/temperature-anomaly>
- Pathania, T., Bottacin-Busolin, A., Rastogi, A. K., & Eldho, T. I. (2019). Simulation of Groundwater Flow in an Unconfined Sloping Aquifer Using the Element-Free Galerkin Method. *Water Resources Management*, 33(8), 2827–2845. <https://doi.org/10.1007/s11269-019-02261-4>
- Patra, S. R., Chu, H.-J., & Tatas. (2023). Regional groundwater sequential forecasting using global and local LSTM models. *Journal of Hydrology: Regional Studies*, 47, 101442.
<https://doi.org/10.1016/j.ejrh.2023.101442>
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., & Cournapeau, D. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830.
- Poursaeid, M., Poursaeid, A. H., & Shabanlou, S. (2022). A Comparative Study of Artificial Intelligence Models and A Statistical Method for Groundwater Level Prediction. *Water Resources Management*, 36(5), 1499–1519. <https://doi.org/10.1007/s11269-022-03070-y>
- Qadir, M., Sharma, B. R., Bruggeman, A., Choukr-Allah, R., & Karajeh, F. (2007). Non-conventional water resources and opportunities for water augmentation to achieve food security in water scarce countries. *Agricultural Water Management*, 87(1), 2–22.
<https://doi.org/10.1016/j.agwat.2006.03.018>
- Radmanesh, F., Golabi, M. R., Khodabakhshi, F., Farzi, S., & Zeinali, M. (2020). Modeling aquifer hydrograph: Performance review of conceptual MODFLOW and simulator models. *Arabian Journal of Geosciences*, 13(5), 240. <https://doi.org/10.1007/s12517-020-5230-2>

- Rajaei, T., Ebrahimi, H., & Nourani, V. (2019). A review of the artificial intelligence methods in groundwater level modeling. *Journal of Hydrology*, 572, 336–351.
<https://doi.org/10.1016/j.jhydrol.2018.12.037>
- Rajeevan, U., & Mishra, B. K. (2020). Sustainable management of the groundwater resource of Jaffna, Sri Lanka with the participation of households: Insights from a study on household water consumption and management. *Groundwater for Sustainable Development*, 10, 100280.
<https://doi.org/10.1016/j.gsd.2019.100280>
- Ricolfi, L., Barbieri, M., Muteto, P. V., Nigro, A., Sappa, G., & Vitale, S. (2020). Potential toxic elements in groundwater and their health risk assessment in drinking water of Limpopo National Park, Gaza Province, Southern Mozambique. *Environmental Geochemistry and Health*, 42(9), 2733–2745. <https://doi.org/10.1007/s10653-019-00507-z>
- Rogojin, V. & Geo. (2024). *Provincial Groundwater Monitoring Network Program* [Technical Report]. Ministry of Environment, Conservation and Parks. <https://www.ontario.ca/page/map-provincial-groundwater-monitoring-network>
- Roshni, T., Jha, M. K., Deo, R. C., & Vandana, A. (2019). Development and Evaluation of Hybrid Artificial Neural Network Architectures for Modeling Spatio-Temporal Groundwater Fluctuations in a Complex Aquifer System. *Water Resources Management*, 33(7), 2381–2397.
<https://doi.org/10.1007/s11269-019-02253-4>
- Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (Eds.). (1986). Learning internal representations by error propagation. In *Parallel Distributed Processing: Explorations in the Microstructures of Cognition, Vol. 1* (pp. 318–362). The MIT Press. <https://doi.org/10.7551/mitpress/4943.001.0001>
- Sahoo, S., Russo, T. A., Elliott, J., & Foster, I. (2017). Machine learning algorithms for modeling groundwater level changes in agricultural regions of the U.S. *Water Resources Research*, 53(5), 3878–3895. <https://doi.org/10.1002/2016WR019933>

- Sahu, R. K., Müller, J., Park, J., Varadharajan, C., Arora, B., Faybishenko, B., & Agarwal, D. (2020a). Impact of Input Feature Selection on Groundwater Level Prediction From a Multi-Layer Perceptron Neural Network. *Frontiers in Water*, 2. <https://doi.org/10.3389/frwa.2020.573034>
- Sahu, R. K., Müller, J., Park, J., Varadharajan, C., Arora, B., Faybishenko, B., & Agarwal, D. (2020b). Impact of Input Feature Selection on Groundwater Level Prediction From a Multi-Layer Perceptron Neural Network. *Frontiers in Water*, 2. <https://doi.org/10.3389/frwa.2020.573034>
- Saltelli, A., Chan, K., & Scott, E. M. (2004). *Sensitivity Analysis in Practice: A Guide to Assessing Scientific Models*. John Wiley & Sons.
- Samuel, A. L. (1959). Some Studies in Machine Learning Using the Game of Checkers. *IBM Journal of Research and Development*, 3(3), 210–229. <https://doi.org/10.1147/rd.33.0210>
- Saroughi, M., Mirzania, E., Vishwakarma, D. K., Nivesh, S., Panda, K. C., & Daneshvar, F. A. (2023). A Novel Hybrid Algorithms for Groundwater Level Prediction. *Iranian Journal of Science and Technology, Transactions of Civil Engineering*, 47(5), 3147–3164. <https://doi.org/10.1007/s40996-023-01068-z>
- Schmidhuber, J. (2015). Deep learning in neural networks: An overview. *Neural Networks*, 61, 85–117. <https://doi.org/10.1016/j.neunet.2014.09.003>
- Secci, D., Giovanna Tanda, M., D’Oria, M., & Todaro, V. (2023). Artificial intelligence models to evaluate the impact of climate change on groundwater resources. *Journal of Hydrology*, 627, 130359. <https://doi.org/10.1016/j.jhydrol.2023.130359>
- Sen, P. K. (1968). Estimates of the Regression Coefficient Based on Kendall’s Tau. *Journal of the American Statistical Association*, 63(324), 1379–1389. <https://doi.org/10.1080/01621459.1968.10480934>
- Şen, Z. (2014). *Practical and Applied Hydrogeology*. Elsevier.
- Sharghi, E., Nourani, V., Zhang, Y., & Ghaneei, P. (2022). Conjunction of cluster ensemble-model ensemble techniques for spatiotemporal assessment of groundwater depletion in semi-arid plains. *Journal of Hydrology*, 610, 127984. <https://doi.org/10.1016/j.jhydrol.2022.127984>

- Sharpe, D. R., Piggott, A., Carter, T., Gerber, R. E., MacRitchie, S. M., De Loë, R. C., Strynka, S., & Zwiers, G. (2014). *Southern Ontario hydrogeological region*. <https://doi.org/10.4095/296947>
- Shen, C. (2017). A trans-disciplinary review of deep learning research for water resources scientists. *Water Resources Research*, 54. <https://doi.org/10.1029/2018WR022643>
- Singer, S. N., Cheng, C. K., & Scafe, M. G. (2003). Hydrogeology of Southern Ontario. *Environmental Monitoring and Reporting Branch, Ministry of the Environment*, (Second Edition).
- Snieder, E., Shakir, R., & Khan, U. T. (2020a). A comprehensive comparison of four input variable selection methods for artificial neural network flow forecasting models. *Journal of Hydrology*, 583, 124299. <https://doi.org/10.1016/j.jhydrol.2019.124299>
- Snieder, E., Shakir, R., & Khan, U. T. (2020b). A comprehensive comparison of four input variable selection methods for artificial neural network flow forecasting models. *Journal of Hydrology*, 583, 124299. <https://doi.org/10.1016/j.jhydrol.2019.124299>
- Snieder, E., Shakir, R., & Khan, U. T. (2020c). A comprehensive comparison of four input variable selection methods for artificial neural network flow forecasting models. *Journal of Hydrology*, 583, 124299. <https://doi.org/10.1016/j.jhydrol.2019.124299>
- Solgi, R., Loáiciga, H. A., & Kram, M. (2021). Long short-term memory neural network (LSTM-NN) for aquifer level time series forecasting using in-situ piezometric observations. *Journal of Hydrology*, 601, 126800. <https://doi.org/10.1016/j.jhydrol.2021.126800>
- Song, Z., Lu, C., Zhang, Y., Chen, J., Liu, W., Liu, B., & Shu, L. (2022). Spatiotemporal Distribution and Statistical Analysis of Abnormal Groundwater Level Rising in Poyang Lake Basin. *Water*, 14(12), Article 12. <https://doi.org/10.3390/w14121906>
- Sophocleous, M. (2002). Interactions between groundwater and surface water: The state of the science. *Hydrogeology Journal*, 10(1), 52–67. <https://doi.org/10.1007/s10040-001-0170-8>
- Statistics Canada. (2016). *Boundary Files, 2016 Census* [Dataset]. <https://open.canada.ca/data/en/dataset/a883eb14-0c0e-45c4-b8c4-b54c4a819edb/resource/12c03de6-c3f7-4f5f-bb5c-d479f2332842#additional-info>

- Steefel, C. I., Appelo, C. A. J., Arora, B., Jacques, D., Kalbacher, T., Kolditz, O., Lagneau, V., Lichtner, P. C., Mayer, K. U., Meeussen, J. C. L., Molins, S., Moulton, D., Shao, H., Šimůnek, J., Spycher, N., Yabusaki, S. B., & Yeh, G. T. (2015). Reactive transport codes for subsurface environmental simulation. *Computational Geosciences*, 19(3), 445–478. <https://doi.org/10.1007/s10596-014-9443-x>
- Sun, M., & Li, X. (2017). Window Length Selection of Singular Spectrum Analysis and Application to Precipitation Time Series. *Global NEST Journal*, 19(2), 306–317.
- Tao, H., Hameed, M. M., Marhoon, H. A., Zounemat-Kermani, M., Heddami, S., Kim, S., Sulaiman, S. O., Tan, M. L., Sa'adi, Z., Mehr, A. D., Allawi, M. F., Abba, S. I., Zain, J. M., Falah, M. W., Jamei, M., Bokde, N. D., Bayatvarkeshi, M., Al-Mukhtar, M., Bhagat, S. K., ... Yaseen, Z. M. (2022). Groundwater level prediction using machine learning models: A comprehensive review. *Neurocomputing*, 489, 271–308. <https://doi.org/10.1016/j.neucom.2022.03.014>
- Toronto and Region Conservation Authority (TRCA). (2024, July 11). *TRCA's Watershed and Ecosystems Reporting Hub*. ArcGIS StoryMaps. <https://storymaps.arcgis.com/collections/8c517b063c81449d8fba71ca02d4278f?item=2>
- Township of Oro-Medonte. (2008). *Lake Simcoe Phosphorus Reduction Strategy: Chapter 11 – Oro-Medonte*.
- Uc Castillo, J. L., Martínez Cruz, D. A., Ramos Leal, J. A., Tuxpan Vargas, J., Rodríguez Tapia, S. A., & Marín Celestino, A. E. (2022). Delineation of Groundwater Potential Zones (GWPZs) in a Semi-Arid Basin through Remote Sensing, GIS, and AHP Approaches. *Water*, 14(13), Article 13. <https://doi.org/10.3390/w14132138>
- Uc-Castillo, J. L., Marín-Celestino, A. E., Martínez-Cruz, D. A., Tuxpan-Vargas, J., & Ramos-Leal, J. A. (2023). A systematic review and meta-analysis of groundwater level forecasting with machine learning techniques: Current status and future directions. *Environmental Modelling & Software*, 168, 105788. <https://doi.org/10.1016/j.envsoft.2023.105788>

- United Nations. (n.d.). *Water and Sanitation* | *Department of Economic and Social Affairs*. Retrieved March 26, 2025, from <https://sdgs.un.org/topics/water-and-sanitation>
- U.S. Geological Survey. (2005). *Boundary condition packages – MODFLOW-2005 guide*. https://water.usgs.gov/ogw/modflow/MODFLOW-2005-Guide/boundary_condition_packages.html
- Vautard, R., Yiou, P., & Ghil, M. (1992). Singular-spectrum analysis: A toolkit for short, noisy chaotic signals. *Physica D: Nonlinear Phenomena*, 58(1), 95–126. [https://doi.org/10.1016/0167-2789\(92\)90103-T](https://doi.org/10.1016/0167-2789(92)90103-T)
- Vu, M. T., Jardani, A., Massei, N., & Fournier, M. (2021). Reconstruction of missing groundwater level data by using Long Short-Term Memory (LSTM) deep neural network. *Journal of Hydrology*, 597, 125776. <https://doi.org/10.1016/j.jhydrol.2020.125776>
- Warner, H. Katherine. (2000). Review: Making the invisible, visible. *Journal of Child Health Care*, 4(3), 123–126. <https://doi.org/10.1177/136749350000400306>
- World Wildlife Fund (WWF) Canada. (n.d.). *Freshwater Habitat* | *WWF-Canada*. WWF.CA. Retrieved March 23, 2025, from <https://wwf.ca/habitat/freshwater/>
- Wunsch, A., Liesch, T., & Broda, S. (2022a). Deep learning shows declining groundwater levels in Germany until 2100 due to climate change. *Nature Communications*, 13(1), 1221. <https://doi.org/10.1038/s41467-022-28770-2>
- Wunsch, A., Liesch, T., & Broda, S. (2022b). Deep learning shows declining groundwater levels in Germany until 2100 due to climate change. *Nature Communications*, 13(1), 1221. <https://doi.org/10.1038/s41467-022-28770-2>
- Xu, Y., & Lofranco, C. (2023). *Water Taking* [Dataset]. Ontario Open Data Catalog. <https://data.ontario.ca/dataset/water-taking>
- Yin, J., Medellín-Azuara, J., & Escrivá-Bou, A. (2022). Groundwater levels hierarchical clustering and regional groundwater drought assessment in heavily drafted aquifers. *Hydrology Research*, 53(7), 1031–1046. <https://doi.org/10.2166/nh.2022.048>

- Zhang, Y., & Yang, Q. (2022). A Survey on Multi-Task Learning. *IEEE Transactions on Knowledge and Data Engineering*, 34(12), 5586–5609. <https://doi.org/10.1109/TKDE.2021.3070203>
- Zhu, H., Liu, J., Zhou, X., Chen, X., Qiu, X., Bello, R. L., & Deng, Z. (2020). The Ontario Climate Data Portal, a user-friendly portal of Ontario-specific climate projections. *Scientific Data*, 7(1), 147. <https://doi.org/10.1038/s41597-020-0489-4>

Appendices

Appendix A

ERA5 vs. PGMN Wells in Southern Ontario

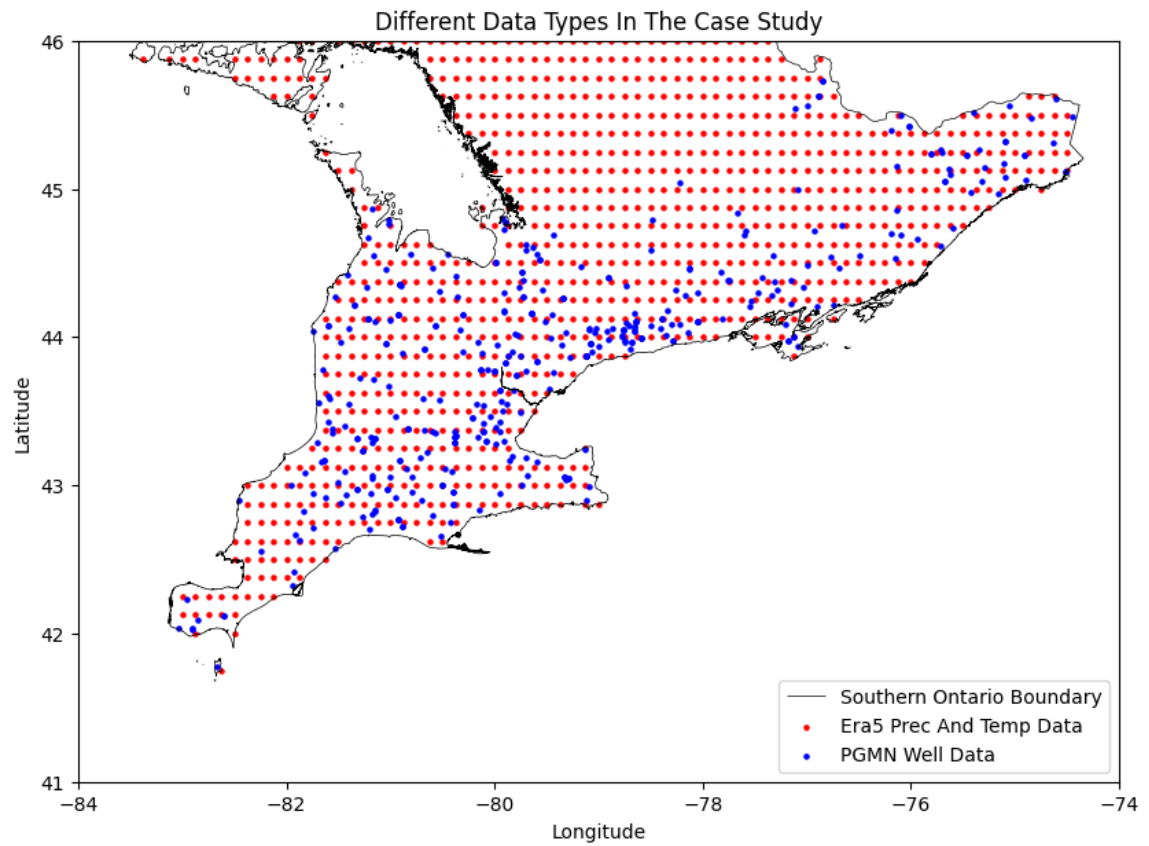


Figure A.1-Spatial distribution of ERA5 grids and PGMN wells in Southern Ontario

Data Inputs for GWL Modelling

293-3:

Table A.1-Descriptive statistics (mean, standard deviation, minimum, 25th percentile, median, 75th percentile, and maximum) of the key input variables and GWL for well 293-3 over the study period (2006–2018).

	Temperature (°C)	Precipitation (mm)	Hydrometric Level (m)	GWL
count	150	150	150	150
mean	6.85	78.72	2.27	308
std	10.13	55.66	0.12	0.39
min	-14.89	0	2.11	306.86
25%	-1.64	40.28	2.18	307.7
50%	6.7	75.9	2.24	308
75%	16.07	111.28	2.33	308.26
max	23.07	277.44	2.87	309.08

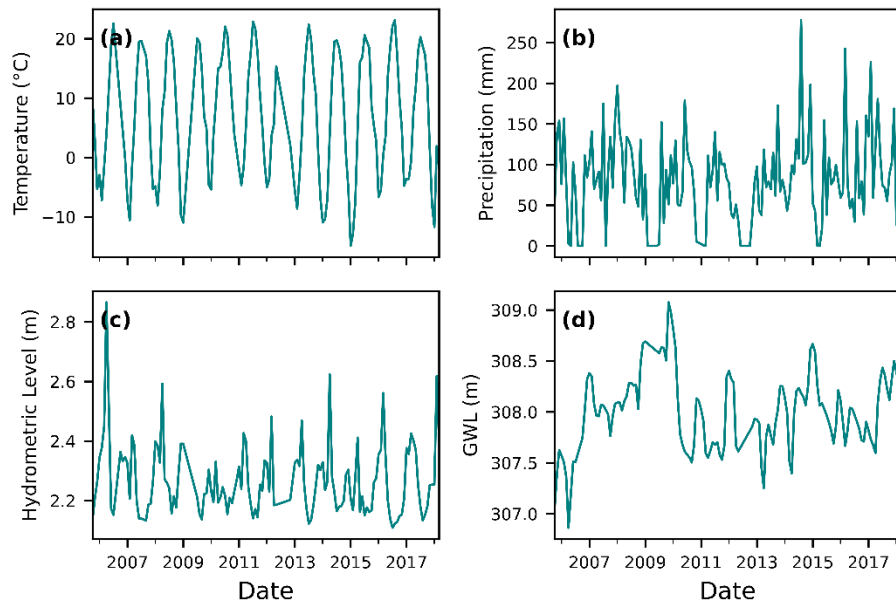


Figure A.2- Time series plots of the monthly (a) temperature, (b) precipitation, (c) hydrometric level, and (d) groundwater level (GWL) at well 293-2, showing seasonal and long-term variations from 2006 to 2017.

439-1:

Table A.2-Descriptive statistics (mean, standard deviation, minimum, 25th percentile, median, 75th percentile, and maximum) of the key input variables and GWL for well 439-1 over the study period (2011–2018).

	Temperature (°C)	Precipitation (mm)	Hydrometric Level (m)	GWL
count	73	73	73	73
mean	7.07	85.75	1.28	308.67
std	10.2	34.69	0.09	0.18
min	-14.94	1.1	1.15	308.28
25%	-2.56	64	1.21	308.56
50%	6.97	80.5	1.27	308.67
75%	17.1	104.2	1.33	308.77
max	22.43	188.2	1.58	309.01

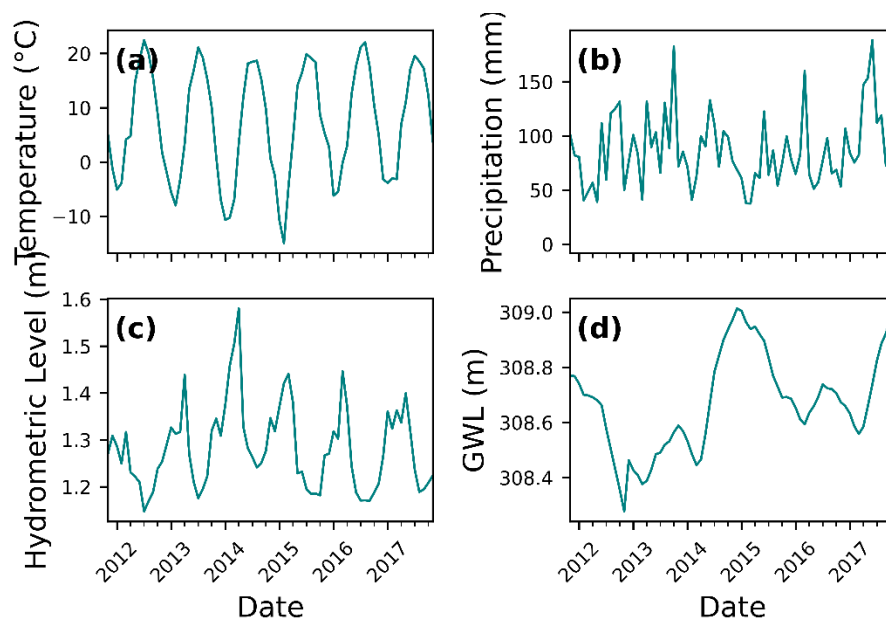


Figure A.3- Time series plots of the monthly (a) temperature, (b) precipitation, (c) hydrometric level, and (d) GWL at well 439-1, showing seasonal and long-term variations from 2011 to 2018.

440-1:

Table 0.3-Descriptive statistics (mean, standard deviation, minimum, 25th percentile, median, 75th percentile, and maximum) of the key input variables and GWL for well 440-1 over the study period (2011–2018).

	Temperature (°C)	Precipitation (mm)	Hydrometric Level (m)	GWL
count	73	73	73	73
mean	7.07	85.75	1.28	219.44
std	10.2	34.69	0.09	0.07
min	-14.94	1.1	1.15	219.31
25%	-2.56	64	1.21	219.39
50%	6.97	80.5	1.27	219.44
75%	17.1	104.2	1.33	219.49
max	22.43	188.2	1.58	219.61

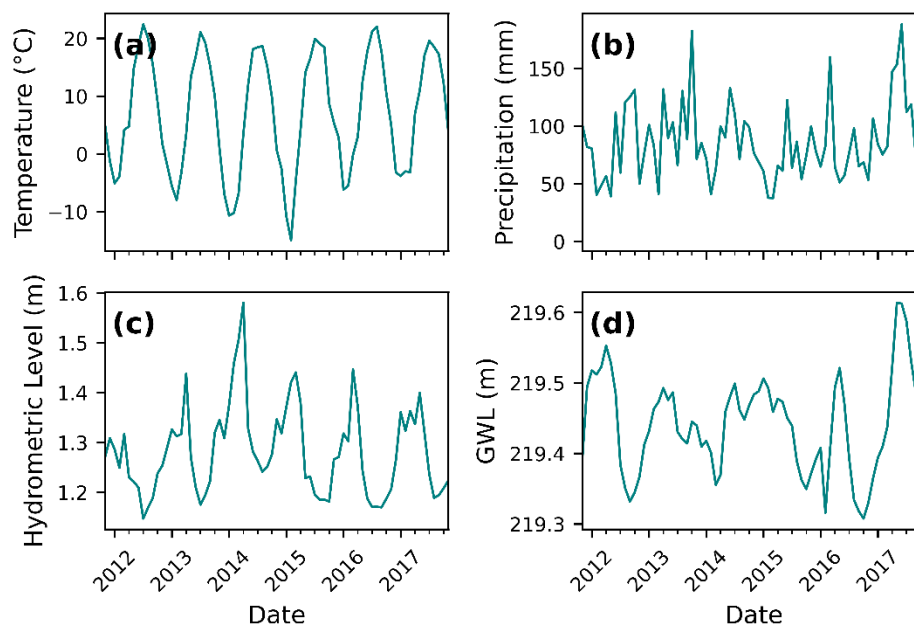


Figure A.4- Time series plots of the monthly (a) temperature, (b) precipitation, (c) hydrometric level, and (d) GWL at well 440-1, showing seasonal and long-term variations from 2011 to 2018.

442-1:

Table 0.4-Descriptive statistics (mean, standard deviation, minimum, 25th percentile, median, 75th percentile, and maximum) of the key input variables and GWL for well 442-1 over the study period (2011–2018).

	Temperature (°C)	Precipitation (mm)	Hydrometric Level (m)	GWL
count	73	73	73	73
mean	7.07	85.75	1.28	248.16
std	10.2	34.69	0.09	0.15
min	-14.94	1.1	1.15	247.91
25%	-2.56	64	1.21	248.06
50%	6.97	80.5	1.27	248.14
75%	17.1	104.2	1.33	248.26
max	22.43	188.2	1.58	248.52

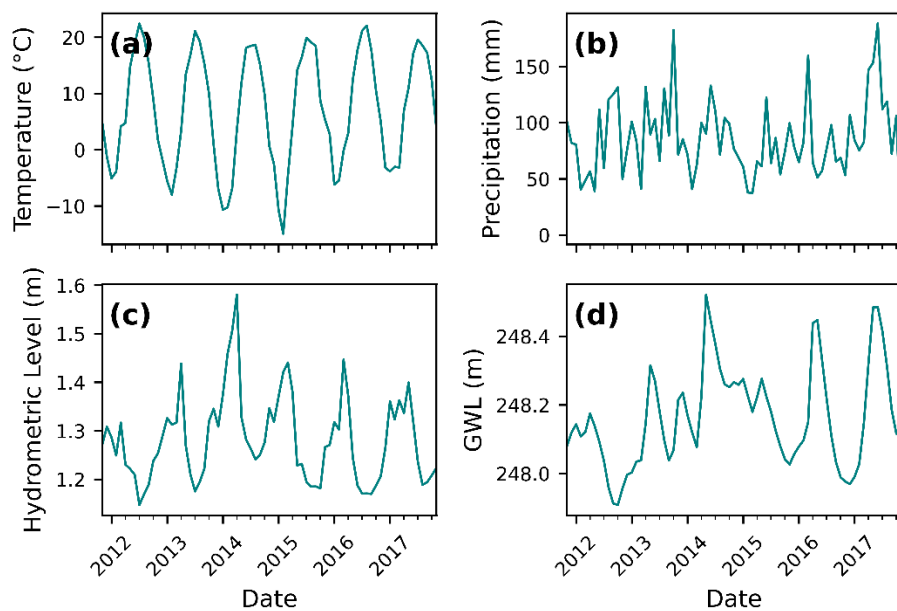


Figure A.5- Time series plots of the monthly (a) temperature, (b) precipitation, (c) hydrometric level, and (d) GWL at well 442-1, showing seasonal and long-term variations from 2011 to 2018.

244-2:

Table 0.5-Descriptive statistics (mean, standard deviation, minimum, 25th percentile, median, 75th percentile, and maximum) of the key input variables and GWL for well 244-2 over the study period (2006–2020).

	Temperature (°C)	Precipitation (mm)	Hydrometric Level (m)	GWL
count	170	170	170	170
mean	7.11	63.51	4.01	230.59
std	9.8	38.92	0.22	0.36
min	-14.66	0	3.67	229.72
25%	-1.27	39.35	3.85	230.39
50%	7.41	63.25	3.97	230.64
75%	16.25	90.79	4.12	230.8
max	22.57	168.8	4.75	231.46

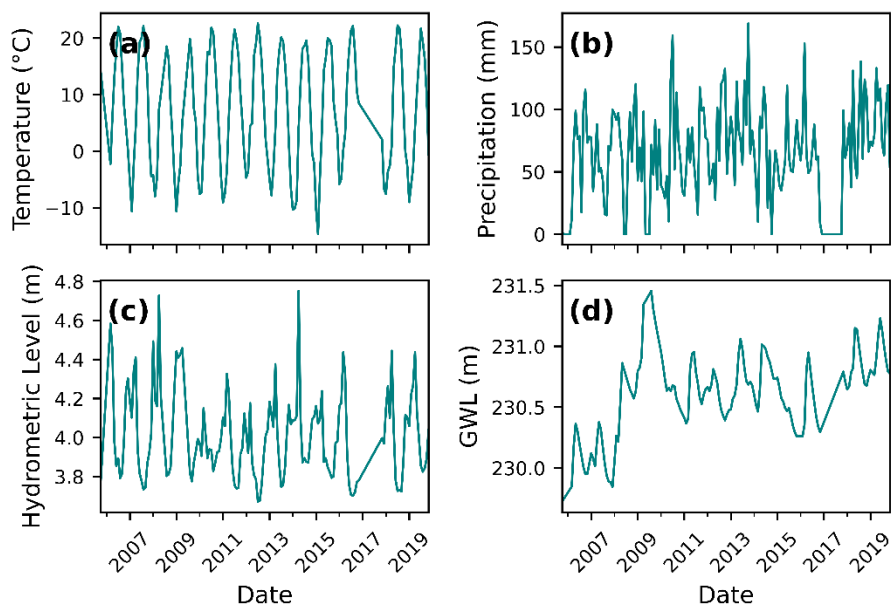


Figure A.6-Time series plots of the monthly (a) temperature, (b) precipitation, (c) hydrometric level, and (d) GWL at well 244-2, showing seasonal and long-term variations from 2006 to 2020.

245-2:

Table 0.6-Descriptive statistics (mean, standard deviation, minimum, 25th percentile, median, 75th percentile, and maximum) of the key input variables and GWL for well 245-2 over the study period (2006–2020).

	Temperature (°C)	Precipitation (mm)	Hydrometric Level (m)	GWL
count	170	170	170	170
mean	7.11	63.51	4.01	227.04
std	9.8	38.92	0.22	0.59
min	-14.66	0	3.67	225.5
25%	-1.27	39.35	3.85	226.67
50%	7.41	63.25	3.97	227.08
75%	16.25	90.79	4.12	227.57
max	22.57	168.8	4.75	227.98

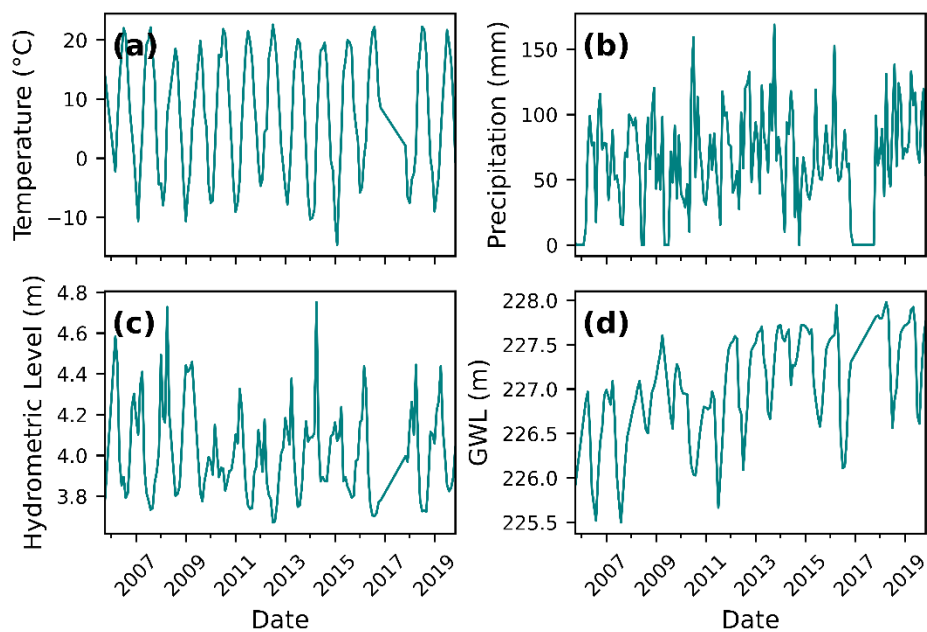


Figure A.7-Time series plots of the monthly (a) temperature, (b) precipitation, (c) hydrometric level, and (d) GWL at well 245-2, showing seasonal and long-term variations from 2006 to 2020.

443-1:

Table A.7-Descriptive statistics (mean, standard deviation, minimum, 25th percentile, median, 75th percentile, and maximum) of the key input variables and GWL for well 443-1 over the study period (2010–2018).

	Temperature (°C)	Precipitation (mm)	Hydrometric Level (m)	GWL
count	101	101	101	101
mean	7.32	81.58	0.54	284.62
std	10.15	33.83	0.08	0.14
min	-14.66	24.7	0.39	284.33
25%	-2.61	58	0.47	284.51
50%	8.32	76.1	0.54	284.62
75%	16.98	101.1	0.6	284.72
max	22.44	178	0.79	284.89

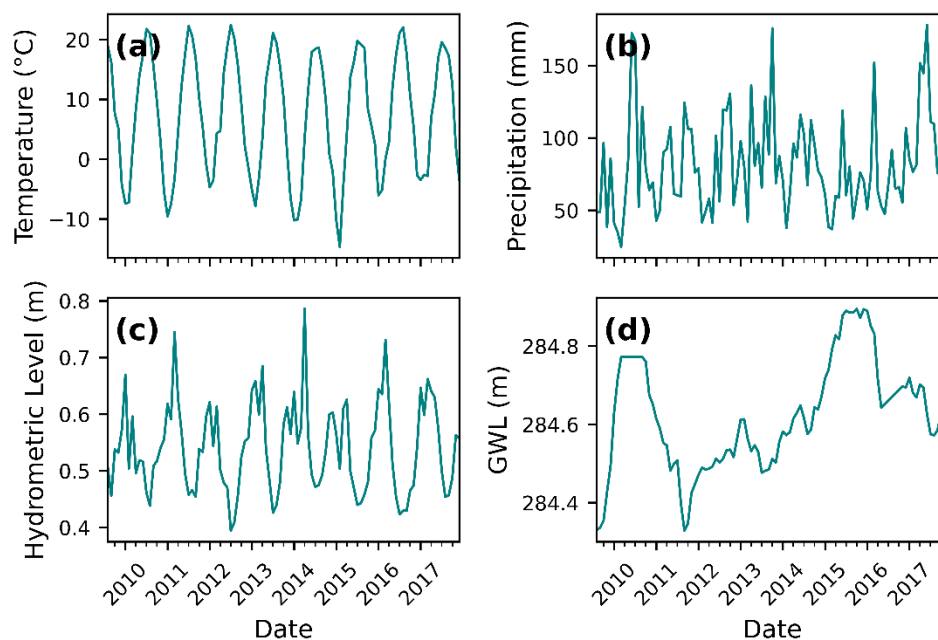


Figure A.8-Time series plots of the monthly (a) temperature, (b) precipitation, (c) hydrometric level, and (d) GWL at well 443-1, showing seasonal and long-term variations from 2010 to 2018.

DMS vs. DSM Inputs

DSM 293-3:

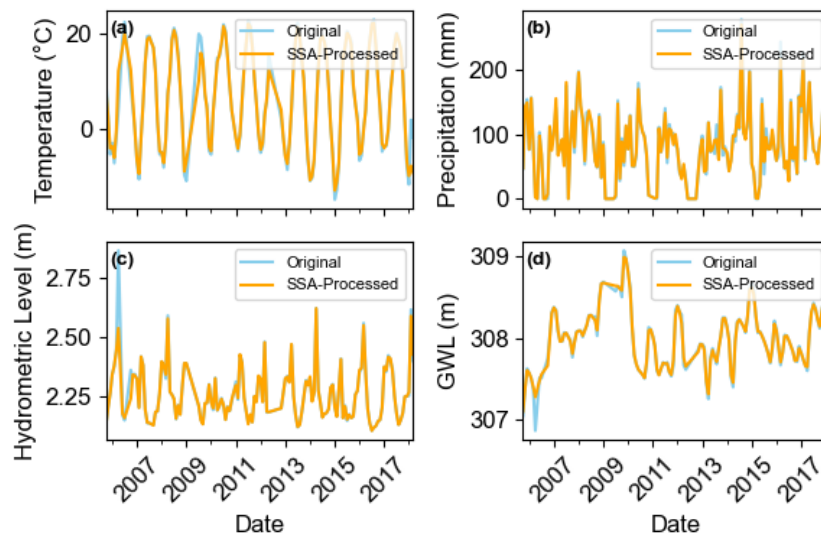


Figure A. 9- Comparison of original monthly data (blue) and SSA-processed data (orange) for well 293-3, where SSA was applied at the daily scale prior to aggregation. The subplots represent (a) air temperature (°C), (b) precipitation (mm), (c) surface water level (m), and (d) GWL (m).

DMS 293-3:

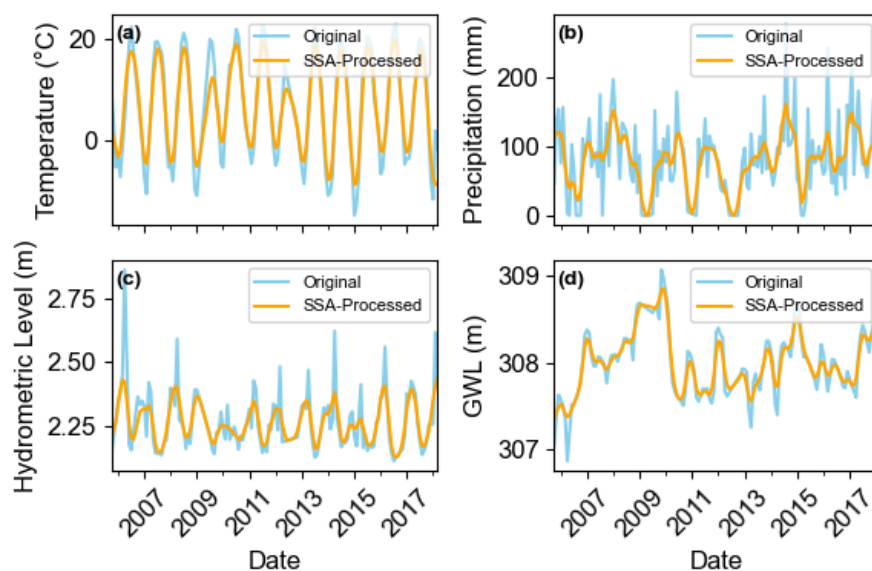


Figure A.10- Comparison of original monthly data (blue) and SSA-processed data (orange) for well 293-3, with SSA applied after aggregation to monthly frequency. The subplots show (a) air temperature (°C), (b) precipitation (mm), (c) surface water level (m), and (d) GWL (m).

DSM 439-1:

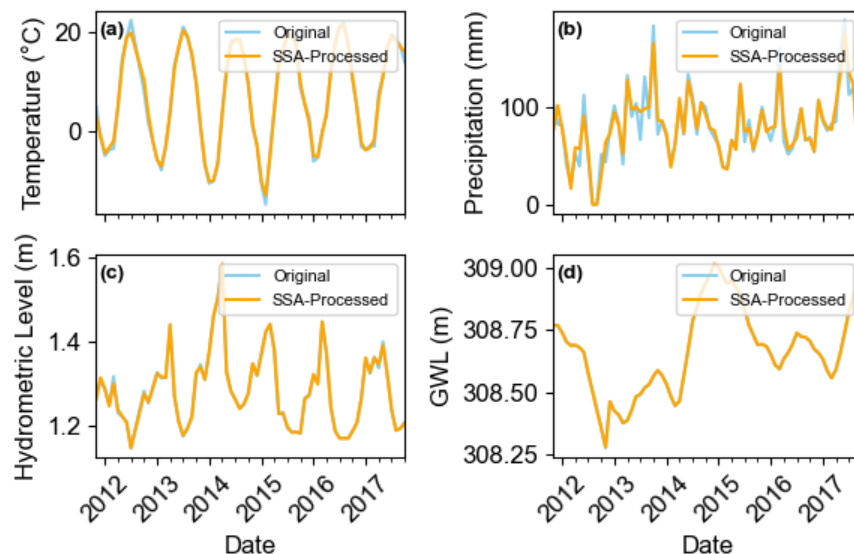


Figure A. 11- Comparison of original monthly data (blue) and SSA-processed data (orange) for well 439-1, where SSA was applied at the daily scale prior to aggregation. The subplots represent (a) air temperature (°C), (b) precipitation (mm), (c) surface water level (m), and (d) GWL (m).

DMS 439-1:

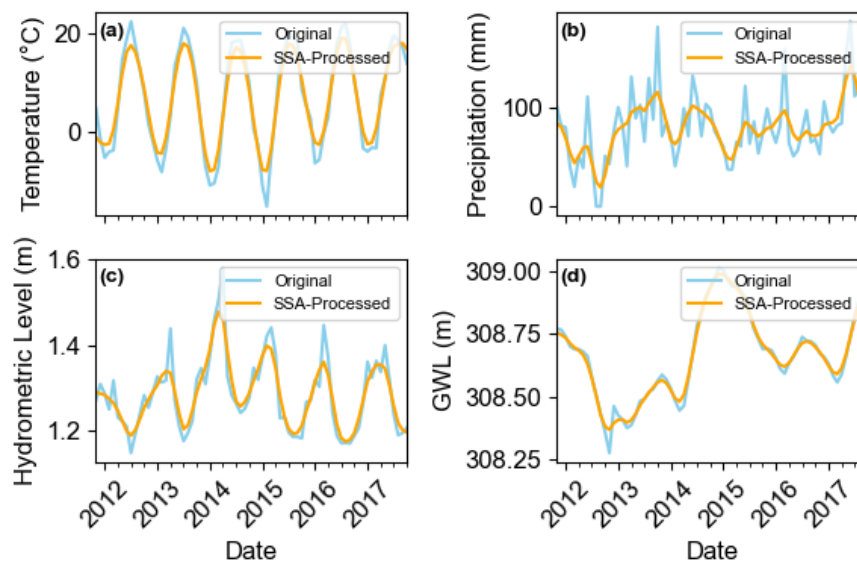


Figure A.12- Comparison of original monthly data (blue) and SSA-processed data (orange) for well 439-1, with SSA applied after aggregation to monthly frequency. The subplots show (a) air temperature (°C), (b) precipitation (mm), (c) surface water level (m), and (d) GWL (m).

DSM 440-1:

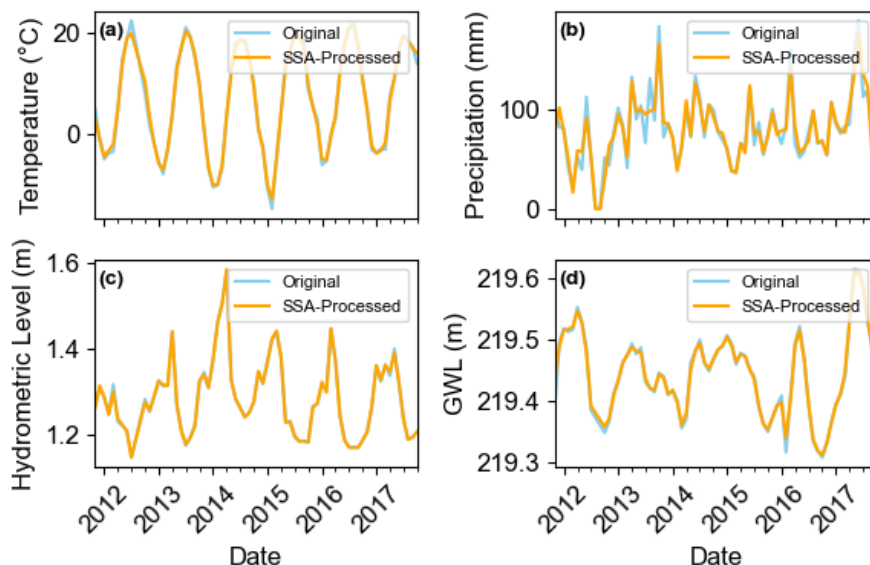


Figure 0.13- Comparison of original monthly data (blue) and SSA-processed data (orange) for well 440-1, where SSA was applied at the daily scale prior to aggregation. The subplots represent (a) air temperature (°C), (b) precipitation (mm), (c) surface water level (m), and (d) GWL (m).

DMS 440-1:

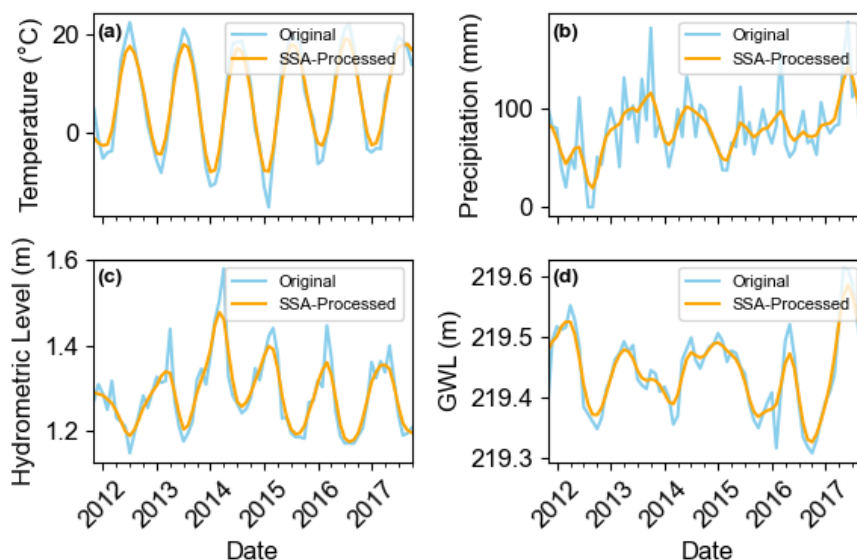


Figure 0.14- Comparison of original monthly data (blue) and SSA-processed data (orange) for well 440-1, with SSA applied after aggregation to monthly frequency. The subplots show (a) air temperature (°C), (b) precipitation (mm), (c) surface water level (m), and (d) GWL (m).

DSM 442-1:

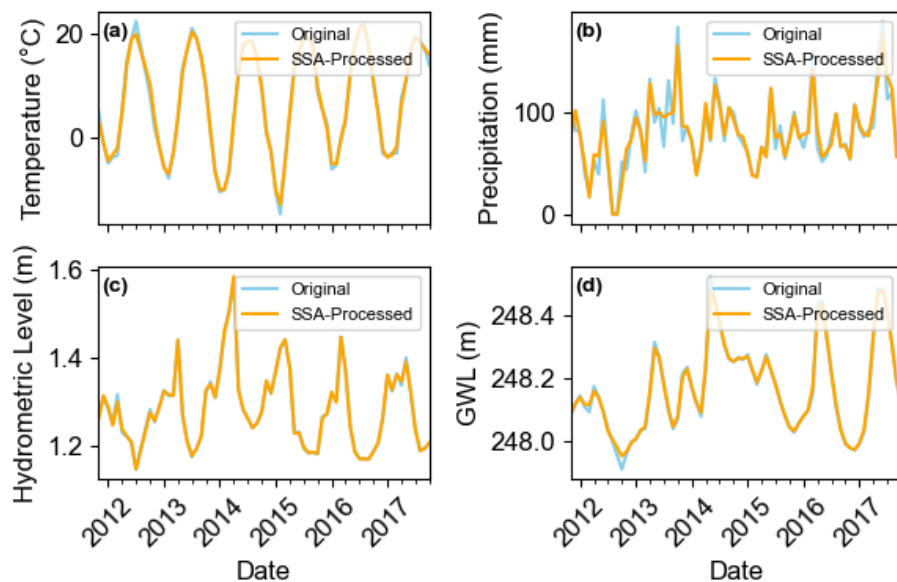


Figure 0.15- Comparison of original monthly data (blue) and SSA-processed data (orange) for well 442-1, where SSA was applied at the daily scale prior to aggregation. The subplots represent (a) air temperature (°C), (b) precipitation (mm), (c) surface water level (m), and (d) GWL (m).

DMS 442-1:

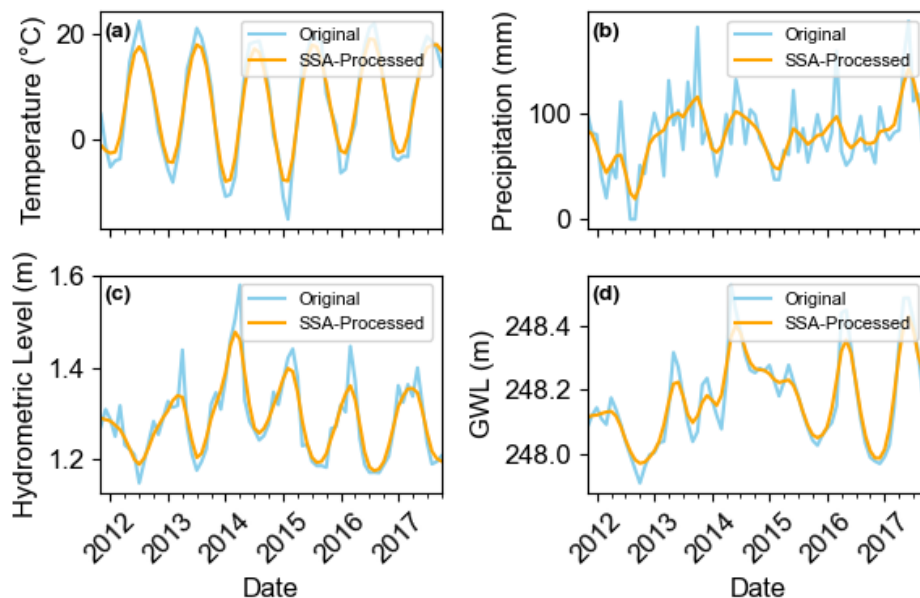


Figure 0.16- Comparison of original monthly data (blue) and SSA-processed data (orange) for well 442-1, with SSA applied after aggregation to monthly frequency. The subplots show (a) air temperature (°C), (b) precipitation (mm), (c) surface water level (m), and (d) GWL (m).

DSM 244-2:

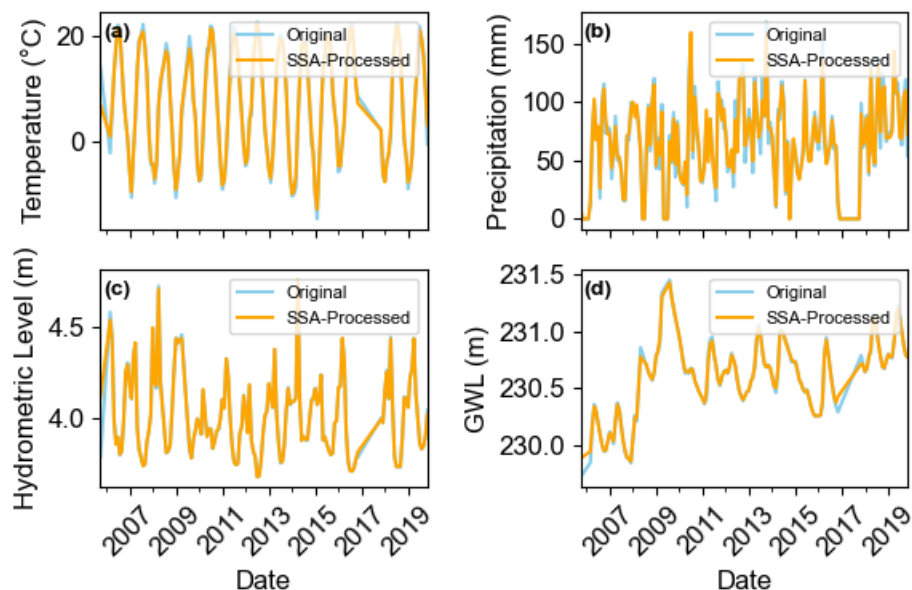


Figure 0.17- Comparison of original monthly data (blue) and SSA-processed data (orange) for well 244-2, where SSA was applied at the daily scale prior to aggregation. The subplots represent (a) air temperature (°C), (b) precipitation (mm), (c) surface water level (m), and (d) GWL (m).

DMS 244-2:

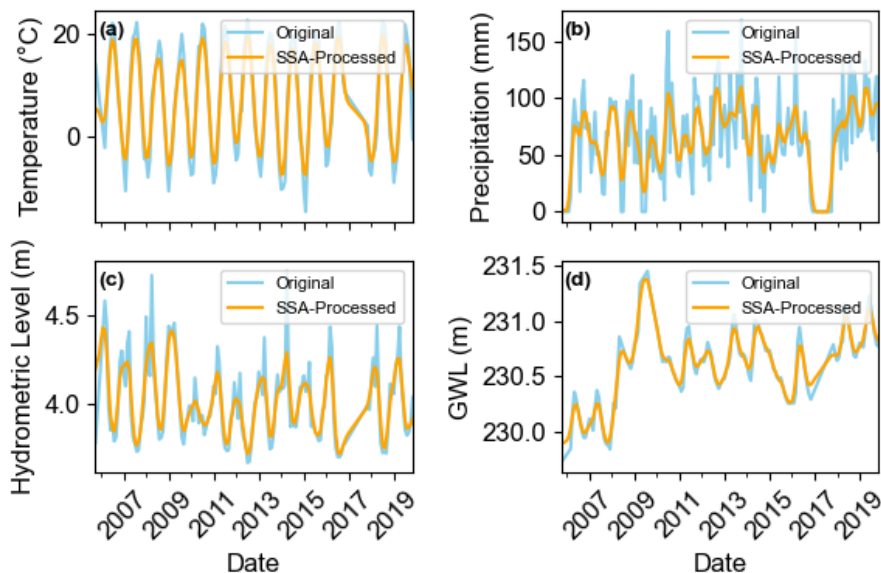


Figure 0.18- Comparison of original monthly data (blue) and SSA-processed data (orange) for well 244-2, with SSA applied after aggregation to monthly frequency. The subplots show (a) air temperature (°C), (b) precipitation (mm), (c) surface water level (m), and (d) GWL (m).

DSM 245-2:

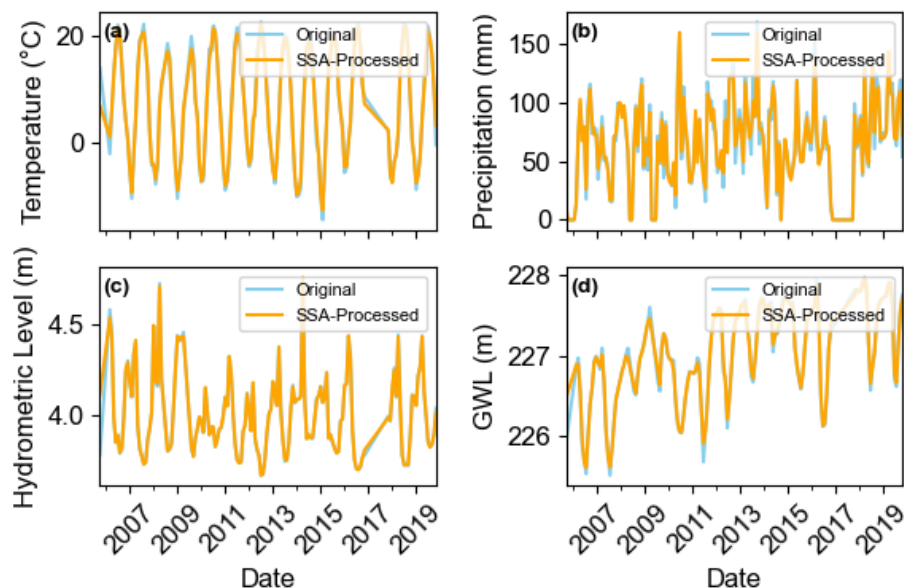


Figure A.19- Comparison of original monthly data (blue) and SSA-processed data (orange) for well 245-2, where SSA was applied at the daily scale prior to aggregation. The subplots represent (a) air temperature ($^{\circ}\text{C}$), (b) precipitation (mm), (c) surface water level (m), and (d) GWL (m).

DMS 245-2:

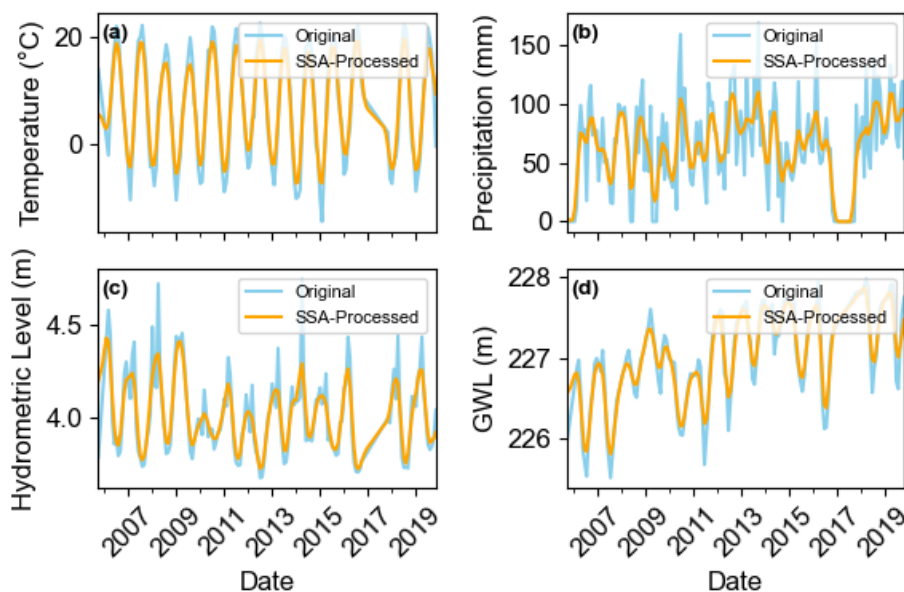


Figure A.20- Comparison of original monthly data (blue) and SSA-processed data (orange) for well 245-2, with SSA applied after aggregation to monthly frequency. The subplots show (a) air temperature ($^{\circ}\text{C}$), (b) precipitation (mm), (c) surface water level (m), and (d) GWL (m).

DSM 443-1:

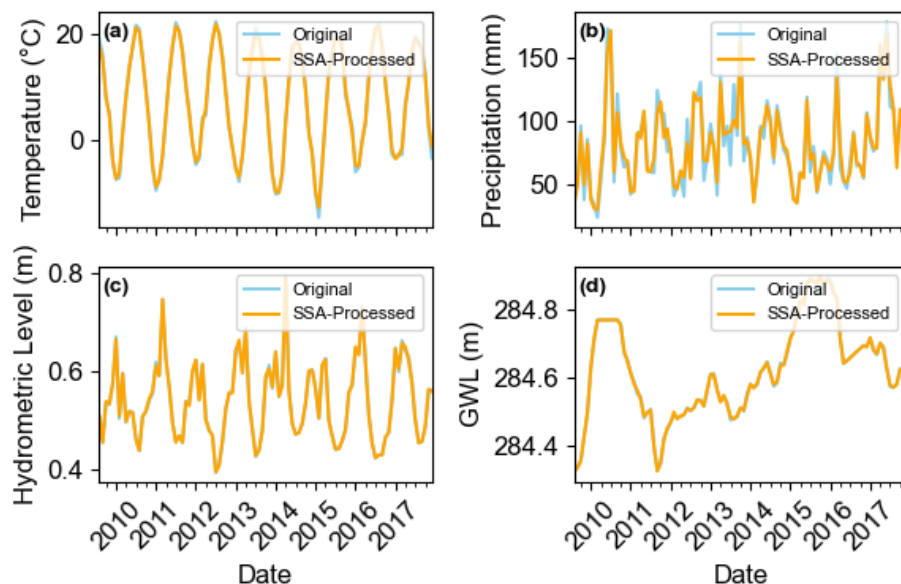


Figure A.21- Comparison of original monthly data (blue) and SSA-processed data (orange) for well 443-1, where SSA was applied at the daily scale prior to aggregation. The subplots represent (a) air temperature ($^{\circ}\text{C}$), (b) precipitation (mm), (c) surface water level (m), and (d) GWL (m).

DMS 443-1:

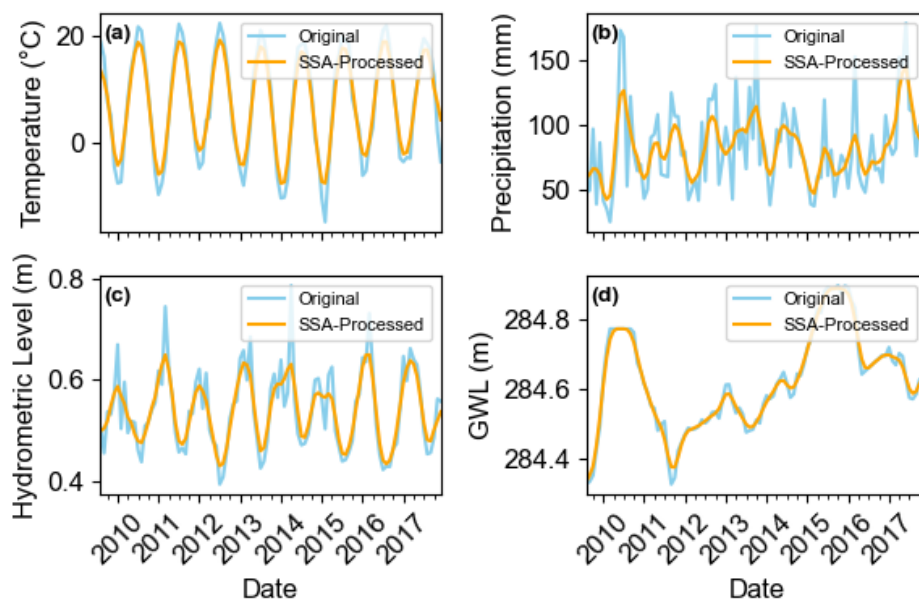


Figure 0.22- Comparison of original monthly data (blue) and SSA-processed data (orange) for well 443-1, with SSA applied after aggregation to monthly frequency. The subplots show (a) air temperature ($^{\circ}\text{C}$), (b) precipitation (mm), (c) surface water level (m), and (d) GWL (m).

Linear Regression Assumptions

Exogeneity

293-3:

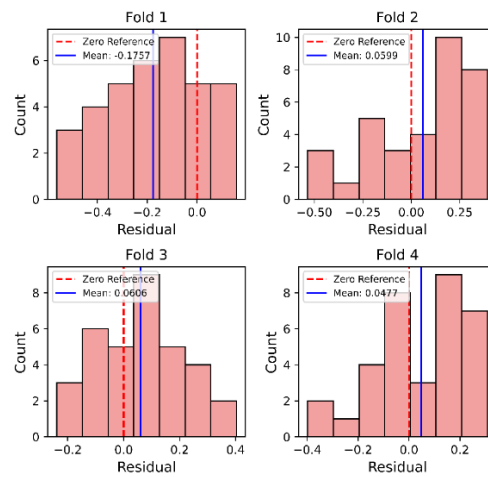


Figure A.23- Histogram plots of test residuals for well 293-3 across four cross-validation folds. The red dashed line in each subplot indicates the zero reference (ideal unbiased residuals), while the solid blue line shows the fold-specific mean residual.

439-1:

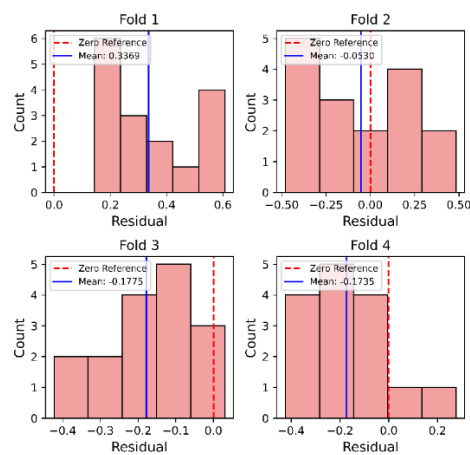


Figure A.24- Histogram plots of test residuals for well 439-1 across four cross-validation folds. The red dashed line in each subplot indicates the zero reference (ideal unbiased residuals), while the solid blue line shows the fold-specific mean residual.

440-1:

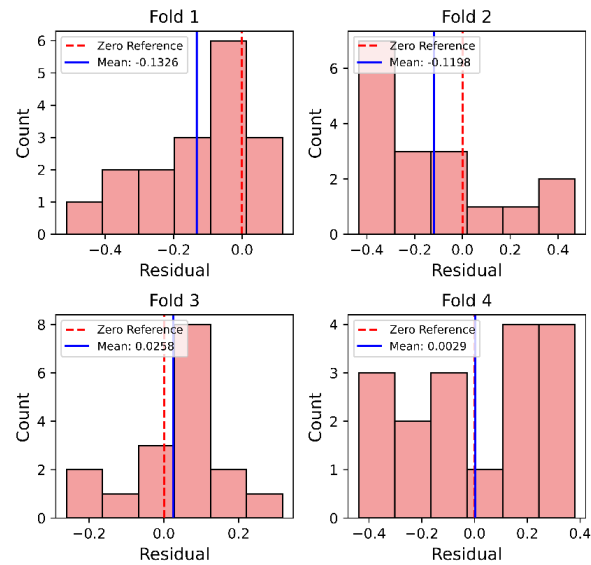


Figure A.25- Histogram plots of test residuals for well 440-1 across four cross-validation folds. The red dashed line in each subplot indicates the zero reference (ideal unbiased residuals), while the solid blue line shows the fold-specific mean residual.

442-1:

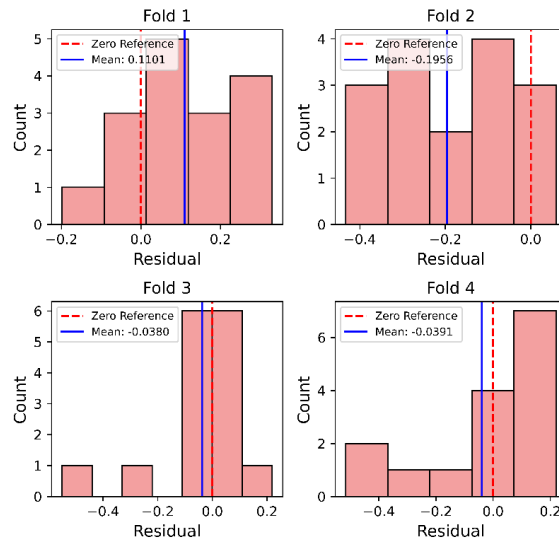


Figure A.26- Histogram plots of test residuals for well 442-1 across four cross-validation folds. The red dashed line in each subplot indicates the zero reference (ideal unbiased residuals), while the solid blue line shows the fold-specific mean residual.

