

ABSTRACT

KOYAMA, YOKO. Machine Learning Models to Predict Early Breakthrough of Recalcitrant Organic Micropollutants in Granular Activated Carbon Treatment of Water (Under the direction of Drs. Detlef Knappe, and Emily Berglund).

Granular activated carbon (GAC) adsorption is frequently considered to control recalcitrant organic micropollutants (MPs) in both drinking water and wastewater. To predict full-scale GAC adsorber performance, bench- and/or pilot- scale studies are widely used. These studies have generated a wealth of MP breakthrough curves. The overarching aim of this research was to develop machine learning (ML) models from these data to predict MP breakthrough from adsorbent, adsorbate, and background water matrix properties. These models provide a simple and fast tool to predict GAC performance. To develop information for model calibration, MP breakthrough curves were collected from the peer-reviewed literature, research reports, and engineering reports. These data sets, which included results from rapid small-scale column tests (RSSCTs) and pilot/full-scale adsorbers, were analyzed to determine the bed volumes of water that could be treated until MP breakthrough reached ten percent of the influent MP concentration (BV10). The data set encompassed 43 MPs (including neutral and ionizable organic compounds), 3 GAC types by base material (18 unique GAC products), and 38 water matrices, including groundwater, surface water, and treated wastewater. Approximately 400 data sets were split into training, validation, and test sets. Seventeen candidate features, such as MP properties (Abraham parameters), background water matrix characteristics, and GAC properties, were explored in ML models to predict log₁₀-transformed BV10 (logBV10). BV10 values obtained from the resulting predictive model were highly correlated with experimentally determined BV10 values (coefficient of determination ~0.89 for logBV10 prediction), and the most effective model predicted BV10 with an absolute mean error of ~ 0.11 log units. Key drivers influencing BV10

prediction included the MP's partitioning coefficient between air and hexadecane (Abraham parameter L); dissolved organic matter concentration in background water matrix; and the adsorbent's point of zero charge (pzc). The model can be used to estimate GAC bed life and select effective GACs for the removal of MPs such as per- and polyfluoroalkyl substances (PFASs), pesticides, pharmaceuticals, and volatile organic compounds (VOCs) in a wide range of water types.

© Copyright 2021 Yoko Koyama

All Rights Reserved

Machine Learning Models to Predict Early Breakthrough of Recalcitrant Organic
Micropollutants in Granular Activated Carbon Treatment of Water

by
Yoko Koyama

A thesis submitted to the Graduate Faculty of
North Carolina State University
in partial fulfillment of the
requirements for the Degree of
Master of Science

Environmental Engineering

Raleigh, North Carolina

2021

APPROVED BY:

Dr. Detlef R. U. Knappe
Chair of Advisory Committee

Dr. Emily Berglund
Chair of Advisory Committee

Dr. Joel Ducoste

DEDICATION

お母さんへ、いつも応援ありがとう。千晴ちゃんと自分を一人でここまで育ててくれたのは大変だったね。お母さんのおかげで今の私の最高のキャリアスタートがあります。これから親孝行していくから、楽しみにしてください。

感谢妈妈一直来对我的支持。你多年在异国他乡单身养育我与妹妹辛苦了。我非常感谢你为我创造的在海外深造的机会。我从此会努力尽孝，敬请期待。

BIOGRAPHY

I was born to a Chinese immigrant family and raised in Chiba, Japan, until high school. My father finished middle school and my mother attended college part-time but none of their education was recognized when they moved to Japan. My mother was responsible for the most part of me and my sister, Chiharu's education. Luckily, Chiharu and I were both admitted to private universities with scholarships to pursue our higher education. I went to a liberal arts college, Washington and Lee University (VA) in the US and finished with a B.S. in Integrated Engineering in May 2019. I immediately entered graduate school to pursue my M.S. of Environmental Engineering degree at North Carolina State University the same year. I will be working for Hazen and Sawyer as an assistant engineer upon successful completion of the MS degree.

ACKNOWLEDGEMENTS

To Detlef, thank you for your mentorship and support. I am especially grateful for your flexibility in letting me pursue a project that was not in the original scope of work. Your work ethic and creativeness have been inspirational to me.

To my peers in my lab and the environmental engineering cohort, I appreciate all your support, encouragement, and fun.

To my friends outside of graduate school, Oliver and Sam, thank you for your continued moral support through both happy and difficult times.

To my family, thank you for your continued support in my career.

To Dr. Berglund, thank you for serving as the co-chair and your research advise.

To Dr. Ducoste, thank you for serving on my committee and offering your research advise.

To Mohammad, thank you for being an awesome mentor guiding me through machine learning.

To Cecile (Dr. Zhi), thank you for inspiring me with your great research ideas and mentoring me.

To Dr. Nelson, thank you for teaching the class, BAE565, so well and offering your research advice.

To the NCSU library data science help desk, thank you for your incredible suggestions and services.

To P.E.O, who awarded me with the International Peace Scholarship and gave me a wonderful support network – thank you, Judy and Peggy!

TABLE OF CONTENTS

LIST OF TABLES.....	viii
LIST OF FIGURES.....	ix
INTRODUCTION	1
METHODS.....	4
<i>Database development</i>	<i>4</i>
<i>Predictor and response variables</i>	<i>5</i>
<i>Model development</i>	<i>6</i>
<i>GBM model interpretation</i>	<i>8</i>
RESULTS AND DISCUSSION	9
<i>Comparison of model performance based on CV and testing prediction error</i>	<i>9</i>
<i>Variable contributions</i>	<i>12</i>
<i>Prediction drivers</i>	<i>13</i>
<i>Effect of MP properties</i>	<i>14</i>
<i>Effect of water quality.....</i>	<i>16</i>
<i>Effect of carbon characteristics</i>	<i>18</i>
<i>Effect of EBCT</i>	<i>20</i>
<i>Sensitivity analysis with GBM model.....</i>	<i>20</i>
<i>Implications and limitations.....</i>	<i>22</i>
REFERENCES	24
APPENDICES.....	31

<i>Appendix A</i>	32
Figures A1~5.....	33
Tables A1~2.....	37
Text A1 Recalcitrancy of MPs.....	41
Text A2 Abraham solvation parameters.....	41
Text A3 Carbon character information.....	41
Text A4 Analysis of breakthrough data to calculate y	44
Text A5 MLR model development.....	46
Text A6 Ensemble tree models development.....	69
Text A7 Permutation variable importance.....	82
Text A8 Partial dependence plots (PDPs) and centered individual contribution expectations (ICEs).....	83
Text A9 Simulation scenarios for sensitivity analysis.....	86
Text A10 MLR model's prediction accuracy by BV10 range.....	87
Text A11 Variable selection between DOC and UV254.....	87
Text A12 Effect of pore size distribution.....	87
Text A13 Positive correlation between L/V and BV10 through GBM model sensitivity analysis.....	91
Text A14 Final GBM and RF models for web application deployment.....	92
References within Appendix A	93
<i>Appendix B</i>	99
<i>Appendix C</i>	100
<i>Appendix D</i>	101
Breakthrough data of Ref ID 1.....	102
Breakthrough data of Ref ID 15.....	114
Breakthrough data of Ref ID 19.....	119
Breakthrough data of Ref ID 35.....	158
Breakthrough data of Ref ID 40.....	175
Breakthrough data of Ref ID 47.....	182

Breakthrough data of Ref ID 50	190
Breakthrough data of Ref ID 57	224
Breakthrough data of Ref ID 60	233
Breakthrough data of Ref ID 61	248
Breakthrough data of Ref ID 62	288
Breakthrough data of Ref ID 63	306
Breakthrough data of Ref ID 69	325

LIST OF TABLES

Table 1	Summary of CV and Testing Errors for MLR, RF, and GBM models.....	10
---------	-------------------------------------------------------------------	----

LIST OF FIGURES

Figure 1	Data distribution by a) compound classes, b) water matrix, c) GAC types and d) column types.....	5
Figure 2	BV10 parity plots for MLR (a), RF (b), and GBM (c) models	12
Figure 3	Permutation Importance Ranking of Variables for the GBM Model	14
Figure 4	Centered individual contribution expectations (c-ICEs) and partial dependence plots (PDPs) for the GBM model of variables (a) L, (b) DOC, (c) pzc and (d) BET.	16
Figure 5	Predicted BV10 (a) and % change in predicted BV10 (b) for all GBM model simulation scenarios.....	21

INTRODUCTION

A crucial step in full-scale granular activated carbon (GAC) adsorber design is assessing the breakthrough behavior of target MPs through bench-scale and pilot-scale column experiments. To predict full-scale GAC adsorber performance, in the past decades, researchers have attempted to utilize modeling approaches to estimate full-scale GAC adsorber performance more efficiently than relying only on physical experiments.

One approach to estimate GAC performance is to use mechanistic models such as the pore-surface-diffusion-model (PSDM, Sontheimer et al., 1988), to scale up bench test such as rapid small-scale column test (RSSCT) results. Scaling up bench scale experimental results or using data from pilot scale experiments results to estimate full-scale GAC column capacity are common physical experiments dependent approaches but can be costly and time-consuming. To overcome these limitations, Kennedy et al (2015) developed a data-driven model (Eq 1) to predict the bed volumes of water that can be treated until MP breakthrough reached ten percent of the influent MP concentration (BV10). However, their model significantly overestimated BV10 by a factor of 4.2 when applied to a wastewater-impacted drinking water source (Kennedy et al., 2015). Furthermore, when the model was applied to a groundwater source, the model (Eq 1) in general under-predicted the results by ~30% and the magnitude of the error increased as the adsorbability of MP increased (Merle et al., 2020). In the same study (Merle et al., 2020), the group also observed a lack of correlation between the predictor variable $\log D$, the octanol-water partition coefficient, and the response variable $\ln BV$ as described in Eq 1.

$$\begin{aligned} \ln BV_{10\%} = & (11.2 \pm 0.2) + (-0.242 \pm 0.052) DOC_0 + (0.138 \pm 0.041) \log D + \\ & (-0.305 \pm 0.093) S + (0.157 \pm 0.069) V \quad (1) \end{aligned}$$

To date, there is no model to accurately predict MP early breakthrough of full scale GAC adsorbers from commonly available data and without in-situ data inputs. Thus, providing cost estimates and full-scale GAC design remains challenging without physical column experiments. This limitation can prevent utilities from timely facility upgrades to adapt to the thousands of new emerging compounds, such as 1,4-dioxane and per- and polyfluoroalkyl substances (PFASs), that can pose severe public health hazards (Kano et al., 2009; Pelch et al., 2019).

The overarching aim of this research was to develop predictive models utilizing machine learning (ML), trained and validated with the abundant breakthrough data generated by previous column studies. ML can be used to exploit the non-linear, complex relationships between input parameters and target response, yielding highly predictive models from surrogate measures that are easy to obtain. In this study, multiple linear regression, a traditional statistical method coupled with ML training algorithm, and models based on tree-based ensemble learning methods are explored and compared. The models served to achieve two goals: 1) to estimate BV10 with practical accuracy that could aid full-scale GAC design, and 2) to shed light on the most influential factors of MP properties, background water matrix, and GAC properties that impact BV10 prediction.

A database of GAC breakthrough data was first constructed and subsequently, predictive models were developed in R (“R: The R Project for Statistical Computing,” 2020). To develop information for model calibration, MP breakthrough curves were collected from the peer-reviewed literature, research reports, and engineering reports. These data sets, which included results from RSSCTs and pilot and full-scale adsorbers, were analyzed to determine BV10. Three modeling approaches to predict BV10— multiple linear regression, random forest, and gradient boosting machine—were compared. An overview of the workflow from database

curation to model deployment is summarized in Appendix A (Appx. A), Figure A1.

METHODS

Database development

GAC column studies were collected to curate a database that served as the foundation for model development. Column studies ranging from bench- to pilot -and full- scale that were collected from peer-reviewed articles, thesis, dissertations, and technical reports via personal correspondences were collected, and their references can be found in Appendix B (Appx. B).

From these GAC studies, only high-quality breakthrough data obtained for recalcitrant organic MPs in RSSCTs and pilot/full- scale adsorbers were selected for construction of the database. The model BIOWIN2 (Howard et al, 1992) was used to determine whether a MP is recalcitrant based on their fast biodegradation probability values (Appx. A, Text A1; Appendix C). Only column studies that exhibited well-defined breakthrough curves and that included sufficient information about GAC and background water characteristics as well as operating conditions were included. Breakthrough data with high quantification limits of the MPs and noisy breakthrough data were excluded. A database of 427 breakthrough curves from 17 studies (Appx. B) was developed. The database included 43 MPs from 6 compound classes, 18 GAC products prepared from 3 base materials, and 38 water matrices from 4 matrix types (Figure 1).

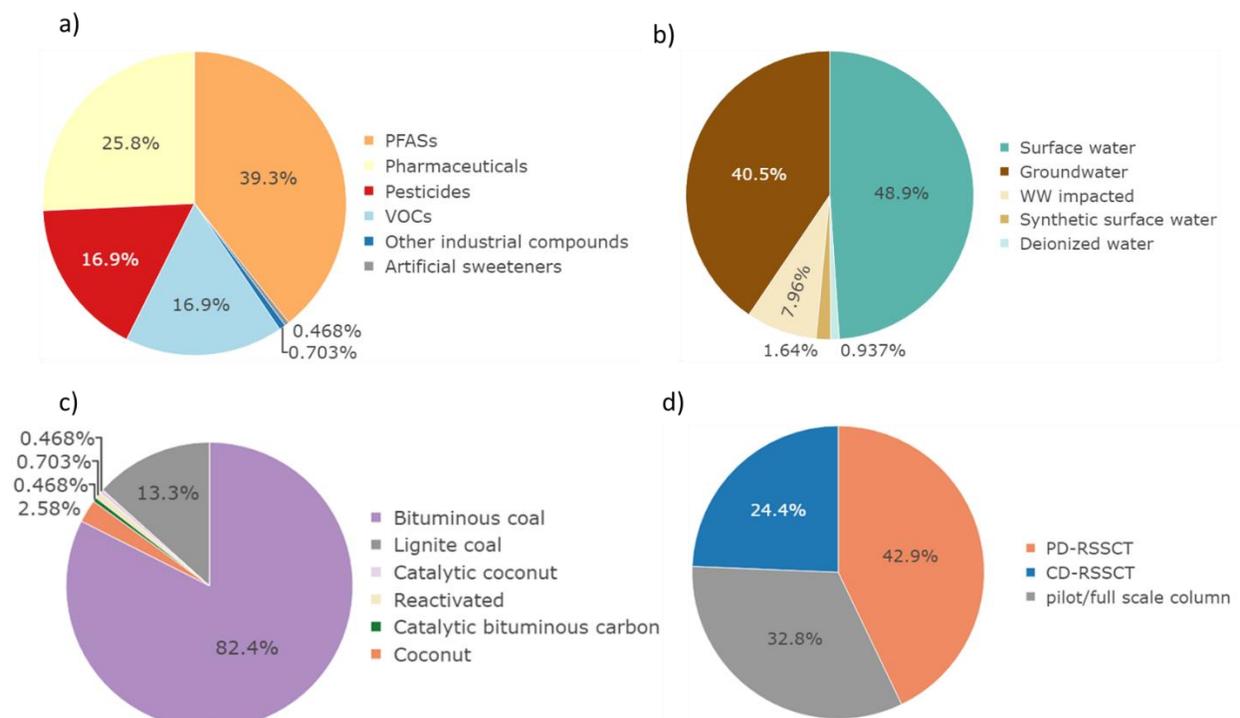


Figure 1. Data distribution by a) compound classes, b) water matrix, c) GAC types and d) column types: Data distribution summary is based on 427 breakthrough curves.

Predictor and response variables

For model development, 17 input variables from four categories— MP properties, water quality, carbon characteristics and operational conditions—were selected as candidate features. Abraham solvation parameters (Appx. A, Text A2), which represent the solute-solvent interactions (Abraham et al., 2004), are commonly employed to model organic MP fate (Brown and Wania, 2009; Zhang et al., 2020) and therefore were included as MP property inputs. Six Abraham solvation parameters, L, B, E, S, A, V, taken from the UFZ-LSER database (Ulrich et al., 2017) were used. Three water quality parameters, pH, DOC, and UV254 absorbance were also included as model inputs. For GAC characteristics, BET, pH_{pzc} and micropore ratio (mp_ratio) were included as inputs. External references containing GAC properties (Appx. A, Text A3) were used when only the GAC product can be identified for a given data set. For

operational conditions, empty bed contact time (EBCT) and 2 dummy variables to represent proportional-diffusivity (PD) and constant-diffusivity (CD) scaled RSSCT data sets were included. The dummy variables were created to account for three different column data types—pilot/full scale column, proportional diffusivity RSSCT (PD-RSSCT) (Crittenden et al., 1987), and constant diffusivity RSSCT (CD-RSSCT) (Crittenden et al., 1986). Additional details for these variables can be found in Appx. A, Table A1.

The response variable y (Appx. A, Table A1), which is the target for ML models to predict, is the \log_{10} transformation of BV10 ($\log_{10}BV10$). BV10 values were acquired by analyzing breakthrough data digitized from figures using WebPlotDigitizer (Rohatgi, 2020). Additional details pertaining to data analysis of differing column scales are described in Appx. A, Text A4 and the results of the analysis are summarized in Appendix D.

Model development

Modeling methods such as multiple linear regression (MLR) and ensemble learning methods were evaluated and compared against one another. All model training and evaluation were performed in R, using packages such as caret, glmnet, ranger, and gbm (Kuhn, 2020; Friedman et al., 2010.; Wright and Ziegler, 2017; Greenwell et al., 2020).

Compared to more complex machine learning algorithms, MLR can produce a model that is both easily interpretable and predictive. Details pertaining to minimizing the effect by collinearity within predictors during feature selection and model training for MLR can be found in Appx. A, Text A5.

Random forest (RF) by Breiman (2001) and gradient boosting machine (GBM) by Friedman (2001) are two popular tree-based ensemble learning methods commonly used in the environmental sciences/engineering disciplines. RF and GBM are expected to yield highly

predictive models while selecting the most predictive variables even under the presence of multicollinearity as these algorithms do not rely on assumptions about data distribution or variable independency (Chang & Chen, 2005). While both methods utilize an aggregate of decision trees, RF involves averaging the regression results of independent trees while GBM involves performing regression with additive, sequential trees.

Prior to applying MLR or ensemble learning methods, the database was divided into training and test sets at a split ratio of 9:1 at random to ensure that data from the same study don't cluster in either the training or test set. In this process, a diversity of data within the test set was ensured to cover different types of compound classes, water matrices, and GAC products, for model development. The training set (n=385) was used to train the models while the testing set (n=42) was set aside to evaluate the trained models.

Training process for MLR and ensemble learning methods differ from each other while they both require hyperparameter tuning through grid-search. For MLR, hyperparameter tuning was done first to perform variable selection with LASSO (Tibshirani, 1996; Text A5). Using the best variable subset found, the final model was built by fitting MLR onto the training set with ordinary least squares (details described in Appx. A Text A5). For ensemble learning methods (GBM and RF), hyperparameter tuning was performed with all candidate variables in the models. Subsequently, we calculated the permutation variable importance for both the tuned GBM and RF models and verified that none of the variables had zero-importance (Appx. A Text A5, Text A6) thus justified the use of all variables.

To compare across different modeling methods, after the models were developed, 10-fold cross-validation (CV) error was calculated for each of the MLR, RF, and GBM models using the training set. Root mean squared error (RMSE) and mean absolute error (MAE) were calculated

for CV error evaluation on logBV10 predictions. In addition to RMSE and MAE, for testing error evaluation, the coefficient of determination (R^2) was also calculated.

GBM model interpretation

The permutation importance for all variables were calculated for the GBM model. To compute permutation importance of a variable, we calculated the percent error increase when the variable is permuted from the original training set (details are included in Appx. A, Text A7).

For the GBM model, variables, L, DOC, pzc and BET's, individual conditional expectations (ICEs) (Goldstein et al, 2015) and partial dependence were calculated and plotted. To compute ICEs of a variable, for each set of the training data, we varied the selected variable and computed the corresponding predicted values of logBV10 (details can be found in Appx .A, Text A8). We then plotted the centered ICEs (c-ICEs) of a few selected variables and the centering procedure can be found in Appx. A, Text A8. For each variable, the average of the c-ICEs was then taken and plotted as the partial dependence plot (PDP) in the centered-scale.

RESULTS AND DISCUSSION

Comparison of model performance based on CV and testing prediction error

In this study, MLR, RF, and GBM models were constructed to predict BV10, the bed volume at which the target MP reaches 10% breakthrough. We compared their prediction performance through cross validation (CV) and test-set prediction in Table 1. In this evaluation process, we set the MLR model's performance as baseline and calculated how the ensemble tree models improved prediction accuracy by computing the percent decrease in CV and testing errors of these models.

Overall, we observed that the ensemble tree models exhibited lower prediction errors compared to the MLR model. For CV performance, the GBM model had the lowest prediction errors (RMSE=0.15 and MAE=0.11, Table 1) followed by the RF model and the MLR model, which had the largest error. For testing performance, the RF model had the lowest errors (RMSE=0.14 and MAE=0.10, Table 1) followed by the GBM model and the MLR model, which again had the largest error. To put the log-unit based error into perspective, for a BV10 value of 10,000, an RMSE value of 0.15 corresponds to a range of 7,100 to 14,100 bed volumes.

Table 1. Summary of 10-fold cross validation (CV) and testing errors for MLR, RF, and GBM models: 10-fold CV error metrics (RMSE and MAE) calculated based on training data and testing error (RMSE, MAE, and R^2) metrics calculated based on testing data. MLR model was used as baseline for comparison across different models since its prediction was the least accurate; error decrease in % is shown for the other two models. Q1 refers to 1st quartile and Q3 refers to 3rd quartile of the 10-fold CV errors. Observed BV10 range from ~800 to 100,000 bed volumes, corresponding to 2.9 to 5 log units.

10-fold Cross Validation (CV) errors						
Model name	RMSE	%decrease	MAE	%decrease	RMSE Q1/Q3	MAE Q1/Q3
MLR model	0.29	baseline	0.23	baseline	0.27/0.31	0.21/0.24
RF model	0.19	35%	0.13	41%	0.17/0.20	0.13/0.14
GBM model	0.15	48%	0.11	52%	0.14/0.16	0.10/0.12
Testing errors						
MLR model	0.27	baseline	0.22	baseline	Within CV range	Within CV range
RF model	0.14	47%	0.10	55%	Below Q1 of CV range	Below Q1 of CV range
GBM model	0.15	45%	0.12	47%	Within CV range	Within CV range

CV error was used to compare prediction accuracy across different algorithms as it is more robust than the testing error. Both ensemble tree models outperformed the MRL model based on the percent decrease in their CV RMSE, which were 35 and 48% respectively for the

RF and GBM models (Table 1). The GBM model had the lowest prediction error in CV—its RMSE and MAE values and both error metrics' corresponding Q1/Q3 range values were the lowest amongst all models (Table 1).

We compared testing RMSE and MAE values to the CV Q1/Q3 error ranges to determine if hyperparameter tuning was performed appropriately to avoid overfitting (Table 1). Overall, the testing errors were similar to the CV errors, meaning the models did not overfit to the training data. Except for the RF model's testing error, all others fell within the range of the CV error. RF model's testing RMSE and MAE fell below their respective Q1 values, indicating an overperformance possibly due to chance associated with the selection of the test data.

For each model, we evaluated its prediction performance across different compound classes, background water matrix types, and a wide range of BV10 (Figure 2a~c) on the test set data. The MLR model (Figure 3a) more effectively predicted treatment scenarios with BV10 values $>20,000$ than those with BV10 values $\leq 20,000$ (for error metrics calculated for both ranges of BV10 prediction by the MLR mode, refer to Appx. A, Text A10). Compared to the MLR model, the ensemble tree models (GBM, RF) effectively predicted GAC performance across the entire range of BV10 (Figure 2a~c). Additionally, the ensemble tree models performed well across different types of water (Figure 2b~c), including wastewater impacted surface water, overcoming a previous challenge faced by the regression model of Kennedy et al (2015), where their model overestimated GAC performance by a factor of 4.2 for wastewater impacted source water. The prediction performance on the same test set across different types of

GAC products (Appx. A, Figure A3) also indicated the ensemble tree models' applicability to various GAC types.

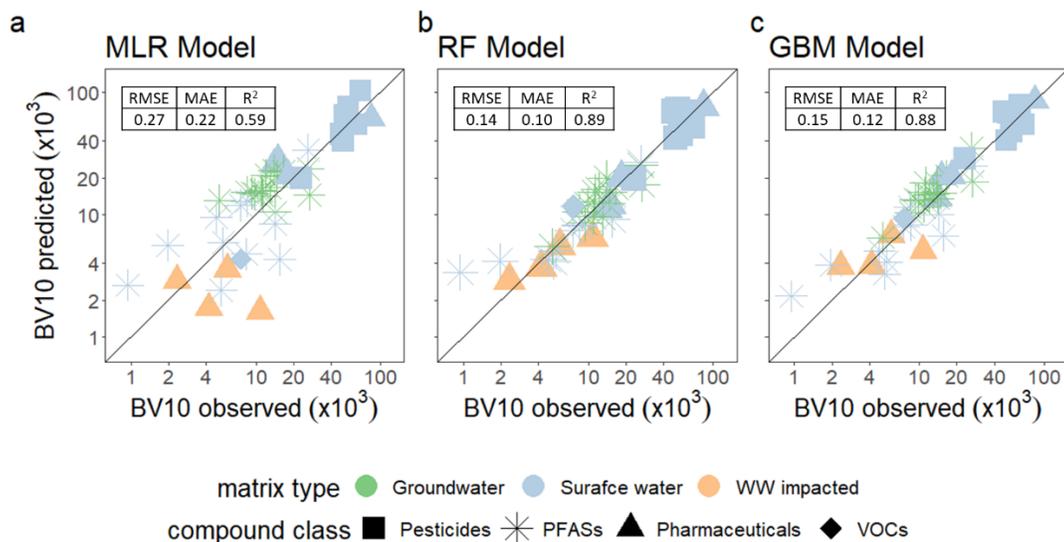


Figure 2. BV10 parity plots for MLR (a), RF (b), and GBM (c) models. Parity plots of BV10 values based on the predictions and observations of test data set (n=42) for (a) MLR, (b) RF, and (c) GBM models: The error metrics shown were calculated in the scale of \log_{10} BV10. Treated WW = treated wastewater.

To summarize, the performances of the two ensemble tree models were similar for the test data set (Figure 2 b~c) and they did better in both CV and testing, than the MLR model (Table 1). Even under the presence of multi-collinearity amongst predictors (Appx. A, Text A5), tree-based ensemble learning methods yielded highly predictive models. The GBM had a smaller CV error than the RF model, and the GBM model is hence explored further in following sections.

Variable contributions

The GBM model was interpreted to assess which variables are influential BV10 predictors and how certain parameters affect BV10 prediction. Two approaches were taken for model interpretation—permutation importance was calculated to rank variables (Figure 3); PDPs

and c-ICEs (a decomposed form of PDPs) were plotted (Figure 4a~c) to evaluate partial dependency and interaction effects of selected variables.

Prediction drivers

We identified BV10 prediction drivers from each category of the parameters—MP properties, water quality and carbon characteristics— respectively, based on their permutation importance of the GBM model (Figure 3). L, an MP property parameter that represents van der Waals interactions (Abraham et al., 2004; Brown and Wania, 2009), was ranked as a top driver (Figure 3).

Closely following L in the ranking is DOC (Figure 3), a water quality parameter, which was the top driver among water quality parameters. While the GBM model ranked L and DOC as two top predictors, placing slightly more weight on L, it is unclear which of them affects the adsorption process more in terms of mechanism. Interestingly, while UV254 was somewhat important in the ranking (Figure 3), the GBM model placed far more weight on DOC (Figure 3). The selection of DOC was surprising because UV254 alone explained slightly more variance than DOC did in terms of their respective semi-partial R^2 (Table A2), and UV254 was selected over DOC in the final MLR model. An extended discussion on variable selection between DOC and UV254 is presented in Appx. A, Text A11.

A GAC characteristic parameter, pzc, was ranked third (Figure 3), which makes it the most important among the carbon characteristic parameters. CD, an operational parameter indicating whether the data is scaled up from a CD-RSSCT, was ranked fourth (Figure 3). This observation indicates that the CD-RSSCT data scaled up with the equation proposed by Kern's et al (2020; Eq S2) do not describe the pilot and full-scale results as well as the scaled-up PD-

RSSCT data. This parameter is only important for incorporating CD-RSSCT data into the modeling process and does not impact end users of the models.

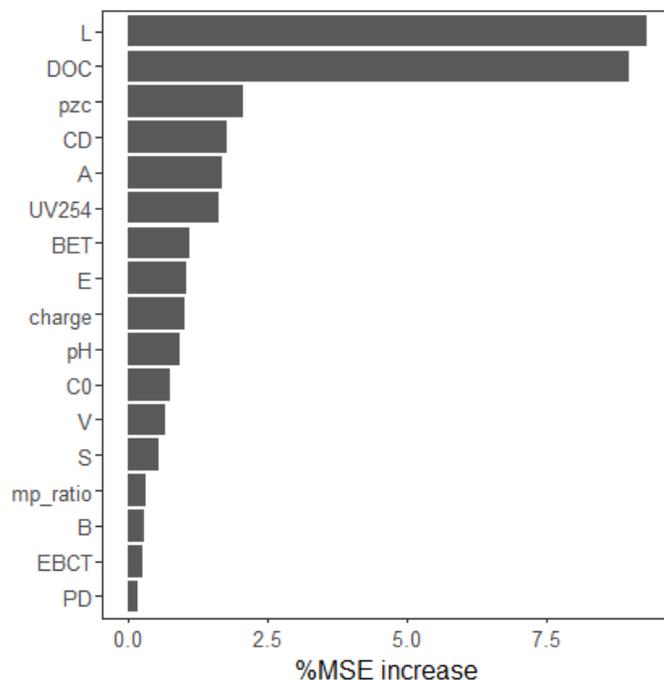


Figure 3. Permutation importance ranking of variables for the GBM model: Percent mean-squared-error (MSE) increase indicating the permutation importance score of each variable is shown. The larger the value is, the more important the variable in the model.

Effect of MP properties

To understand the physical meaning behind the importance of Abraham parameters, we referred to a process-based model (Brown and Wania, 2009) involving Abraham parameters that simulated MP environmental fate and transport. Brown and Wania (2009) found that L, which represents van der Waals interaction in the solute-adsorbent interaction (Abraham et al., 2004; Götz et al., 2007), was the most important Abraham descriptor in environmental fate modeling of organic MPs across multiple media in their study. The agreement between this process-based model and our GBM model suggest that L is a strong predictor based on mechanism. Since L is

the most important Abraham parameter, it is reasonable to state that van der Waals interactions are the most important intermolecular interaction in GAC adsorption processes.

Because L was selected as a top prediction driver, we evaluated its marginal effect on logBV10 prediction (Fig 4a). While normally compounds tend to have larger V values with increasing L values, for the purpose of the marginal effect analysis, V and all other variables were held constant while L was increased. We found that the positive effect by L in the data-dense region of L, $1 < L < 6$, is rather pronounced, which can drive up BV10 prediction by up to 0.8 log units (Figure 4a), which translates to a factor of ~6 on a linear scale. However, a heterogeneity of L's partial dependency was observed and is evident based on the vertical spread of its c-ICEs (Figure 4a). When $BET < 800$ and $DOC > 4$, logBV10 does not change much as L changes (Figure 4a). This can be explained by pore blockage effects as DOM adsorbs on GAC with low BET surface area, adversely affecting MP removal such that the effect by increasing L is masked.

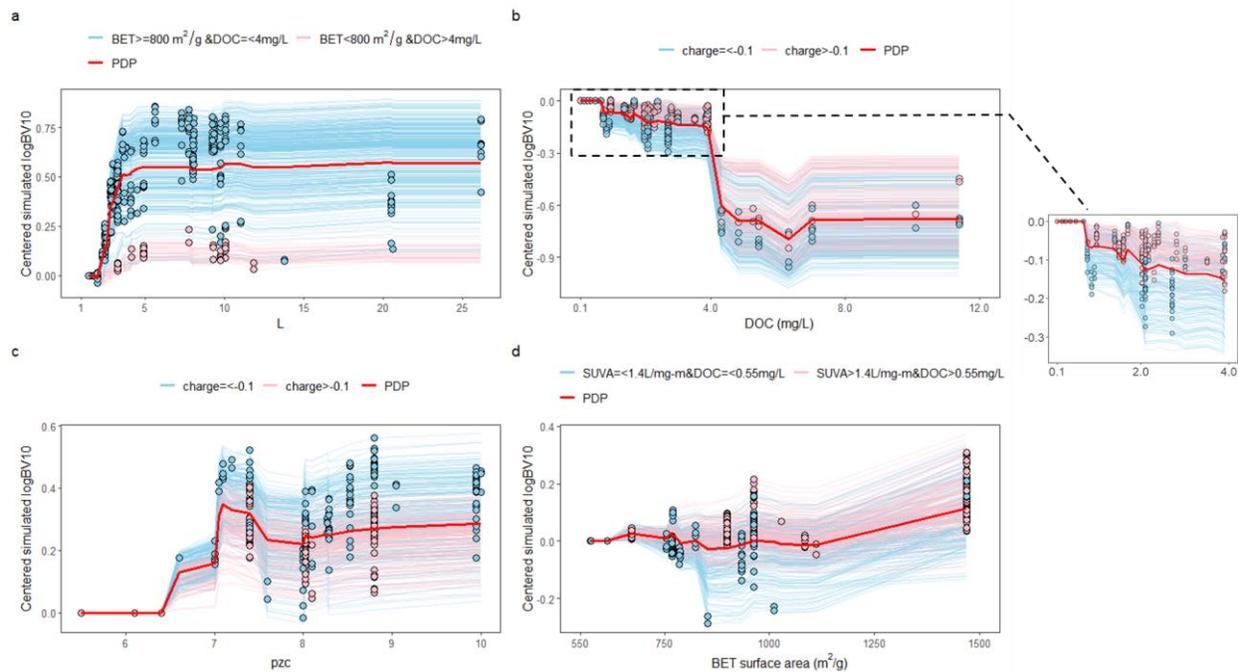


Figure 4 Centered individual contribution expectations (c-ICEs) and partial dependence plots (PDPs) for the GBM model of variables (a) L, (b) DOC, (c) pzc and (d) BET: In a, blue lines/dots represent the c-ICEs/predictions for data entries with $BET \geq 800 \text{ m}^2/\text{g}$ and $DOC \leq 4 \text{ mg/L}$ and pink lines/dots represent those of the complimentary set. In b, blue lines/dots represent the c-ICEs/predictions for data entries with $charge \leq -0.1$ and pink line/dot represent those of the complimentary set. Additionally, in b, an insert to highlight the DOC effect in the range of $0.1 < DOC < 4.0 \text{ mg/L}$ is included. In c, blue lines/dots represent the c-ICEs/predictions for data entries with $charge \leq -0.1$ and pink line/dot represent those of the complimentary set. In d, blue lines/dots represent c-ICEs/predictions for data entries with $SUVA \leq 1.4 \text{ L/mg-m}$ and $DOC \leq 0.55 \text{ mg/L}$; pink lines/dots represent the those of the complimentary set. Across a~d, red lines represent PDPs.

Variable importance ranking (Figure 3) suggested that C0 is of lower importance, which confirmed observations regarding MP removal made in previous studies (Matsui et al., 2002; Corwin and Summers, 2012; Zietzschmann et al., 2016)—initial MP concentration does not impact early breakthrough on a normalized basis.

Effect of water quality

We evaluated the marginal effect by DOC (Figure 4b), the top driver in water quality parameters, on BV10 prediction. The negative marginal effect is most pronounced in the 0.1~4 mg/L range, where DOC data are dense and BV10 prediction decreased up to 0.3 log units (-50%

on a linear scale) in this region (Figure 5b). Interestingly, the DOC effect in the data-dense region (DOC=0.1~4mg/L) drives the BV10 prediction down (-0.3 log units) less than does L drive up the BV10 prediction (+0.8 log units) in the data dense region of L (L=1~6). The weaker effect of DOC alone can be explained in part by the collinearity between DOC and UV254 (Pearson's r value ≥ 0.9 ; Table A2) – i.e., UV254 values increase with increasing DOC values. However, to assess the marginal effect of DOC, we held UV254 constant. We evaluated the effect of DOM more comprehensively in the section entitled *Sensitivity analysis with GBM model*.

The partial dependence of BV10 on DOC is stratified vertically by MP charge (Figure 4b) —MPs with charge ≤ -0.1 decreased more in logBV10 as DOC increased, than did MPs with charge > -0.1 (Figure 5b). This result can be explained by repulsive electrostatic interactions between negatively charged DOM and negatively charged MPs, which adds to the fouling effect and thus further inhibits the adsorption of negatively charged MPs. Subsequently, we conducted the same analysis for UV254, and a similar negative effect to that observed for DOC was observed from the c-ICEs and PDP for UV254 (Appx. A, Figure A4). A possible strength of the GBM model is that it separately considers DOC and UV254 and thus may be able to predict the effect of DOM on BV10 for waters with a range of DOM characteristics and concentrations. Overall, our results agree with numerous previous studies (Crittenden et al., 1991; Matsui et al., 2002; Quinlivan et al., 2005; Shimabuku et al., 2017; Sgroi et al., 2018) that highlighted the adverse effect of dissolved organic matter (DOM) concentration on the extent of MP adsorption by GAC.