244-2:

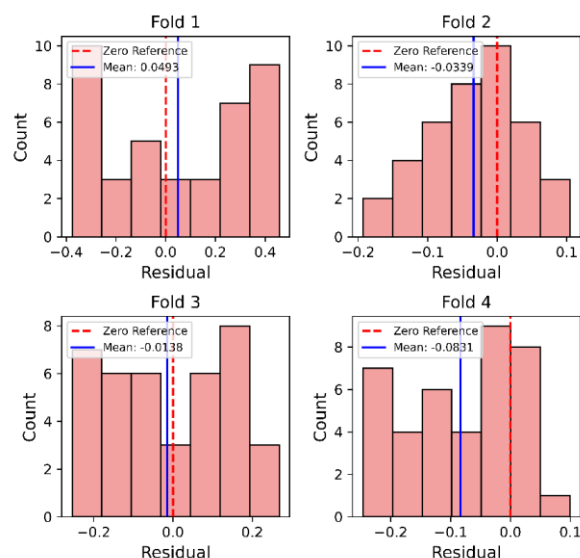


Figure A.27- Histogram plots of test residuals for well 244-2 across four cross-validation folds. The red dashed line in each subplot indicates the zero reference (ideal unbiased residuals), while the solid blue line shows the fold-specific mean residual.

245-2:

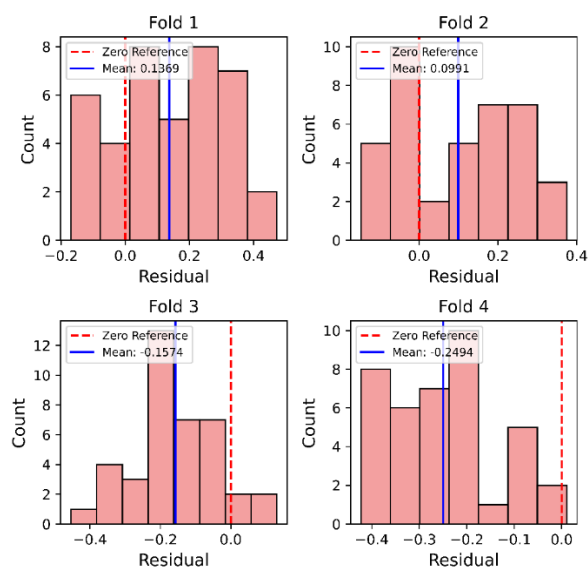


Figure A.28- Histogram plots of test residuals for well 245-2 across four cross-validation folds. The red dashed line in each subplot indicates the zero reference (ideal unbiased residuals), while the solid blue line shows the fold-specific mean residual.

443-1:

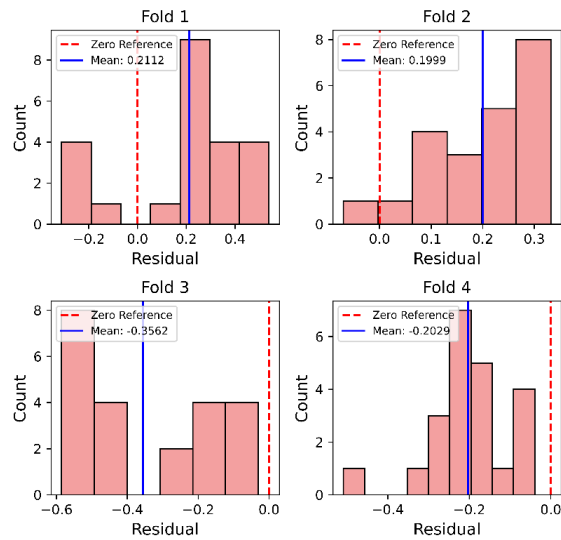


Figure A.29- Histogram plots of test residuals for well 443-1 across four cross-validation folds. The red dashed line in each subplot indicates the zero reference (ideal unbiased residuals), while the solid blue line shows the fold-specific mean residual.

Residual Distribution

293-3

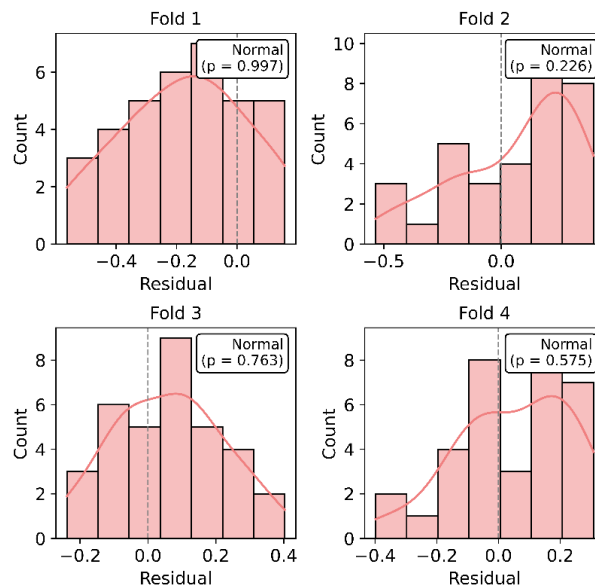


Figure A.30- Histograms of test residuals for Well 293-3 across four cross-validation folds (a-d), overlaid with Kernel Density Estimate (KDE) curves. Each subplot reports the p-value from the Kolmogorov-Smirnov (K-S) test for normality.

439-1:

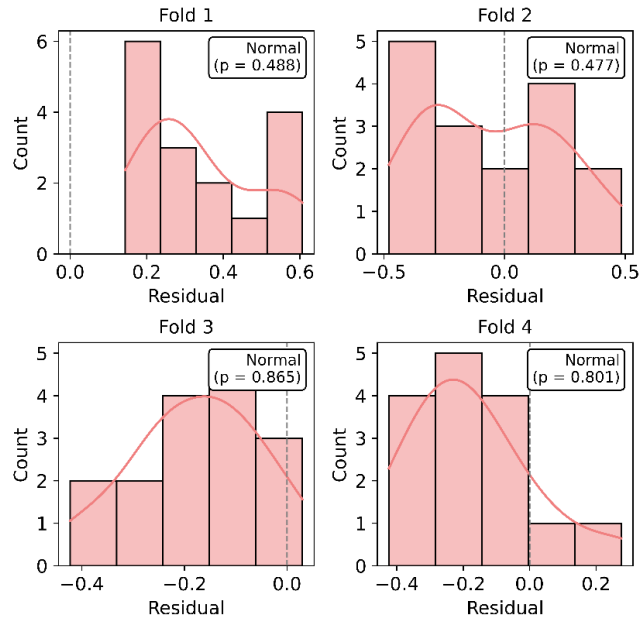


Figure A.31- Histograms of test residuals for Well 439-1 across four cross-validation folds (a-d), overlaid with Kernel Density Estimate (KDE) curves. Each subplot reports the p-value from the Kolmogorov-Smirnov (K-S) test for normality.

440-1:

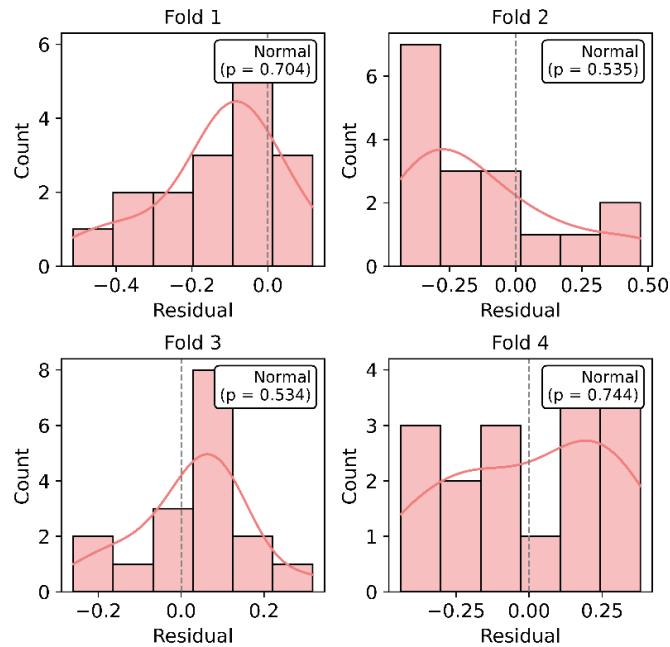


Figure A.32- Histograms of test residuals for Well 440-1 across four cross-validation folds (a-d), overlaid with Kernel Density Estimate (KDE) curves. Each subplot reports the p-value from the Kolmogorov-Smirnov (K-S) test for normality.

442-1:

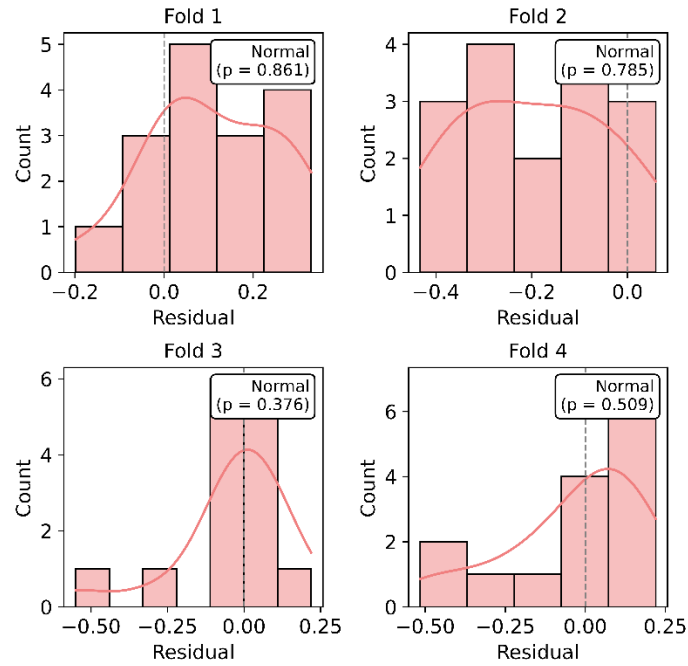


Figure A.33- Histograms of test residuals for Well 442-1 across four cross-validation folds (a–d), overlaid with Kernel Density Estimate (KDE) curves. Each subplot reports the p -value from the Kolmogorov–Smirnov (K–S) test for normality.

244-2:

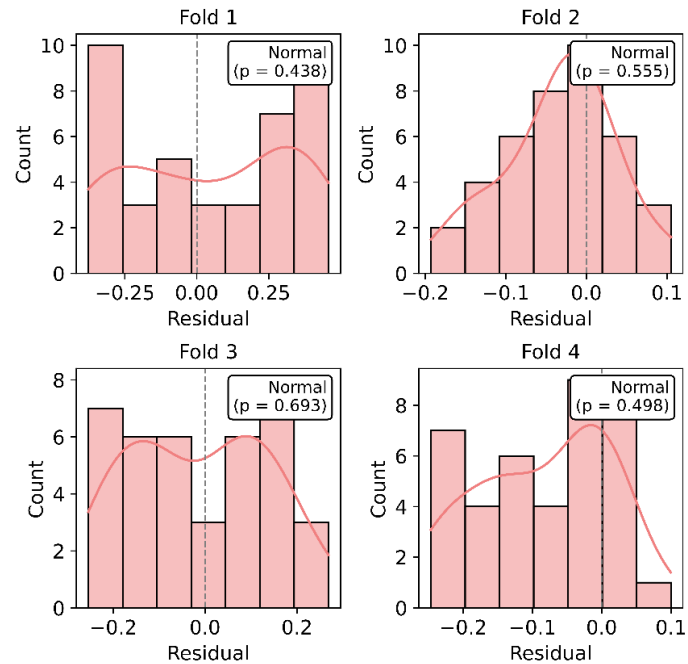


Figure A.34- Histograms of test residuals for Well 244-2 across four cross-validation folds (a–d), overlaid with Kernel Density Estimate (KDE) curves. Each subplot reports the p -value from the Kolmogorov–Smirnov (K–S) test for normality.

245-2:

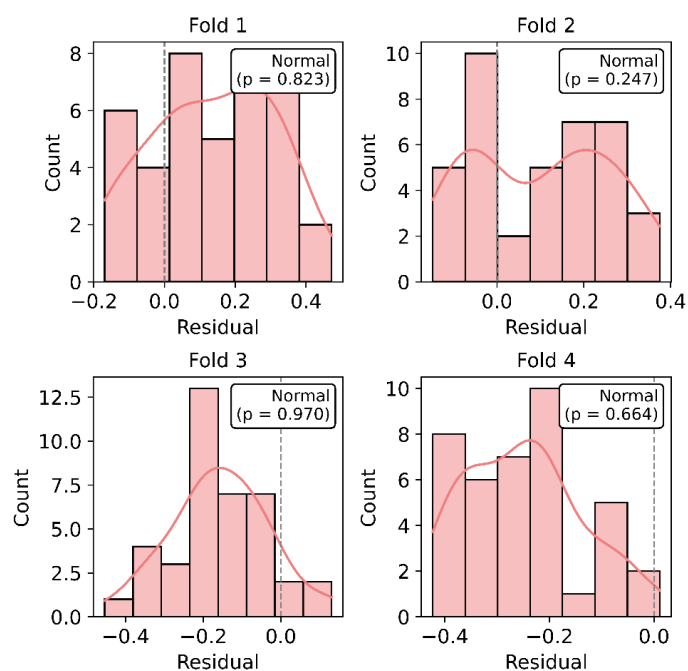


Figure A.35- Histograms of test residuals for Well 245-2 across four cross-validation folds (a-d), overlaid with Kernel Density Estimate (KDE) curves. Each subplot reports the p-value from the Kolmogorov-Smirnov (K-S) test for normality.

443-1:

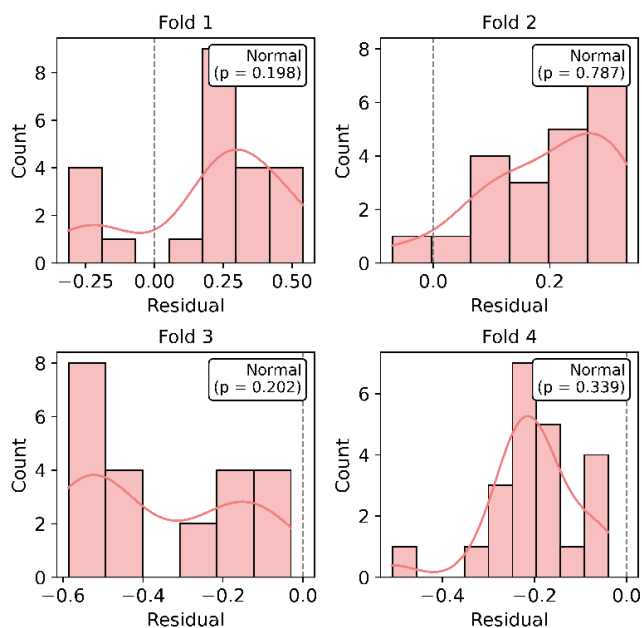


Figure A.36- Histograms of test residuals for Well 443-1 across four cross-validation folds (a-d), overlaid with Kernel Density Estimate (KDE) curves. Each subplot reports the p-value from the Kolmogorov-Smirnov (K-S) test for normality.

Homoscedasticity

293-3:

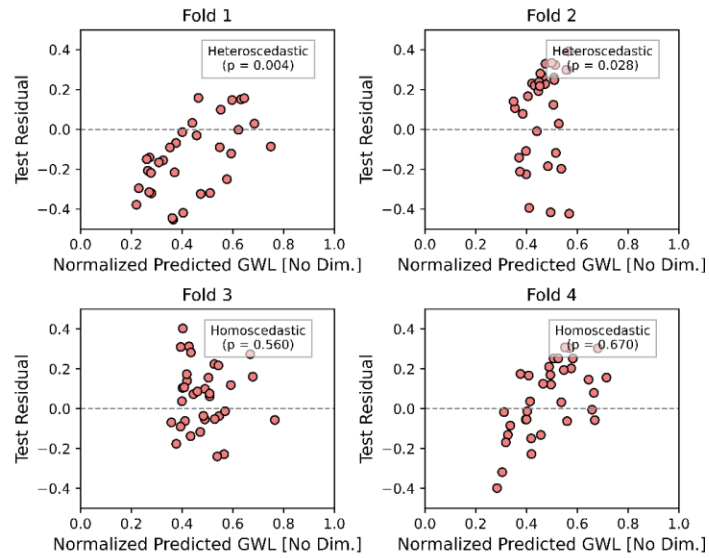


Figure A.37- Residuals versus normalized predicted GWL values across four cross-validation folds for well 293-3.

439-1:

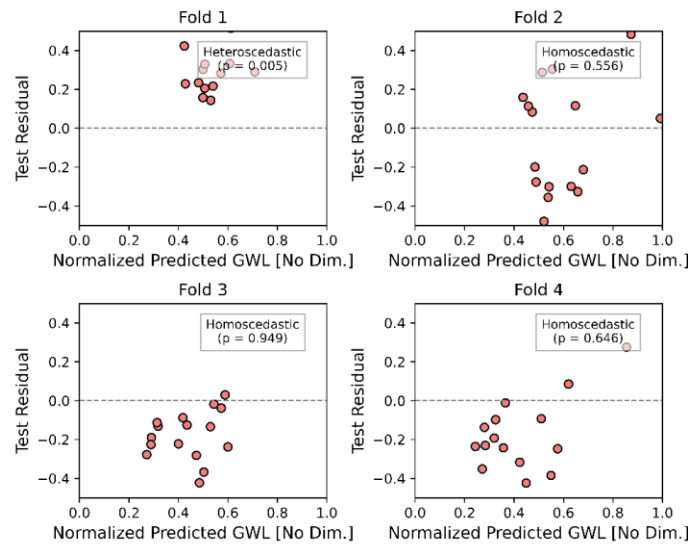


Figure A.38- Residuals versus normalized predicted GWL values across four cross-validation folds for well 439-1.

440-1:

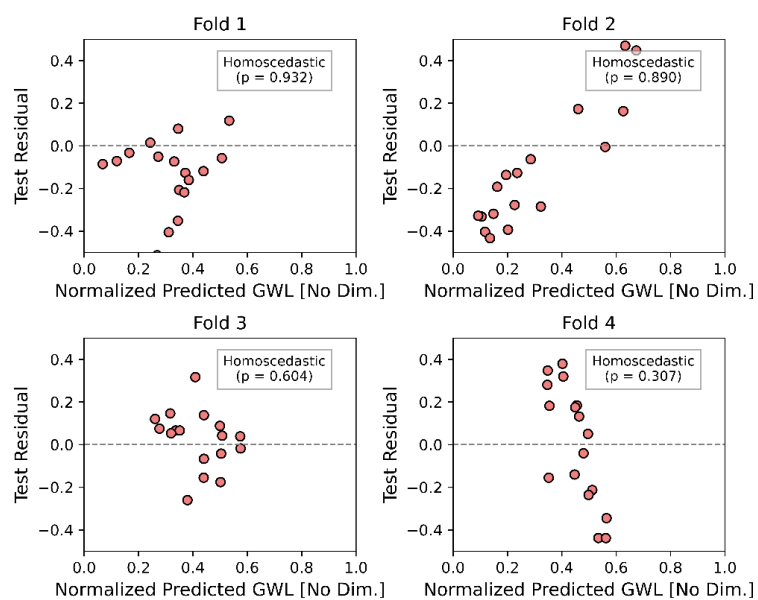


Figure A.39- Residuals versus normalized predicted GWL values across four cross-validation folds for well 440-1.

442-1:

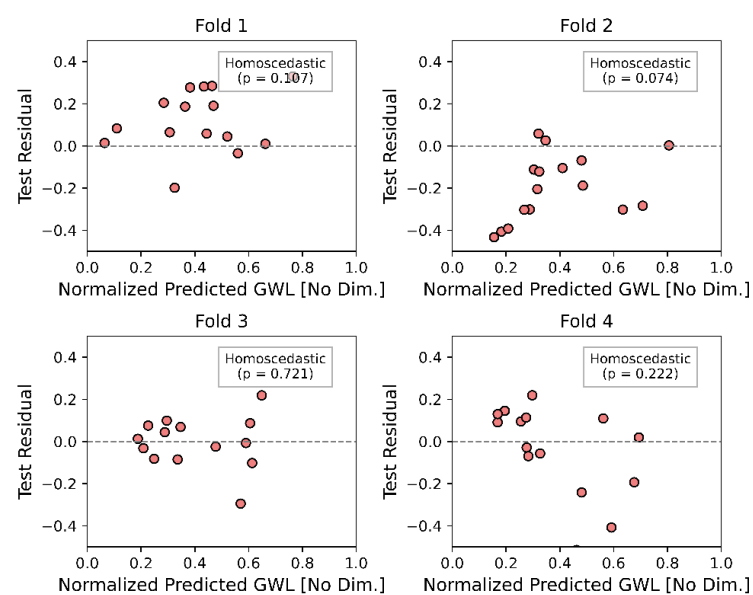


Figure A.40- Residuals versus normalized predicted GWL values across four cross-validation folds for well 442-1.

244-2:

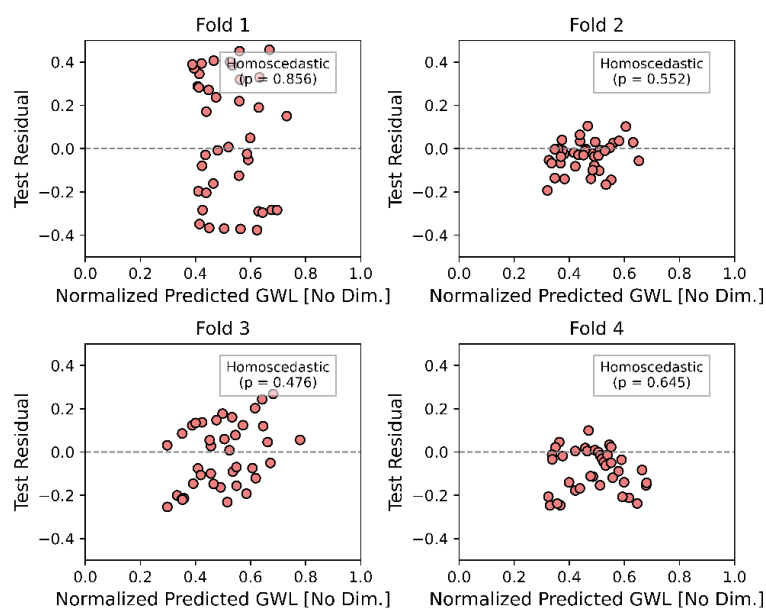


Figure A.41- Residuals versus normalized predicted GWL values across four cross-validation folds for well 244-2.

245-2:

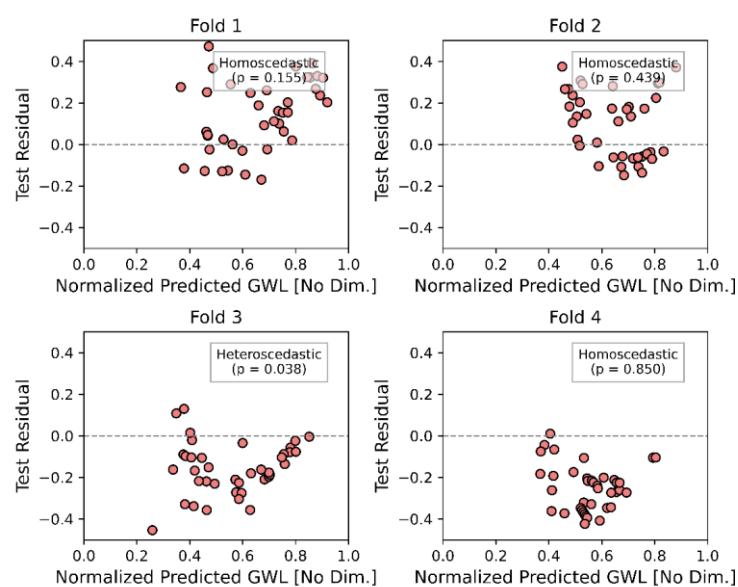


Figure A.42- Residuals versus normalized predicted GWL values across four cross-validation folds for well 245-2.

443-1:

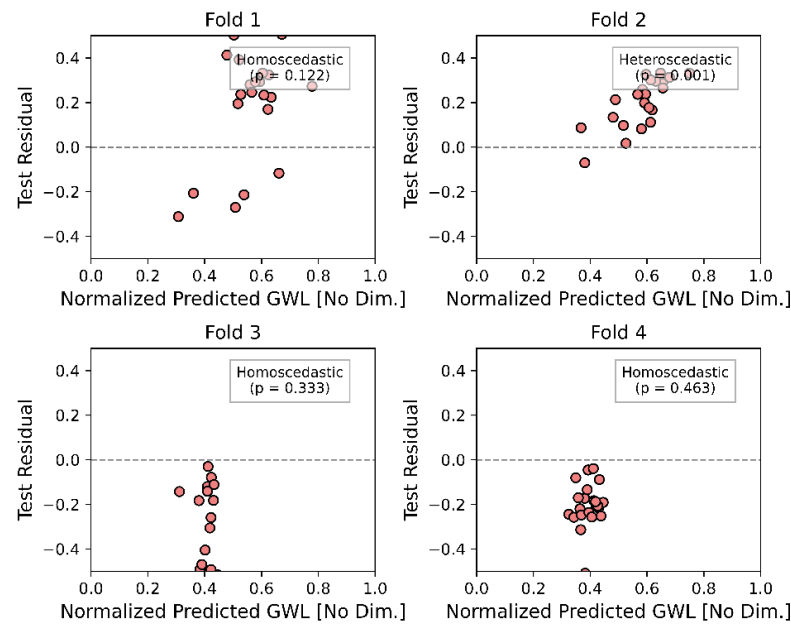


Figure A.43- Residuals versus normalized predicted GWL values across four cross-validation folds for well 443-1.

Autocorrelation in Residuals

293-3

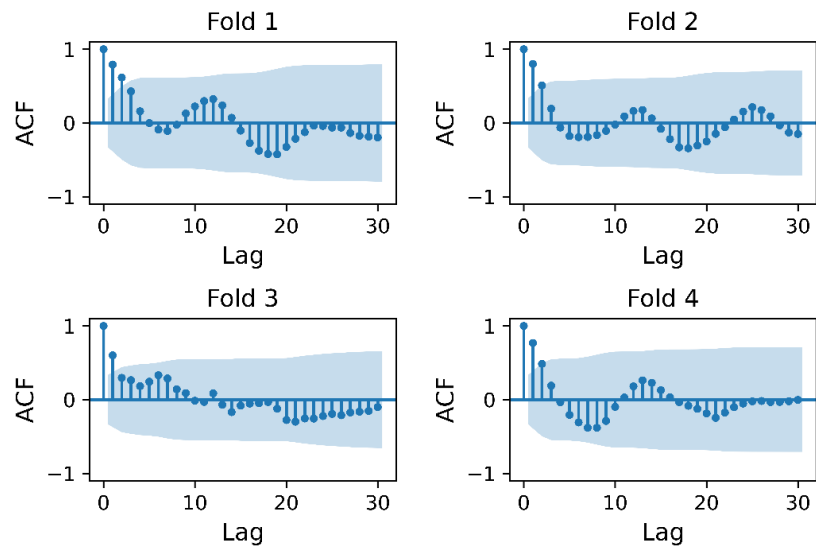


Figure A.44- Autocorrelation function (ACF) plots of residuals for each cross-validation fold for well 293-3.

439-1:

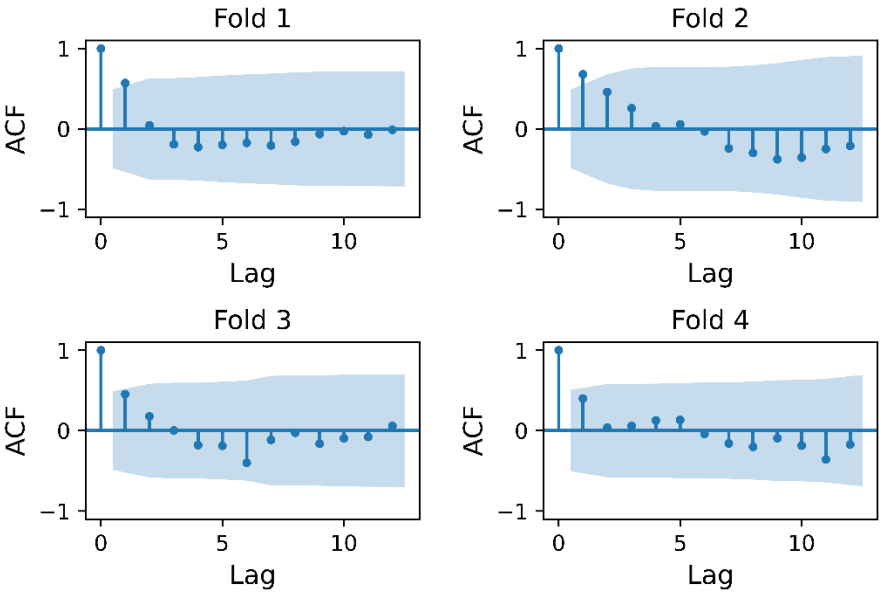


Figure A.45- Autocorrelation function (ACF) plots of residuals for each cross-validation fold for well 439-1.

440-1:

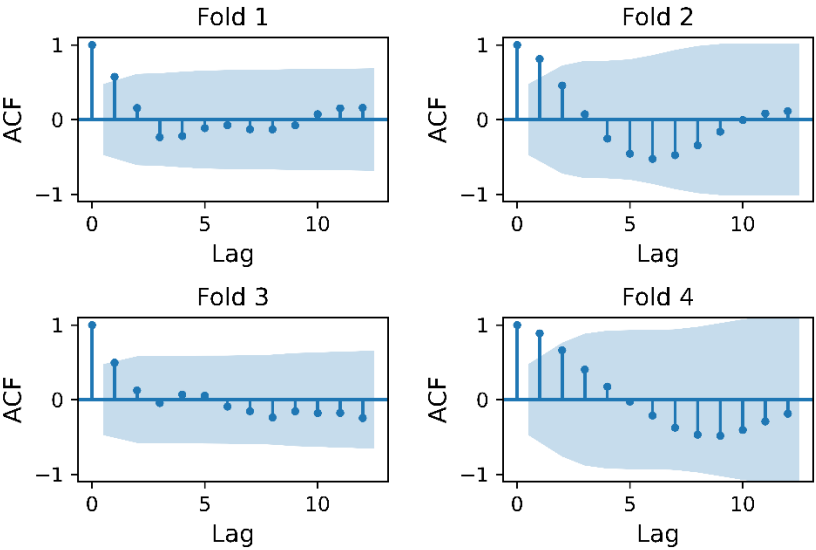


Figure A.46- Autocorrelation function (ACF) plots of residuals for each cross-validation fold for well 440-1.

442-1:

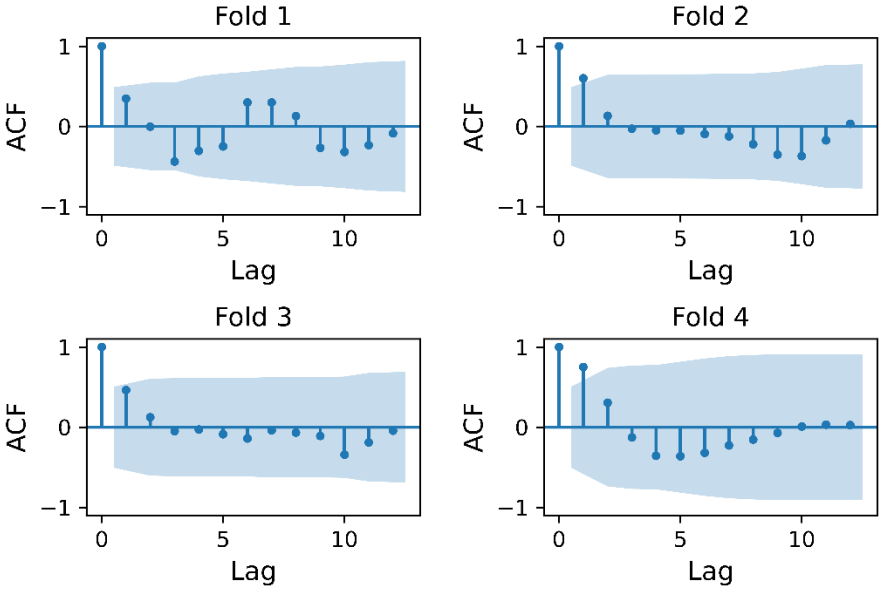


Figure A.47- Autocorrelation function (ACF) plots of residuals for each cross-validation fold for well 442-1.

244-2:

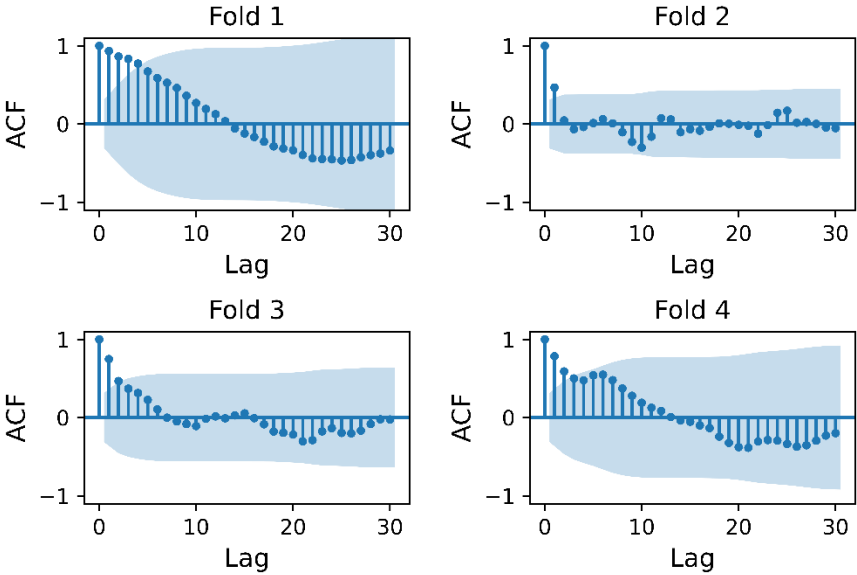


Figure A.48- Autocorrelation function (ACF) plots of residuals for each cross-validation fold for well 244-2.

245-2:

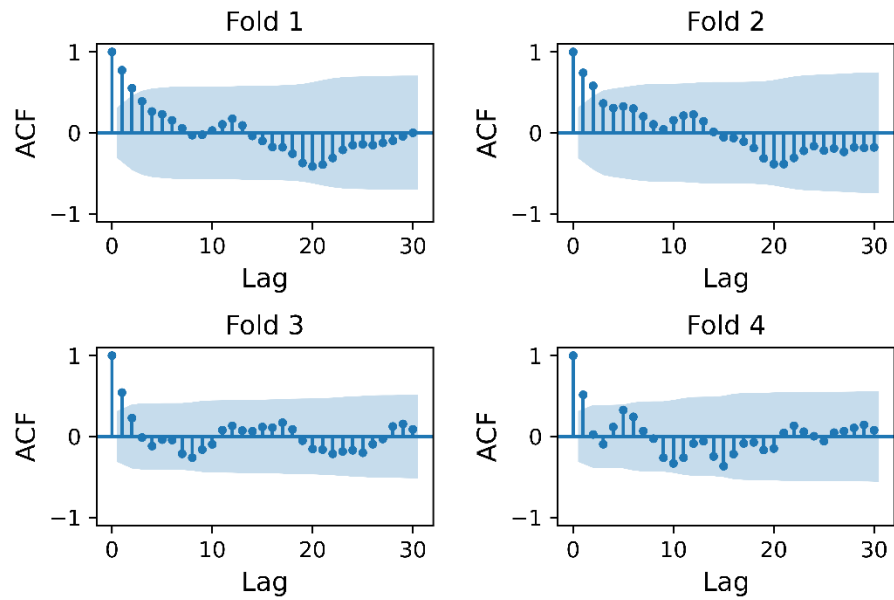


Figure A.49- Autocorrelation function (ACF) plots of residuals for each cross-validation fold for well 245-2.

443-1:

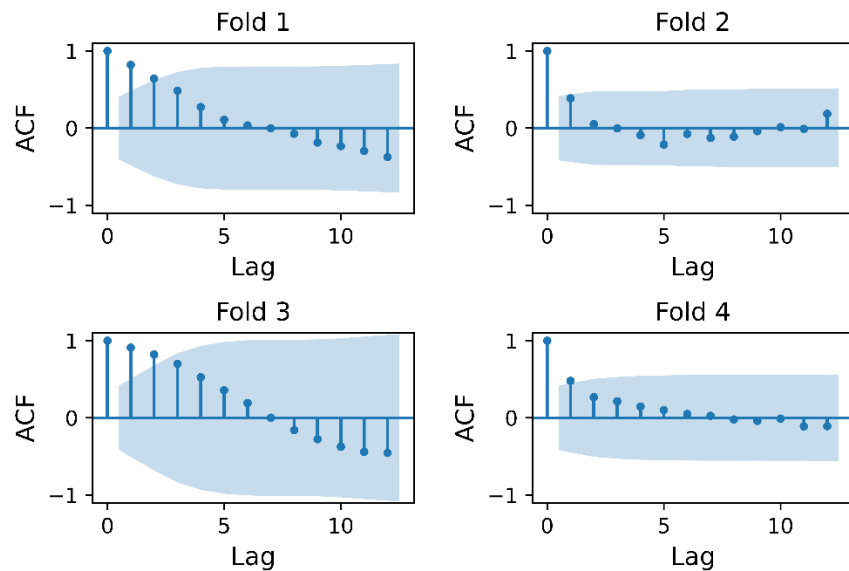
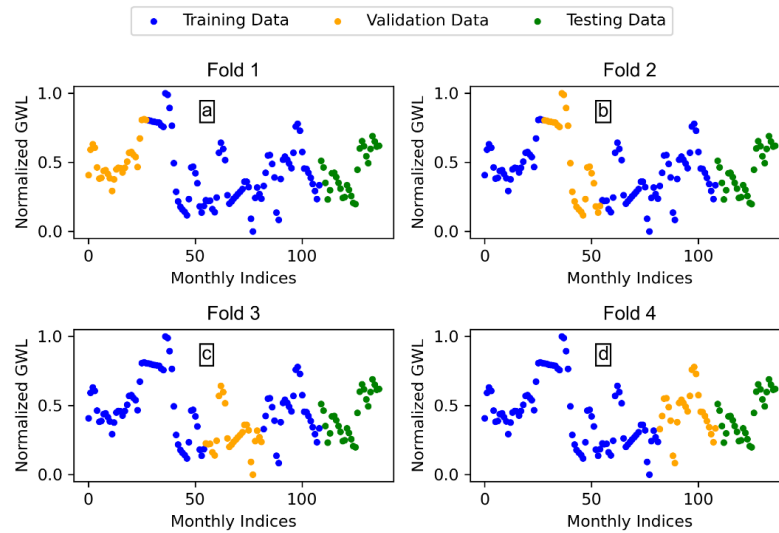


Figure A.50- Autocorrelation function (ACF) plots of residuals for each cross-validation fold for well 443-1.

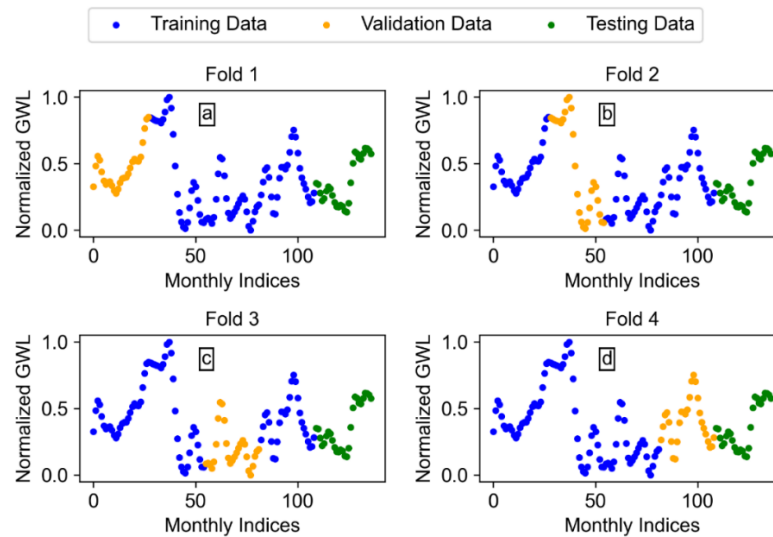
Appendix B

K-Fold Cross-Validation for All Well Datasets

DSM-FFNN 293-3:



DMS-FFNN 293-3



DSM-FFNN 439-1

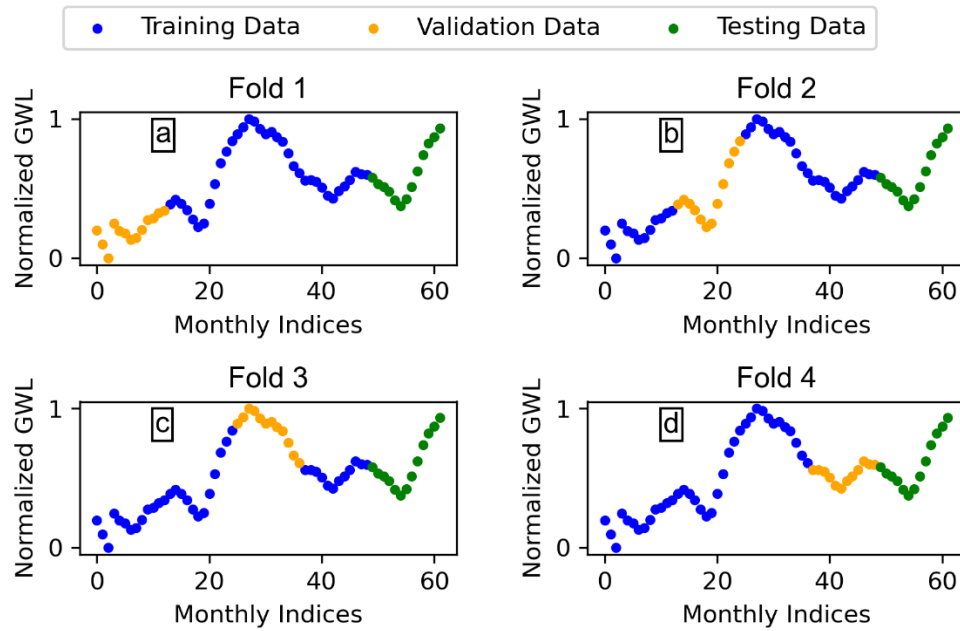


Figure B.3- KFCV used to partition the normalized monthly GWL data from well 439-1, DSM scenario.

DMS-FFNN 439-1

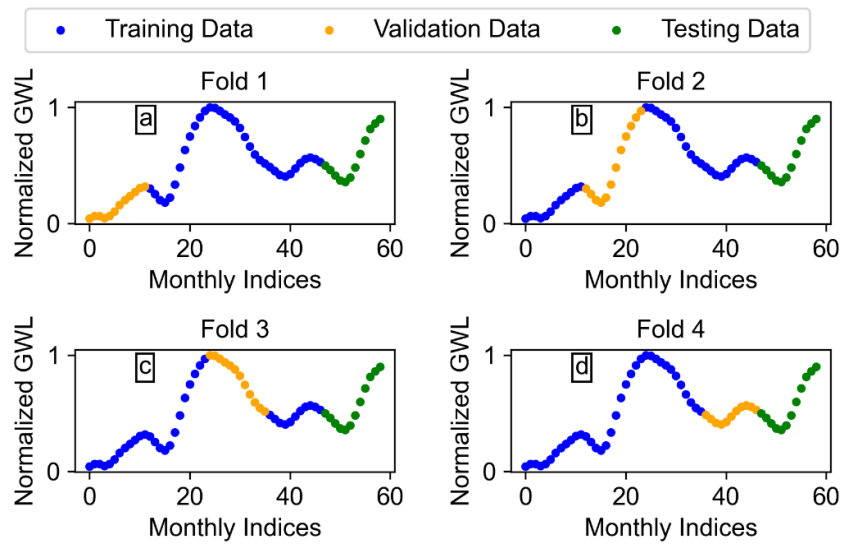


Figure B. 4- KFCV used to partition the normalized monthly GWL data from well 439-1, DMS scenario.

DSM-FFNN 440-1:

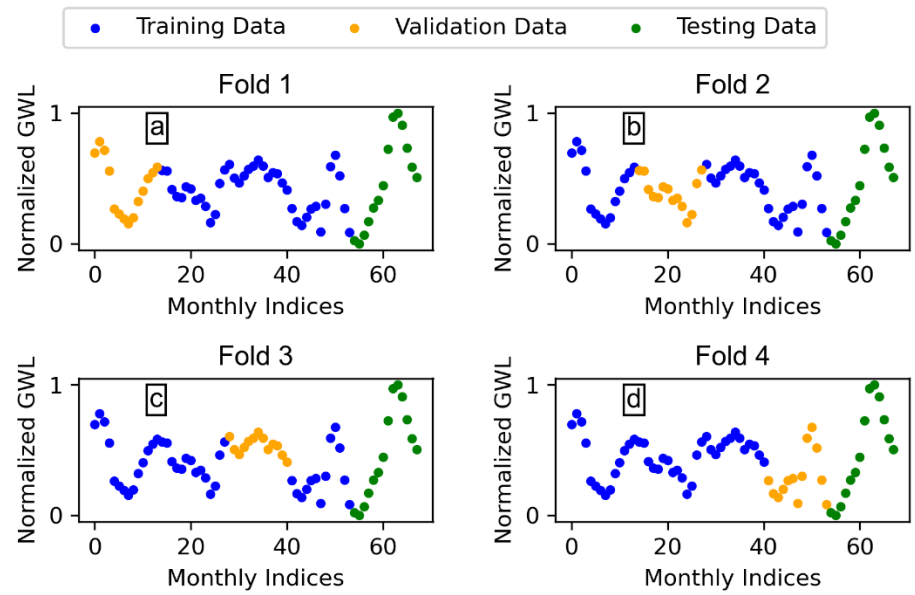


Figure B.5- KFCV used to partition the normalized monthly GWL data from well 440-1, DSM scenario.

DMS-FFNN 440-1

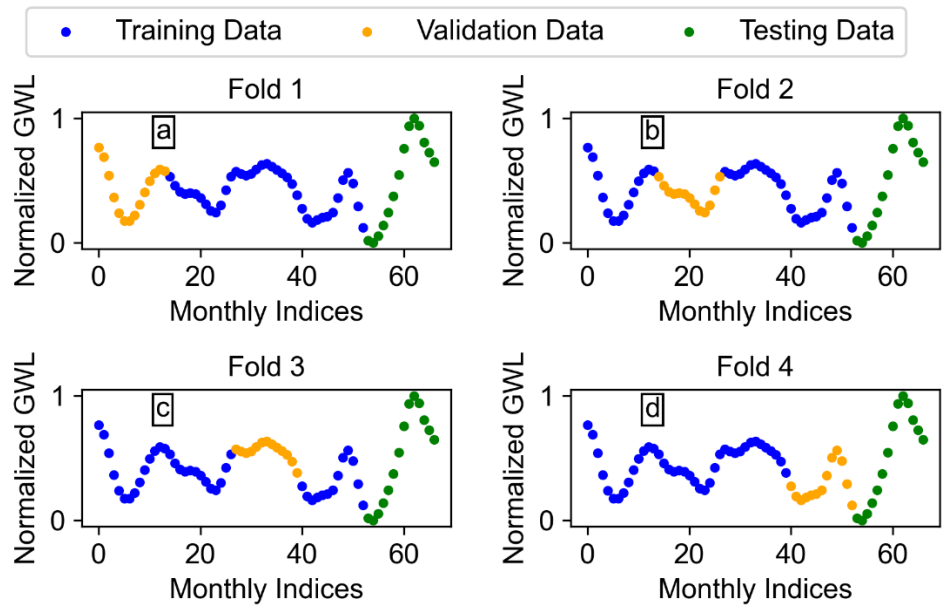


Figure B.6-KFCV used to partition the normalized monthly GWL data from well 440-1, DMS scenario.

DSM-FFNN 442-1:

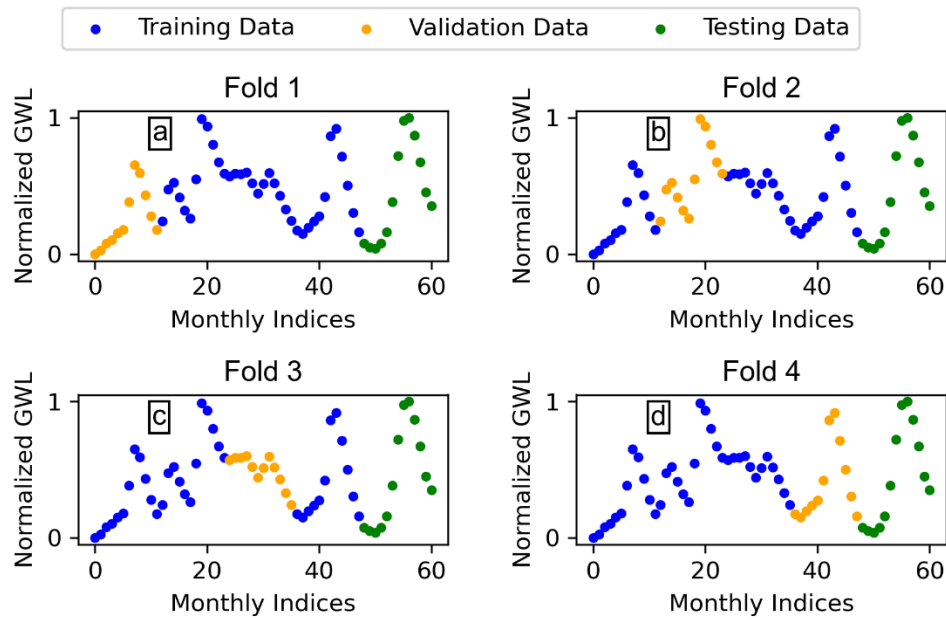


Figure B.7-KFCV used to partition the normalized monthly GWL data from well 442-1, DSM scenario.

DMS-FFNN 442-1:

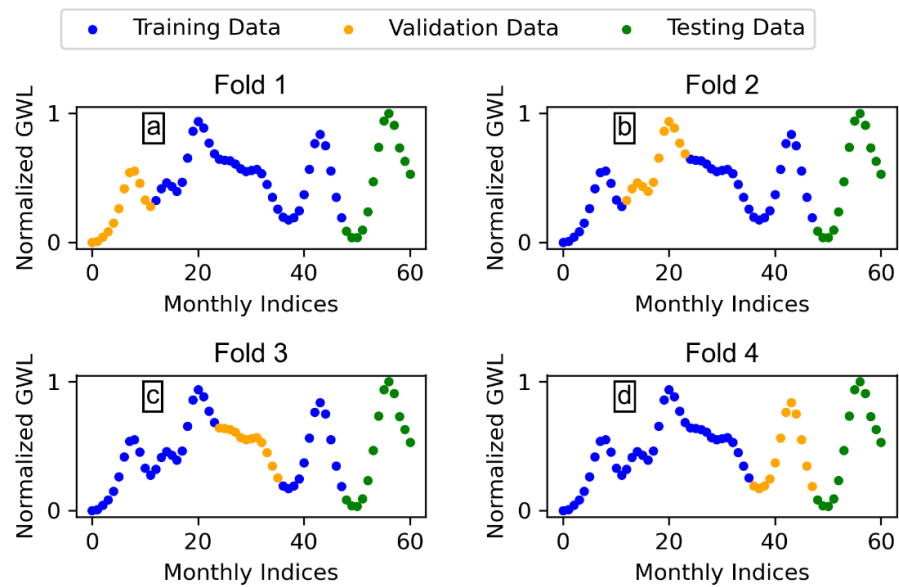
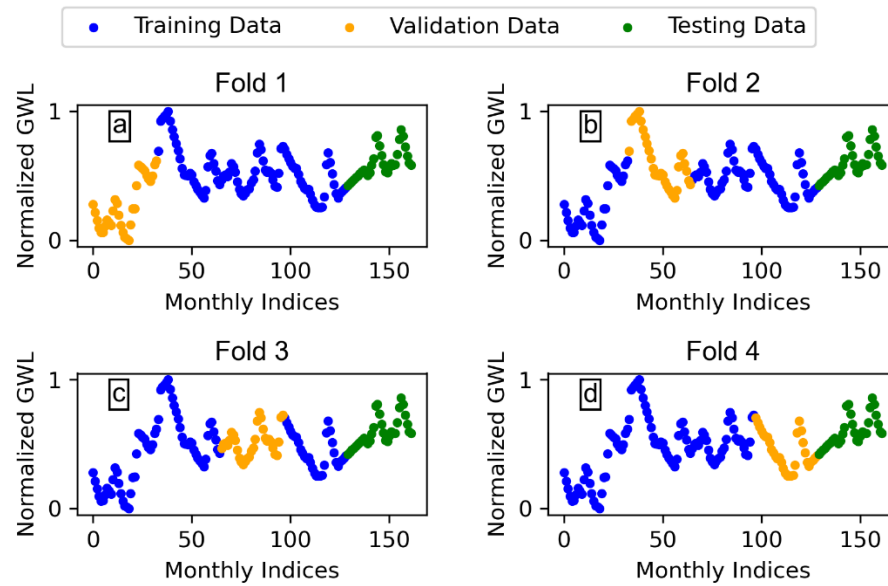
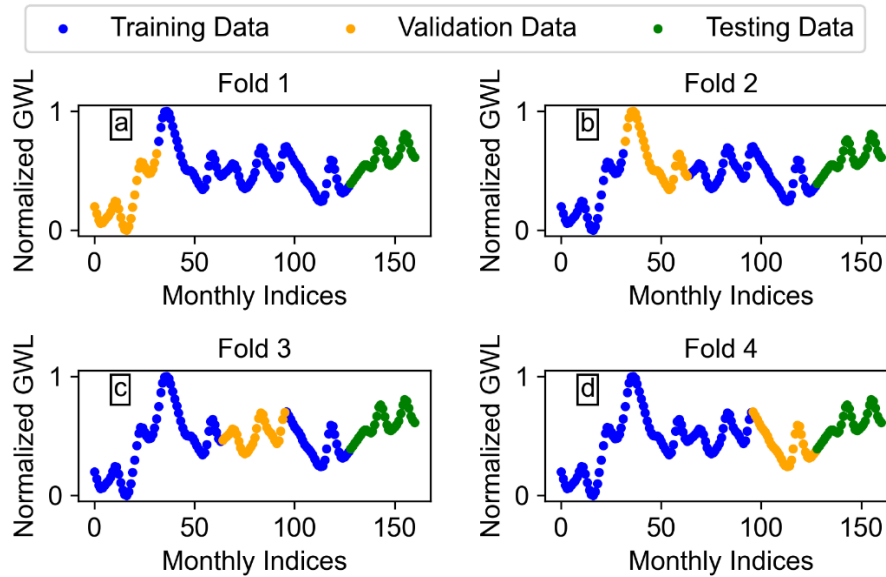


Figure B.8-KFCV used to partition the normalized monthly GWL data from well 442-1, DMS scenario.

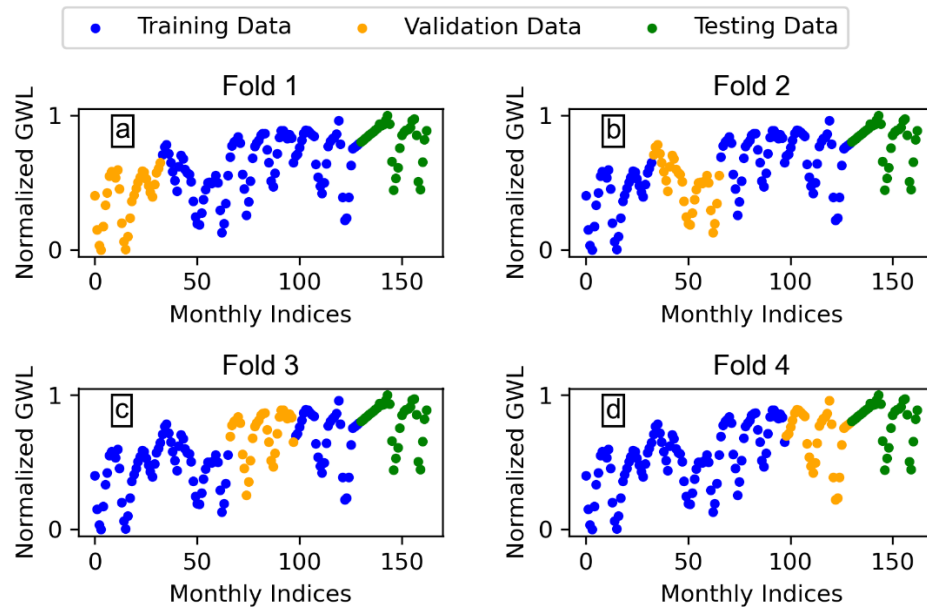
DSM-FFNN 244-2:



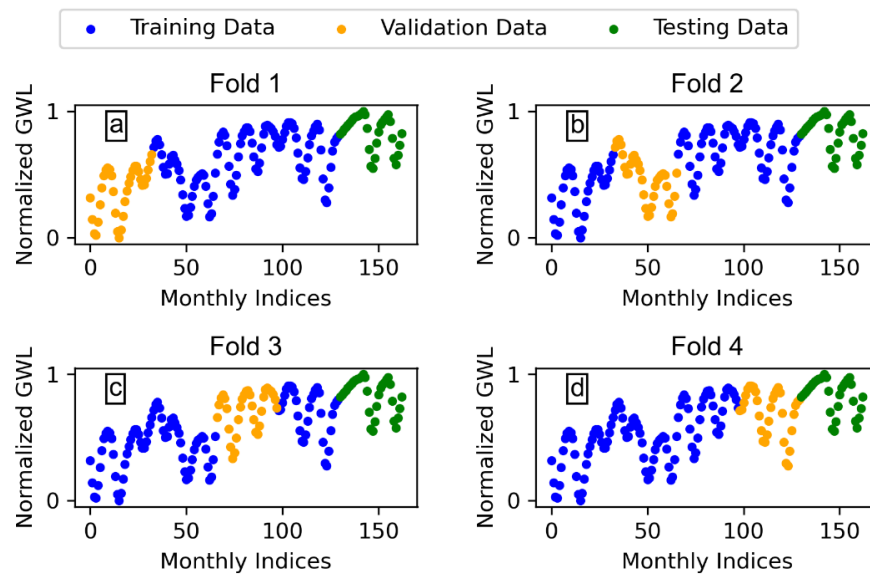
DMS-FFNN 244-2:



DSM-FFNN 245-2:



DMS-FFNN 245-2:



DSM-FFNN 443-1:

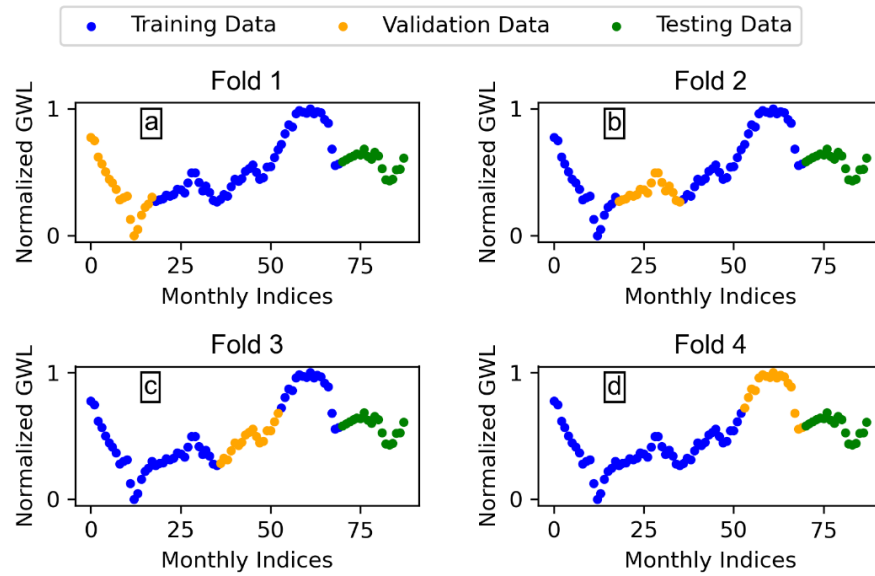


Figure B.13-KFCV used to partition the normalized monthly GWL data from well 443-1, DSM scenario.

DMS-FFNN 443-1:

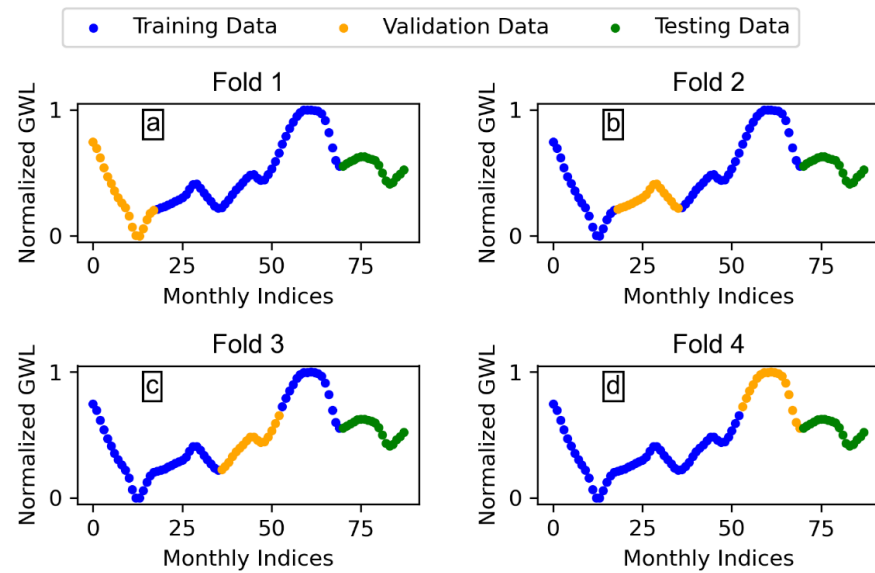


Figure B.14- KFCV used to partition the normalized monthly GWL data from well 443-1, DMS scenario.

Performance Plateau By-Fold Plots (R^2)

DSM-FFNN model by-fold graph for 293-3:

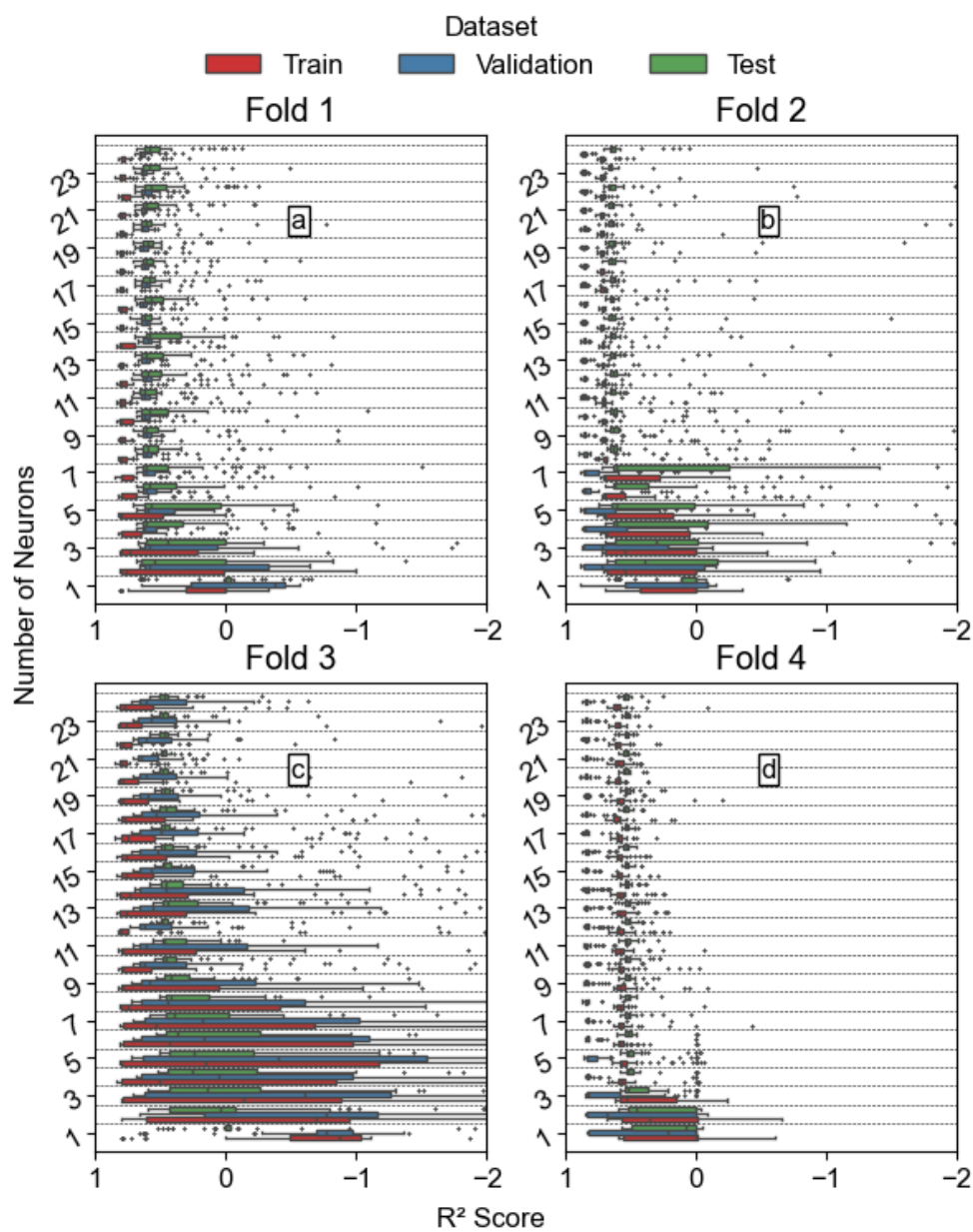


Figure B.15- R^2 score distributions across different neuron counts for well 293-3 in each cross-validation fold, DSM scenario.

DMS-FFNN model by-fold graph for 293-3:

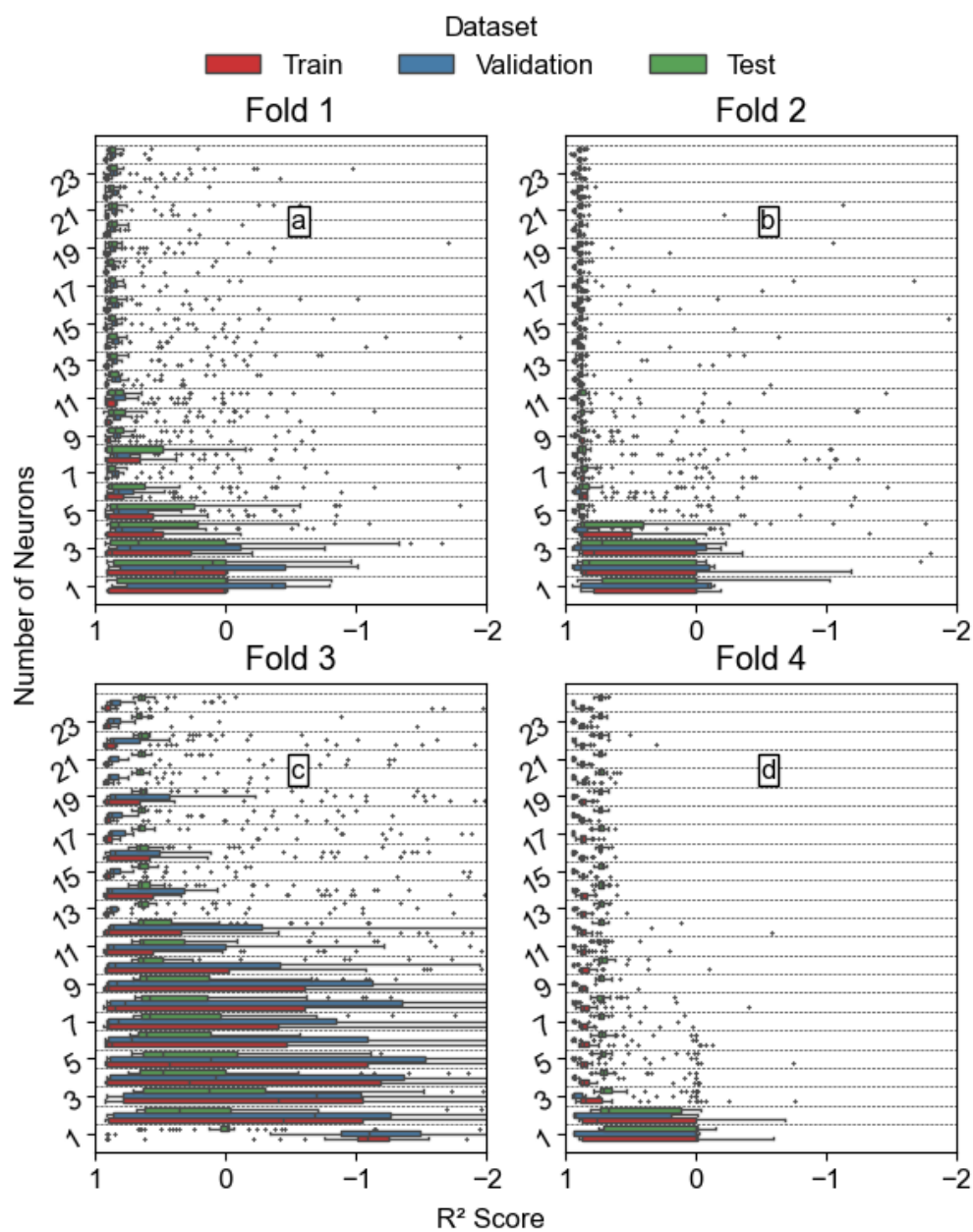


Figure B.16- R^2 score distributions across different neuron counts for well 293-3 in each cross-validation fold, DMS scenario.

DSM-FFNN model by-fold graph for W1:

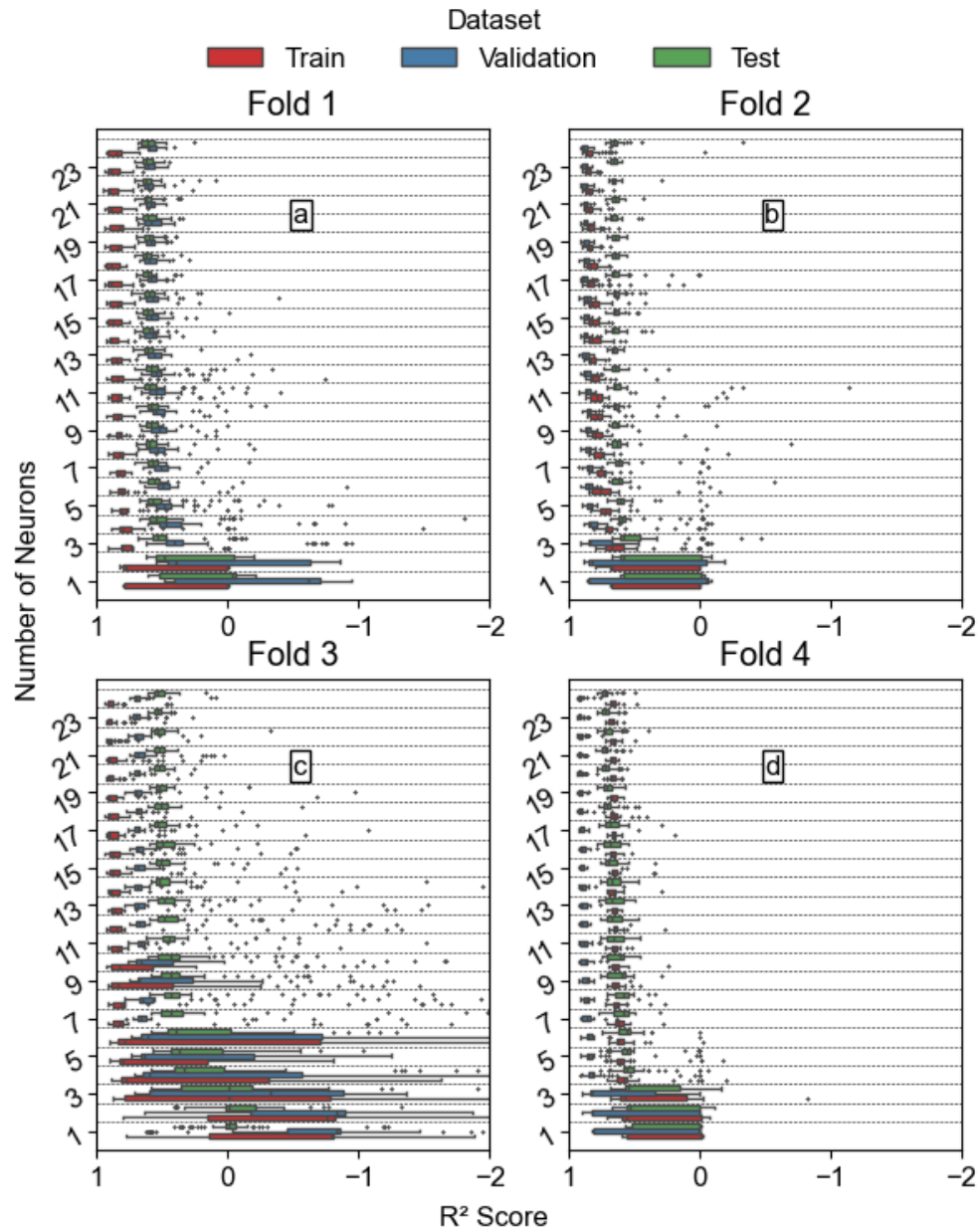


Figure B.17- R² score distributions across different neuron counts for W1 in each cross-validation fold, DSM scenario.

DMS-FFNN model by-fold graph for W1:

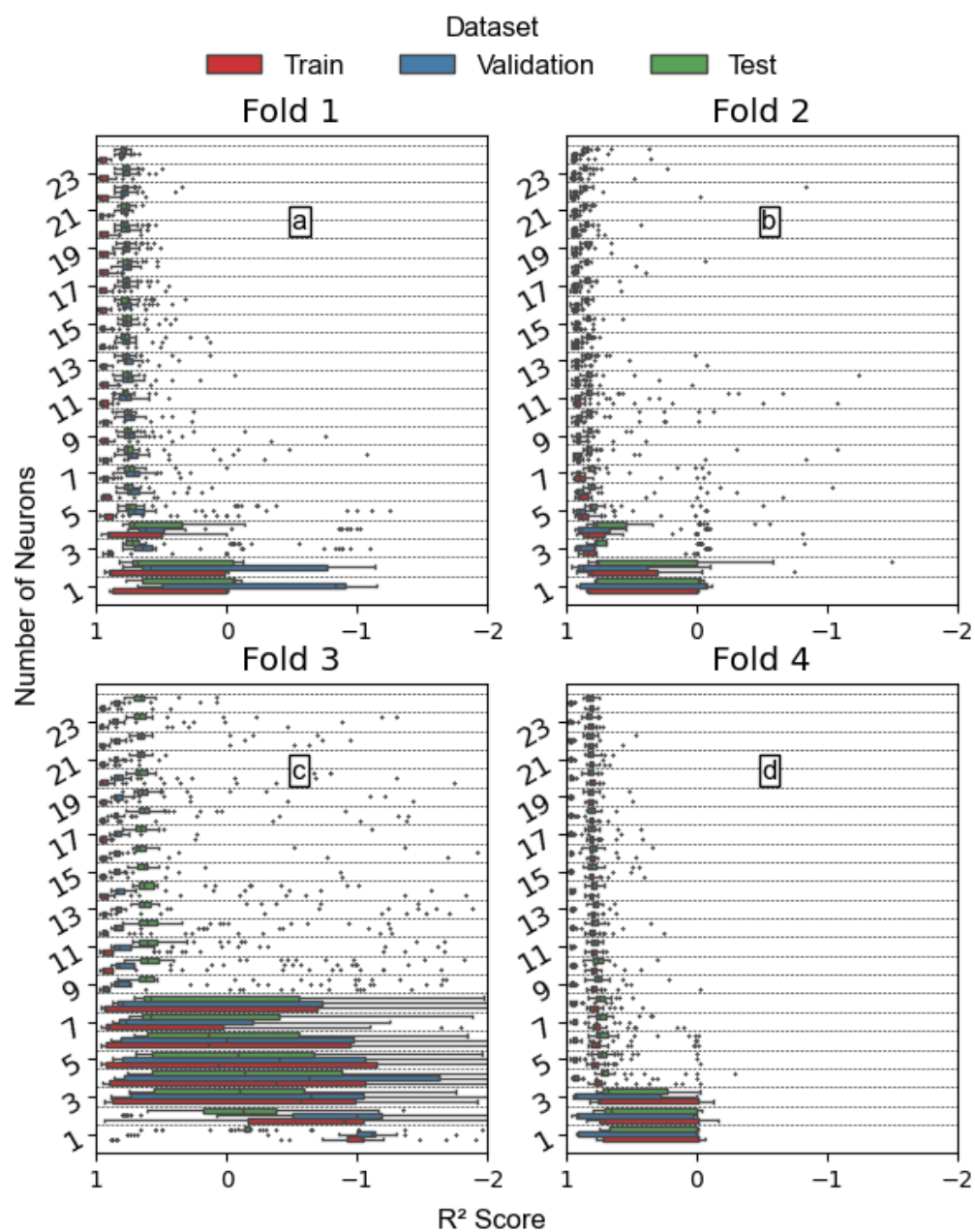


Figure B.18- R² score distributions across different neuron counts for W1 in each cross-validation fold, DMS scenario.

DSM-FFNN model by-fold graph for 439-1:

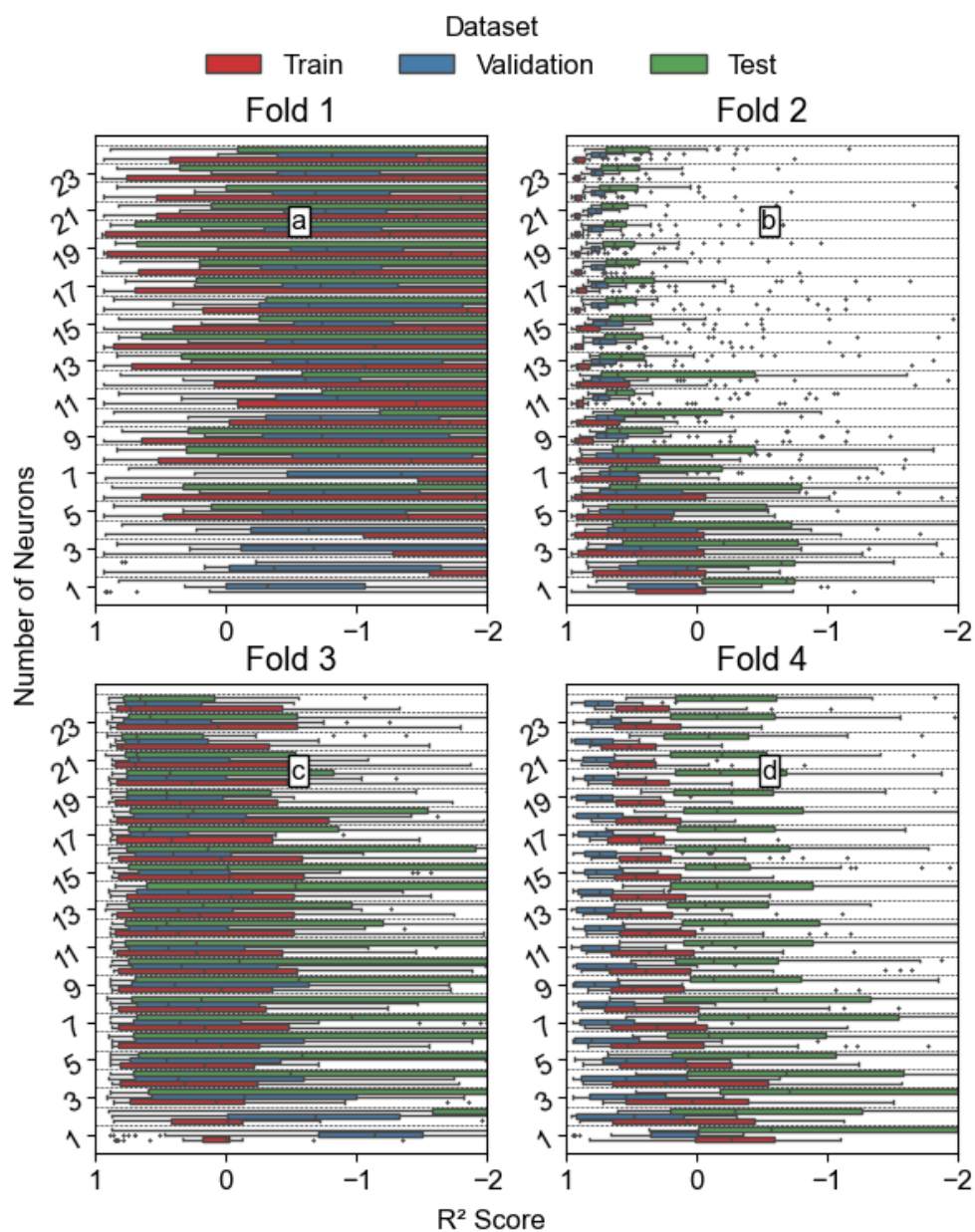


Figure B.19- R² score distributions across different neuron counts for well 439-1 in each cross-validation fold, DSM scenario.

DMS-FFNN model by-fold graph for 439-1:

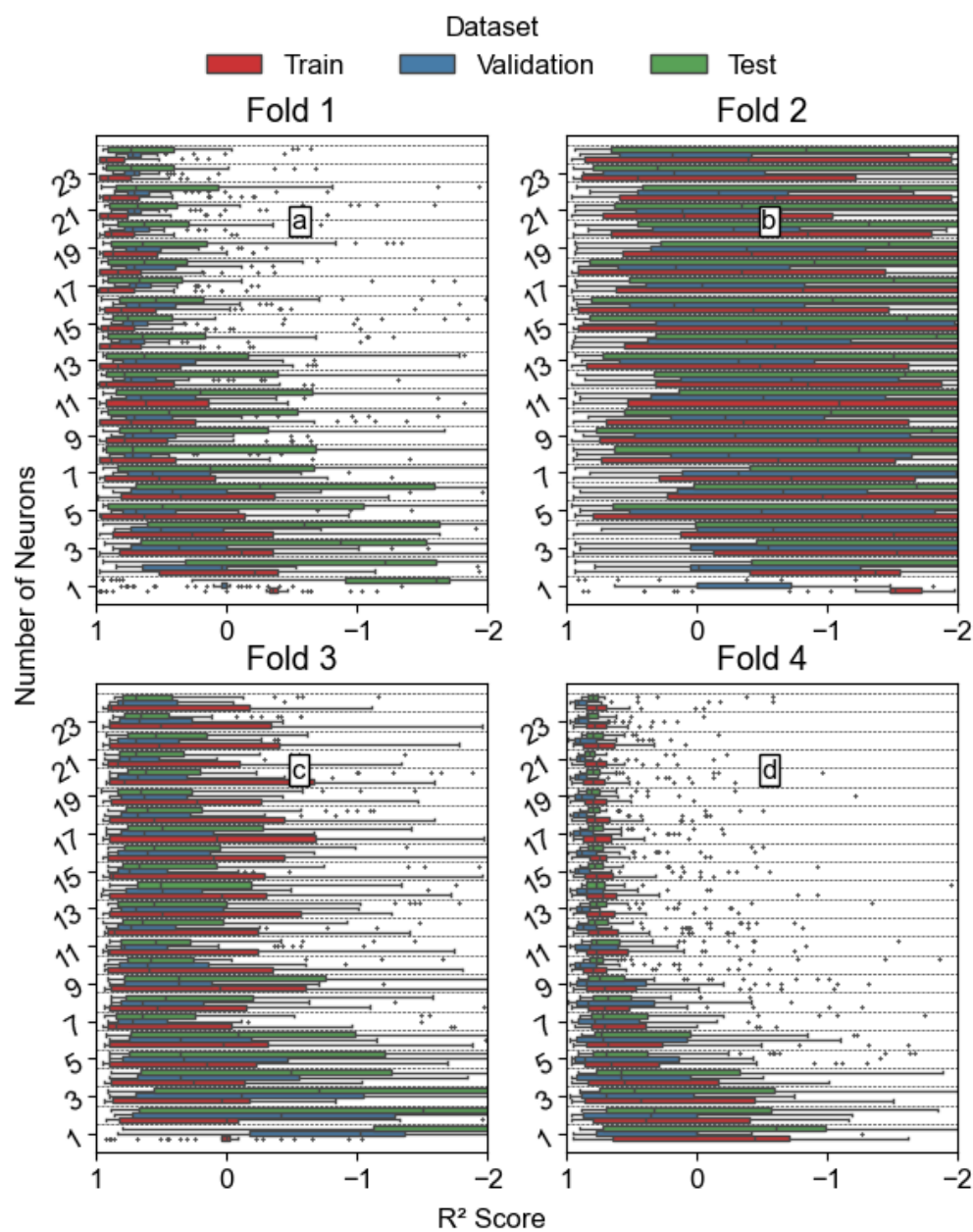


Figure B.20- R^2 score distributions across different neuron counts for well 439-1 in each cross-validation fold, DMS scenario.

DSM-FFNN model saturation by fold graph for 440-1:

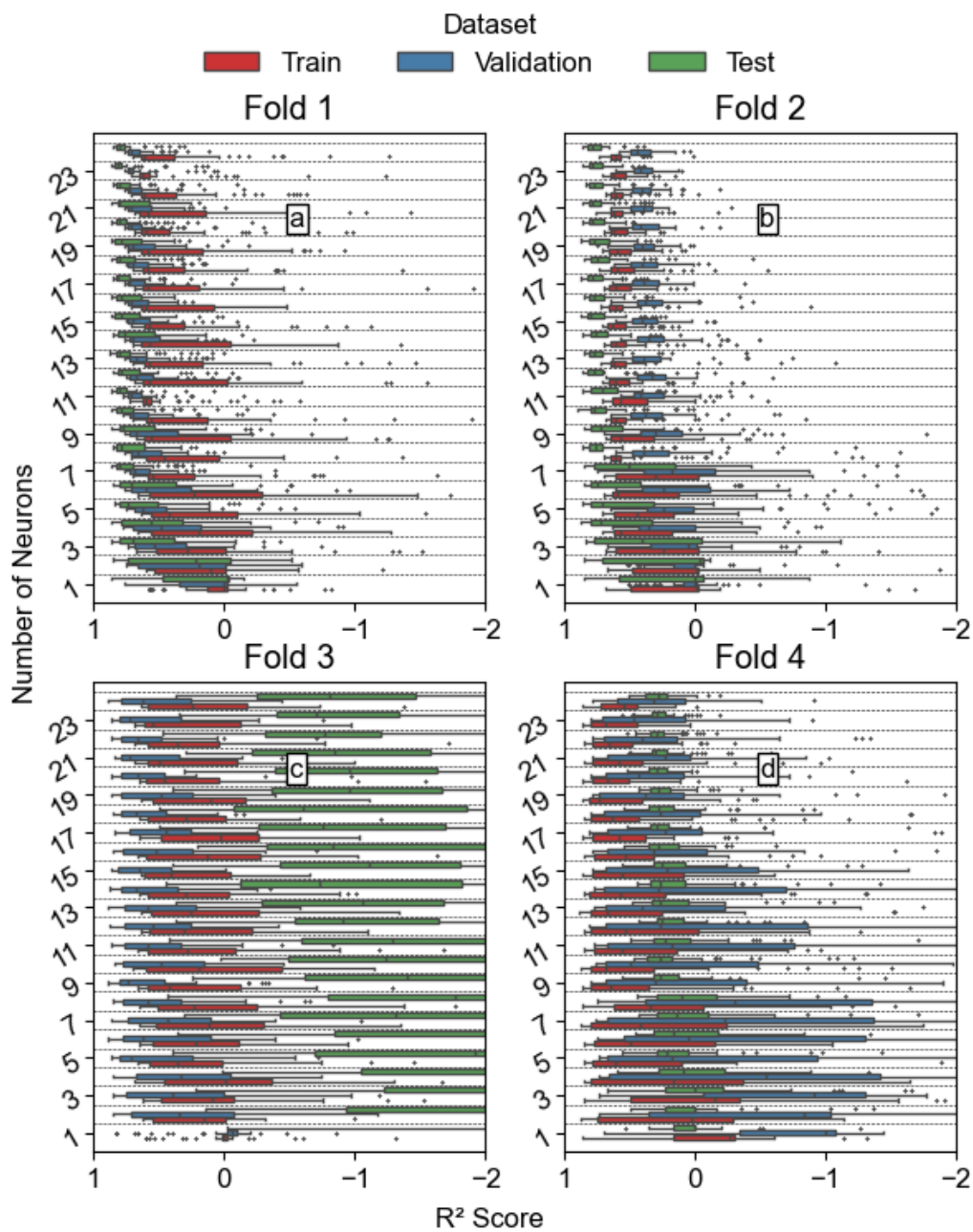


Figure B.21- R² score distributions across different neuron counts for well 440-1 in each cross-validation fold, DSM scenario.

DMS-FFNN model saturation by fold graph for 440-1:

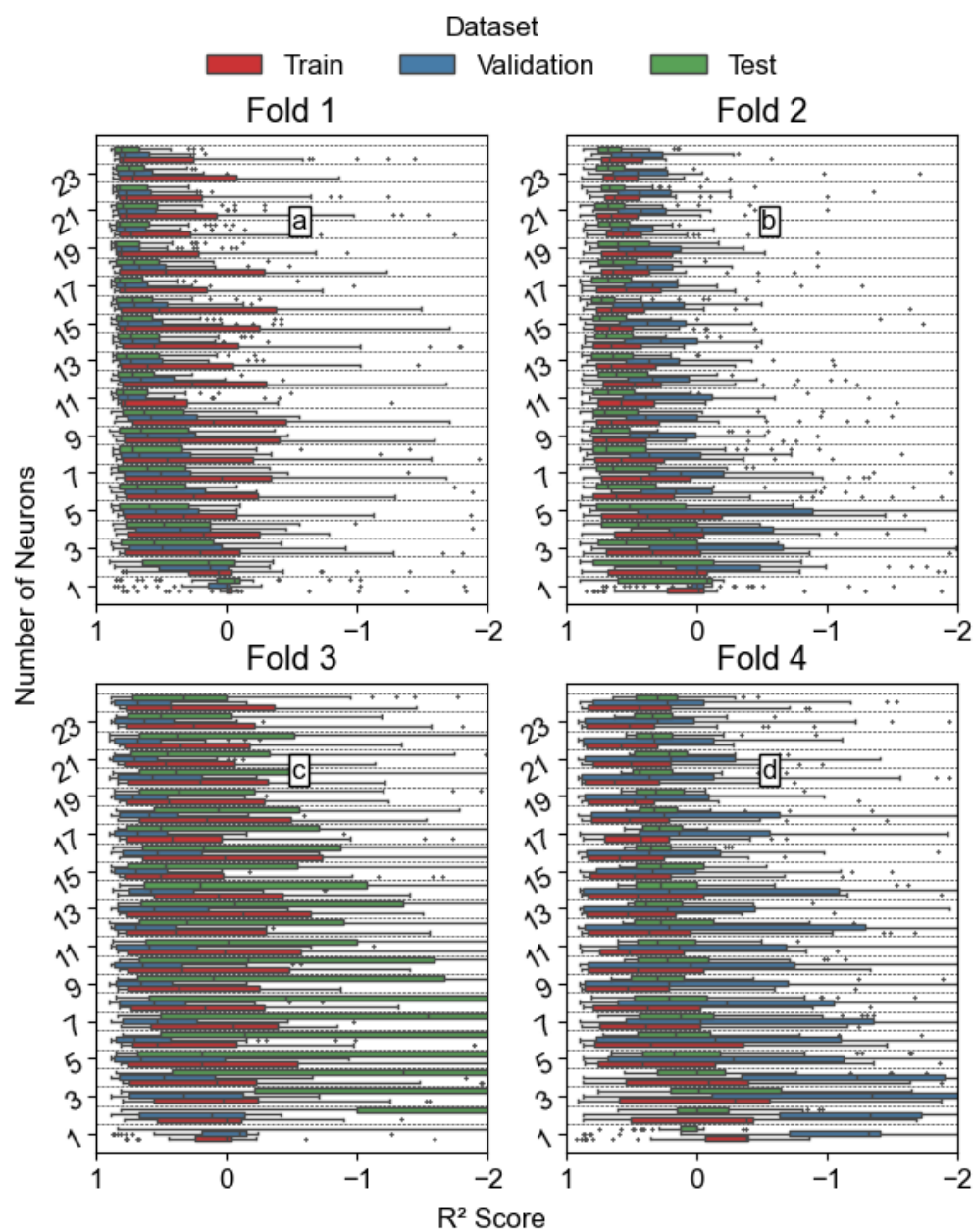


Figure B.22- R² score distributions across different neuron counts for well 440-1 in each cross-validation fold, DMS scenario.

DSM-FFNN model saturation by fold graph for 442-1:

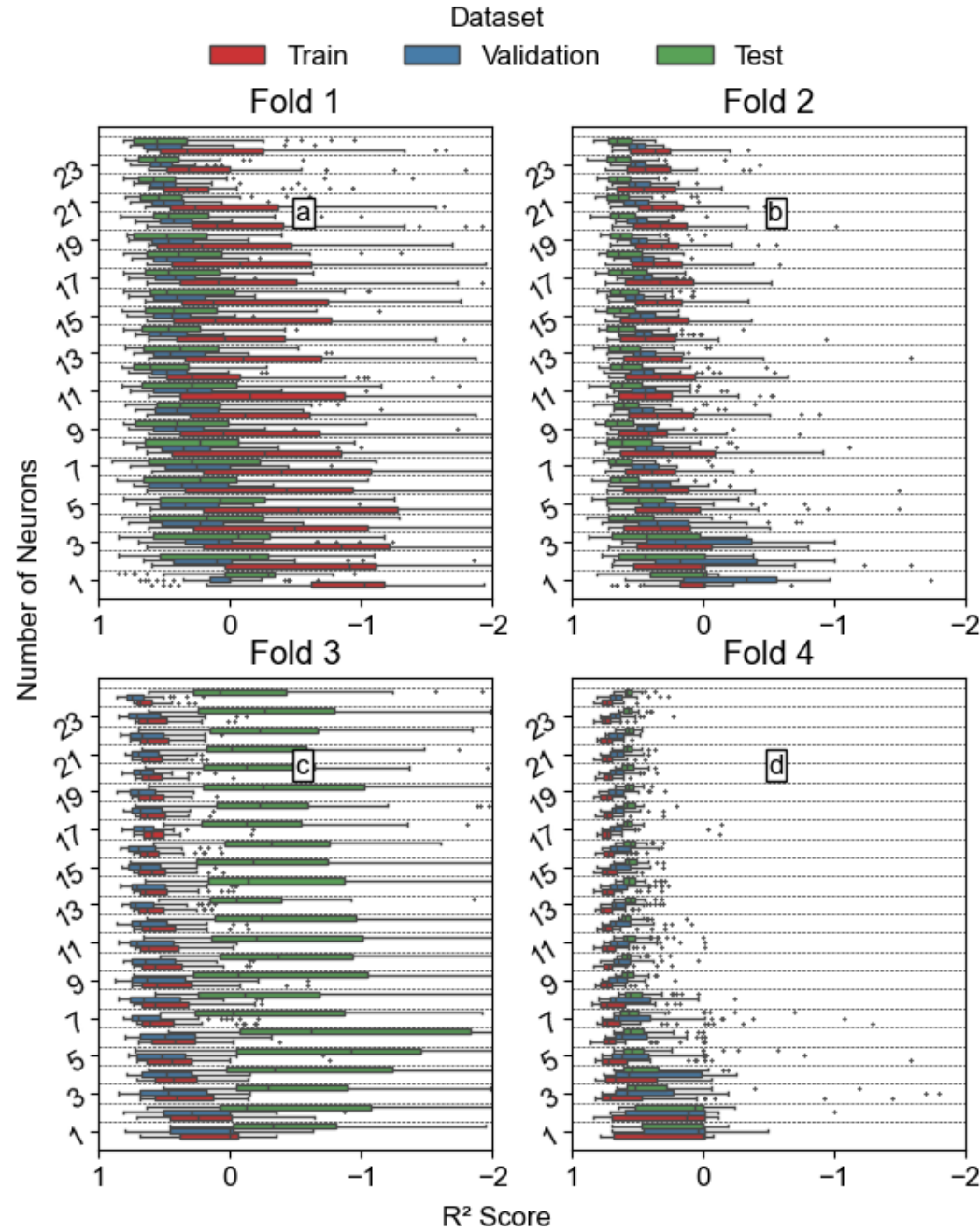


Figure B.23- R^2 score distributions across different neuron counts for well 442-1 in each cross-validation fold, DSM scenario.

DMS-FFNN model saturation by fold graph for 442-1:

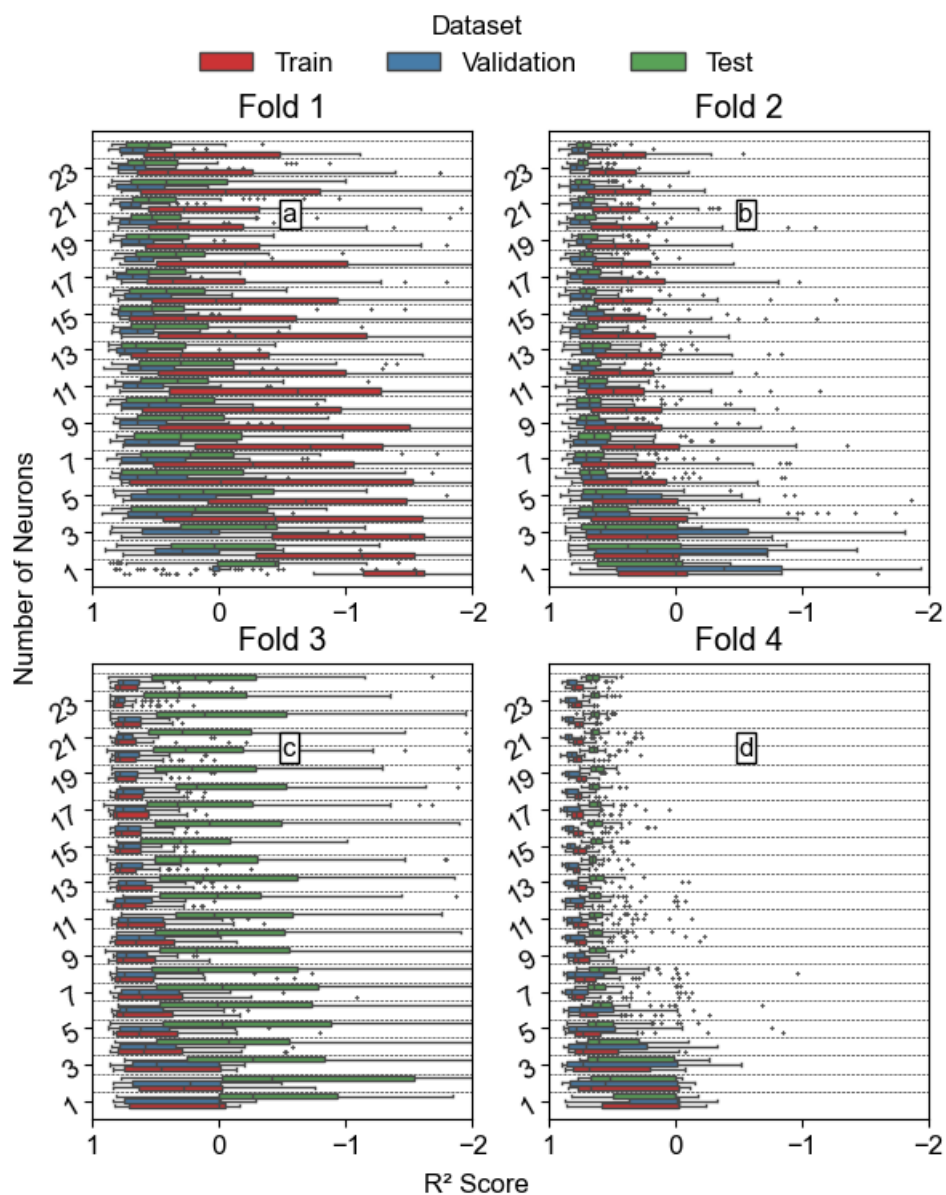


Figure B.24- R^2 score distributions across different neuron counts for well 442-1 in each cross-validation fold, DMS scenario.

DSM-FFNN plateau by fold graph for W2:

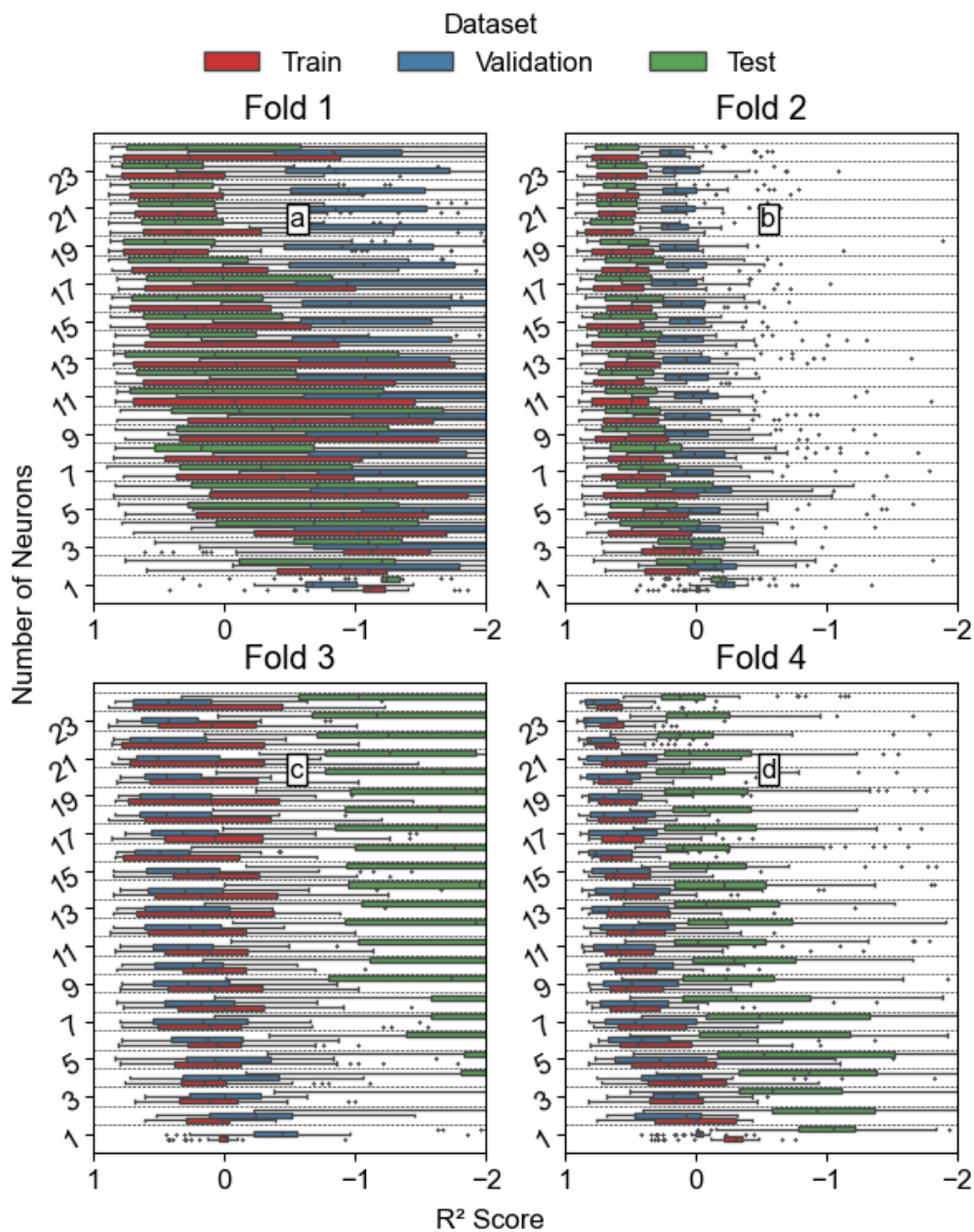


Figure B.25- R^2 score distributions across different neuron counts for W2 combined in each cross-validation fold, DSM scenario.

DMS-FFNN plateau by fold graph for W2:

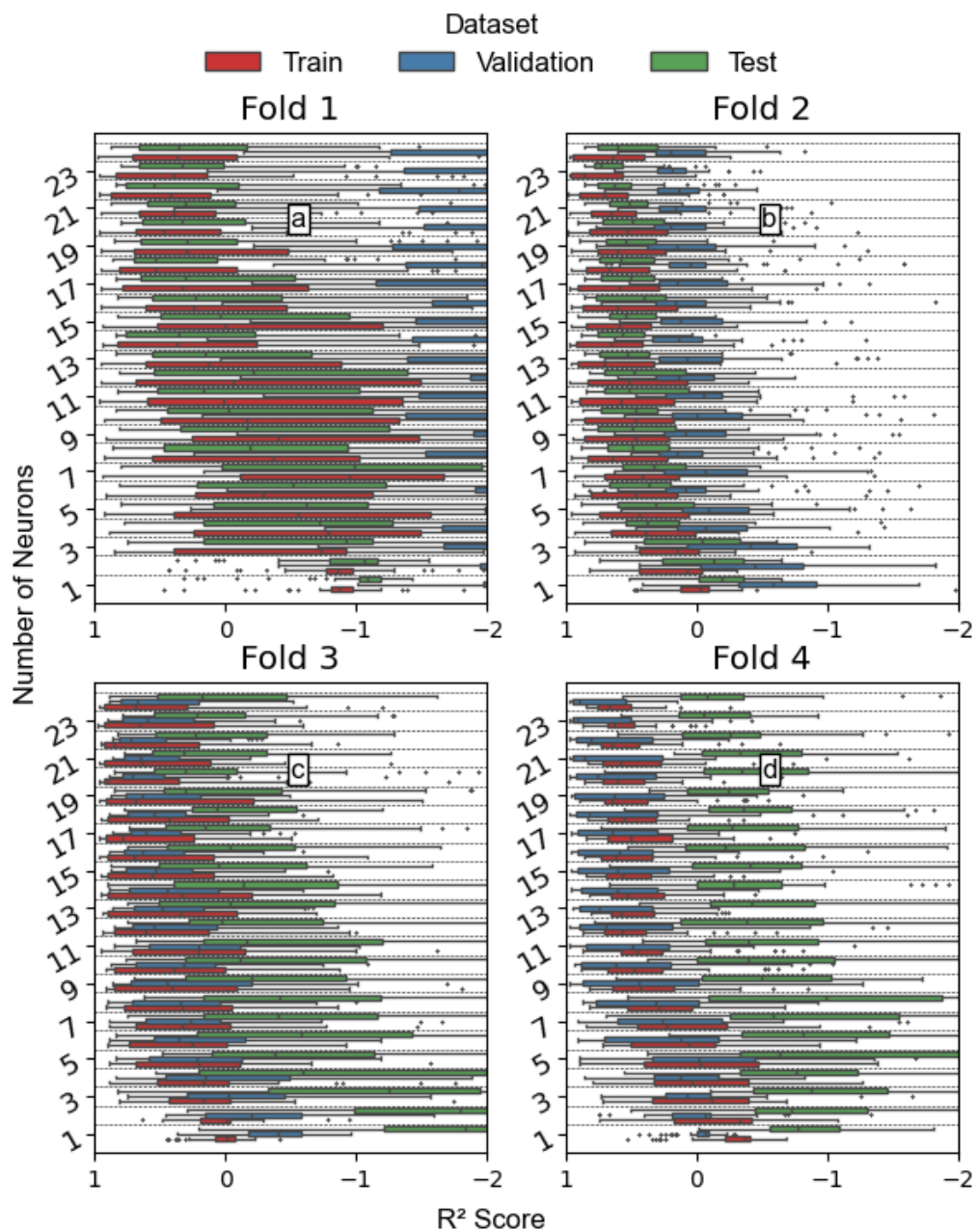


Figure B.26- R^2 score distributions across different neuron counts for W2 combined in each cross-validation fold, DMS scenario.

DSM-FFNN model saturation by fold graph for 244-2:

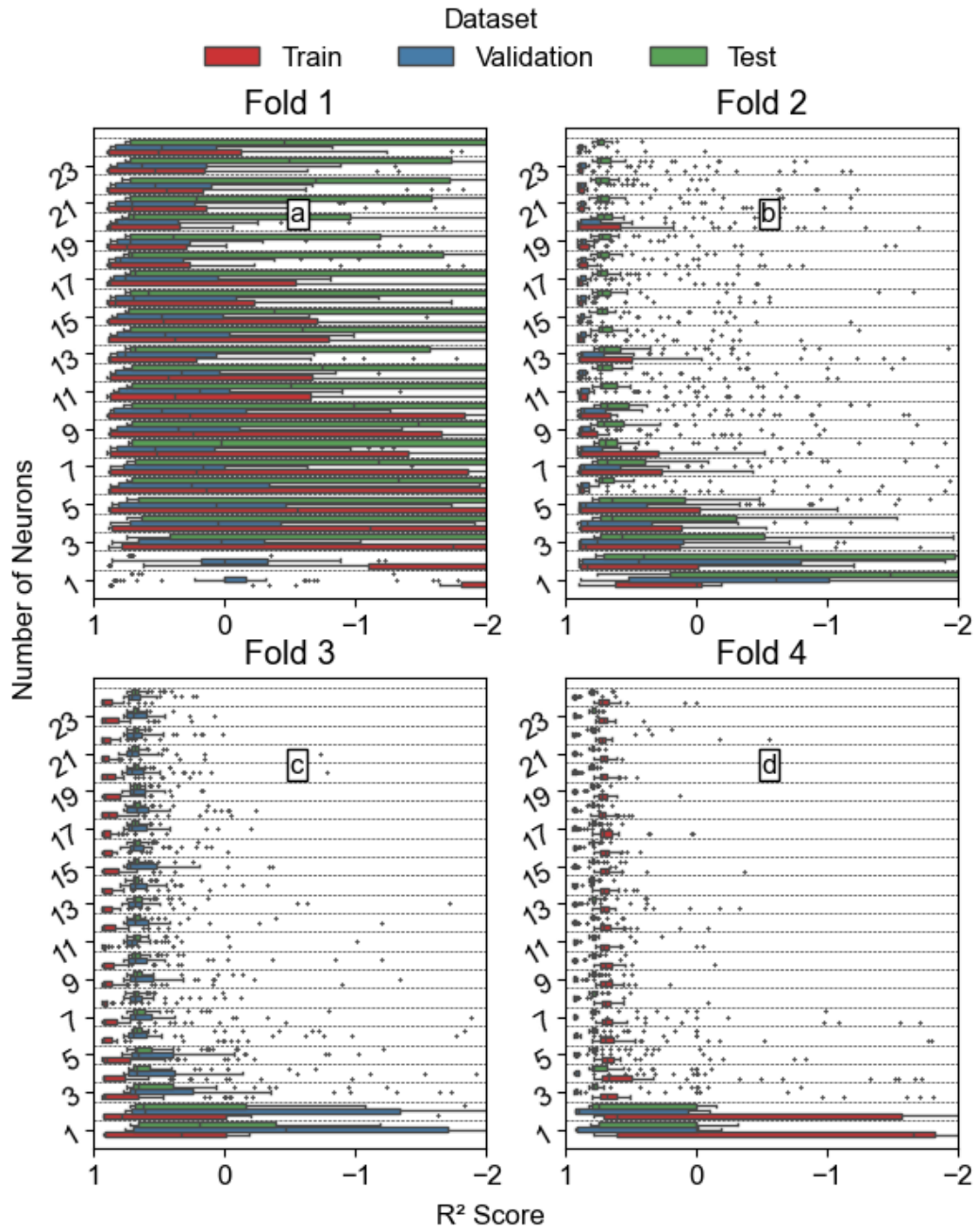


Figure B.27- R^2 score distributions across different neuron counts for well 244-2 in each cross-validation fold, DSM scenario.

DMS-FFNN model saturation by fold graph for 244-2:

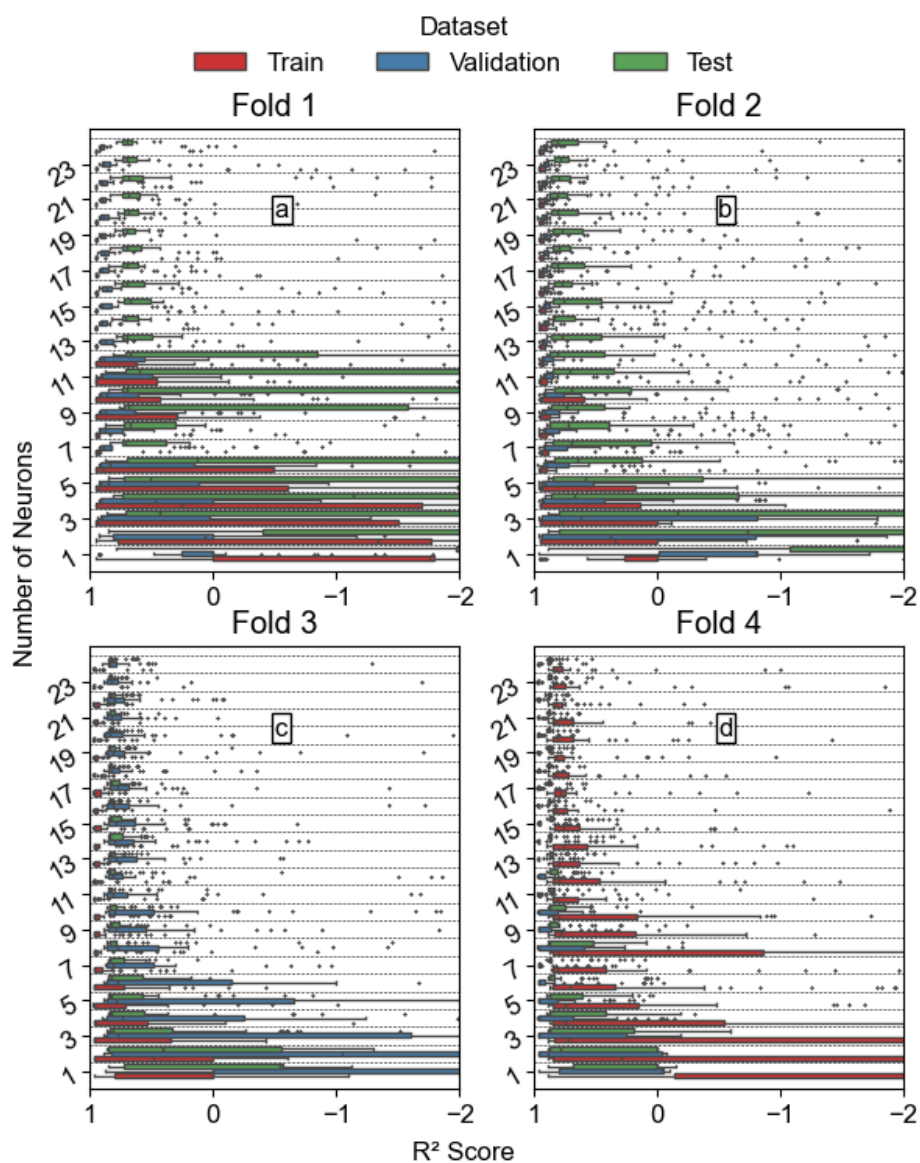


Figure B.28- R^2 score distributions across different neuron counts for well 244-2 in each cross-validation fold, DMS scenario.

DSM-FFNN model saturation by fold graph for 245-2:

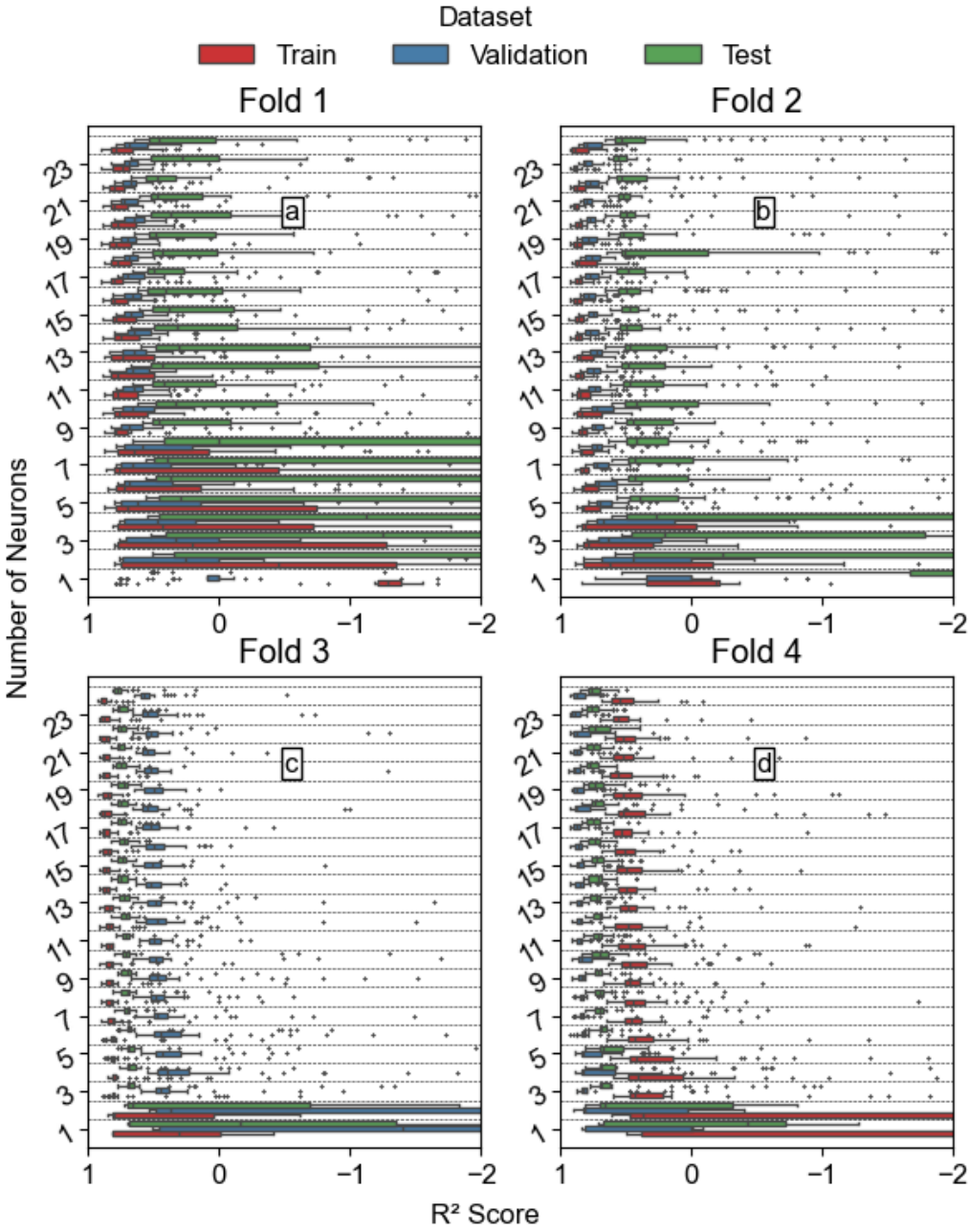


Figure B.29- R^2 score distributions across different neuron counts for well 245-2 in each cross-validation fold, DSM scenario.

DMS-FFNN model saturation by fold graph for 245-2:

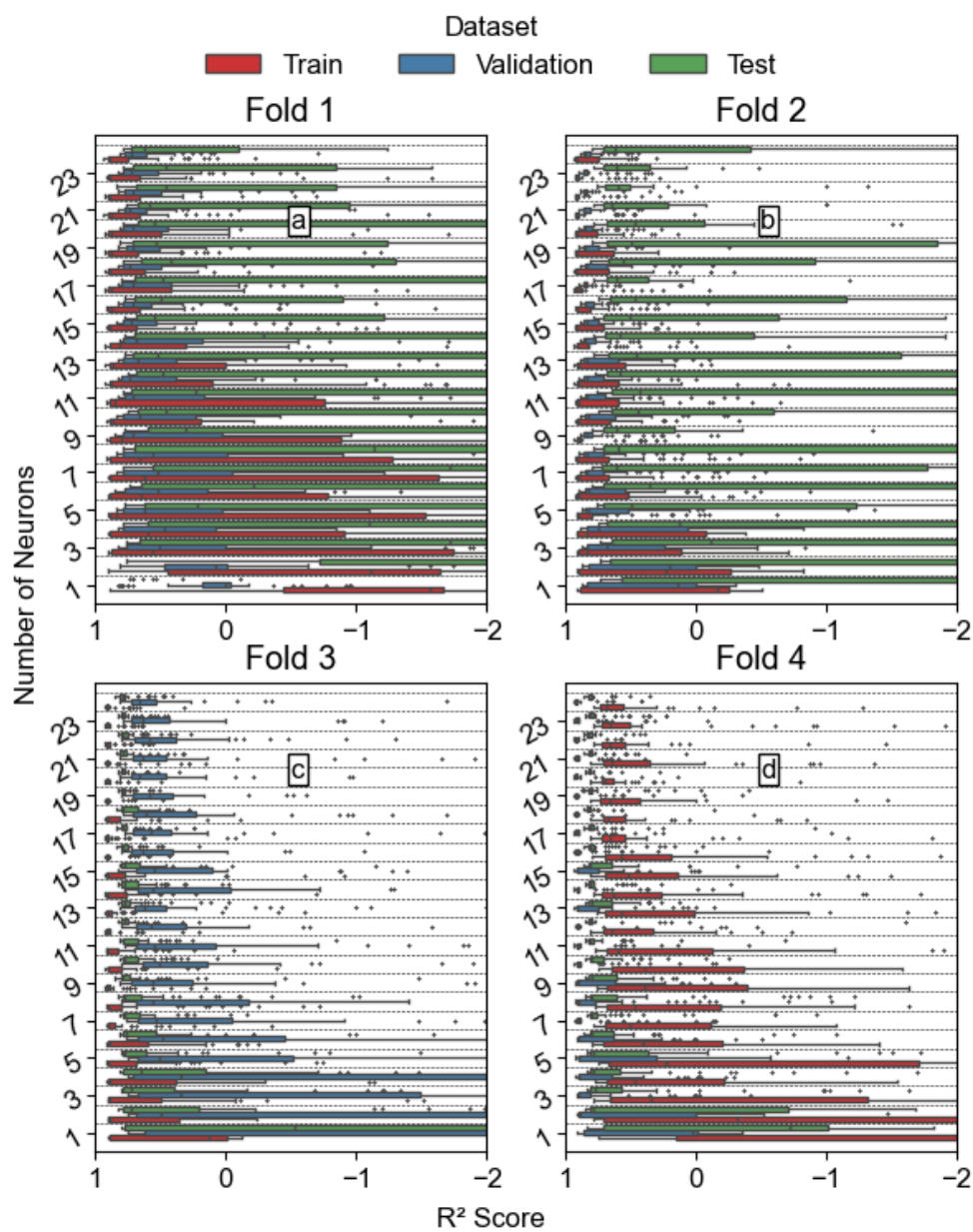


Figure B.30- R^2 score distributions across different neuron counts for well 245-2 in each cross-validation fold, DMS scenario.

DSM-FFNN model plateau by fold graph for W3:

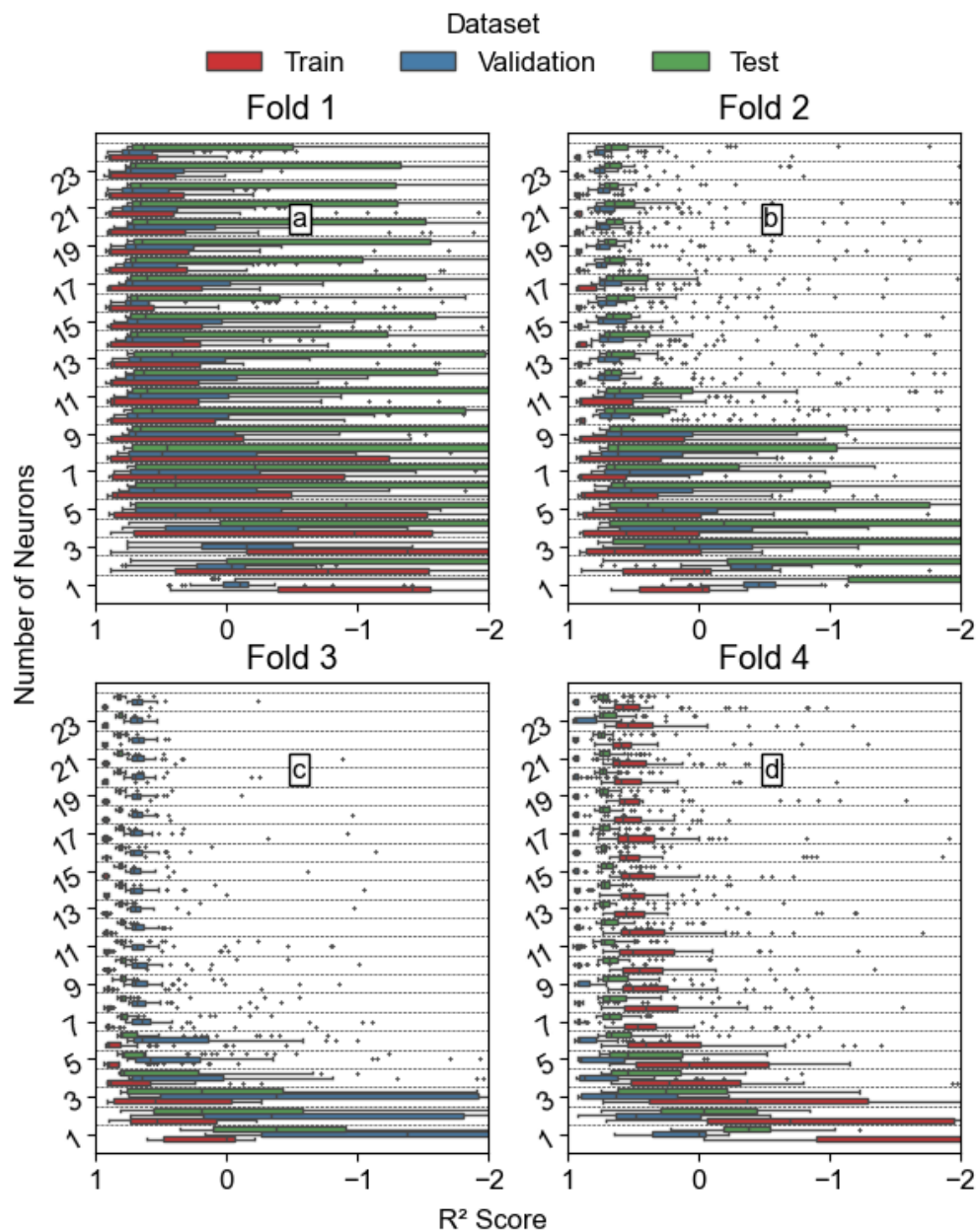


Figure B.31- R^2 score distributions across different neuron counts for W3 combined in each cross-validation fold, DSM scenario.

DMS-FFNN model plateau by fold graph for W3:

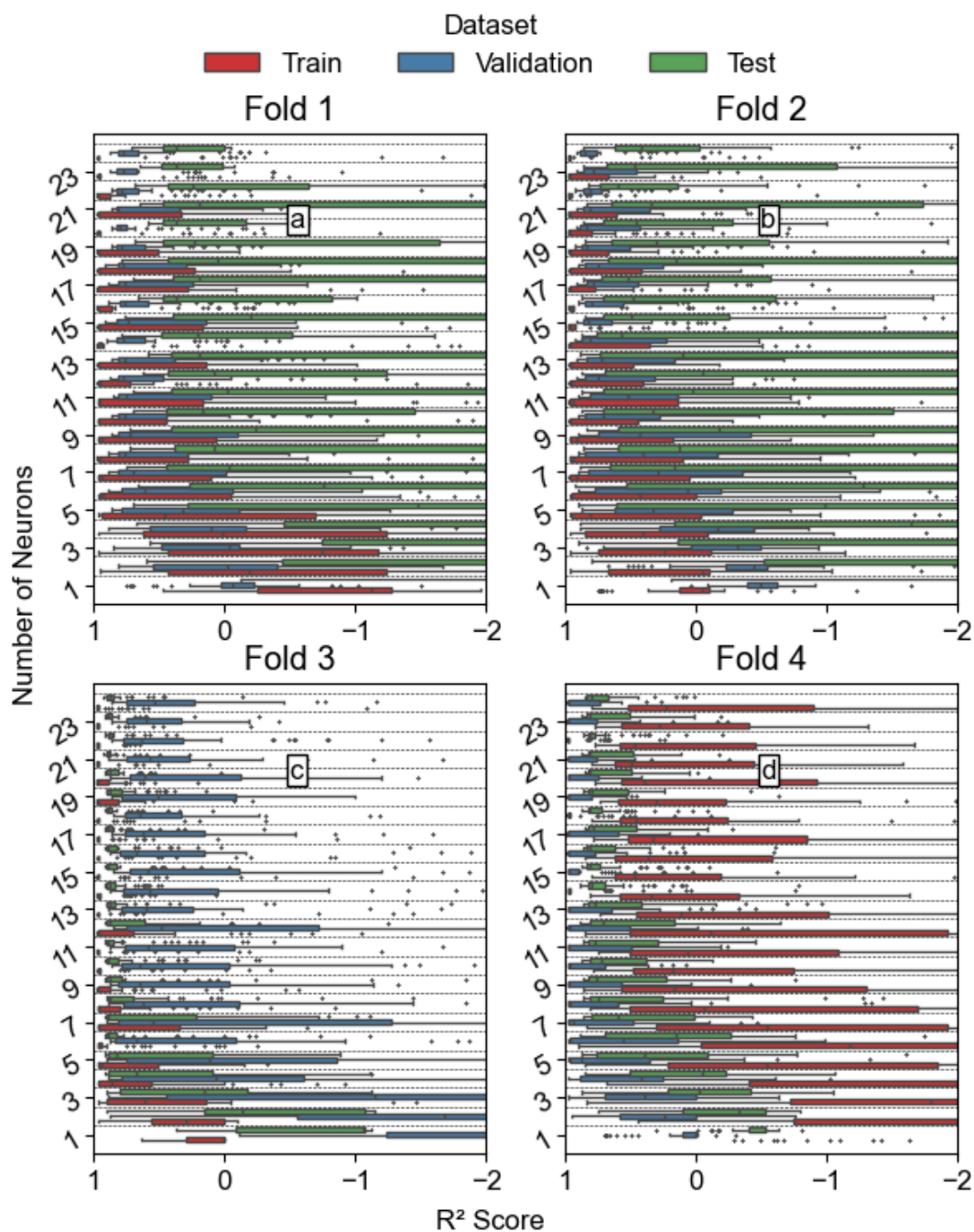


Figure B.32- R^2 score distributions across different neuron counts for W3 combined in each cross-validation fold, DMS scenario.

DSM-FFNN model saturation by fold graph for 443-1:

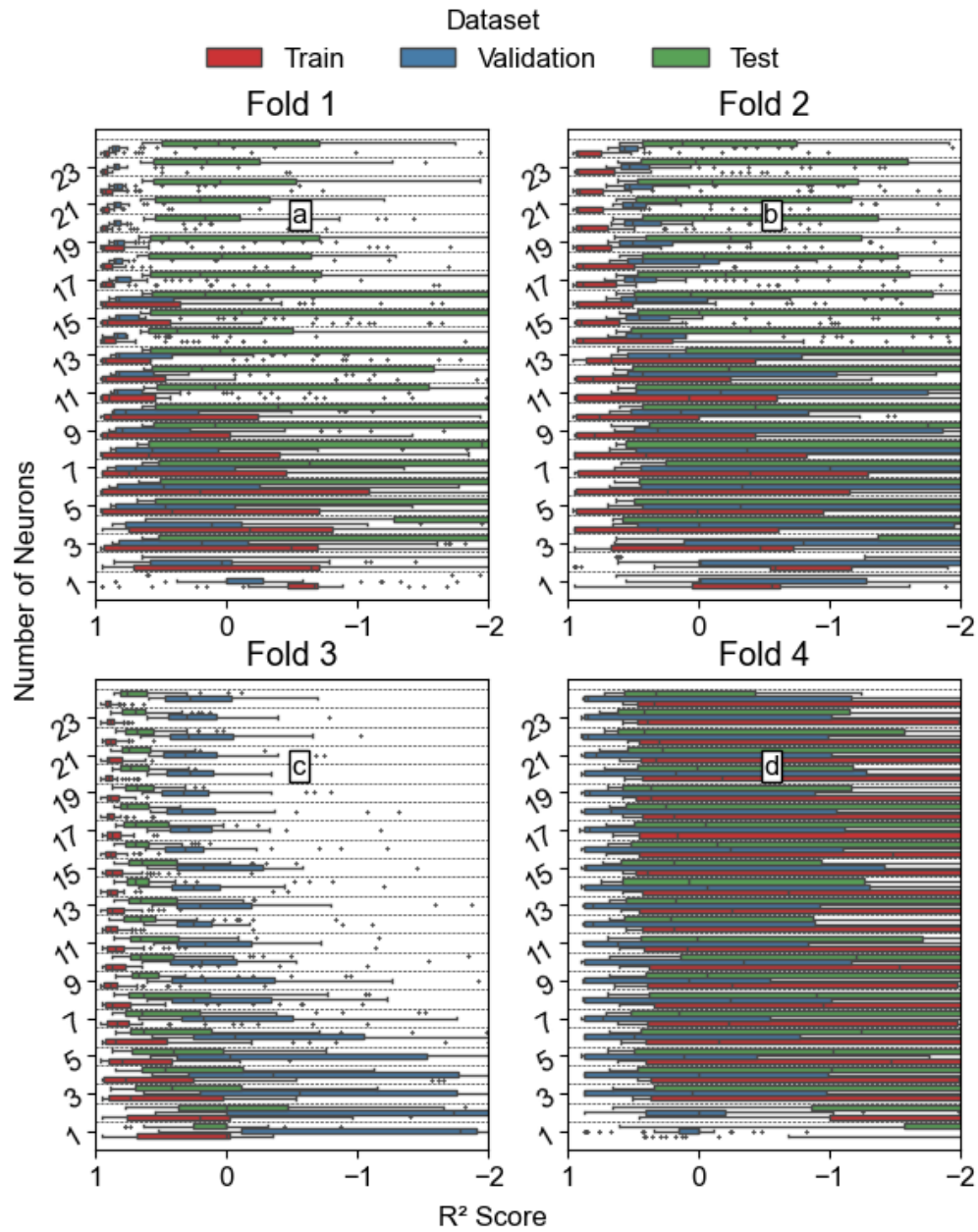


Figure B.33- R² score distributions across different neuron counts for well 443-1 in each cross-validation fold, DSM scenario.