Effect of carbon characteristics

The pzc, i.e. the pH corresponding to the point of zero charge, emerged as the most important GAC characteristic parameter in the variable importance ranking (Figure 3), and we evaluated its marginal effect on logBV10 (Figure 4c). Interestingly, ICEs for pzc were spread out vertically over a wide range – in some cases, an increase in pzc (especially for pzc >6) led to larger BV10 values (~0.5 log units increase) but in others, BV10 values did not change much (Figure 4c). In Figure 4c, MPs with charge ≤ -0.1 showed a larger increase in logBV10 as the pzc of GAC increased in the range of $6 < \text{pzc} < 7.2$. This is reasonable because increasing pzc leads to increased positive (or less negative) charge on the GAC surface, which in turn promotes adsorption of negatively charged MPs through electrostatic attraction.

BET surface area was ranked second within the GAC characteristics category (Figure 3); hence we explored its marginal effects on logBV10 prediction (Figure 4d). Overall, higher BET surface area did not necessarily lead to larger BV10 prediction. PDP showed a small effect of increasing BET on predicted logBV10, but individual c-ICEs showed that the marginal effect of BET surface area is in some instances positive, and in others, negative (Figure 4d). The c-ICEs associated with $\text{SUVA} \leq 1.4 \text{ L/mg-m}$ and $\text{DOC} \leq 0.55 \text{ mg/L}$ showed a more pronounced decrease in logBV10 with increasing BET surface area than those with higher SUVA and DOC values (Figure 4d). Entries associated with lower SUVA and DOC values, in other words, DOM with more hydrophilic nature (Karanfil et al., 2002) at a relatively low level of concentration would be susceptible to negative effect by increasing BET on BV10 prediction. This provides explanation as to why in single-solute isotherm experiments (Wu et al., 2005; Kearns et al., 2014) positive correlation between BET and MP removal was observed while in isotherm and column experiments with co-adsorbing DOM (Quinlivan et al., 2005; Benstoem and Pinnekamp,

2017), such correlation was absent. The presence of DOM adds complexity to the adsorption system and thus result in interactions between different parameters of the system, one of which is the combined interaction effect by SUVA and DOC (Figure 4d).

Additionally, we observed pronounced negative impact by increasing BET for data obtained with two GAC product with BET surface areas of 852 and 1011 m²/g, where predicted BV10 decreased over 0.2 log units relative to the c-ICE (Figure 4d). These 4 entries correspond to PFAS removal by two coconut-based GAC products. Other data entries corresponding to PFASs removal by coconut-based GAC products did not cluster with these 4 entries within the c-ICEs.

Although the micropore ratio (mp_ratio) was ranked low for variable importance (Figure 3), we evaluated its partial dependence (details in Text A12), where similar patterns to the marginal effect of BET surface area were found for mp_ratio. Data with lower SUVA value (<1.5 L/mg-m; corresponding DOC values ranged from ~0.3 to 2.1 mg/L) exhibited decreasing logBV10 as mp_ratio increased (Figure A5), thus indicating that more hydrophilic DOM with lower SUVA values (Karanfil et al., 2002) is associated with negative impact on BV10 prediction when mp_ratio is increased. Furthermore, entries corresponding to PFAS removal by coconut-based GAC products suffered the largest decrease in logBV10 as mp_ratio increased.

A possible explanation for the observed similar negative effect of increasing mp_ratio and BET values for waters containing DOM with low SUVA values is blockage of micropores caused by hydrophilic DOM. Earlier studies have shown that the presence of mesopore is crucial to alleviating the pore blockage effect by DOM (Pelekani & Snoeyink, 2000, 2001; Li et al., 2003). The marginal effect of BET surface area, which is

correlated to micropore volume as well as micropore volume ratio (Table A12-2), can thus also be explained by pore blockage of the micropores. Together, the similarity in marginal effects of BET surface area and mp_ratio suggests that the presence and nature of DOM, in conjunction with physical GAC characteristics, affect MP adsorption by GAC in a complex manner that was effectively captured by the GBM model.

Effect of EBCT

We evaluated the importance of EBCT (range: 4.6 to 24 minutes), an important parameter for the design of GAC adsorbers. The effect of EBCT on BV10 prediction was found to be less important than effects of other variables (Figure 3), and the effect of EBCT was found to be ambiguous (Appx. A, Figure A6). Statistics based on semi-partial R^2 (Table A2) also suggest that EBCT only describes a small portion of the variance of the entire data set. These observations are consistent with previous studies that found EBCT has only small effects on GAC use rates for MP removal when EBCT was varied between 10 and 20 minutes (Corwin and Summers, 2012; Kennedy et al., 2015).

Sensitivity analysis with GBM model

To demonstrate how the GBM model can be used to simulate changes in BV10 when varying MP properties, water quality, and carbon characteristics, we conducted a sensitivity analysis. First, we predicted BV10 values for the removal of different perfluoroalkyl carboxylic acids (PFCAs) by the same GAC product in the same water matrix to see how varying L impacts BV10 prediction (Simulation ID 1~7). Second, we varied DOC, pzc, and BET one parameter at a time to see how these variables impact BV10 predictions, respectively (Simulation ID 8~13). Finally, we changed both BET and DOC to evaluate their combined effect on BV10 prediction (Simulation ID 14). Resulting BV10 values from these simulations are summarized in Figure 5a

and the % change in BV10 of Simulation ID 2~14 from Simulation ID 1 can be found in Figure 5b.

a.

Simulation ID	BV10
1	20804
2	6191
3	10248
4	14934
5	19026
6	22156
7	23015
8	28993
9	2936
10	8486
11	14407
12	23199
13	31786
14	53126

b.

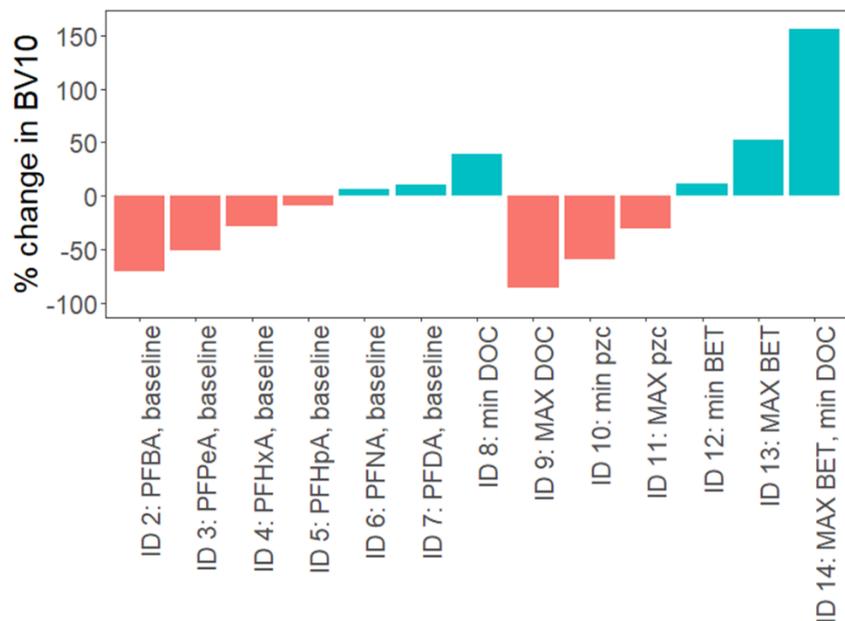


Figure 5. Predicted BV10 (a) and % change in predicted BV10 (b) for all GBM model simulation scenarios: Simulation ID 1 is an entry taken from the test data set which describes PFOA removal from groundwater. Baseline MP removal conditions (ID 1) are water matrix with pH 7.56, DOC = 2.7 mg/L, and UV254 = 0.0259 abs/cm and carbon made of bituminous coal with pzc = 7.4, mp_ratio = 0.167, and BET = 963 m²/g. Simulation ID 1~7 describe treatment scenarios of carboxylic perfluoroalkyl acids of different chain lengths, treated with the same removal conditions as in ID 1. Simulation ID 8~13 are treatment scenarios of PFOA when pzc, BET, DOC was changed one at a time, to either the minimum or the maximum value of the training set. In Simulations ID 8~9, SUVA was kept at 0.96 L/mg-m. Simulation ID 14 is a PFOA removal scenario where BET was maximized while DOC was minimized. Observed BV10 value for ID 1 was 16600 bed volumes and the predicted value was 20800 bed volumes.

By predicting BV10 values for PFCAs of different chain lengths with L values in the 1<L<5 range (Simulation ID 1~7), we further elucidated the impact of L on GAC performance. Results for simulation ID 1~ 7 suggest that when GAC and background water matrix parameters are fixed, larger L corresponds to larger BV10 (Text A13). Relative to results for PFOA, BV10 for PFBA, the smallest among the PFCAs experienced the largest percent decrease while BV10

for PFDA, the largest among the considered PFCAs experienced the largest percent increase. (Figure 5b). Larger values of L correspond to larger compounds (Text A13), for which van der Waals interactions are stronger.

When DOC was varied (Simulation ID 8~9), we observed the expected negative impact of DOC on BV10 prediction. Note that during these simulations, the SUVA value was kept the same (0.96 L/mg-m) to avoid introducing any effects associated with DOM character as expressed by SUVA. We observed that minimizing DOC (0.1 mg/L) led to a 39% increase of BV10 while maximizing DOC (11.4 mg/L) led to a ~86% decrease on BV10 (Figure 5b).

Subsequently, we varied carbon characteristics such as pzc and BET surface area (Simulation ID 10~14) to evaluate their impact on BV10 values. Simulation ID 10 led to a 59% decrease while ID 11 led to a 31% decrease in BV10 prediction, suggesting there is an optimum pzc value between the minimum and maximum pzc values used in these simulations. Minimizing BET (573 m²/g, Simulation ID12) led to a positive change of 12% and maximizing BET (1468 m²/g, Simulation ID13) led to a positive change of 53% in BV10 prediction. These results highlight that the predictor-response relationship between GAC properties and BV10 is highly complex and may be a reflection of the properties of specific GACs for which data were available. Simulation ID 14 showed that maximizing BET (1468 m²/g) while minimizing DOC (0.1 mg/L) simultaneously increased BV10 up to 150% over the base case (Figure 5b).

Implications and limitations

Our ensemble tree models which have an average testing MAE of ~0.11 log-units (Table 1), equivalent to an average of -22% to +29% of error in the linear BV10 scale, are more effective than previous BV10 estimation methods that were based on MLR (e.g. Kennedy et al., 2017). Furthermore, through our sensitivity analysis, we demonstrated that by changing input

values of pzc and BET, an ML model can be used to select effective GAC products to meet treatment goals for a wide range of MPs and background water quality. Also, simulation results indicated that reducing DOC/UV254 during pre-treatment can prolong the service life of GAC bed and reduce carbon use rate (CUR).

The ensemble tree models were proven to be effective in this study, but they possess certain limitations. One limitation is that the models in this study are static— that is, they cannot adapt to sudden changes in influent water quality. To this end, when simulating BV10, the average of influent water quality year-round should be used. Additionally, ML models are only as good as the training data used to construct them. Hence, more data covering a wider range of distribution of MPs, GAC products, and background water qualities can help improve the prediction accuracy and applicability domain.

To conclude, we recommend the use of two models, the RF and GBM models, for predictions of BV10 of recalcitrant organic MPs to develop initial GAC adsorber design and its associated cost estimation. For wide distribution, the test data set was merged with the training data set to tune the final models, which will be made available as a web-based application (Appx. A, Text A14).

REFERENCES

- Abraham, M. H., Ibrahim, A., & Zissimos, A. M. (2004). Determination of sets of solute descriptors from chromatographic measurements. *Journal of Chromatography A*, *1037*(1), 29–47. <https://doi.org/10.1016/j.chroma.2003.12.004>
- Benstoem, F., & Pinnekamp, J. (2017). Characteristic numbers of granular activated carbon for the elimination of micropollutants from effluents of municipal wastewater treatment plants. *Water Science and Technology*, *76*(2), 279–285. <https://doi.org/10.2166/wst.2017.199>
- Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 5–32. <https://doi.org/10.1023/A:1010933404324>
- Brown, T. N., & Wania, F. (2009). Development and Exploration of an Organic Contaminant Fate Model Using Poly-Parameter Linear Free Energy Relationships. *Environmental Science & Technology*, *43*(17), 6676–6683. <https://doi.org/10.1021/es901205j>
- Corwin, C. J., & Summers, R. S. (2012). Controlling trace organic contaminants with GAC adsorption. *Journal - AWWA*, *104*(1), E36–E47. <https://doi.org/10.5942/jawwa.2012.104.0004>
- Crittenden, J. C., Berrigan, J. K., & Hand, D. W. (1986). Design of Rapid Small-Scale Adsorption Tests for a Constant Diffusivity. *Journal (Water Pollution Control Federation)*, *58*(4), 312–319. JSTOR.
- Crittenden, J. C., Reddy, P. S., Arora, H., Trynoski, J., Hand, D. W., Perram, D. L., & Summers, R. S. (1991). Predicting GAC Performance With Rapid Small-Scale Column Tests. *Journal (American Water Works Association)*, *83*(1), 77–87.

- Crittenden John C., Berrigan John K., Hand David W., & Lykins Ben. (1987). Design of Rapid Fixed-Bed Adsorption Tests for Nonconstant Diffusivities. *Journal of Environmental Engineering*, 113(2), 243–259. [https://doi.org/10.1061/\(ASCE\)0733-9372\(1987\)113:2\(243\)](https://doi.org/10.1061/(ASCE)0733-9372(1987)113:2(243))
- Friedman, J. H. (2001). Greedy Function Approximation: A Gradient Boosting Machine. *The Annals of Statistics*, 29(5), 1189–1232.
- Götz, C. W., Scheringer, M., MacLeod, M., Roth, C. M., & Hungerbühler, K. (2007). Alternative Approaches for Modeling Gas–Particle Partitioning of Semivolatile Organic Chemicals: Model Development and Comparison. *Environmental Science & Technology*, 41(4), 1272–1278. <https://doi.org/10.1021/es060583y>
- Hastie, T., Qian, J., & Tay, K. (n.d.). *An Introduction to glmnet*. 38.
- Howard, P. H., Stiteler, W. M., Meylan, W. M., Hueber, A. E., Beauman, J. A., Larosche, M. E., & Boethling, R. S. (1992). Predictive model for aerobic biodegradability developed from a file of evaluated biodegradation data. *Environmental Toxicology and Chemistry*, 11(5), 593–603. <https://doi.org/10.1002/etc.5620110502>
- Inglis, A., Parnell, A., Hurley, C., 2021. Visualizing Variable Importance and Variable Interaction Effects in Machine Learning Models.
- Kano, H., Umeda, Y., Kasai, T., Sasaki, T., Matsumoto, M., Yamazaki, K., Nagano, K., Arito, H., & Fukushima, S. (2009). Carcinogenicity studies of 1,4-dioxane administered in drinking-water to rats and mice for 2years. *Food and Chemical Toxicology*, 47(11), 2776–2784. <https://doi.org/10.1016/j.fct.2009.08.012>

- Karanfil, T., & Kilduff, J. E. (1999). Role of Granular Activated Carbon Surface Chemistry on the Adsorption of Organic Compounds. 1. Priority Pollutants. *Environmental Science & Technology*, 33(18), 3217–3224. <https://doi.org/10.1021/es981016g>
- Karanfil, T., Kitis, M., Kilduff, J. E., & Wigton, A. (1999). Role of Granular Activated Carbon Surface Chemistry on the Adsorption of Organic Compounds. 2. Natural Organic Matter. *Environmental Science & Technology*, 33(18), 3225–3233. <https://doi.org/10.1021/es9810179>
- Karanfil, T., Schlautman, M. A., & Erdogan, I. (2002). Survey of DOC and UV measurement practices with implications for SUVA determination. *Journal AWWA*, 94(12), 68–80. <https://doi.org/10.1002/j.1551-8833.2002.tb10250.x>
- Kearns, J., Dickenson, E., & Knappe, D. (2020). Enabling Organic Micropollutant Removal from Water by Full-Scale Biochar and Activated Carbon Adsorbers Using Predictions from Bench-Scale Column Data. *Environmental Engineering Science, Journal Article*. <https://doi.org/10.1089/ees.2019.0471>
- Kearns, J. P., Wellborn, L. S., Summers, R. S., & Knappe, D. R. U. (2014). 2,4-D adsorption to biochars: Effect of preparation conditions on equilibrium adsorption capacity and comparison with commercial activated carbon literature data. *Water Research*, 62, 20–28. <https://doi.org/10.1016/j.watres.2014.05.023>
- Kennedy, A. M., Reinert, A. M., Knappe, D. R. U., Ferrer, I., & Summers, R. S. (2015a). Full- and pilot-scale GAC adsorption of organic micropollutants. *Water Research*, 68, 238–248. <https://doi.org/10.1016/j.watres.2014.10.010>

- Kennedy, A. M., Reinert, A. M., Knappe, D. R. U., Ferrer, I., & Summers, R. S. (2015b). Full- and pilot-scale GAC adsorption of organic micropollutants. *WATER RESEARCH*, 68, 238–248. <https://doi.org/10.1016/j.watres.2014.10.010>
- Kennedy, A. M., Reinert, A. M., Knappe, D. R. U., & Summers, R. S. (2017). Prediction of Full-Scale GAC Adsorption of Organic Micropollutants. *ENVIRONMENTAL ENGINEERING SCIENCE*, 34(7), 496–507. <https://doi.org/10.1089/ees.2016.0525>
- Kuhn, M. (2008). Building Predictive Models in R Using the caret Package. *Journal of Statistical Software*, 28(1), 1–26. <https://doi.org/10.18637/jss.v028.i05>
- Li, Q., Snoeyink, V. L., Mariñas, B. J., & Campos, C. (2003). Elucidating competitive adsorption mechanisms of atrazine and NOM using model compounds. *Water Research*, 37(4), 773–784. [https://doi.org/10.1016/S0043-1354\(02\)00390-1](https://doi.org/10.1016/S0043-1354(02)00390-1)
- Liu, C. J., Werner, D., & Bellona, C. (2019). Removal of per- and polyfluoroalkyl substances (PFASs) from contaminated groundwater using granular activated carbon: A pilot-scale study with breakthrough modeling. *ENVIRONMENTAL SCIENCE-WATER RESEARCH & TECHNOLOGY*, 5(11), 1844–1853. <https://doi.org/10.1039/c9ew00349e>
- Matsui, Y., Knappe, D. R. U., Iwaki, K., & Ohira, H. (2002). Pesticide Adsorption by Granular Activated Carbon Adsorbers. 2. Effects of Pesticide and Natural Organic Matter Characteristics on Pesticide Breakthrough Curves. *Environmental Science & Technology*, 36(15), 3432–3438. <https://doi.org/10.1021/es011366u>
- Merle, T., Knappe, D. R. U., Pronk, W., Vogler, B., Hollender, J., & Gunten, U. von. (2020). Assessment of the breakthrough of micropollutants in full-scale granular activated carbon adsorbers by rapid small-scale column tests and a novel pilot-scale sampling approach.

Environmental Science: Water Research & Technology.

<https://doi.org/10.1039/D0EW00405G>

Nam, S.-W., Choi, D.-J., Kim, S.-K., Her, N., & Zoh, K.-D. (2014). Adsorption characteristics of selected hydrophilic and hydrophobic micropollutants in water using activated carbon.

Journal of Hazardous Materials, 270, 144–152.

<https://doi.org/10.1016/j.jhazmat.2014.01.037>

Newcombe, G., & Drikas, M. (1997). Adsorption of NOM onto activated carbon: Electrostatic and non-electrostatic effects. *Carbon*, 35(9), 1239–1250. [https://doi.org/10.1016/S0008-](https://doi.org/10.1016/S0008-6223(97)00078-X)

[6223\(97\)00078-X](https://doi.org/10.1016/S0008-6223(97)00078-X)

Pelch, K. E., Reade, A., Wolffe, T. A. M., & Kwiatkowski, C. F. (2019). PFAS health effects database: Protocol for a systematic evidence map. *Environment International*, 130, 104851.

<https://doi.org/10.1016/j.envint.2019.05.045>

Pelekani, C., & Snoeyink, V. L. (2000). Competitive adsorption between atrazine and methylene blue on activated carbon: The importance of pore size distribution. *Carbon*, 38(10), 1423–

1436. [https://doi.org/10.1016/S0008-6223\(99\)00261-4](https://doi.org/10.1016/S0008-6223(99)00261-4)

Pelekani, C., & Snoeyink, V. L. (2001). A kinetic and equilibrium study of competitive adsorption between atrazine and Congo red dye on activated carbon: The importance of pore size distribution. *Carbon*, 39(1), 25–37. [https://doi.org/10.1016/S0008-6223\(00\)00078-](https://doi.org/10.1016/S0008-6223(00)00078-6)

[6](https://doi.org/10.1016/S0008-6223(00)00078-6)

- Quinlivan, P. A., Li, L., & Knappe, D. R. U. (2005). Effects of activated carbon characteristics on the simultaneous adsorption of aqueous organic micropollutants and natural organic matter. *Water Research*, *39*(8), 1663–1673. <https://doi.org/10.1016/j.watres.2005.01.029>
- R Core team. (2020). *R: The R Project for Statistical Computing*. <https://www.r-project.org/>
- Ridgeway, G. (2020). *Generalized Boosted Models: A guide to the gbm package*. 15.
- Rohatgi, A. (2020). *WebPlotDigitizer User Manual Version 4.3*. 23.
- Sgroi, M., Anumol, T., Roccaro, P., Vagliasindi, F. G. A., & Snyder, S. A. (2018). Modeling emerging contaminants breakthrough in packed bed adsorption columns by UV absorbance and fluorescing components of dissolved organic matter. *Water Research*, *145*, 667–677. <https://doi.org/10.1016/j.watres.2018.09.018>
- Shimabuku, K. K., Kennedy, A. M., Mulhern, R. E., & Summers, R. S. (2017). Evaluating Activated Carbon Adsorption of Dissolved Organic Matter and Micropollutants Using Fluorescence Spectroscopy. *Environmental Science & Technology*, *51*(5), 2676–2684. <https://doi.org/10.1021/acs.est.6b04911>
- Sontheimer, H., Crittenden, J. C., & Summers, R. S. (1988). *Activated carbon for water treatment*. DVGW-Forschungsstelle, Engler-Bunte-Institut, Universitat Karlsruhe (TH).
- Tibshirani, R. (1996). Regression Shrinkage and Selection Via the Lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, *58*(1), 267–288. <https://doi.org/https://doi.org/10.1111/j.2517-6161.1996.tb02080.x>

- Ulrich, N., Endo, S., Brown, T. N., Watanabe, N., Bronner, G., Abraham, M. H., & Gross, K.-U. (2017). *UFZ-LSER database v 3.2.1*. Leipzig, Germany, Helmholtz Centre for Environmental Research-UFZ. <http://www.ufz.de/lserd>
- Wright, M. N., & Ziegler, A. (2017). ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R. *Journal of Statistical Software*, 77(1).
<https://doi.org/10.18637/jss.v077.i01>
- Wu, F.-C., Tseng, R.-L., & Hu, C.-C. (2005). Comparisons of pore properties and adsorption performance of KOH-activated and steam-activated carbons. *Microporous and Mesoporous Materials*, 80(1), 95–106. <https://doi.org/10.1016/j.micromeso.2004.12.005>
- Zhang, K., Zhong, S., & Zhang, H. (2020). Predicting Aqueous Adsorption of Organic Compounds onto Biochars, Carbon Nanotubes, Granular Activated Carbons, and Resins with Machine Learning. *Environ. Sci. Technol.*, 11.
<https://pubs.acs.org/doi/10.1021/acs.est.0c02526?ref=pdf>
- Zietzschmann, F., Stuetzer, C., & Jekel, M. (2016). Granular activated carbon adsorption of organic micro-pollutants in drinking water and treated wastewater—Aligning breakthrough curves and capacities. *WATER RESEARCH*, 92, 180–187.
<https://doi.org/10.1016/j.watres.2016.01.056>

APPENDICES

Appendix A

Appendix A includes Figures A1~6, and Tables A1~2, and Text A1~14.

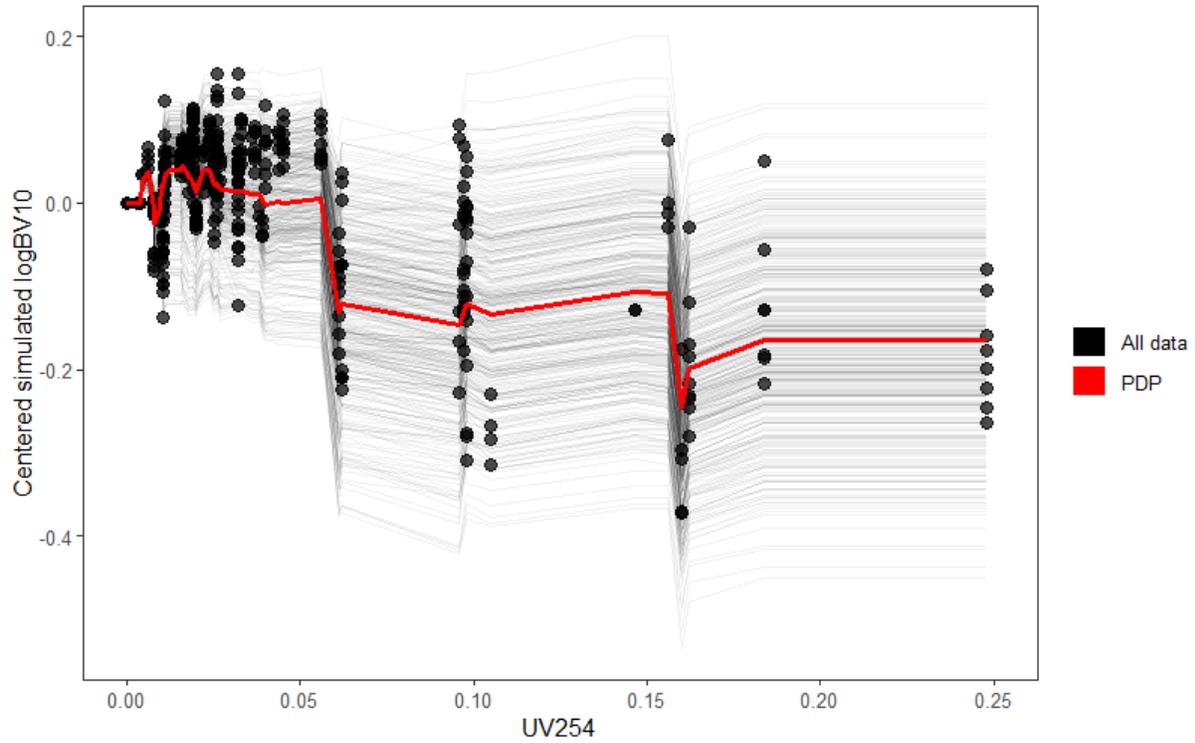


Figure A 3. PDP and c-ICEs of UV254 on logBV10. While there is some vertical spread of the c-ICEs, PDP suggest that EBCT affects logBV10 prediction on average by up to 0.1 units in the range of EBCT of data set.

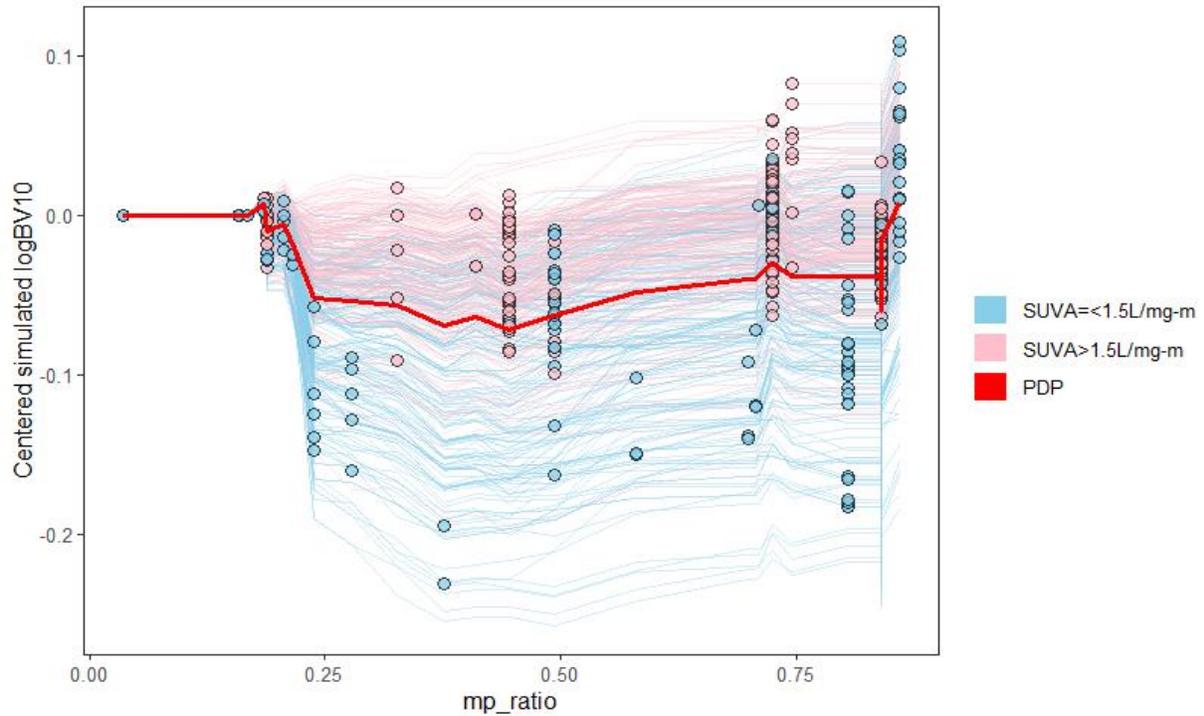


Figure A 4. PDP and c-ICEs of EBCT on logBV10. While there is some vertical spread of the c-ICEs, PDP suggest that EBCT affects logBV10 prediction on average by up to 0.1 units in the range of EBCT of data set. Blue lines/dots represent c-ICEs/predictions for data entries with $SUVA < 1.5 \text{ L/mg-m}$; pink lines/dots represent the those of the complimentary set. Red line represents PDP.

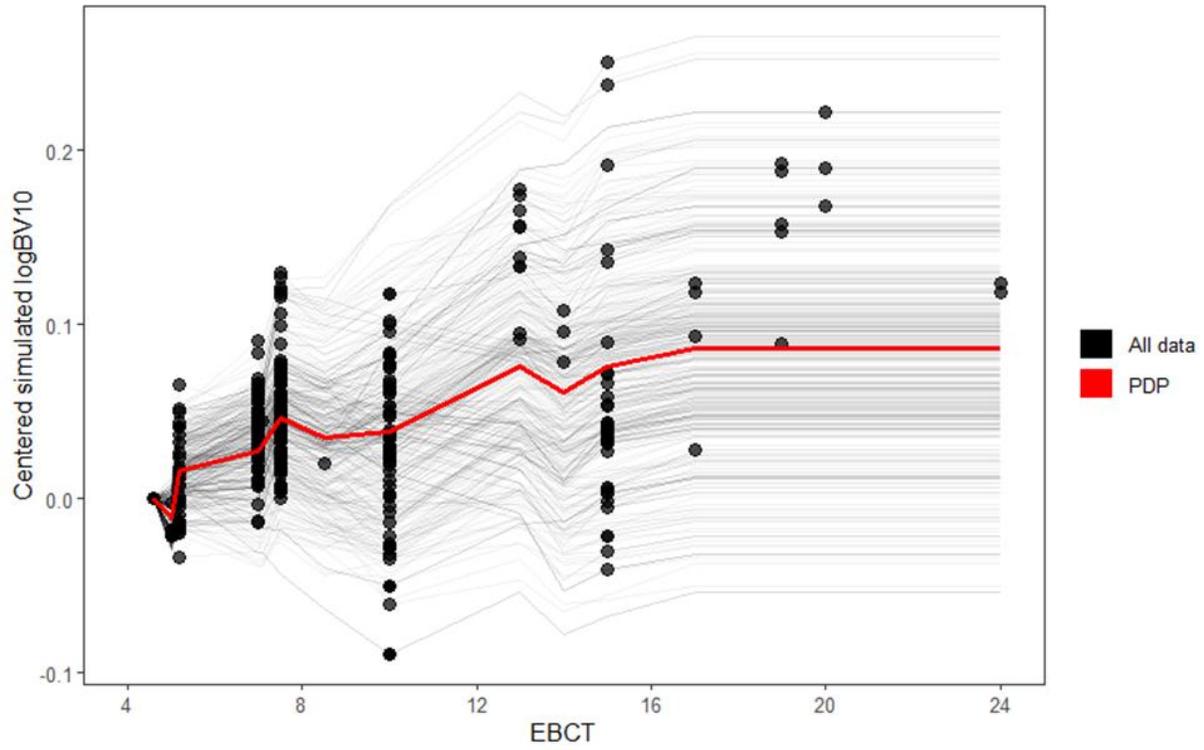


Figure A 5. PDP and c-ICEs of EBCT on logBV10. While there is some vertical spread of the c-ICEs, PDP suggest that EBCT affects logBV10 prediction on average by up to 0.1 units in the range of EBCT of data set.

Tables A1~2

Table A 1. Summary of the 17 input and one response variables. Eight chemical property parameters, 4 water quality parameters, 3 carbon property parameters and 3 (of which 2 are dummy variables converted from categorical variables) operational condition parameters are summarized by their abbreviation, numerical value range and variable category.

Variable	Abbreviation	Variable type, Range	Variable category
Initial concentration (ug/L)	C0	Continuous, 0.0012 ~ 1000	MP properties
Table A 1 (Continued)			
Charge of MP, dependent on pH	charge	Continuous, -1 ~1	MP properties
Partitioning coefficient between gas phase and hexadecane ($\log L^{16}$)	L	Continuous, 1.404 ~26.130	MP properties
Partitioning coefficient between gas phase and hexadecane ($\log L^{16}$)	B	Continuous, 0.00~ 4.96	MP properties
Excessive molar refraction (R2... ($\text{cm}^3 \text{mol}^{-1}$)/10)	E	Continuous, -1.49 ~ 4.60	MP properties

Table A 1. (Continued)

Polarity/polarizability (π_2^H)	S	Continuous, -0.88 ~5.02	MP properties
Hydrogen bonding acidity of solute (α_2^H)	A	Continuous, 0.00 ~1.55	MP properties
McGowan volume characteristic ($V_x \dots$ ($\text{cm}^3 \text{mol}^{-1}$)/100)	V	Continuous, 0.4698 ~11.0250	MP properties
pH of water	pH	Continuous, 5.7 ~8.5	Water quality
Dissolved organic carbon concentration(mg/L)	DOC	Continuous, 0.1 ~11.4	Water quality
UV absorbance at 254nm (abs/cm)	UV254	Continuous, 0.000 ~0.248	Water quality
Micropore volume to total pore volume ratio	mp_ratio	Continuous, 0.035 ~0.86	Carbon characteristics
Brunauer, Emmett and Teller surface area (m^2/g)	BET	Continuous, 573 ~1468	Carbon characteristics

Table A 1. (Continued)

pH at which the carbon reaches point of zero charge	pzc	Continuous, 5.5 ~10	Carbon characteristics
Empty bed contact time of column (min)	EBCT	Continuous, 4.6 ~24	Operational conditions
Proportional diffusivity RSSCT	PD (dummy variable)	Binary, 0/1	Operational conditions
Constant diffusivity RSSCT	CD (dummy variable)	Binary, 0/1	Operational conditions
Log ₁₀ of bed volume to 10% breakthrough	y	3 ~5	Response to predict

Table A 2. Semi-partial R^2 of all input variables are summarized below. Variables with p values of <0.05 are considered statistically significant. Semi-partial R^2 of each variable was calculated by fitting them one at a time into a linear regression to predict logBV10.

Variable	R²	p value
C0	0.004	0.184
charge	0.095	<0.05
L	0.027	<0.05
B	0.017	<0.05
E	0.022	<0.05
S	0.010	<0.05
A	0.038	<0.05
V	0.022	<0.05
pH	0.009	<0.05
DOC	0.146	<0.05
UV254	0.157	<0.05
mp_ratio	0.038	<0.05
BET	0.281	<0.05
pzc	0.183	<0.05
EBCT	0.006	0.102
CD	0.143	<0.05
PD	0.080	<0.05

Text A1 Recalcitrancy of MPs

A non-linear model, Biowin 2 (Boethling et al., 2003), was used to determine recalcitrancy of MPs as it exhibited a training classification accuracy of 86.2% for slowly degrading compounds, which was the highest amongst all the models available in BioWIN (Boethling et al., 2003) within the EPI Suite™ 4.10 (US EPA, 2021). An MP's fast biodegradation probability was calculated with this model, and the MP was classified as recalcitrant if the probability was less than 0.5. Calculation results of biodegradation probabilities of the MPs included in this study can be found in Appendix C, which is available in a zip format.

Text A2 Abraham solvation parameters

Abraham solvation parameters were taken from the UFZ-LSER data base (Ulrich et al., 2017) where experimental values of Abraham parameters, L, S, E, A, B, V, were compiled from past studies and where experimental values were not available, the database also provided a tool to calculate theoretical values. The source of individual Abraham parameters corresponding to individual compound can be found in Supplemental data 2.

Text A3 Carbon character information

External references for carbon character inputs were utilized when the subject GAC study does not contain information on pzc, mp_ratio, and BET. The values for these variables of each carbon and their corresponding references are summarized in Table A3-1.