DMS-FFNN model saturation by fold graph for 443-1:

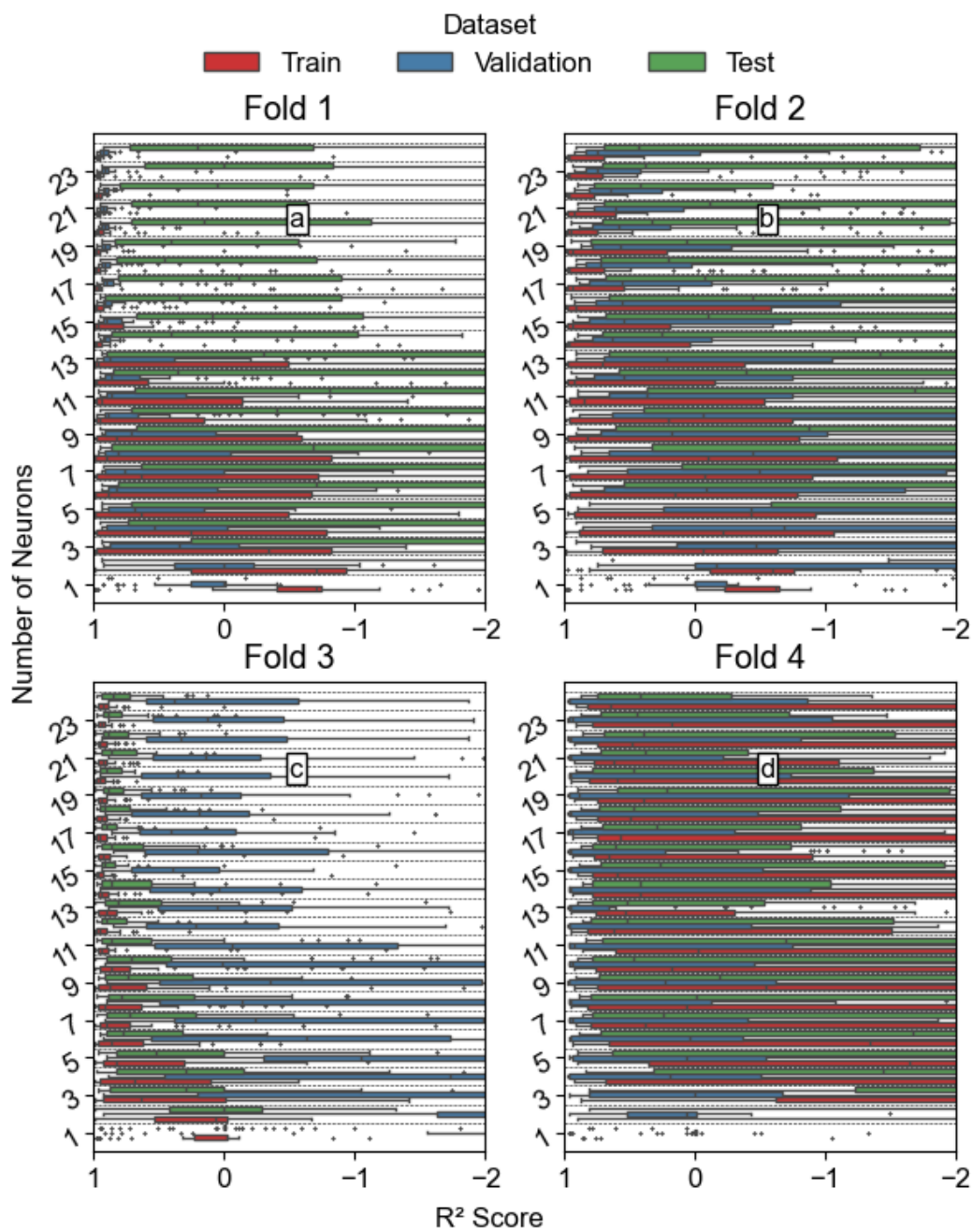


Figure B.34- R² score distributions across different neuron counts for well 443-1 in each cross-validation fold, DMS scenario.

DSM-FFNN model for the aggregated model:

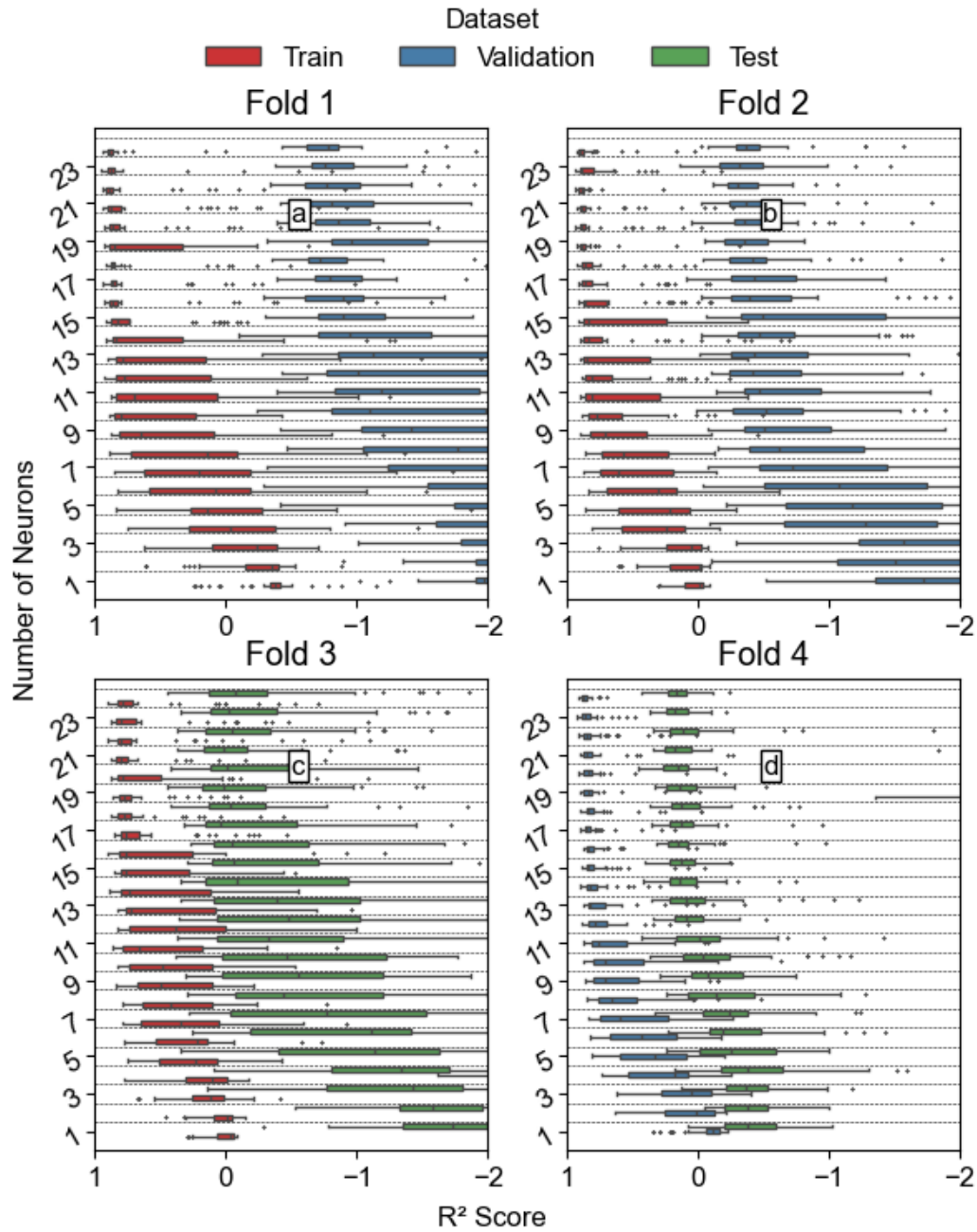


Figure B.35- R^2 score distributions across different neuron counts for the aggregated model in each cross-validation fold, DSM scenario.

DMS-FFNN model for the aggregated model:

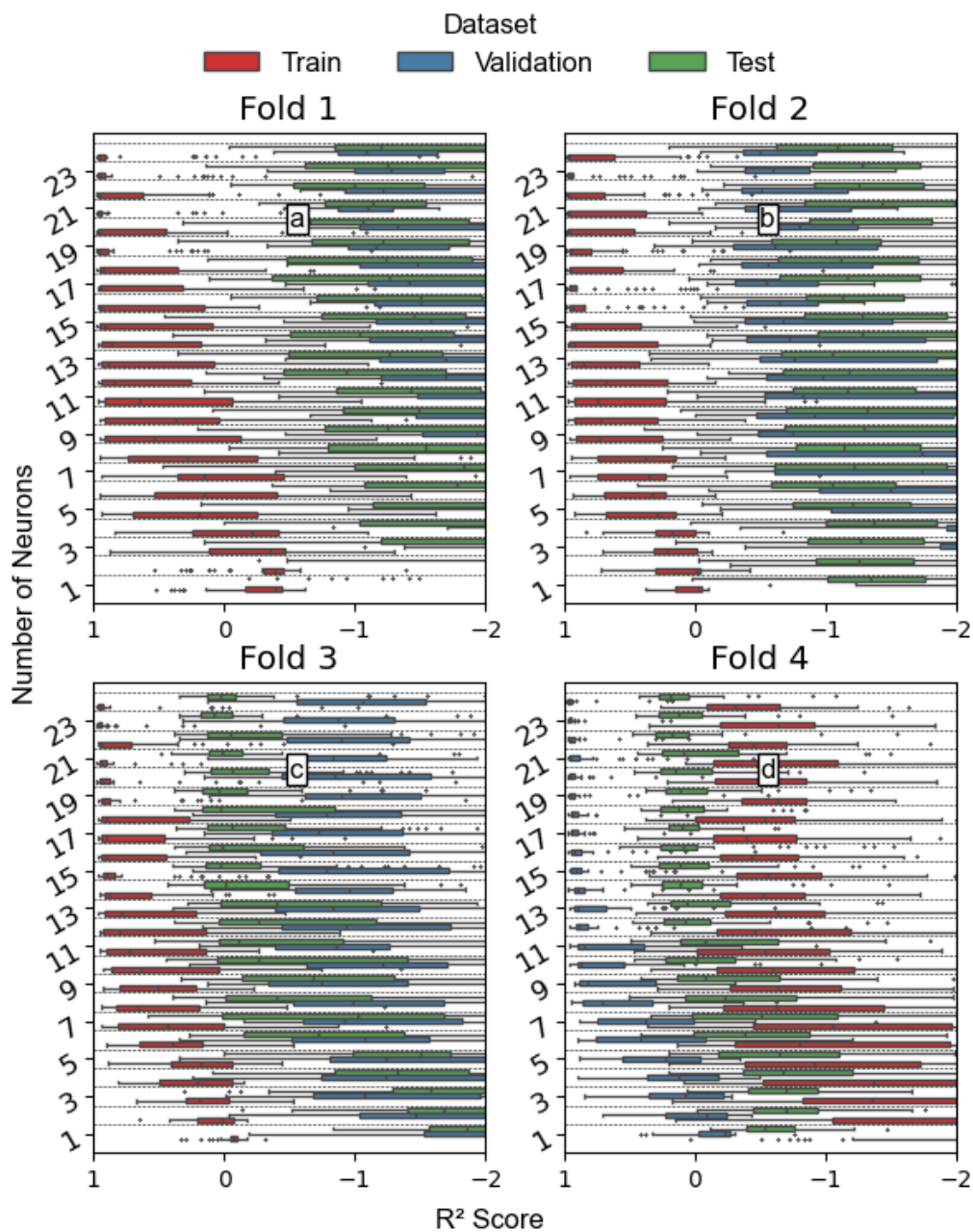


Figure B.36- R^2 score distributions across different neuron counts for the aggregated model in each cross-validation fold, DMS scenario.

Performance Plateau Overall Plots (R^2)

DSM-FFNN model plateau overall graph for 293-3:

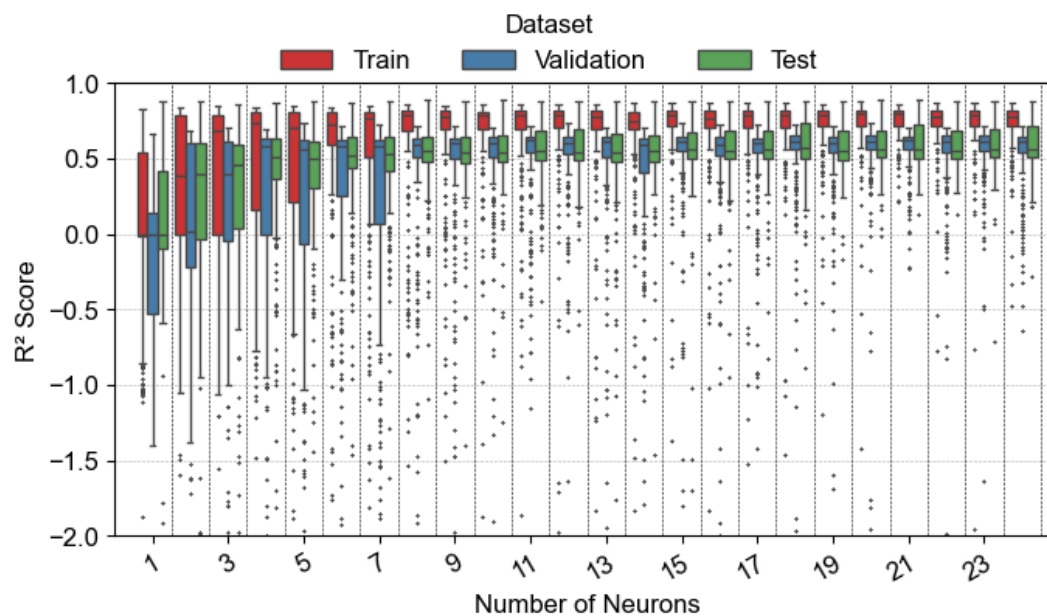


Figure B.37- Combined R^2 score distributions for well 293-3, DSM scenario.

DMS-FFNN model plateau overall graph for 293-3:

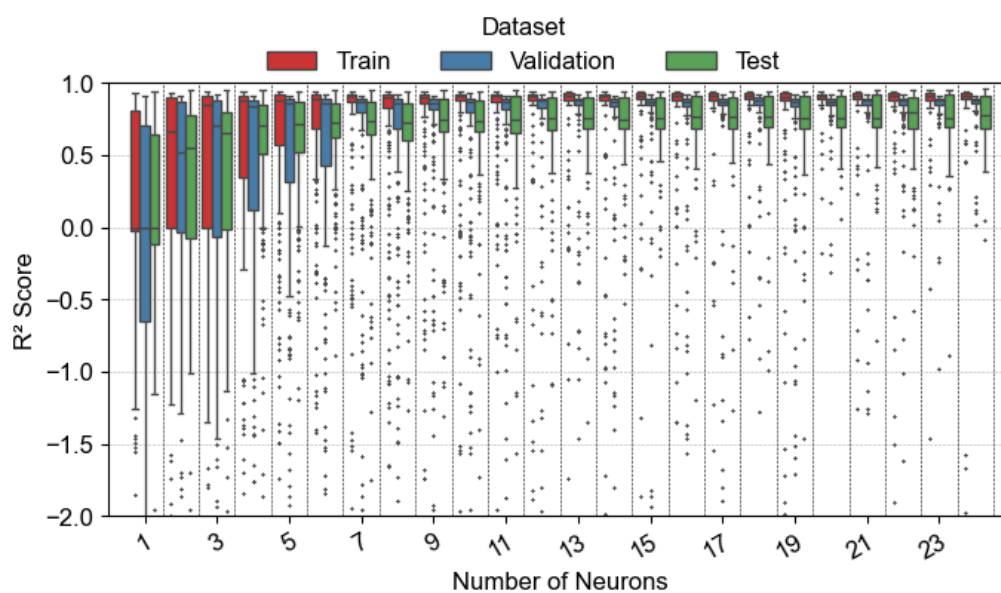


Figure B.38- Combined R^2 score distributions for well 293-3, DMS scenario.

DSM-FFNN model plateau overall graph for W1:

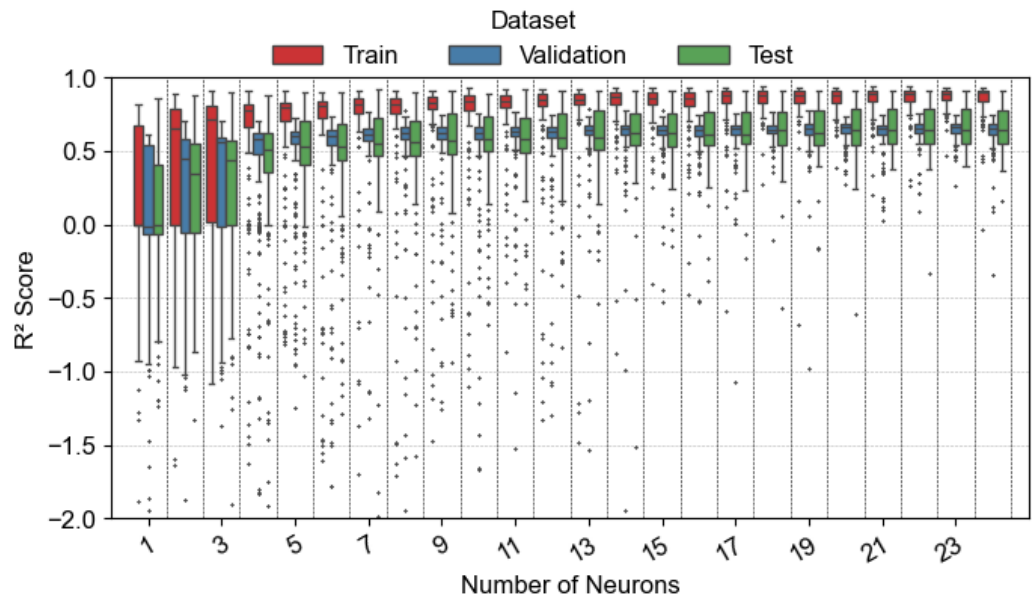


Figure B.39- Combined R^2 score distributions for W1 combined, DSM scenario.

DMS-FFNN model plateau overall graph for W1:

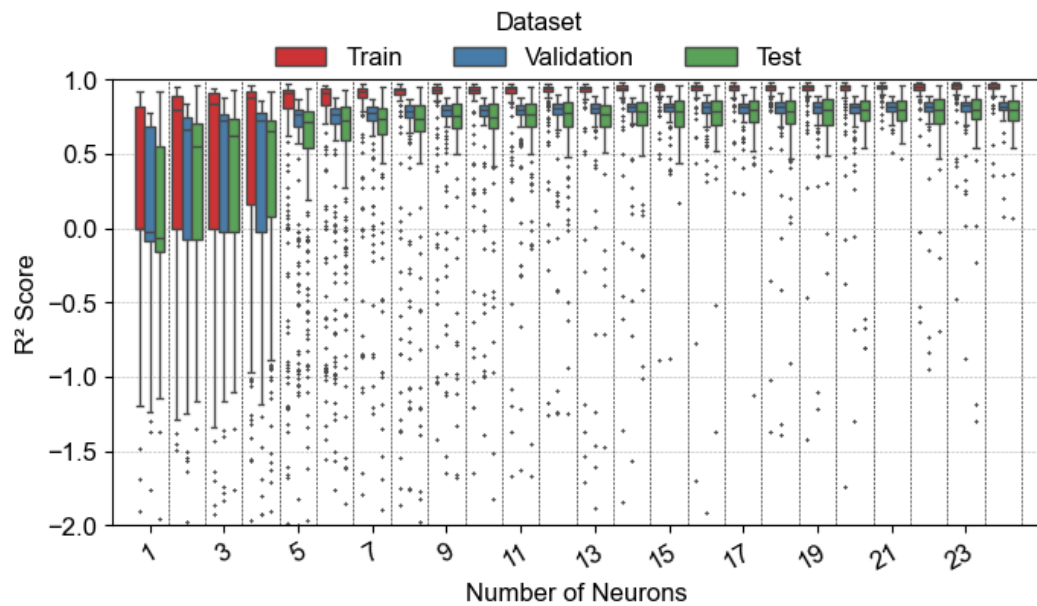


Figure B.40- Combined R^2 score distributions for W1 combined, DMS scenario.

DSM-FFNN model plateau overall graph for 439-1:

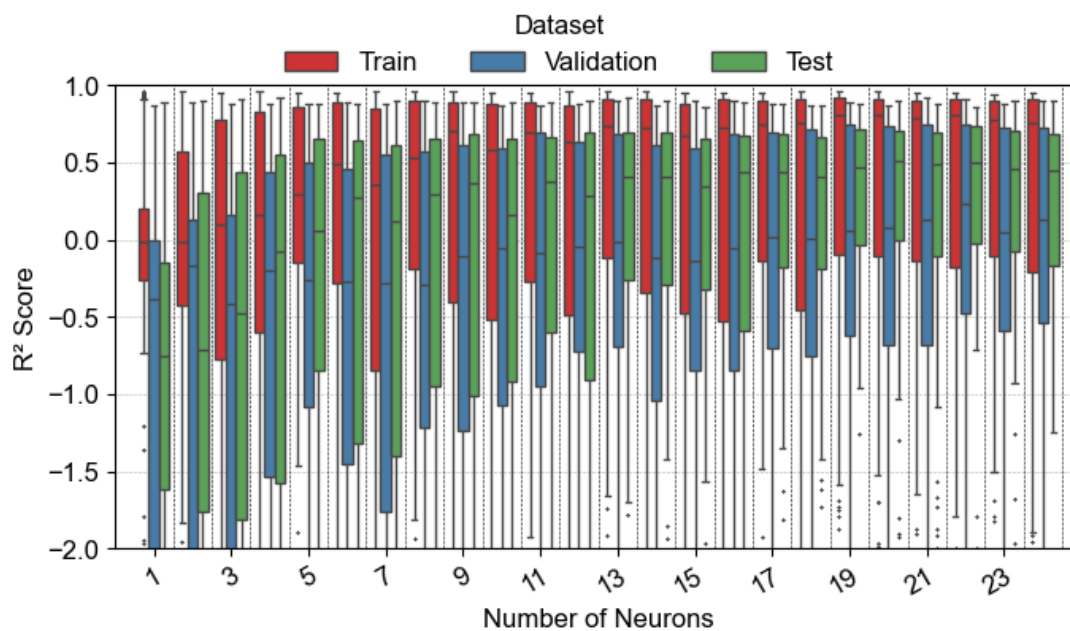


Figure B.41- Combined R^2 score distributions for well 439-1, DSM scenario.

DMS-FFNN model plateau overall graph for 439-1:

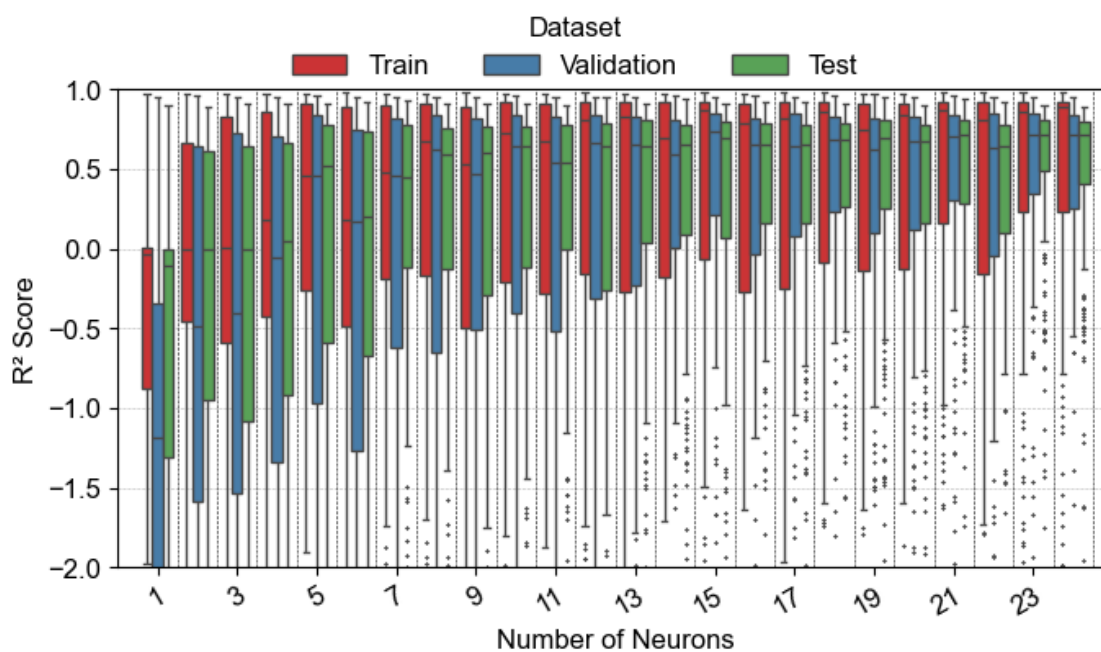


Figure B.42- Combined R^2 score distributions for well 439-1, DMS scenario.

DSM-FFNN model plateau overall graph for 440-1:

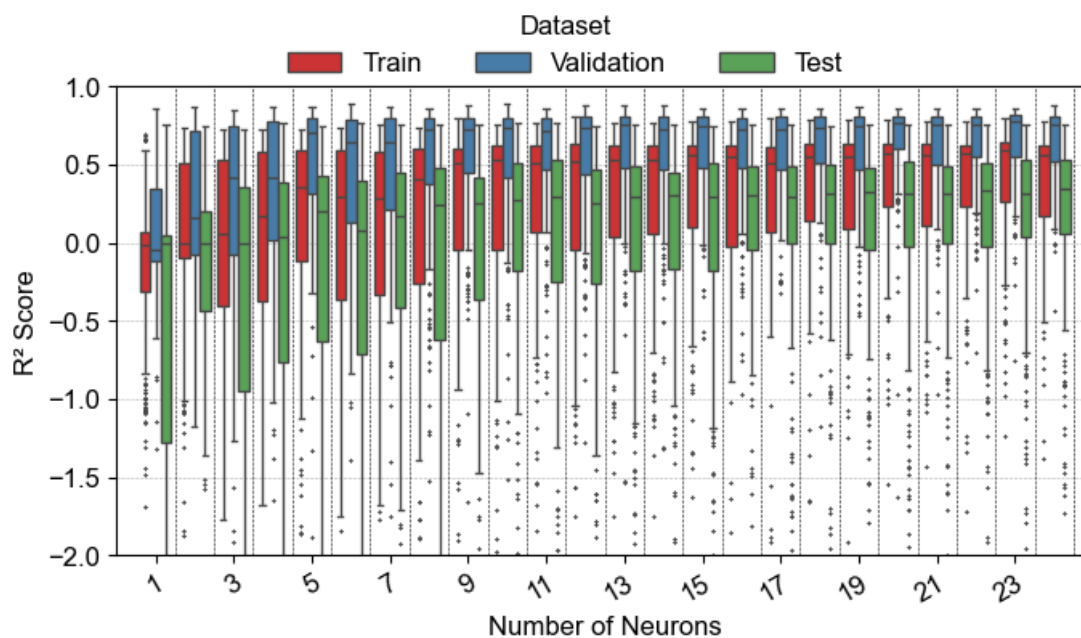


Figure B.43- Combined R^2 score distributions for well 440-1, DSM scenario.

DMS-FFNN model plateau overall graph for 440-1:

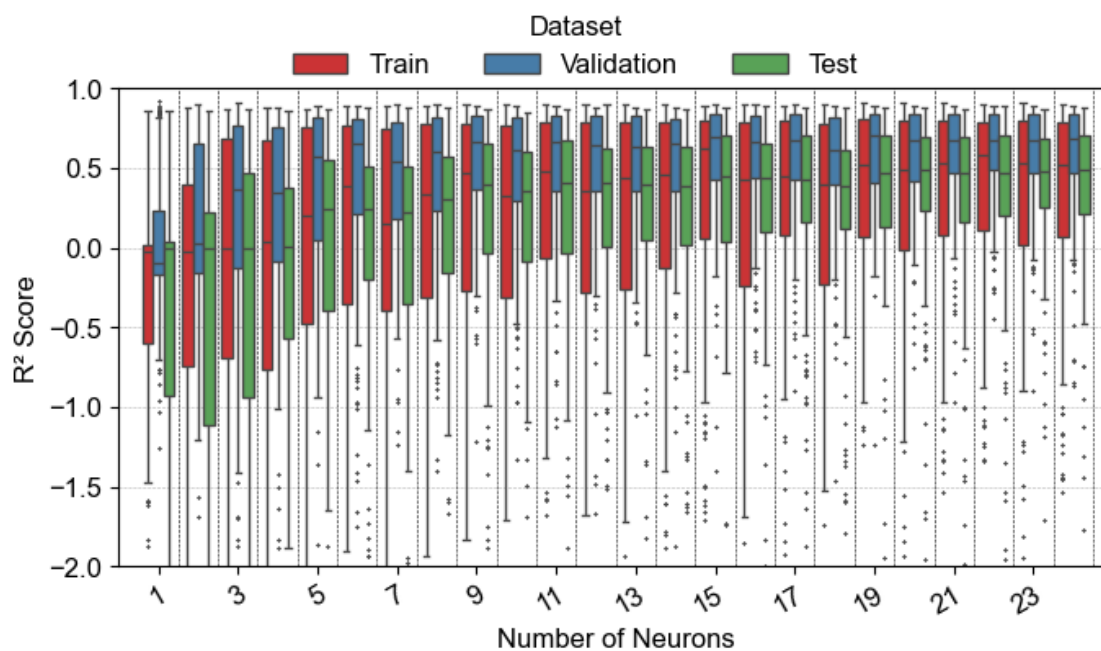


Figure B.44- Combined R^2 score distributions for well 440-1, DMS scenario.

DSM-FFNN model plateau overall graph for 442-1:

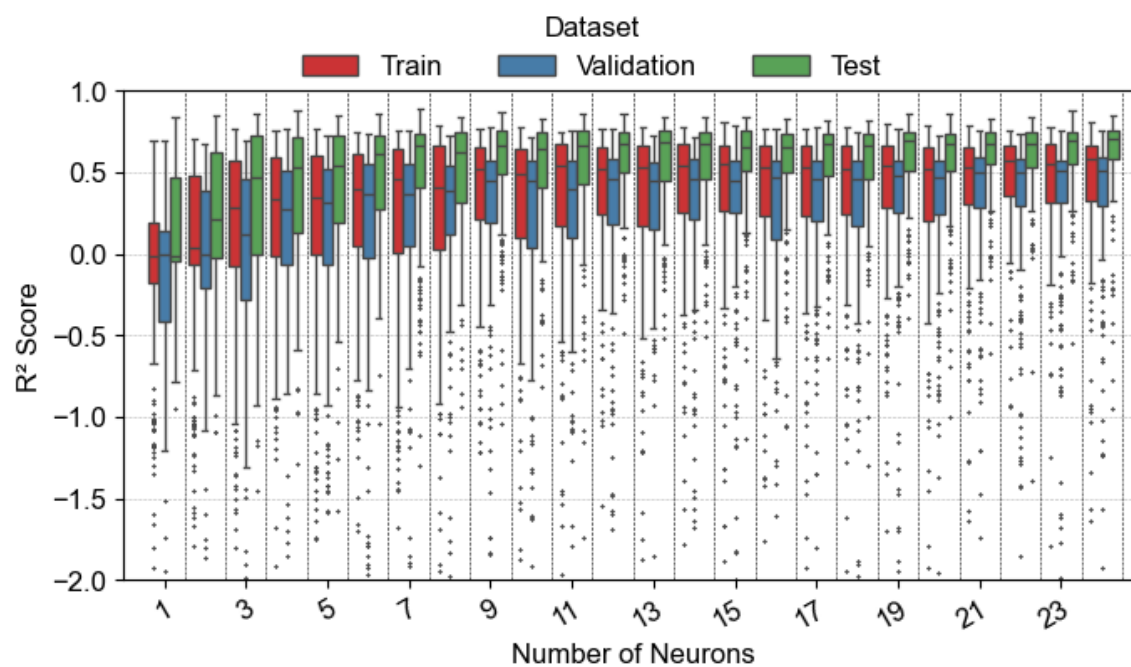


Figure B.45- Combined R^2 score distributions for well 442-1, DSM scenario.

DMS-FFNN model plateau overall graph for 442-1:

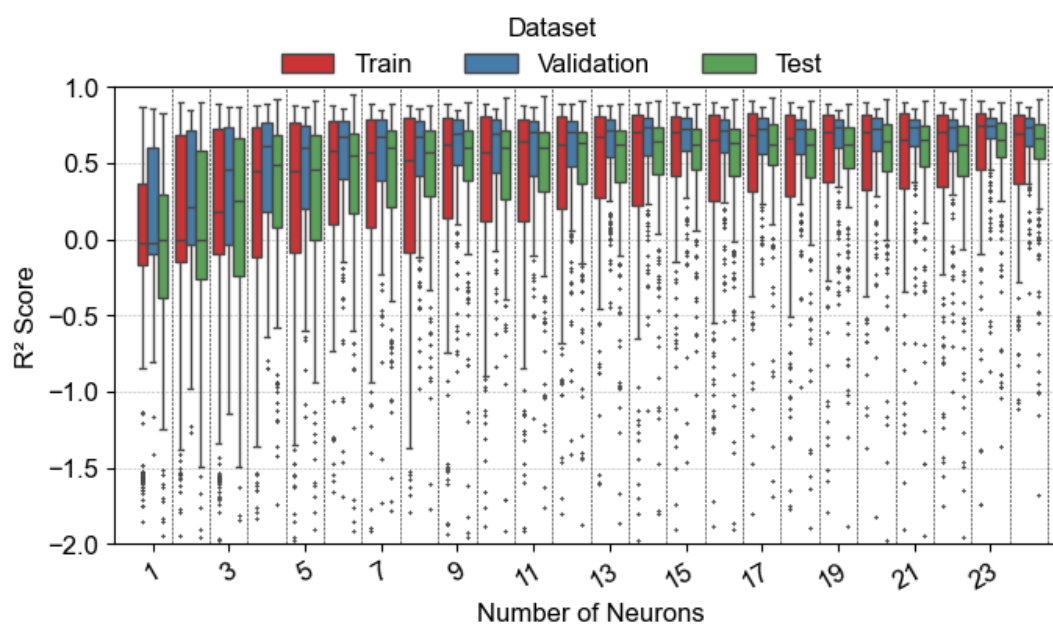


Figure B.46- Combined R^2 score distributions for well 442-1, DMS scenario.

DSM-FFNN model plateau overall graph for W2:

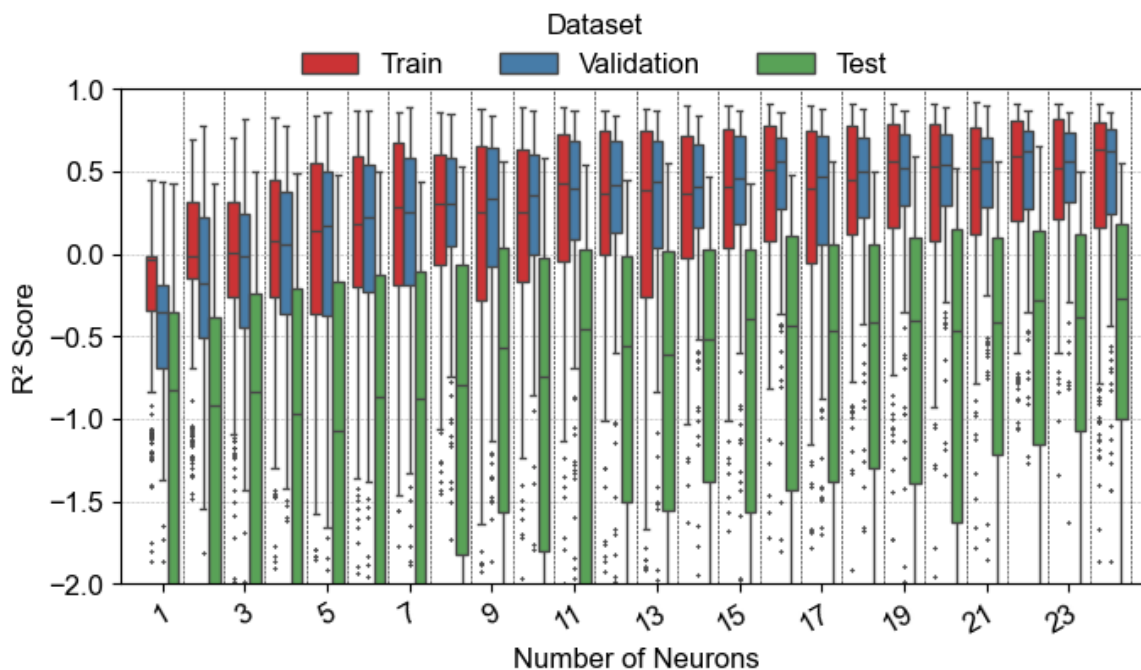


Figure B.47- Combined R^2 score distributions for W2 combined, DSM scenario.

DMS-FFNN model plateau overall graph for W2:

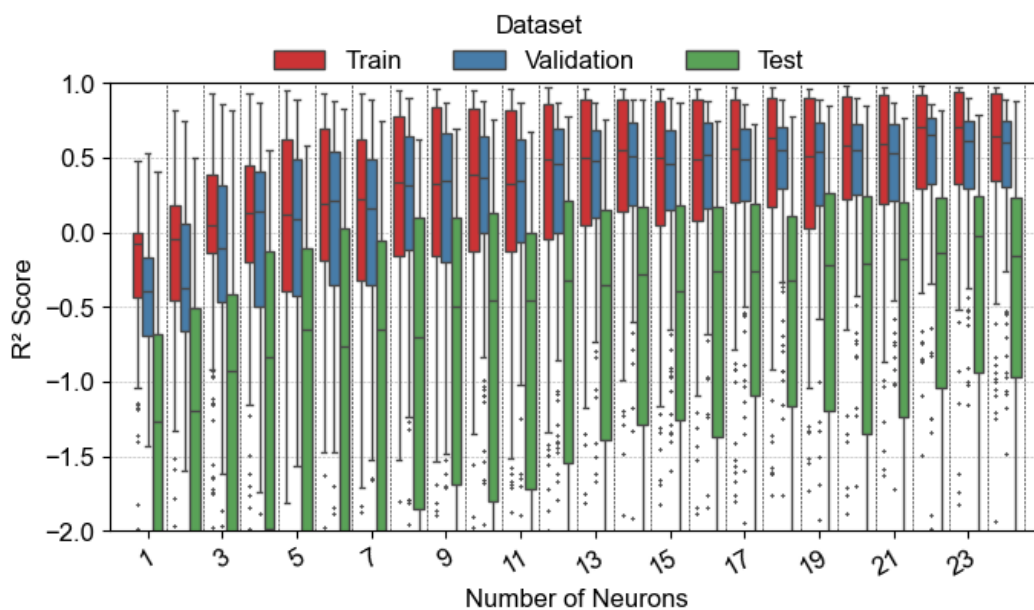


Figure B.48- Combined R^2 score distributions for W2 combined, DMS scenario.

DSM-FFNN model plateau overall graph for 244-2:

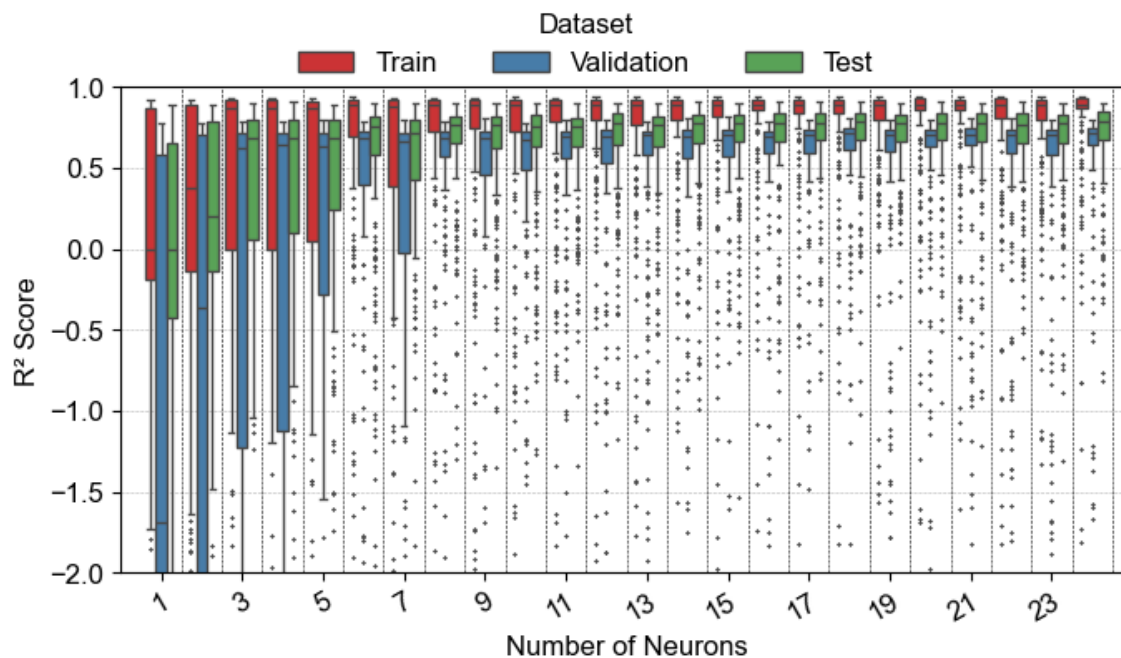


Figure B.49- Combined R^2 score distributions for well 244-2, DSM scenario.

DMS-FFNN model plateau overall graph for 244-2:

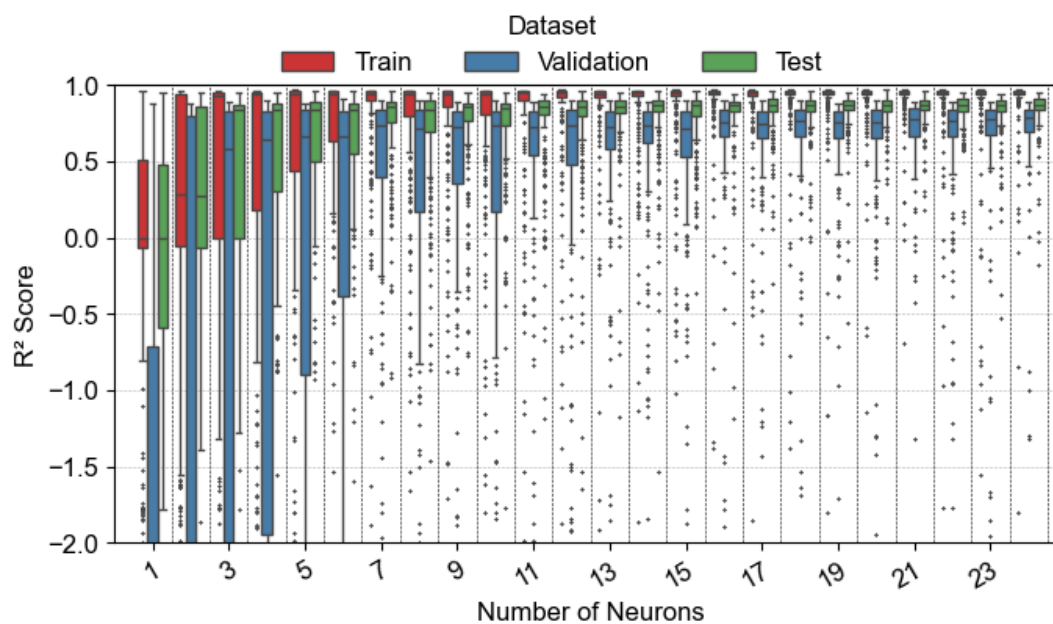


Figure B.50- Combined R^2 score distributions for well 244-2, DMS scenario.

DSM-FFNN model plateau overall graph for 245-2:

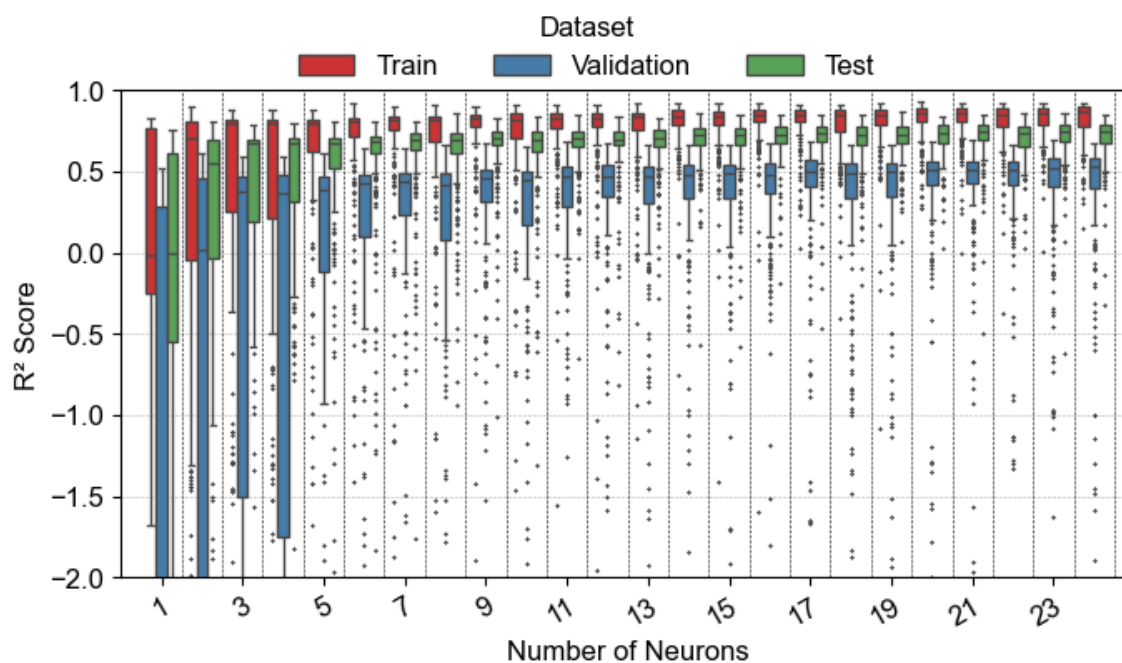


Figure B.51- Combined R^2 score distributions for well 245-2, DSM scenario.

DMS-FFNN model plateau overall graph for 245-2:

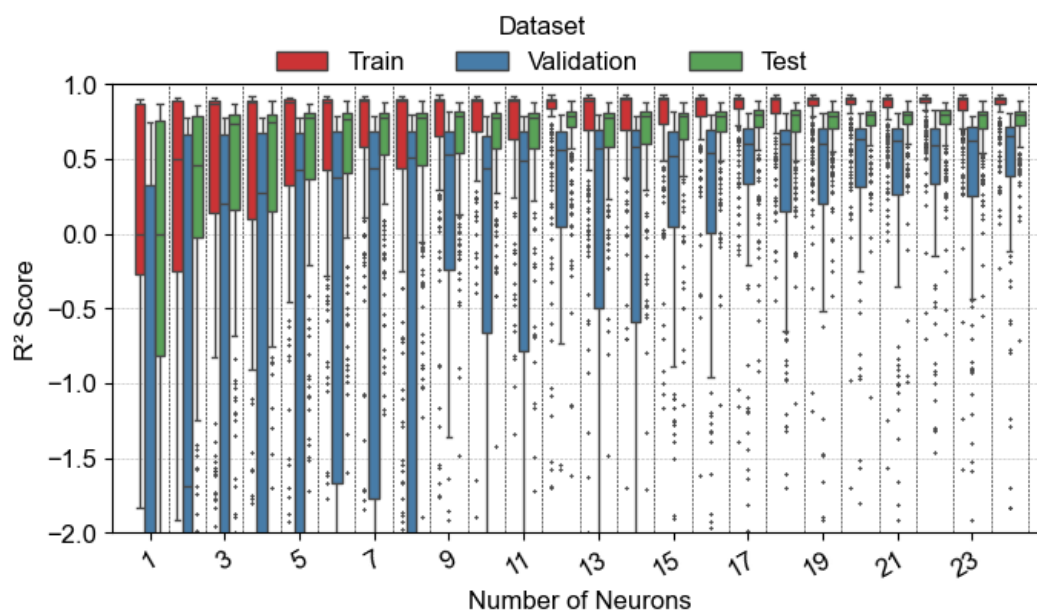


Figure B.52- Combined R^2 score distributions for well 245-2, DMS scenario.

DSM-FFNN model plateau overall graph for W3:

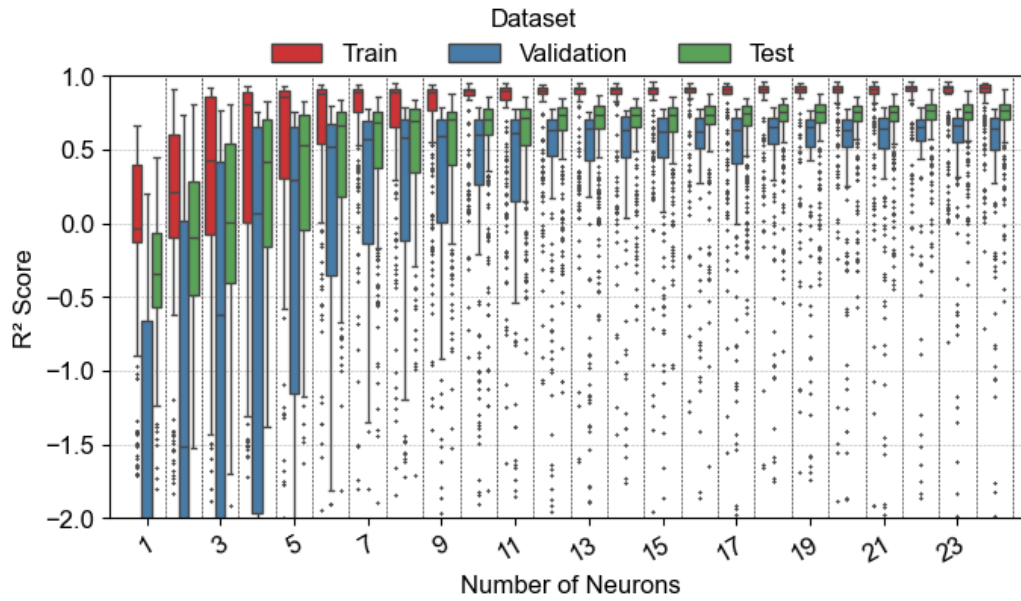


Figure B.53- Combined R^2 score distributions for W3 combined, DSM scenario.

DMS-FFNN model plateau overall graph for W3:

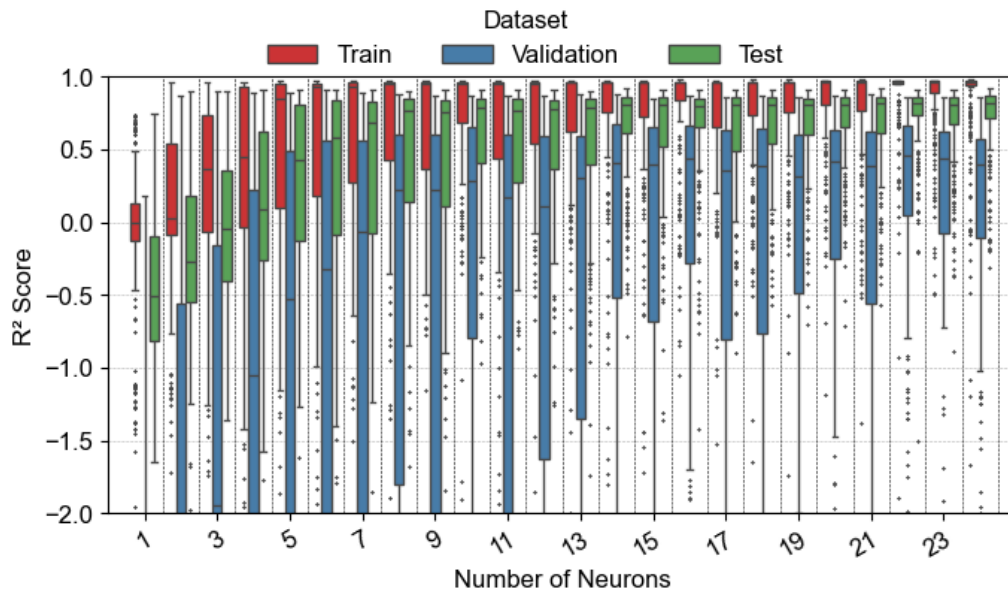


Figure B.54- Combined R^2 score distributions for W3 combined, DMS scenario.

DSM-FFNN model plateau overall graph for 443-1:

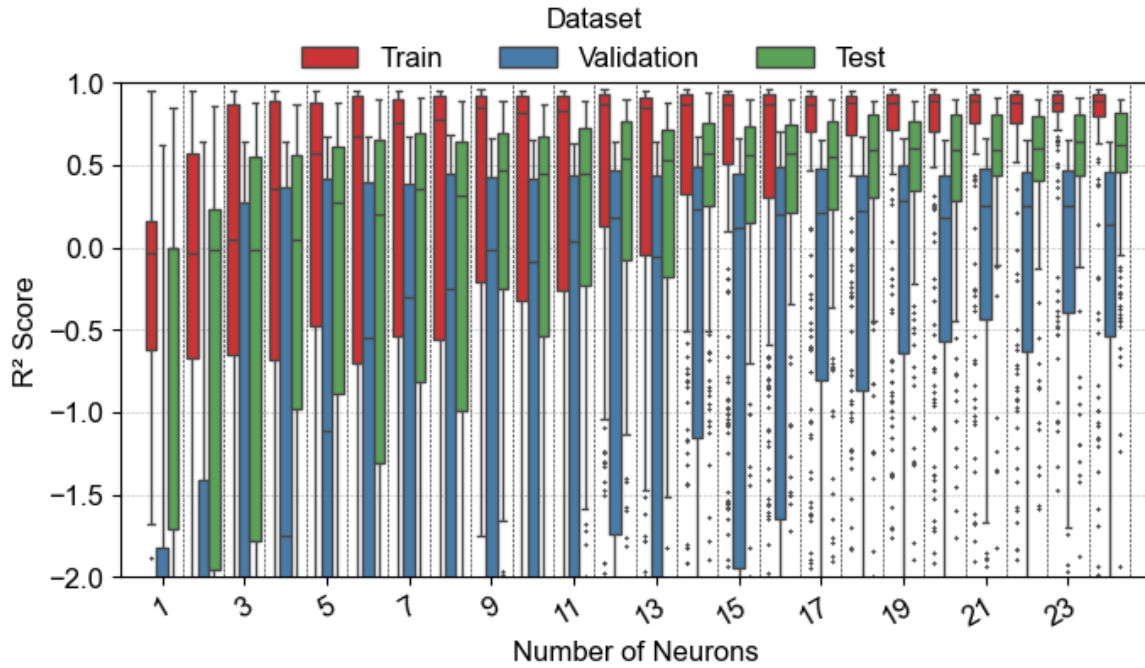


Figure B.55- Combined R^2 score distributions for well 443-1, DSM scenario.

DMS-FFNN model plateau overall graph for 443-1:

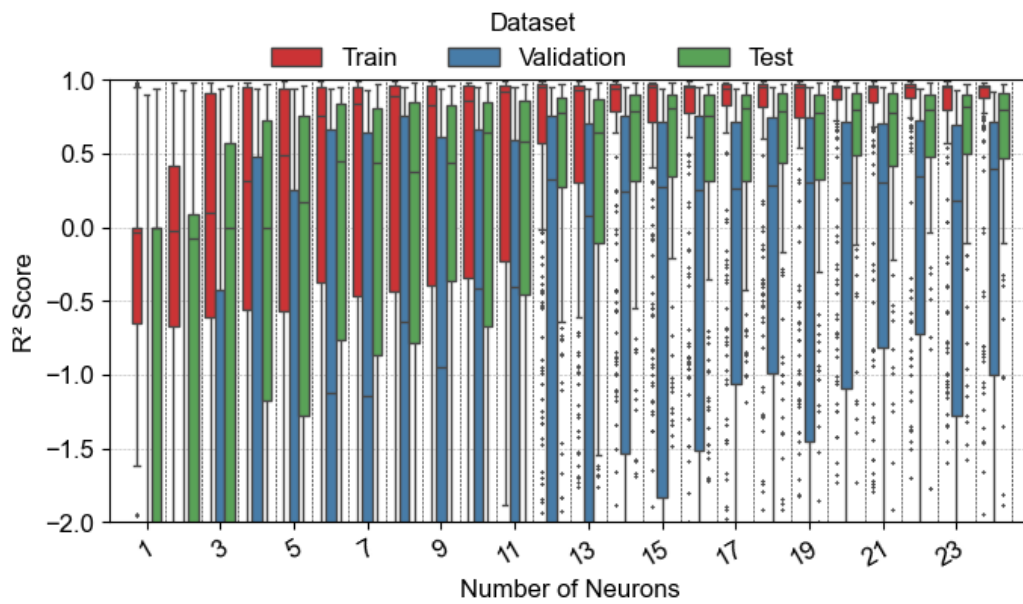


Figure B.56- Combined R^2 score distributions for well 443-1, DMS scenario.

DSM-FFNN model plateau overall graph for the aggregated model:

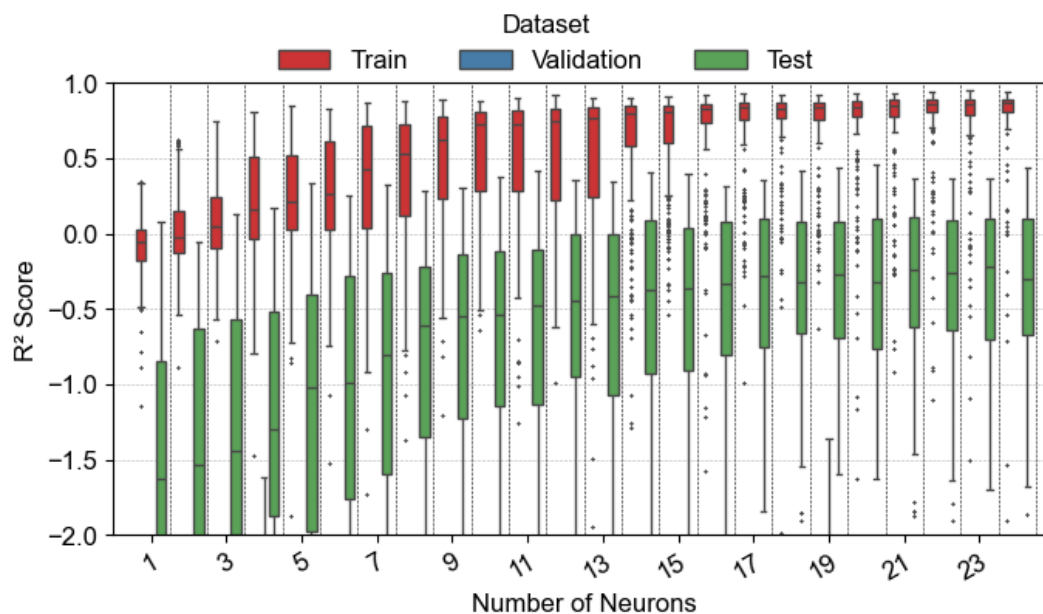


Figure B.57- Combined R^2 score distributions for the aggregated model, DSM scenario.

DMS-FFNN model plateau overall graph for the aggregated model:

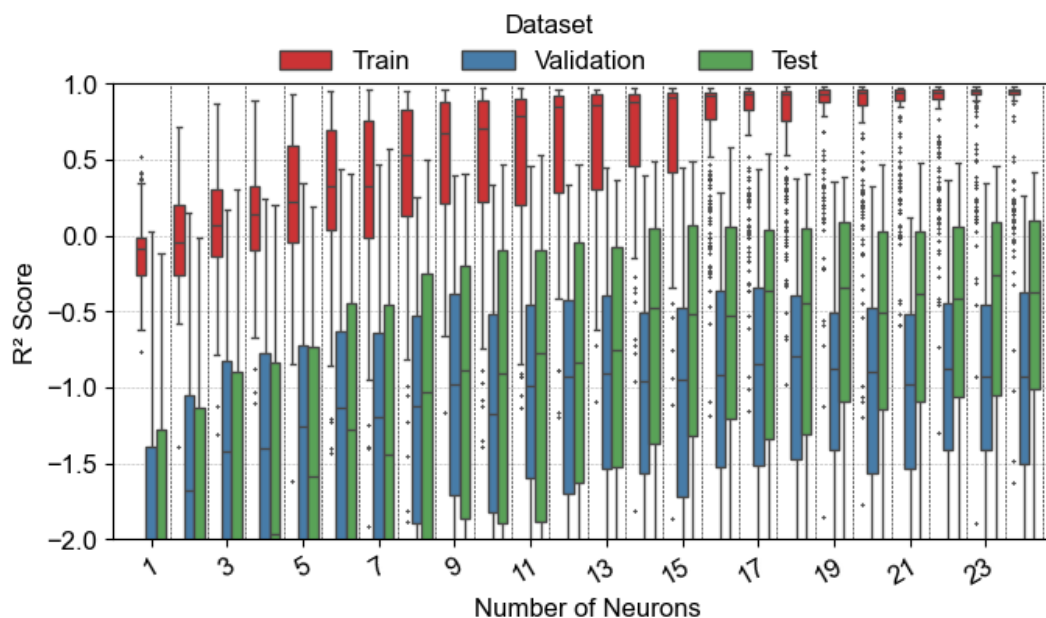


Figure B.58- Combined R^2 score distributions for the aggregated model, DMS scenario.

Performance Plateau By-Fold Plots (MSE)

DMS-FFNN model plateau by fold graph for 293-2:

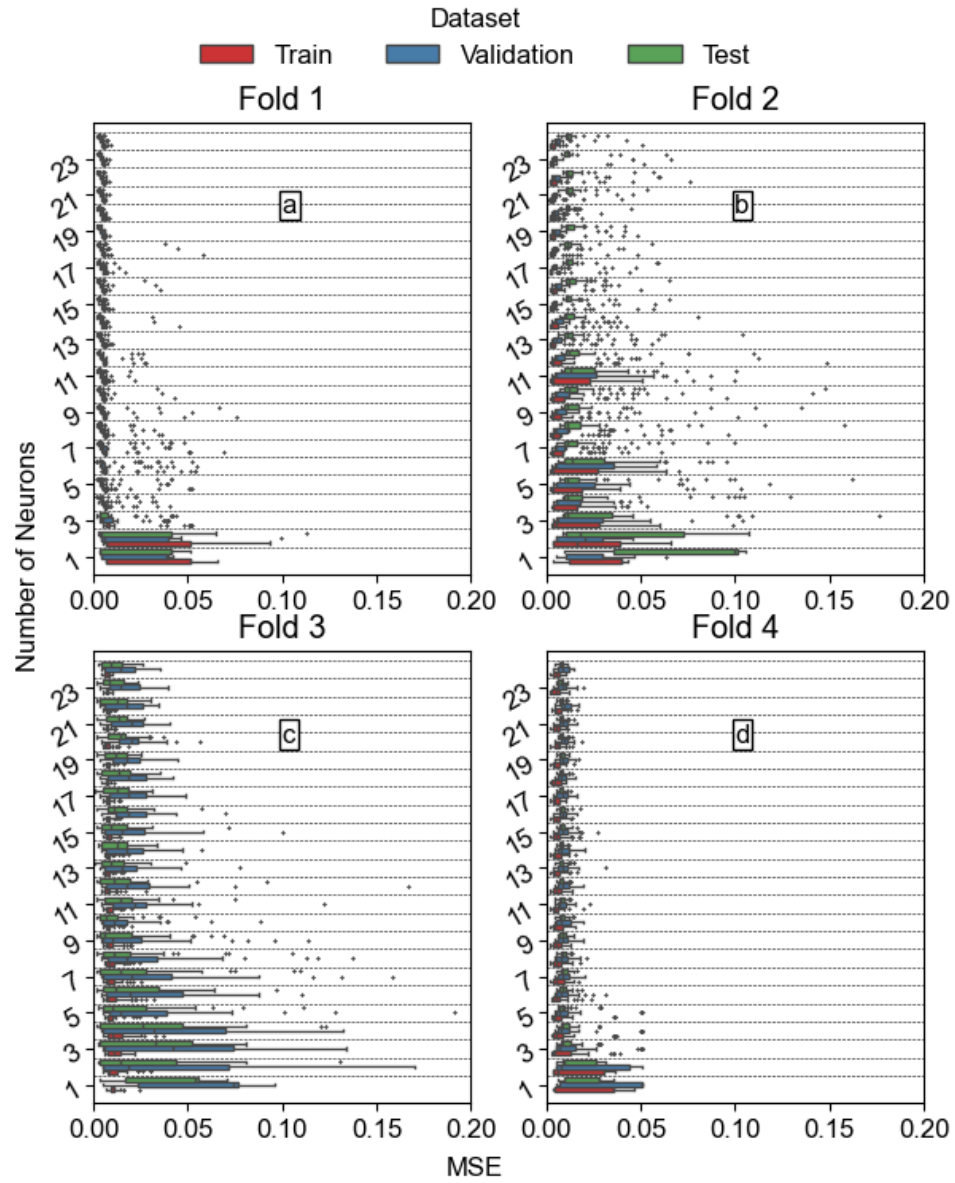


Figure B.59- MSE distributions across different neuron counts for well 293-2 in each cross-validation fold, DMS scenario.

DSM-FFNN model plateau by fold graph for 293-3:

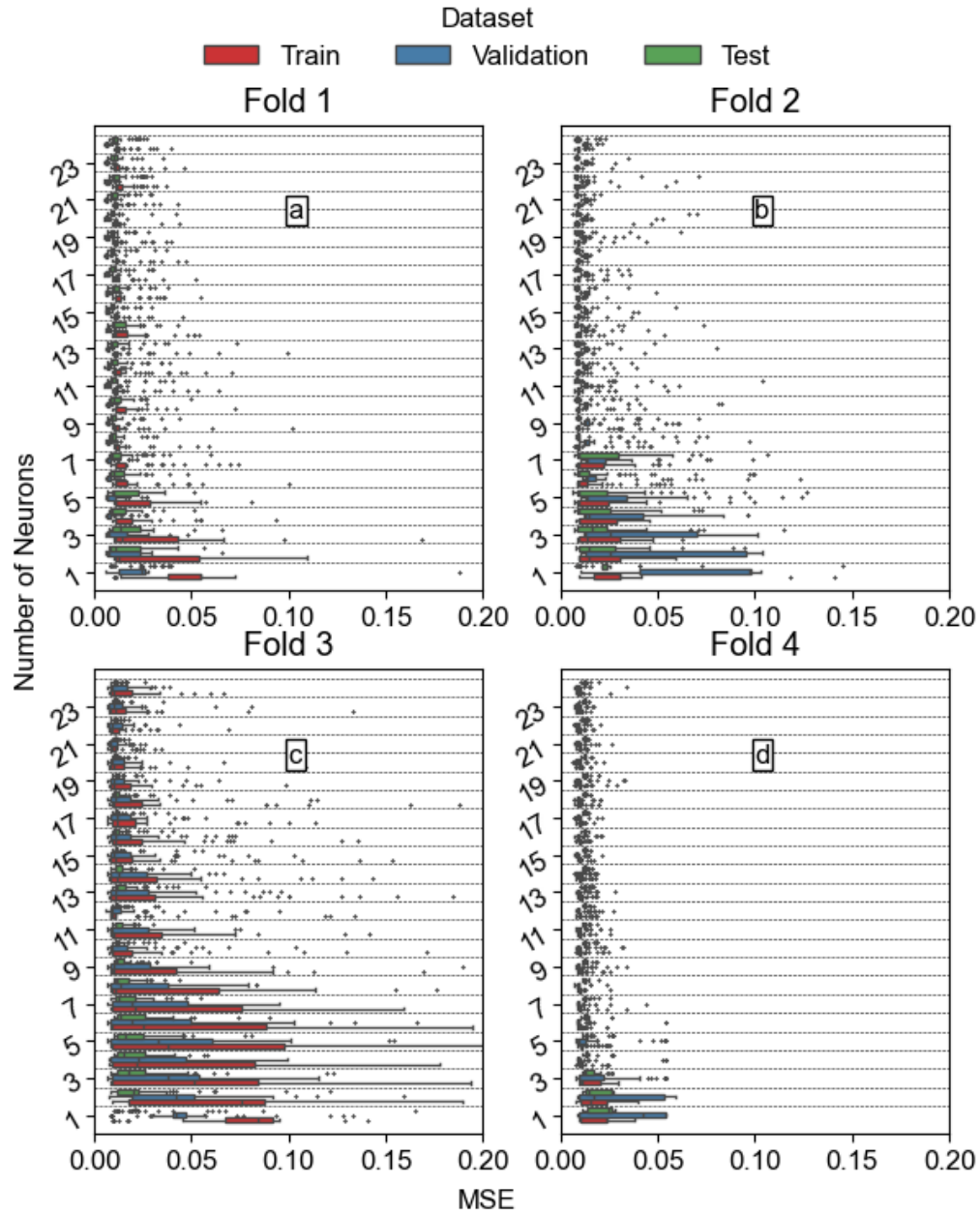


Figure B.60- MSE distributions across different neuron counts for well 293-3 in each cross-validation fold, DSM scenario.

DMS-FFNN model plateau by fold graph for 293-3:

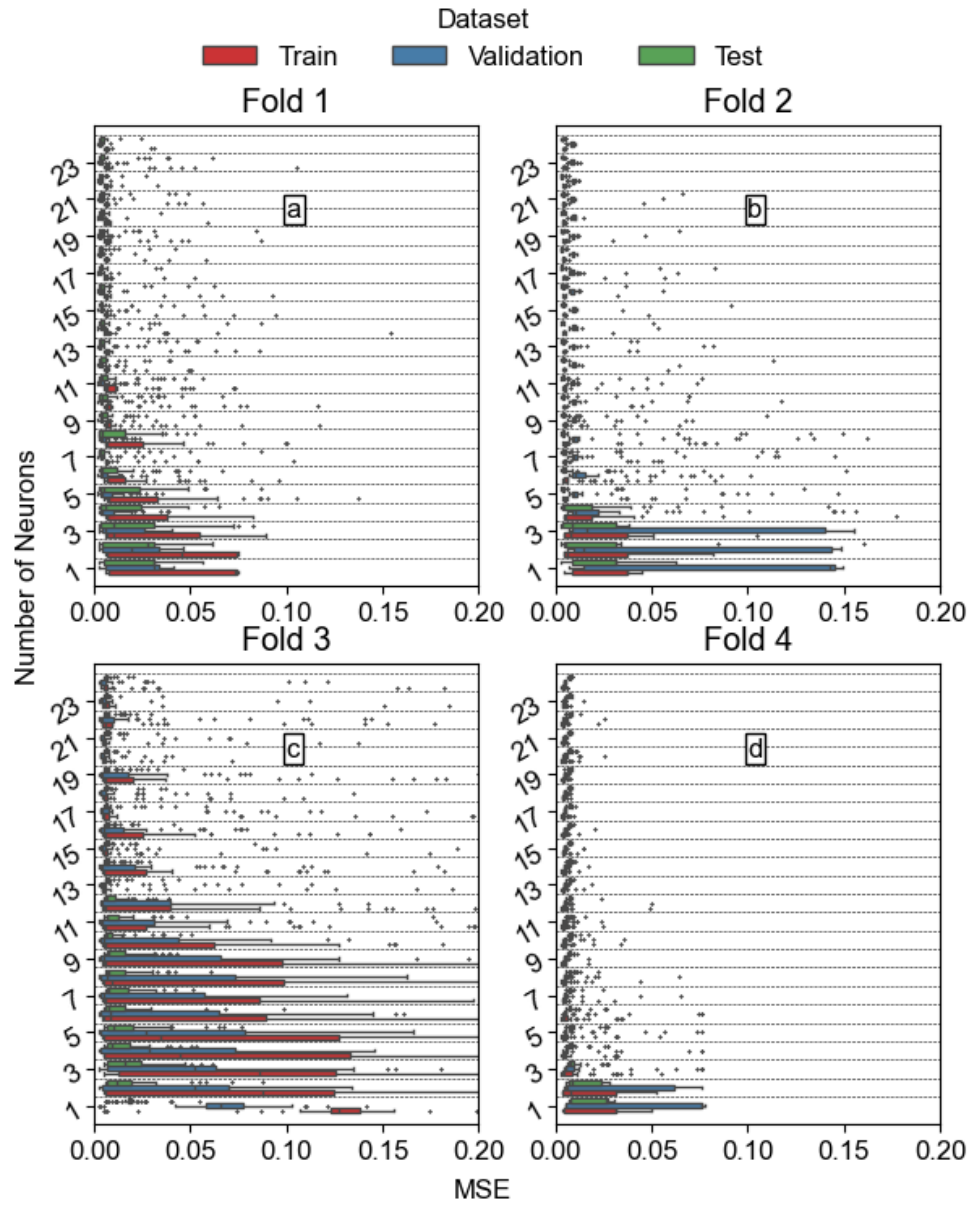


Figure B.61- MSE distributions across different neuron counts for well 293-3 in each cross-validation fold, DMS scenario.

DSM-FFNN model plateau by fold graph for W1:

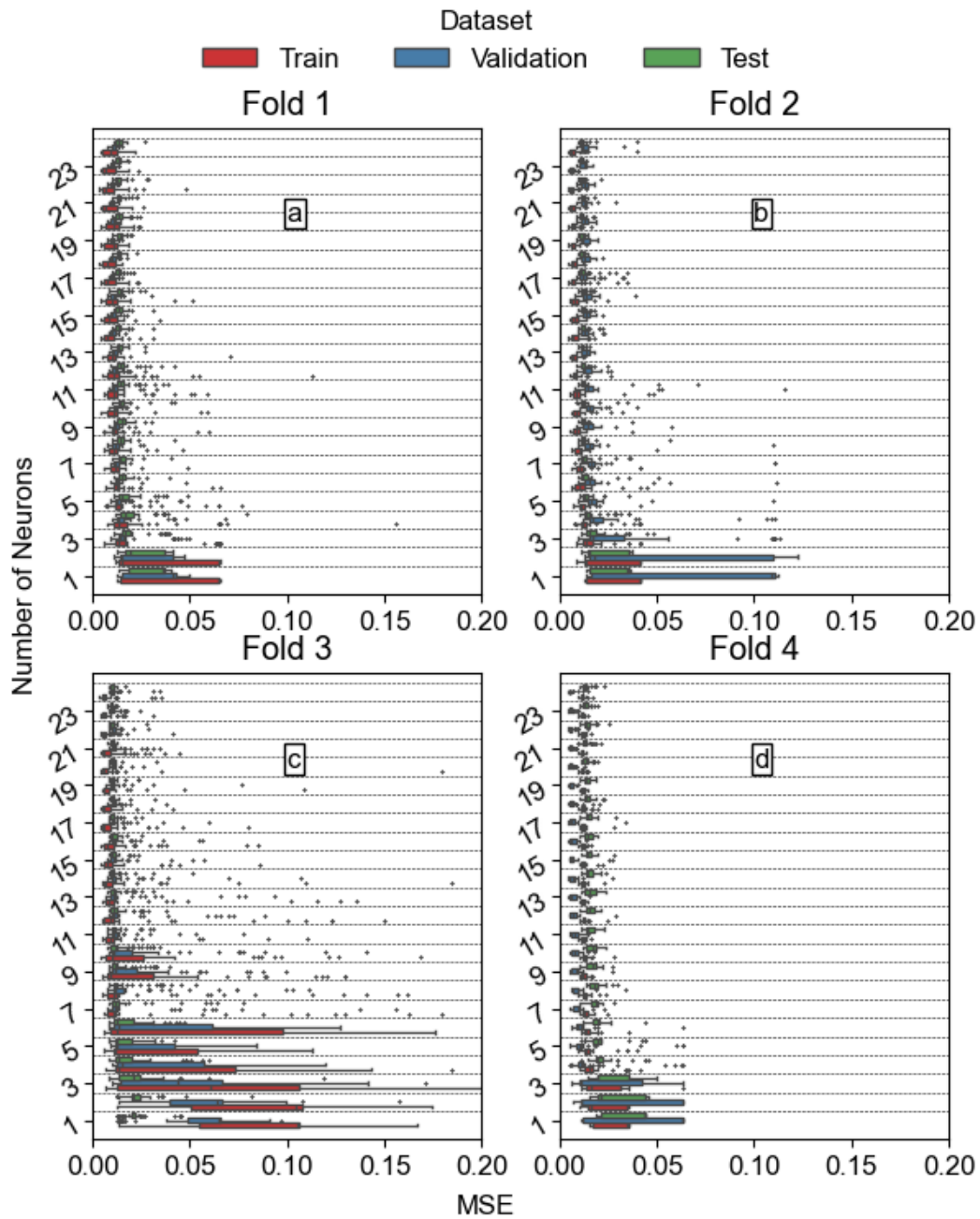


Figure B.62- MSE distributions across different neuron counts for W1 combined in each cross-validation fold, DSM scenario.

DMS-FFNN model plateau by fold graph for W1:

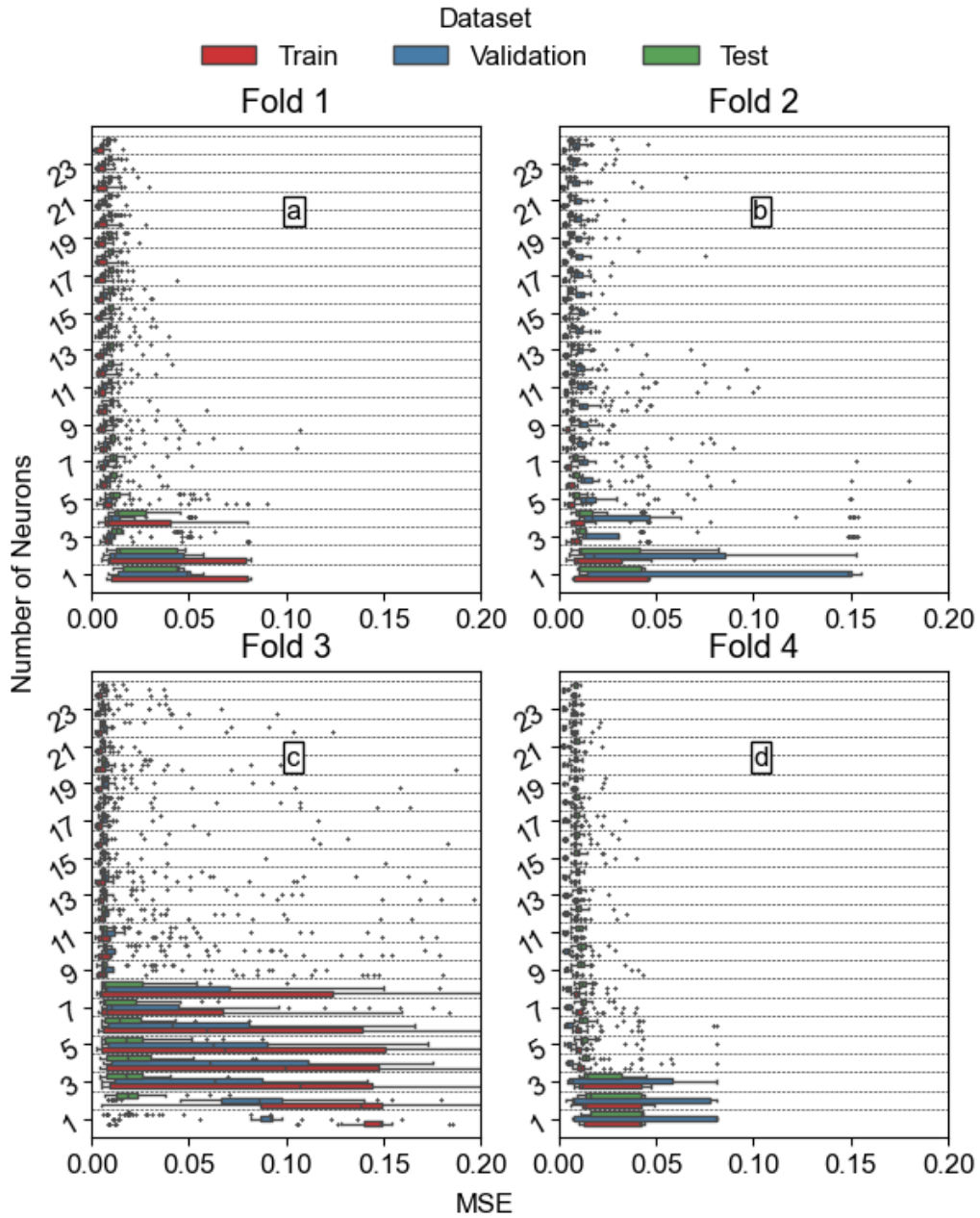


Figure B.63- MSE distributions across different neuron counts for W1 combined in each cross-validation fold, DMS scenario.

DSM-FFNN model plateau by fold graph for 439-1:

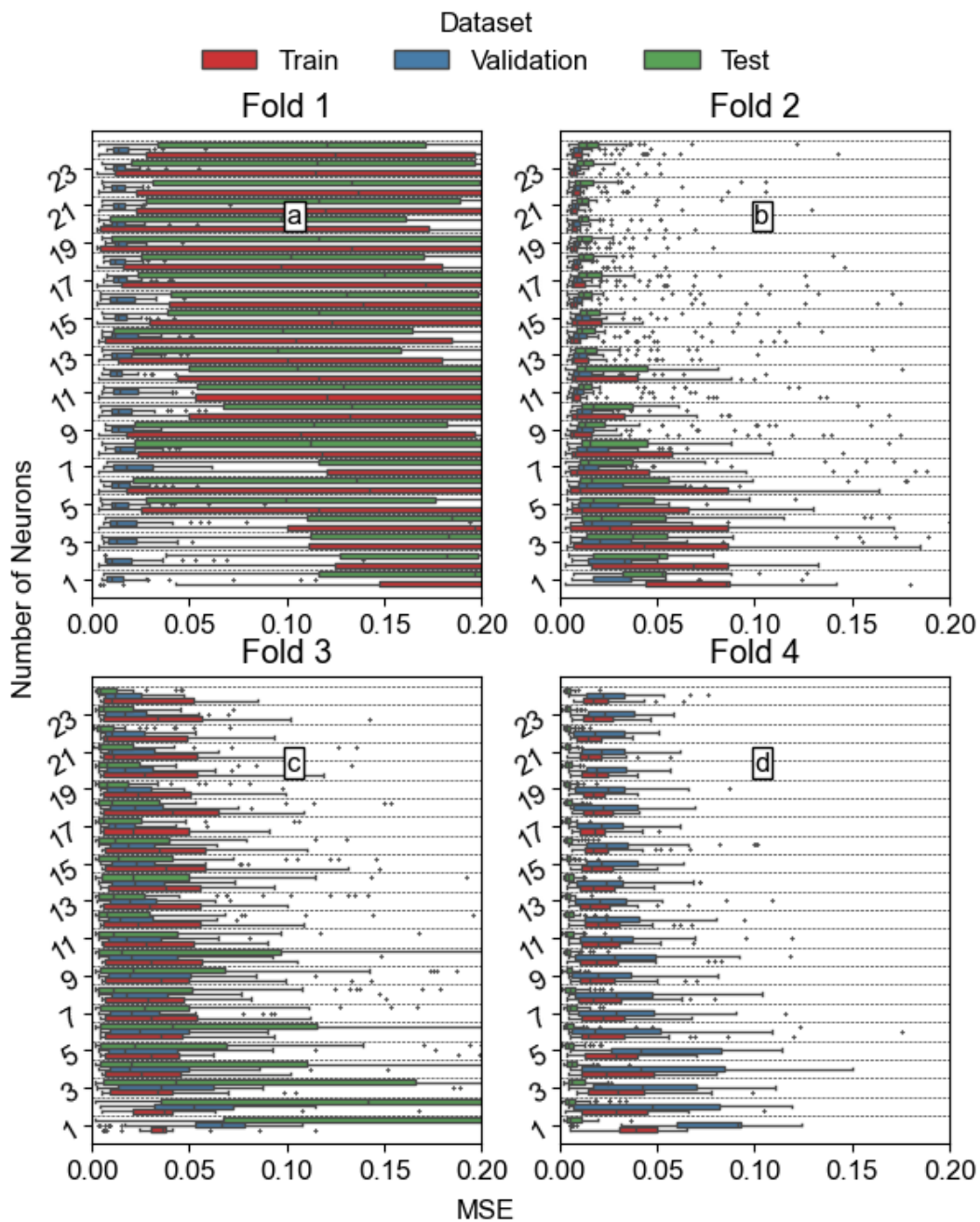


Figure B.64- MSE distributions across different neuron counts for well 439-1 in each cross-validation fold, DSM scenario.

DMS-FFNN model plateau by fold graph for 439-1:

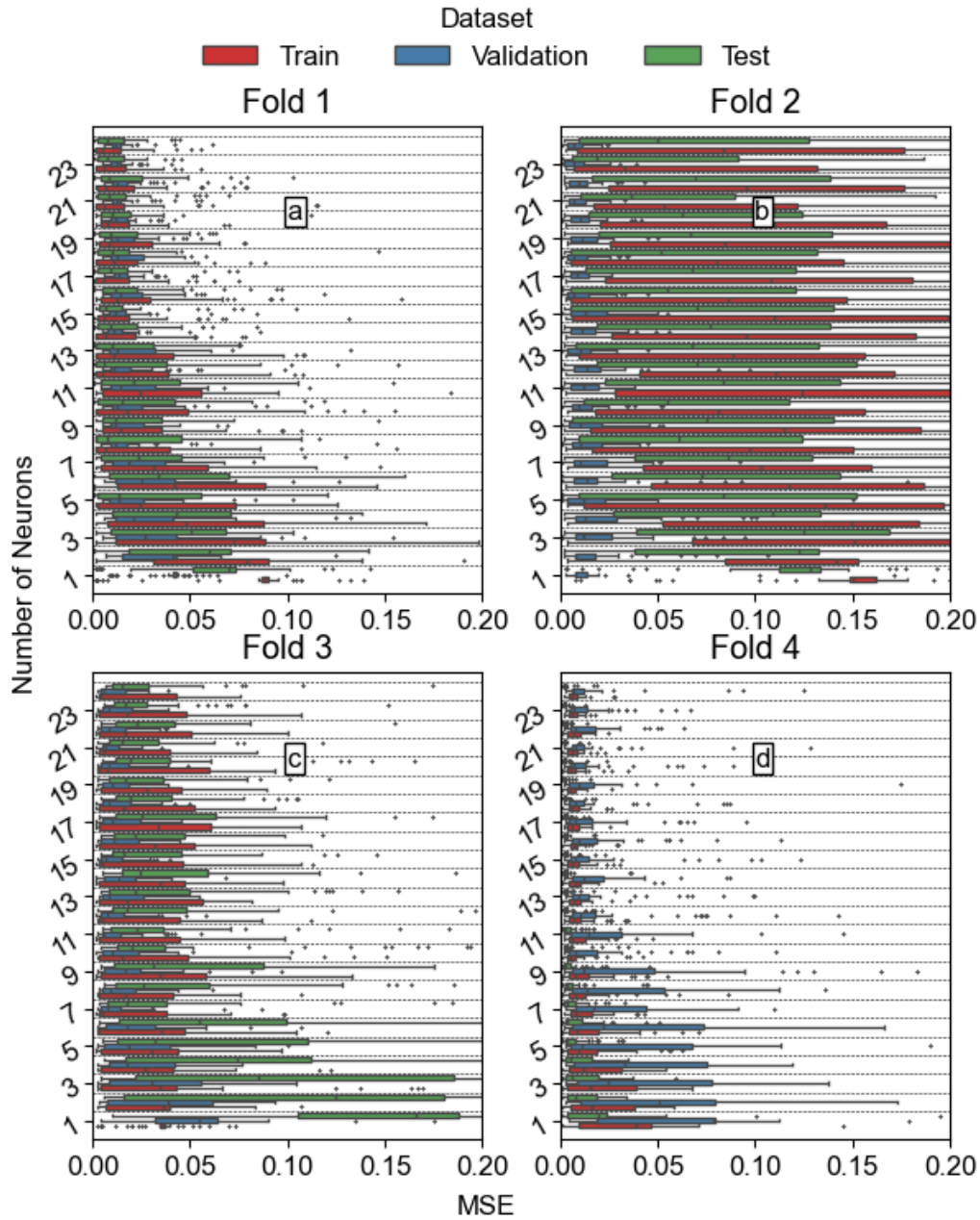


Figure B.65- MSE distributions across different neuron counts for well 439-1 in each cross-validation fold, DMS scenario.

DSM-FFNN model plateau by fold graph for 440-1:

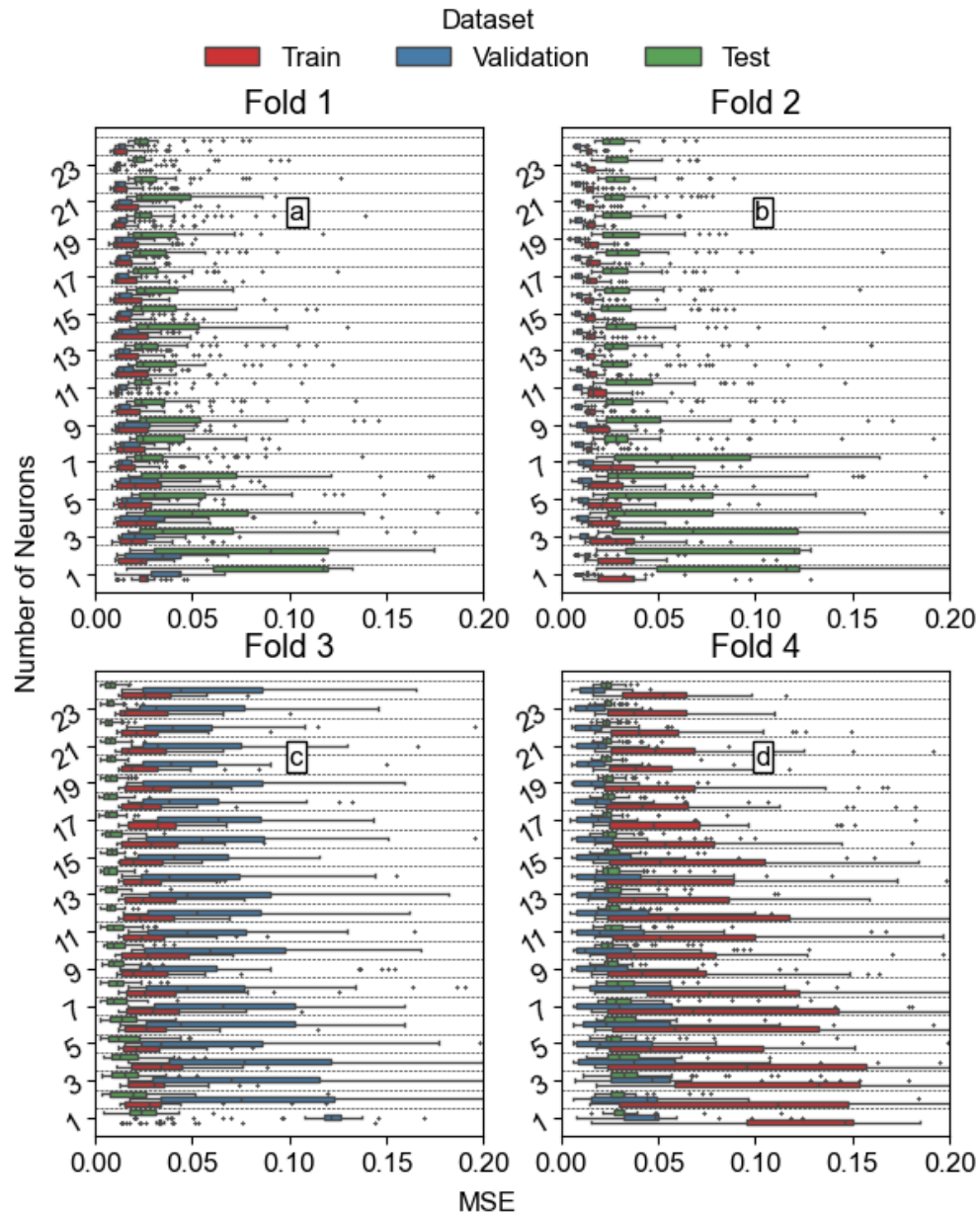


Figure B.66- MSE distributions across different neuron counts for well 440-1 in each cross-validation fold, DSM scenario.

DMS-FFNN model plateau by fold graph for 440-1:

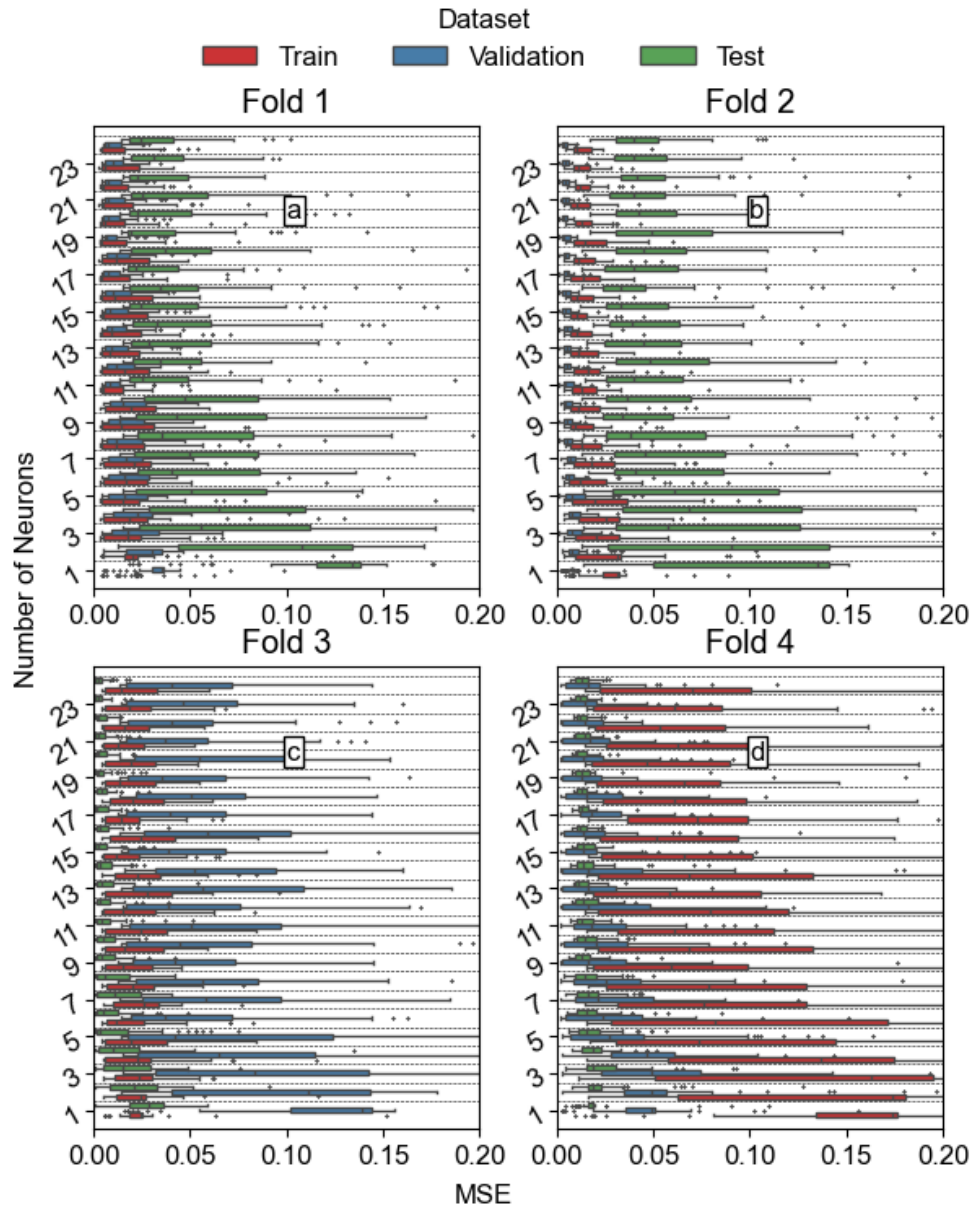


Figure B.67- MSE distributions across different neuron counts for well 440-1 in each cross-validation fold, DMS scenario.

DSM-FFNN model plateau by fold graph for 442-1:

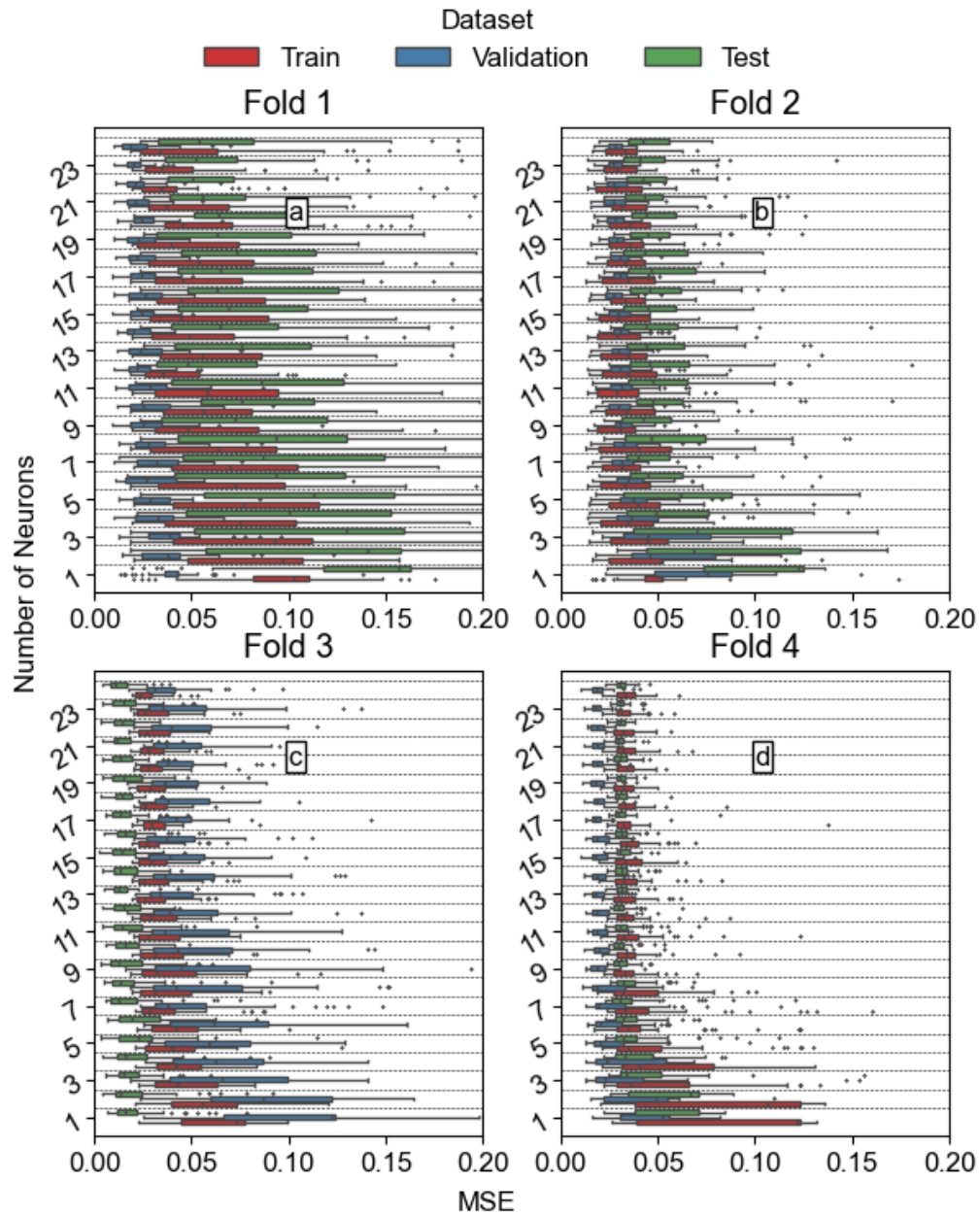


Figure B.68- MSE distributions across different neuron counts for well 442-1 in each cross-validation fold, DSM scenario.

DMS-FFNN model plateau by fold graph for 442-1:

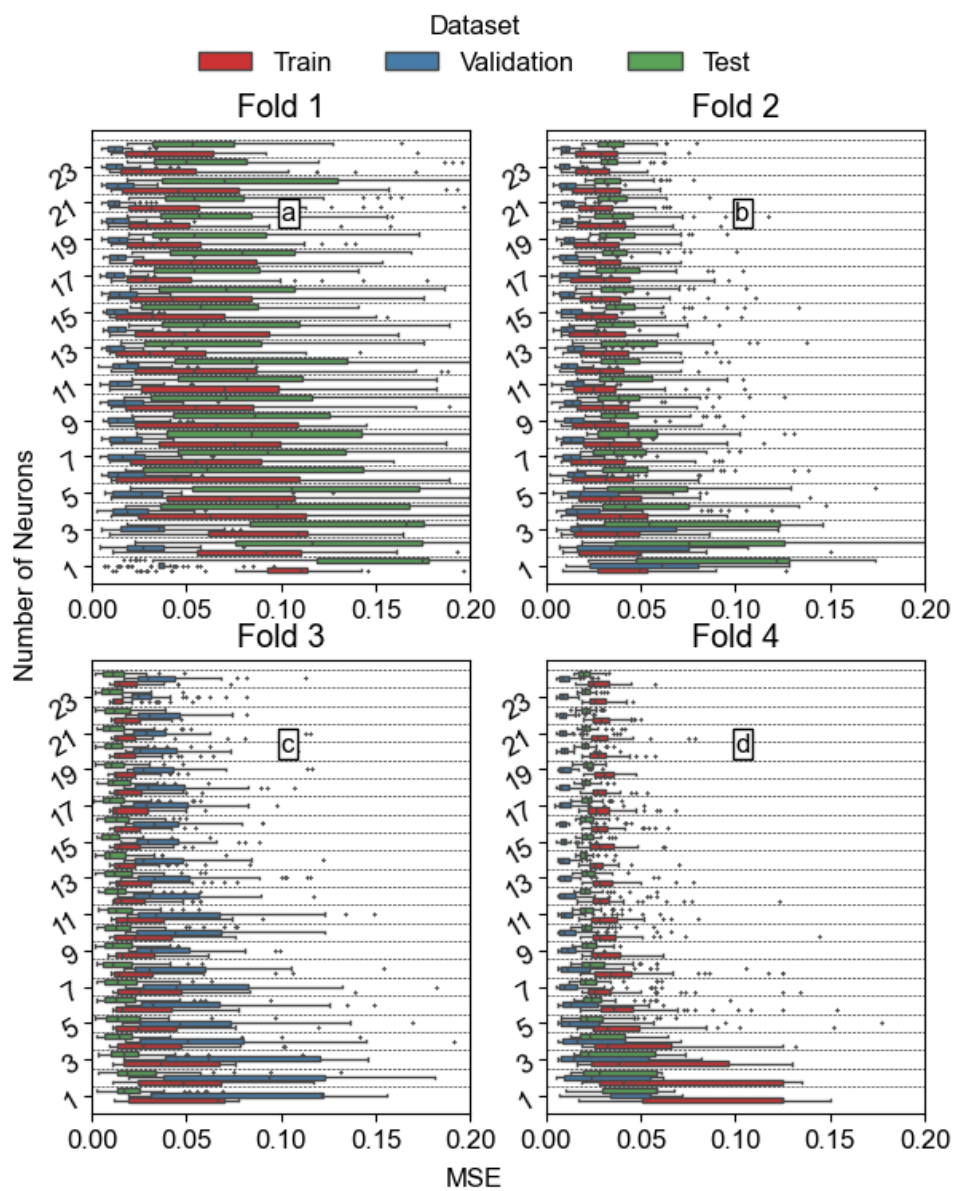


Figure B.69- MSE distributions across different neuron counts for well 442-1 in each cross-validation fold, DMS scenario.

DSM-FFNN model plateau by fold graph for W2:

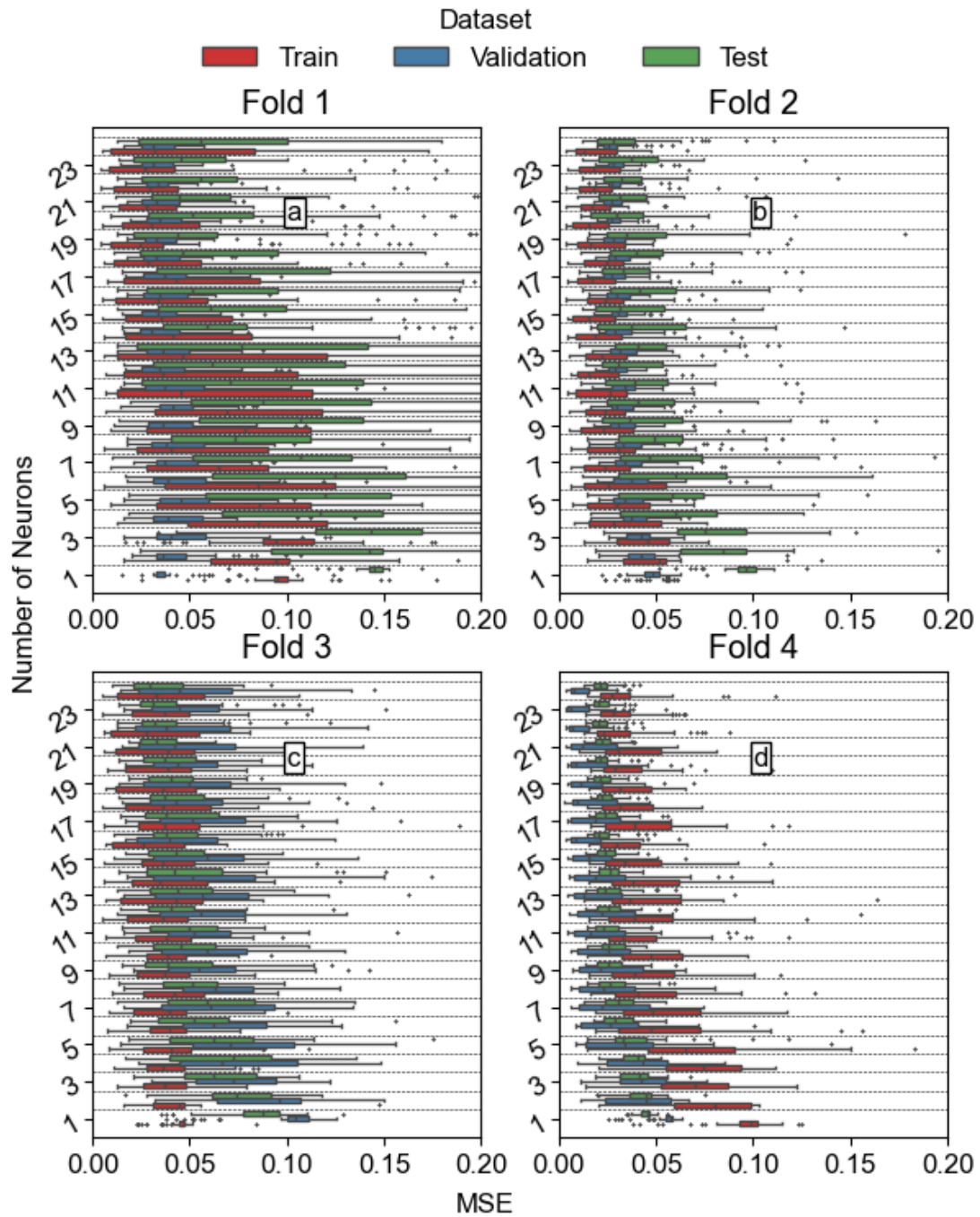


Figure B.70- MSE distributions across different neuron counts for W2 combined in each cross-validation fold, DSM scenario.

DMS-FFNN model plateau by fold graph for W2:

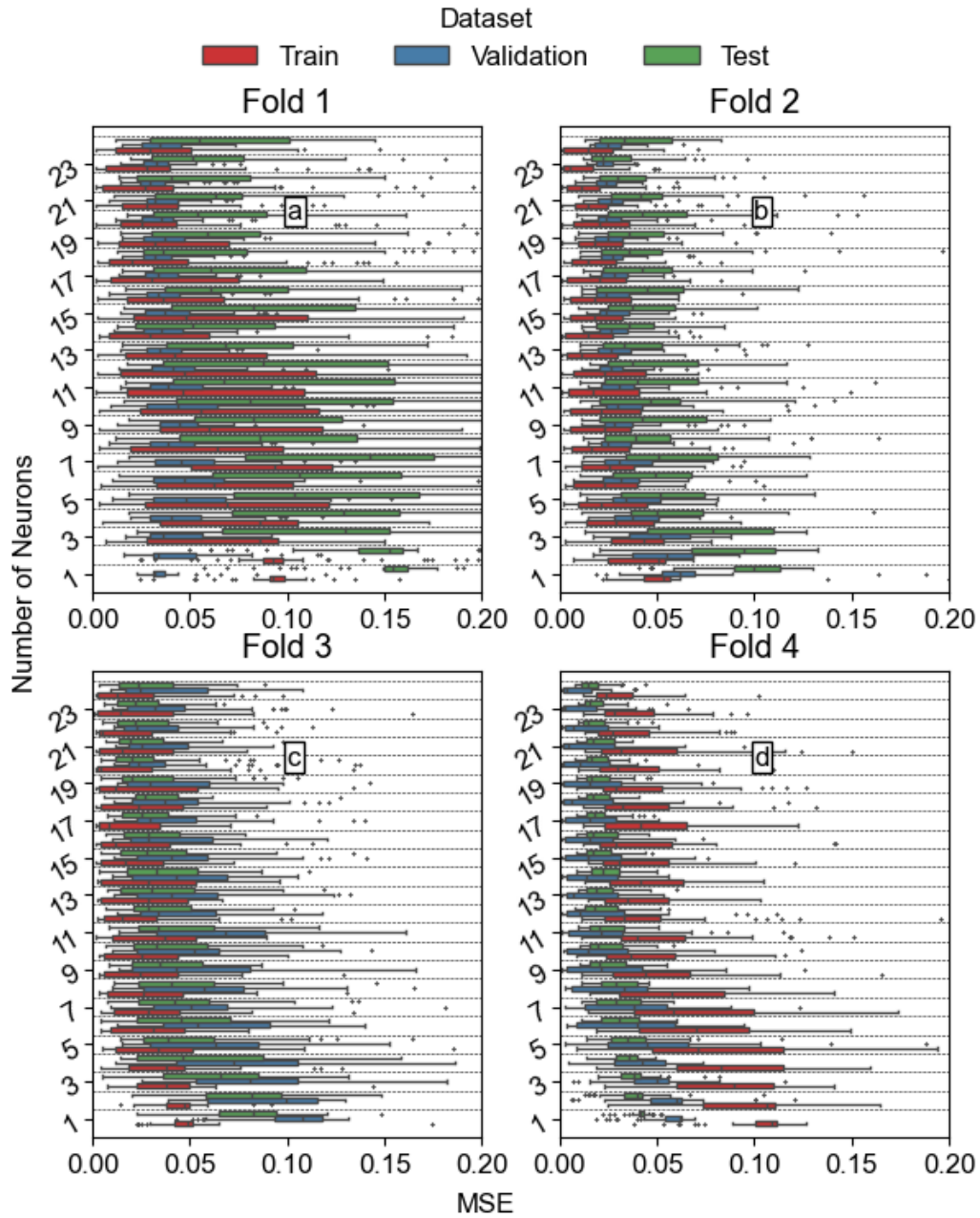


Figure B.71- MSE distributions across different neuron counts for W2 combined in each cross-validation fold, DMS scenario.

DSM-FFNN model plateau by fold graph for 244-2:

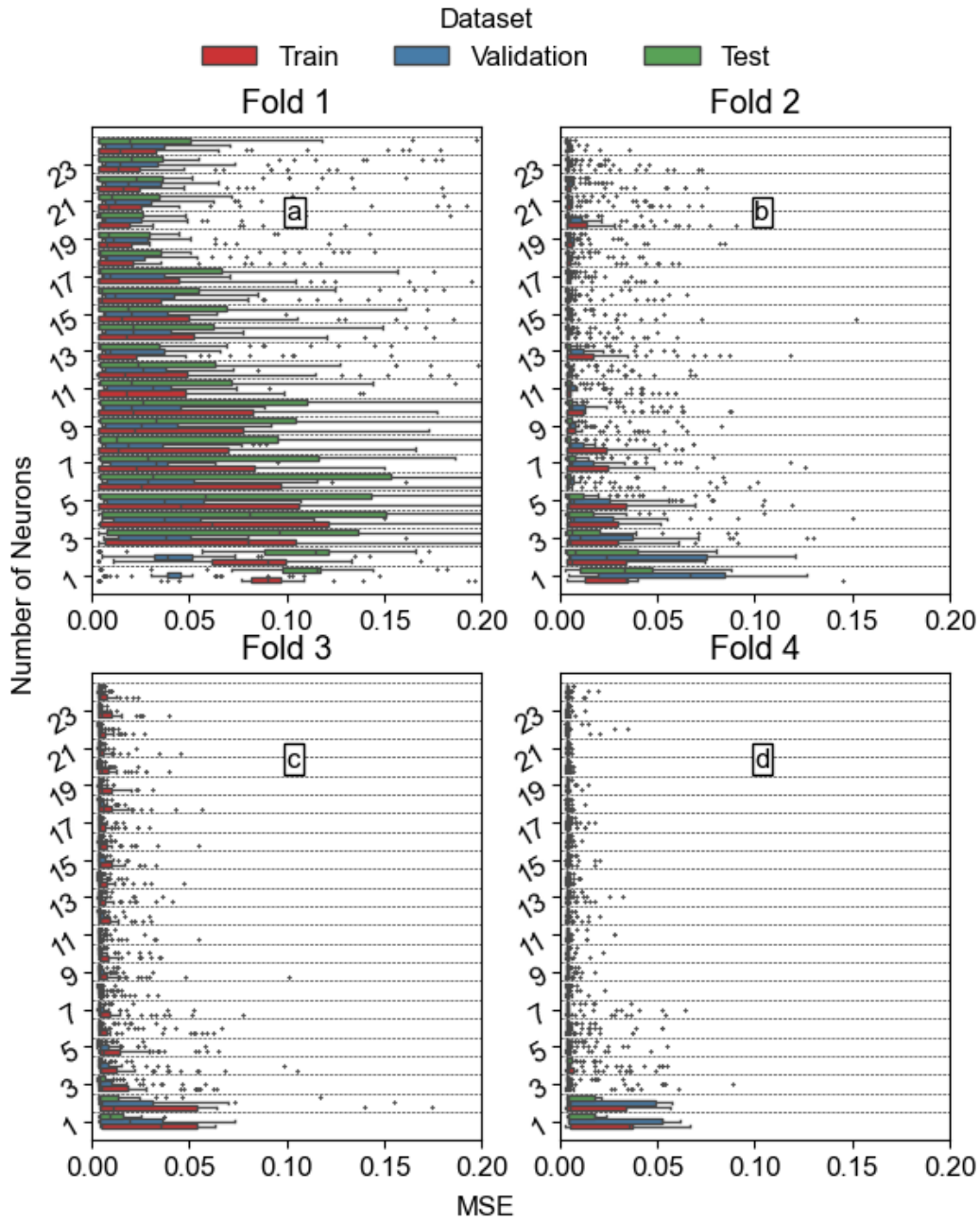


Figure B.72- MSE distributions across different neuron counts for well 244-2 in each cross-validation fold, DSM scenario.

DMS-FFNN model plateau by fold graph for 244-2:

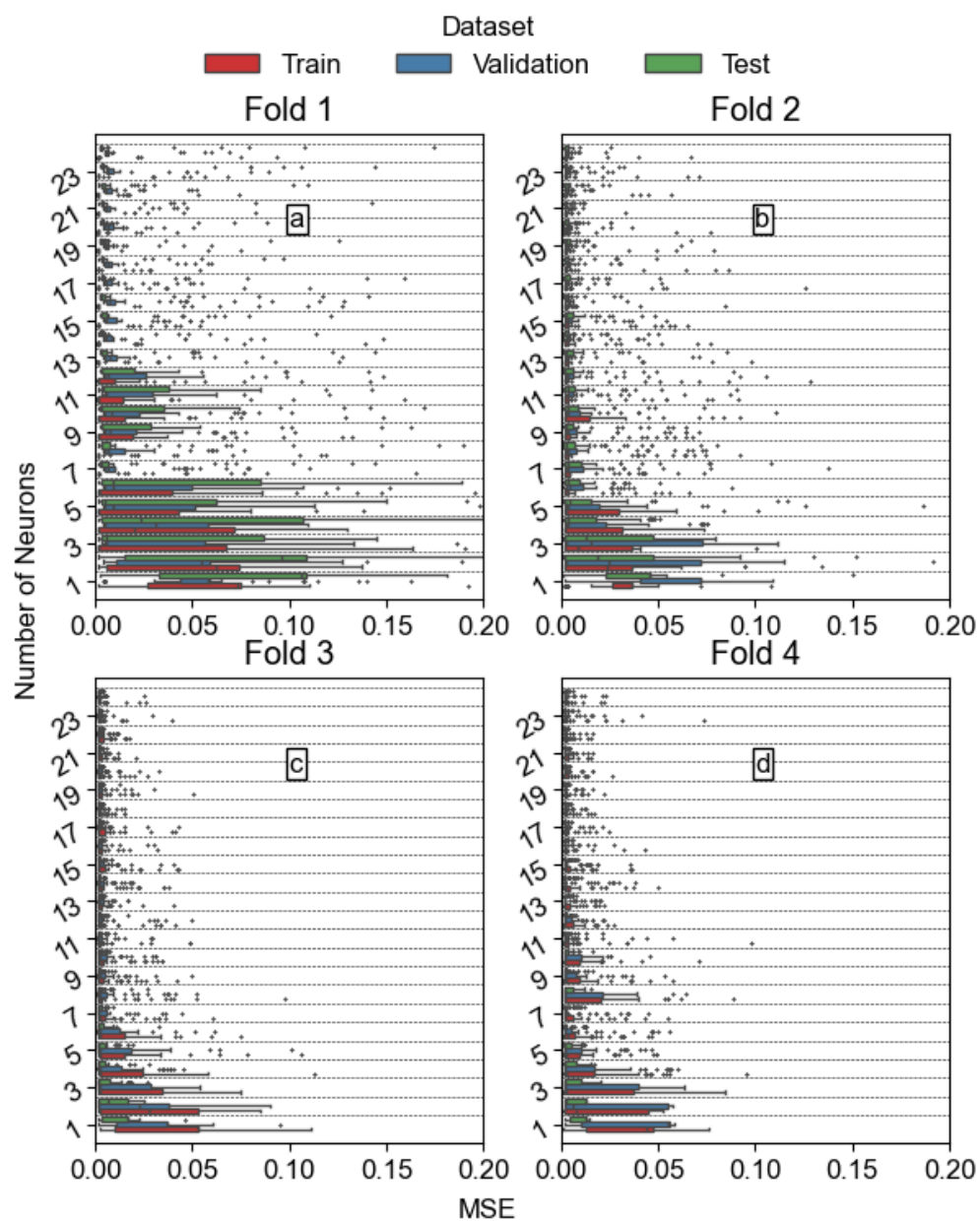


Figure B.73- MSE distributions across different neuron counts for well 244-2 in each cross-validation fold, DMS scenario.

DSM-FFNN model plateau by fold graph for 245-2:

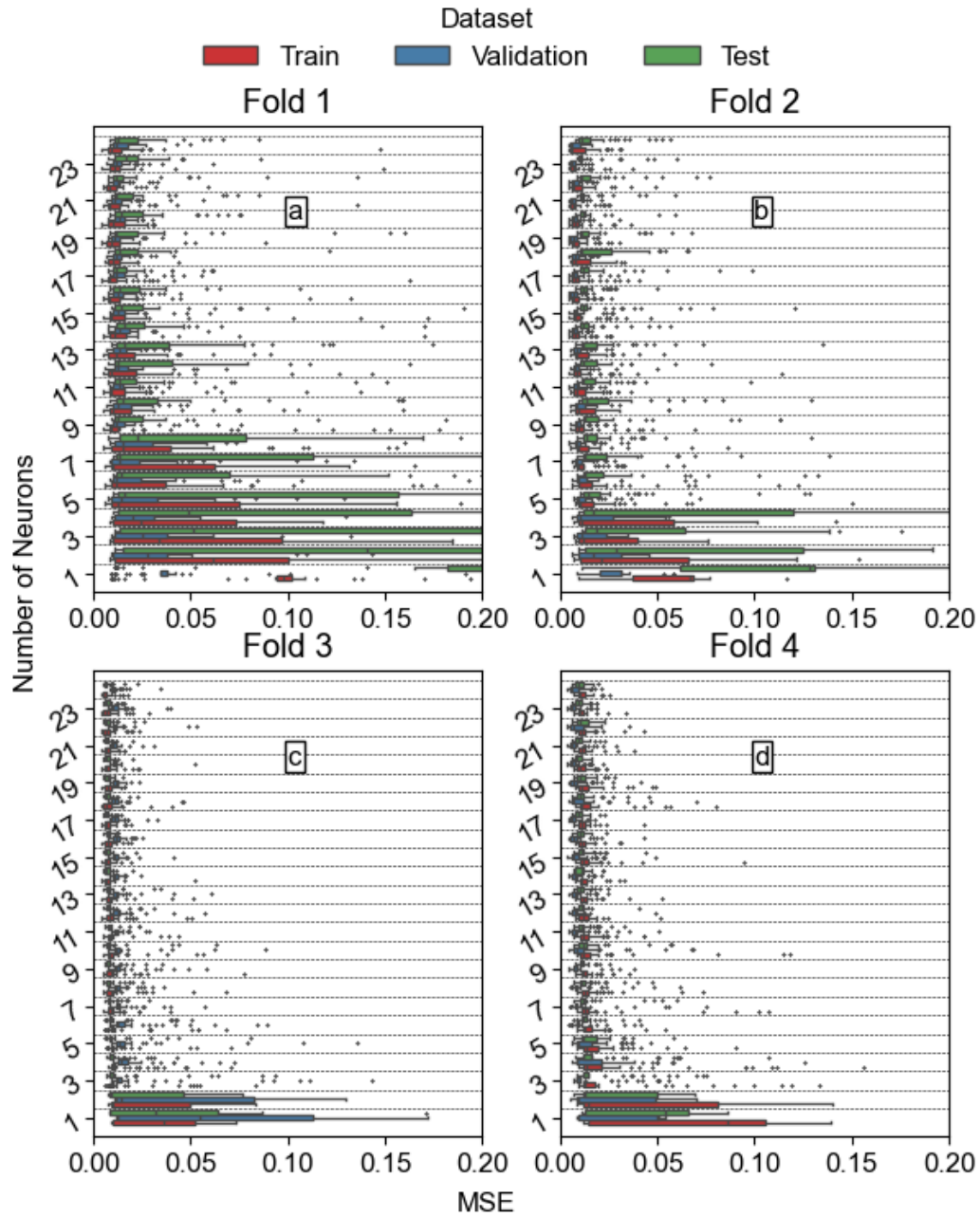


Figure B.74- MSE distributions across different neuron counts for well 245-2 in each cross-validation fold, DSM scenario.

DMS-FFNN model plateau by fold graph for 245-2:

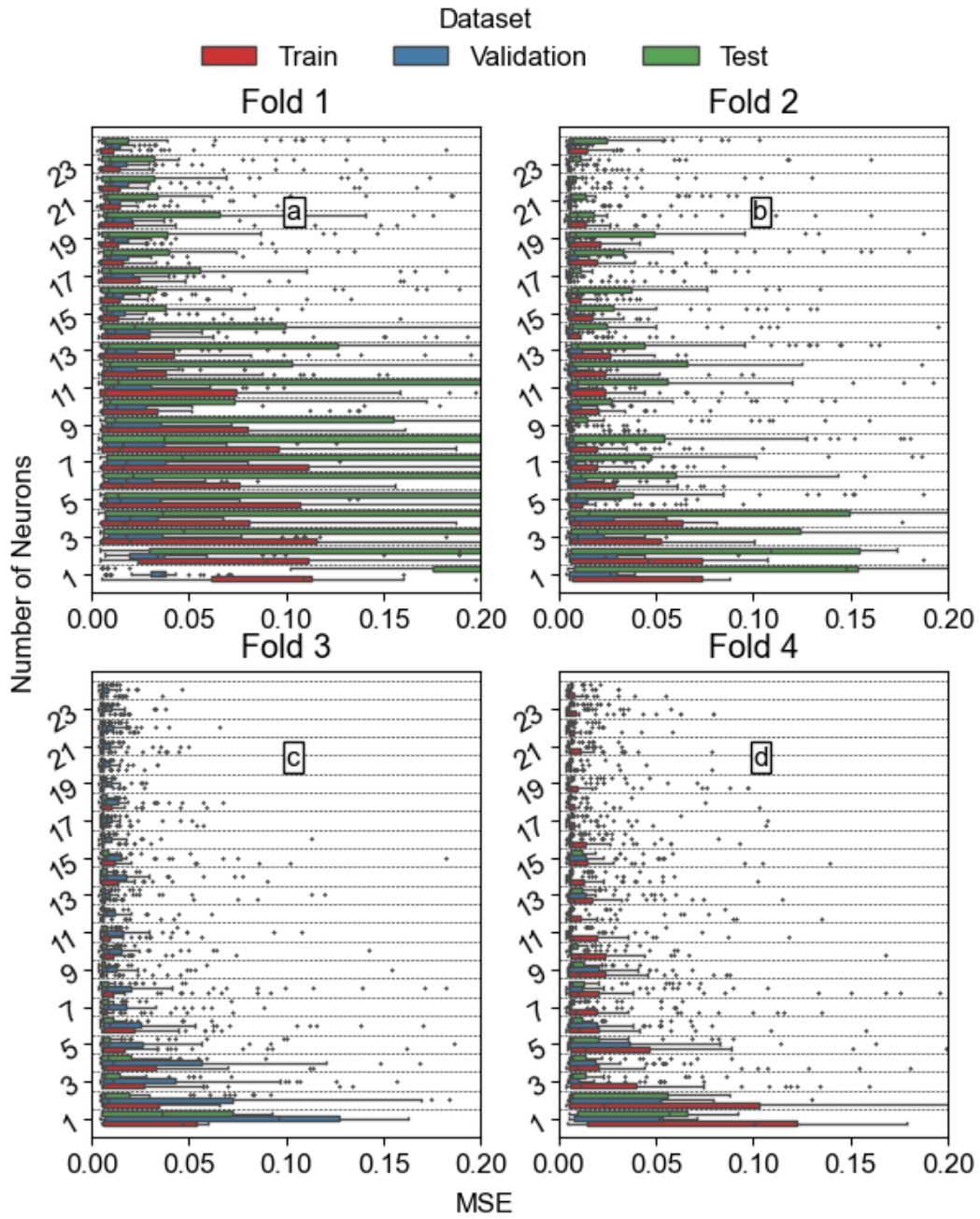


Figure B.75- MSE distributions across different neuron counts for well 245-2 in each cross-validation fold, DMS scenario.

DSM-FFNN model plateau by fold graph for W3:

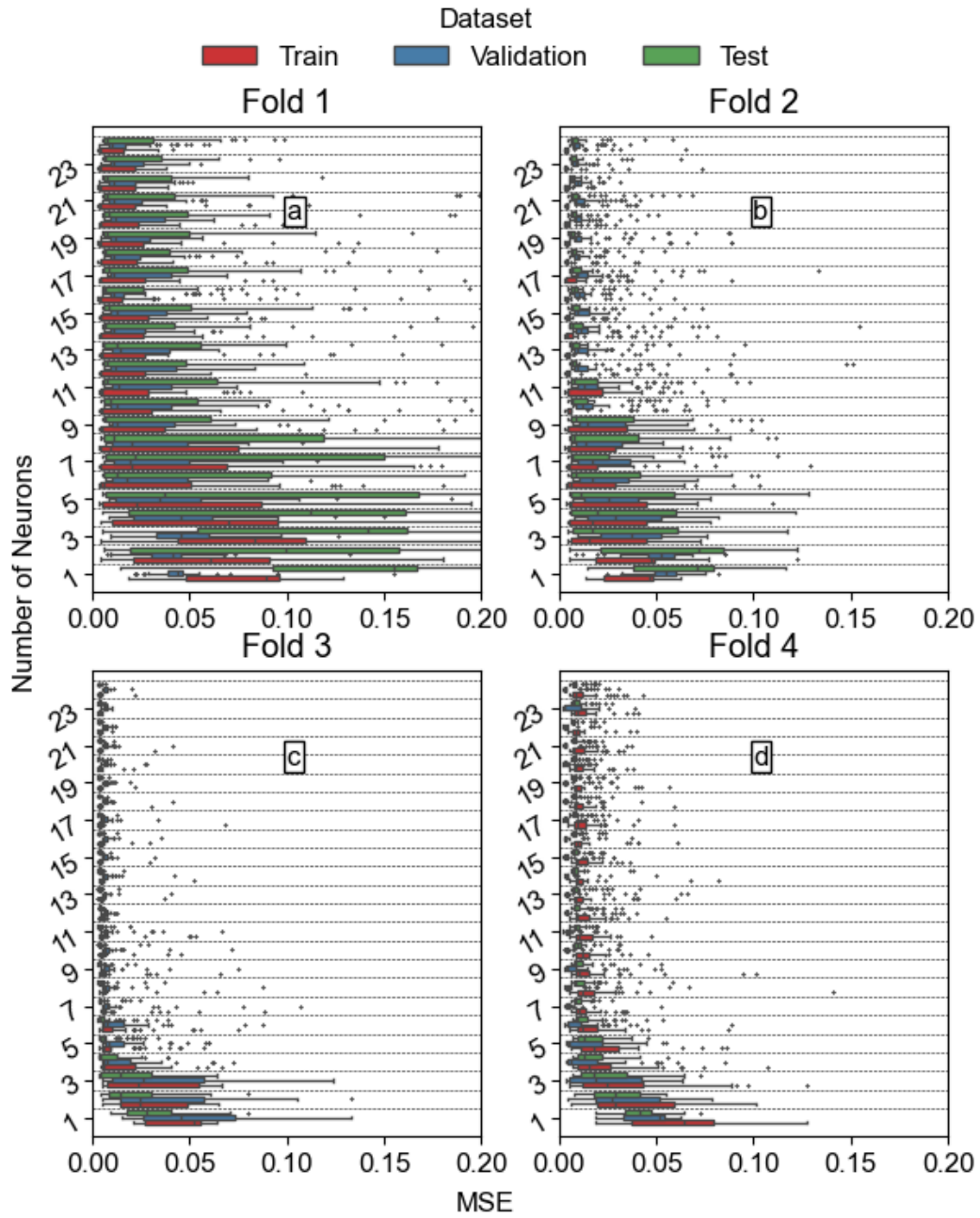


Figure B.76- MSE distributions across different neuron counts for W3 combined in each cross-validation fold, DSM scenario.

DMS-FFNN model plateau by fold graph for W3:

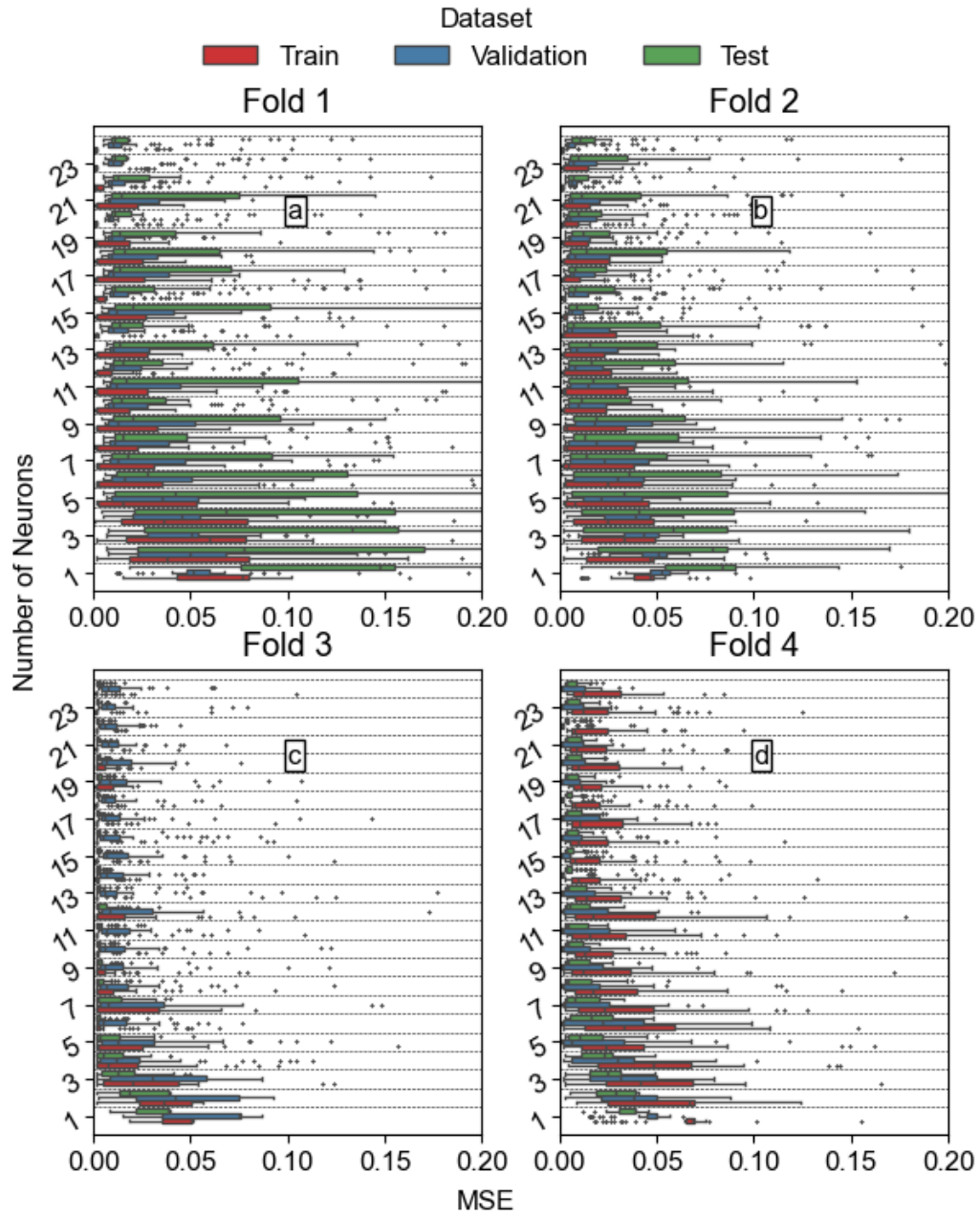


Figure B.77- MSE distributions across different neuron counts for W3 combined in each cross-validation fold, DMS scenario.