Table A3-1. Summary of external reference values of GAC characteristics, pzc, mp_ratio, and BET.

Product Name	pzc	Citation	mp_ratio	Citation	BET	Citation
Calgon F300	8.03	(Selmi et al., 2020)	0.84	Calgon Carbon, personal communication, 2016	899	Calgon Carbon, personal communication, 2016
Calgon F400	8.29	(Zhi and Liu, 2016)	0.86	Calgon Carbon, personal communication, 2016	933	Calgon Carbon, personal communication, 2016
Calgon F600	8.53	(Yan et al., 2016)	0.54	(Yan et al., 2016)	824	(Yan et al., 2016)
Calgon F820	6.1	(Chae et al., 2013)	0.71	(Chae et al., 2013)	1027	(Chae et al., 2013)
Donau carbon epibon a 8*30	8.1	(Aschermann et al., 2018)	0.75	Calgon Carbon, personal communication, 2016	1083	Calgon Carbon, personal communication, 2016

Table A3- 1 (Continued)

Hydrodarco 4000	7	(Reed, 1995)	0.28	(Redding et al., 2009)	575	(Reed, 1995)
Norit 1240	8.8	(Yapsakli et al., 2009)	0.72	(Vanderhey den et al., 2016)	1468	(Vanderheyden et al., 2016)
Norit darco 1240	6.4	(Anumol et al., 2015)	0.45	(Karanjkar et al., 2016)	671	(Karanjkar et al., 2016)
Norit GAC 400	0.17	Calgon Carbon, personal communication , 2016	0.05	Calgon Carbon, personal communica tion, 2016	963	Calgon Carbon, personal communication, 2016
Norit GCN 1240	8	(Erto et al., 2010)	0.94	Calgon Carbon, personal communica tion, 2016	1118	Calgon Carbon, personal communication, 2016
Norit N1240	5.5	(Hnatukova et al., 2011)	0.41	(Hnatukova et al., 2011)	1110	(Hnatukova et al., 2011)

Text A4 Analysis of breakthrough data to calculate y

Here, y is either the log transformation of BV10 itself in the case of pilot/full scale data or scaled up logBV10 results in the case of RSSCT data. In either case, we begin the process by analyzing breakthrough curve to acquire BV10.

First, breakthrough curves in graphical format were digitized and tabulated in Excel, using WebPlotDigitizer (Rohatgi, 2020). The validity of this digitization approach has been studied by Drevon et al. (2017). Only parts of the breakthrough curves that were needed to capture the behavior around BV10 were extracted.

To get BV10, a process-based model (Sontheimer, 1988), Pore Surface Diffusion Model (PSDM), was first fitted to the digitized breakthrough data, by calibrating both the kinetic and capacity parameters (model outputs along with digitized breakthrough data can be found in Supplemental data 3). AdDesignS (Michigan Technological University, 2005) was used for this step of calibrating the PSDM and outputting the numerical solutions. Subsequently, BV10 was interpolated from the fitted PSDM model output. For instance, the figure below illustrates this interpolation step, where BV10 was interpolated from the PSDM fitted to a specific set of breakthrough data. A summary of all breakthrough data acquired from graphs and fitted with PSDM can be found in Appendix C.

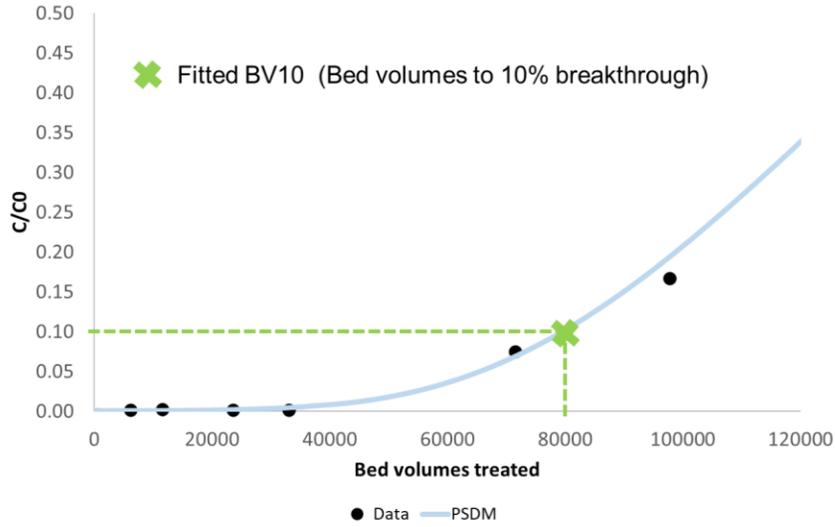


Figure A4- 1 Example breakthrough data with fitted Pore Surface Diffusion Model (PSDM) of carbamazepine in pilot scale column.

For data with author fitted PDSM, BV10 value was directly copied from the study. For pilot/full scale column data, the interpolated values were directly log-transformed as y . In the figure above for instance, y would equal to $\log_{10}(80000)$. For bench-scale (RSSCT) data, the interpolated BV10 was first log-transformed then scaled-up using an empirical equation (Eq S2, which was re-written from Eq S1) developed by Kearns et al (2020) as y .

Original Kearns equation:

$$\ln BV_{10\%}^{full\ scale} = (0.8 \pm 0.039) \ln BV_{10\%}^{RSSCT} + (1.21 \pm 0.41) \quad (S1)$$

\log_{10} transformation of Kearns equation

$$\log BV_{10\%}^{full} = 0.8 \times \log BV_{10\%}^{RSSCT} + \frac{1.21}{1n(10)} \quad (S2)$$

Text A5 MLR model development

MLR model was developed in a two-step manner— variable selection and fitting of MLR with ordinary least squares onto the training set using selected variables.

LASSO (Tibshirani, 1996) was selected as a tool for variable selection as the algorithm (Eq S3) can select predictive models even under the presence of collinearity (Dormann et al, 2013) and penalizes useless variables and shrink their coefficients to zero (Eq S3). The hyperparameter lambda, λ , was tuned via 10-fold CV in concurrent with feeding a combination of input variables. To perform feature selection via LASSO, a combination of input variables was standardized and fed into the algorithm and as a result, a combination of features was returned while lambda was set to the largest possible value that minimized the RMSE during CV. The feature selection results are shown in Table A 3 and the lambda values obtained from CV are summarized in Table A 4.

$$\sum_{i=1}^n (y_i - \beta_0 - \sum_{j=1}^p \beta_j x_{ij})^2 + \lambda \sum_{j=1}^p |\beta_j| = RSS + \lambda \sum_{j=1}^p |\beta_j| \quad (S3)$$

A recent study by Leeuwenberg et al (2021) suggested that although collinearity within variables wouldn't affect predictive outcomes, strong predictor selection methods such as LASSO can result in unstable variables selection. Collinearity by itself does not affect the predictive power of the model but model interpretability could decrease (Midi et al, 2010).

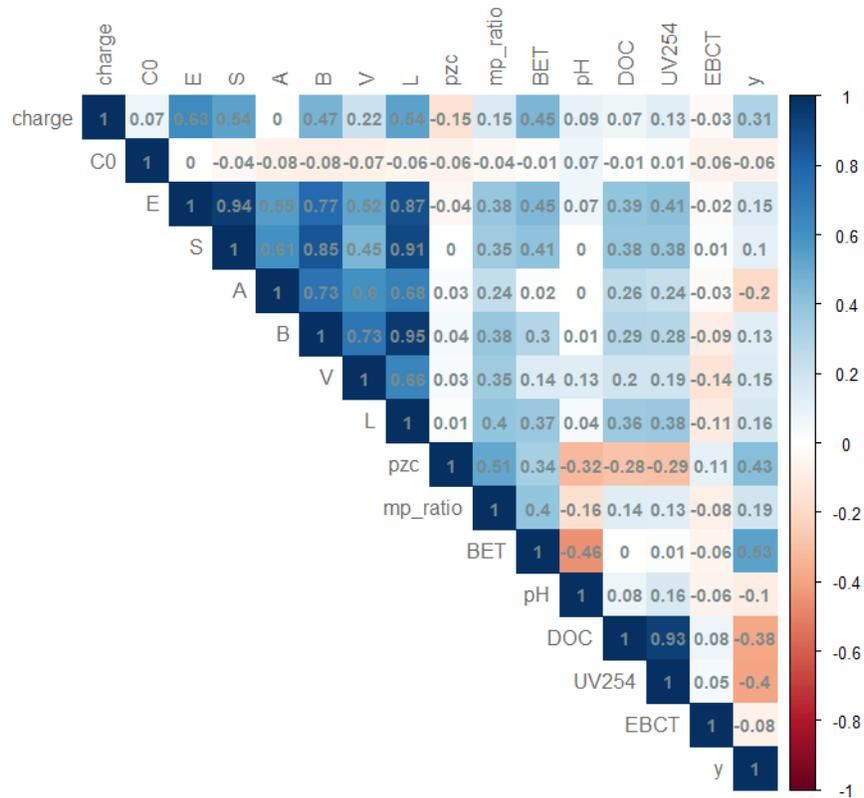


Figure A5- 1. Pair-wise Pearson correlation coefficient between predictor variables and the response variable. Pearson coefficients between each pair of the 16 variables excluding the binary variables CD and PD are summarized up to 2 digits after decimal in a matrix with color contour.

To this end, an effort to minimize the collinearity impact while preserving the maximum amount of information as all input variables were selected based on domain knowledge was made. First, Pearson correlation coefficient (Figure A5-1) was calculated between each possible pair of variables amongst both the response and the predictor variables (Fig S 1). Also, semi-partial least square, in words, how much variance one variable alone explains, were calculated for each variable (Table A 2). Subsequently, three modes of collinearity removal were conducted. In the first mode, we removed variables that are highly collinear, $|r| > 0.9$, as this is the threshold for highly problematic (Midi et al, 2010). Two pairs of variables, L/B, DOC/UV254, were identified as highly

collinear and only one from each pair was included in a model at a time (Table A5-1, ID 1, 2, 7, 8). Although S is collinear with some variables, its partial R^2 is lower than its collinear pairs and its p-value is higher than its collinear pairs (Table A 2); therefore, the collinear effect associated with S was ignored. In the second mode, only collinearity between DOC and UV254 was accounted for (Table A5-1, ID 3, 9). In the third mode, all candidate variables were fed into LASSO, ignoring any collinearity effects (Table A5-1, ID 10, 14) as previous studies have included such collinear variables to model contaminant fate (Brown and Wania, 2009). These 3 modes yielded 14 combinations of input variables to feed into LASSO for variables selection (Table A5-1). We also explored models that excluded S completely in the three modes (Table A5-1, ID 4, 5, 6, 11, 12, 13). During this feature selection process, an additional 51 data points were merged with the existing training set when the input combination did not include UV254 (Table A5-1, ID 1~6).

During LASSO feature selection, a hyperparameter, λ (Eq S1), was tuned through grid search via 10-fold cross-validation (CV). A λ value that minimized the mean-squared-error during CV was chosen and the CV results for the selection of λ are shown below (Figure A5-2~15). The R package “glmnet” (Hastie et al., 2010) was used for this procedure. The package by default computes two special lambda values— a value λ that gives minimum mean cross-validated error and a value of λ that gives the most regularized model such that the cross-validated error is within one standard error of the minimum. We chose the former since it would give a less parsimonious model.

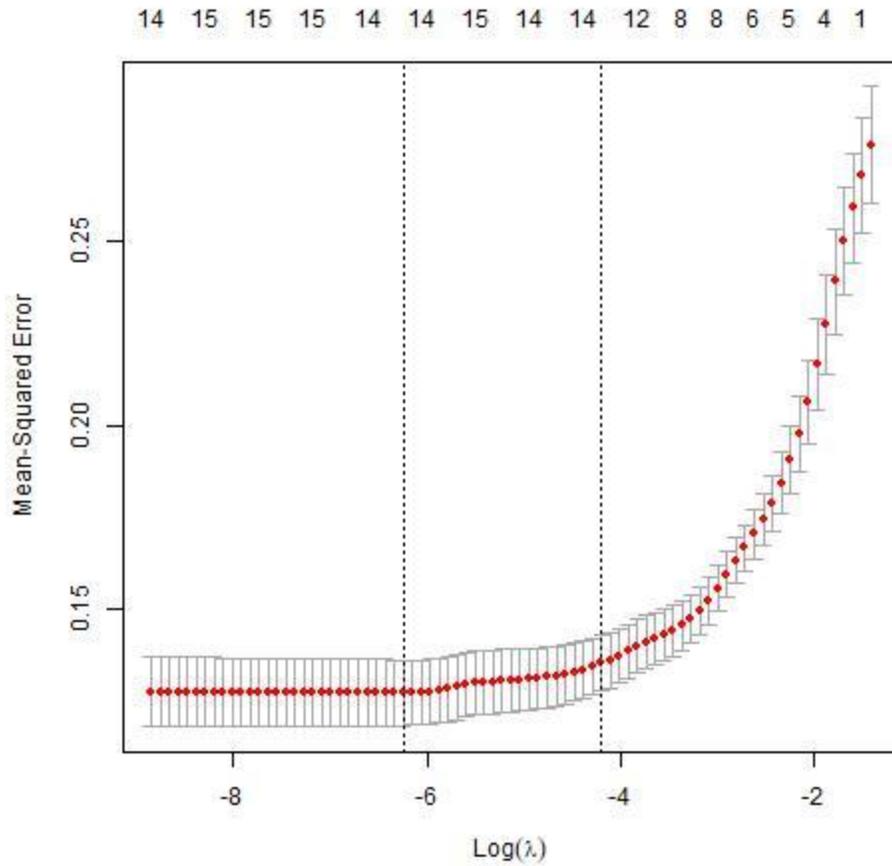


Figure A5- 2. CV error that results from applying LASSO with a series of values of λ for model ID 1.

The primary x-axis (bottom) is λ in natural log scale while the secondary x-axis (top) is the number of features selected at a given λ value. The first dashed line is the λ value that minimized the CV MSE while the second dashed line is the largest λ at which the MSE is within one standard error of the smallest MSE. The first value was taken for LASSO feature selection.

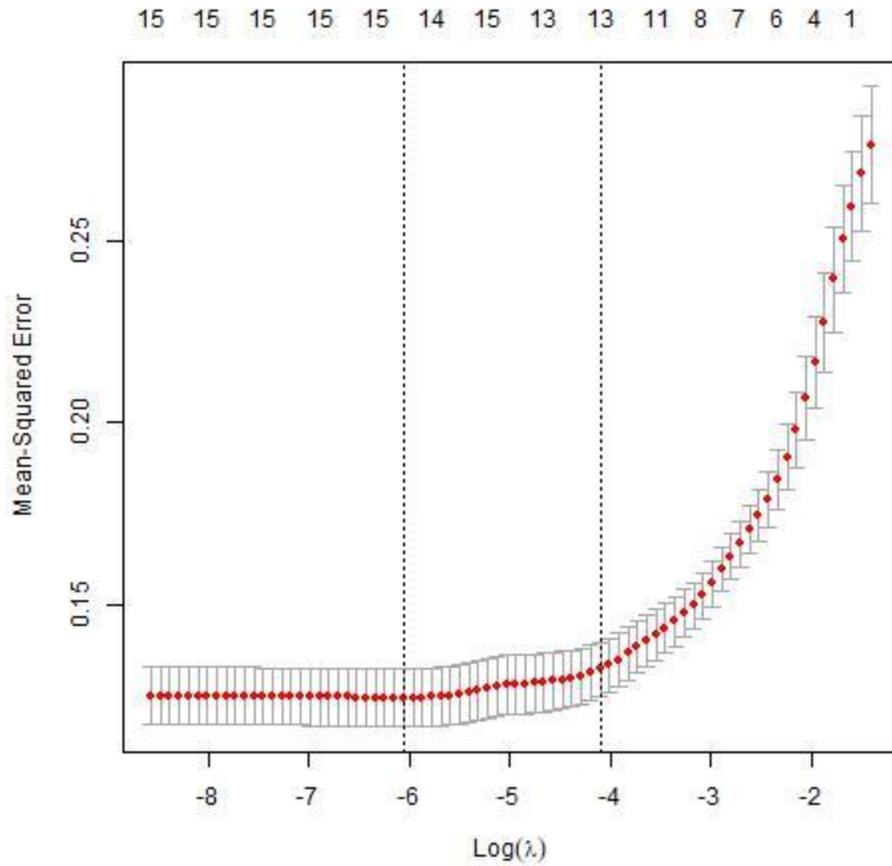


Figure A5- 3. CV error that results from applying LASSO with a series of values of λ for model ID 2.

The primary x-axis (bottom) is λ in natural log scale while the secondary x-axis (top) is the number of features selected at a given λ value. The first dashed line is the λ value that minimized the CV MSE while the second dashed line is the largest λ at which the MSE is within one standard error of the smallest MSE. The first value was taken for LASSO feature selection.

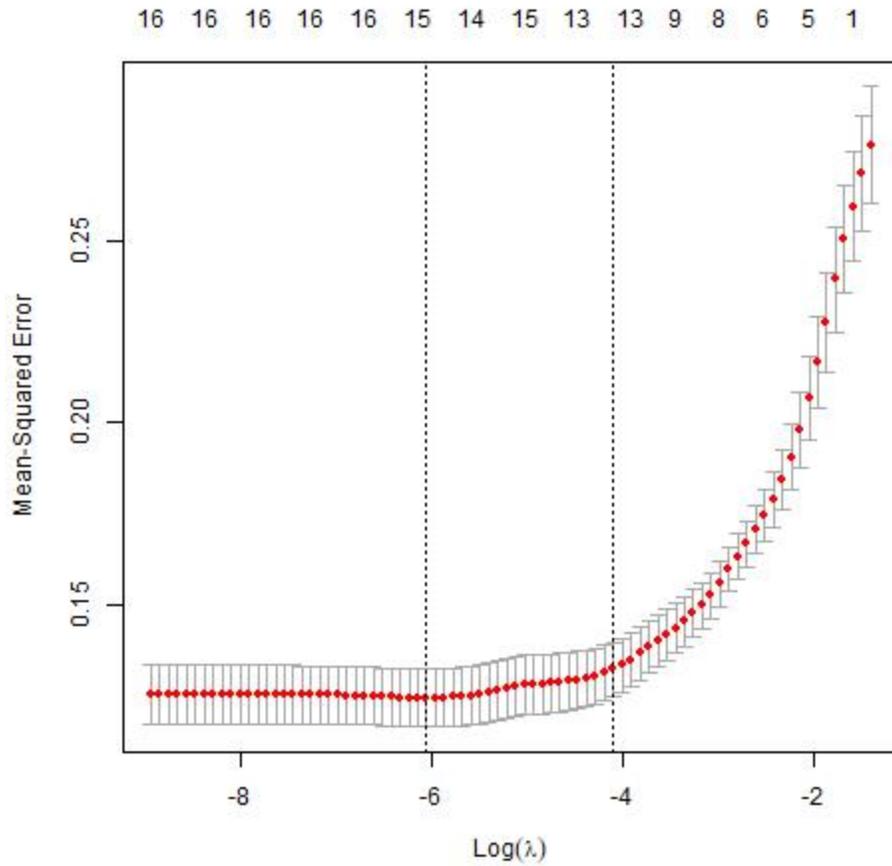


Figure A5- 4. CV error that results from applying LASSO with a series of values of λ for model ID 3.

The primary x-axis (bottom) is λ in natural log scale while the secondary x-axis (top) is the number of features selected at a given λ value. The first dashed line is the λ value that minimized the CV MSE while the second dashed line is the largest λ at which the MSE is within one standard error of the smallest MSE. The first value was taken for LASSO feature selection.

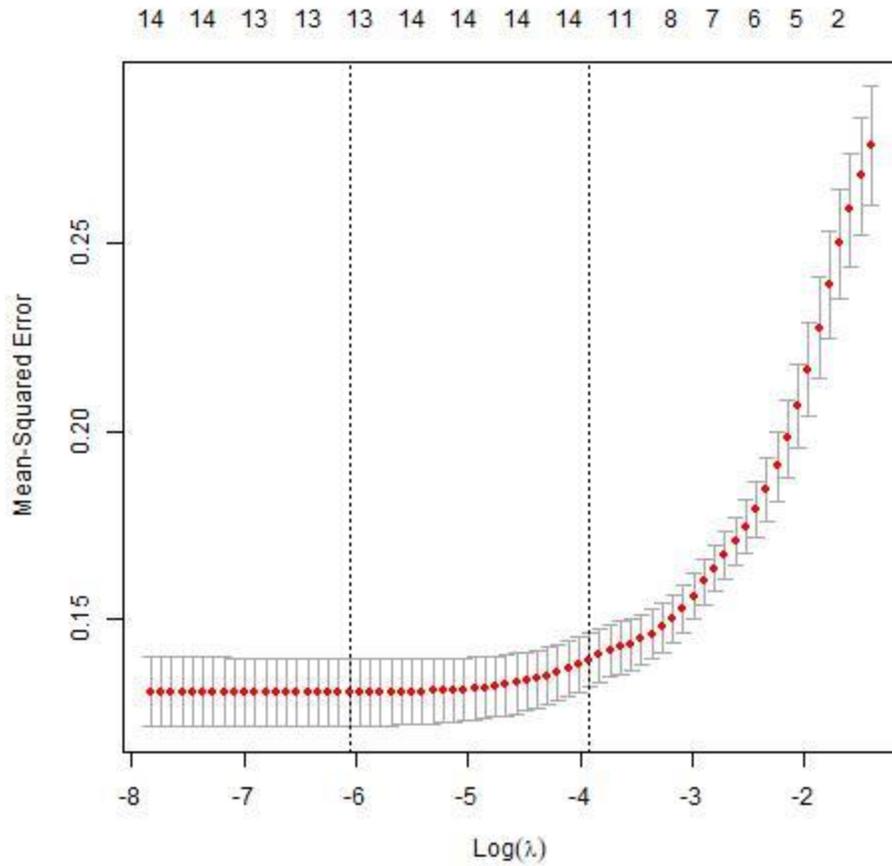


Figure A5- 5. CV error that results from applying LASSO with a series of values of λ for model ID 4.

The primary x-axis (bottom) is λ in natural log scale while the secondary x-axis (top) is the number of features selected at a given λ value. The first dashed line is the λ value that minimized the CV MSE while the second dashed line is the largest λ at which the MSE is within one standard error of the smallest MSE. The first value was taken for LASSO feature selection.

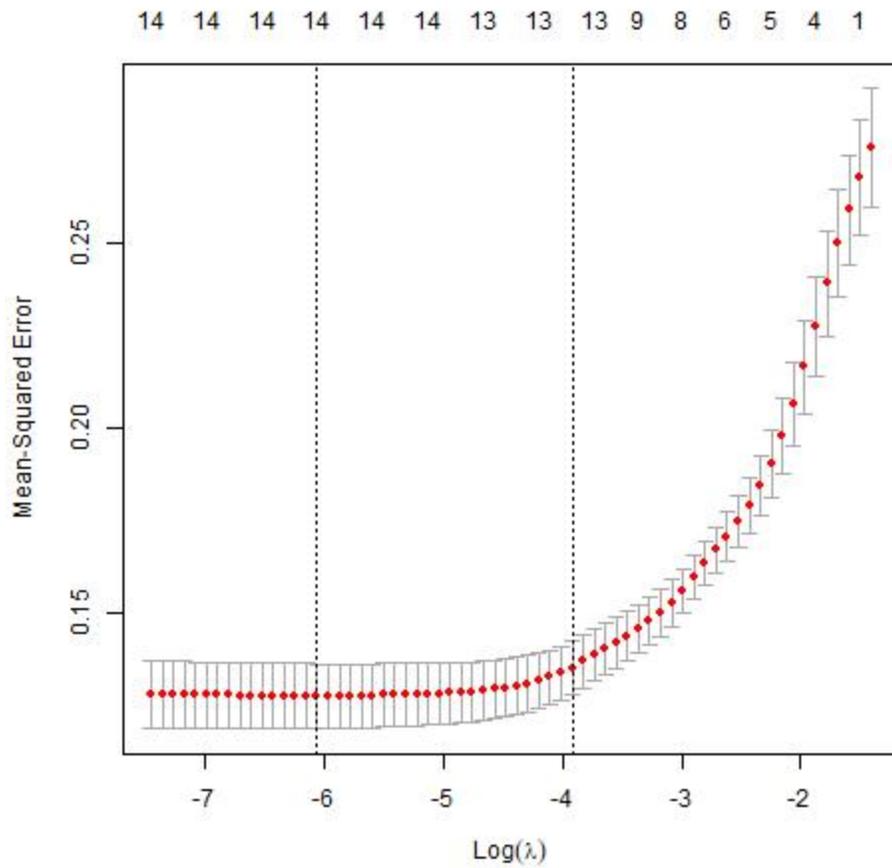


Figure A5- 6. CV error that results from applying LASSO with a series of values of λ for model ID 5.

The primary x-axis (bottom) is λ in natural log scale while the secondary x-axis (top) is the number of features selected at a given λ value. The first dashed line is the λ value that minimized the CV MSE while the second dashed line is the largest λ at which the MSE is within one standard error of the smallest MSE. The first value was taken for LASSO feature selection.

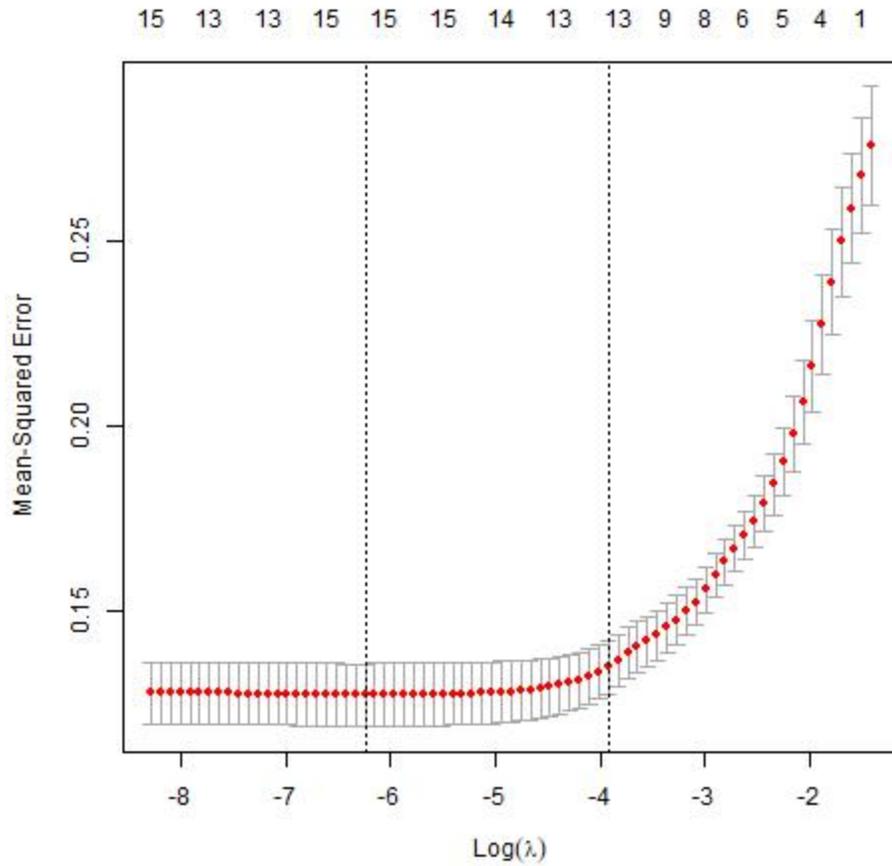


Figure A5- 7. CV error that results from applying LASSO with a series of values of λ for model ID 6.

The primary x-axis (bottom) is λ in natural log scale while the secondary x-axis (top) is the number of features selected at a given λ value. The first dashed line is the λ value that minimized the CV MSE while the second dashed line is the largest λ at which the MSE is within one standard error of the smallest MSE. The first value was taken for LASSO feature selection.

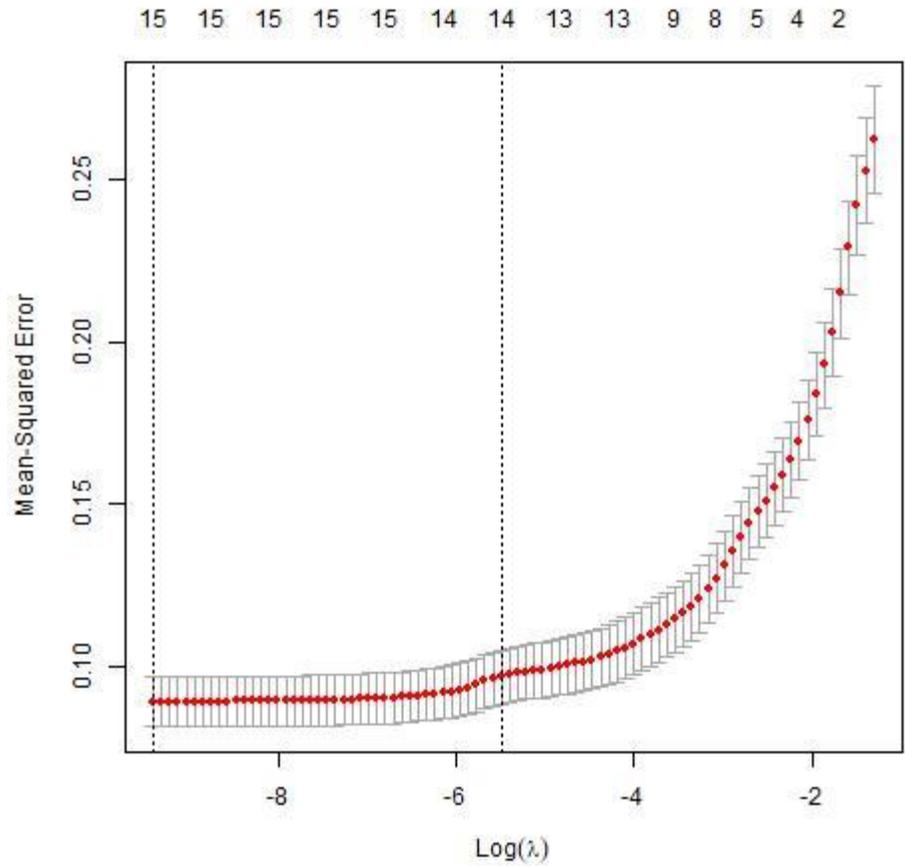


Figure A5- 8. CV error that results from applying LASSO with a series of values of λ for model ID 7.

The primary x-axis (bottom) is λ in natural log scale while the secondary x-axis (top) is the number of features selected at a given λ value. The first dashed line is the λ value that minimized the CV MSE while the second dashed line is the largest λ at which the MSE is within one standard error of the smallest MSE. The first value was taken for LASSO feature selection.

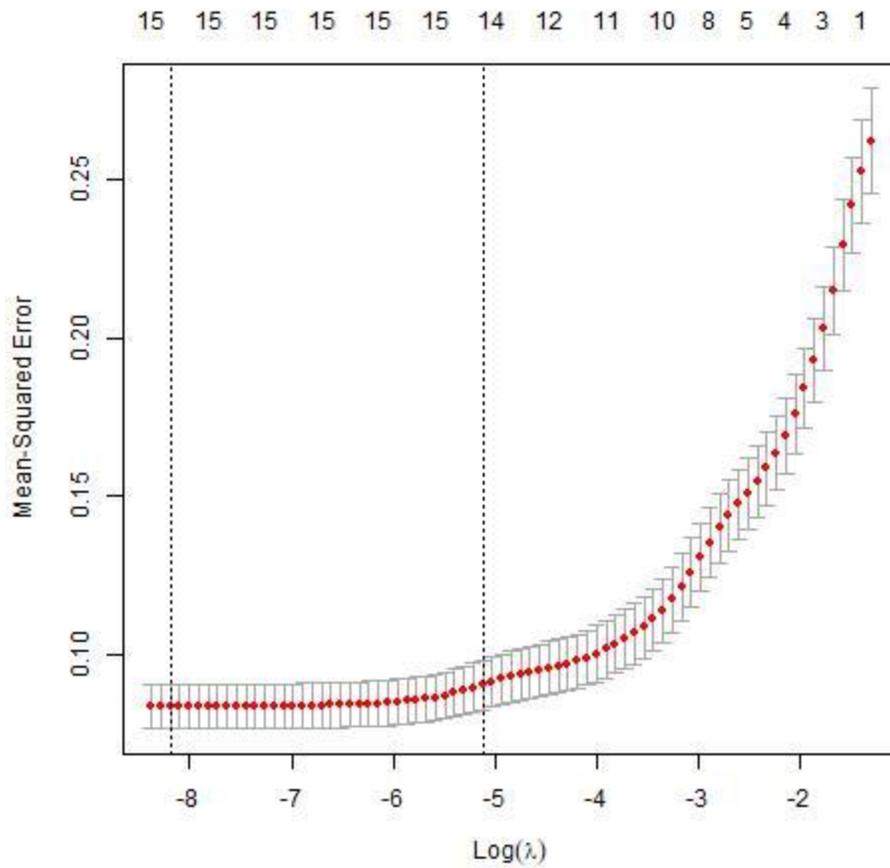


Figure A5- 9. CV error that results from applying LASSO with a series of values of λ for model ID 8.

The primary x-axis (bottom) is λ in natural log scale while the secondary x-axis (top) is the number of features selected at a given λ value. The first dashed line is the λ value that minimized the CV MSE while the second dashed line is the largest λ at which the MSE is within one standard error of the smallest MSE. The first value was taken for LASSO feature selection.

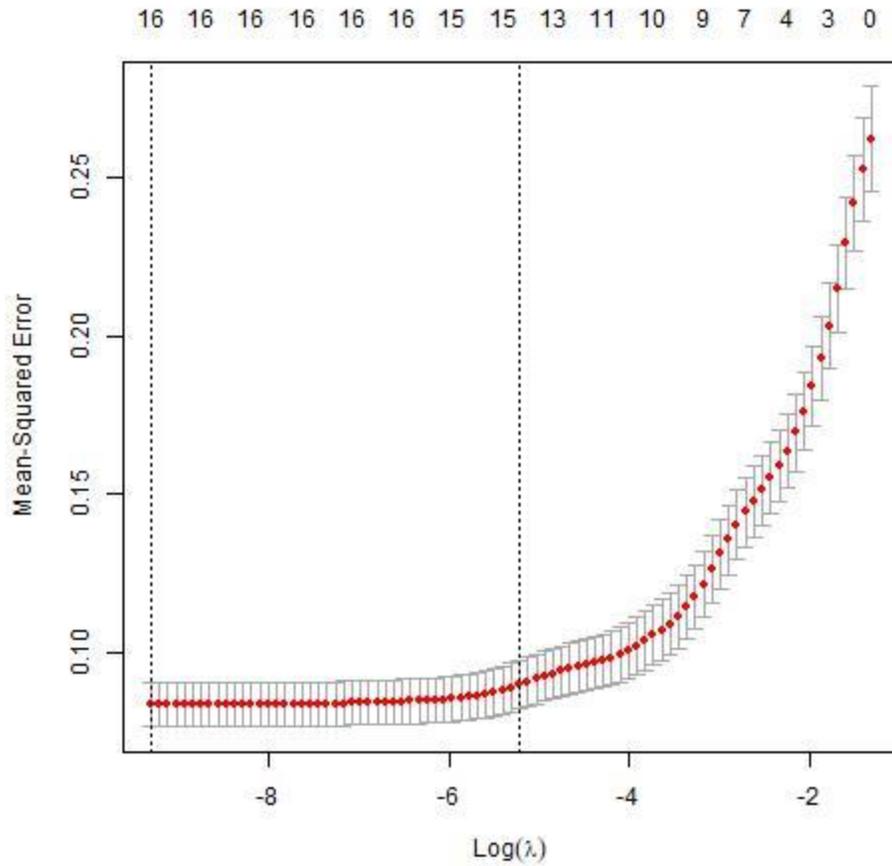


Figure A5- 10. CV error that results from applying LASSO with a series of values of λ for model ID 9.

The primary x-axis (bottom) is λ in natural log scale while the secondary x-axis (top) is the number of features selected at a given λ value. The first dashed line is the λ value that minimized the CV MSE while the second dashed line is the largest λ at which the MSE is within one standard error of the smallest MSE. The first value was taken for LASSO feature selection.

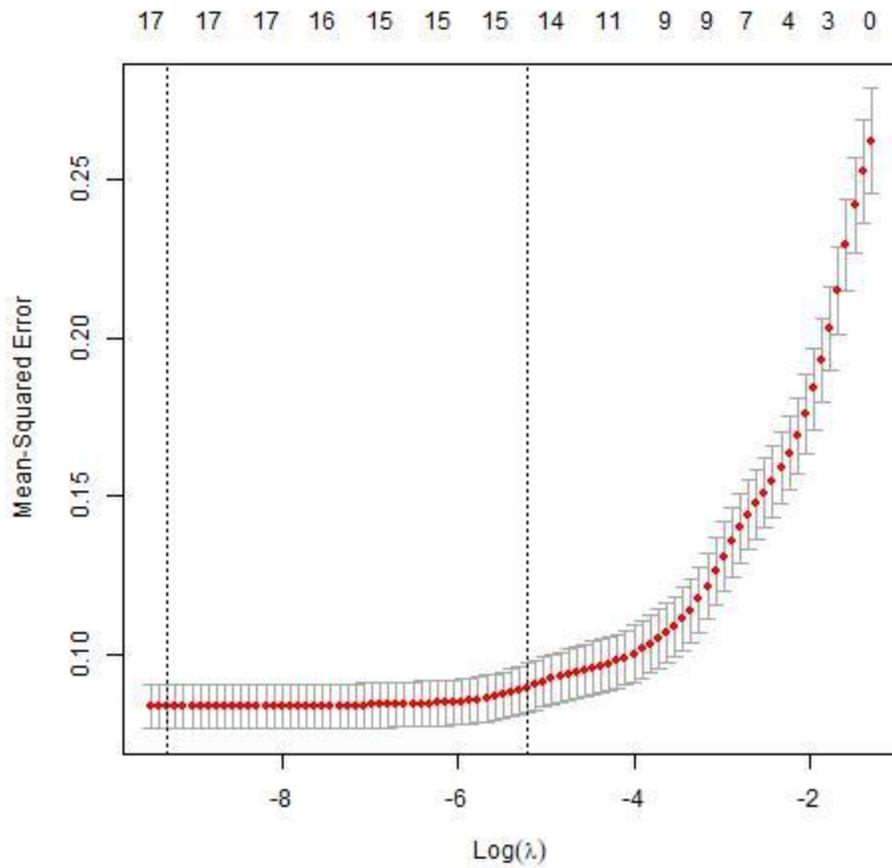


Figure A5- 11. CV error that results from applying LASSO with a series of values of λ for model ID 10.

The primary x-axis (bottom) is λ in natural log scale while the secondary x-axis (top) is the number of features selected at a given λ value. The first dashed line is the λ value that minimized the CV MSE while the second dashed line is the largest λ at which the MSE is within one standard error of the smallest MSE. The first value was taken for LASSO feature selection.

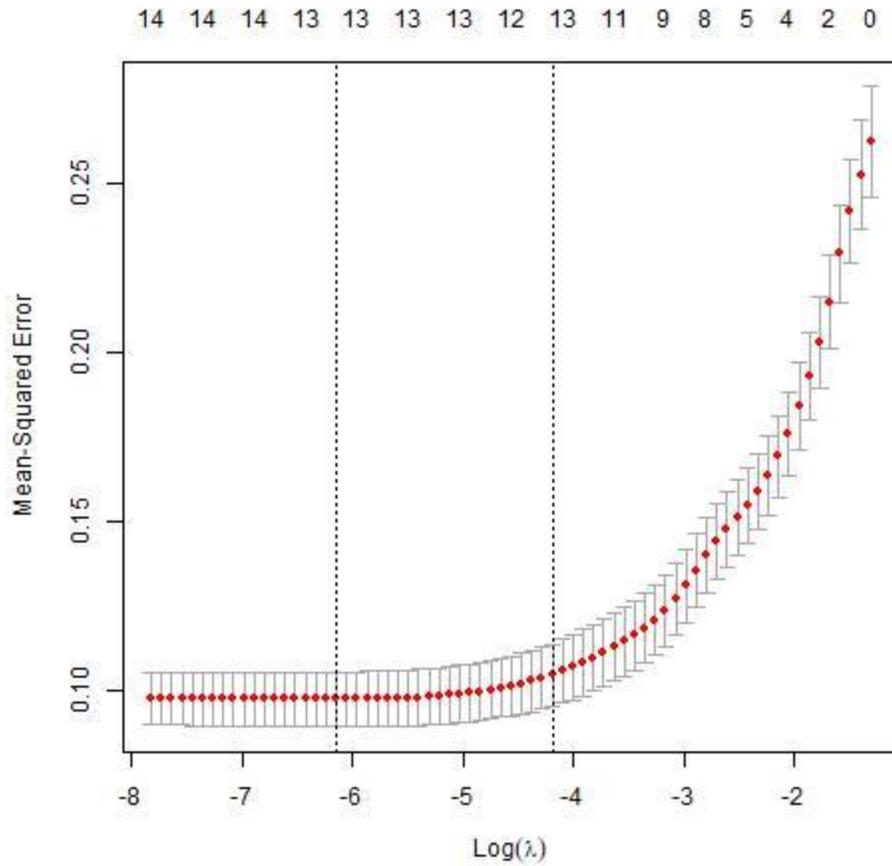


Figure A5- 12. CV error that results from applying LASSO with a series of values of λ for model ID 11.

The primary x-axis (bottom) is λ in natural log scale while the secondary x-axis (top) is the number of features selected at a given λ value. The first dashed line is the λ value that minimized the CV MSE while the second dashed line is the largest λ at which the MSE is within one standard error of the smallest MSE. The first value was taken for LASSO feature selection.

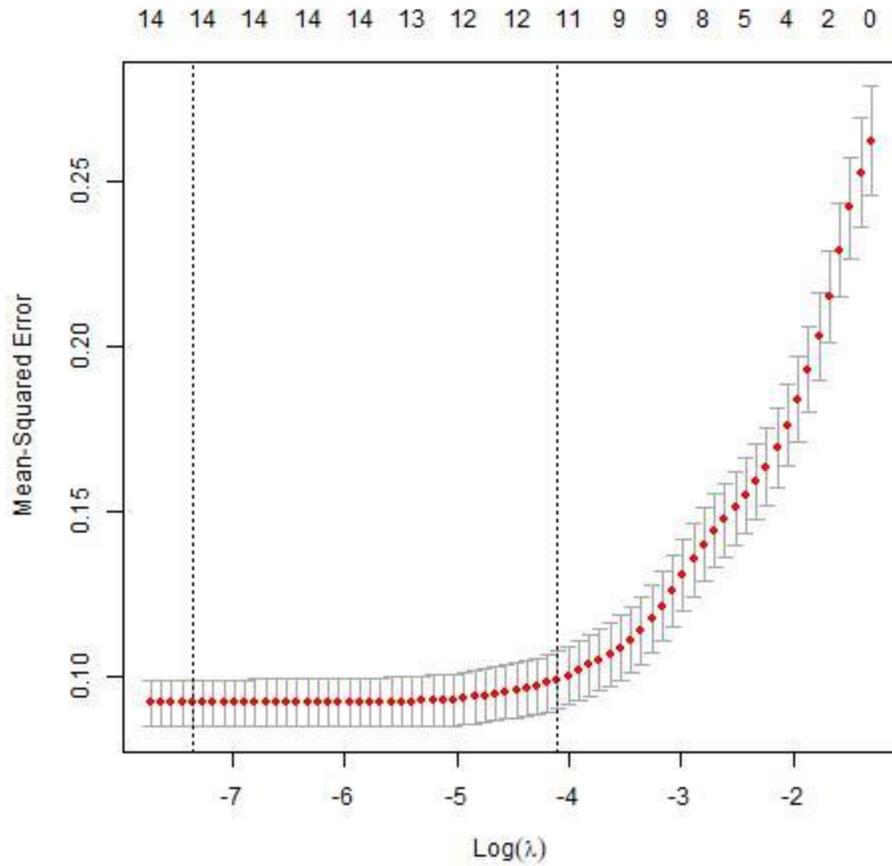


Figure A5- 13. CV error that results from applying LASSO with a series of values of λ for model ID 12.

The primary x-axis (bottom) is λ in natural log scale while the secondary x-axis (top) is the number of features selected at a given λ value. The first dashed line is the λ value that minimized the CV MSE while the second dashed line is the largest λ at which the MSE is within one standard error of the smallest MSE. The first value was taken for LASSO feature selection.

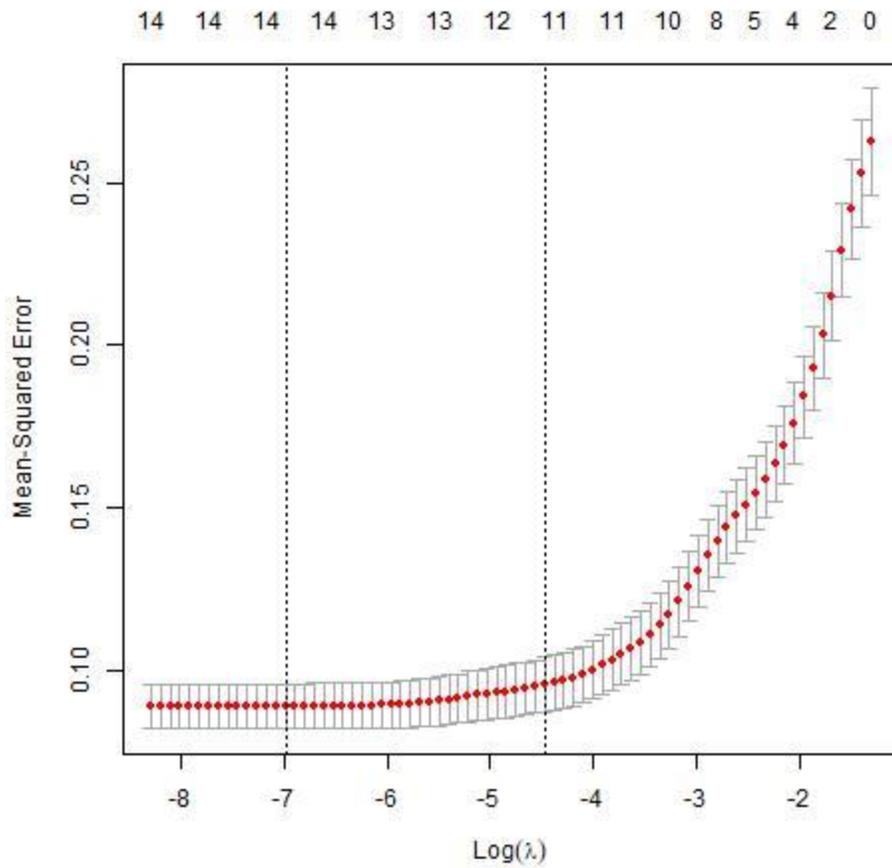


Figure A5- 14. CV error that results from applying LASSO with a series of values of λ for model ID 13.

The primary x-axis (bottom) is λ in natural log scale while the secondary x-axis (top) is the number of features selected at a given λ value. The first dashed line is the λ value that minimized the CV MSE while the second dashed line is the largest λ at which the MSE is within one standard error of the smallest MSE. The first value was taken for LASSO feature selection.

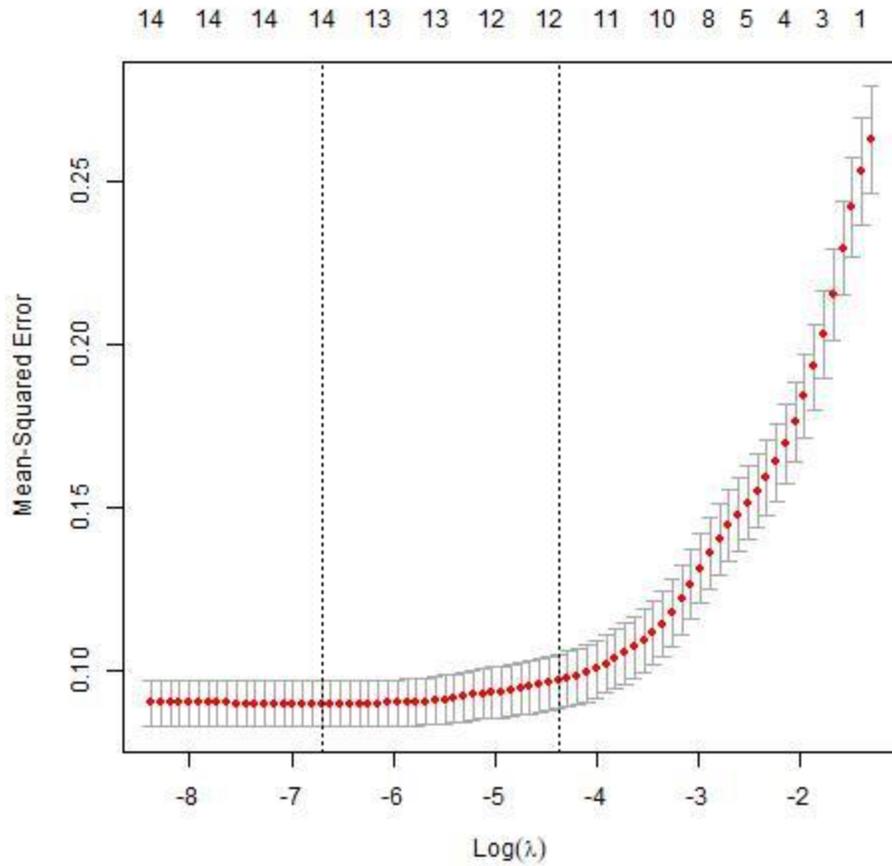


Figure A5- 15. CV error that results from applying LASSO with a series of values of λ for model ID 14.

The primary x-axis (bottom) is λ in natural log scale while the secondary x-axis (top) is the number of features selected at a given λ value. The first dashed line is the λ value that minimized the CV MSE while the second dashed line is the largest λ at which the MSE is within one standard error of the smallest MSE. The first value was taken for LASSO feature selection.

Table A5- 1. Summary of input variable combinations prior to feeding into LASSO, output combinations from LASSO, and the variables dropped by LASSO.

ID	Input variable combination fed into LASSO	Output variable combination from LASSO	Dropped variable
1	C0 + A +V +pH+mp_ratio+charge+ BET + pzc +CD+PD +EBCT +DOC+B+E+S	C0 + A + V + pH + mp_ratio + charge + BET + pzc + CD + PD + EBCT + DOC + B + E + S	-
2	C0 + A +V +pH+mp_ratio+charge+ BET + pzc +CD+PD +EBCT +DOC+L+E+S	C0 + A + V + pH + mp_ratio + +BET + pzc + CD + +EBCT + DOC + L + E + S	PD
3	C0 + A +V +pH+mp_ratio+charge+ BET + pzc +CD+PD +EBCT +DOC+L+E+S+B	C0 + A + V + pH + mp_ratio + +BET + pzc + CD + +EBCT + DOC + L + E + S	B,PD
4	C0 + A +V +pH+mp_ratio+charge+ BET + pzc +CD+PD +EBCT +DOC+B+E	C0 + A + V + pH + mp_ratio + charge + BET + pzc + CD + +EBCT + DOC + B + E	PD
5	C0 + A +V +pH+mp_ratio+charge+ BET + pzc +CD+PD +EBCT +DOC+L+E	C0 + A + V + pH + mp_ratio + charge + BET + pzc + CD + PD + EBCT + DOC + L + E	-
6	C0 + A +V +pH+mp_ratio+charge+ BET + pzc +CD+PD +EBCT +DOC+L+E+B	C0 + A + V + pH + mp_ratio + charge + BET + pzc + CD + PD + EBCT + DOC + L + E + B	-

Table A5- 1 (Continued)

7	C0 + A +V +pH+mp_ratio+charge+ BET + pzc +CD+PD +EBCT +UV254+B+E+S	C0 + A + V + pH + mp_ratio + charge + BET + pzc + CD + PD + EBCT + UV254 + B + E + S	-
8	C0 + A +V +pH+mp_ratio+charge+ BET + pzc +CD+PD +EBCT +UV254+L+E+S	C0 + A + V + pH+mp_ratio + charge + BET + pzc + CD + PD + EBCT + UV254 + L + E + S	-
9	C0 + A +V +pH+mp_ratio+charge+ BET + pzc +CD+PD +EBCT +UV254+L+E+S+B	C0 + A + V + pH + mp_ratio + charge + BET + pzc + CD + PD + EBCT + UV254 + L + E + S + B	-
10	C0 + A +V +pH+mp_ratio+charge+ BET + pzc +CD+PD +EBCT +UV254+L+S+B+E+DOC	C0 + A + V + pH + mp_ratio + charge + BET + pzc + CD + PD + EBCT + UV254 + L + S + B + E + DOC	-
11	C0 + A +V +pH+mp_ratio+charge+ BET + pzc +CD+PD +EBCT +UV254+B+E	C0 + A + V + pH + mp_ratio + +BET + pzc + CD + PD + EBCT + UV254 + B + E	charge
12	C0 + A +V +pH+mp_ratio+charge+ BET + pzc +CD+PD +EBCT +UV254+L+E	C0 + A + V + pH + mp_ratio + charge + BET + pzc + CD + PD + EBCT + UV254 + L + E	-
13	C0 + A +V +pH+mp_ratio+charge+ BET + pzc +CD+PD +EBCT +UV254+L+E+B	C0 + A + V + pH + mp_ratio + +BET + pzc + CD + PD + EBCT + UV254 + L + E + B	charge

Table A5- 1 (Continued)

14	C0 + A + V + pH + mp_ratio + charge + BET + pzc + CD + PD + EBCT + UV254 + L + B + E + DOC	C0 + A + V + pH + mp_ratio + +BET + pzc + CD + PD + EBCT + UV254 + L + B + E	charge, DOC
-----------	--------------------------------------------------------------------------------------------------	------------------------------------------------------------------------------------	----------------

After LASSO, we used each of the output variable combinations from Table A 3 to construct 14 models and compared them via 10-fold 10-repeat CV. Note that input variables were standardized in the training process. The 14 models were then ranked by RMSE in ascending order and the Bayesian Information Criterion (BIC) of the top 2 models were assessed and compared. The model with the lowest BIC (Neath and Cavanaugh, 2012), ID 9’s variable subset was then selected to perform MLR with ordinary least squares to fit the training set, which yielded the final MLR model (Table A5 -3). CV and variable pre-processing were performed using the R package “caret” (Kuhn, 2008). A summary of error metrics and lambda values obtained from CV corresponding to each variable ID combination is shown below (ranked in ascending order of RMSE).

Table A5- 2. Summary of tuning results of 10-fold CV λ corresponding to each model ID and 10-fold 10 repeat CV error of the resulting model using the λ value that has been cross-validated.

ID	λ	RMSE	R²	MAE	BIC
9	9.06E-05	0.320	0.676	0.238	211
10	9.06E-05	0.321	0.675	0.238	217
8	2.77E-04	0.326	0.674	0.239	
7	8.25E-05	0.329	0.655	0.249	
13	9.27E-04	0.350	0.655	0.254	
14	1.23E-03	0.350	0.655	0.254	
12	6.39E-04	0.351	0.644	0.260	
2	2.34E-03	0.357	0.570	0.276	
3	2.34E-03	0.357	0.570	0.276	
6	1.94E-03	0.364	0.554	0.282	
5	2.34E-03	0.365	0.552	0.283	
1	1.94E-03	0.366	0.552	0.283	
11	2.14E-03	0.368	0.624	0.270	
4	2.34E-03	0.371	0.543	0.289	

The MLR model development results showed that the model which only had the collinearity between DOC and UV254 removed was the best at prediction (ID 9). The final MLR model's coefficients are summarized in Table A5-3, and results showed that UV254 was selected over DOC although the two collinear variables exhibited similar semi-partial least square values (Table A2).

Table A5- 3. Final MLR model fitted to 385 entries excluding the additional 51 entries missing UV254 data. Coefficient, confidence interval (CI) and p-value (p) of each variable selected for the final MLR model are summarized below.

<i>Predictors</i>	<i>Estimates</i>	<i>CI</i>	<i>p</i>
(Intercept)	4.15	4.12 – 4.18	<0.001
C0	-0.03	-0.06 – -0.00	0.02
A	-0.3	-0.37 – -0.24	<0.001
V	-0.07	-0.17 – 0.02	0.111
pH	0.11	0.07 – 0.15	<0.001
mp_ratio	-0.11	-0.15 – -0.07	<0.001
charge	-0.14	-0.22 – -0.07	<0.001
BET	0.17	0.12 – 0.22	<0.001
pzc	0.19	0.14 – 0.23	<0.001
CD	-0.13	-0.18 – -0.09	<0.001
PD	-0.07	-0.11 – -0.02	0.01
EBCT	-0.02	-0.06 – 0.01	0.184
UV254	-0.22	-0.26 – -0.18	<0.001
L	0.38	0.24 – 0.53	<0.001
E	0.51	0.30 – 0.72	<0.001
S	-0.64	-0.87 – -0.40	<0.001
B	0.21	-0.01 – 0.43	0.059
R² / R² adjusted	0.709 / 0.696		

Finally, a few diagnostic plots of the MLR model are presented in panel a through c (Figure A5-16) to verify the linear assumption made prior to performing linear regression. The fitted values in a $\sim c$ refers to the predicted values of the training set. In a, the assumption that the residuals are randomly distributed was confirmed. In b, the assumption that the standardized residuals fall into the theoretical quantiles was verified. In c, since no points were outside of Cook's distance, there were no influential outliers needed to be taken out of the training set.

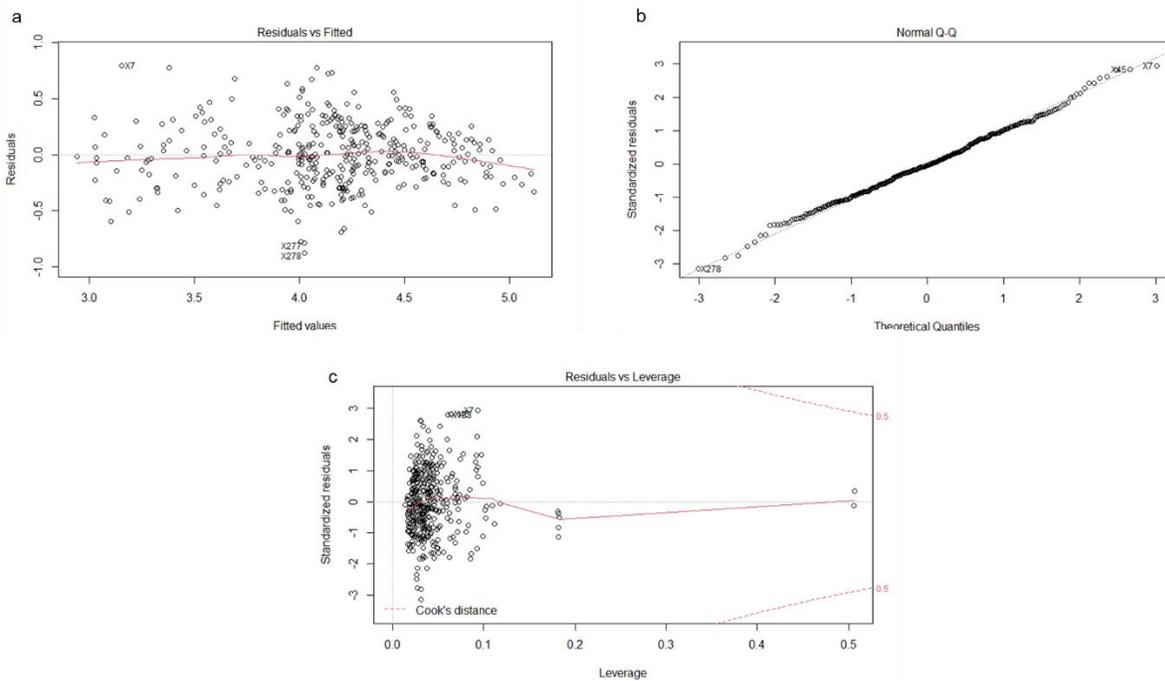


Figure A5- 16. Diagnostic plots of MLR model in panels a through c. In a, residuals were plotted against fitted BV10 values. In b, observed standardized residuals were plotted against theoretical quantiles of residuals. In c, standardized residuals plotted against leverage.

Text A6 Ensemble tree models development

The ensemble tree models were tuned first using all candidate variables and later we verified the relative importance of all variables.

The tuning procedure of ensemble tree models closely followed “*Hands-On Machine Learning with R*” (Boehmke and Greenwell, 2019). Out-of-bag (OOB) RMSE was minimized when choosing hyperparameters for RF; meanwhile, RMSE during CV was minimized when choosing hyperparameters for GBM. The hyperparameters in RF, `mtry` (number of random variables considered at split, 3~17), `min.node.size` (minimum number of observations at terminal node), and `sample.fraction` (size of the bootstrapped dataset in ratio to the full training dataset), were tuned concurrently while `n.trees` (number of trees grown) was set to constant. The hyperparameters in GBM, `shrinkage`, `interaction.depth`, `n.minobsinnode`, `bag.fraction`, and `n.trees`, were tuned sequentially. Shrinkage (learning rate at each iteration) was tuned first, then concurrently, `interaction.depth` (the depth of each tree), `n.minobsinnode` (minimum number of observations at the terminal node), and `bag.fraction` (subsampling rate) were tuned; finally, `n.trees` (number of iterations) was tuned.

For tuning the RF model, we set `replace` to `FALSE` and `n.trees` to p (number of input variables) $\times 10$, where $p = 17$, while the following parameters were tuned concurrently: `mtry`, `min.node.size`, and `sample.fraction`. Note that p denotes the number of candidate features included in the model.

Grid search, which is a common approach for finding optimal hyperparameter combinations, was performed when tuning the hyperparameters for both RF and GBM models. During a grid search, different hyperparameters of a ML algorithm are combined

to construct different models. For tuning the RF model, grid search between values: `mtry` = 3, 4, 5, 6, 7, 17, `min.node.size` = 1, 3, 5, 10, `sample.fraction` = 0.6, 0.7, 0.8 was performed. A total of 72 grid points were assessed using R package, “`ranger`” (Wright & Ziegler, 2017), and results in ascending RMSE order are shown here:

Table A6- 1. Tuning results of 72 grid points for RF model.

rank	mtry	min.node.size	sample.fraction	rmse
1	4	1	0.8	0.161627
2	5	1	0.8	0.164307
3	4	3	0.8	0.165038
4	3	1	0.8	0.165284
5	5	3	0.8	0.16539
6	5	1	0.7	0.16539
7	7	3	0.7	0.165445
8	7	1	0.7	0.165683
9	4	1	0.7	0.166361
10	6	1	0.8	0.1664
11	6	3	0.8	0.167047
12	7	1	0.8	0.167304
13	3	3	0.8	0.167591
14	3	1	0.7	0.167763
15	7	1	0.6	0.1679
16	5	5	0.8	0.16805

17	6	1	0.7	0.168114
-----------	---	---	-----	----------

Table A6- 1 (Continued)

18	4	1	0.6	0.168367
-----------	---	---	-----	----------

19	7	3	0.8	0.168506
-----------	---	---	-----	----------

20	4	3	0.7	0.169158
-----------	---	---	-----	----------

21	5	3	0.7	0.169469
-----------	---	---	-----	----------

22	3	3	0.7	0.169565
-----------	---	---	-----	----------

23	5	1	0.6	0.16959
-----------	---	---	-----	---------

24	4	3	0.6	0.169848
-----------	---	---	-----	----------

25	4	5	0.8	0.169982
-----------	---	---	-----	----------

26	3	1	0.6	0.17118
-----------	---	---	-----	---------

27	6	3	0.7	0.171303
-----------	---	---	-----	----------

28	6	5	0.8	0.171383
-----------	---	---	-----	----------

29	7	3	0.6	0.171502
-----------	---	---	-----	----------

30	5	3	0.6	0.171588
-----------	---	---	-----	----------

31	5	5	0.7	0.171867
-----------	---	---	-----	----------

32	4	5	0.7	0.172685
-----------	---	---	-----	----------

33	6	3	0.6	0.172714
-----------	---	---	-----	----------

34	4	5	0.6	0.172842
-----------	---	---	-----	----------

35	7	5	0.7	0.173514
-----------	---	---	-----	----------

36	3	3	0.6	0.173725
-----------	---	---	-----	----------

37	6	1	0.6	0.174031
-----------	---	---	-----	----------

38	6	5	0.7	0.174056
-----------	---	---	-----	----------

39	3	5	0.7	0.174168
-----------	---	---	-----	----------

Table A6- 1 (Continued)

40	7	5	0.8	0.174173
-----------	---	---	-----	----------

41	7	5	0.6	0.174197
-----------	---	---	-----	----------

42	3	5	0.8	0.174459
-----------	---	---	-----	----------

43	5	5	0.6	0.175374
-----------	---	---	-----	----------

44	17	1	0.6	0.176538
-----------	----	---	-----	----------

45	5	10	0.8	0.176995
-----------	---	----	-----	----------

46	6	5	0.6	0.177374
-----------	---	---	-----	----------

47	17	3	0.6	0.177523
-----------	----	---	-----	----------

48	17	1	0.7	0.177969
-----------	----	---	-----	----------

49	17	1	0.8	0.178707
-----------	----	---	-----	----------

50	17	3	0.7	0.178738
-----------	----	---	-----	----------

51	7	10	0.8	0.178877
-----------	---	----	-----	----------

52	4	10	0.8	0.179202
-----------	---	----	-----	----------

53	6	10	0.8	0.179653
-----------	---	----	-----	----------

54	17	3	0.8	0.179787
-----------	----	---	-----	----------

55	7	10	0.7	0.180412
-----------	---	----	-----	----------

56	3	5	0.6	0.180458
-----------	---	---	-----	----------

57	5	10	0.7	0.181407
-----------	---	----	-----	----------

58	3	10	0.8	0.181954
-----------	---	----	-----	----------

59	17	5	0.6	0.181962
-----------	----	---	-----	----------

60	6	10	0.7	0.182216
-----------	---	----	-----	----------

61	17	5	0.7	0.182226
Table A6- 1 (Continued)				
62	4	10	0.7	0.183245
63	7	10	0.6	0.183697
64	17	5	0.8	0.184073
65	6	10	0.6	0.184639
66	4	10	0.6	0.185393
67	3	10	0.7	0.186645
68	5	10	0.6	0.187608
69	17	10	0.7	0.190839
70	3	10	0.6	0.191453
71	17	10	0.8	0.192048
72	17	10	0.6	0.192392

For tuning the GBM model, we used R package “gbm” (Ridgeway, 2020). The following were set to constant while tuning shrinkage: n.trees = 5000, interaction.depth = 3, n.minobsinnode = 10, cv.folds = 10, train.fraction=1 (default by the gbm package), bag.fraction=0.5 (default by gbm package). The following values were tried for

shrinkage: 0.05,0.06,0.07,0.08,0.09, 0.1,0.3, and results are summarized in ascending RMSE order in Table A6-2.

Table A6- 2. Tuning results of shrinkage for GBM model.

rank	learning_rate	RMSE	trees
1	0.09	0.144543	2125
2	0.08	0.145104	3805
3	0.05	0.145173	2894
4	0.1	0.145274	2162
5	0.06	0.145525	2894
6	0.07	0.146233	2770
7	0.3	0.149967	550

After the optimum shrinkage was found to be 0.075, a grid search of the following values were performed: interaction.depth = 4,5,6, n.minobsinnode = 3,4,5,6,7,8,9,10, bag.fraction = 0.5,0.6,0.7,0.8. During this grid search, n.tree was set to 3000 since the optimum number of trees were 2125 for the model with shrinkage = 0.075 (Table A 7). A total of 128 grid points were assessed, and results are shown in ascending RMSE order here:

Table A6- 3. Tuning results of 128 grid points for GBM model.

rank	n.trees	shrinkage	interaction.depth	n.minobsinnode	bag.fraction	rmse
1	3000	0.09	6	10	0.5	0.141533
2	3000	0.09	5	10	0.5	0.141868
3	3000	0.09	6	7	0.5	0.142127
4	3000	0.09	6	6	0.5	0.1425

Table A6- 3 (Continued)

5	3000	0.09	6	9	0.5	0.143779
6	3000	0.09	4	9	0.7	0.143859
7	3000	0.09	4	6	0.6	0.143957
8	3000	0.09	4	10	0.5	0.144393
9	3000	0.09	4	7	0.7	0.144434
10	3000	0.09	3	10	0.5	0.144543
11	3000	0.09	3	9	0.6	0.144708
12	3000	0.09	6	8	0.5	0.144858
13	3000	0.09	5	9	0.5	0.145007
14	3000	0.09	5	6	0.6	0.145045
15	3000	0.09	5	8	0.6	0.145166
16	3000	0.09	6	10	0.6	0.145239
17	3000	0.09	5	8	0.7	0.145244
18	3000	0.09	4	7	0.6	0.145393
19	3000	0.09	3	8	0.6	0.145421
20	3000	0.09	3	10	0.6	0.145441
21	3000	0.09	5	7	0.5	0.14545

22	3000	0.09	4	8	0.7	0.145561
23	3000	0.09	5	10	0.6	0.145584
24	3000	0.09	4	9	0.6	0.145677
25	3000	0.09	4	6	0.5	0.145687
26	3000	0.09	3	10	0.8	0.145743

Table A6- 3 (Continued)

27	3000	0.09	3	3	0.5	0.145757
28	3000	0.09	5	8	0.5	0.145761
29	3000	0.09	3	5	0.6	0.145783
30	3000	0.09	5	3	0.6	0.145849
31	3000	0.09	6	5	0.5	0.145942
32	3000	0.09	3	6	0.5	0.145944
33	3000	0.09	6	4	0.5	0.145963
34	3000	0.09	3	3	0.7	0.145977
35	3000	0.09	4	10	0.7	0.146107
36	3000	0.09	4	8	0.6	0.146152
37	3000	0.09	6	3	0.5	0.146168
38	3000	0.09	5	9	0.7	0.14622
39	3000	0.09	5	6	0.5	0.14626
40	3000	0.09	6	7	0.7	0.146291
41	3000	0.09	5	7	0.7	0.146327
42	3000	0.09	6	8	0.6	0.146348
43	3000	0.09	5	4	0.5	0.146394

44	3000	0.09	5	4	0.6	0.146428
45	3000	0.0s9	5	9	0.6	0.146492
46	3000	0.09	3	7	0.5	0.146508
47	3000	0.09	4	7	0.5	0.146541
48	3000	0.09	3	5	0.8	0.146553

Table A6- 3 (Continued)

49	3000	0.09	6	7	0.6	0.146632
50	3000	0.09	3	9	0.7	0.14664
51	3000	0.09	4	8	0.5	0.146653
52	3000	0.09	3	6	0.6	0.146741
53	3000	0.09	6	9	0.6	0.146743
54	3000	0.09	3	8	0.5	0.14677
55	3000	0.09	3	4	0.6	0.146833
56	3000	0.09	3	10	0.7	0.146896
57	3000	0.09	4	10	0.6	0.146969
58	3000	0.09	3	7	0.6	0.14698
59	3000	0.09	4	5	0.6	0.146993
60	3000	0.09	4	3	0.6	0.147019
61	3000	0.09	6	6	0.6	0.147045
62	3000	0.09	5	10	0.7	0.147055
63	3000	0.09	6	3	0.6	0.147202
64	3000	0.09	3	8	0.7	0.147243
65	3000	0.09	4	9	0.5	0.147248

66	3000	0.09	3	5	0.7	0.147261
67	3000	0.09	3	9	0.5	0.147281
68	3000	0.09	5	3	0.7	0.147328
69	3000	0.09	3	4	0.7	0.147366
70	3000	0.09	6	9	0.7	0.147454

Table A6- 3 (Continued)

71	3000	0.09	3	4	0.5	0.14748
72	3000	0.09	3	5	0.5	0.147481
73	3000	0.09	6	8	0.7	0.147568
74	3000	0.09	5	5	0.6	0.147688
75	3000	0.09	4	5	0.5	0.147689
76	3000	0.09	6	4	0.6	0.147722
77	3000	0.09	3	6	0.8	0.147751
78	3000	0.09	5	7	0.6	0.14792
79	3000	0.09	4	4	0.7	0.14798
80	3000	0.09	6	5	0.6	0.148052
81	3000	0.09	5	4	0.7	0.148106
82	3000	0.09	4	3	0.7	0.148232
83	3000	0.09	6	10	0.8	0.148281
84	3000	0.09	3	7	0.7	0.14835
85	3000	0.09	3	9	0.8	0.148409
86	3000	0.09	6	10	0.7	0.148453
87	3000	0.09	5	3	0.5	0.148526

88	3000	0.09	3	6	0.7	0.148542
89	3000	0.09	4	6	0.7	0.148568
90	3000	0.09	4	3	0.5	0.148602
91	3000	0.09	6	8	0.8	0.148662
92	3000	0.09	4	5	0.7	0.148803

Table A6- 3 (Continued)

93	3000	0.09	5	6	0.7	0.148821
94	3000	0.09	3	7	0.8	0.14883
95	3000	0.09	6	6	0.7	0.148901
96	3000	0.09	5	5	0.8	0.148993
97	3000	0.09	5	8	0.8	0.148998
98	3000	0.09	4	4	0.5	0.149067
99	3000	0.09	4	7	0.8	0.149138
100	3000	0.09	3	4	0.8	0.149181
101	3000	0.09	5	5	0.7	0.149251
102	3000	0.09	4	10	0.8	0.149372
103	3000	0.09	4	5	0.8	0.149548
104	3000	0.09	5	9	0.8	0.149619
105	3000	0.09	4	8	0.8	0.149749
106	3000	0.09	4	4	0.6	0.149876
107	3000	0.09	5	6	0.8	0.149893
108	3000	0.09	5	5	0.5	0.149909
109	3000	0.09	3	8	0.8	0.149909

110	3000	0.09	3	3	0.8	0.150145
111	3000	0.09	5	10	0.8	0.150289
112	3000	0.09	6	7	0.8	0.150323
113	3000	0.09	4	6	0.8	0.150422
114	3000	0.09	6	6	0.8	0.150462
Table A6- 3 (Continued)						
115	3000	0.09	3	3	0.6	0.150529
116	3000	0.09	6	5	0.7	0.150548
117	3000	0.09	4	9	0.8	0.150715
118	3000	0.09	4	3	0.8	0.150727
119	3000	0.09	5	7	0.8	0.150993
120	3000	0.09	6	9	0.8	0.151129
121	3000	0.09	6	3	0.7	0.15126
122	3000	0.09	6	5	0.8	0.151509
123	3000	0.09	5	3	0.8	0.152276
124	3000	0.09	6	4	0.7	0.152715
125	3000	0.09	6	4	0.8	0.15319
126	3000	0.09	5	4	0.8	0.153254
127	3000	0.09	6	3	0.8	0.153373
128	3000	0.09	4	4	0.8	0.15341

Lastly, we will use 10-fold CV to find the number of trees, n_{tree} , that will minimize the CV mean-squared-error. Results for this step are shown in Fig S 19 and n_{tree} was found to be

613 (Figure A6-1).