DSM-FFNN model plateau by fold graph for 443-1:

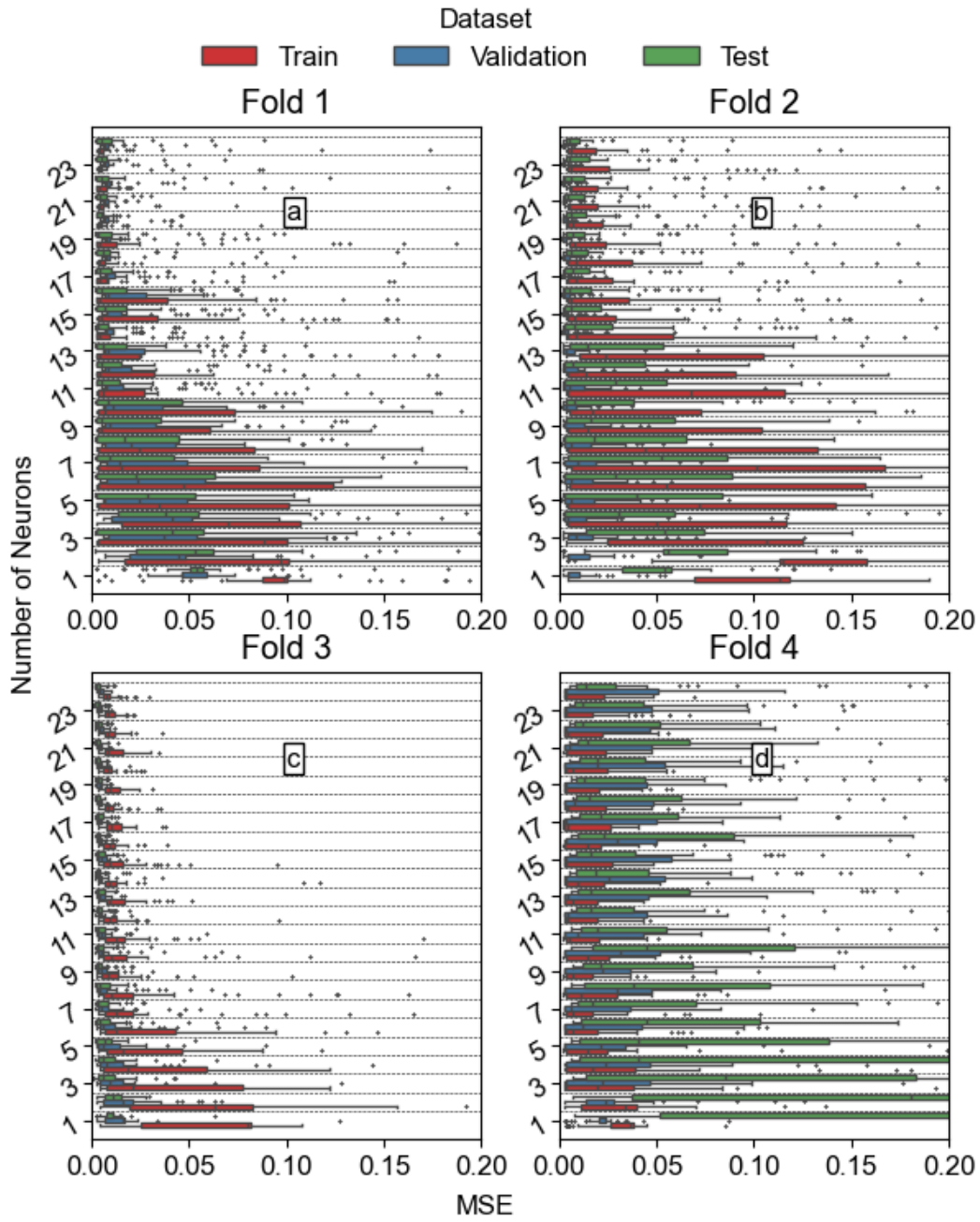


Figure B.78- MSE distributions across different neuron counts for well 443-1 in each cross-validation fold, DSM scenario.

DMS-FFNN model plateau by fold graph for 443-1:

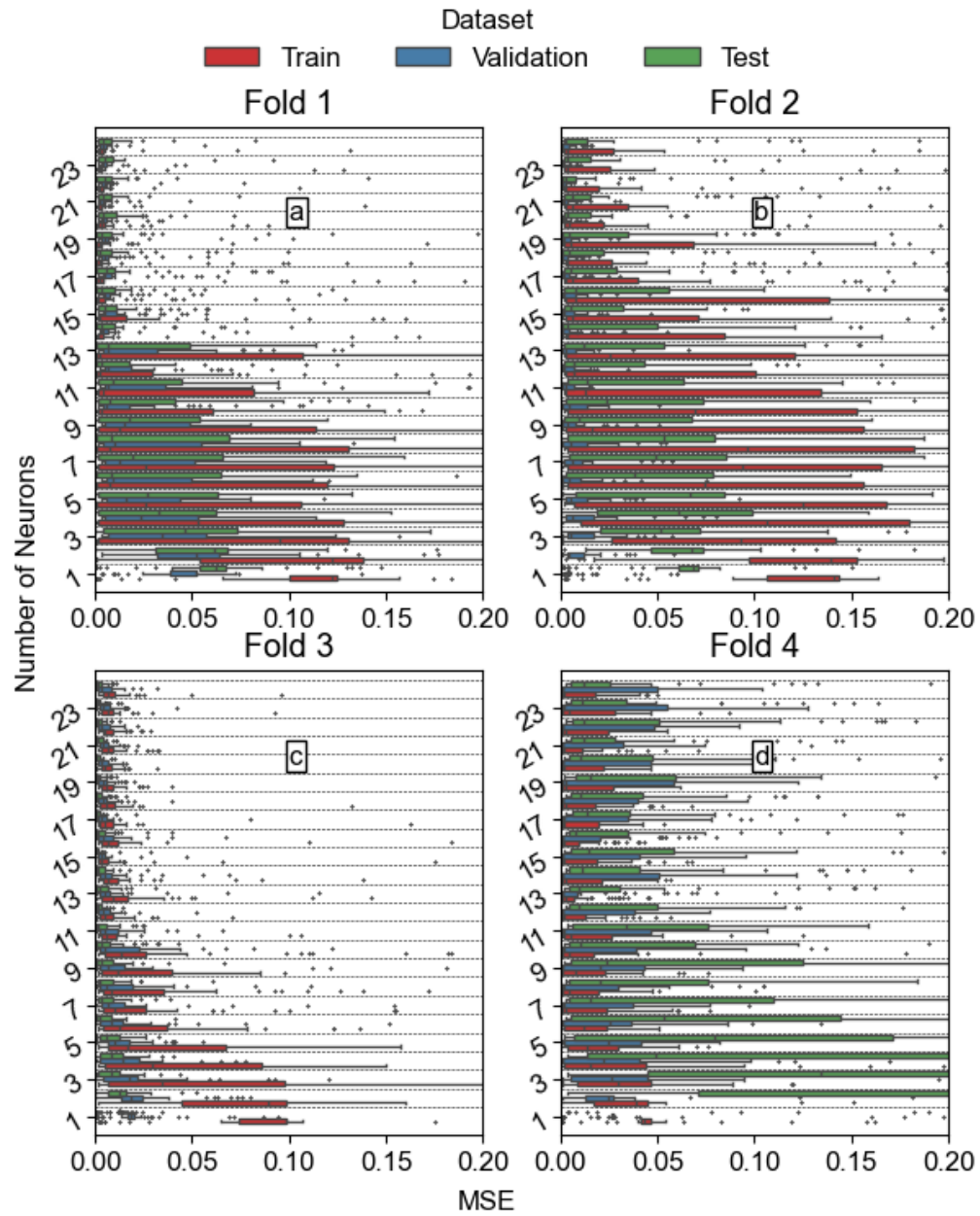


Figure B.79- MSE distributions across different neuron counts for well 443-1 in each cross-validation fold, DMS scenario.

DSM-FFNN model plateau by fold graph for the aggregated model:

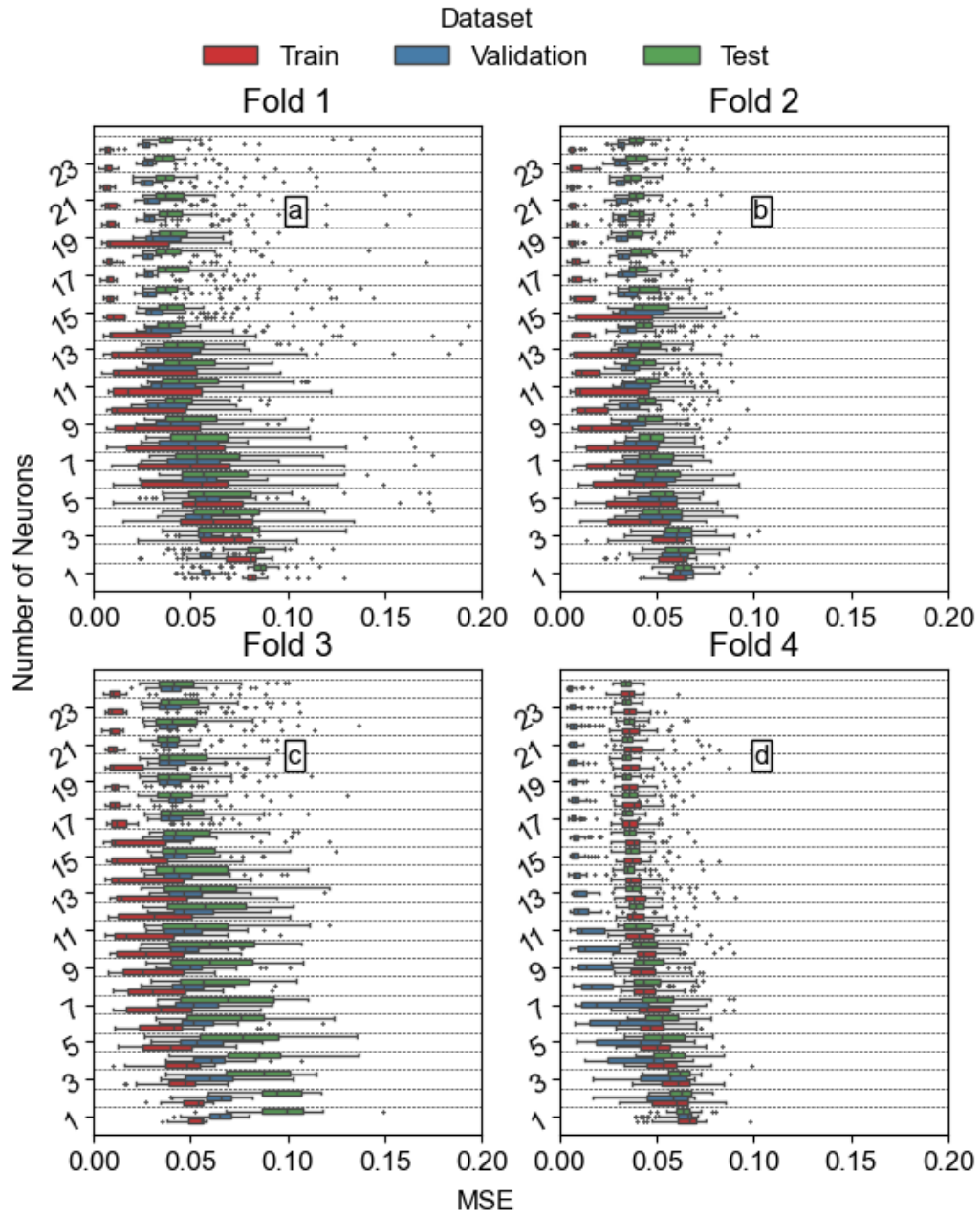


Figure B.80- MSE distributions across different neuron counts for the aggregated model in each cross-validation fold, DSM scenario.

DMS-FFNN model plateau by fold graph for the aggregated model:

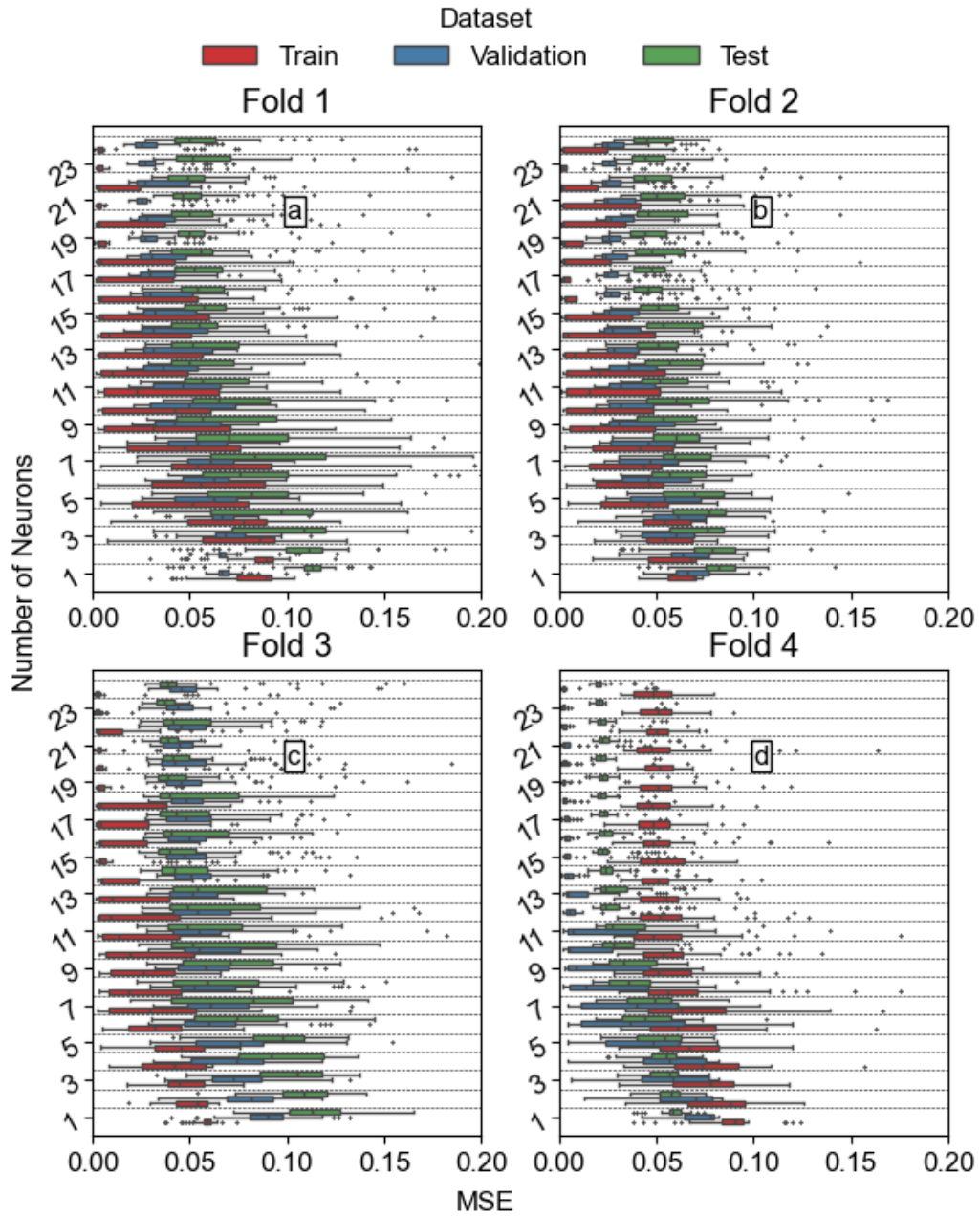


Figure B.81-MSE distributions across different neuron counts for the aggregated model in each cross-validation fold, DMS scenario.

Performance Plateau Overall Plots (MSE)

DSM-FFNN model plateau overall graph for 293-3:

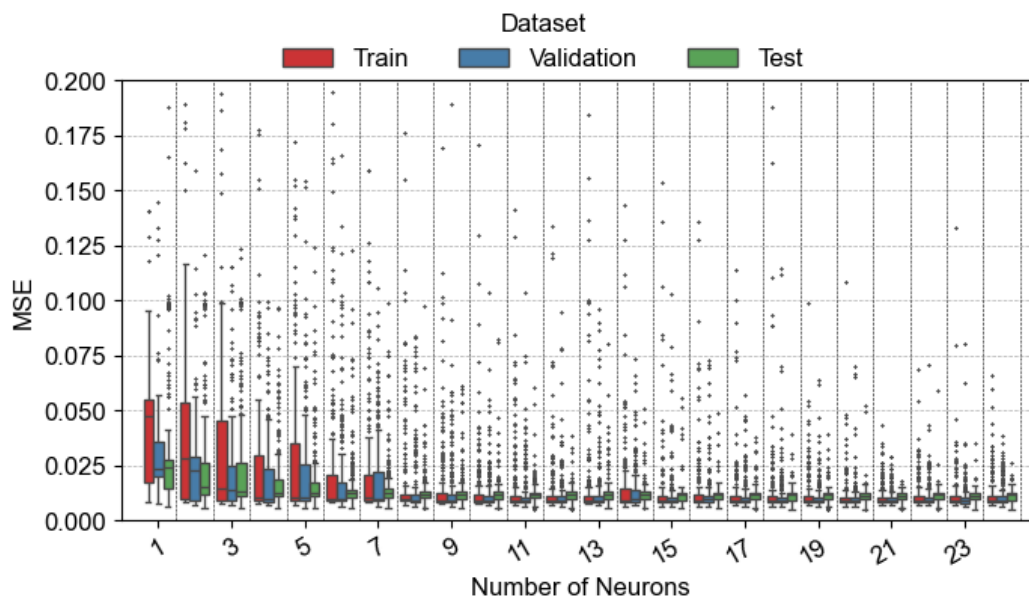


Figure B.82- Combined MSE distributions for well 293-3, DSM scenario.

DMS-FFNN model plateau overall graph for 293-3:

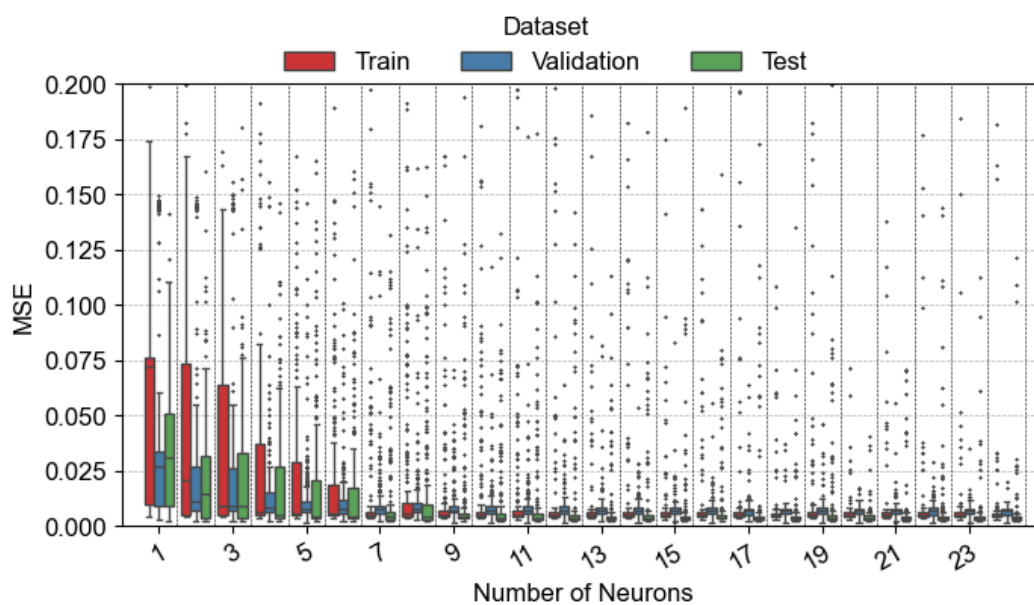


Figure B.83- Combined MSE distributions for well 293-3, DMS scenario.

DSM-FFNN model plateau overall graph for W1:

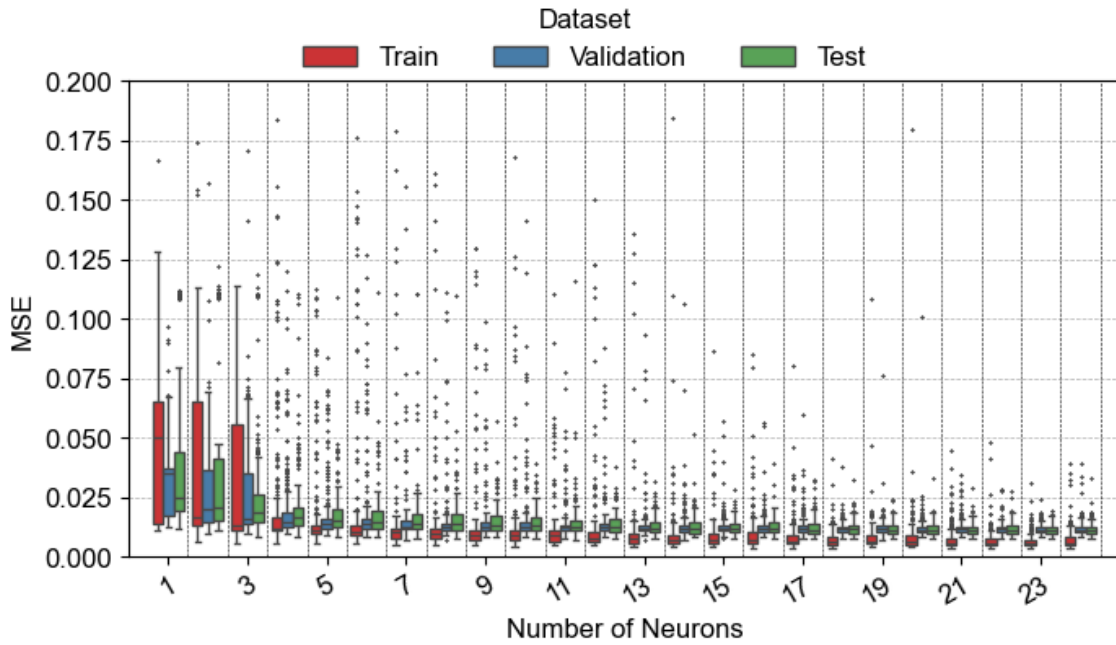


Figure B.84- Combined MSE distributions for W1 combined model, DSM scenario.

DMS-FFNN model plateau overall graph for W1:

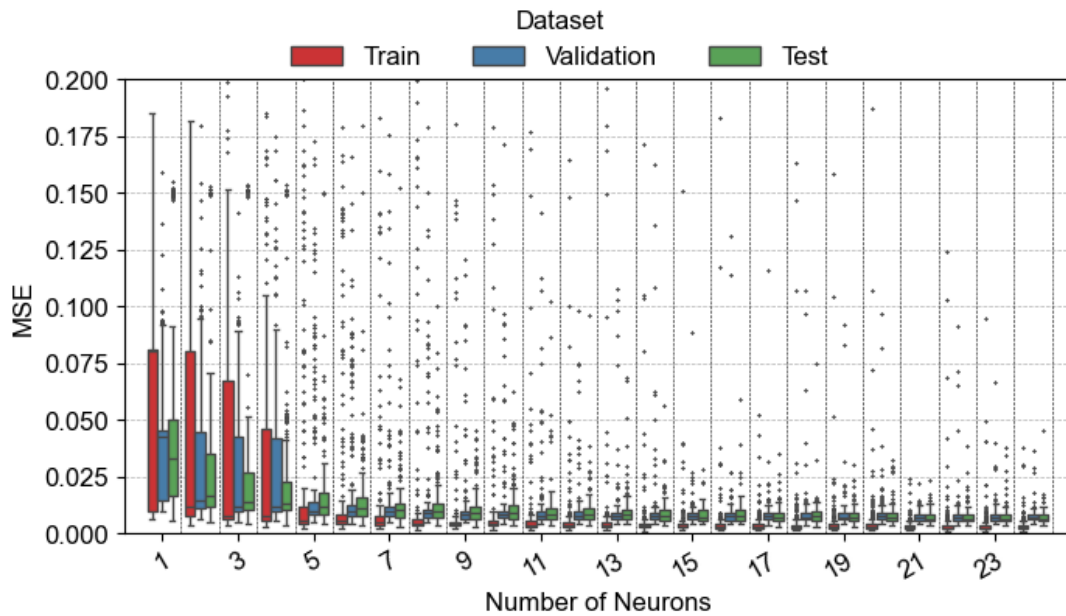


Figure B.85- Combined MSE distributions for W1 combined model, DMS scenario.

DSM-FFNN model plateau overall graph for 439-1:

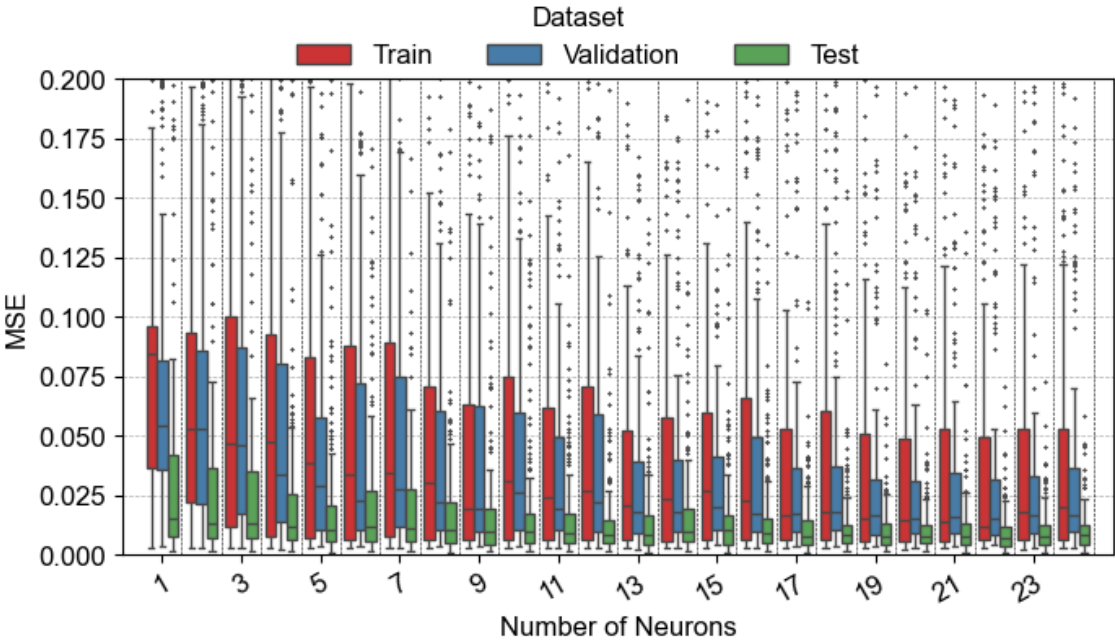


Figure B.86- Combined MSE distributions for well 439-1, DSM scenario.

DMS-FFNN model plateau overall graph for 439-1:

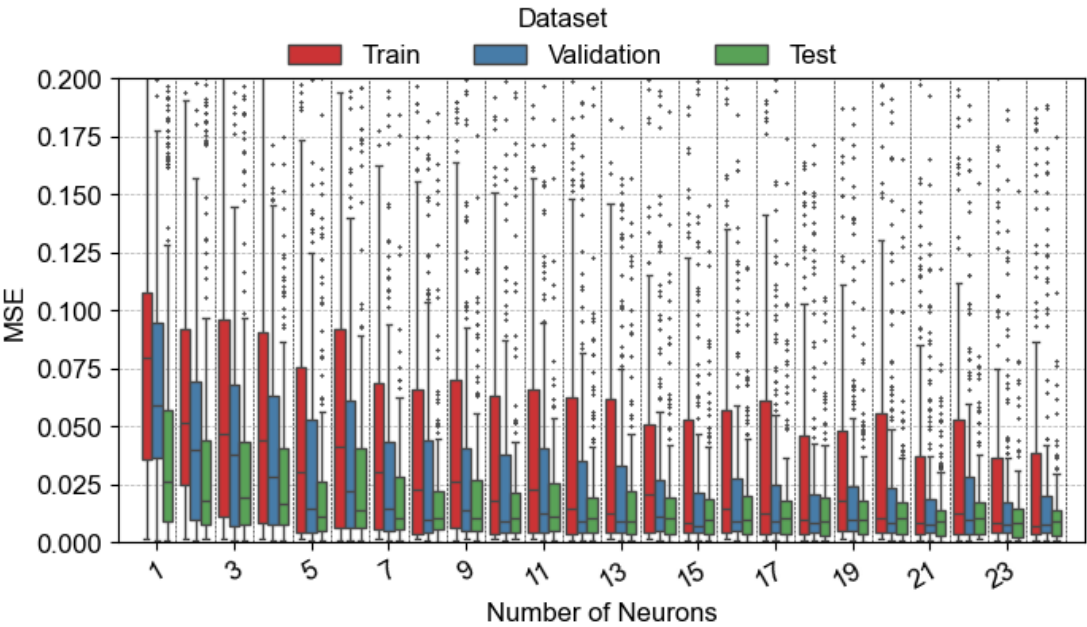


Figure B.87- Combined MSE distributions for well 439-1, DMS scenario.

DSM-FFNN model plateau overall graph for 440-1:

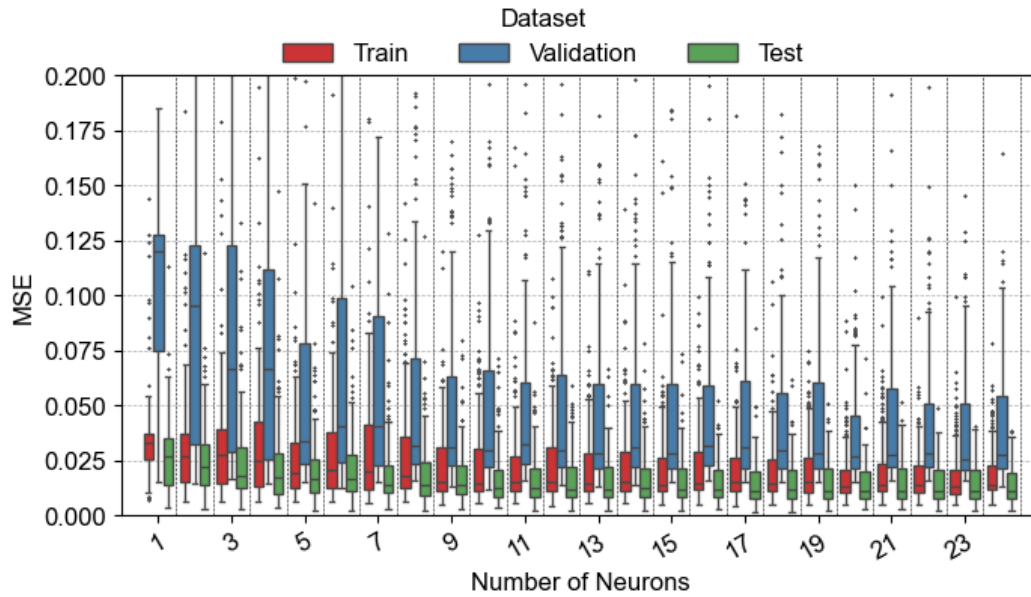


Figure B.88- Combined MSE distributions for well 440-1, DSM scenario.

DMS-FFNN model plateau overall graph for 440-1:

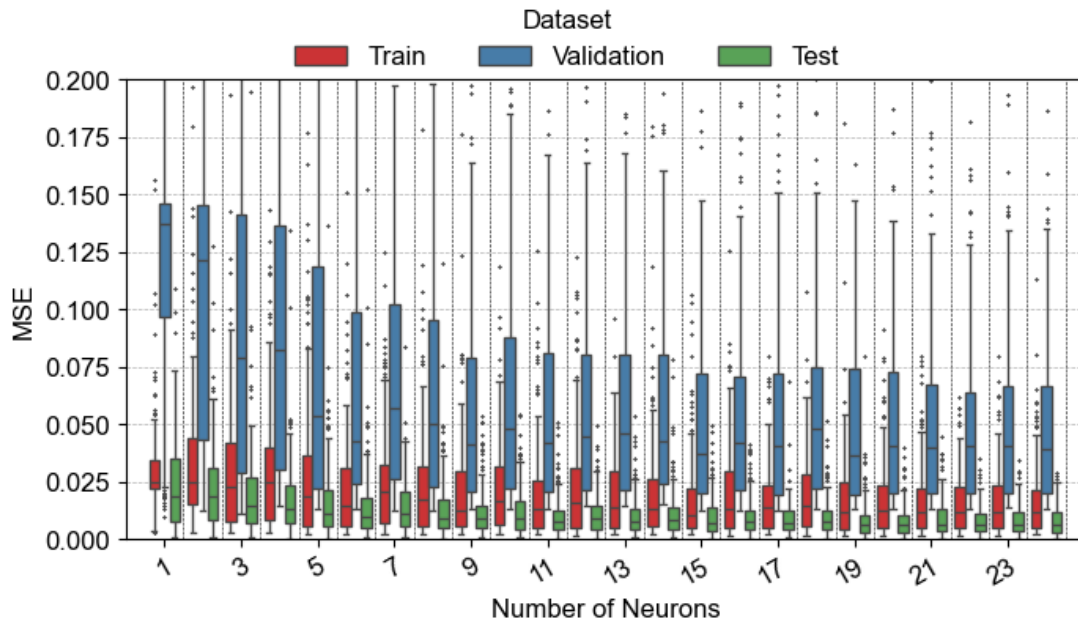


Figure B.89- Combined MSE distributions for well 440-1, DMS scenario.

DSM-FFNN model plateau overall graph for 442-1:

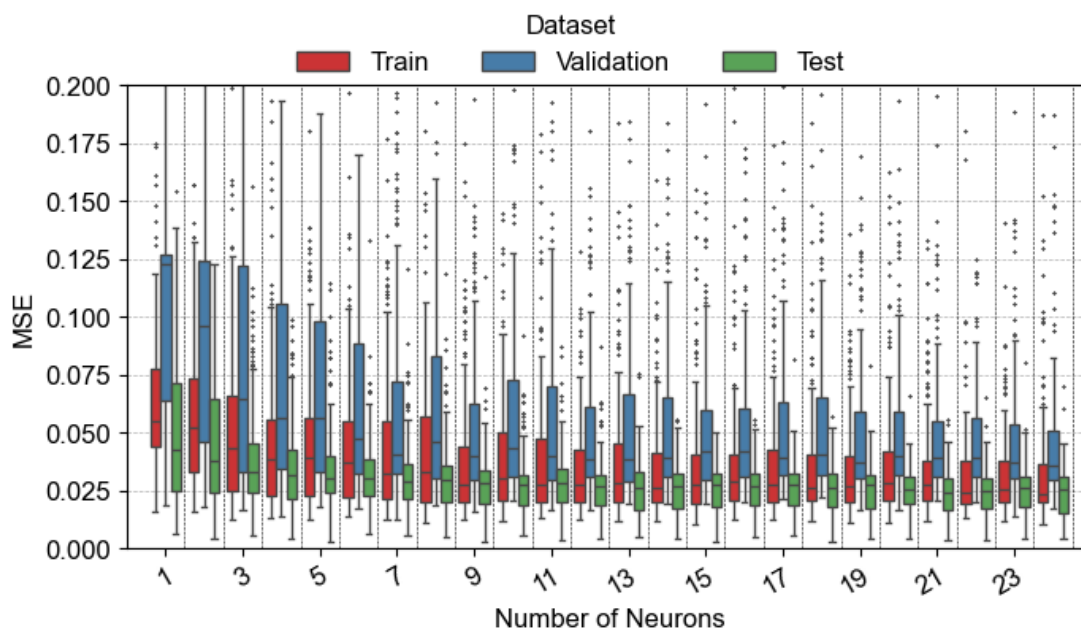


Figure B.90- Combined MSE distributions for well 442-1, DSM scenario.

DMS-FFNN model plateau overall graph for 442-1:

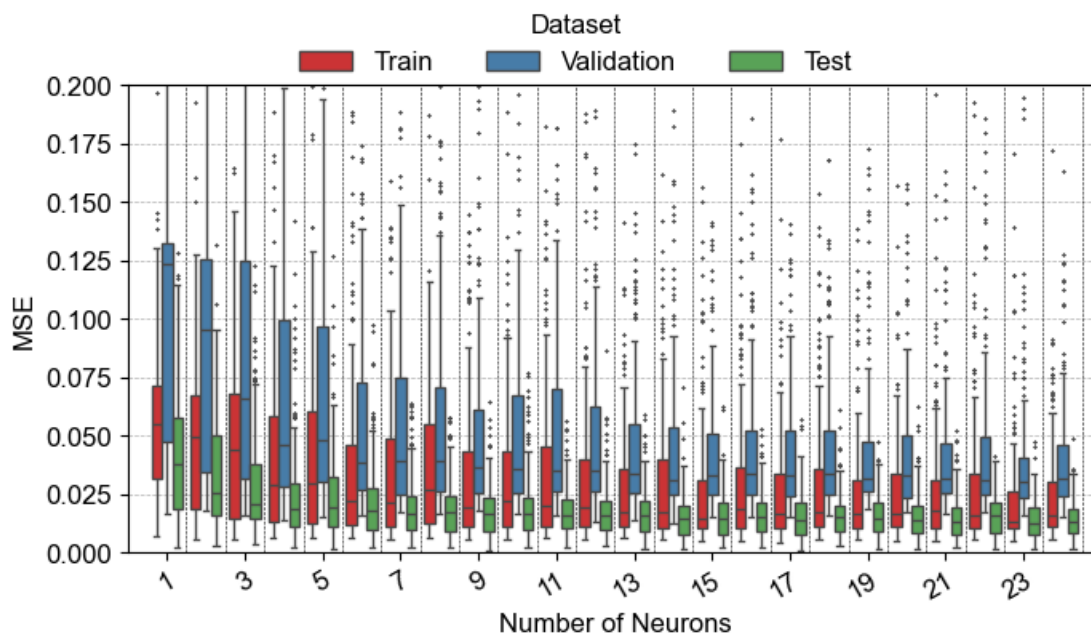


Figure B.91- Combined MSE distributions for well 442-1, DMS scenario.

DSM-FFNN model plateau overall graph for W2:

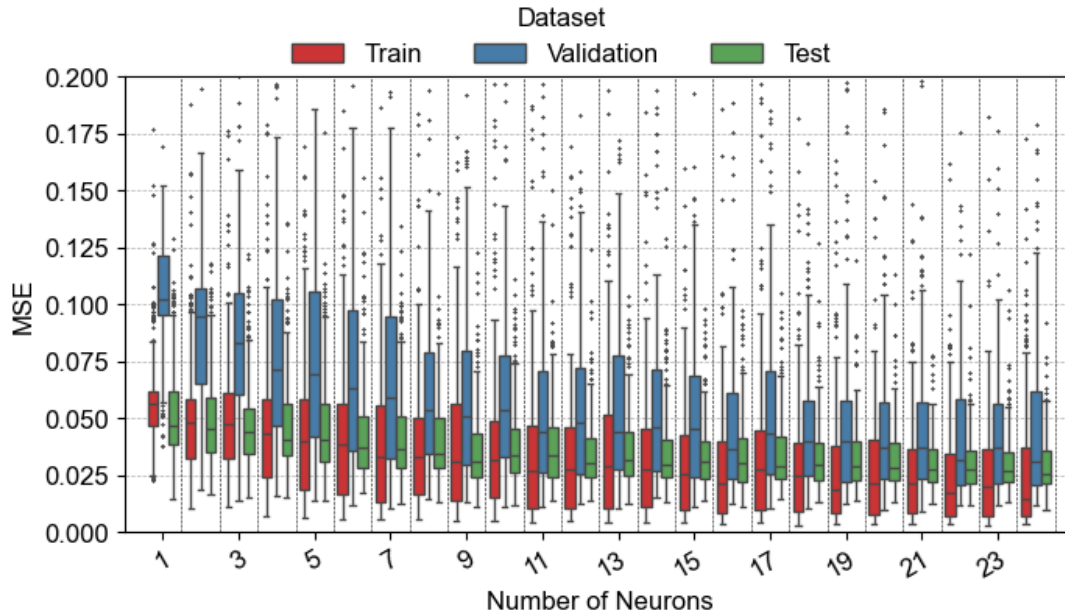


Figure B.92- Combined MSE distributions for W2 combined model, DSM scenario.

DMS-FFNN model plateau overall graph for W2:

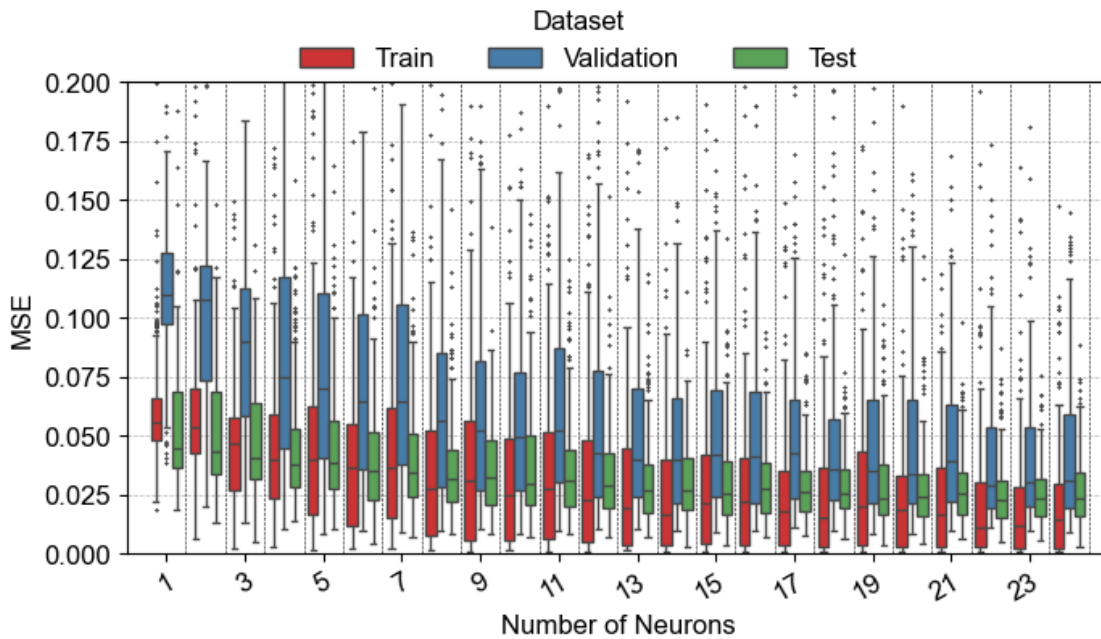


Figure B.93- Combined MSE distributions for W2 combined model, DMS scenario.

DSM-FFNN model plateau overall graph for 244-2:

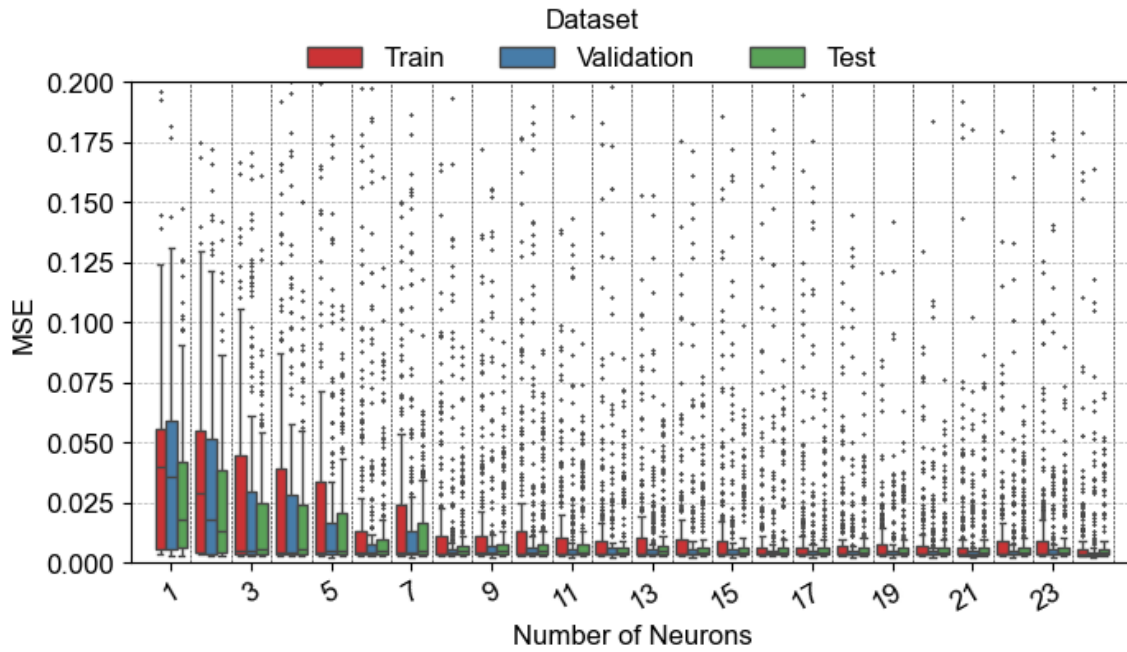


Figure B.94- Combined score distributions for well 244-2, DSM scenario.

DMS-FFNN model plateau overall graph for 244-2:

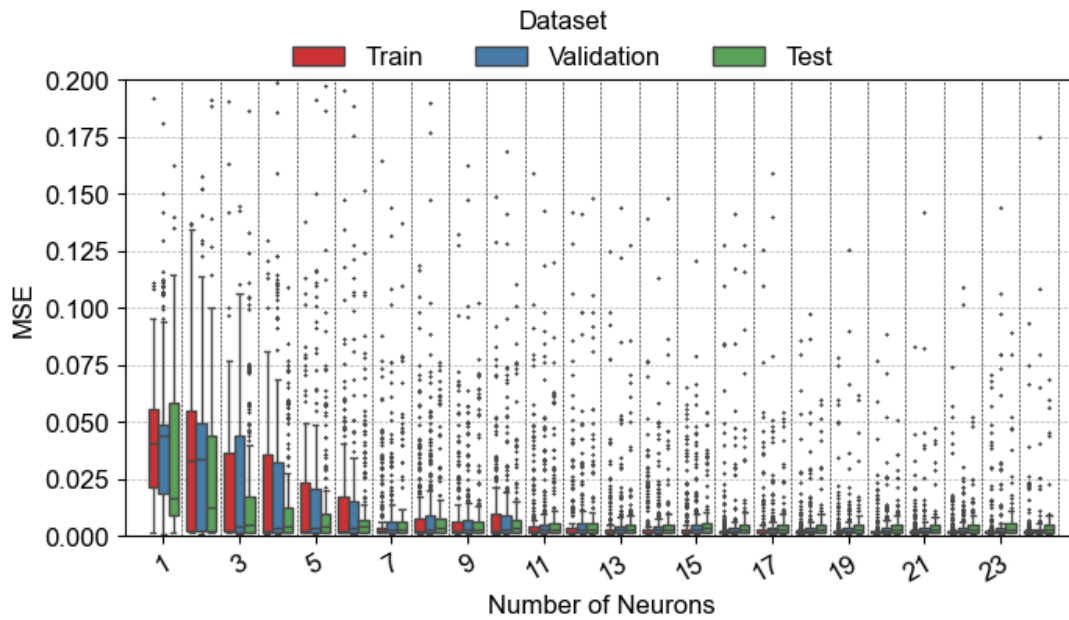


Figure B.95- Combined MSE distributions for well 244-2, DMS scenario.

DSM-FFNN model plateau overall graph for 245-2:

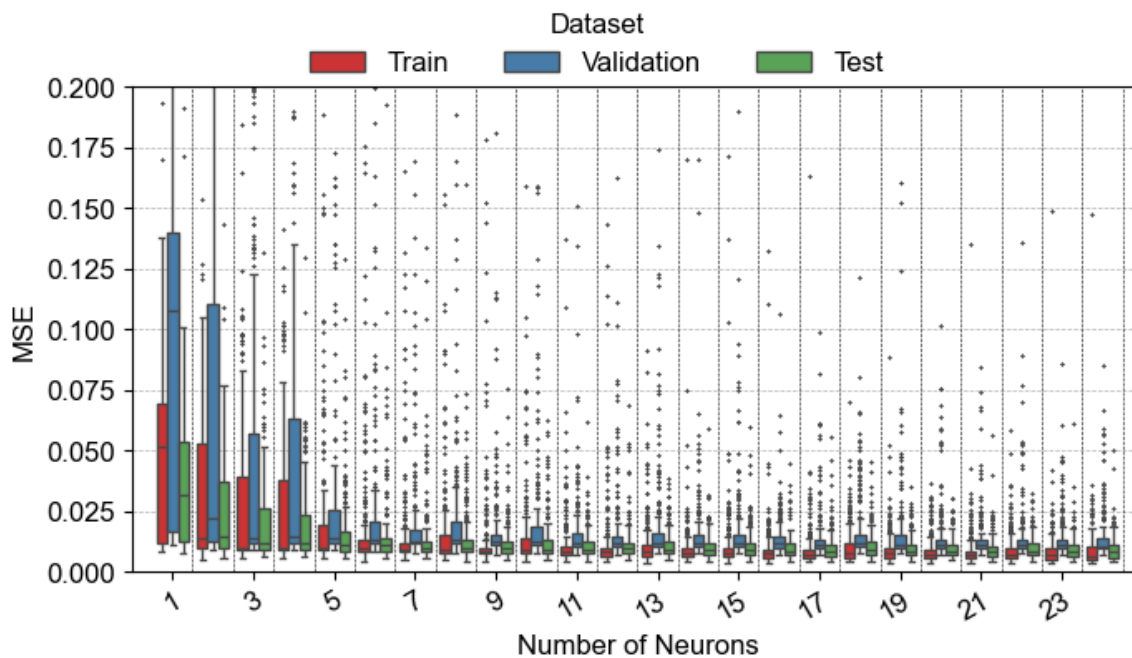


Figure B.96- Combined MSE distributions for well 245-2, DSM scenario.

DMS-FFNN model plateau overall graph for 245-2:

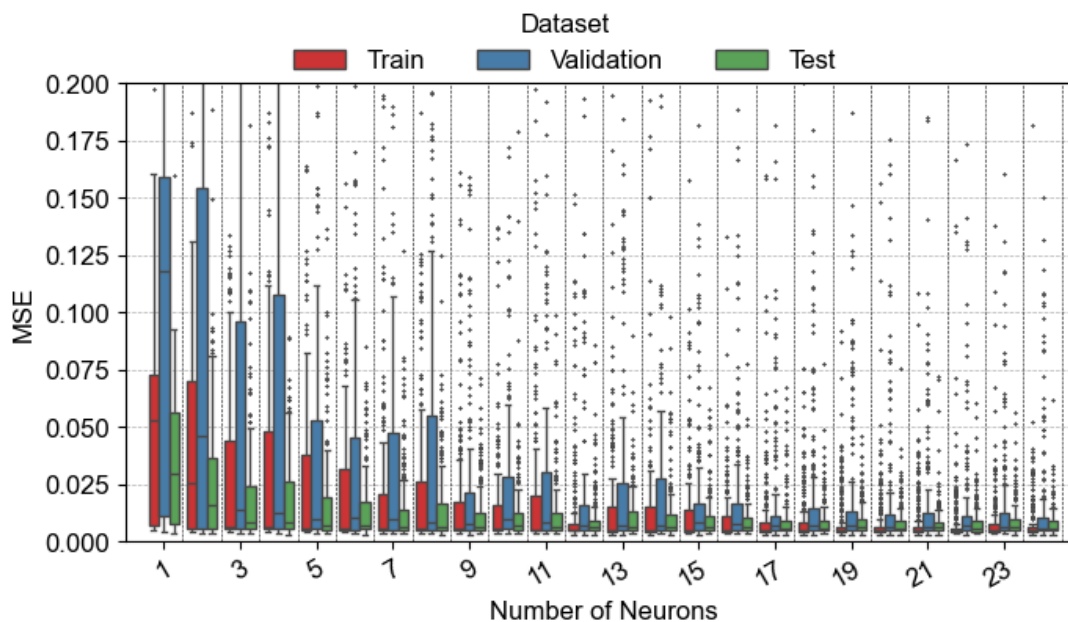


Figure B.97- Combined MSE distributions for well 245-2, DMS scenario.

DSM-FFNN model plateau overall graph for W3:

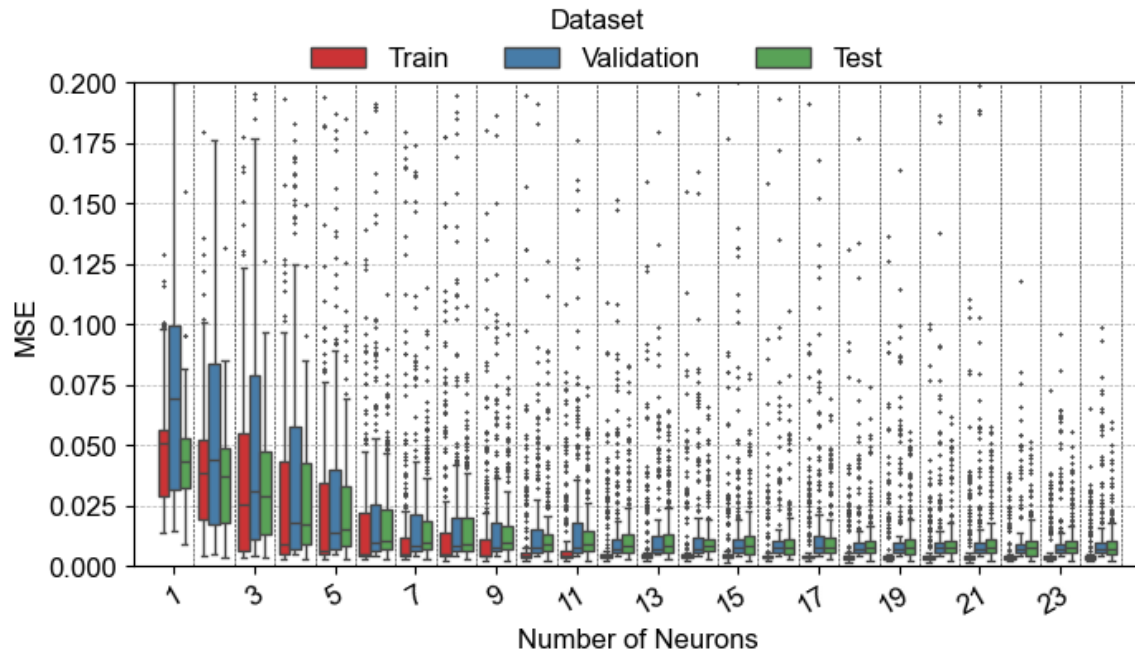


Figure B.98- Combined MSE distributions for W3 combined model, DSM scenario.

DMS-FFNN model plateau overall graph for W3:

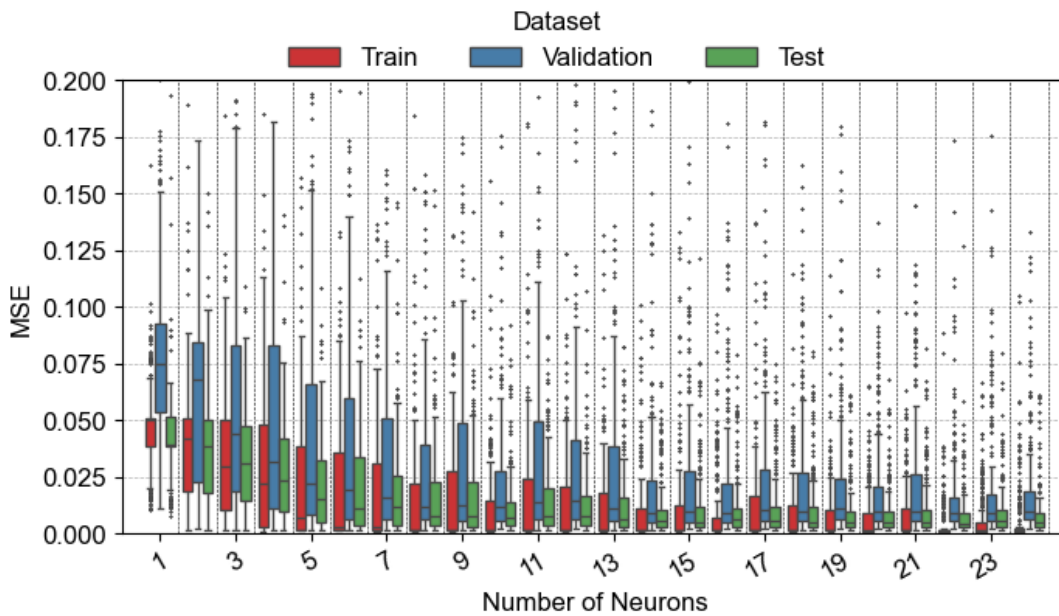


Figure B.99- Combined MSE distributions for W3 combined, DMS scenario.

DSM-FFNN model plateau overall graph for 443-1:

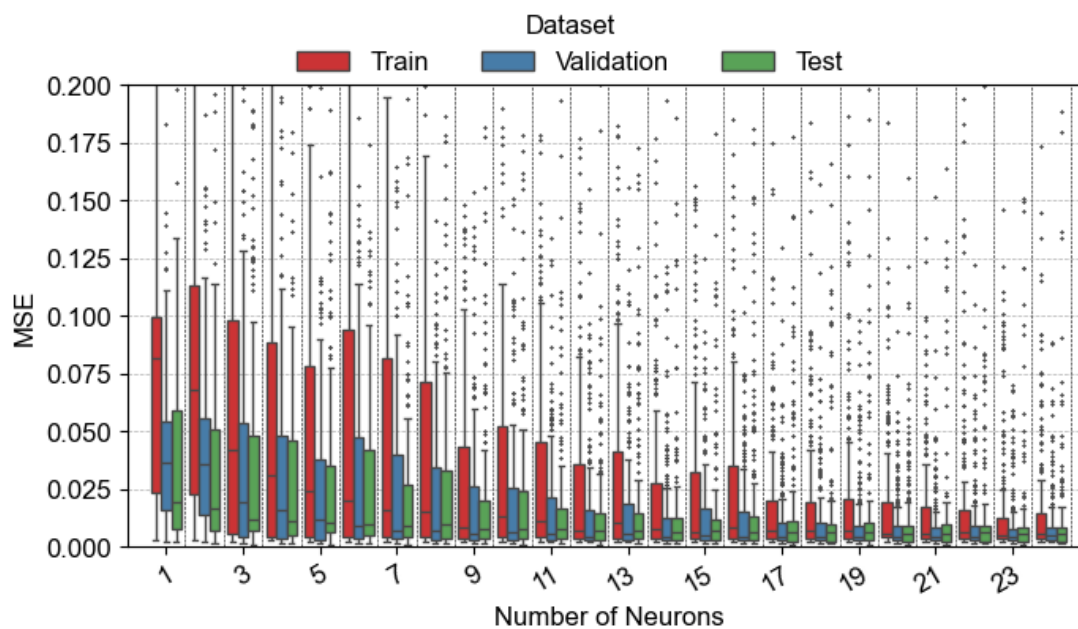


Figure B.100- Combined MSE distributions for well 443-1, DSM scenario.

DMS-FFNN model plateau overall graph for 443-1:

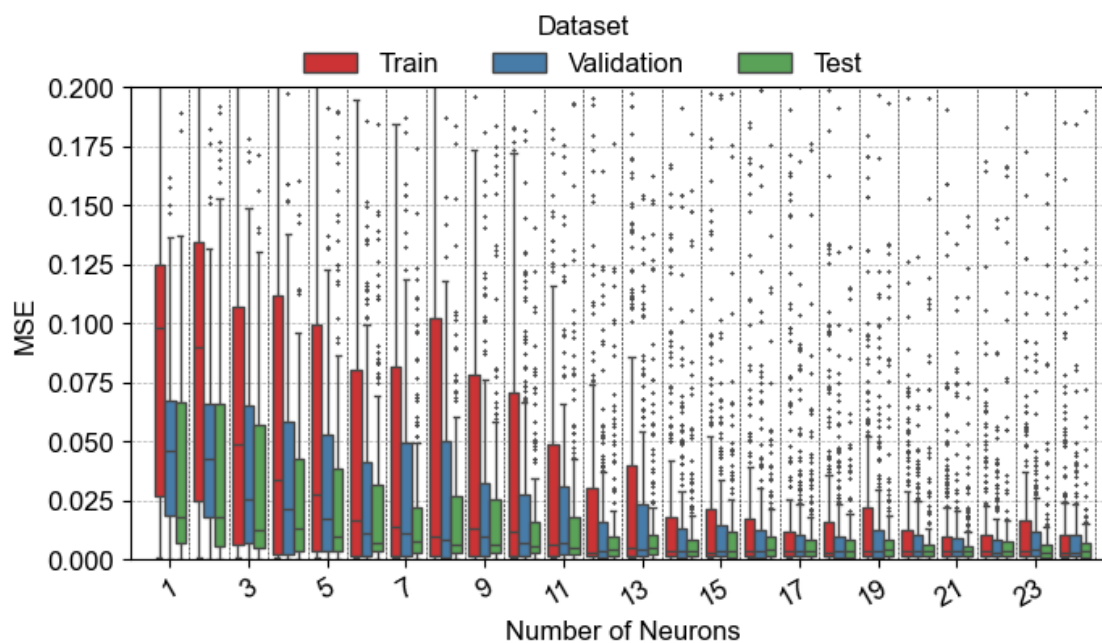


Figure B.101- Combined MSE distributions for well 443-1, DMS scenario.

DSM-FFNN model plateau overall graph for the aggregated model:

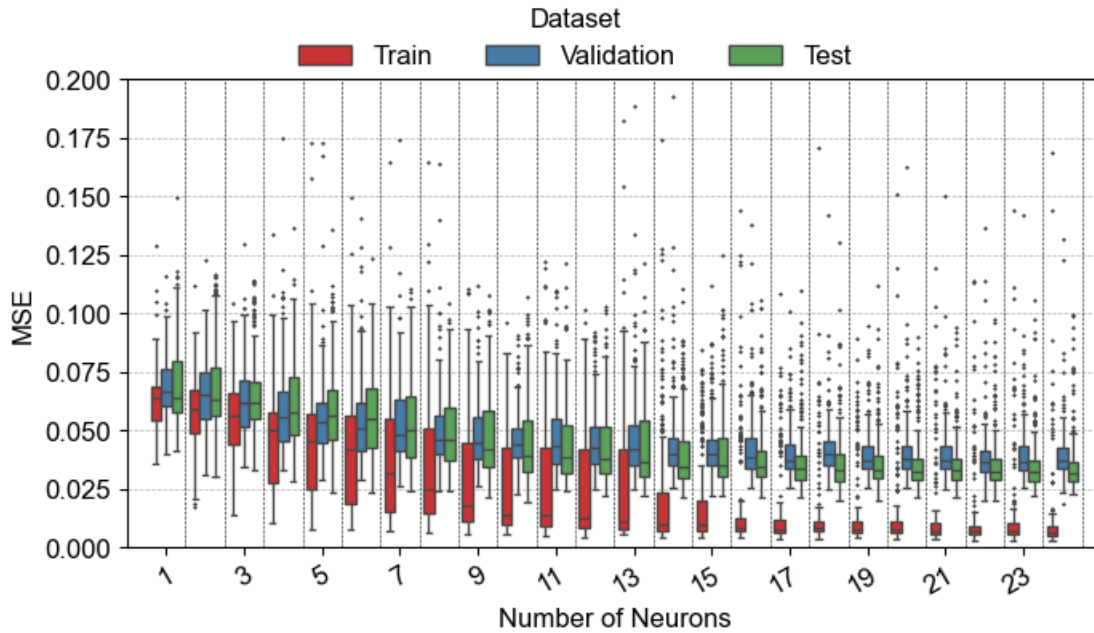


Figure B.102- Combined MSE distributions for the aggregated model, DSM scenario.

DMS-FFNN model plateau overall graph for the aggregated model:

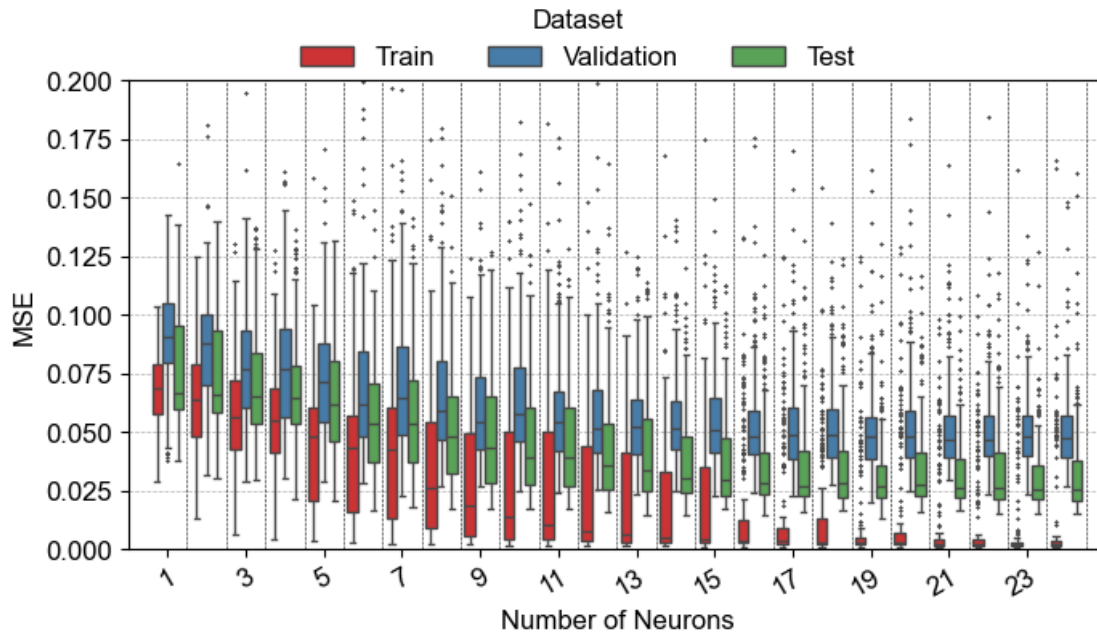


Figure B.103- Combined MSE distributions for the aggregated model, DMS scenario.

Performance Plateau By-Fold Plots (AIC)

DMS-FFNN model plateau by fold graph for 293-2:

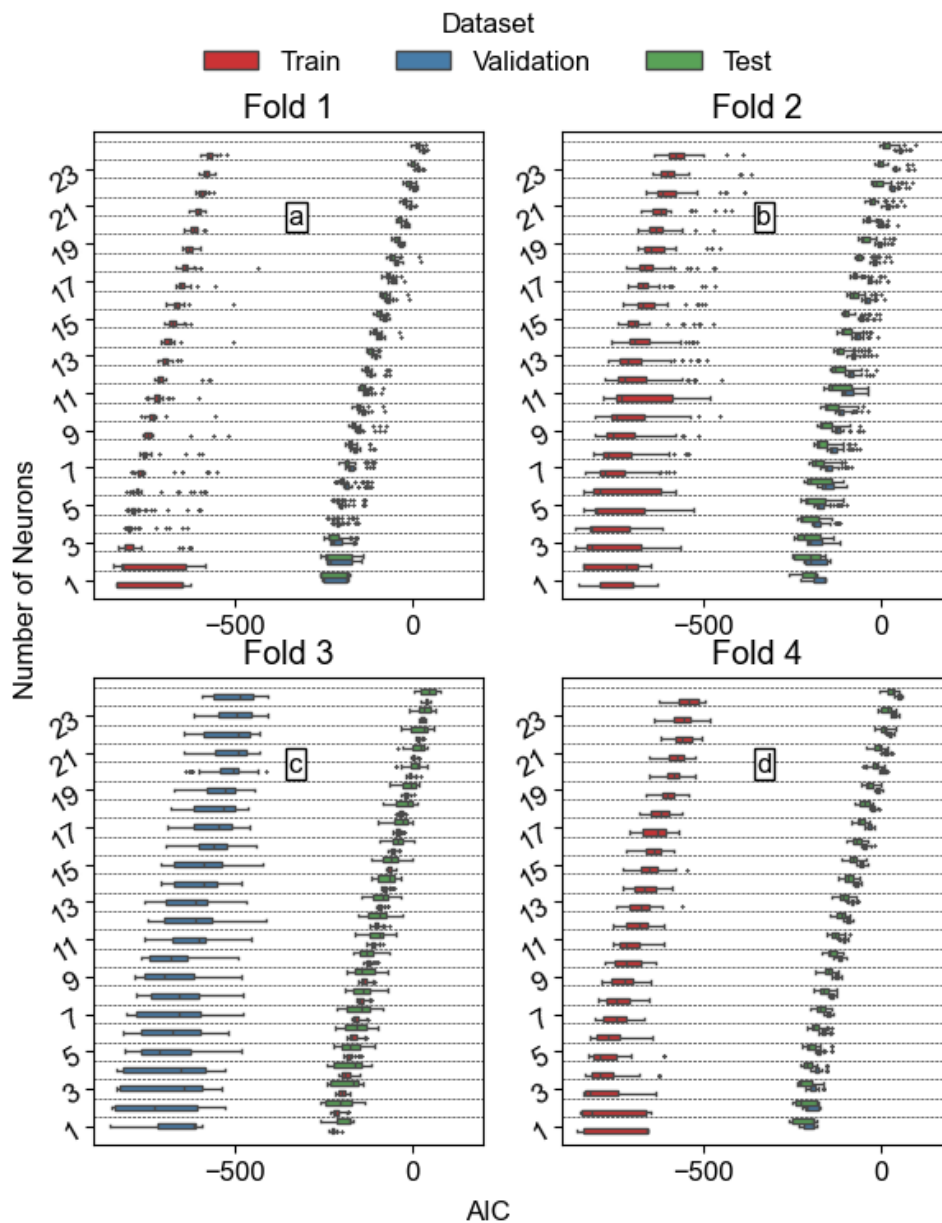


Figure B.104- AIC distributions across different neuron counts for well 293-2 in each cross-validation fold, DMS scenario.