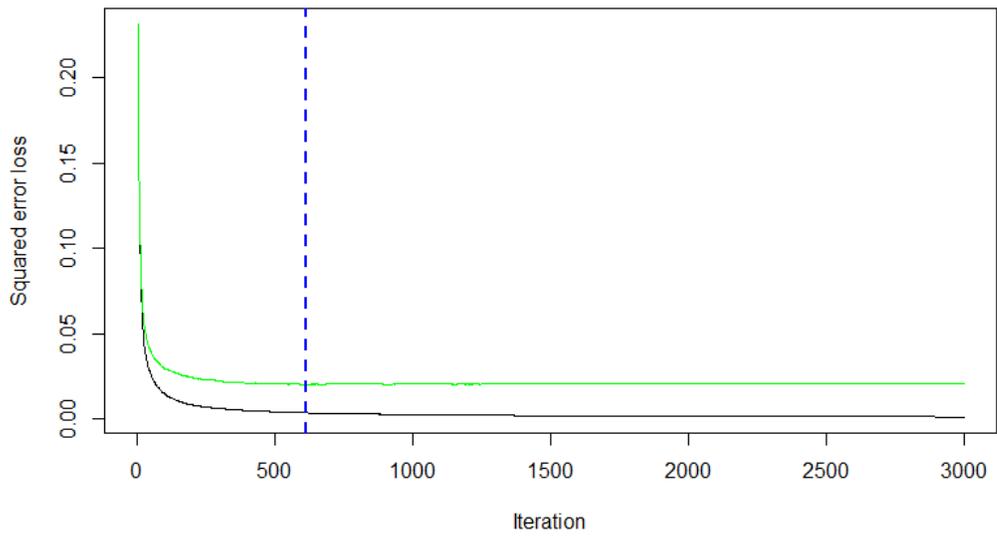


Figure A6- 1 Summary of training error (black line) and validation error (green line) over varying number of iterations (n.tree).

Finally, we verified the usefulness of all candidate variables in the ensemble tree models via the computation of permutation variable importance score (computation

details in Text A7). Results (Figure A6 -2) indicate that since none of the variable had scores of zero, the use of all variables was justified.

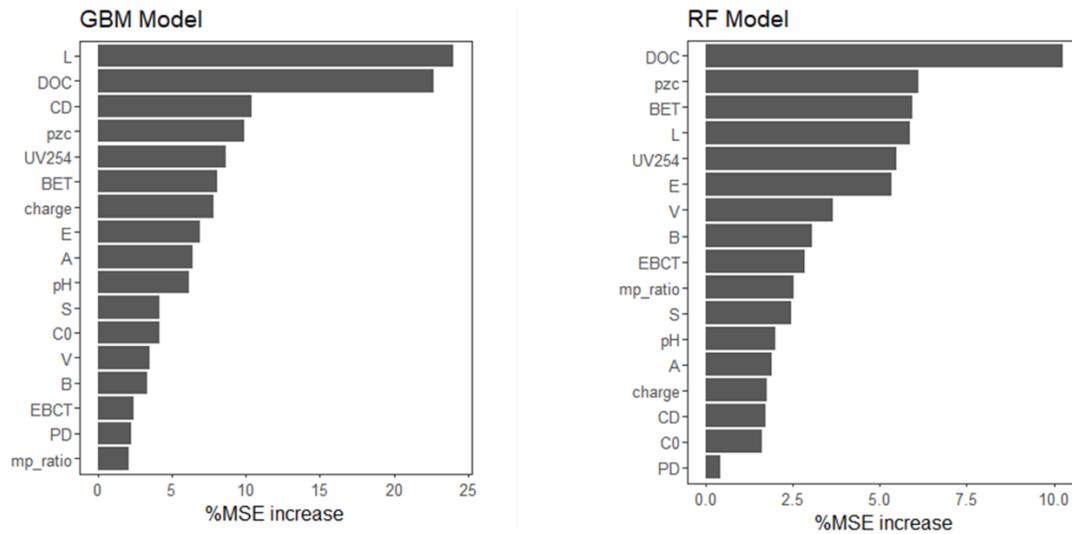


Figure A6- 2 Permutation importance ranking of variables calculated for the GBM (left) and RF (right) models.

Text A7 Permutation variable importance

The permutation importance accounts for the percent increase in mean squared error (MSE) when the data for a specific variable in the training dataset was shuffled, from the MSE calculated from fitting the model to the unshuffled dataset. Consider a dataset in a table format of the following:

Name.of.	compound	charge	C0.ng.L.	C0	E	S	A	B	V	L
PFOA	PFASs	-1	21.6	0.0216	-0.88	-0.39	0.46	0.33	1.5757	3.297
PFOA	PFASs	-1	21.6	0.0216	-0.88	-0.39	0.46	0.33	1.5757	3.297
PFOA	PFASs	-1	21.6	0.0216	-0.88	-0.39	0.46	0.33	1.5757	3.297
PFOA	PFASs	-1	21.6	0.0216	-0.88	-0.39	0.46	0.33	1.5757	3.297
PFOA	PFASs	-1	21.6	0.0216	-0.88	-0.39	0.46	0.33	1.5757	3.297
PFOA	PFASs	-1	21.6	0.0216	-0.88	-0.39	0.46	0.33	1.5757	3.297
PFOA	PFASs	-1	21.6	0.0216	-0.88	-0.39	0.46	0.33	1.5757	3.297
PFOA	PFASs	-1	21.6	0.0216	-0.88	-0.39	0.46	0.33	1.5757	3.297
PFOA	PFASs	-1	21.6	0.0216	-0.88	-0.39	0.46	0.33	1.5757	3.297
PFDA	PFASs	-1	21.6	0.0216	-1.19	-0.64	0.46	0.33	1.9283	4.14
PFNA	PFASs	-1	21.6	0.0216	-1.03	-0.51	0.46	0.33	1.752	3.719

Shuffle this column

Figure A7-1 Schematic of training data in tabular form.

To evaluate the permutation importance of the variable C0 in this example, we would first calculate the MSE of the model fitted to the unshuffled data set. Subsequently, we calculate the MSE of the model fit to the data set once the column containing the variable of interest (indicated by blue arrow) is shuffled. We then calculate the percent change in MSE for the latter from the former scenario. We repeated this procedure for all 17 variables to obtain the percent change in MSE and ranked the variables with ascending order.

Permutation importance is selected as the variable importance metric over another popular metric, the impurity importance, as it is a robust measure when dealing with variables of differing scales (Grömping, 2009), and when collinear predictors are present (Nicodemus et al., 2010), which is the case for this study (collinearity amongst predictors is illustrated in Fig S1).

Text A8 Partial dependence plots (PDPs) and centered individual contribution expectations (ICEs)

PDP (Friedman, 2001) plots the change in the average predicted value of the response variable as the specified input variable vary over its distribution within the training data set (Goldstein, 2015). ICE disaggregates the PDPs—for each of N observations in the training data set, ICE plots the change in predicted value of the response variable as the specified input variable vary over its distribution, while holding all other conditions fixed, generating N curves as a result. We then used a centering procedure (Goldstein, 2015) to change the scale of ICEs, to illustrate the heterogeneity or

the lack thereof of within the simulated changes in the response variable—the altered ICEs are termed c-ICEs (Goldstein, 2015).

To illustrate the construction of c-ICEs, we used L as an example input variable. For L, both the original scale and the centered scale ICEs are plotted in Fig A8- 1. a~b. In Figure A8- 1 panel a, simulated logBV10 at each of the 42 observed distinctive L values, were plotted for a training data set of N = 385, generating a total of 385 curves as a result. PDP, which is the average of these 385 curves, was plotted in Figure A8- 1 panel a as well.

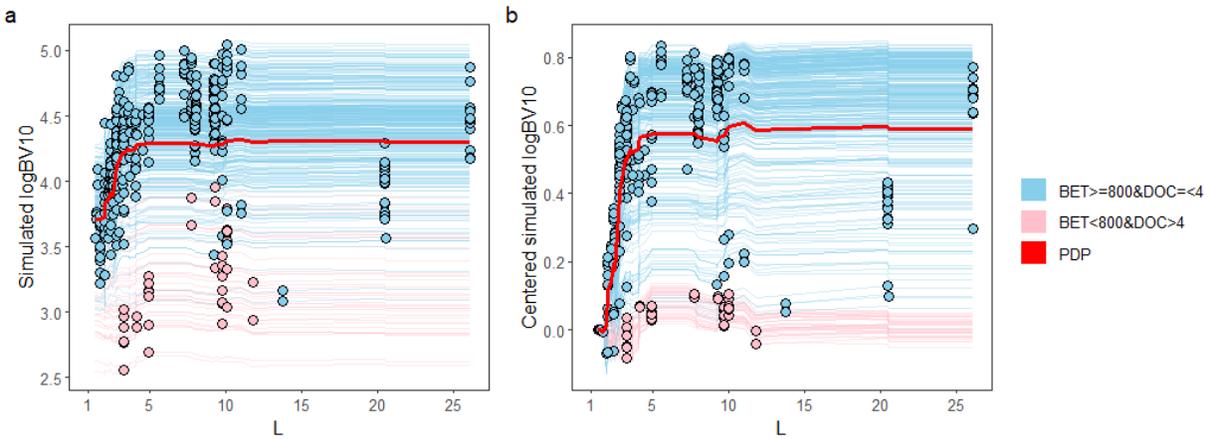


Figure A8- 1 ICEs PDP by L on simulated logBV10 in original scale (a) and centered scale (b). Blue dots/lines represent simulated logBV10/ICEs of data entries with $BET \geq 800$ and $DOC \leq 4$; while pink dots/lines represent those of the complimentary data set. Red lines represent the PDPs of L.

In Figure A8- 1 panel b, c-ICE plot is shown, and the construction procedure is as follows, supported by Goldstein et al (2015)’s work. For $i = 1, 2, \dots, 385$, first, we chose the first point of each curve $\hat{f}^{(i)}$ in the ICE plot as $\hat{f}(x^*, x_{ci})$, the “base case”, for each of the 385 curves. For each curve $\hat{f}^{(i)}$ in the ICE plot, the corresponding c-ICE curve is given by:

$$\hat{f}_{cent}^{(i)} = \hat{f}^{(i)} - \hat{f}(x^*, x_{ci}) \quad (S4)$$

where $\hat{f}_{cent}^{(1)} = 0$, $x^* = 1.446$, and x_{Ci} is a vector of all other input variables besides L whose values are fixed. The PDP in the Figure A8- 1 panel b is the average change in centered scale logBV10. Similarly, c-ICEs for other input variables can be constructed in this manner. Results of UV254 and EBCT are shown in Figure A8- 2 and Figure A8- 3.

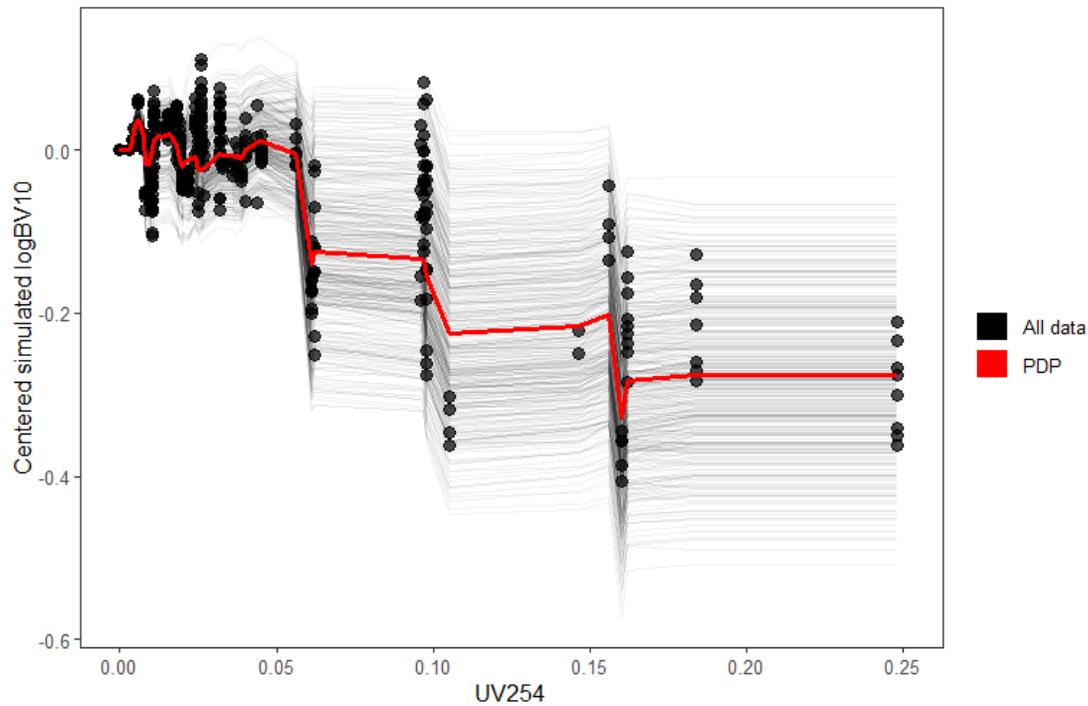


Figure A8- 2 Individual contribution expectations (ICEs) and partial dependence plot (PDP) by UV254 for the GBM model.

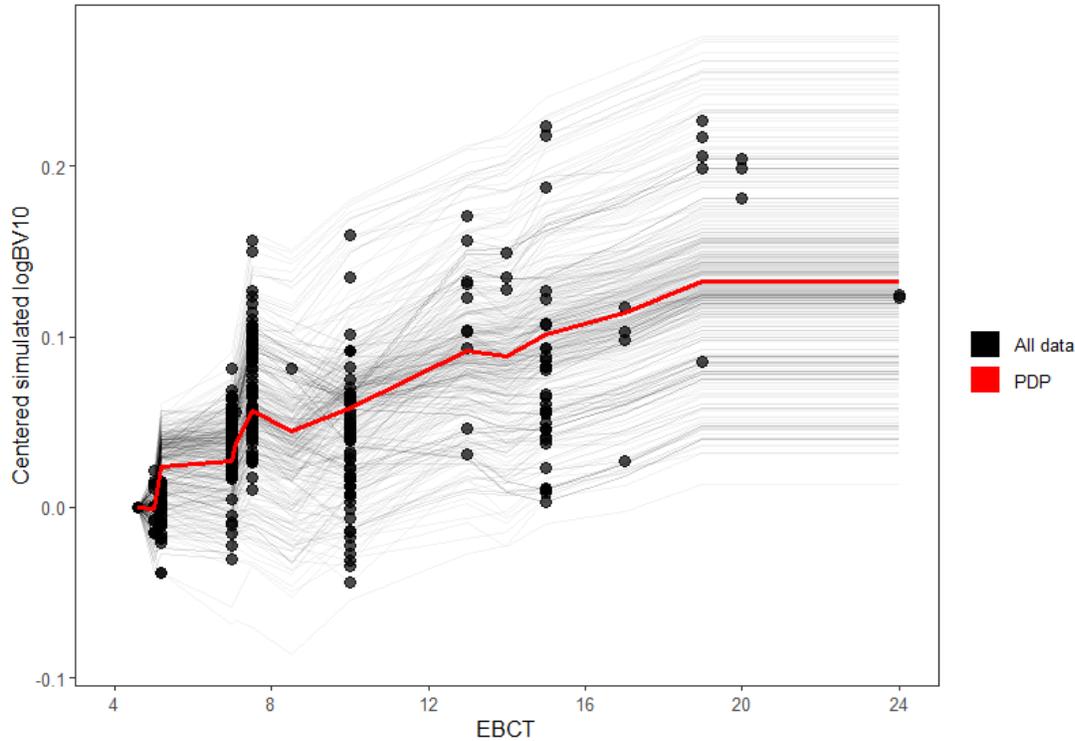


Figure A8- 3 Individual contribution expectations (ICEs) and partial dependence plot (PDP) by UV254 for the GBM model.

Text A9 Simulation scenarios for sensitivity analysis

A list of descriptions for the 14 simulations follows:

1. PFOA removal with baseline conditions (Liu et al, 2019).
2. PFBA removal with baseline conditions.
3. PFPeA removal with baseline conditions.
4. PFHxA removal with baseline conditions.
5. PFHpA removal with baseline conditions.
6. PFNA removal with baseline conditions.
7. PFDA removal with baseline conditions.
8. DOC was set to the minimum value observed in the dataset.
9. DOC was set to the maximum value observed in the dataset.
10. pzc was set to the minimum value observed in the dataset.
11. pzc was set to the maximum value observed in the dataset.
12. BET was set to the minimum value observed in the dataset.
13. BET was set to the maximum value observed in the dataset.
14. BET was set to maximum value while DOC was set to minimum value observed in the dataset.

Text A10 MLR model's prediction accuracy by BV10 range

The testing RMSE of the data in the range of $BV10 > 20,000$ was calculated to be 0.14 whereas for the data range of $BV10 \leq 20,000$, RMSE was 0.30 with the MLR model.

Text A11 Variable selection between DOC and UV254

The comparison of the ensemble tree models to the MLR model helped elucidate how the selection between two collinear water quality parameters, DOC and UV254, affect BV10 prediction. Interestingly, while UV254 was chosen over DOC in the MLR's final model, GBM and RF models ranked (See Figure 3 in main text and Figure A-2 for permutation importance rankings) DOC as the more predictive variable between the two while incorporating both. During the MLR model's feature selection, only one variable, between DOC and UV254, was selected in its final model, as the two variables are highly co-linear (Table A2) and either of them alone explained a large amount of variance (Table A2). This limitation in variable selection of MLR was posed by the assumption that all predictors in the MLR model are independent. Unlike the MLR model, ensemble tree models do not make assumptions about predictor variables and thus can utilize both collinear predictors to make accurate predictions at a level of, on average $R^2 > 0.88$, against unforeseen data.

Text A12 Effect of pore size distribution

Although there have been numerous studies on isotherms regarding how micropore volume relates to adsorption extent of MPs (Pelekani and Snoeyink, 2001, 2000; Quinlivan et al., 2005; Zhi and Liu, 2016), no systematic evaluation of the effect of micropore proportion on GAC bed life has been conducted. Thus, in our study, we

analyzed logBV10's partial dependency on mp_ratio, micropore volume/ total pore volume.

The c-ICEs of mp_ratio associated with lower SUVA values, $SUVA \leq 1.5$ L/mg-m, had more decrease in logBV10 than ones associated with higher SUVA values, $SUVA > 1.5$ L/mg-m (Figure A5). Lower SUVA value corresponds to more hydrophilic nature of the DOM (Karanfil et al., 2002), thus the observation indicates that more hydrophilic DOM with $SUVA \leq 1.5$ L/mg-m negatively impacts BV10 prediction.

Interestingly, in previous study on activated carbon isotherms of pseudo-single solute experiments with co-adsorbing DOM (Quinlivan et al., 2005), positive correlation between adsorption extent and micropore volume ratio was found. For Quinlivan et al (2005)'s data, we conducted a regression analysis to evaluate the significance of the positive correlation as this analysis was not provided in their original publication (Table A12-1). In this data set, we found that micropore volume ratio had a statistically significant positive correlation to q^{e10} mg/g, the equilibrium solid-phase concentration at an equilibrium liquid-phase concentration of 10 μ g/L, for both trichloroethene TCE and methyl tertiary-butyl ether (MTBE). In single-solute isotherm experiments, such relationship between micropore proportion and adsorption extent of MP was observed as well. (Quinlivan et al., 2005; Zhi and Liu, 2016). In fact, in Quinlivan et al (2005)'s data, we observed a more pronounced positive correlation (Table A12-1) between micropore proportion and adsorption extent in the single-solute experiments. These observations along with mp_ratio's c-ICEs indicate that the presence of DOM complicates the adsorption process and thus makes the relationship between GAC bed life and micropore proportion nonlinear.

Table A12- 1. Carbon characteristics and isotherm results of three activated carbon fiber products (ACF-10, ACF-15, ACF-20, Nippon Kynol, Inc., Osaka, Japan) which were chemically modified by acid-washing (AW), H₂O₂ oxidation following acid-washing (OAW), heat treatment in a hydrogen atmosphere following acid-washing (HAW), and heat treatment in an ammonia atmosphere following acid-washing (AAW) in addition to those of three commercially available GACs (F600, G-219, Picazine) are summarized (Quinlivan et al.,2005). The Pearson coefficient and the significance of the correlation between micropore volume ratio and q^{e10} is summarized at the end of the table.

Carbon	BET surface area (m ² /g)	Micro- pore volume (cm ³ /g)	Meso- pore volume (cm ³ /g)	Micro- pore volume/ +meso- pore volume)	<i>TCE</i>	<i>MTBE</i>	<i>TCE</i>	<i>MTBE</i>
					q^{e10} (mg/g)	q^{e10} (mg/g)	Single- solute q^e ¹⁰ (mg/g)	Single- solute q^{e10} (mg/g)
AW10	760	0.387	0.075	0.838	7.76	0.11	16	0.355
OAW10	750	0.34	0.089	0.793	3.01	0.107	4.77	0.175
HAW10	830	0.414	0.074	0.848	10.8	0.405	18	0.648
AAW10	720	0.371	0.061	0.859	7.09	0.198	14.1	0.597
AW15	1480	0.567	0.099	0.851	3.46	0.435	7.53	0.607
OAW15	1350	0.439	0.154	0.74	1.27	0.11	2.45	0.16
HAW15	1530	0.593	0.085	0.875	5.14	0.592	8.64	0.814
AAW15	1510	0.573	0.098	0.854	3.31	0.45	7.29	0.694
AW20	1670	0.594	0.12	0.832	2.37	0.375	6.12	0.484
OAW20	1530	0.504	0.144	0.778	0.78	0.11	2.67	0.139
HAW20	1710	0.622	0.096	0.866	3.91	0.452	8.2	0.805
AAW20	1700	0.599	0.13	0.822	3.36	0.405	6.66	0.492

Table A12-1 (Continued)

F600	820	0.353	0.085	0.806	6.04	0.286	11.7	0.46
G219	1270	0.438	0.118	0.788	4.09	0.349	8.02	0.439
Picazine	1680	0.377	0.252	0.599	0.32	0.011	0.409	0.0134
Correlation for TCE					Correlation for MTBE			
DOM present	Pearson's coefficient	0.57	p-value	0.026	Pearson's coefficient	0.71	p-value	0.003
Single solute	Pearson's coefficient	0.63	p-value	0.012	Pearson's coefficient	0.83	p-value	0.00011

Finally, mp_ratio's marginal effect was similar to that of BET surface area, in that instances with low SUVA values had negative impact on logBV10 prediction with increasing mp_ratio or BET values. One explanation for the observed similarity between the marginal effects of mp_ratio and BET surface area is the pore blockage of micropores by hydrophilic DOM. Earlier studies (Pelekani & Snoeyink, 2000, 2001; Li et al., 2003) have shown that the presence of mesopore is crucial to alleviating the pore block effect as highly microporous GAC would suffer more from reduction in adsorption extent of MPs caused by DOM. This explains why for instances with lower SUVA values, increasing mp_ratio value had negative effect on logBV10 prediction. The correlation between mp_ratio and BET was found to be positive ($r=0.40$, Table A12-2) and significant ($p\text{-value}<0.05$, Table A12-2), which in turn explains the marginal effect of BET surface area. Higher BET values indicate high microporosity and thus imply increasing vulnerability to pore blockage of micropores by DOM.

Table A12-2. Summary of Pearson's correlation coefficient between mp_ratio and BET, and significance of the correlation.

Statistical measure	Value
<i>r</i> (Pearson's coefficient)	0.40
t-value	8.9
Degree of freedom	425
p-value	< 2.2e-16

Text A13 Positive correlation between L/V and BV10 through GBM model sensitivity analysis

In addition to the pronounced positive effect in the data-dense region of L on BV10, positive effect by V was also observed. Larger V correlates to larger BV10 as compounds with large L typically has large V as well and the values of V and L for the PFASs compounds included in the sensitivity analysis are summarized in Table A13-1.

Table A13- 1. Abraham solvation parameters of PFDA, PFNA, PFOA, PFHpA, PFHxA, PFPeA, and PFBA.

Simulation ID	Compound name	Chain Length	E	S	A	B	V	L
2	PFDA	10	-1.19	-0.64	0.46	0.33	1.93	4.14
3	PFNA	9	-1.03	-0.51	0.46	0.33	1.75	3.72
1	PFOA	8	-0.88	-0.39	0.46	0.33	1.58	3.30
4	PFHpA	7	-0.7	-0.27	0.46	0.33	1.40	2.88
Table A13-1 (Continued)								
5	PFHxA	6	-0.77	-0.15	0.46	0.33	1.22	2.45
6	PFPeA	5	-0.62	-0.02	0.46	0.33	1.05	2.03

7	PFBA	4	-0.47	0.1	0.46	0.33	0.87	1.74
---	------	---	-------	-----	------	------	------	------

Text A14 Final GBM and RF models for web application deployment

Test set data was merged with training set data to tune the hyperparameters of the final GBM and RF models for deployment. These models are deployed as a web application, and link follows:

https://yoko-koyamas-gac-mdls.shinyapps.io/GAC_bv10_estimation_beta/

References within Appendix A

Anumol, T., Sgroi, M., Park, M., Roccaro, P., & Snyder, S. A. (2015). Predicting trace organic compound breakthrough in granular activated carbon using fluorescence and UV absorbance as surrogates. *Water Research*, 76, 76–87.

<https://doi.org/10.1016/j.watres.2015.02.019>

Aschermann, G., Zietzschmann, F., & Jekel, M. (2018). Influence of dissolved organic matter and activated carbon pore characteristics on organic micropollutant desorption. *Water Research*, 133, 123–131. <https://doi.org/10.1016/j.watres.2018.01.015>

Boehmke, B., & Greenwell, B. M. (2019). *Hands-On Machine Learning with R*. CRC Press.

Boethling, R. S., Lynch, D. G., & Thom, G. C. (2003). Predicting ready biodegradability of premanufacture notice chemicals. *Environmental Toxicology and Chemistry*, 22(4), 837–844. <https://doi.org/10.1002/etc.5620220423>

Breiman, L. (2001). Random Forests. *Machine Learning*, 45(1), 5–32.

<https://doi.org/10.1023/A:1010933404324>

Burkhardt, J., & T, S. (n.d.). *Contact Us AdDesignS: Modeling Drinking Water Treatment with Granular Activated Carbon* [Conference presentation].

https://cfpub.epa.gov/si/si_public_file_download.cfm?p_download_id=541152&Lab=CESE

R

Chae, S.-H., Kim, S.-S., Jeong, W., & Park, N.-S. (2013). Evaluation of physical properties and adsorption capacity of regenerated granular activated carbons (GACs). *Korean Journal of Chemical Engineering*, 30(4), 891–897. <https://doi.org/10.1007/s11814-012-0179-9>

- Dormann, C. F., Elith, J., Bacher, S., Buchmann, C., Carl, G., Carré, G., Marquéz, J. R. G., Gruber, B., Lafourcade, B., Leitão, P. J., Münkemüller, T., McClean, C., Osborne, P. E., Reineking, B., Schröder, B., Skidmore, A. K., Zurell, D., & Lautenbach, S. (2013). Collinearity: A review of methods to deal with it and a simulation study evaluating their performance. *Ecography*, *36*(1), 27–46. <https://doi.org/10.1111/j.1600-0587.2012.07348.x>
- Drevon, D., Fursa, S. R., & Malcolm, A. L. (2017). Intercoder Reliability and Validity of WebPlotDigitizer in Extracting Graphed Data. *Behavior Modification*, *41*(2), 323–339. <https://doi.org/10.1177/0145445516673998>
- Ebie, K., Li, F., Azuma, Y., Yuasa, A., & Hagishita, T. (2001). Pore distribution effect of activated carbon in adsorbing organic micropollutants from natural water. *Water Research*, *35*(1), 167–179. [https://doi.org/10.1016/S0043-1354\(00\)00257-8](https://doi.org/10.1016/S0043-1354(00)00257-8)
- Erto, A., Lancia, A., & Musmarra, D. (2013). Fixed-bed Adsorption of Trichloroethylene onto Activated Carbon. In Pierucci, S and Klemes, JJ (Ed.), *ICHEAP-11: 11TH INTERNATIONAL CONFERENCE ON CHEMICAL AND PROCESS ENGINEERING, PTS 1-4* (Vol. 32, pp. 1969–1974). <https://doi.org/10.3303/CET1332329>
- Friedman, J. H. (2001). Greedy Function Approximation: A Gradient Boosting Machine. *The Annals of Statistics*, *29*(5), 1189–1232.
- Goldstein, A., Kapelner, A., Bleich, J., & Pitkin, E. (2015). Peeking Inside the Black Box: Visualizing Statistical Learning With Plots of Individual Conditional Expectation. *Journal of Computational and Graphical Statistics*, *24*(1), 44–65. <https://doi.org/10.1080/10618600.2014.907095>

- Grömping, U. (2009). Variable Importance Assessment in Regression: Linear Regression versus Random Forest. *The American Statistician*, 63(4), 308–319.
<https://doi.org/10.1198/tast.2009.08199>
- Hastie, T., Qian, J., & Tay, K. (n.d.). *An Introduction to glmnet*. 38.
- Hnatukova, P., Kopecka, I., & Pivokonsky, M. (2011). Adsorption of cellular peptides of *Microcystis aeruginosa* and two herbicides onto activated carbon: Effect of surface charge and interactions. *Water Research*, 45(11), 3359–3368.
<https://doi.org/10.1016/j.watres.2011.03.051>
- Karanfil, T., Schlautman, M. A., & Erdogan, I. (2002). Survey of DOC and UV measurement practices with implications for SUVA determination. *Journal AWWA*, 94(12), 68–80.
<https://doi.org/10.1002/j.1551-8833.2002.tb10250.x>
- Karanjkar, P. U., Burt, S. P., Chen, X., Barnett, K. J., Ball, M. R., Kumbhalkar, M. D., Wang, X., Miller, J. B., Hermans, I., Dumesic, J. A., & Huber, G. W. (2016). Effect of carbon supports on RhRe bifunctional catalysts for selective hydrogenolysis of tetrahydropyran-2-methanol. *Catalysis Science & Technology*, 6(21), 7841–7851. <https://doi.org/10.1039/C6CY01763K>
- Kuhn, M. (2008). Building Predictive Models in R Using the caret Package. *Journal of Statistical Software*, 28(1), 1–26. <https://doi.org/10.18637/jss.v028.i05>
- Leeuwenberg, A. M., van Smeden, M., Langendijk, J. A., Mauer, M. E., Moons, K. G. M., Reitsma, J. B., & Schuit, E. (n.d.). *Comparing methods addressing multi-collinearity when developing prediction models*. 32.

- Li, Q., Snoeyink, V. L., Mariñas, B. J., & Campos, C. (2003). Elucidating competitive adsorption mechanisms of atrazine and NOM using model compounds. *Water Research*, 37(4), 773–784. [https://doi.org/10.1016/S0043-1354\(02\)00390-1](https://doi.org/10.1016/S0043-1354(02)00390-1)
- Midi, H., Sarkar, S. K., & Rana, S. (2010). Collinearity diagnostics of binary logistic regression model. *Journal of Interdisciplinary Mathematics*, 13(3), 253–267. <https://doi.org/10.1080/09720502.2010.10700699>
- Nicodemus, K. K., Malley, J. D., Strobl, C., & Ziegler, A. (2010). The behaviour of random forest permutation-based variable importance measures under predictor correlation. *BMC Bioinformatics*, 11(1), 110. <https://doi.org/10.1186/1471-2105-11-110>
- Pelekani, C., & Snoeyink, V. L. (2000). Competitive adsorption between atrazine and methylene blue on activated carbon: The importance of pore size distribution. *Carbon*, 38(10), 1423–1436. [https://doi.org/10.1016/S0008-6223\(99\)00261-4](https://doi.org/10.1016/S0008-6223(99)00261-4)
- Pelekani, C., & Snoeyink, V. L. (2001). A kinetic and equilibrium study of competitive adsorption between atrazine and Congo red dye on activated carbon: The importance of pore size distribution. *Carbon*, 39(1), 25–37. [https://doi.org/10.1016/S0008-6223\(00\)00078-6](https://doi.org/10.1016/S0008-6223(00)00078-6)
- Quinlivan, P. A., Li, L., & Knappe, D. R. U. (2005). Effects of activated carbon characteristics on the simultaneous adsorption of aqueous organic micropollutants and natural organic matter. *Water Research*, 39(8), 1663–1673. <https://doi.org/10.1016/j.watres.2005.01.029>
- Redding, A. M., Cannon, F. S., Snyder, S. A., & Vanderford, B. J. (2009). A QSAR-like analysis of the adsorption of endocrine disrupting compounds, pharmaceuticals, and personal care

- products on modified activated carbons. *Water Research*, 43(15), 3849–3861.
<https://doi.org/10.1016/j.watres.2009.05.026>
- Reed, B. E. (1995). Identification of Removal Mechanisms for Lead in Granular Activated Carbon (GAC) Columns. *Separation Science and Technology*, 30(1), 101–116.
<https://doi.org/10.1080/01496399508012216>
- Ridgeway, G. (2020). *Generalized Boosted Models: A guide to the gbm package*. 15.
- Rohatgi, A. (2020). *WebPlotDigitizer User Manual Version 4.3*. 23.
- Selmi, T., Seffen, M., Celzard, A., & Fierro, V. (2020). Effect of the adsorption pH and temperature on the parameters of the Brouers–Sotolongo models. *Environmental Science and Pollution Research*, 27(19), 23437–23446. <https://doi.org/10.1007/s11356-018-3835-8>
- Tibshirani, R. (1996). Regression Shrinkage and Selection Via the Lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1), 267–288.
<https://doi.org/10.1111/j.2517-6161.1996.tb02080.x>
- US EPA. (2021). *Estimation Programs Interface Suite™ for Microsoft® Windows* (4.10) [Windows]. <https://www.epa.gov/tsca-screening-tools/epi-suitetm-estimation-program-interface#citing>
- Vanderheyden, S. R. H., Ammel, R. V., Sobiech-Matura, K., Vanreppelen, K., Schreurs, S., Schroeyers, W., Yperman, J., & Carleer, R. (2016). Adsorption of cesium on different types of activated carbon. *Journal of Radioanalytical and Nuclear Chemistry*, 310(1), 301–310.
<https://doi.org/10.1007/s10967-016-4807-4>

Wright, M. N., & Ziegler, A. (2017). ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R. *Journal of Statistical Software*, 77(1).

<https://doi.org/10.18637/jss.v077.i01>

Yan, L., Lv, D., Huang, X., Shi, H., & Zhang, G. (2016). Adsorption characteristics of Bisphenol-A on tailored activated carbon in aqueous solutions. *Water Science and Technology*, 74(7), 1744–1751. <https://doi.org/10.2166/wst.2016.325>

Yapsakli, K., Cecen, F., Aktaş, Ö., & Can, Z. (2009). Impact of Surface Properties of Granular Activated Carbon and Preozonation on Adsorption and Desorption of Natural Organic Matter. *Environmental Engineering Science - ENVIRON ENG SCI*, 26, 489–500.

<https://doi.org/10.1089/ees.2008.0005>

Zhi, Y., & Liu, J. (2016). Surface modification of activated carbon for enhanced adsorption of perfluoroalkyl acids from aqueous solutions. *Chemosphere*, 144, 1224–1232.

<https://doi.org/10.1016/j.chemosphere.2015.09.097>

Appendix B

All GAC column studies used in this research and their corresponding references are summarized in Appendix B, as an external attachment in the format of 'xlsx'.

Training and test sets of the data as well as brief descriptions of column names of the tables are included.

Appendix C

Fast biodegradation probabilities of all recalcitrant organic MPs are summarized in Appendix C, included in an external zip folder.

Appendix D

A summary of breakthrough data extracted from figures from each study and fitted PSDM follows. Black dots represent data extracted using WebPlotDigitizer; red lines represent PSDM fit; light blue envelope represent the range of anticipated uncertainty of the PSDM output. Curve ID which corresponds to the row entries in Supplemental data 2 and maximum reporting limit (MRL) for the micropollutant (MP) when applicable is written in the title of each breakthrough figure. The MRL is also depicted in the graphing area in blue whenever applicable. For certain compounds in the study corresponding to Ref ID 15, the author fitted PSDM curves for those compounds were extracted instead of the raw data.

Breakthrough data of Ref ID 1

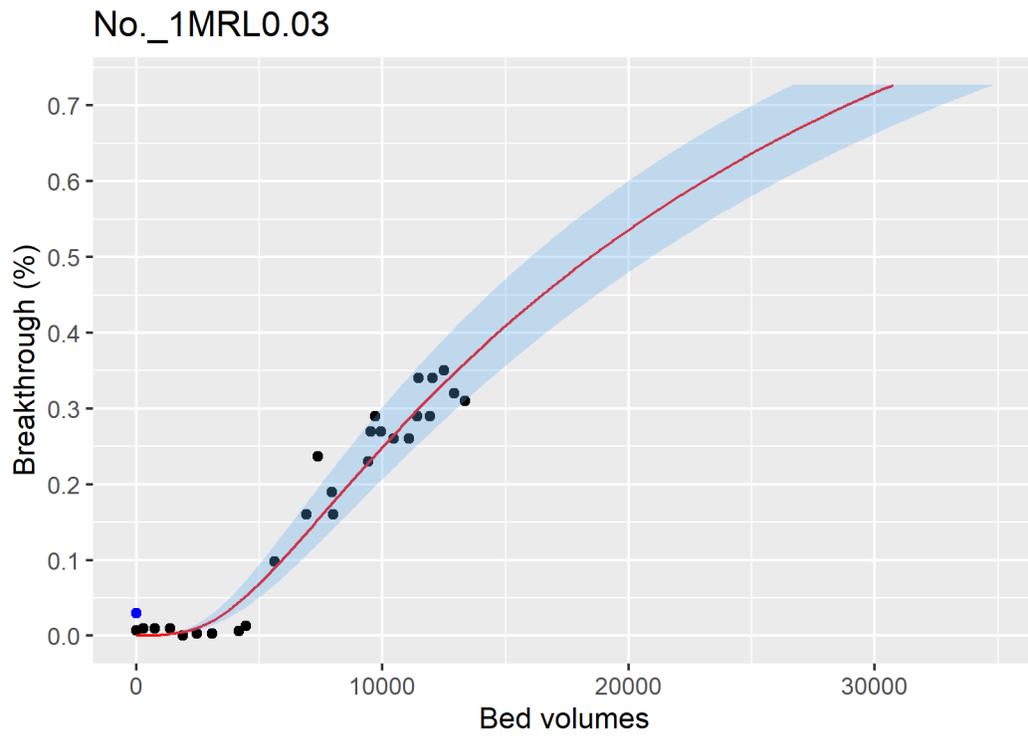


Figure ID 1- 1

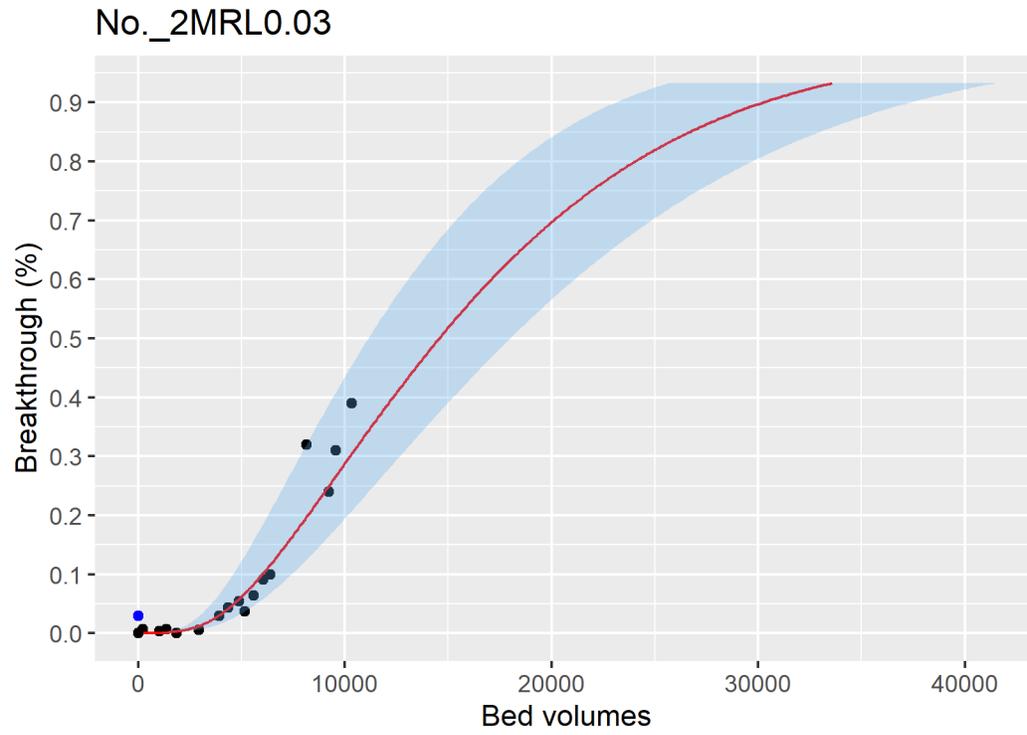


Figure ID 1- 2

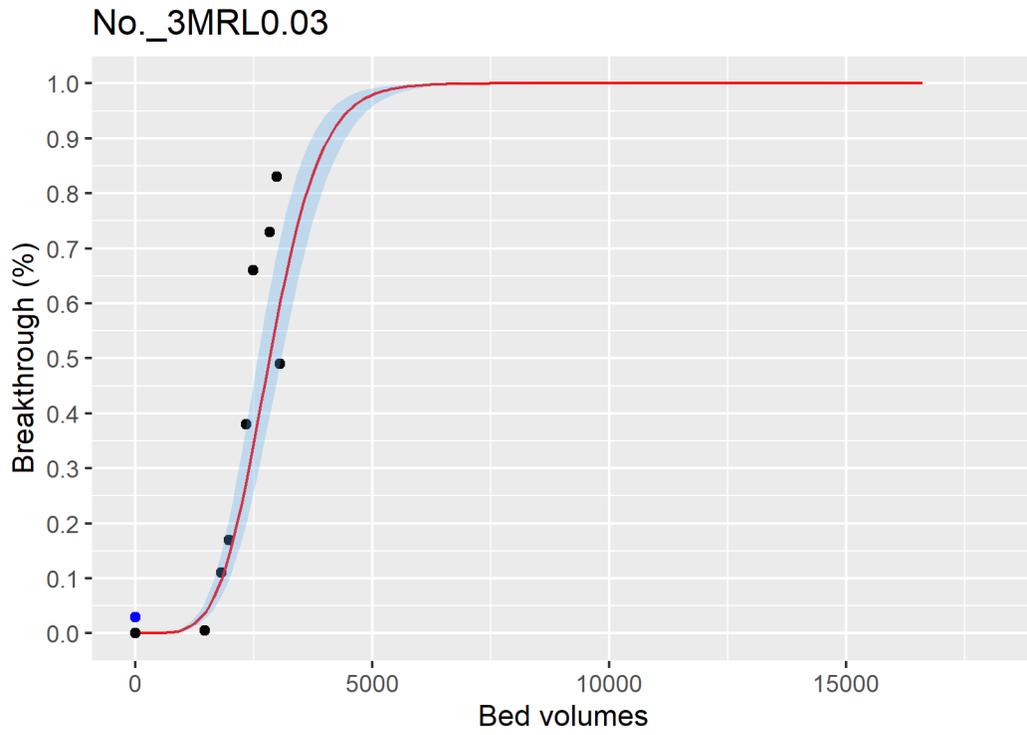


Figure ID 1- 3

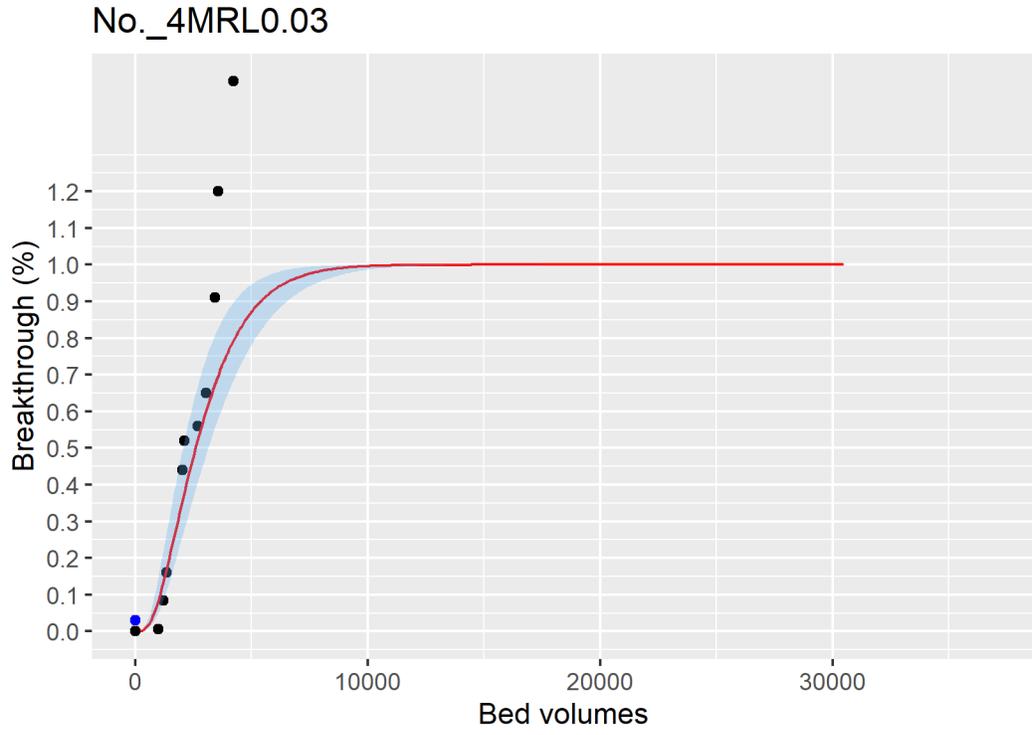


Figure ID 1- 4

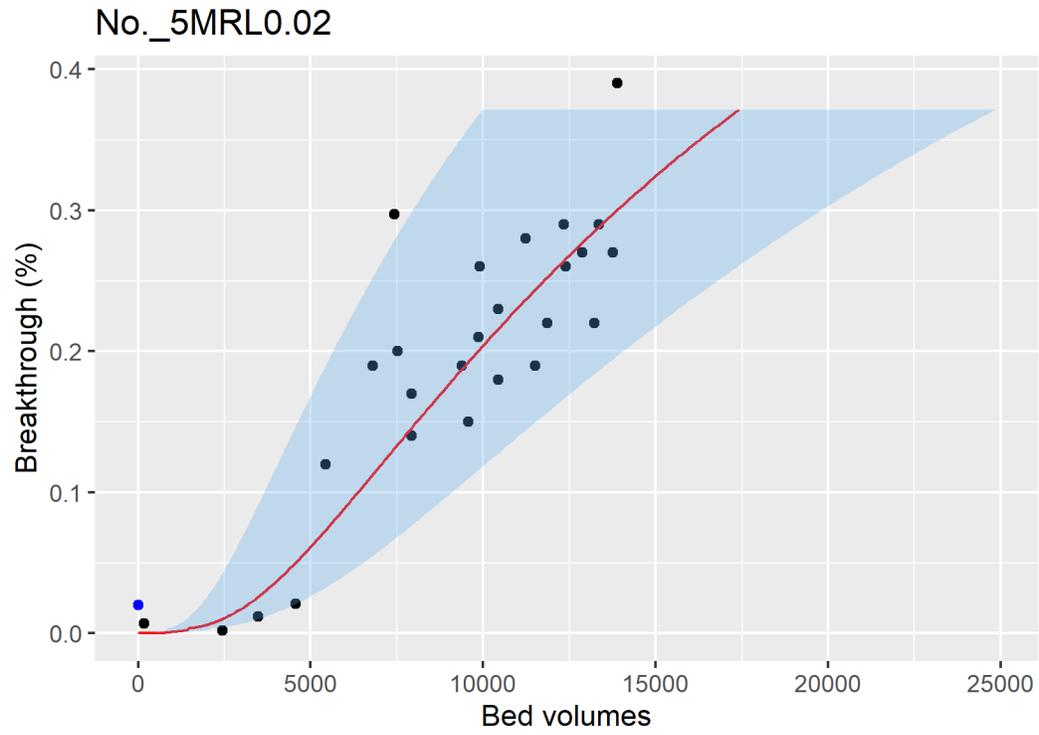


Figure ID 1- 5

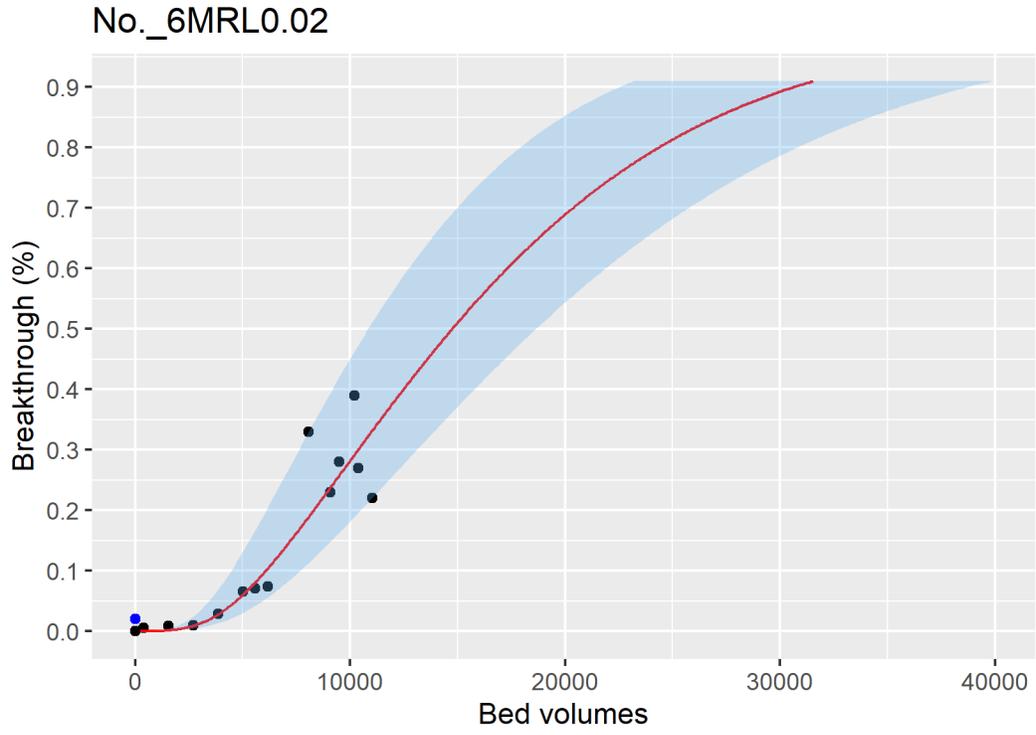


Figure ID 1- 6

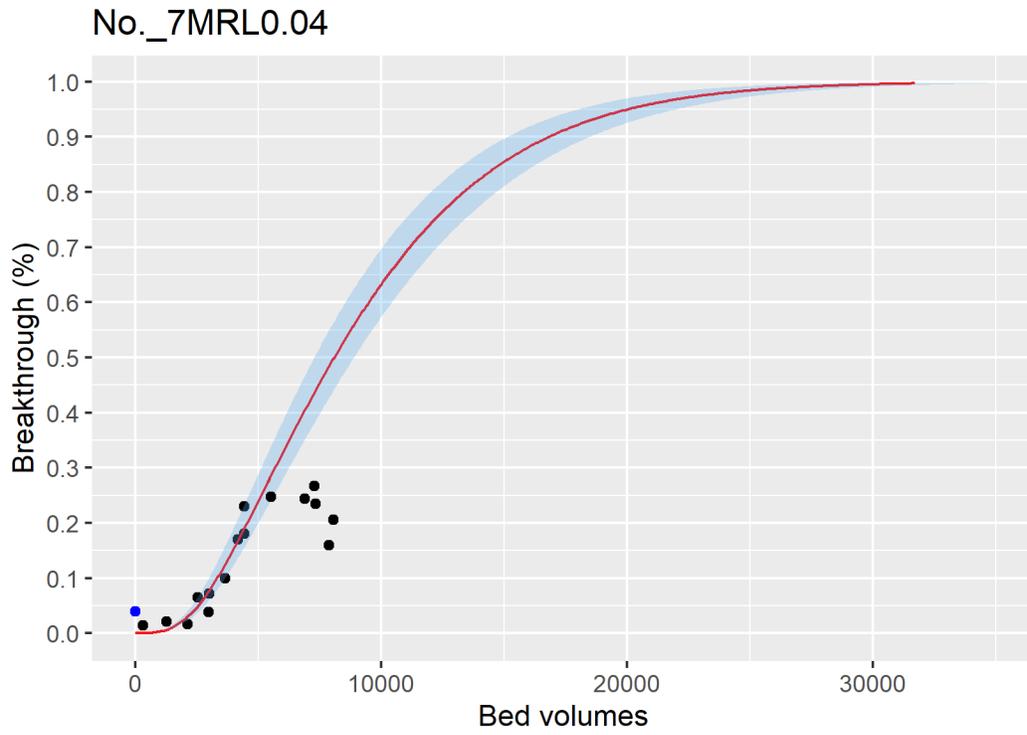


Figure ID 1- 7

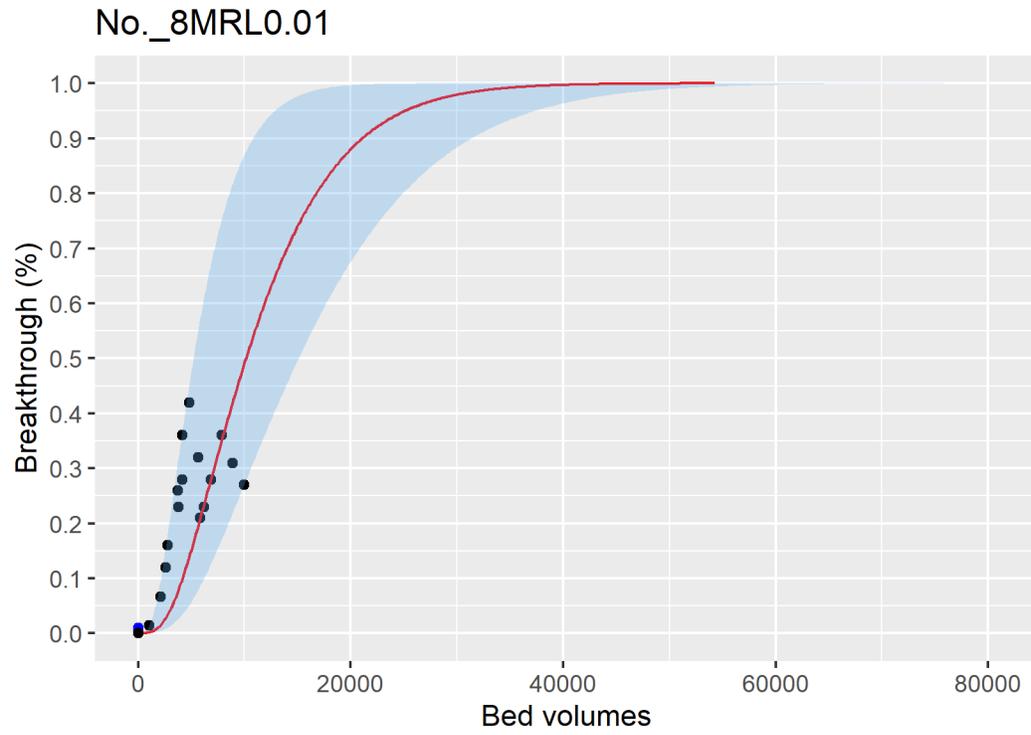


Figure ID 1- 8

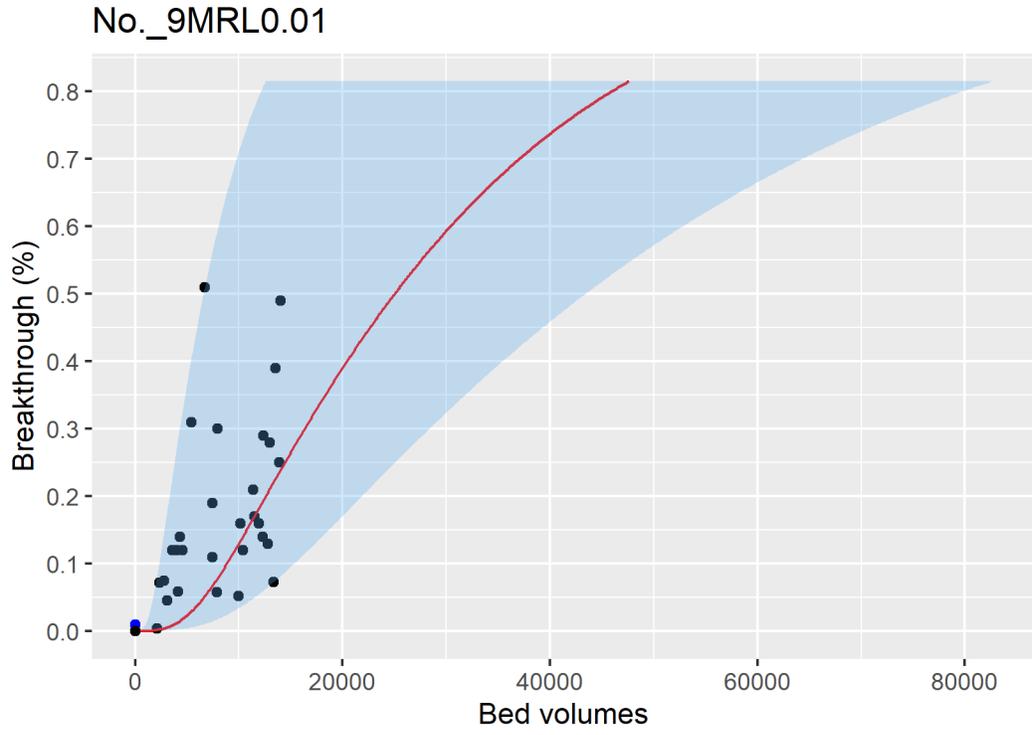


Figure ID 1-9

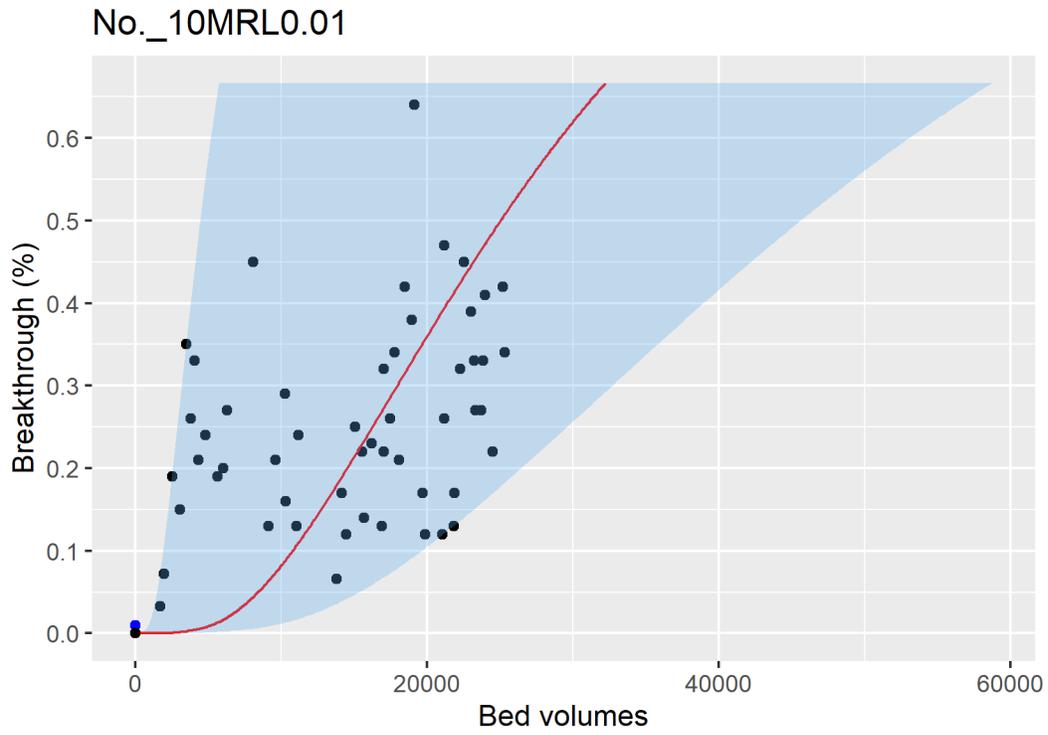


Figure ID 1- 10

No._13MRL0.17

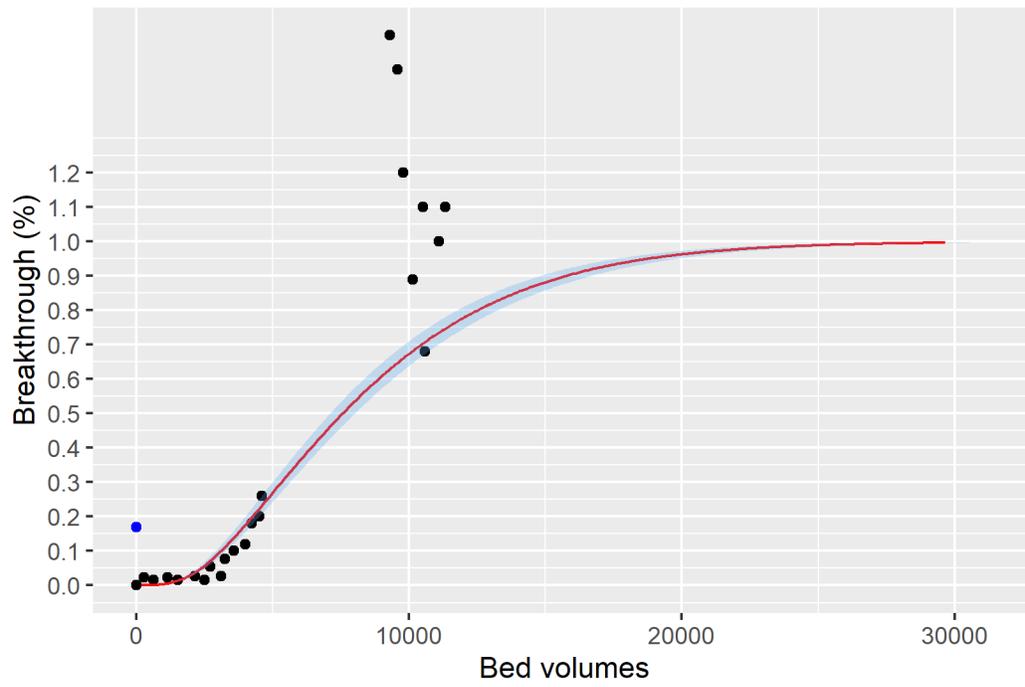


Figure ID 1- 11

No._14MRL0.17

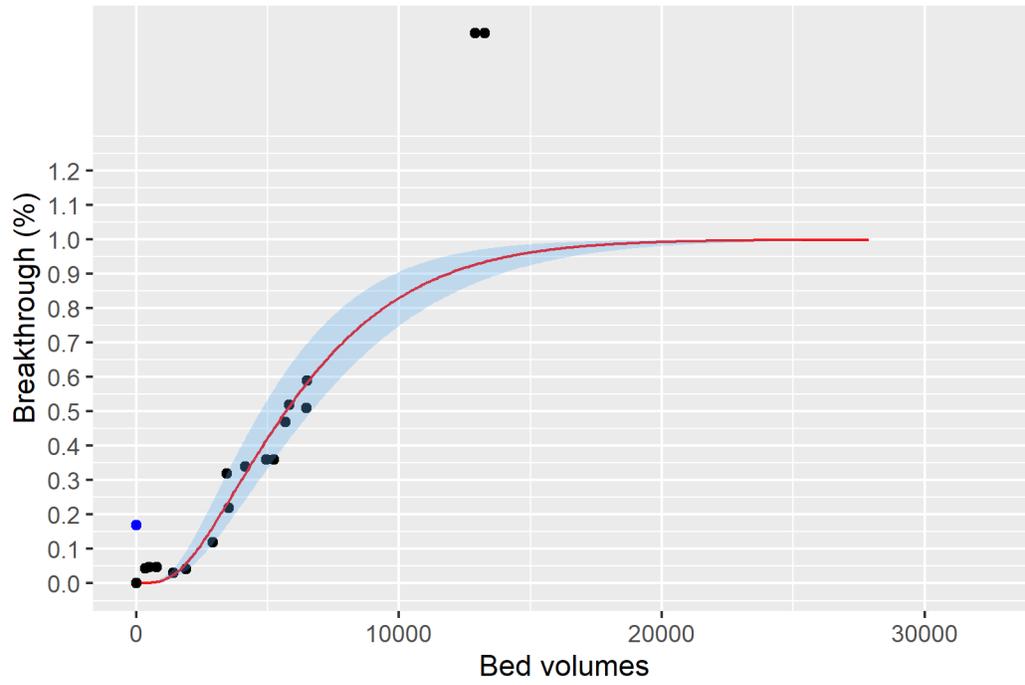


Figure ID 1- 12

Breakthrough data of Ref ID 15

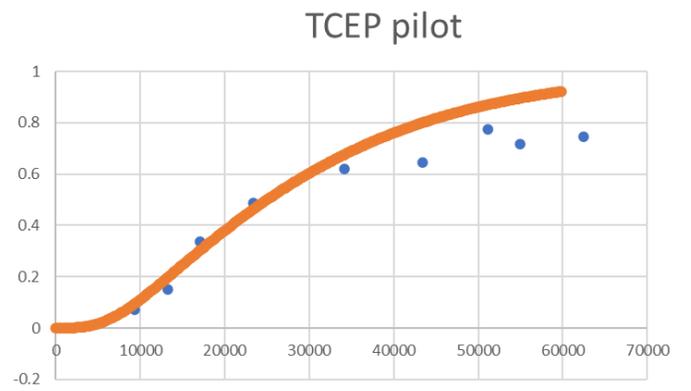
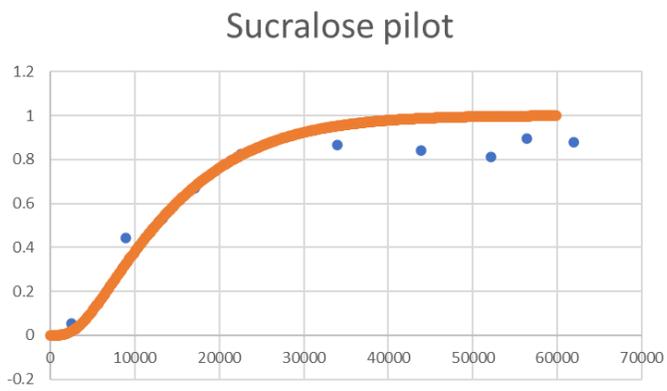
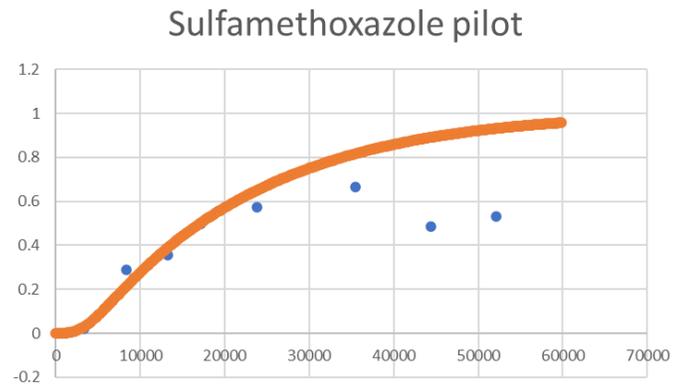
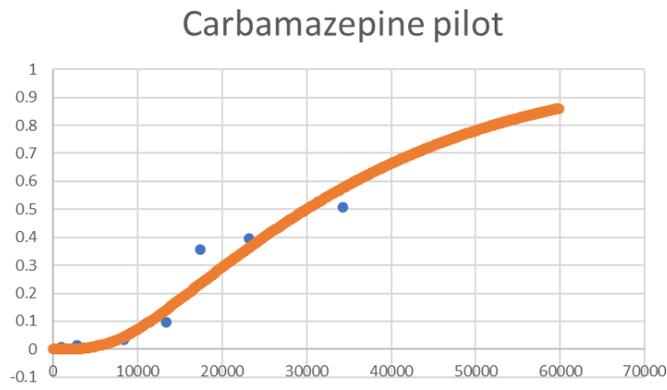