DSM-FFNN model plateau by fold overall graph for 293-3:

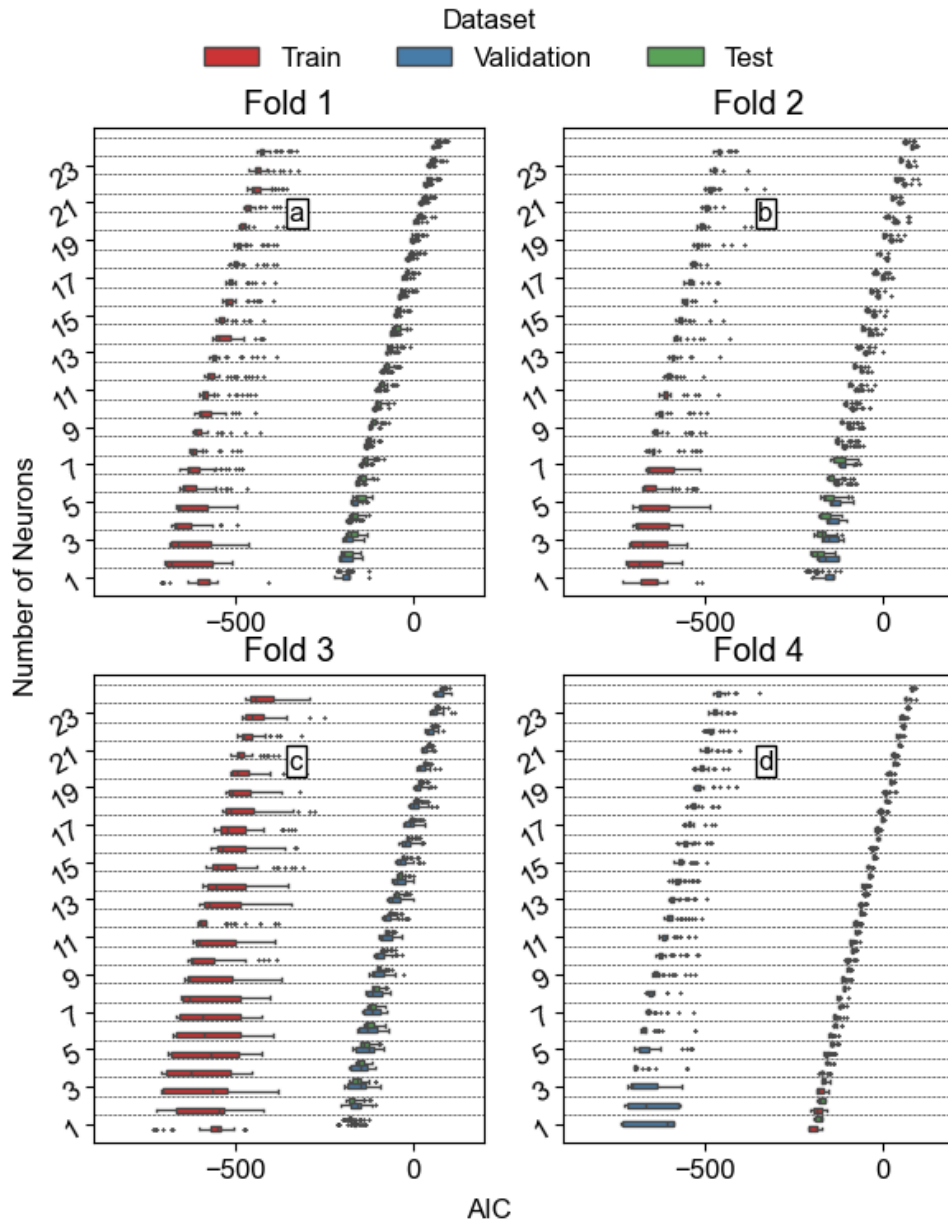


Figure B.105- AIC distributions across different neuron counts for well 293-3 in each cross-validation fold, DSM scenario.

DMS-FFNN model plateau by fold overall graph for 293-3:

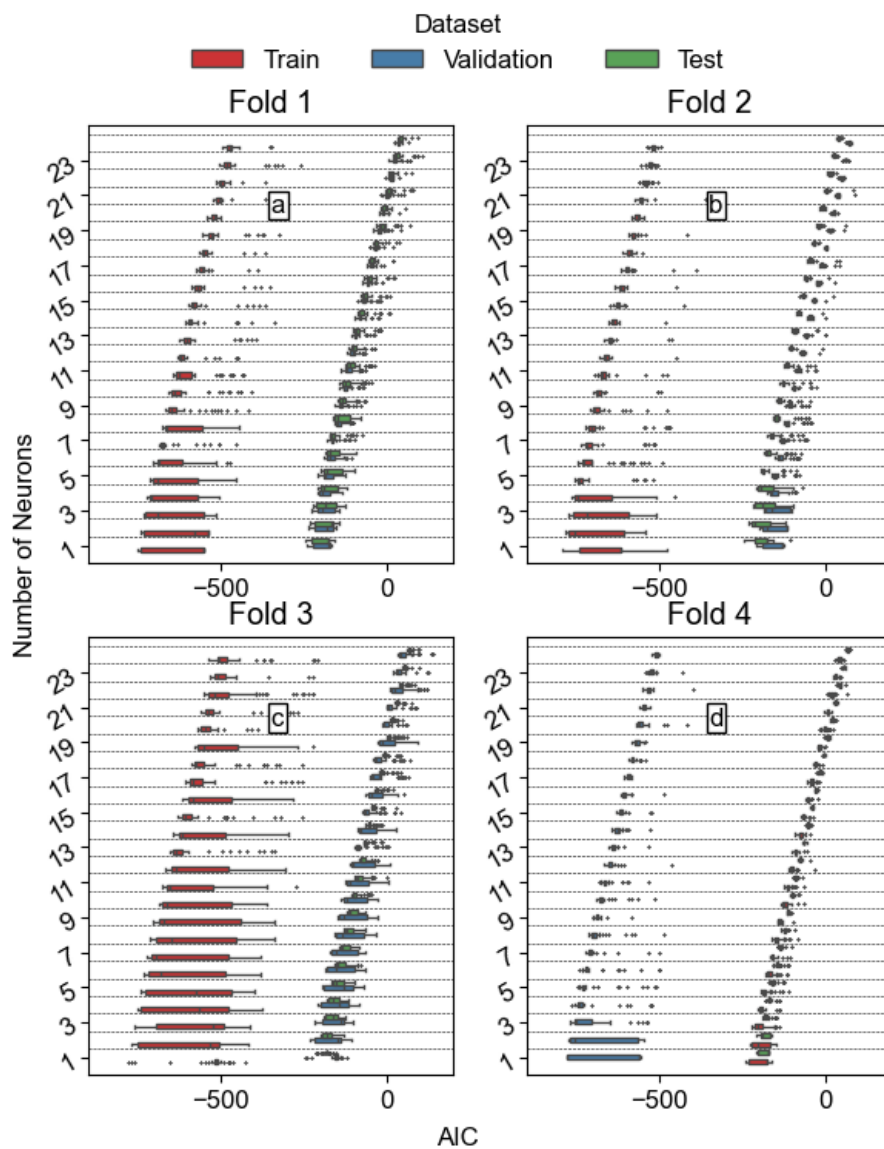


Figure B.106- AIC distributions across different neuron counts for well 293-3 in each cross-validation fold, DMS scenario.

DSM-FFNN model plateau by fold graph for W1:

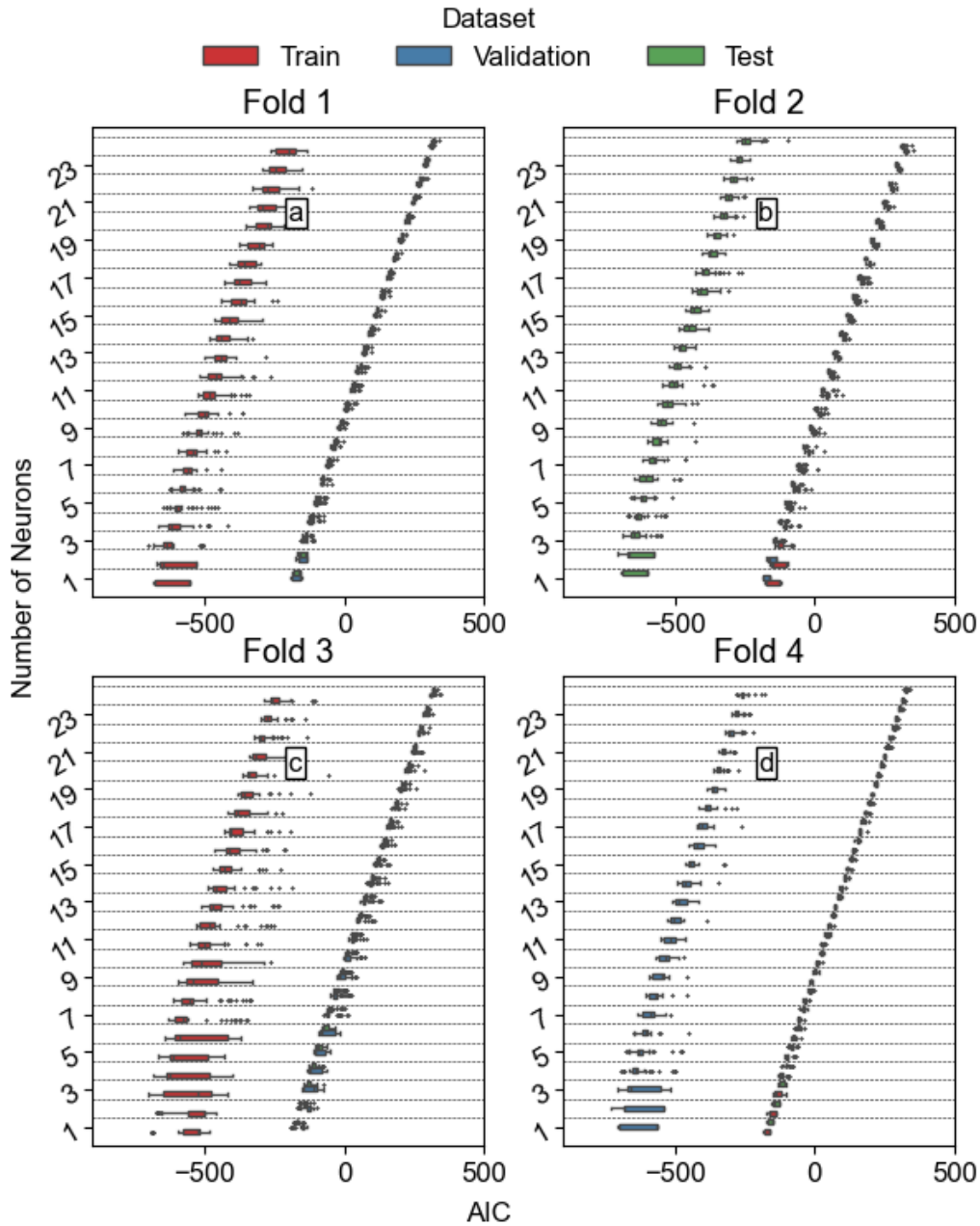


Figure B.107- AIC distributions across different neuron counts for W1 combined model in each cross-validation fold, DSM scenario.

DMS-FFNN model plateau by fold graph for W1:

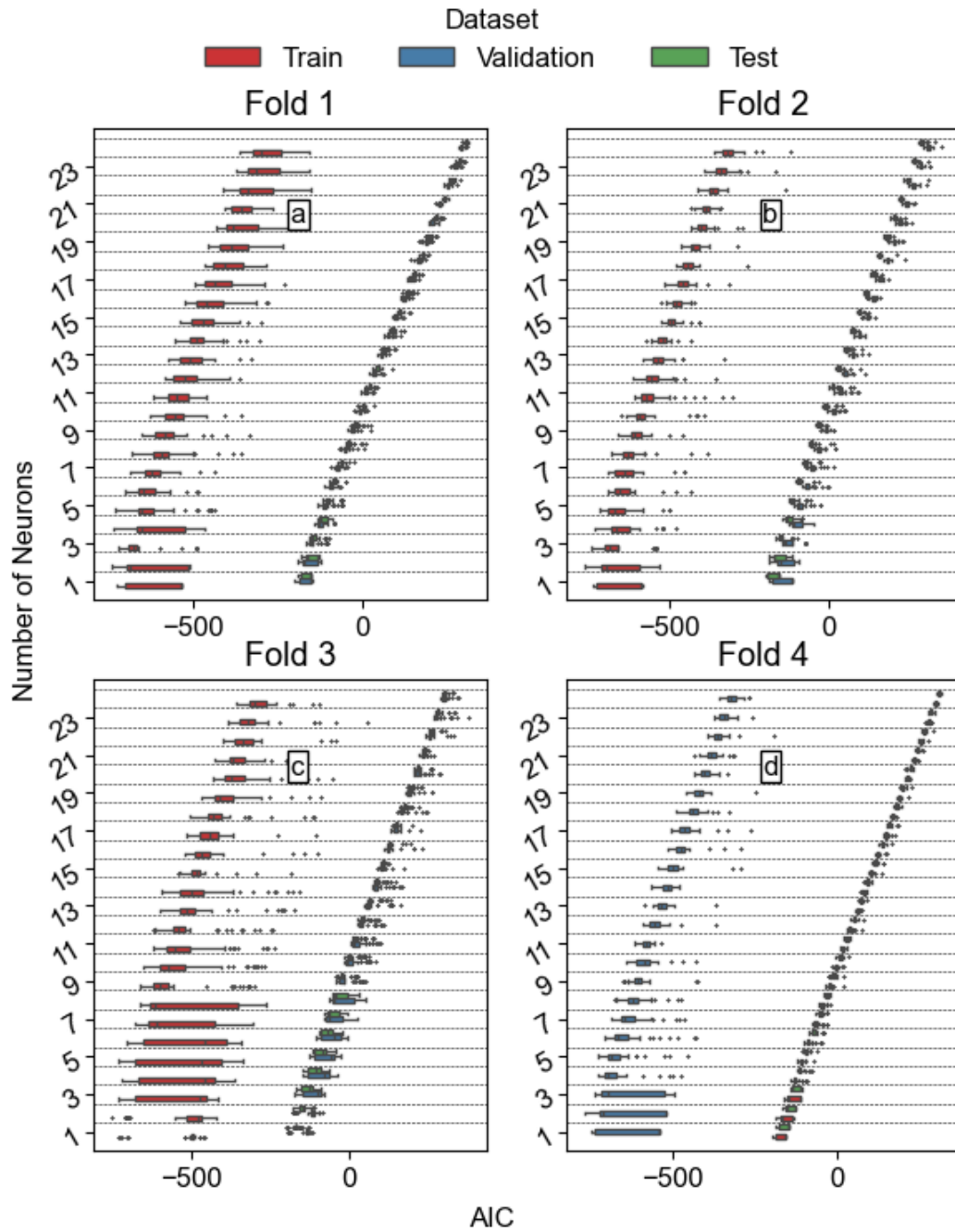


Figure B.108- AIC distributions across different neuron counts for W1 combined model in each cross-validation fold, DMS scenario.

DSM-FFNN model plateau by fold graph for 439-1:

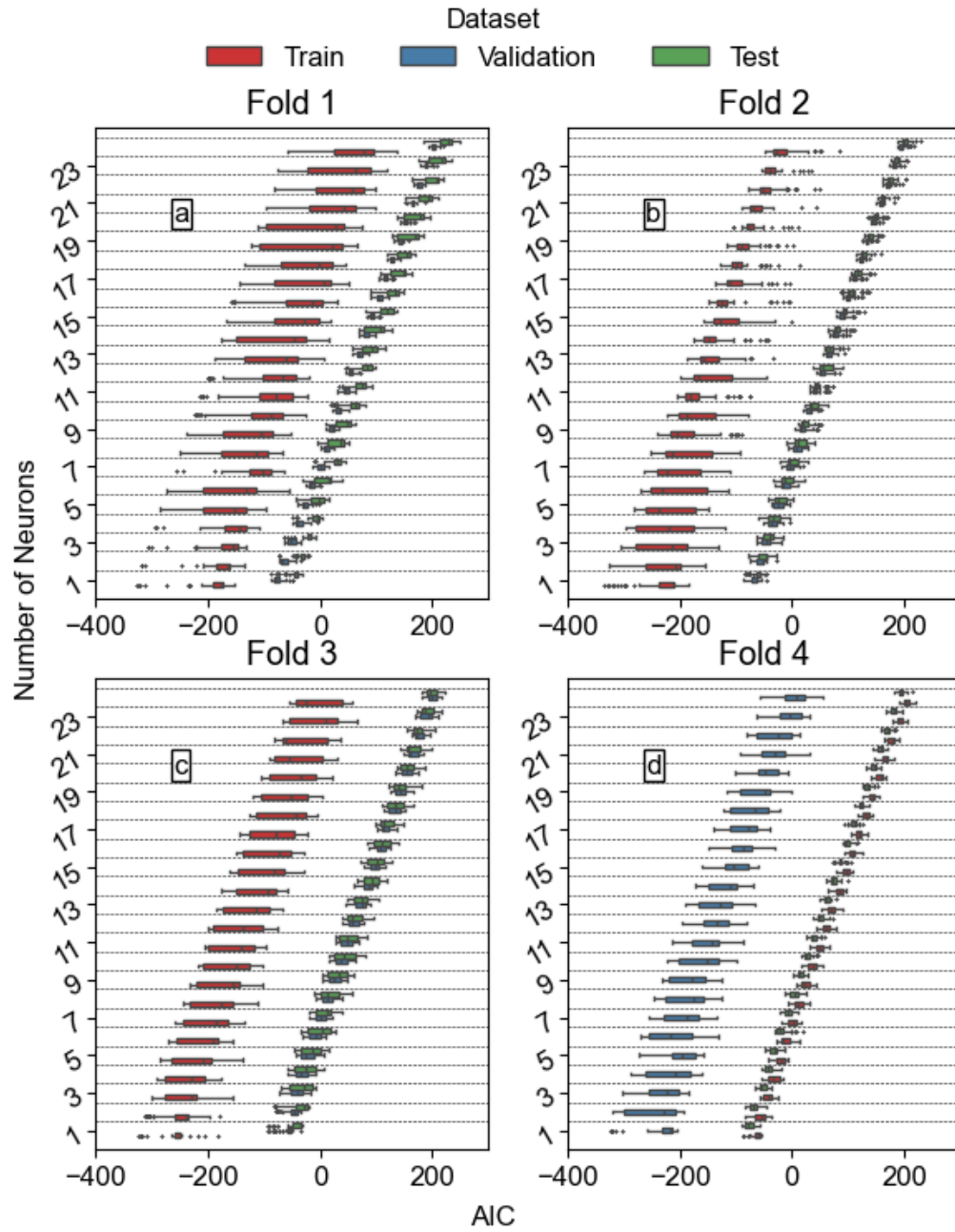


Figure B.109- AIC distributions across different neuron counts for well 439-1 in each cross-validation fold, DSM scenario.

DMS-FFNN model plateau by fold graph for 439-1:

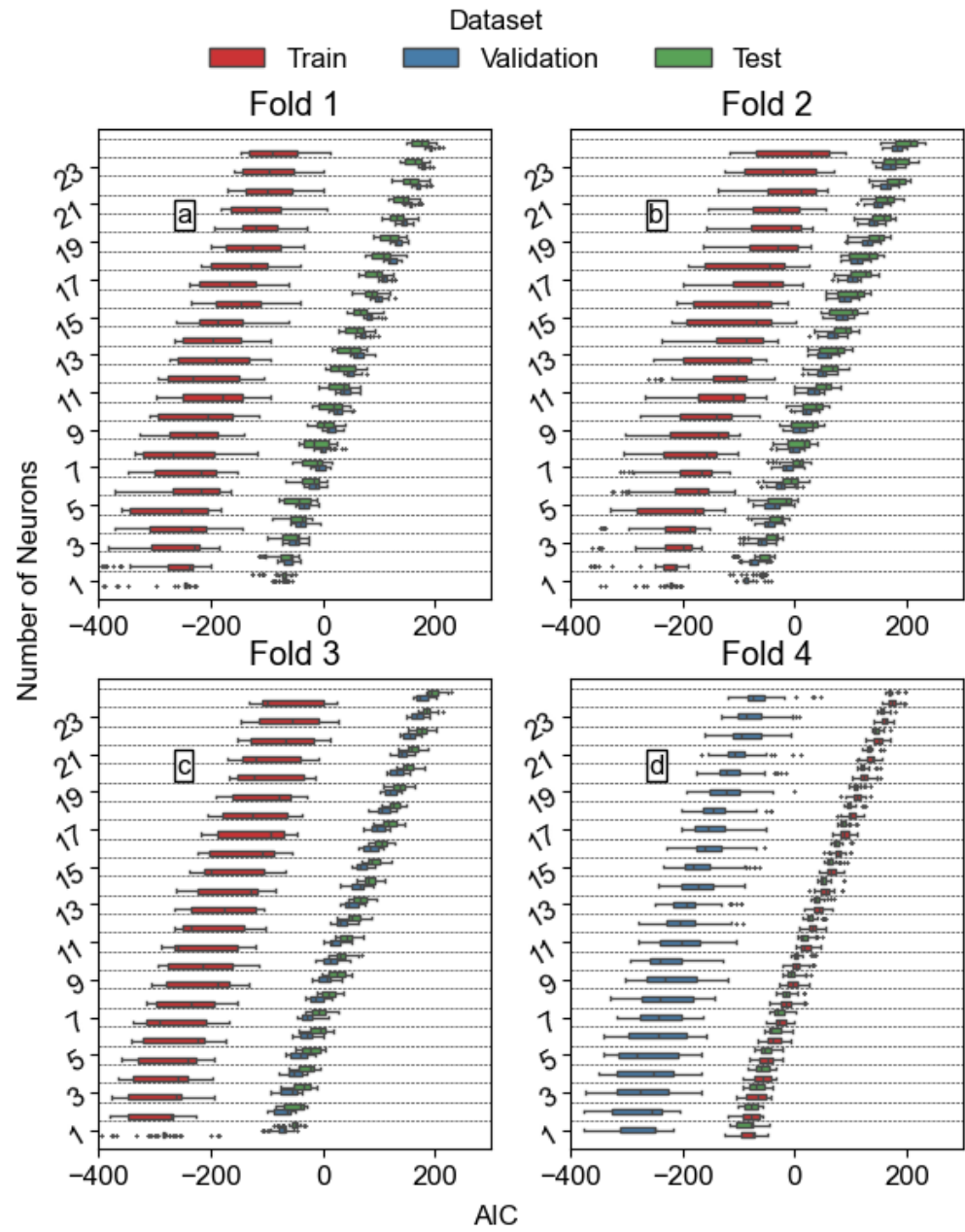


Figure B.110- AIC distributions across different neuron counts for well 439-1 in each cross-validation fold, DMS scenario.

DSM-FFNN model plateau by fold graph for 440-1:

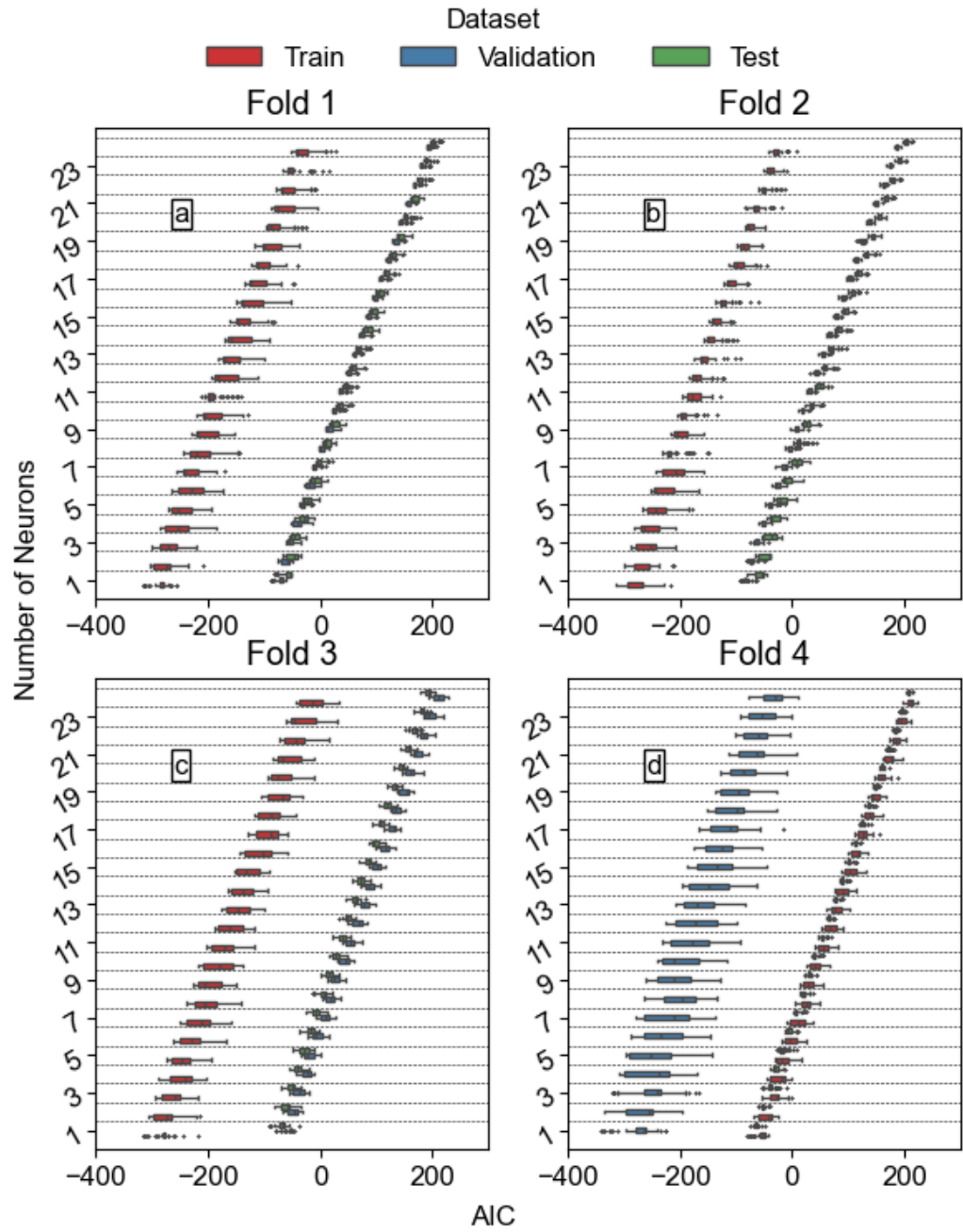


Figure B.111- AIC distributions across different neuron counts for well 440-1 in each cross-validation fold, DSM scenario.

DMS-FFNN model plateau by fold graph for 440-1:

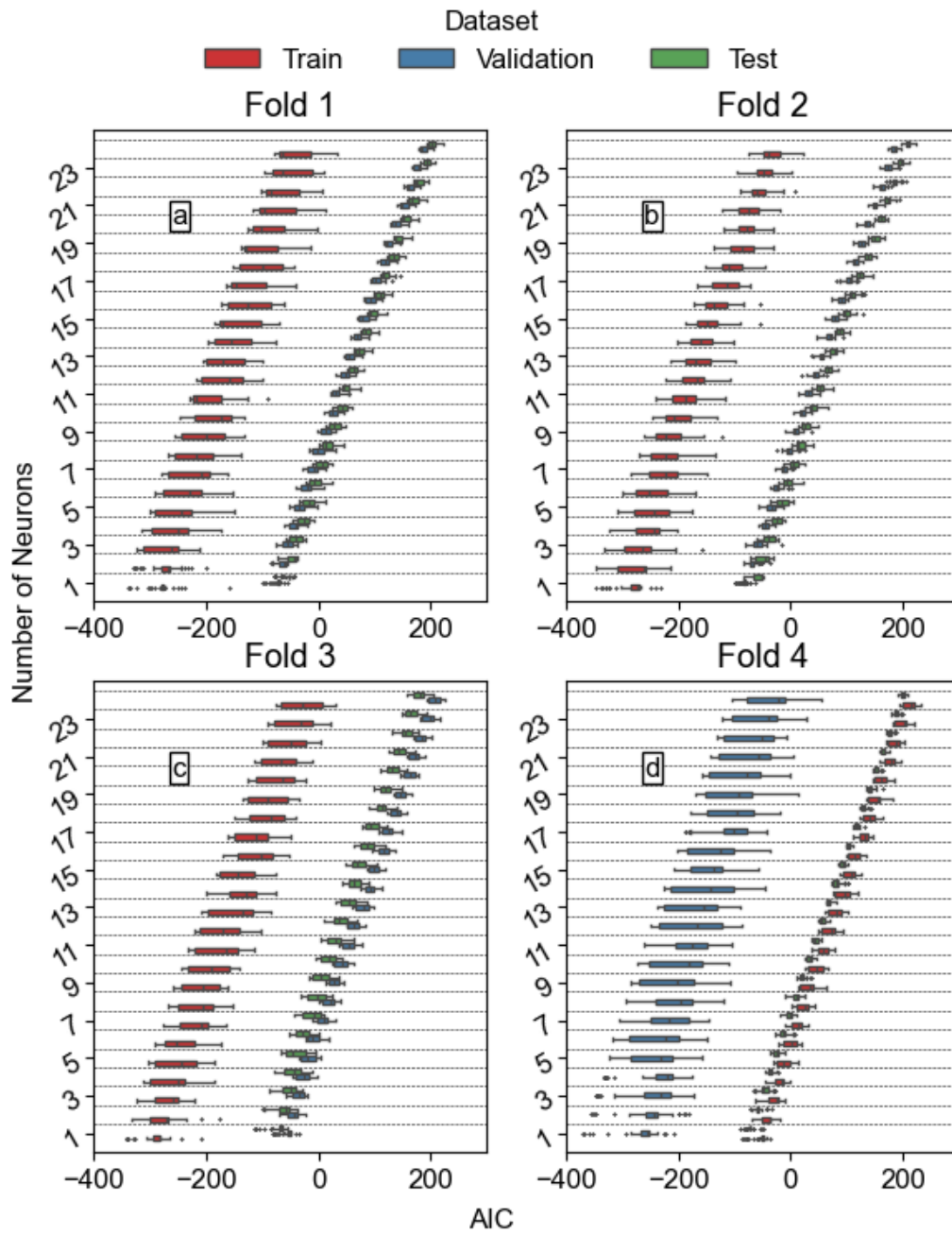


Figure B.112- AIC distributions across different neuron counts for well 440-1 in each cross-validation fold, DMS scenario.

DSM-FFNN model plateau by fold graph for 442-1:

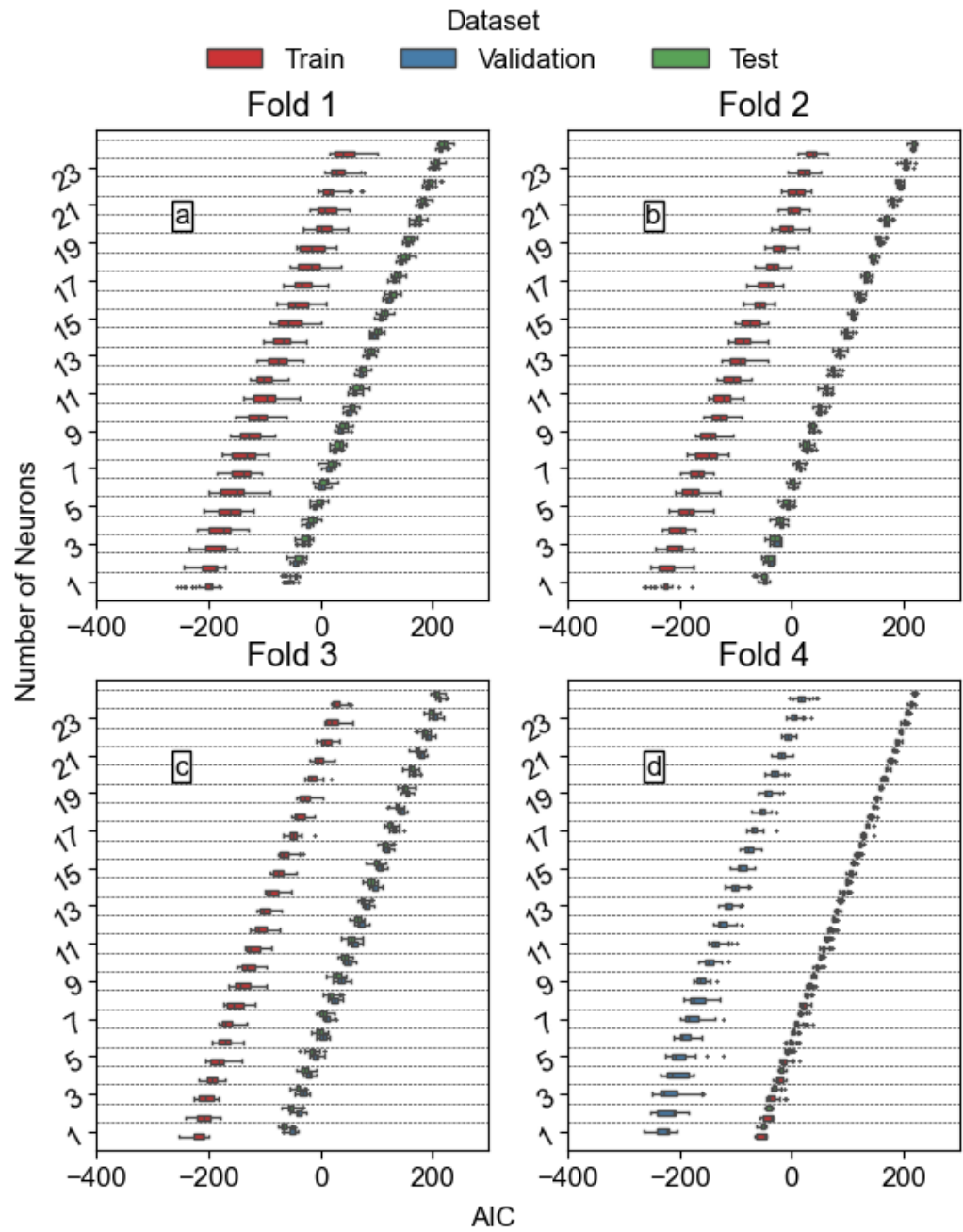


Figure B.113- AIC distributions across different neuron counts for well 442-1 in each cross-validation fold, DSM scenario.

DMS-FFNN model plateau by fold graph for 442-1:

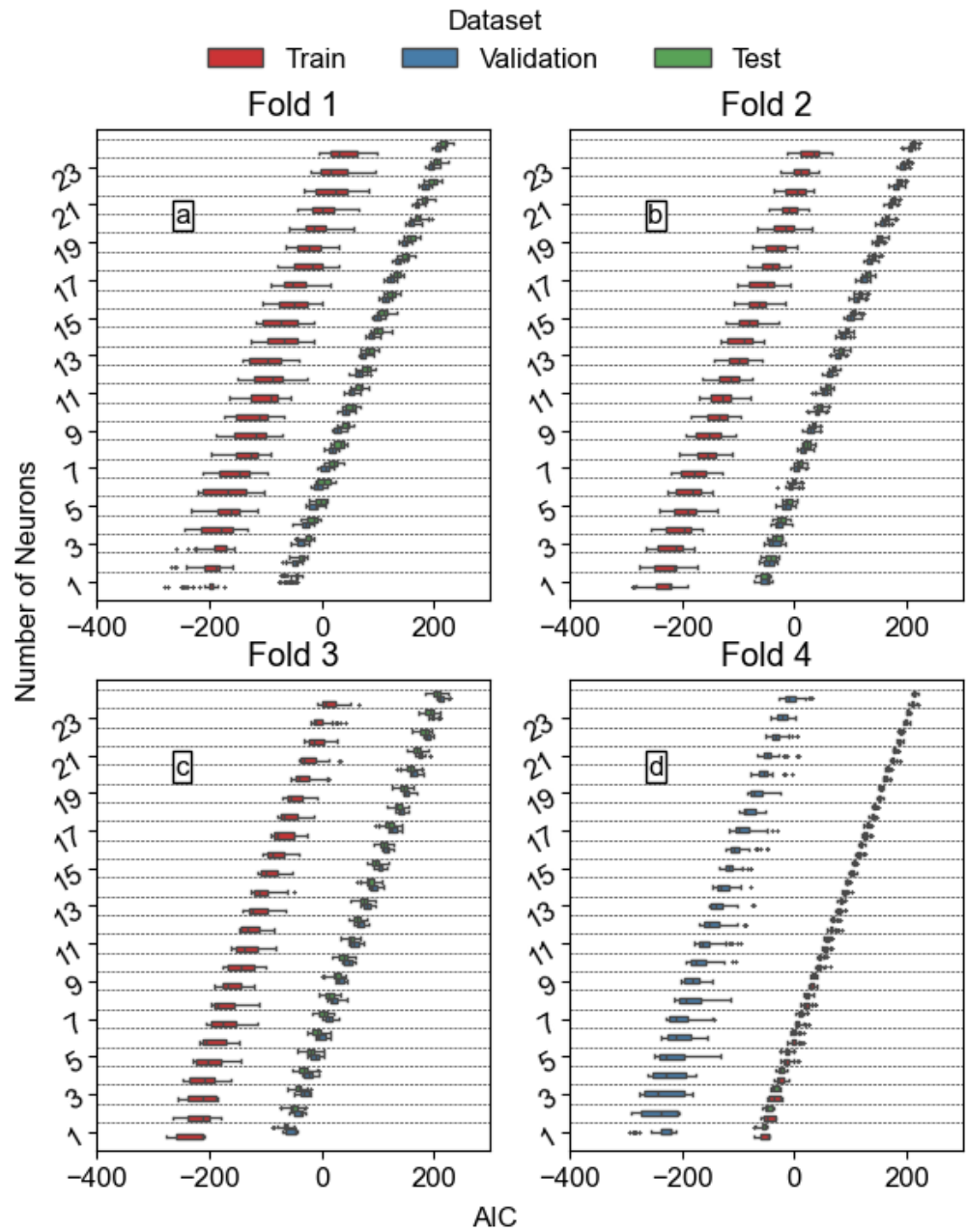


Figure B.114- AIC distributions across different neuron counts for well 442-1 in each cross-validation fold, DMS scenario.

DSM-FFNN model plateau by fold graph for W2:

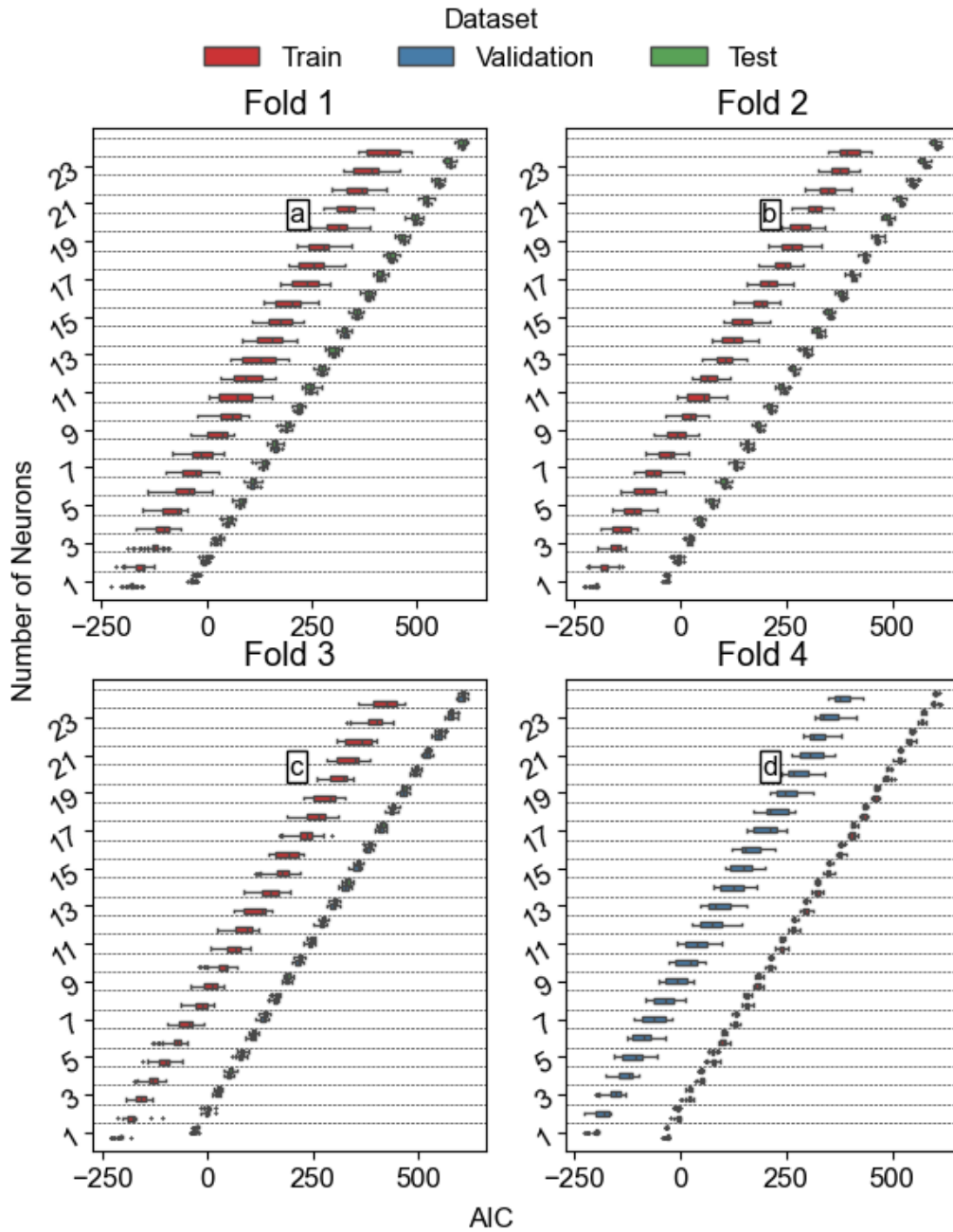


Figure B.115- AIC distributions across different neuron counts for W2 combined in each cross-validation fold, DSM scenario.

DMS-FFNN model plateau by fold graph for W2:

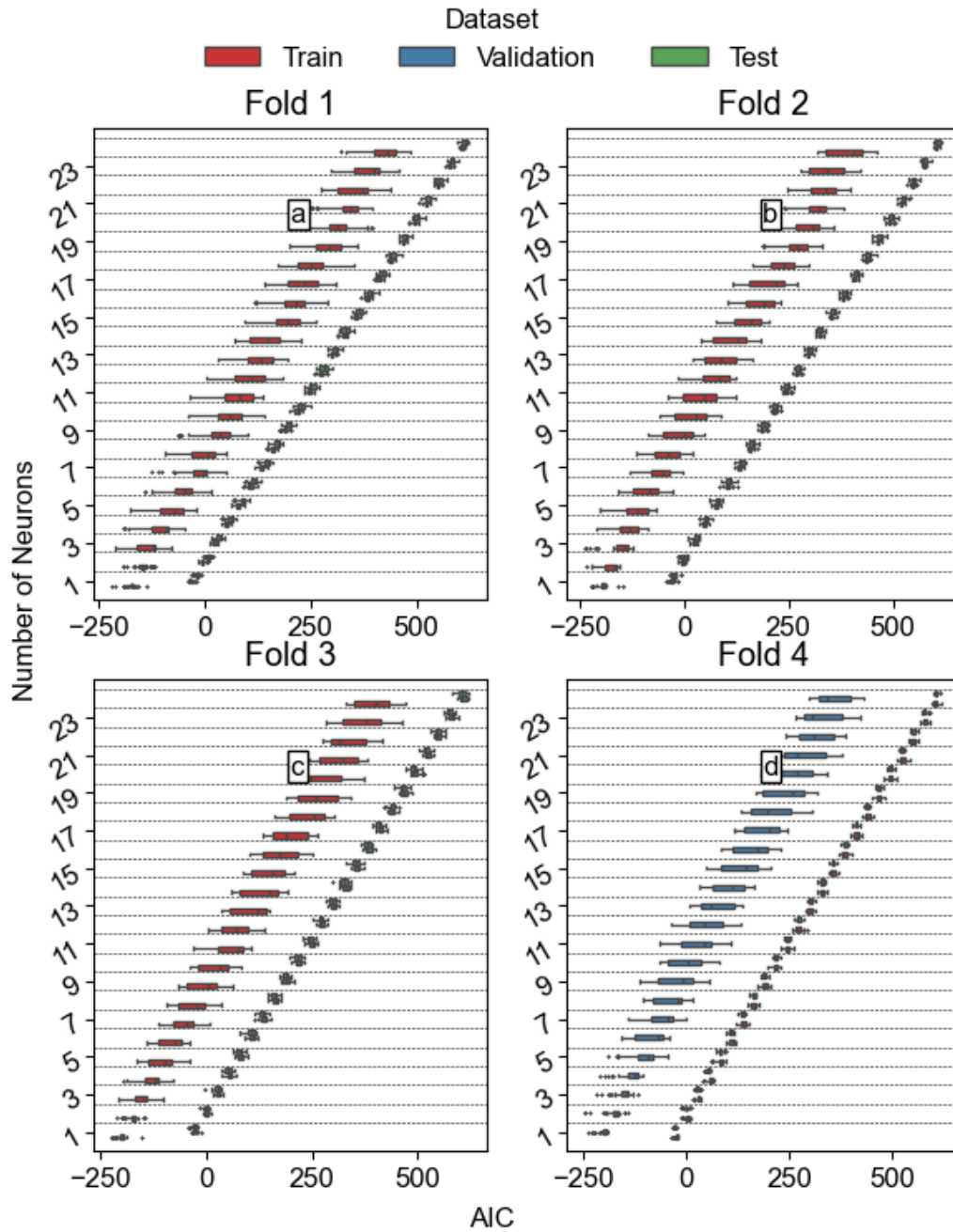


Figure B.116- AIC distributions across different neuron counts for W2 combined in each cross-validation fold, DMS scenario.

DSM-FFNN model plateau by fold graph for 244-2:

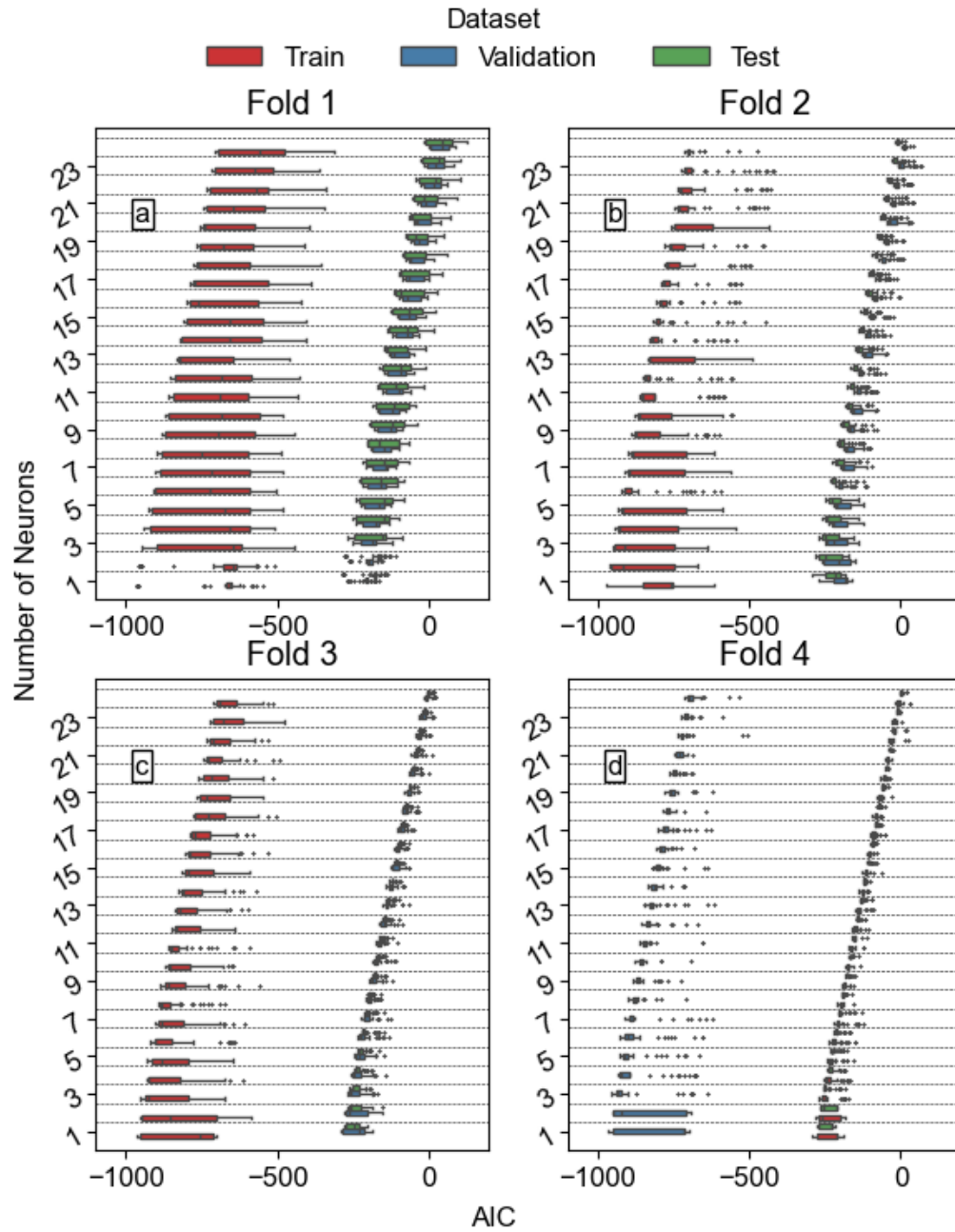


Figure B.117- AIC distributions across different neuron counts for well 244-2 in each cross-validation fold, DSM scenario.

DMS-FFNN model plateau by fold graph for 244-2:

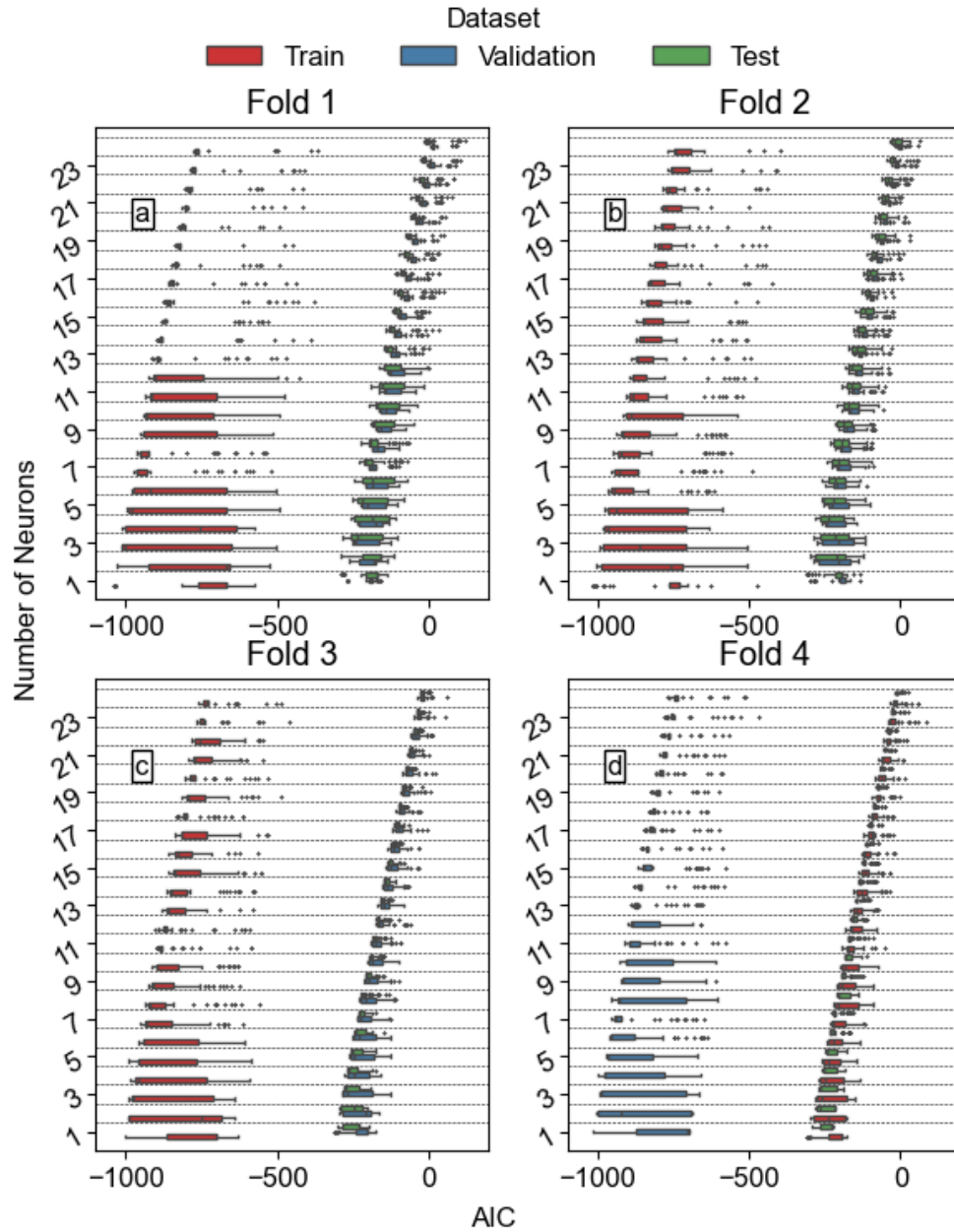


Figure B.118- AIC distributions across different neuron counts for well 244-2 in each cross-validation fold, DMS scenario.

DSM-FFNN model plateau by fold graph for 245-2:

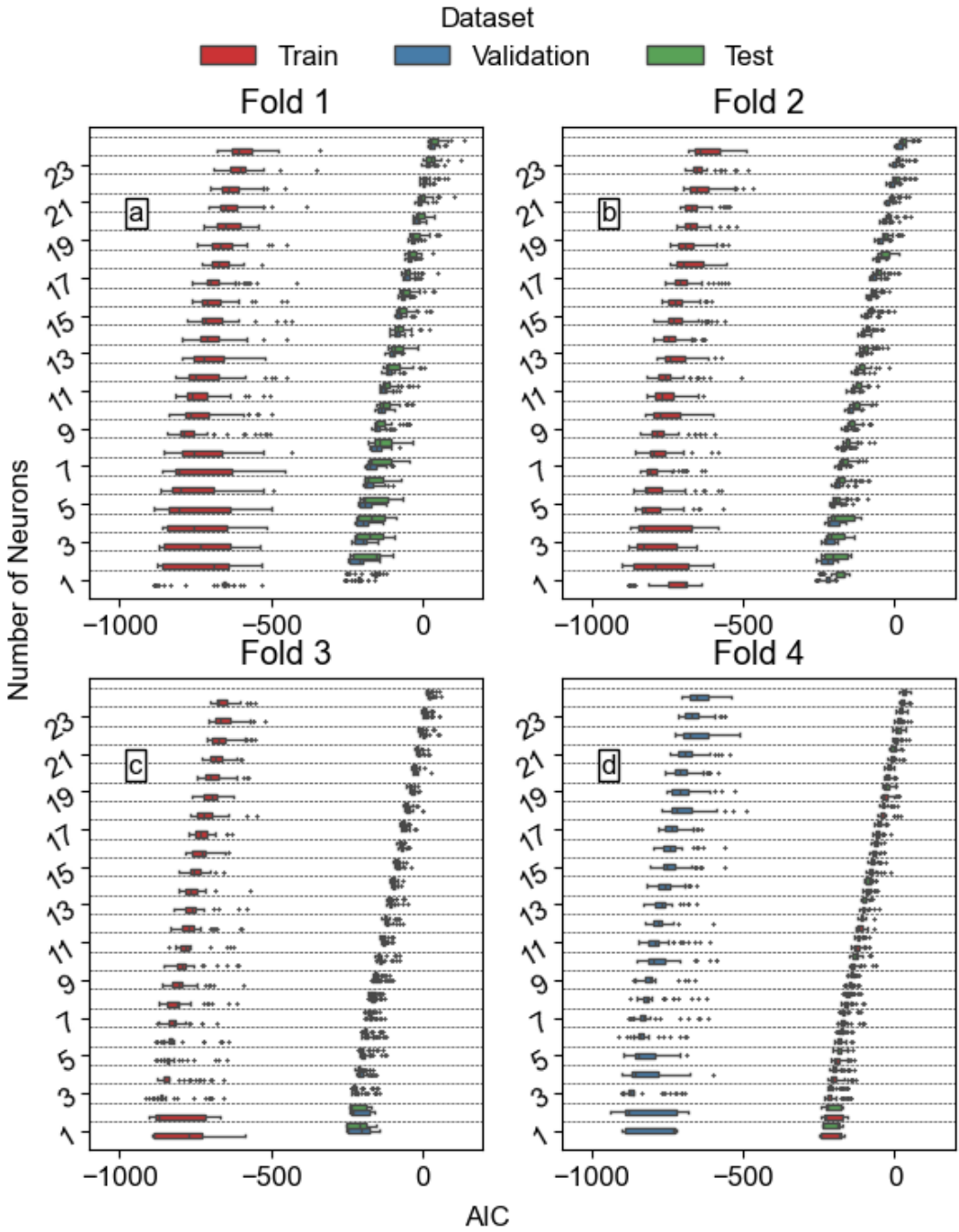


Figure B.119- AIC distributions across different neuron counts for well 245-2 in each cross-validation fold, DSM scenario.

DMS-FFNN model plateau by fold graph for 245-2:

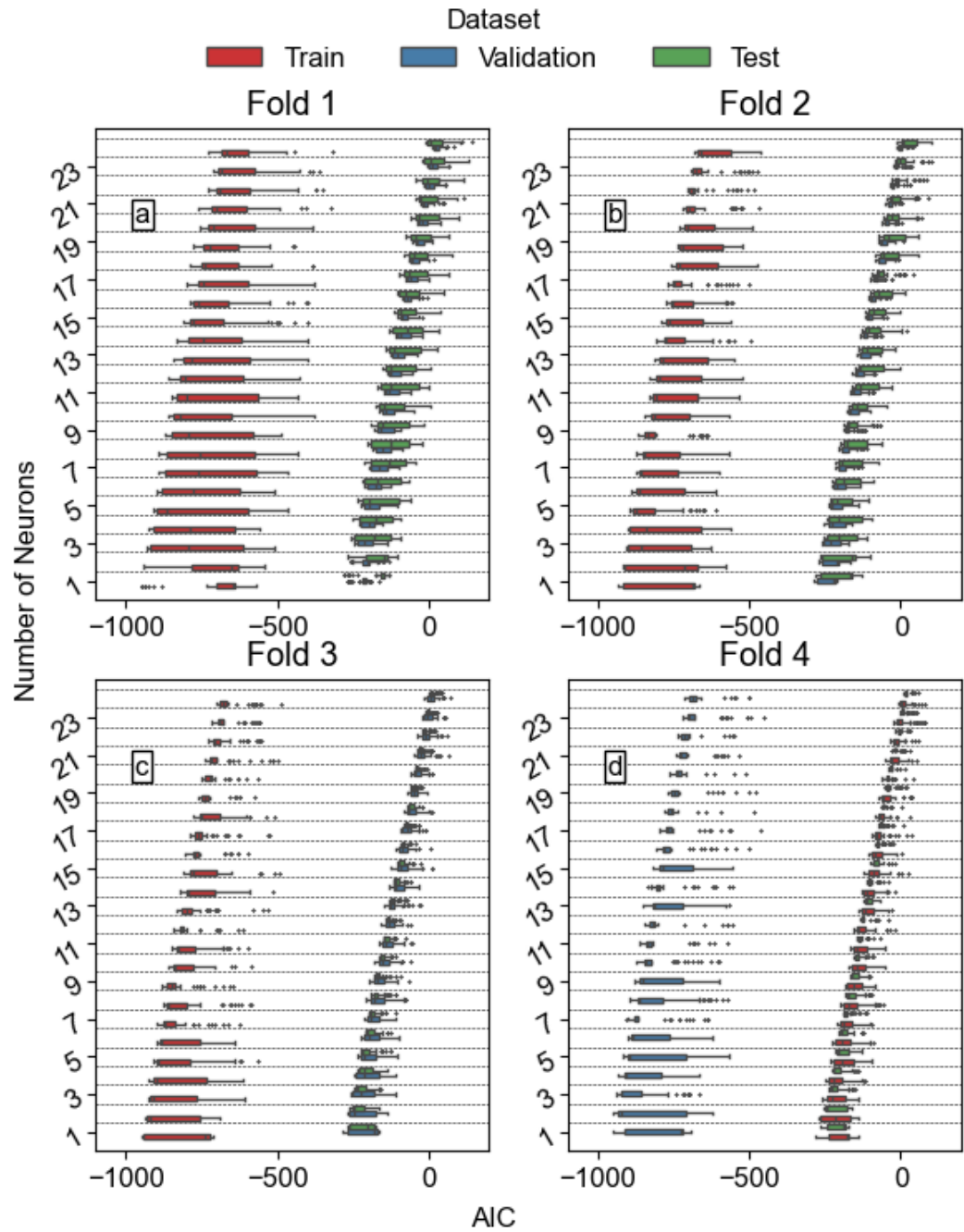


Figure B.120- AIC distributions across different neuron counts for well 245-2 in each cross-validation fold, DMS scenario.

DSM-FFNN model plateau by fold graph for W3:

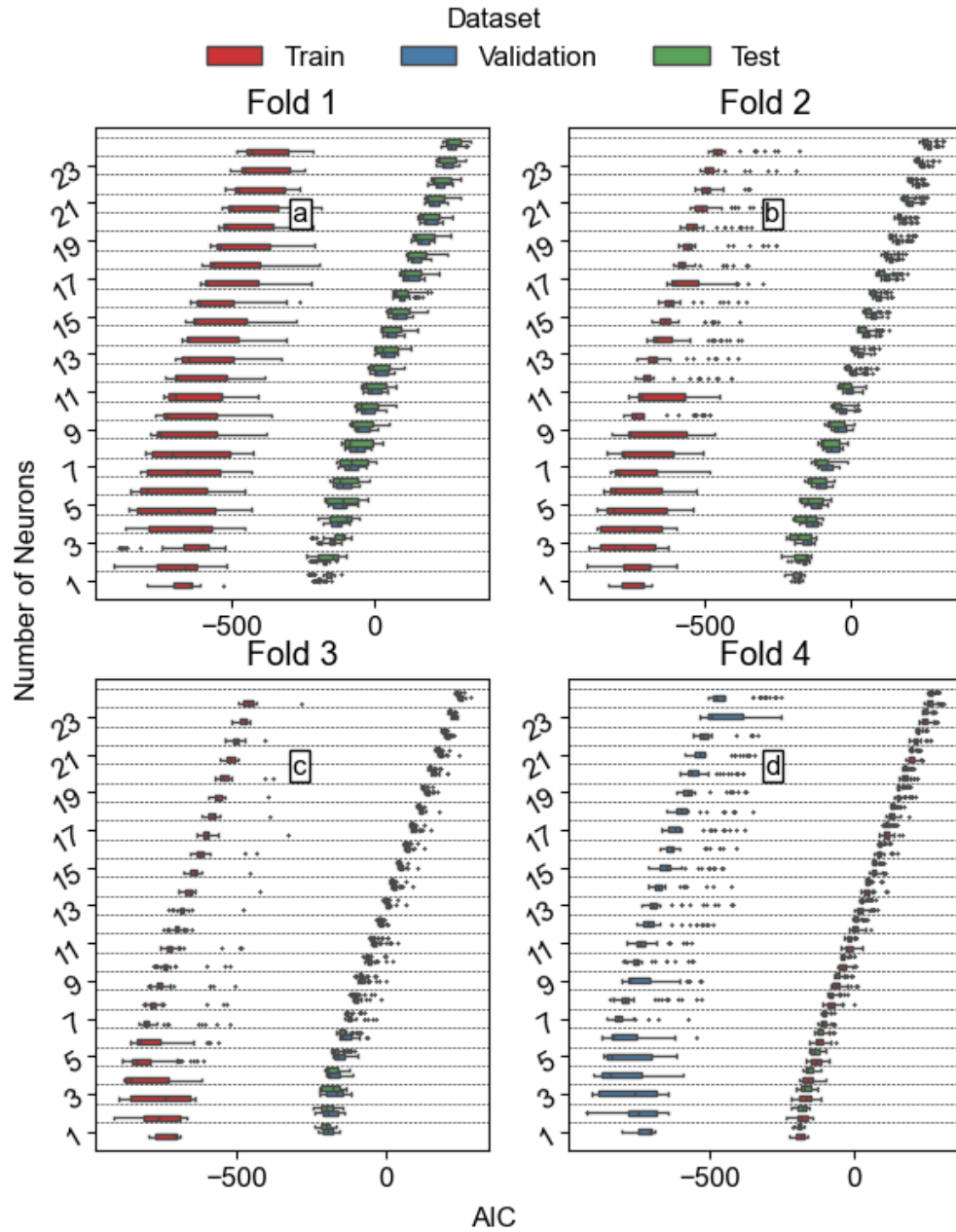


Figure B.121- AIC distributions across different neuron counts for W3 combined model in each cross-validation fold, DSM scenario.