Figure ID 15- 1

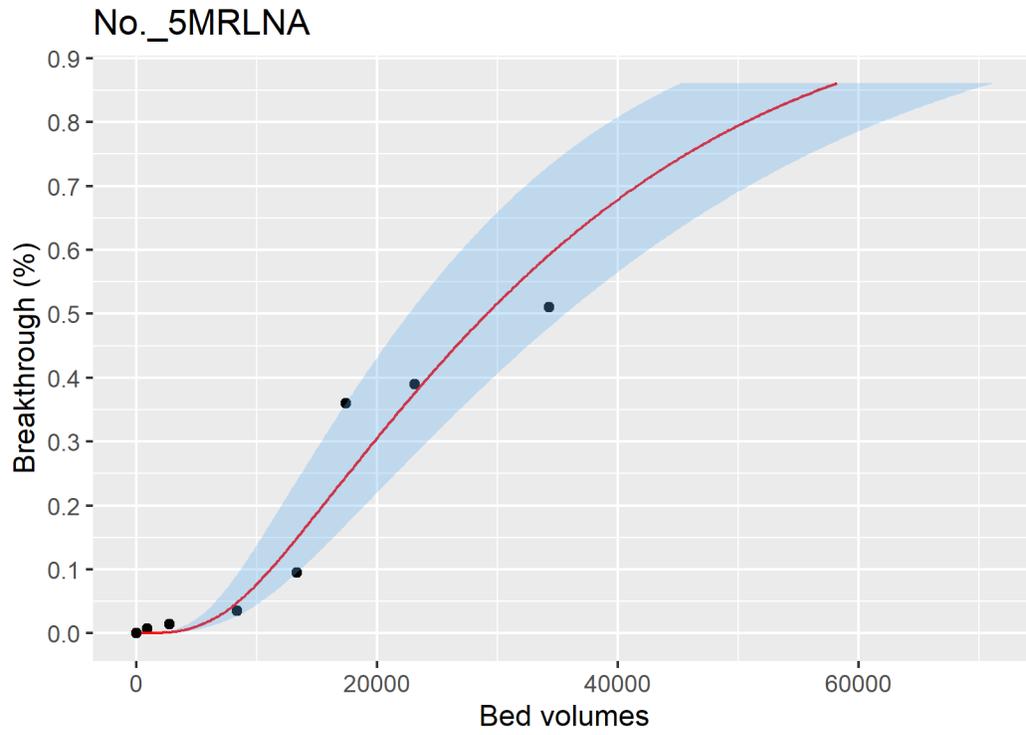


Figure ID 15- 2

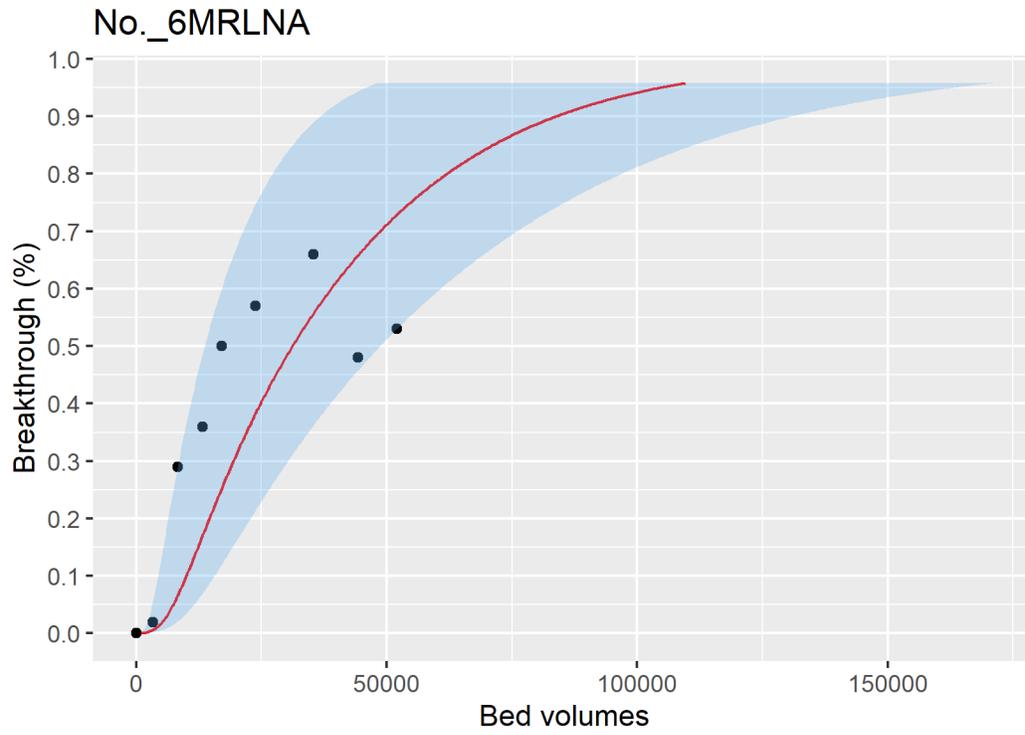


Figure ID 15- 3

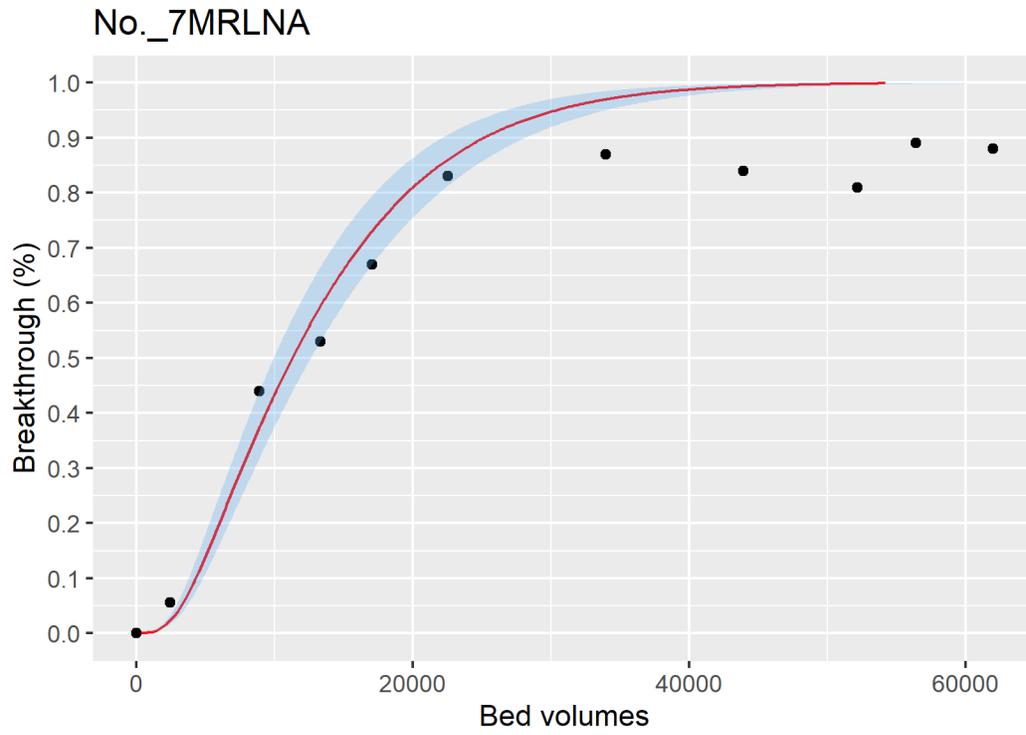


Figure ID 15- 4

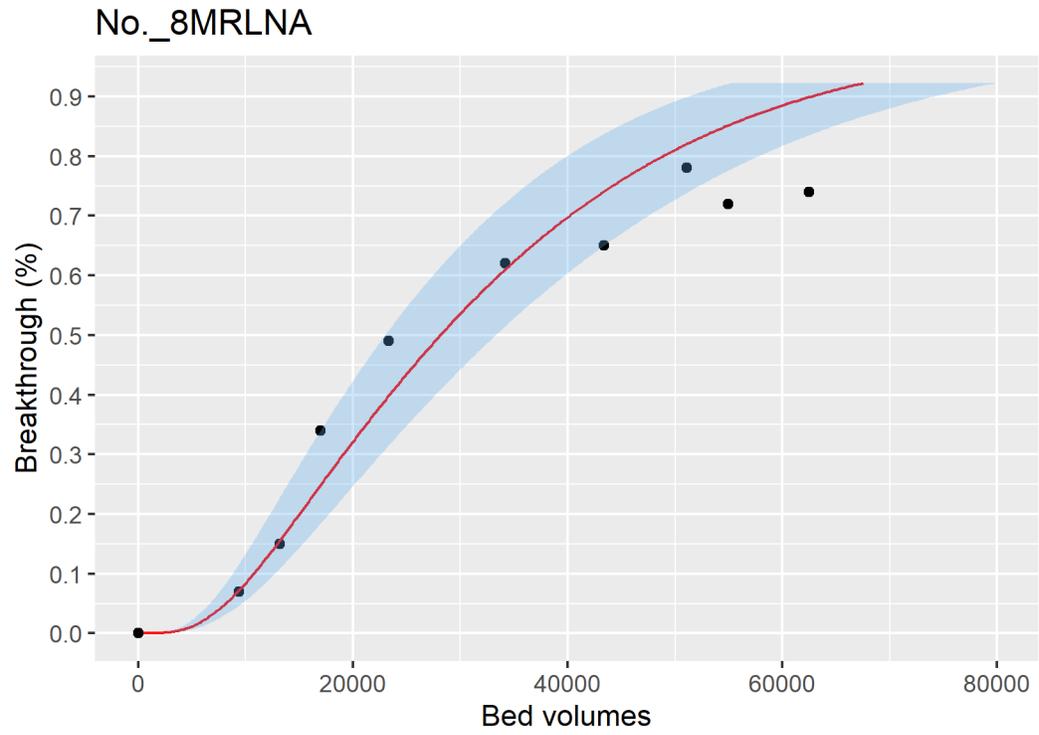


Figure ID 15- 5

Breakthrough data of Ref ID 19

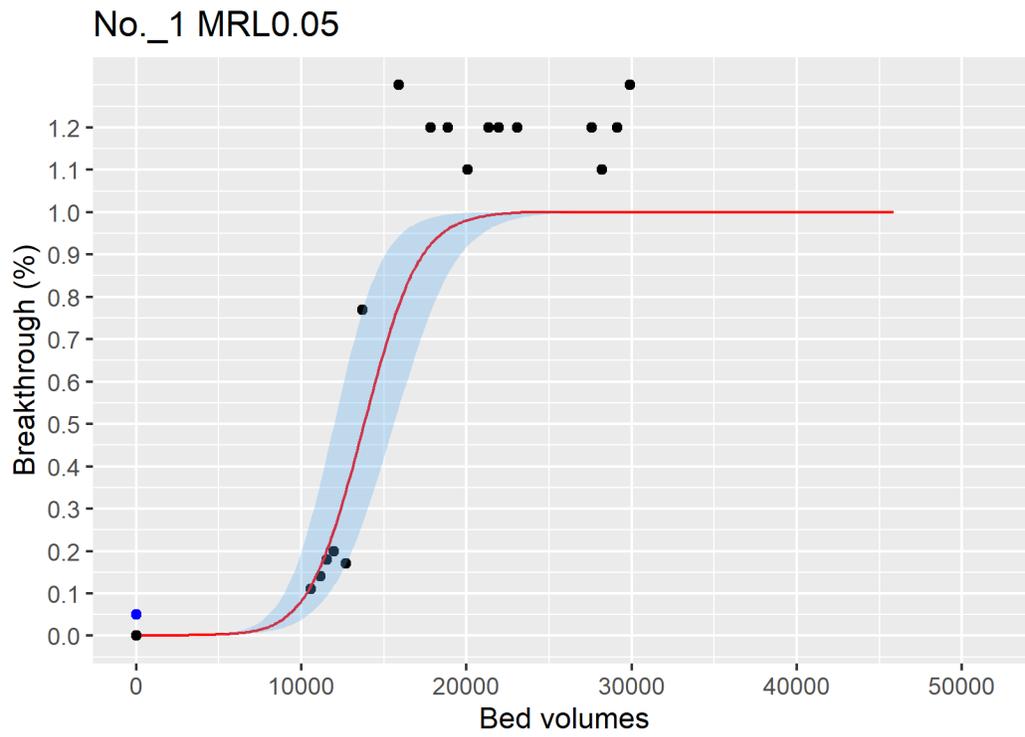


Figure ID 19- 1

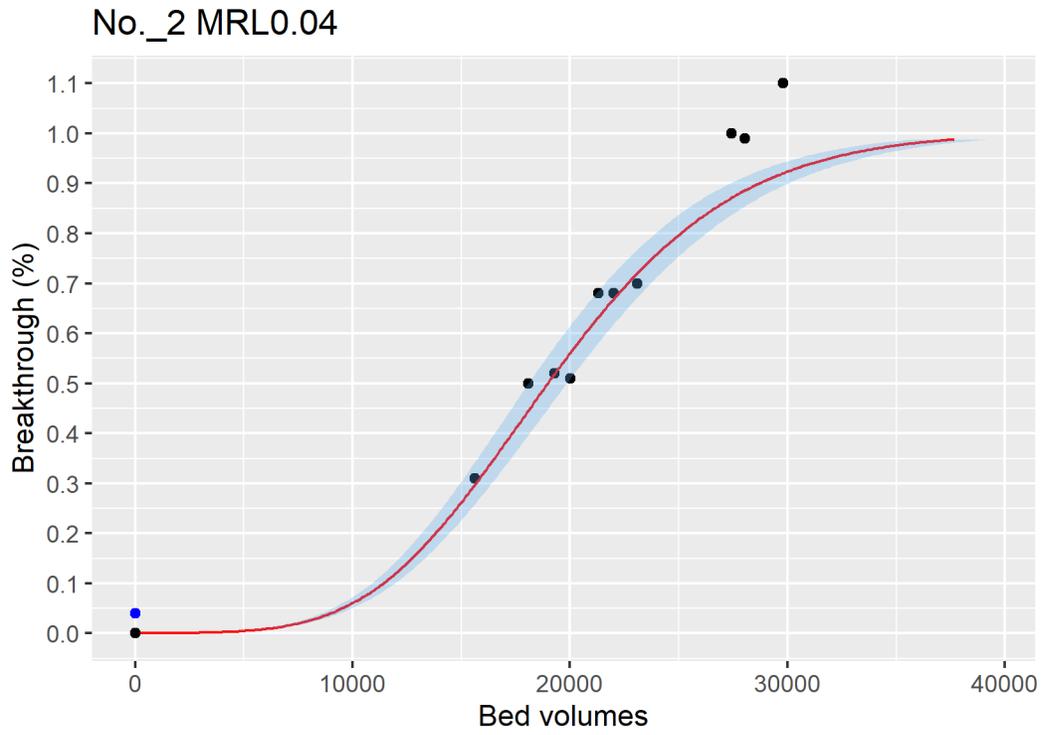


Figure ID 19- 2

No._3 MRL0.11

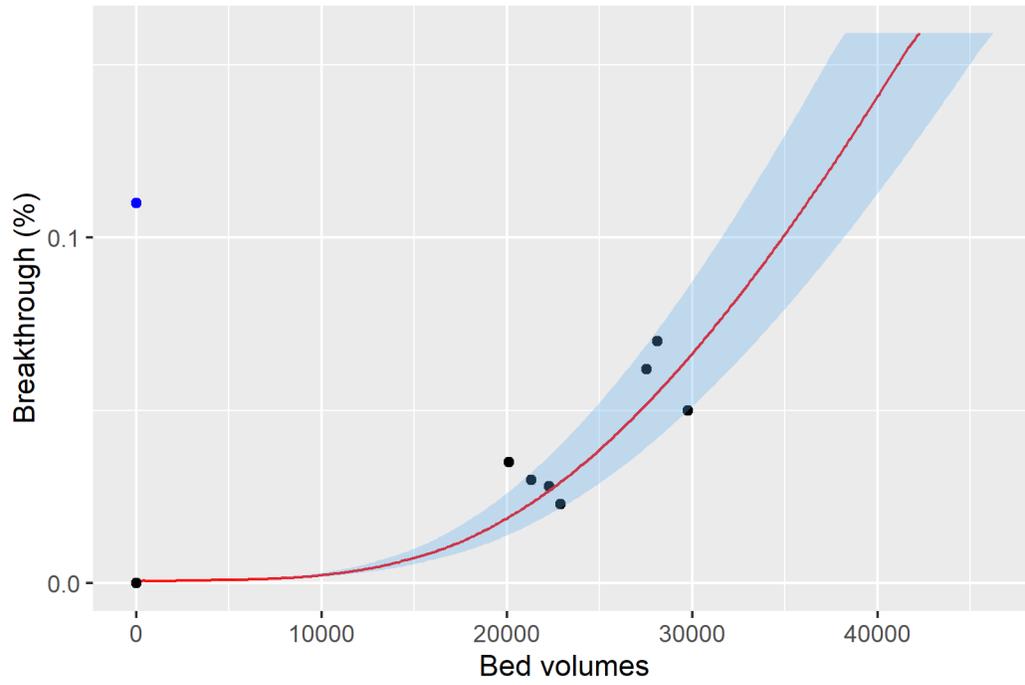


Figure ID 19- 3

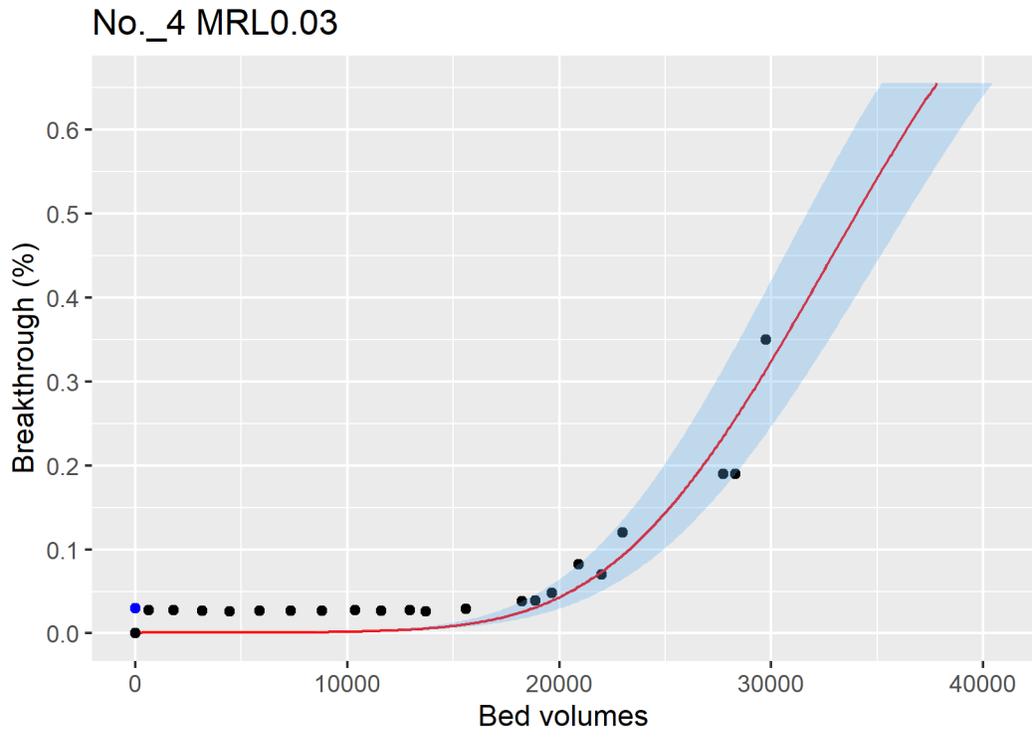


Figure ID 19- 4

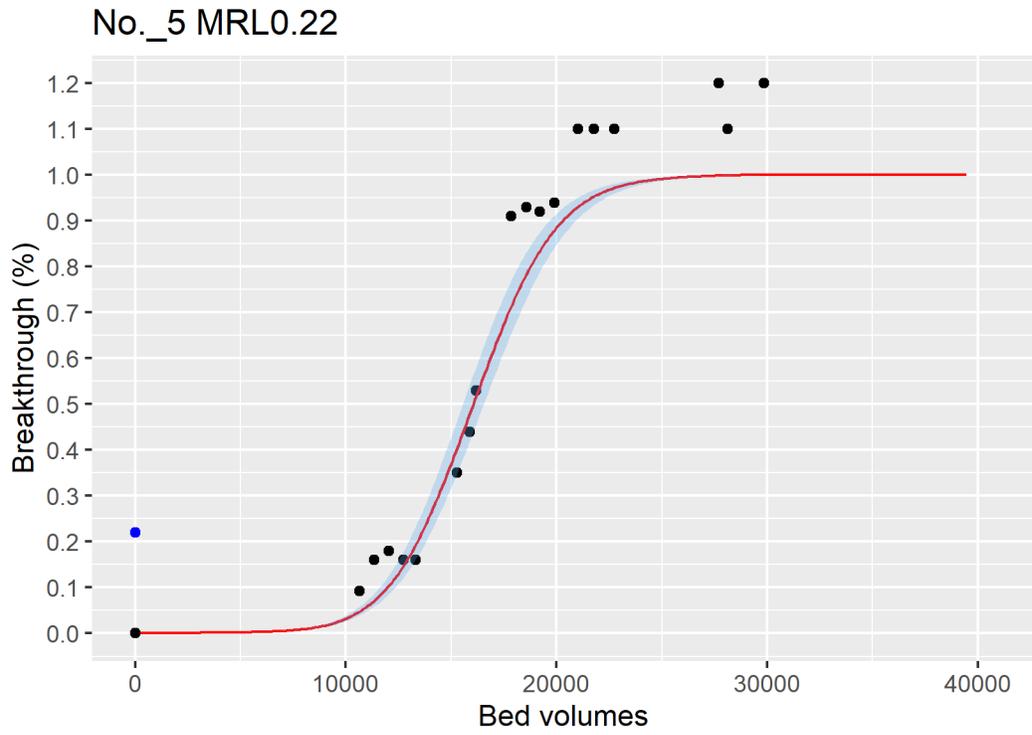


Figure ID 19- 5

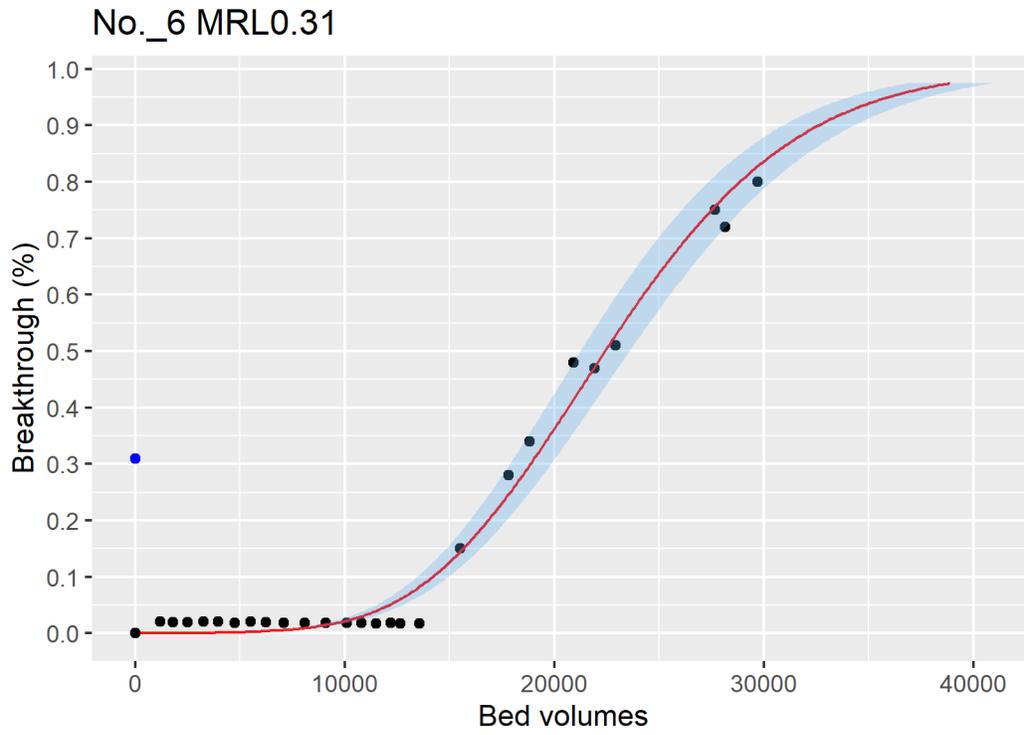


Figure ID 19- 6

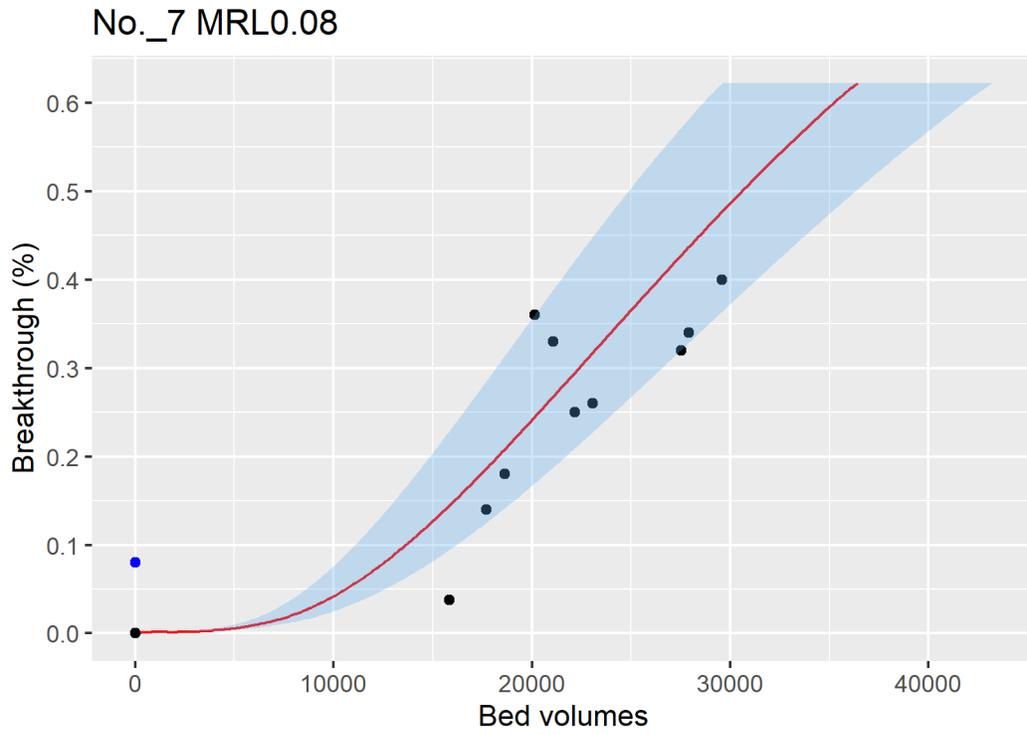


Figure ID 19- 7

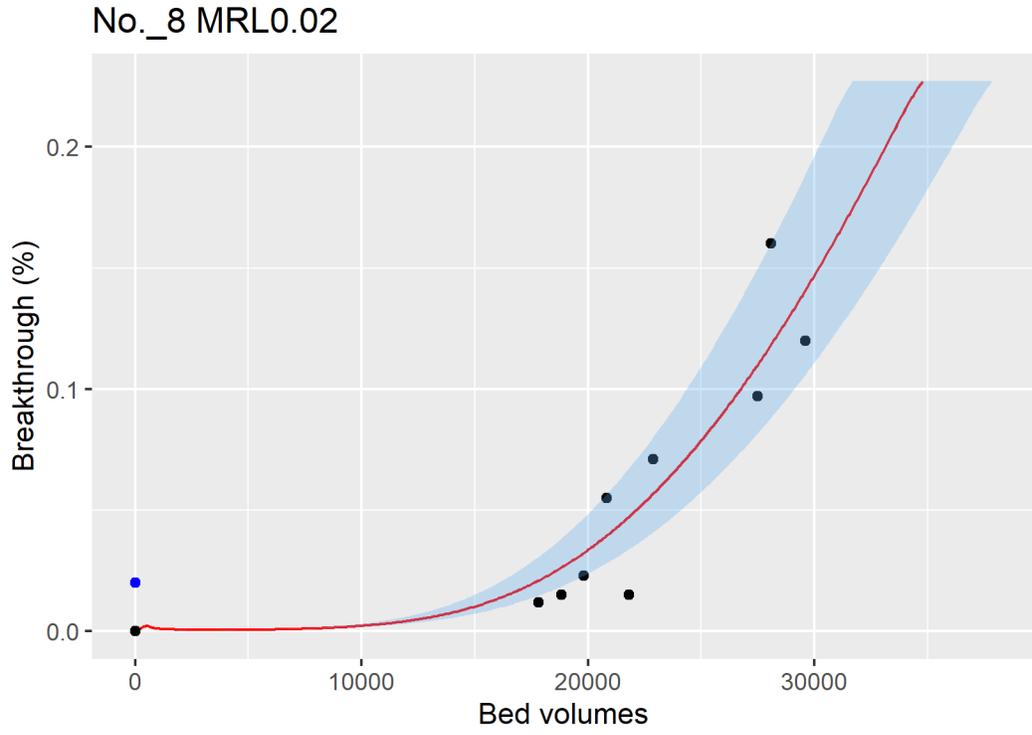


Figure ID 19- 8

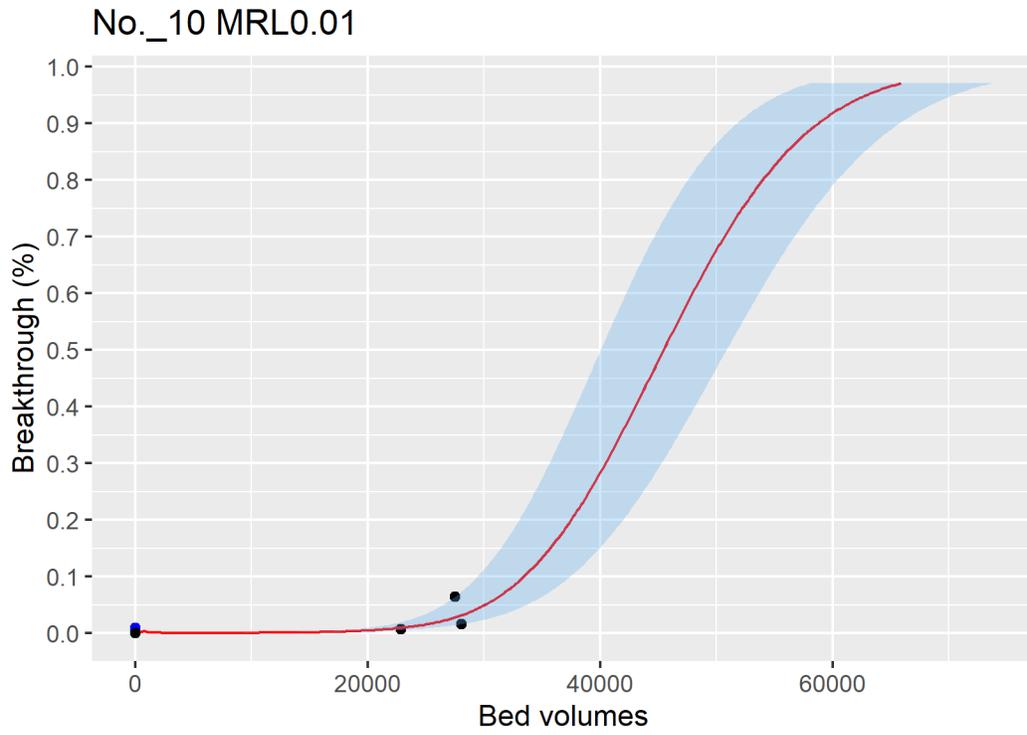


Figure ID 19- 9

No._11 MRL0.05

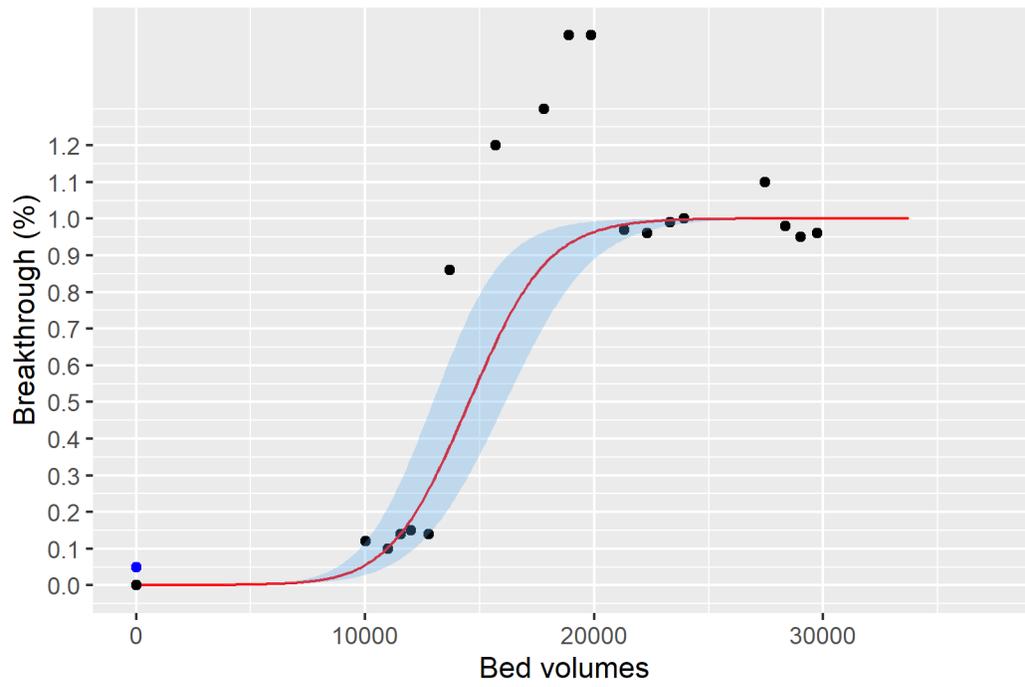


Figure ID 19- 10

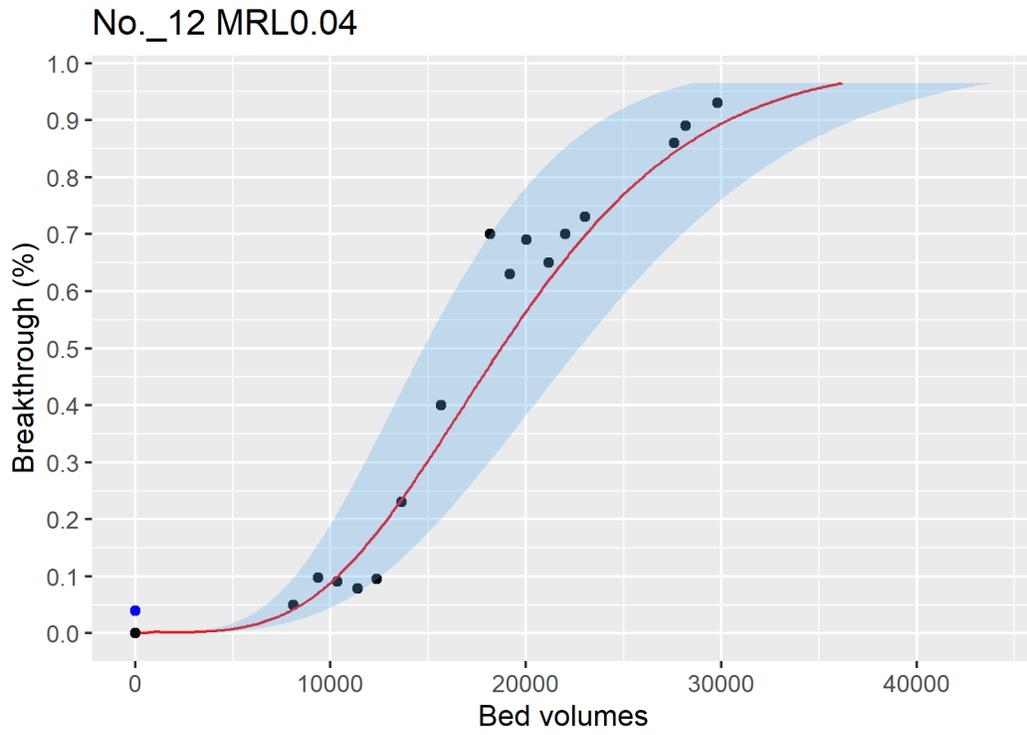


Figure ID 19- 11

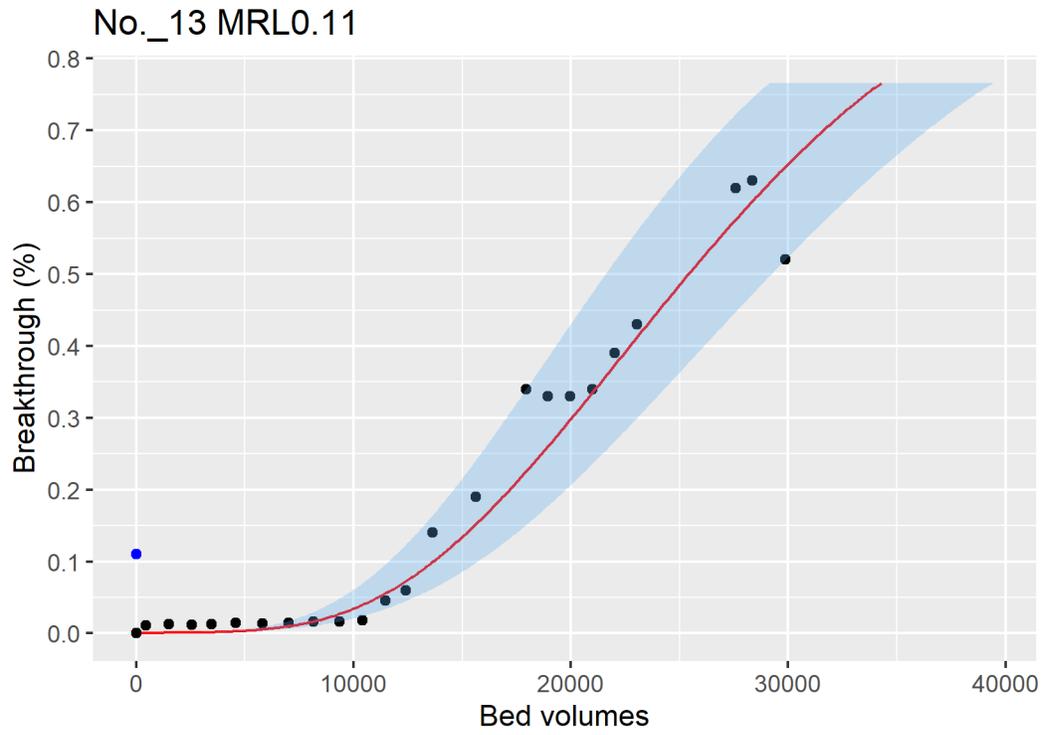


Figure ID 19- 12