DMS-FFNN model plateau by fold graph for W3:

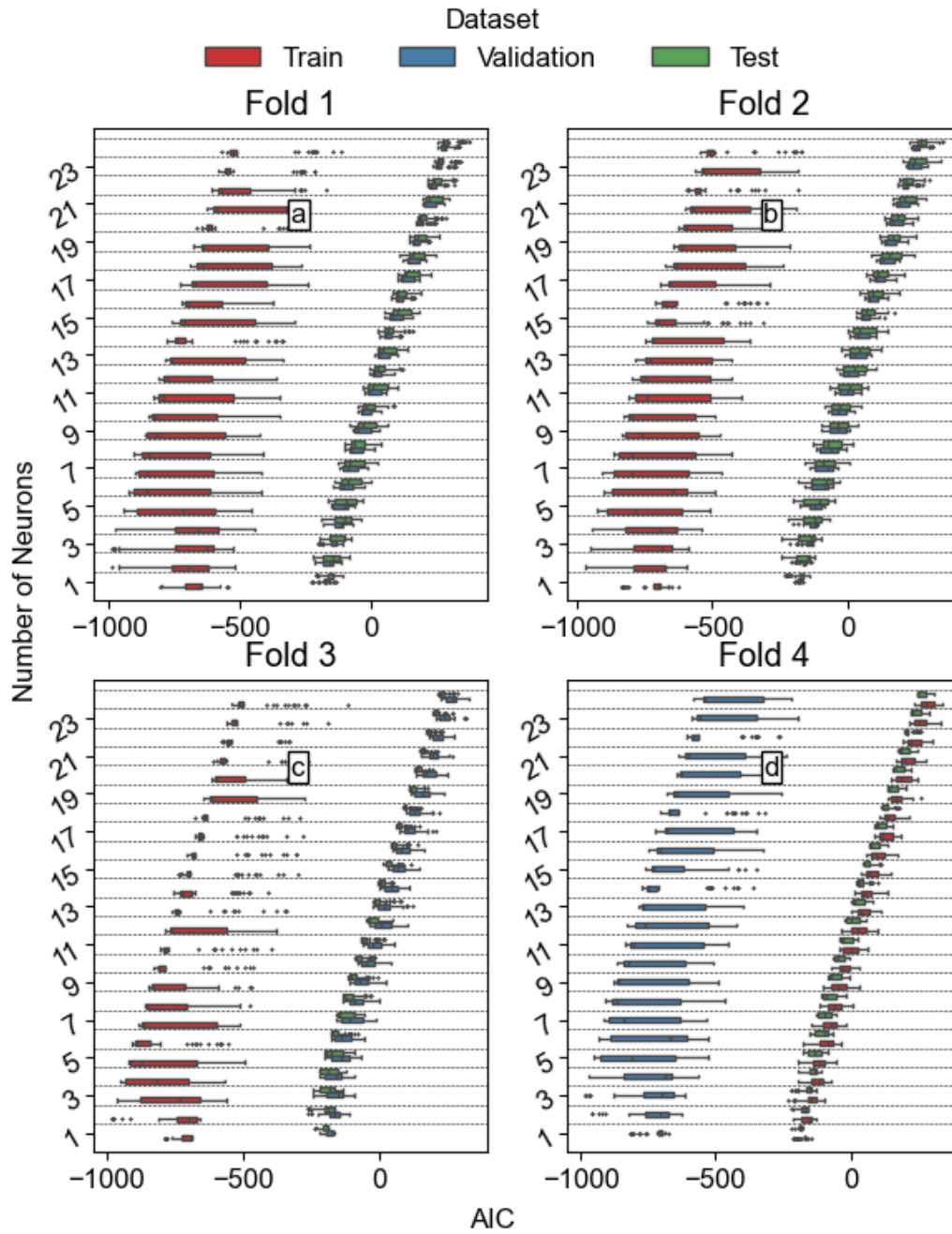


Figure B.122- AIC distributions across different neuron counts for W3 combined model in each cross-validation fold, DMS scenario.

DSM-FFNN model plateau by fold graph for 443-1:

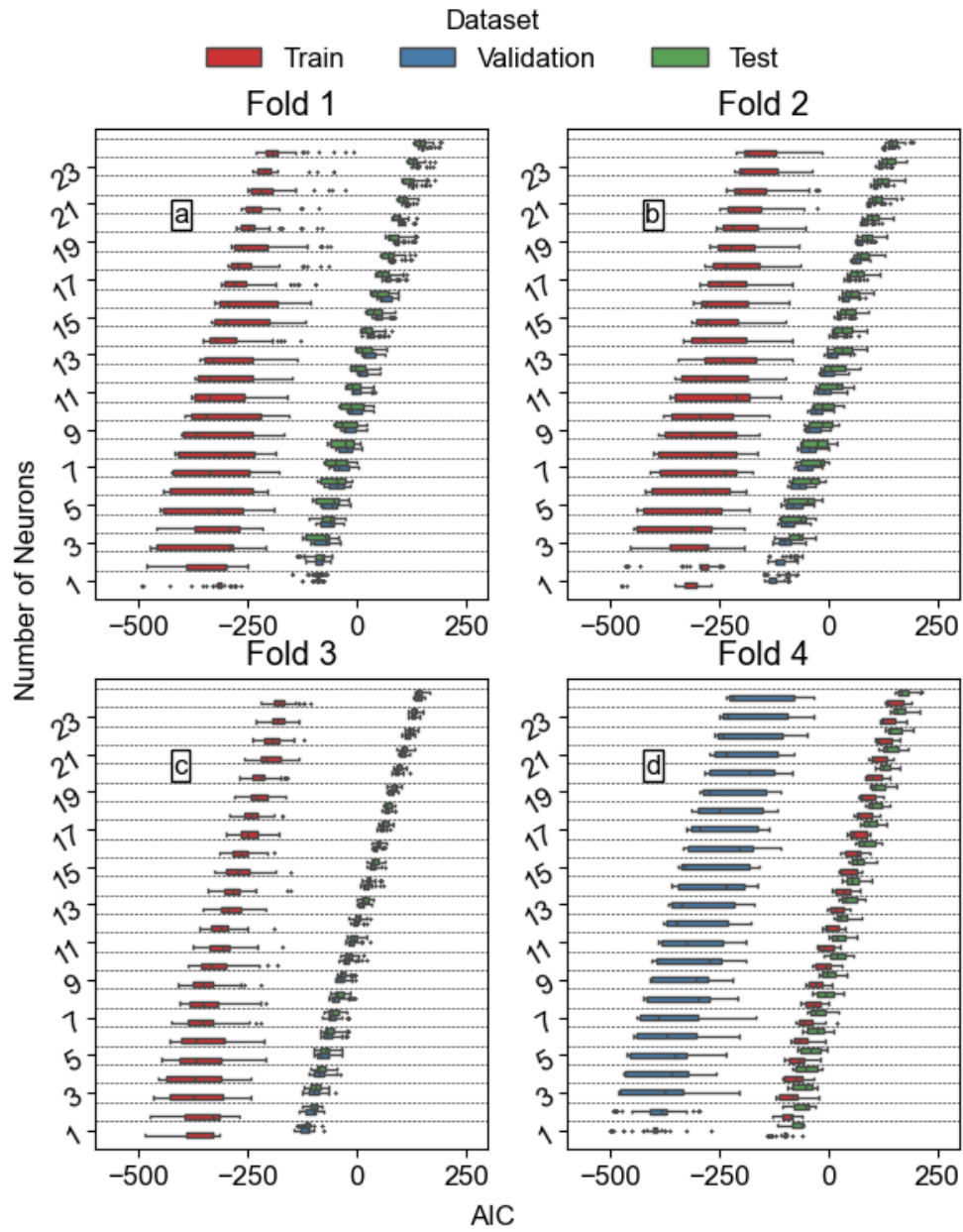


Figure B.123- AIC distributions across different neuron counts for well 443-1 in each cross-validation fold, DSM scenario.

DMS-FFNN model saturation overall graph for 443-1:

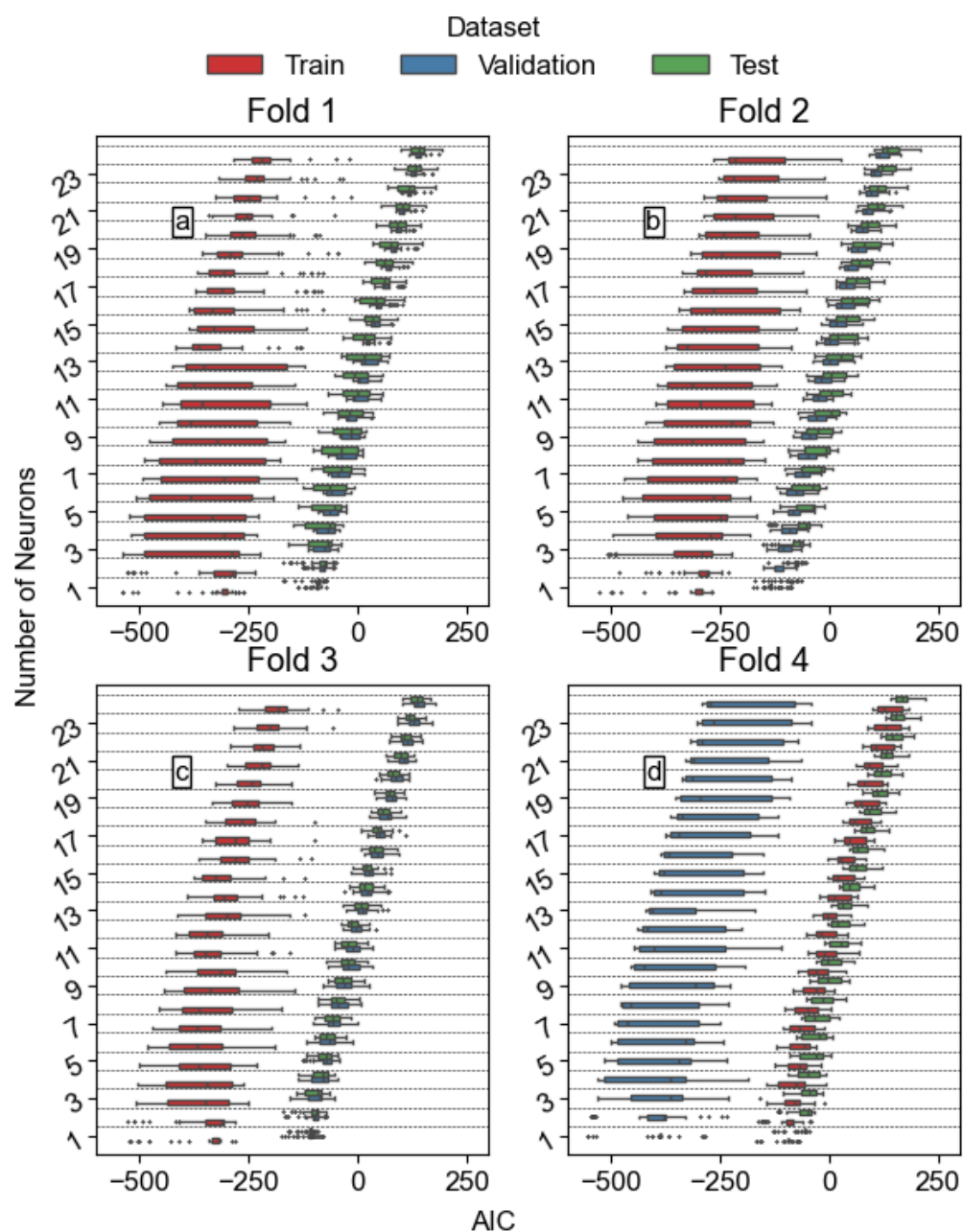


Figure B.124- AIC distributions across different neuron counts for well 443-1 in each cross-validation fold, DMS scenario.

DSM-FFNN model saturation overall graph for the aggregated model:

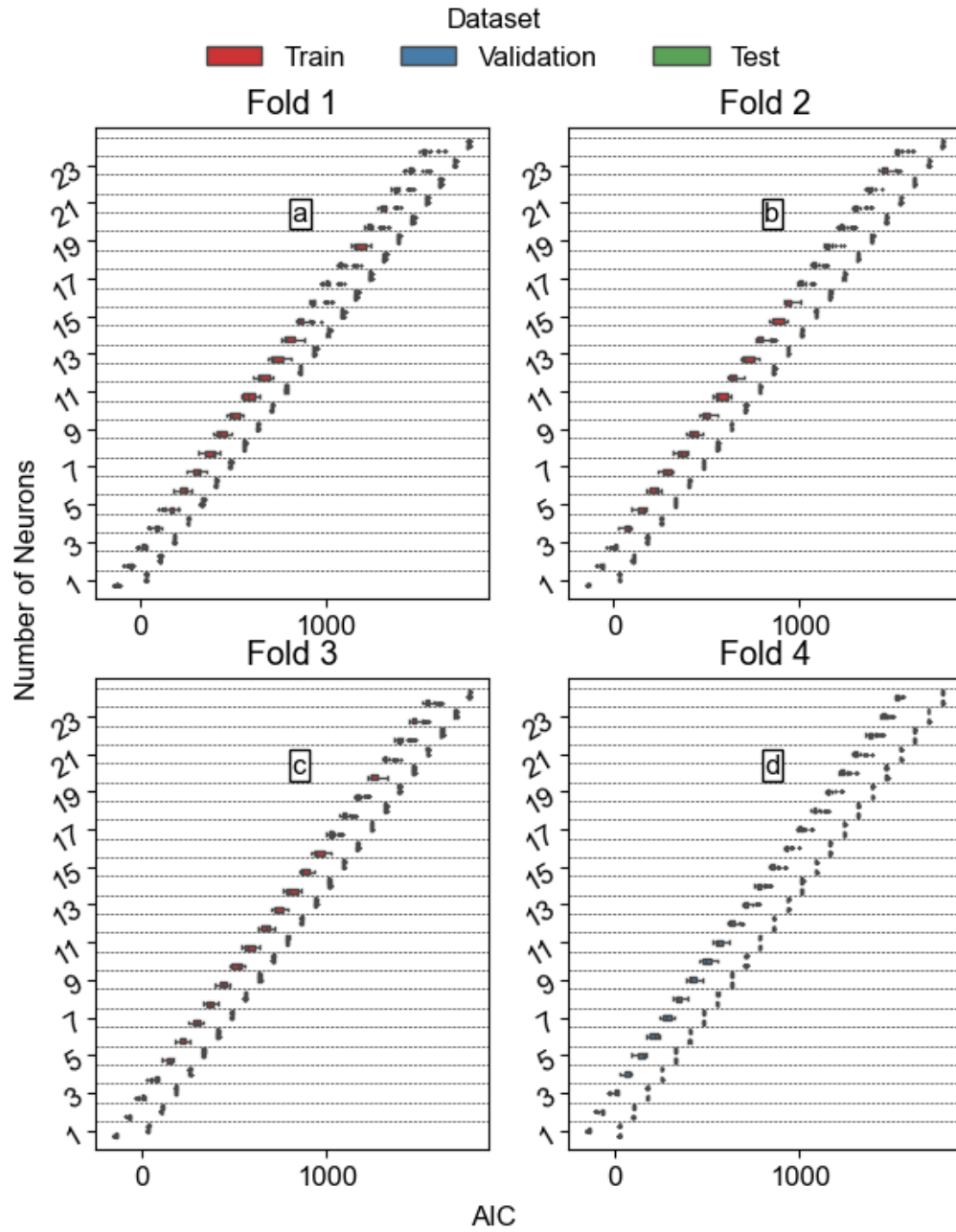


Figure B.125- AIC distributions across different neuron counts for the aggregated model in each cross-validation fold, DSM scenario.

DMS-FFNN model saturation overall graph for the aggregated model:

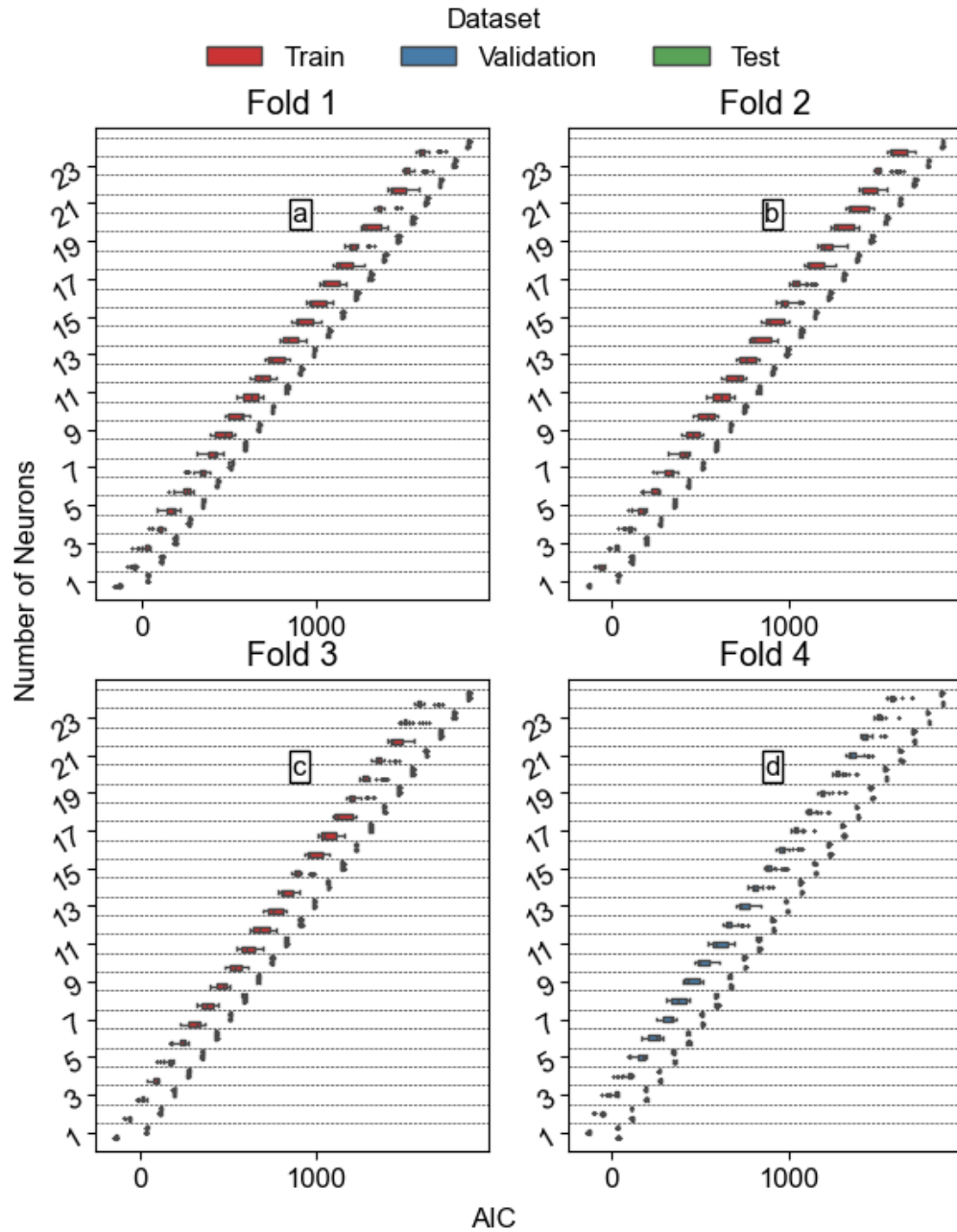


Figure B.126- AIC distributions across different neuron counts for the aggregated model in each cross-validation fold, DMS scenario.

Performance Plateau Overall Plots (AIC)

DSM-FFNN model plateau overall graph for 293-3:

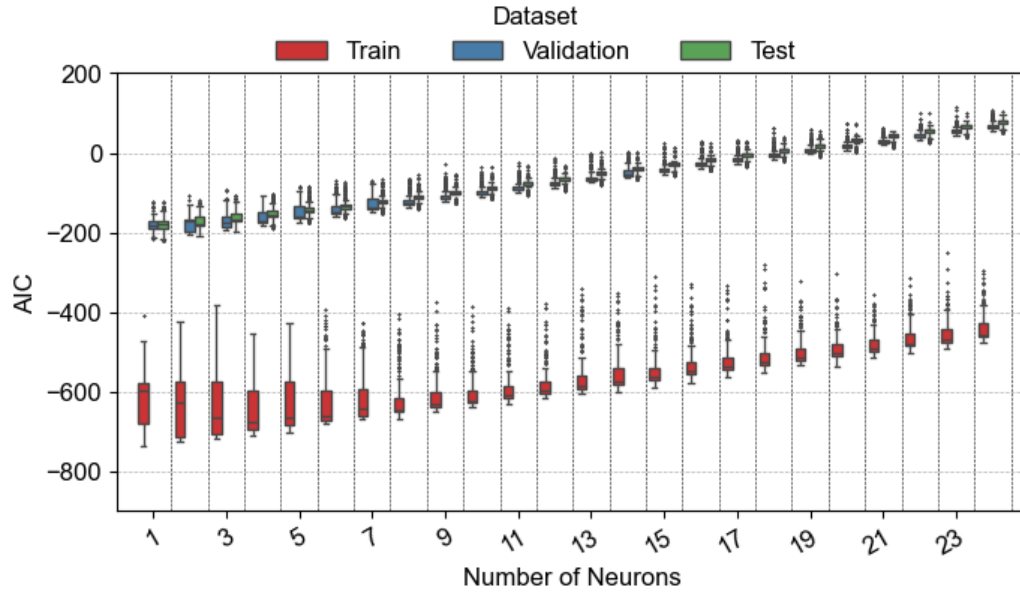


Figure B.127- Combined AIC distributions for well 293-3, DSM scenario.

DMS-FFNN model plateau overall graph for 293-3:

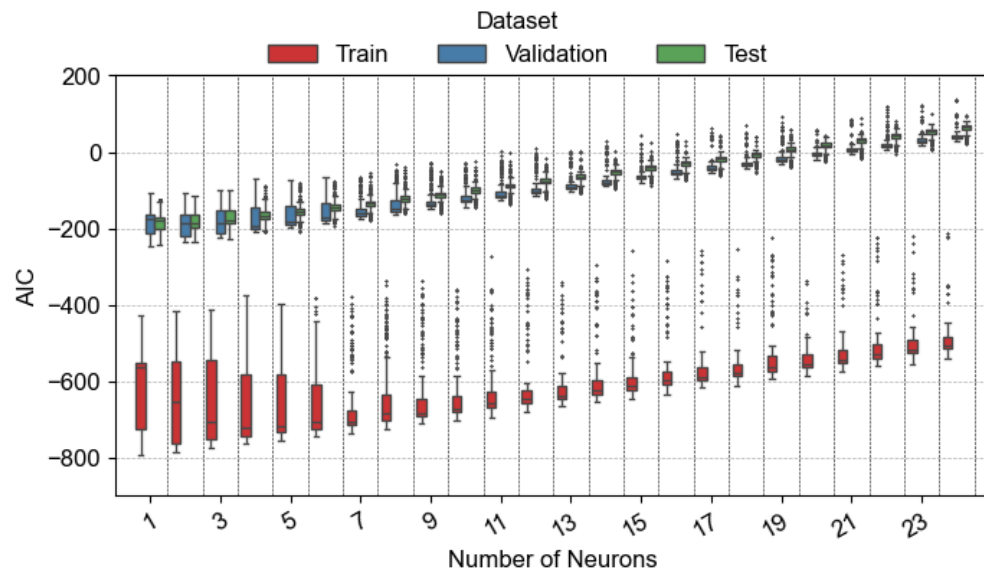


Figure B.128- Combined AIC distributions for well 293-3, DMS scenario.

DSM-FFNN model plateau overall graph for W1:

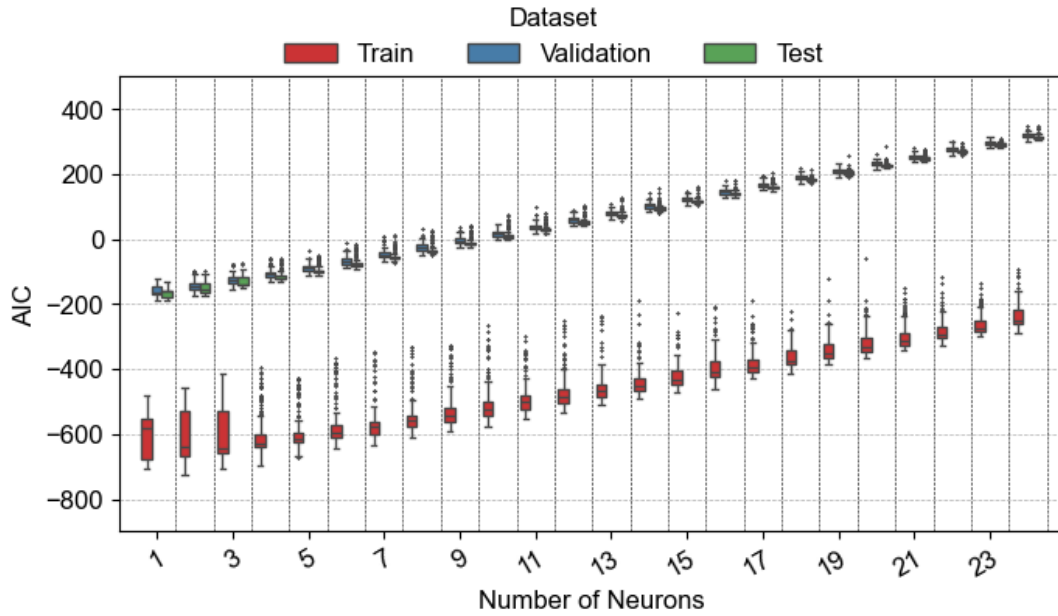


Figure B.129- Combined AIC distributions for W1 combined model, DSM scenario.

DMS-FFNN model plateau overall graph for W1:

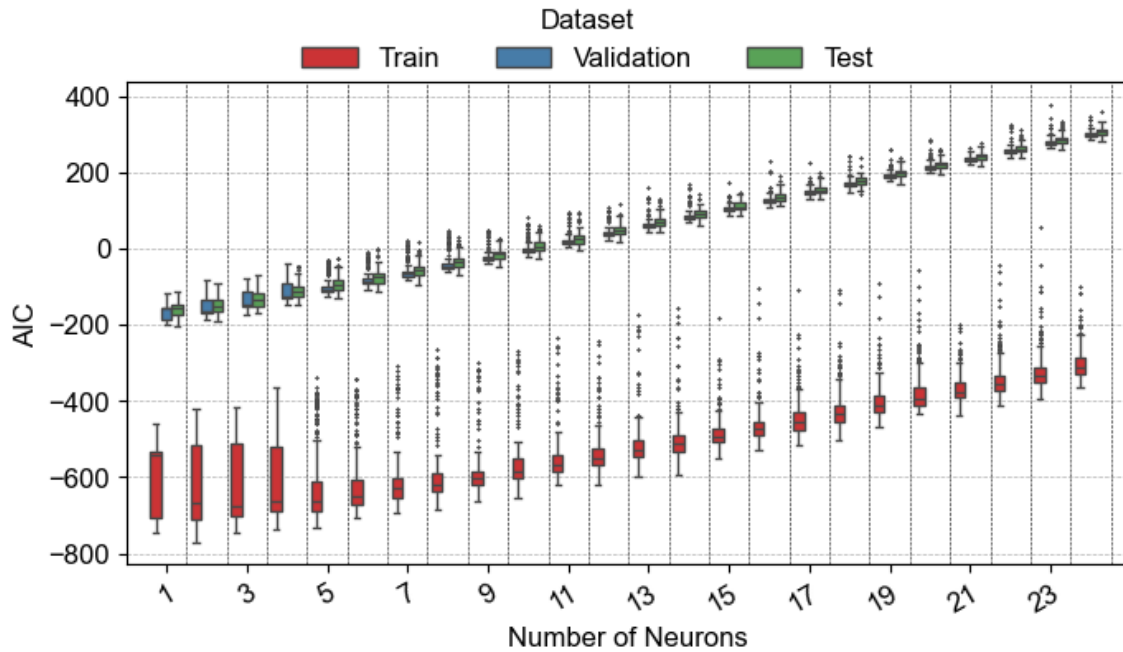


Figure B.130- Combined AIC distributions for W1 combined model, DMS scenario.

DSM-FFNN model plateau overall graph for 439-1:

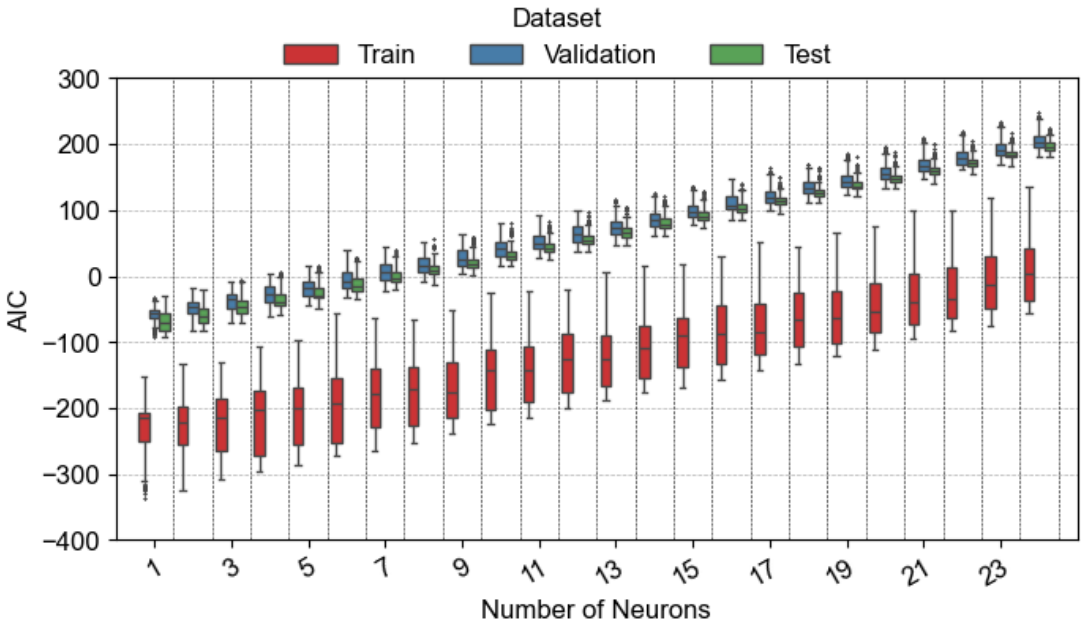


Figure B.131- Combined AIC distributions for well 439-1, DSM scenario.

DMS-FFNN model plateau overall graph for 439-1:

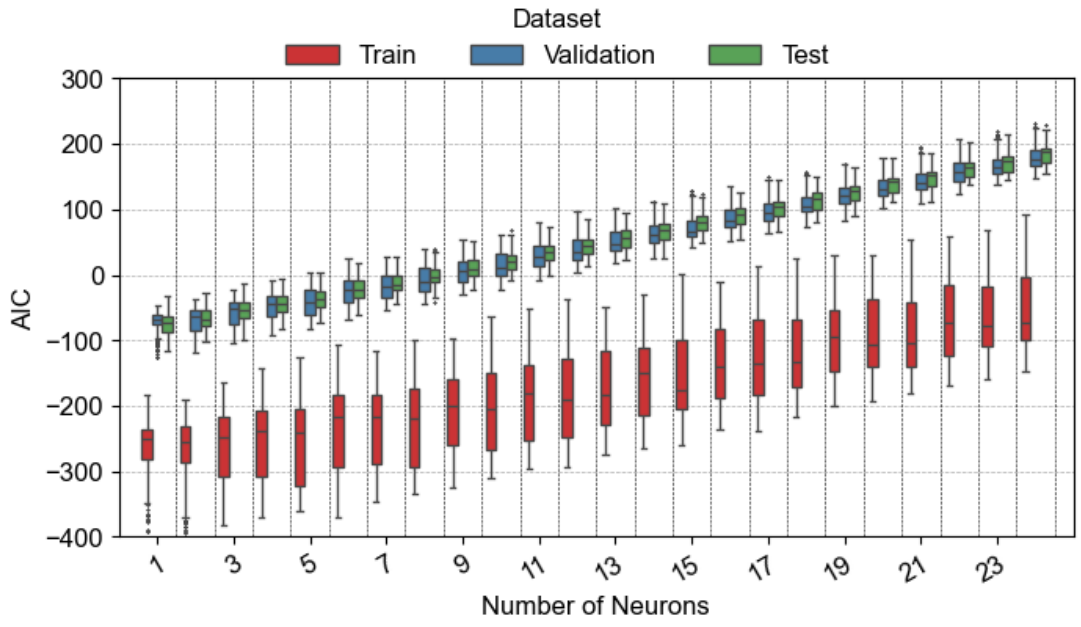


Figure B.132- Combined AIC distributions for well 439-1, DMS scenario.

DSM-FFNN model plateau overall graph for 440-1:

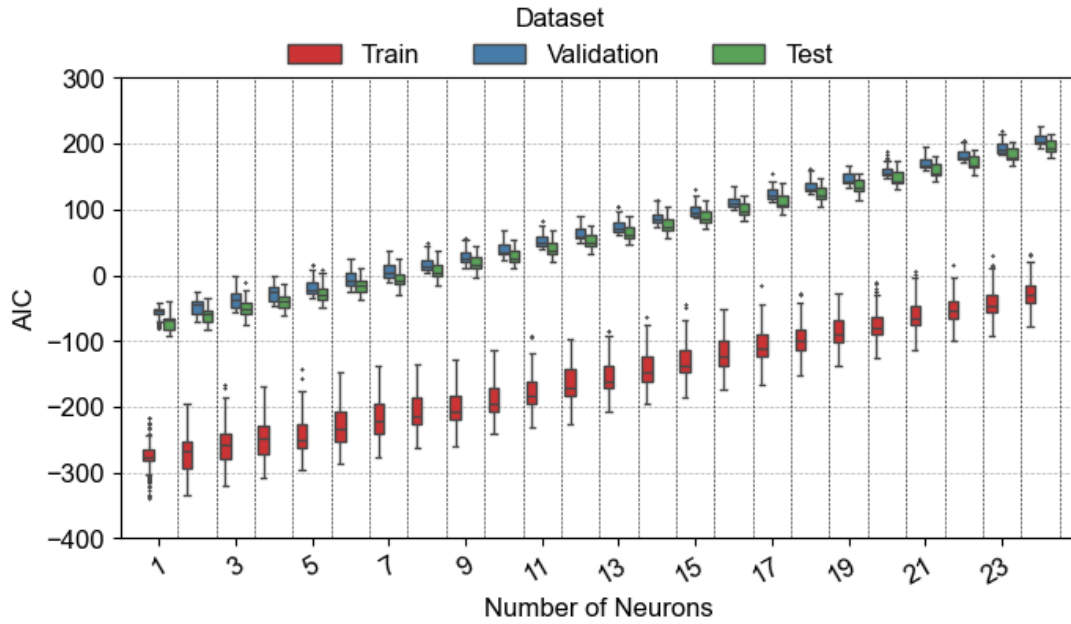


Figure B.133- Combined AIC distributions for well 440-1, DSM scenario.

DMS-FFNN model plateau overall graph for 440-1:

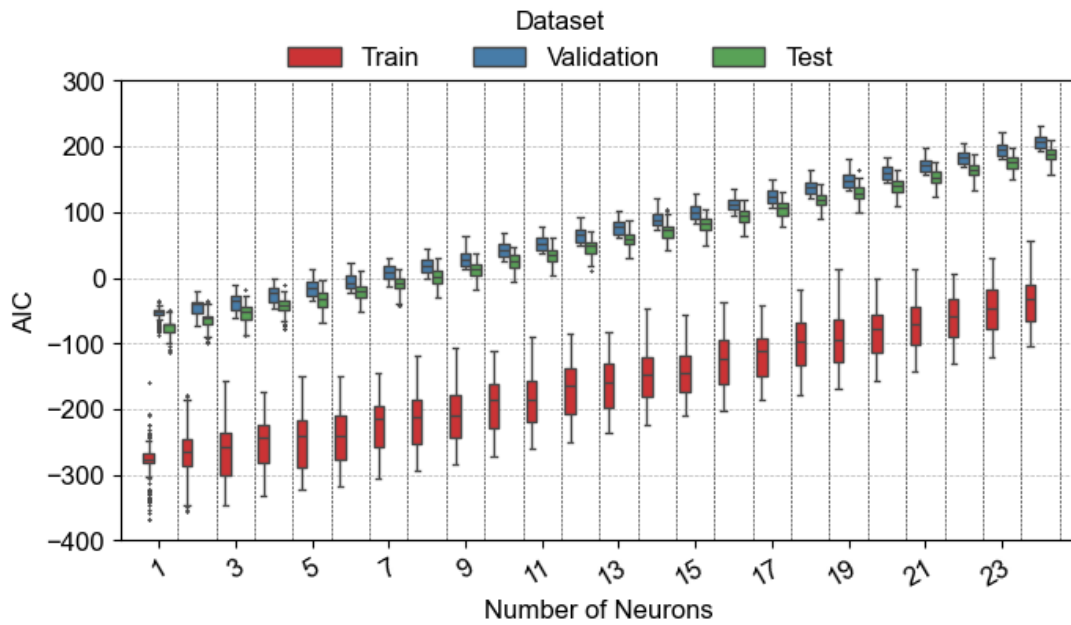


Figure B.134- Combined AIC distributions for well 440-1, DMS scenario.

DSM-FFNN model plateau overall graph for 442-1:

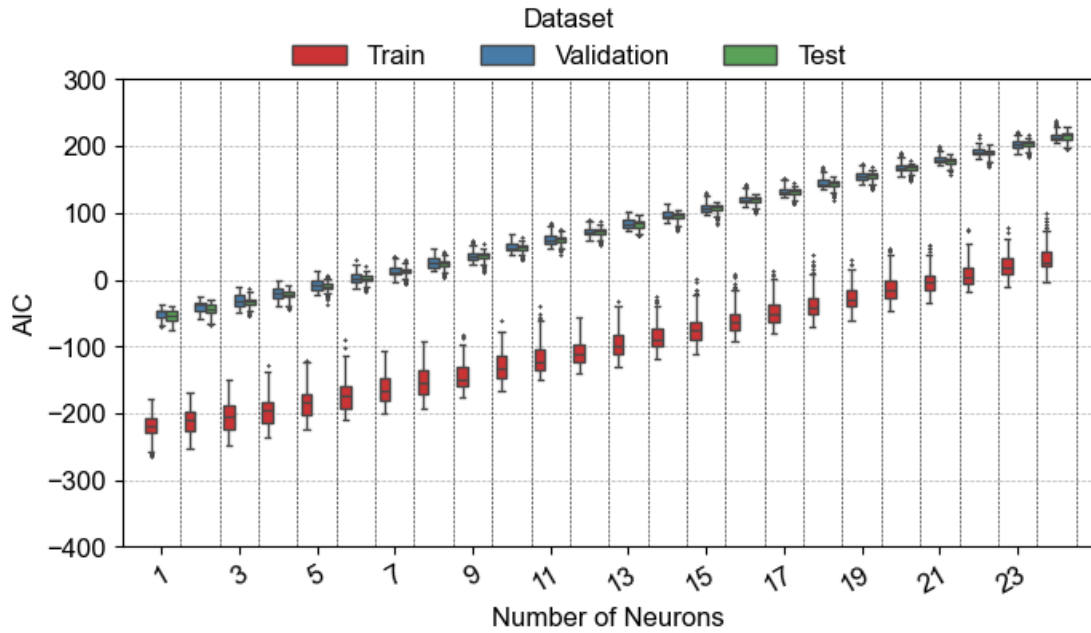


Figure B.135- Combined AIC distributions for well 442-1, DSM scenario.

DMS-FFNN model plateau overall graph for 442-1:

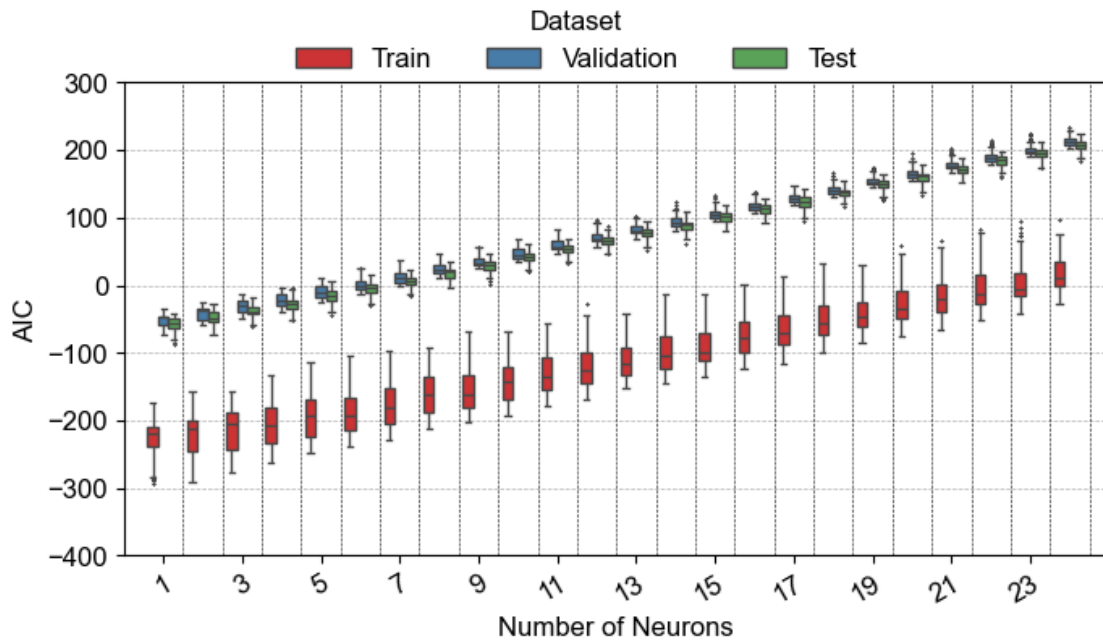


Figure B.136- Combined AIC distributions for well 442-1, DMS scenario.

DSM-FFNN model plateau overall graph for W2:

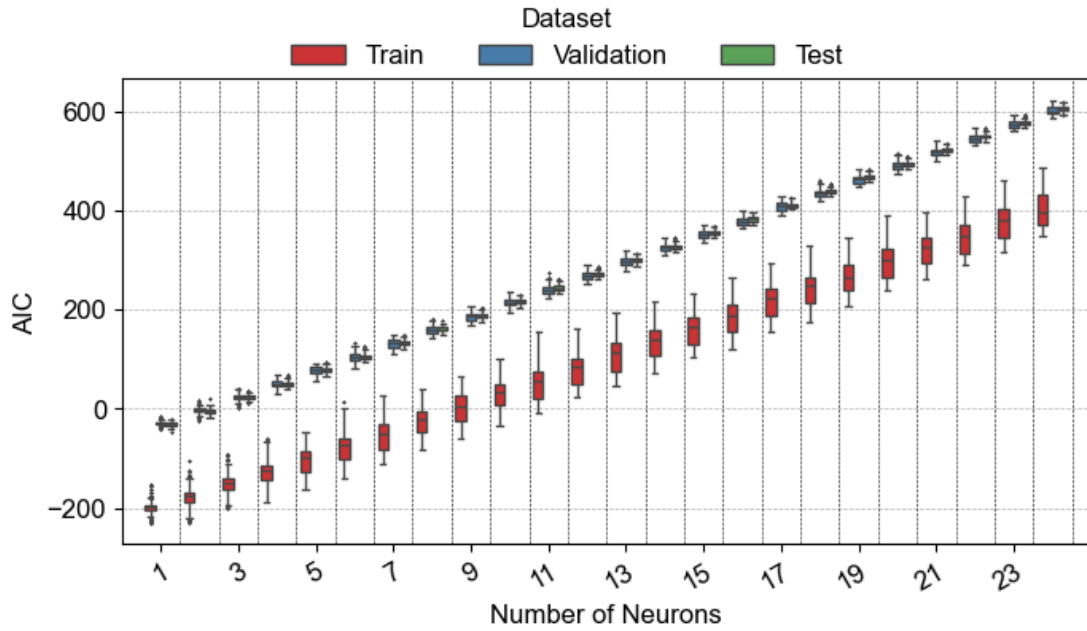


Figure B.137- Combined AIC distributions for W2 combined model, DSM scenario.

DMS-FFNN model plateau overall graph for W2:

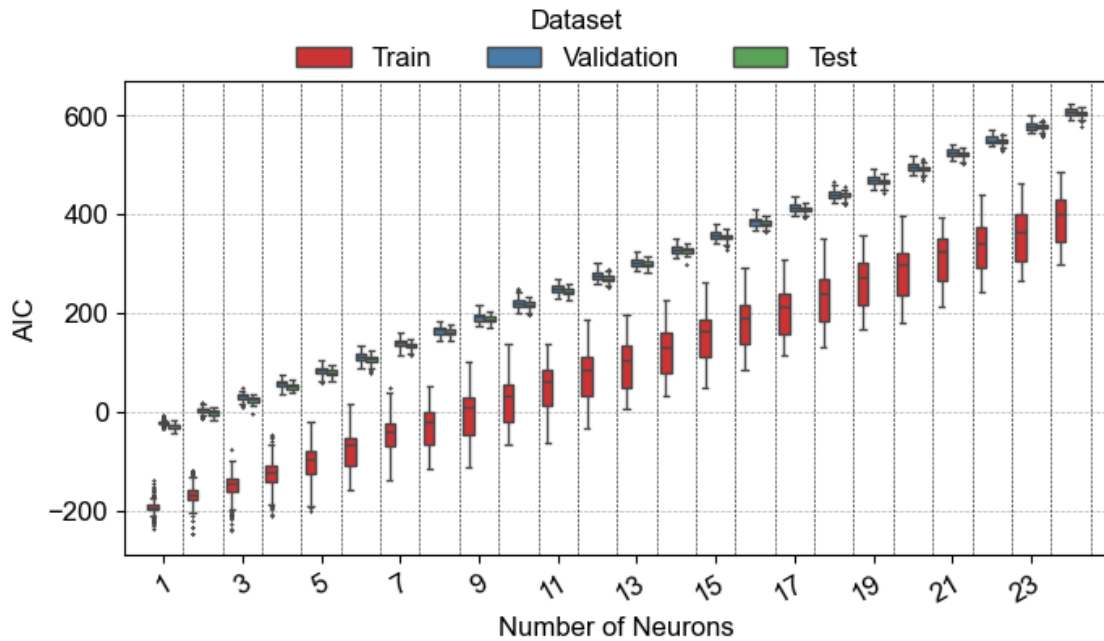


Figure B.138- Combined AIC distributions for W2 combined model, DMS scenario.

DSM-FFNN model plateau overall graph for 244-2:

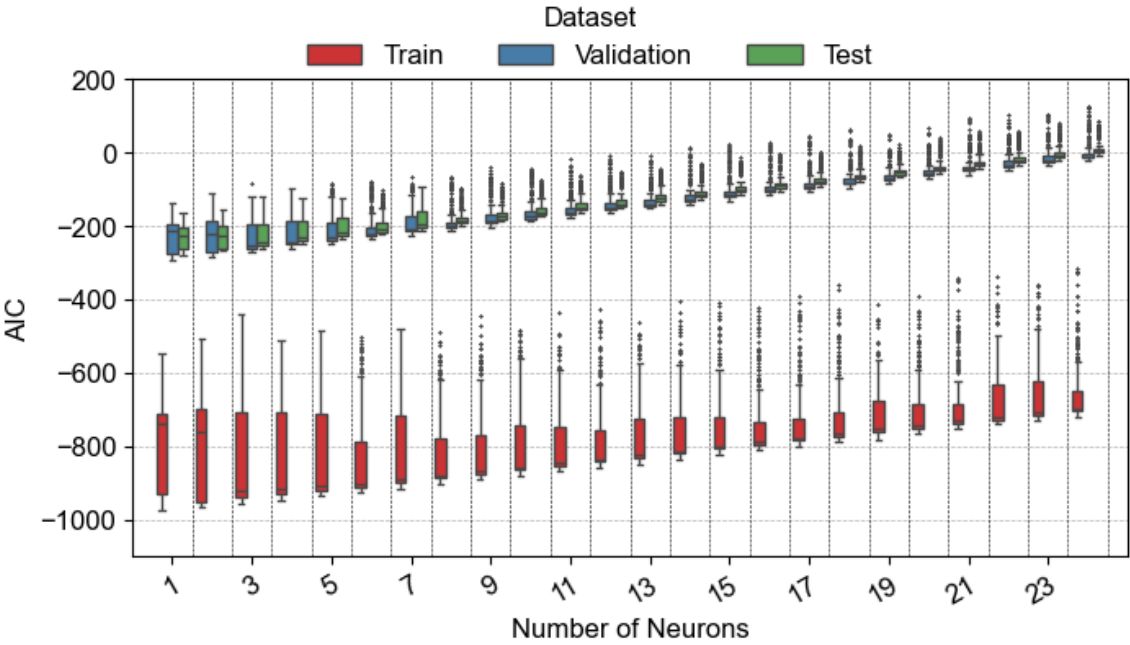


Figure B.139- Combined AIC distributions for well 244-2, DSM scenario.

DMS-FFNN model plateau overall graph for 244-2:

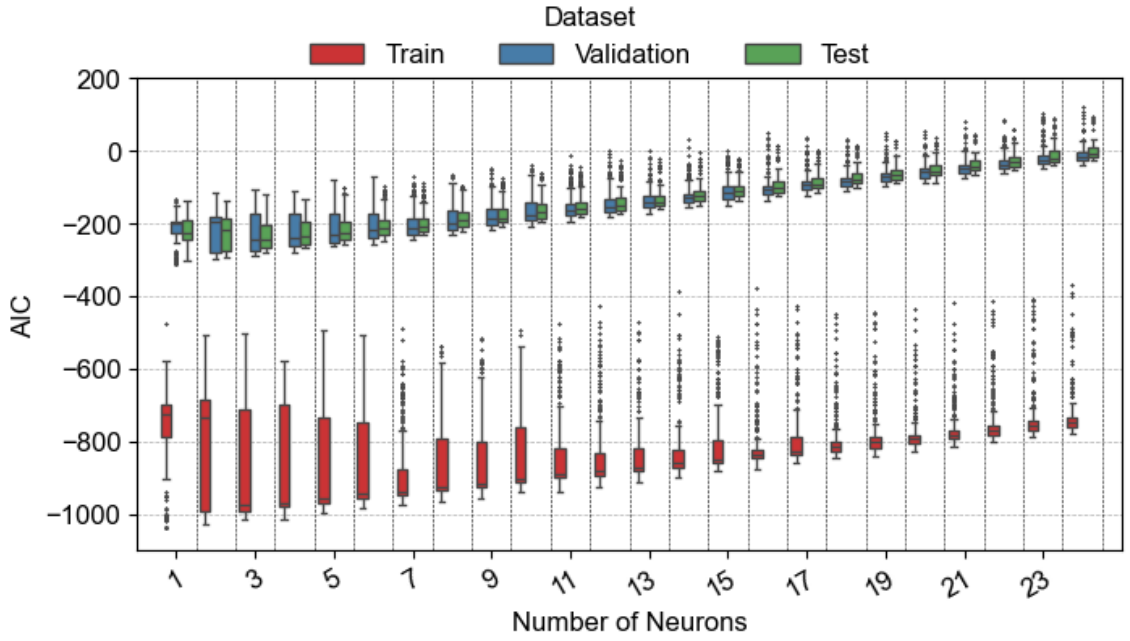


Figure B.140- Combined AIC distributions for well 244-2, DMS scenario.

DSM-FFNN model plateau overall graph for 245-2:

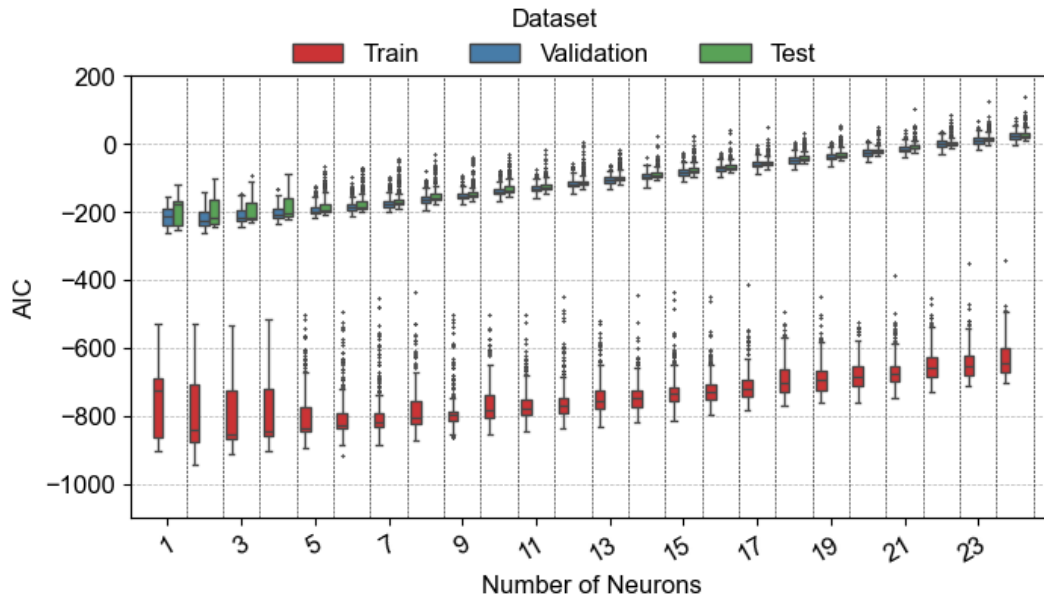


Figure B.141- Combined AIC distributions for well 245-2, DSM scenario.

DMS-FFNN model plateau overall graph for 245-2:

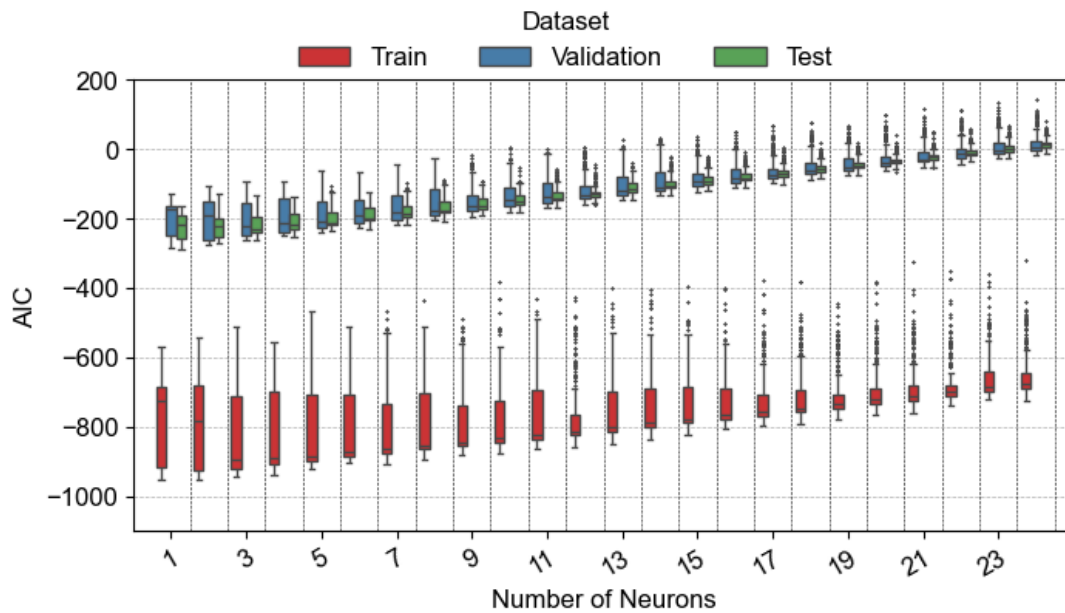


Figure B.142- Combined AIC distributions for well 245-2, DMS scenario.

DSM-FFNN model plateau overall graph for W3:

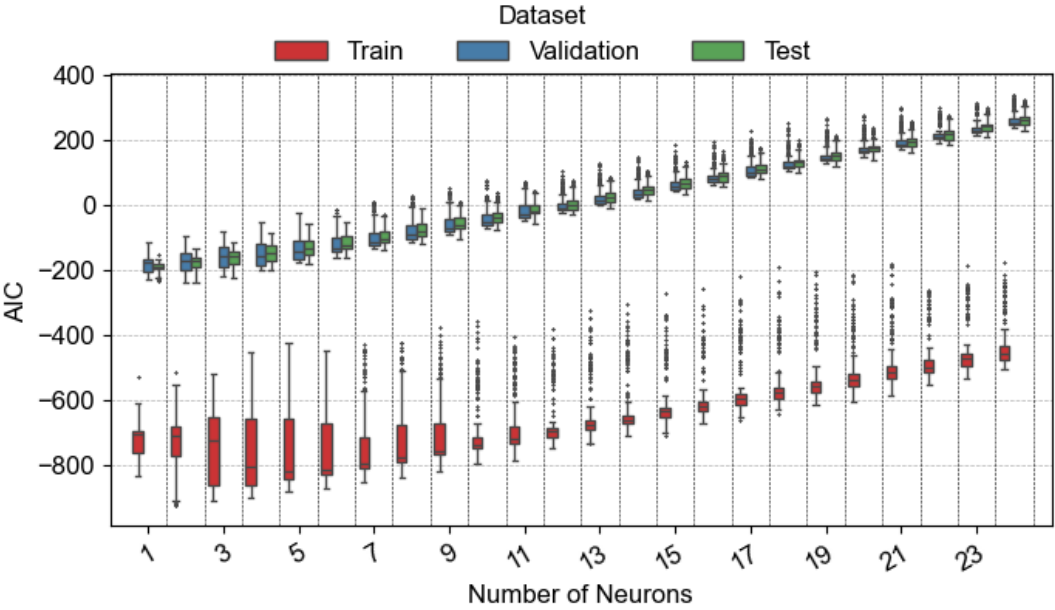


Figure B.143- Combined AIC distributions for W3 combined model, DSM scenario.

DMS-FFNN model plateau overall graph for W3:

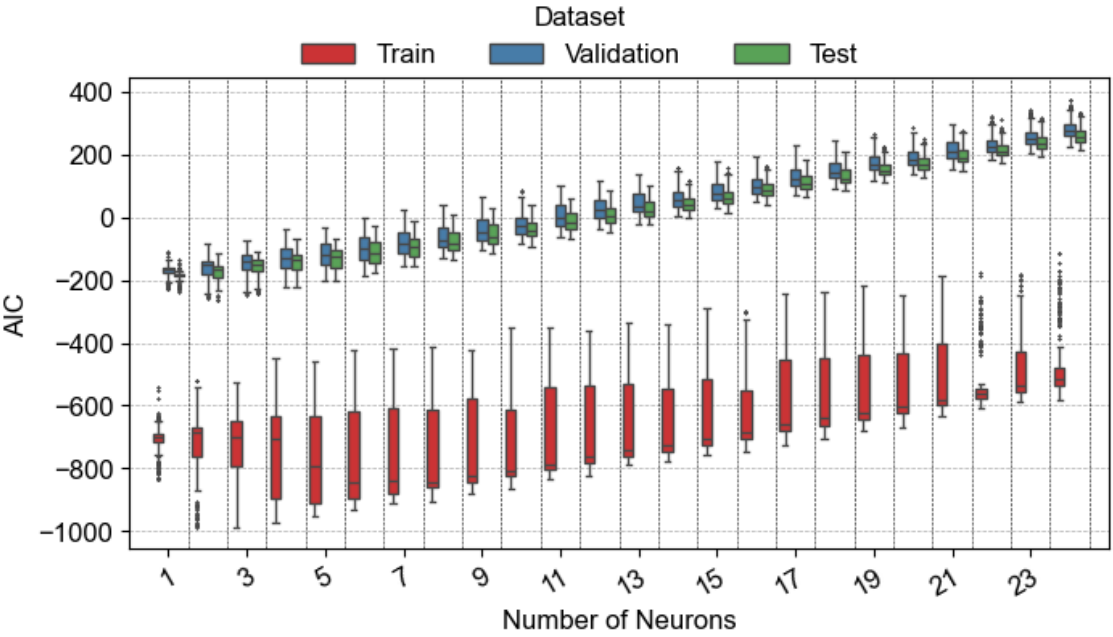


Figure B.144- Combined AIC distributions for W3 combined model, DMS scenario.

DSM-FFNN model plateau overall graph for 443-1:

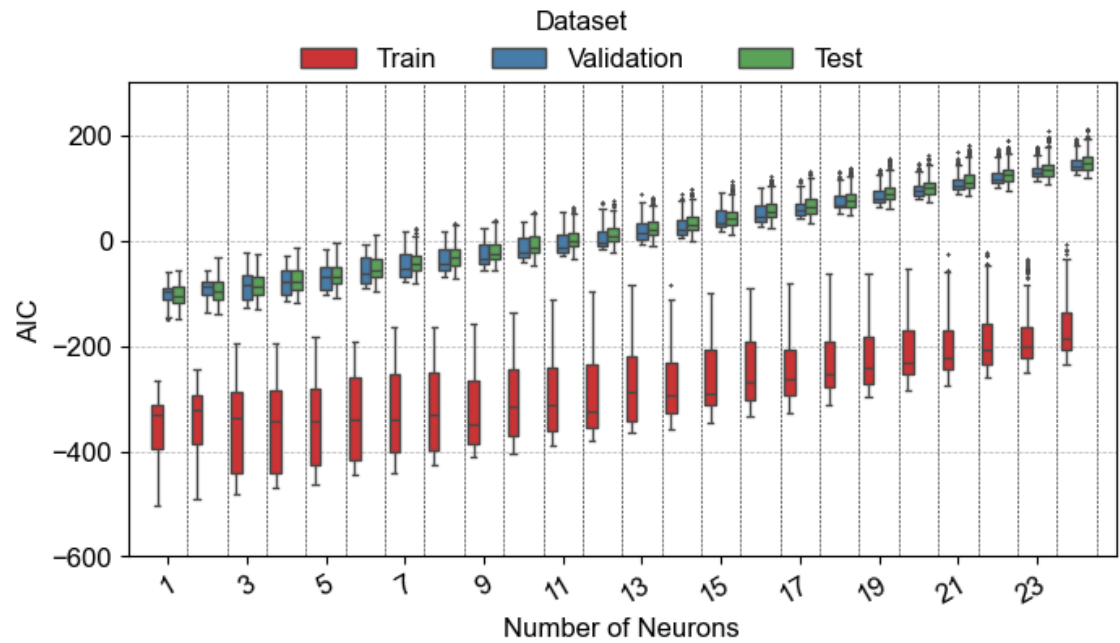


Figure B.145- Combined AIC distributions for well 443-1, DSM scenario.

DMS-FFNN model plateau overall graph for 443-1:

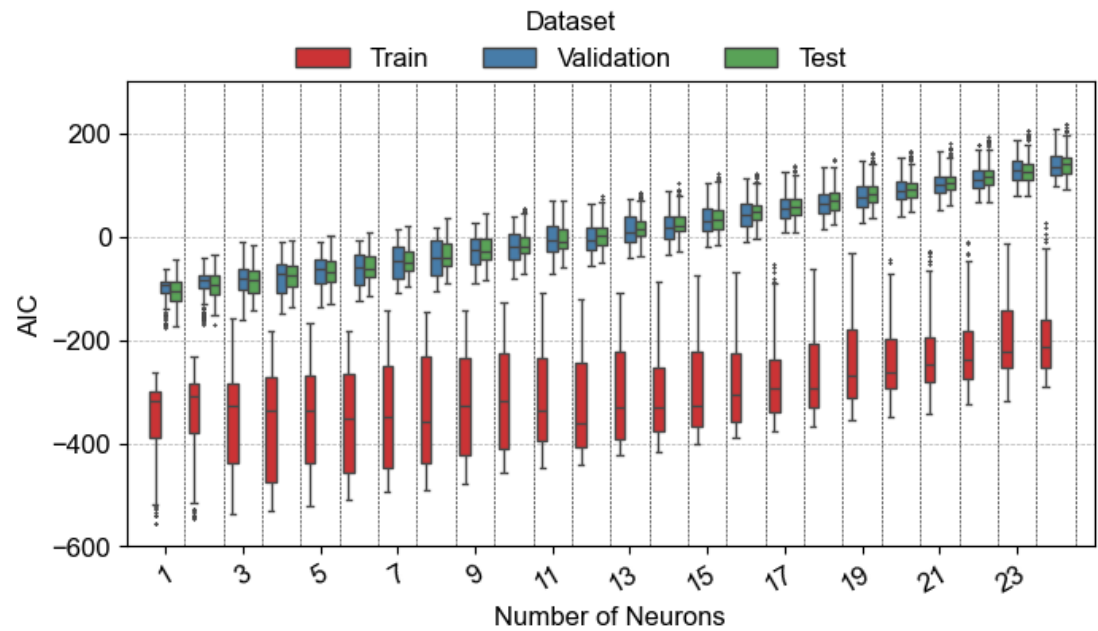


Figure B.146- Combined AIC distributions for well 443-1, DMS scenario.

DSM-FFNN model plateau overall graph for the aggregated model:

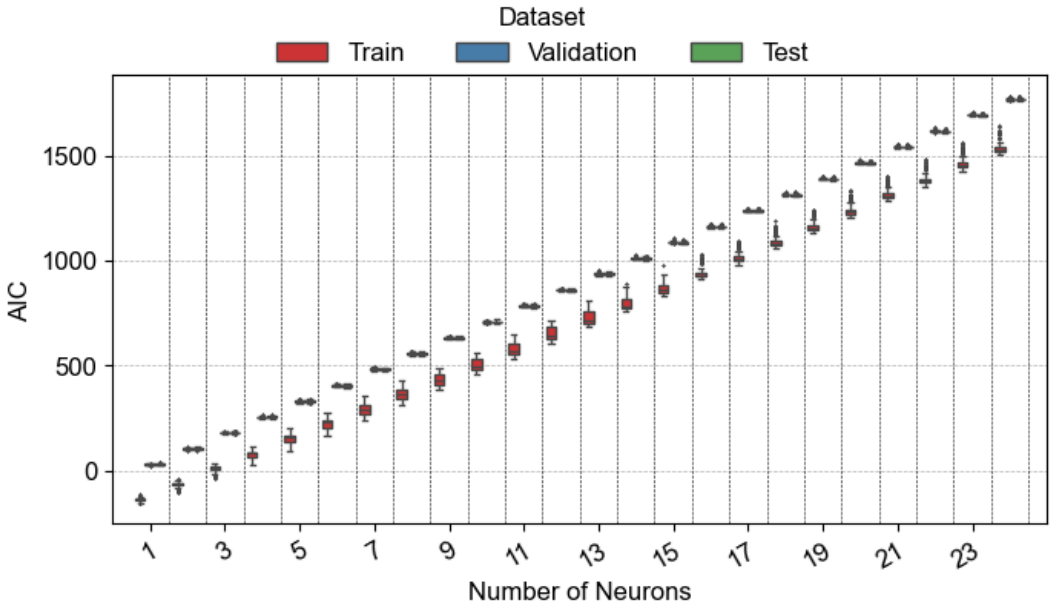


Figure B.147- Combined AIC distributions for the aggregated model, DSM scenario.

DMS-FFNN model plateau overall graph for the aggregated model:

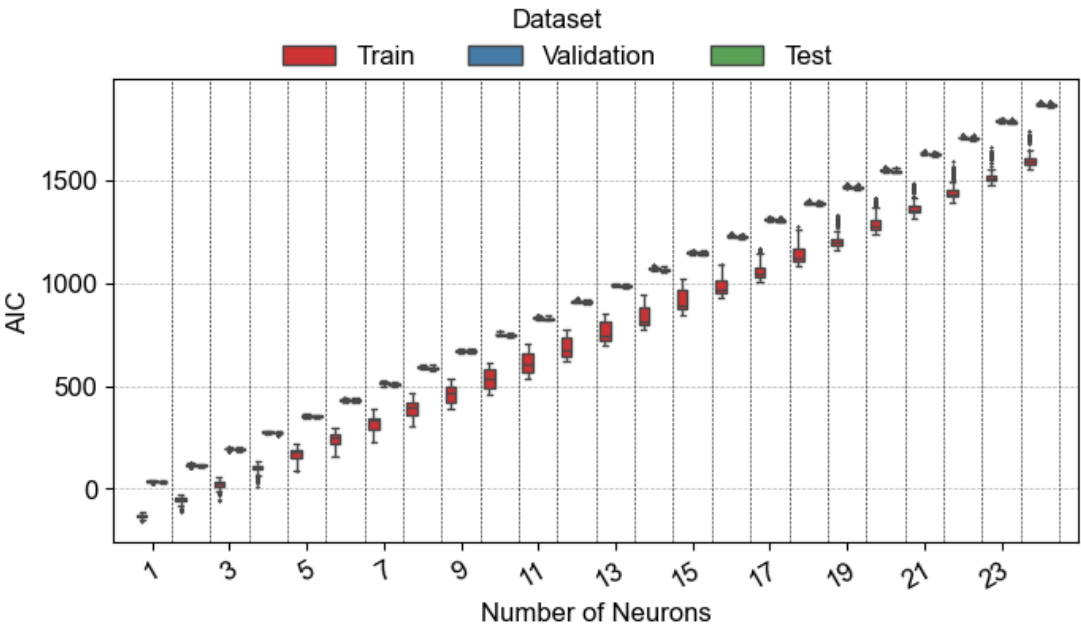


Figure B.148- Combined AIC distributions for the aggregated model, DMS scenario.