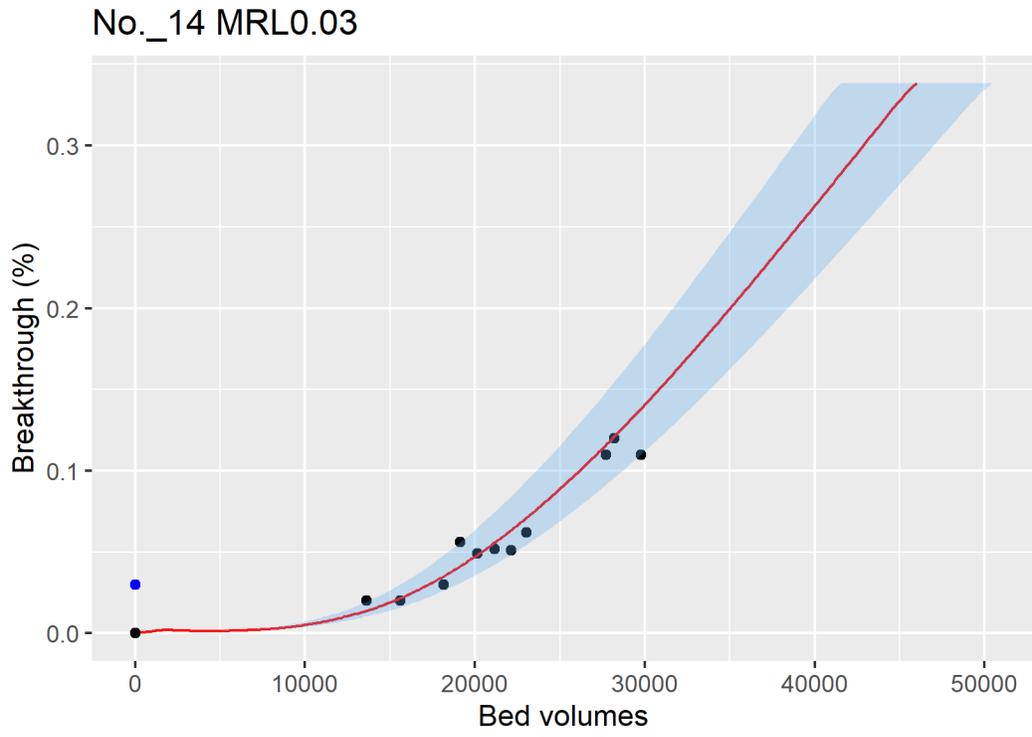


Figure ID 19- 13

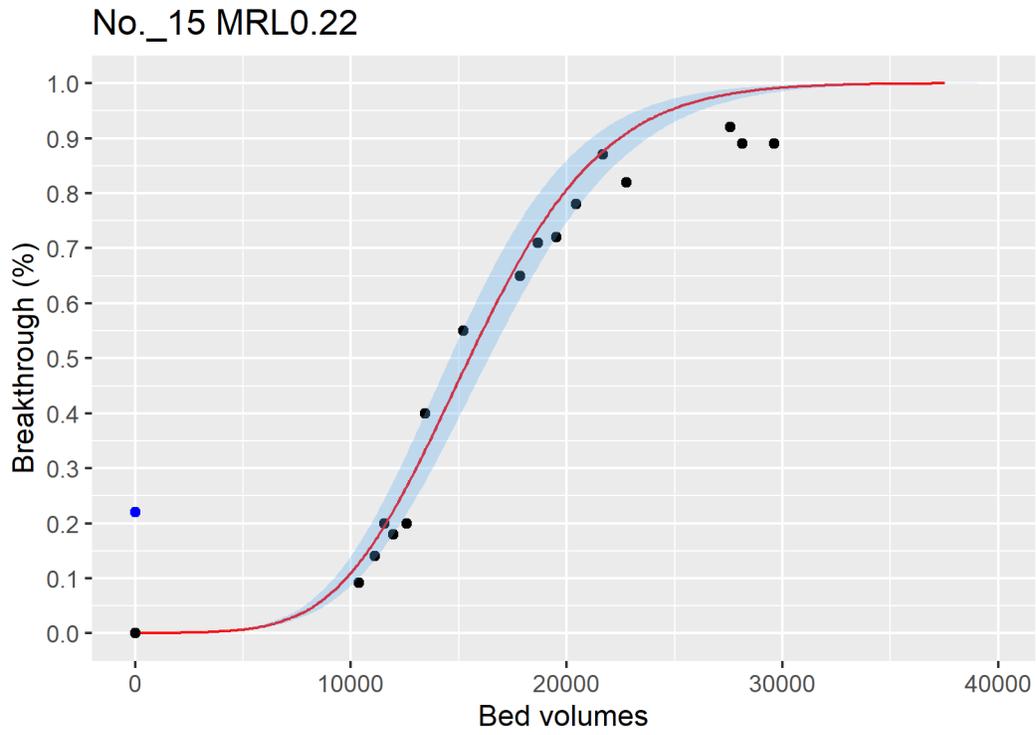


Figure ID 19- 14

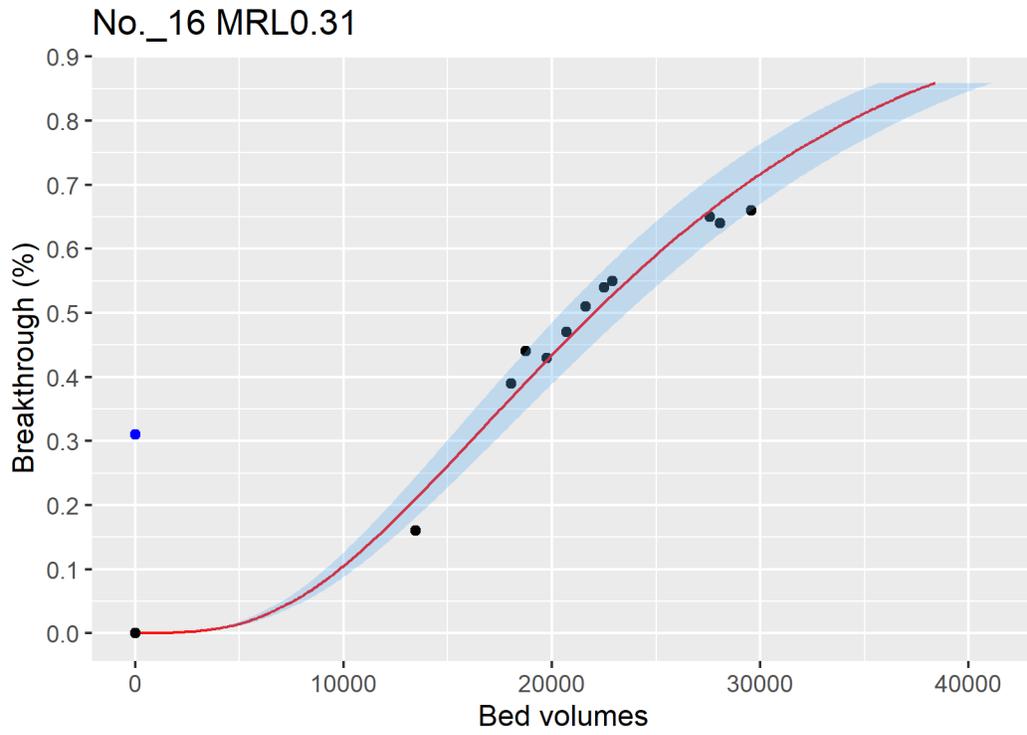


Figure ID 19- 15

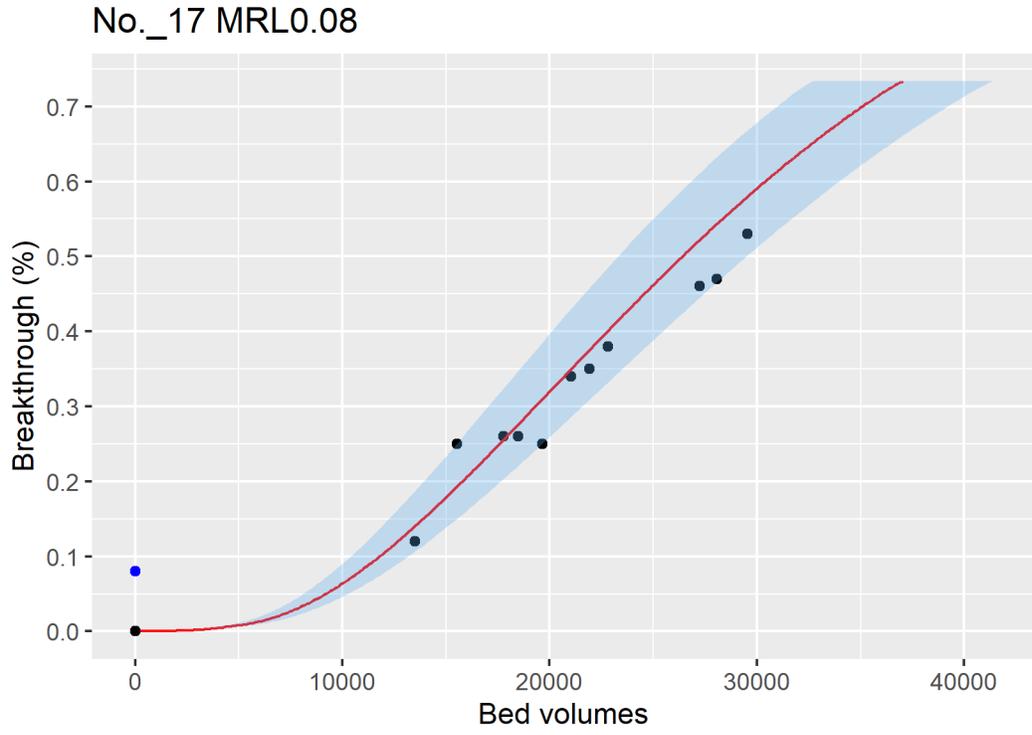


Figure ID 19- 16

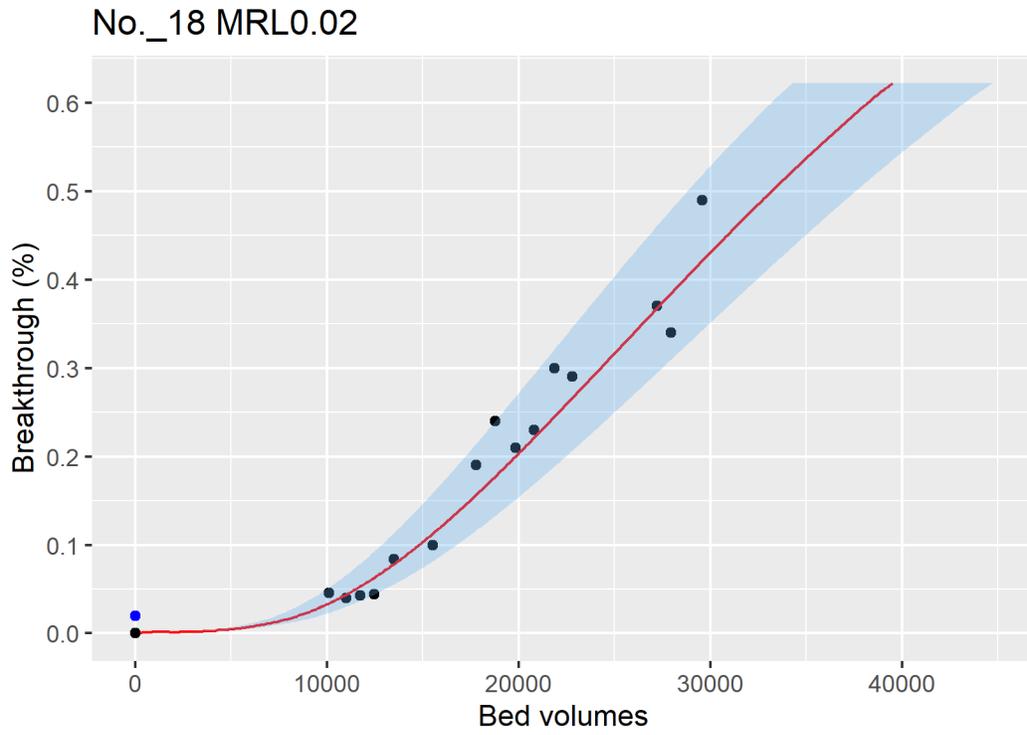


Figure ID 19- 17

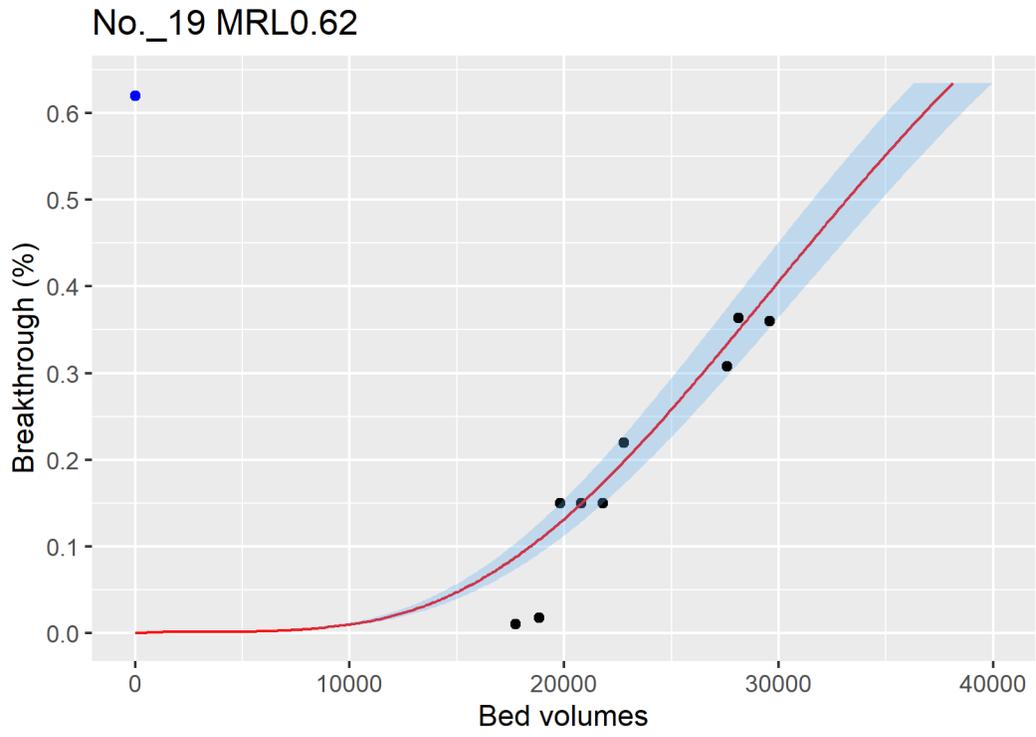


Figure ID 19- 18

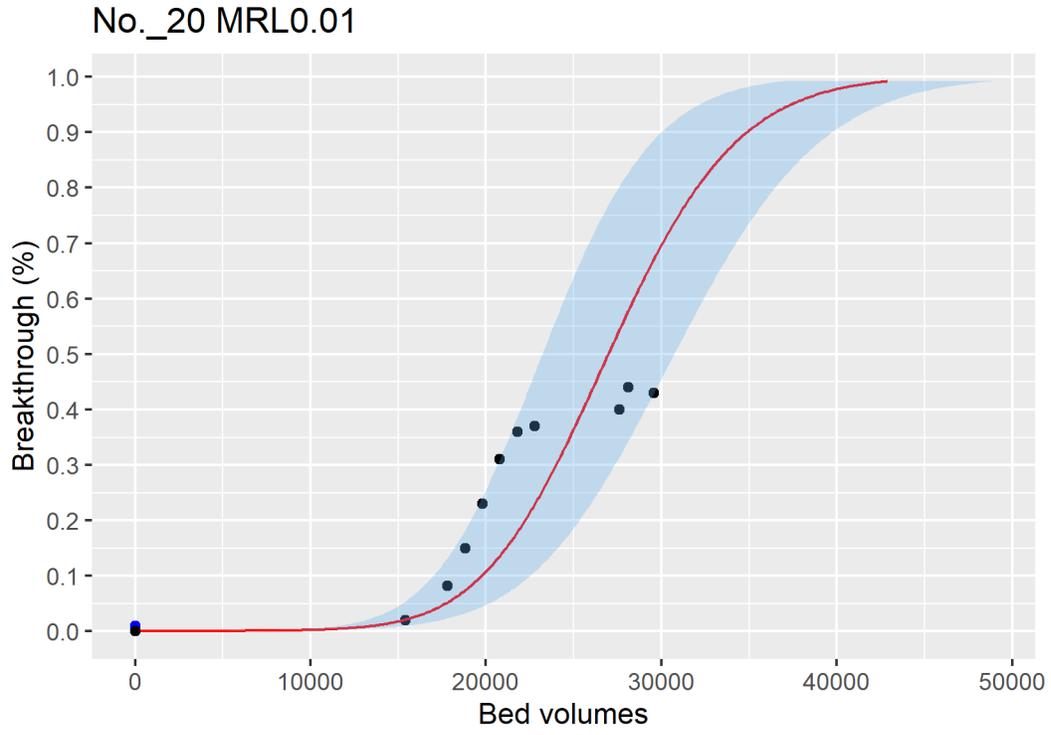


Figure ID 19- 19

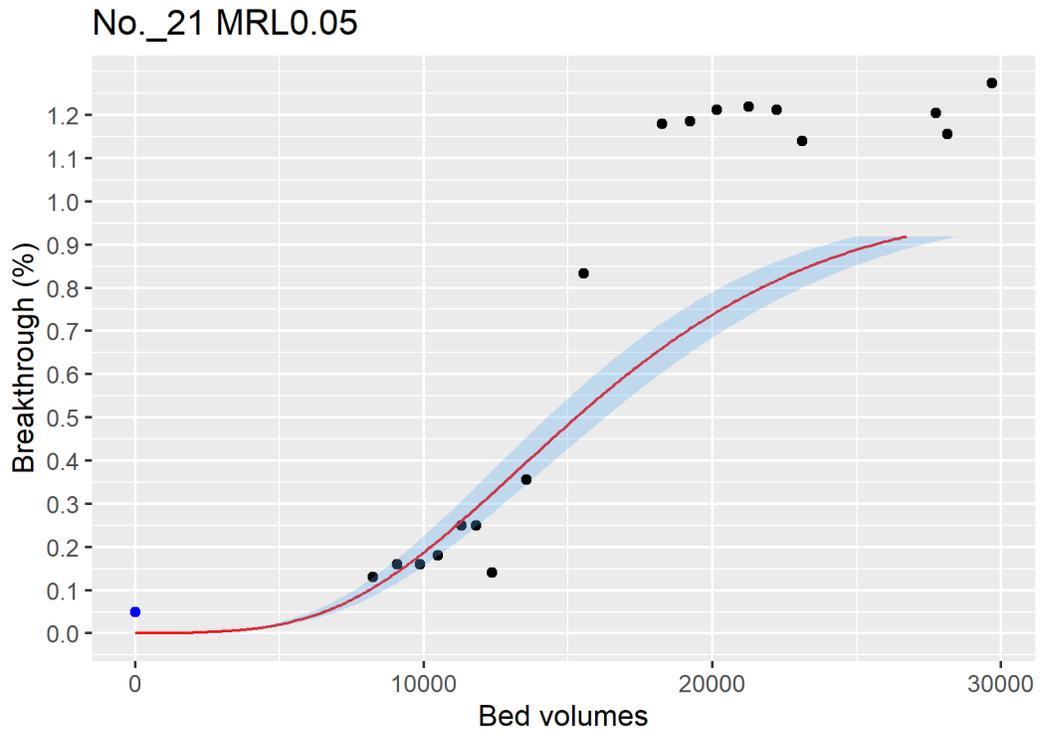


Figure ID 19- 20

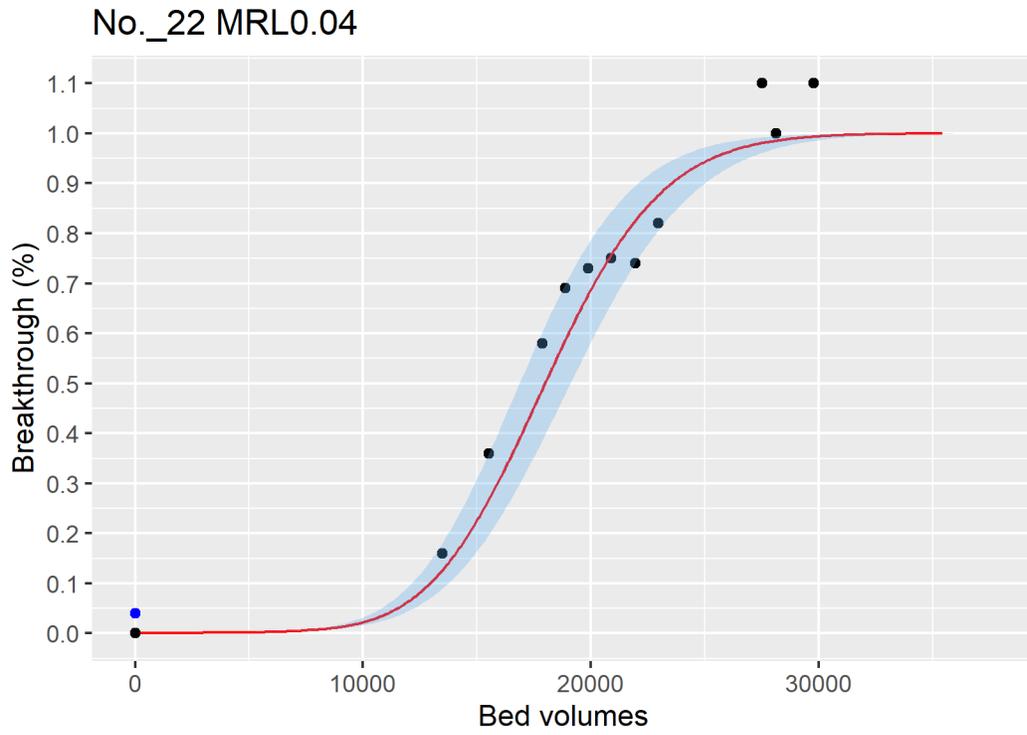


Figure ID 19- 21

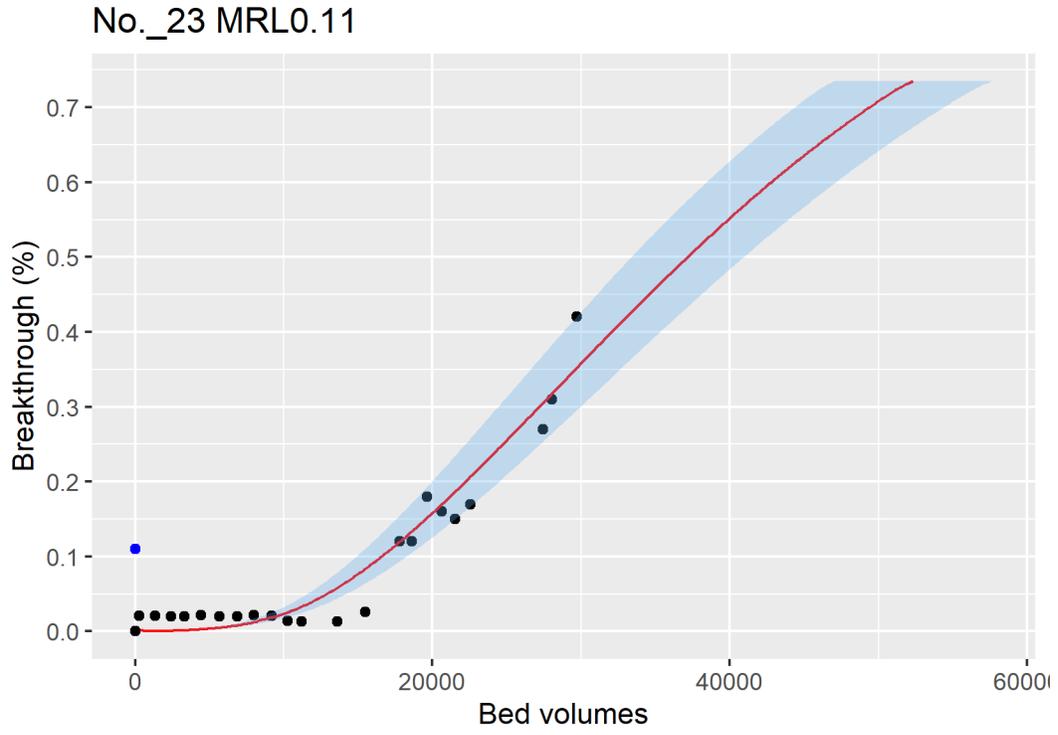


Figure ID 19- 22

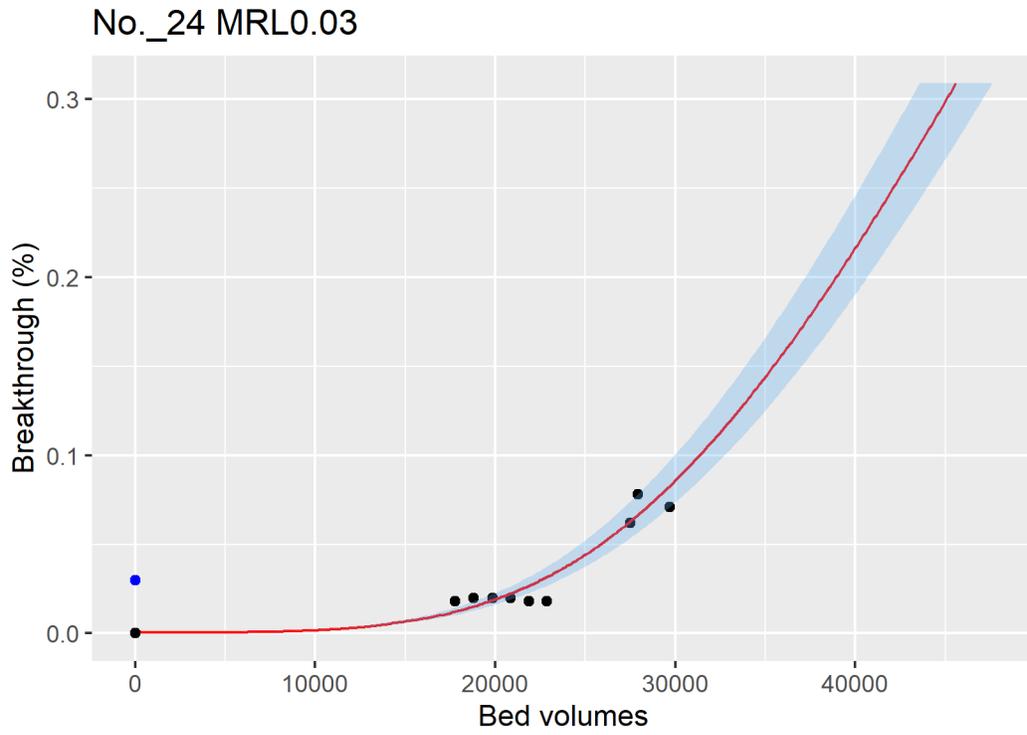


Figure ID 19- 23

No._25 MRL0.22

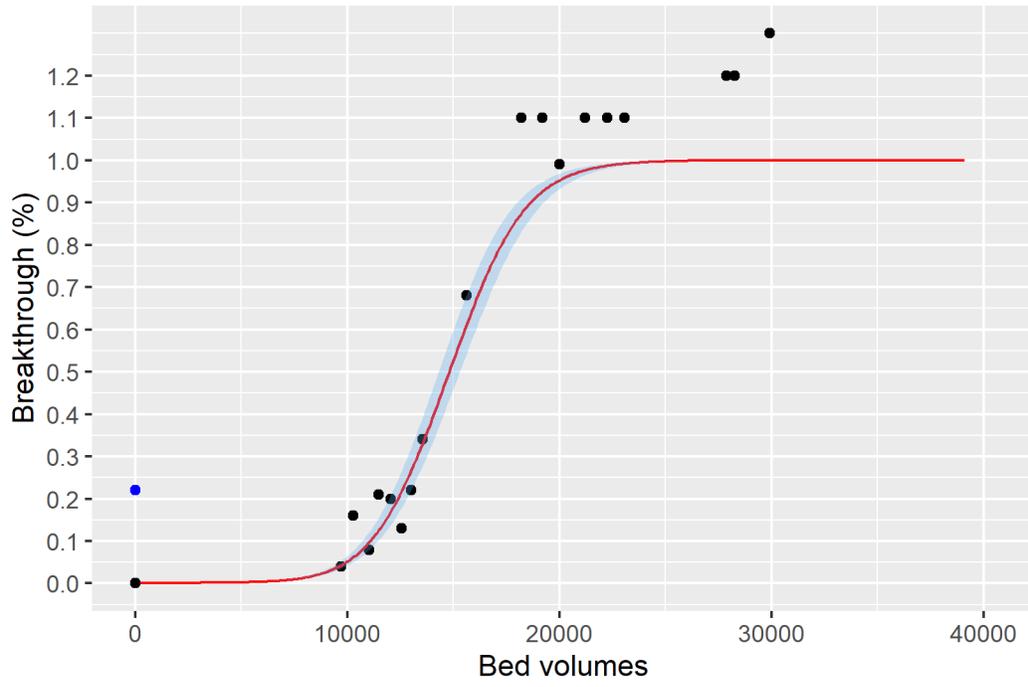


Figure ID 19- 24

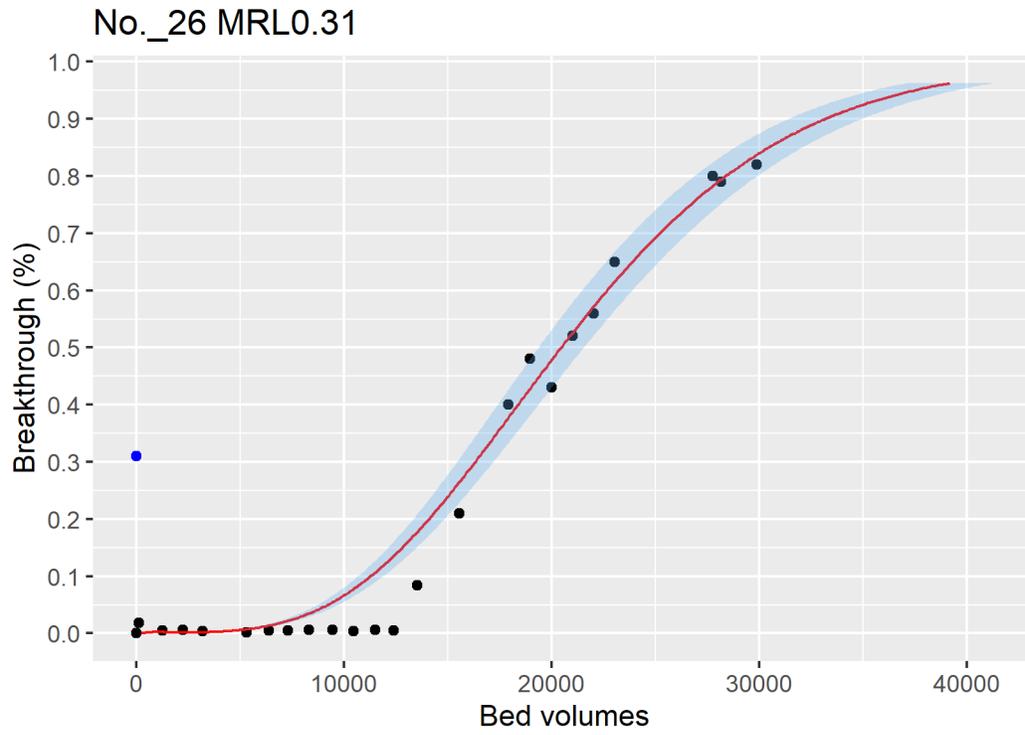


Figure ID 19- 25

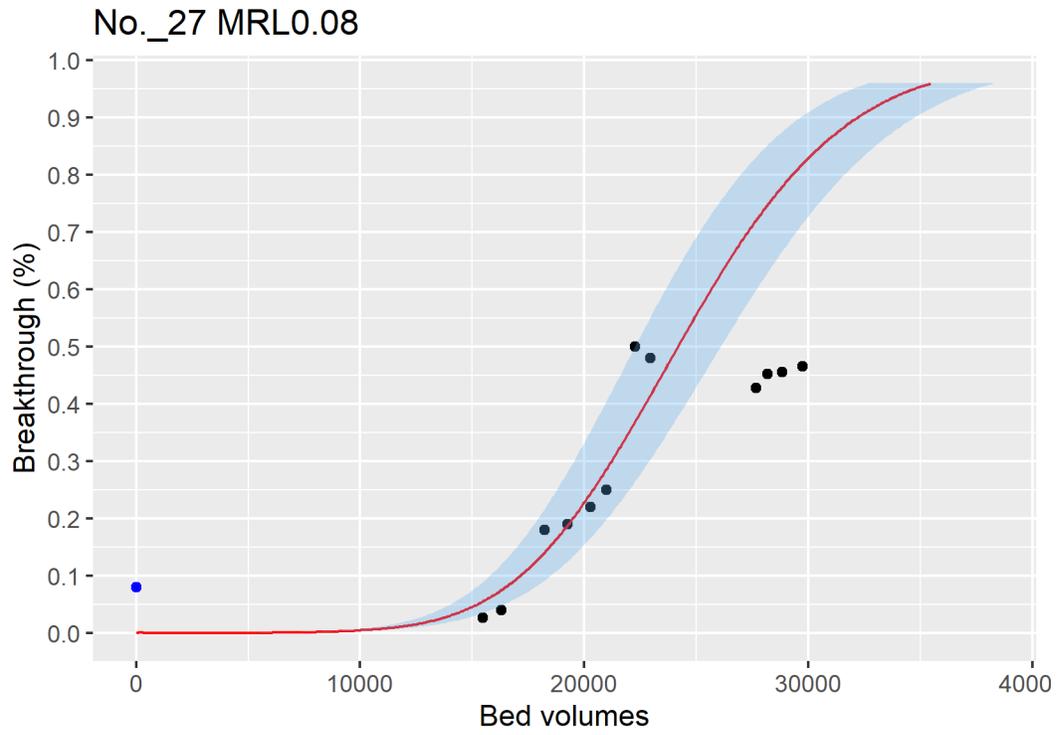


Figure ID 19- 26

No._28 MRL0.02

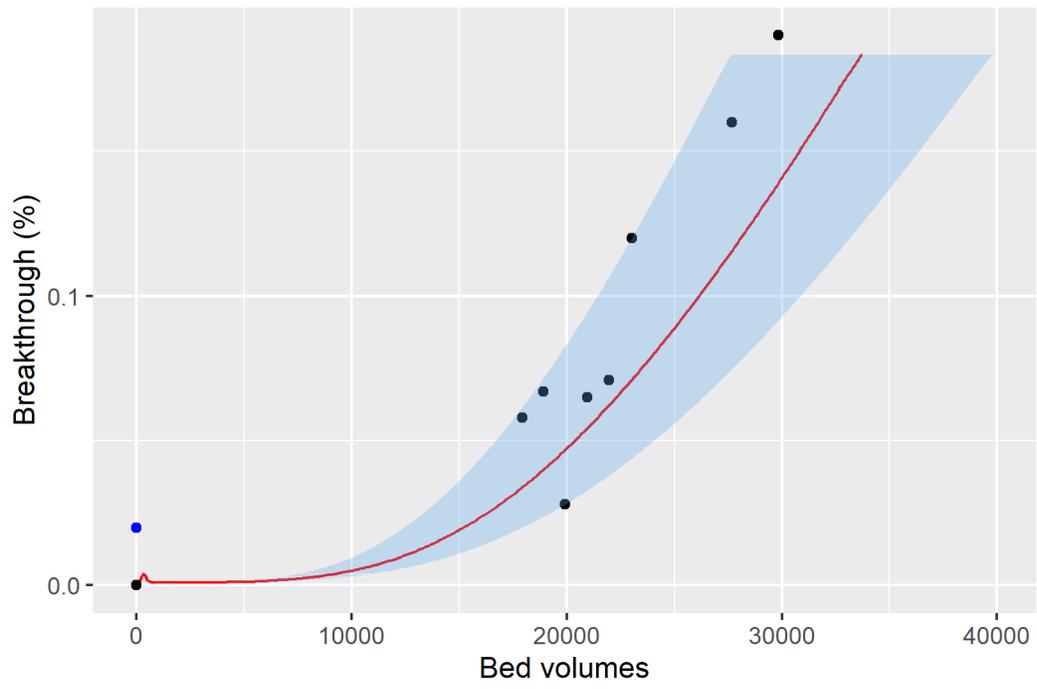


Figure ID 19- 27

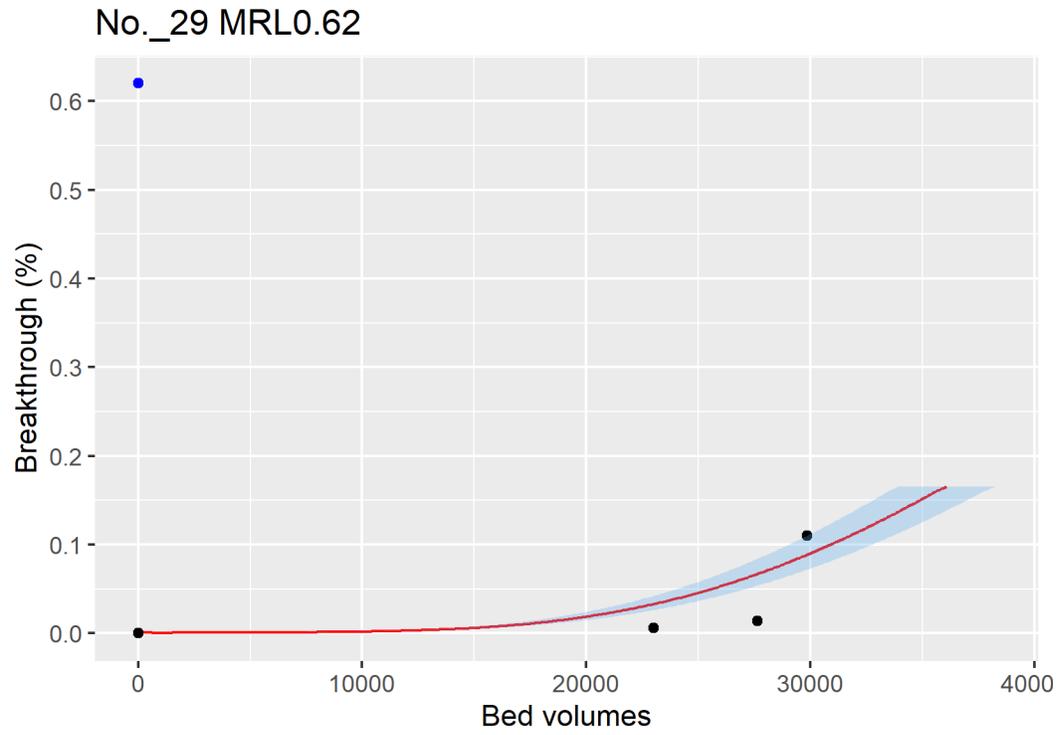


Figure ID 19- 28

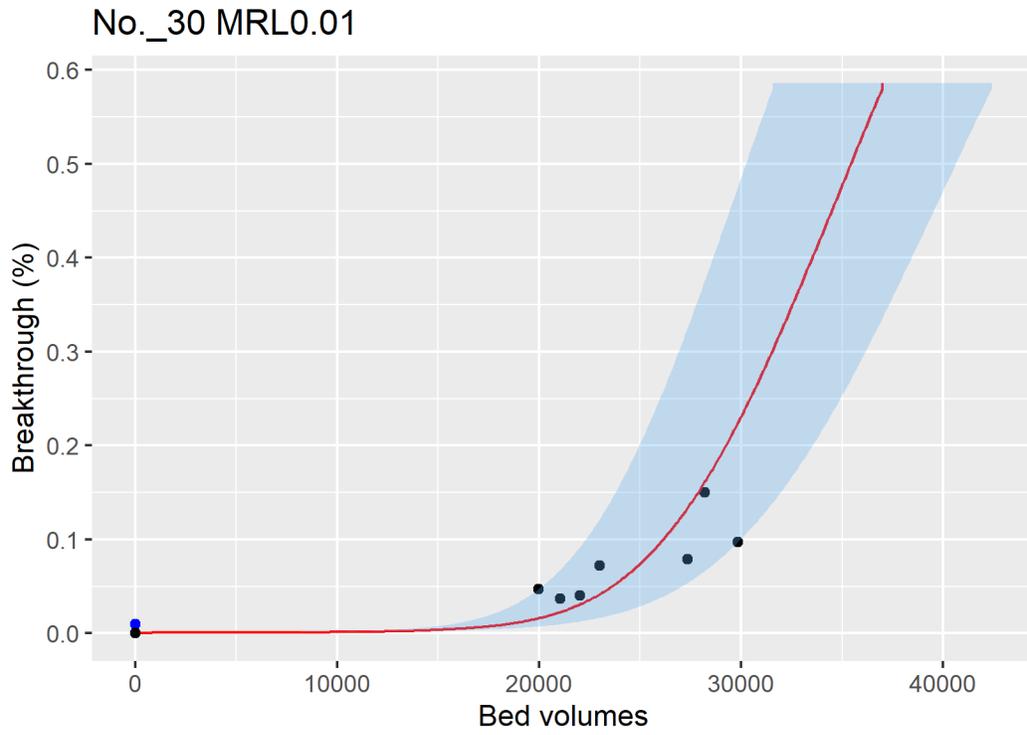


Figure ID 19- 29

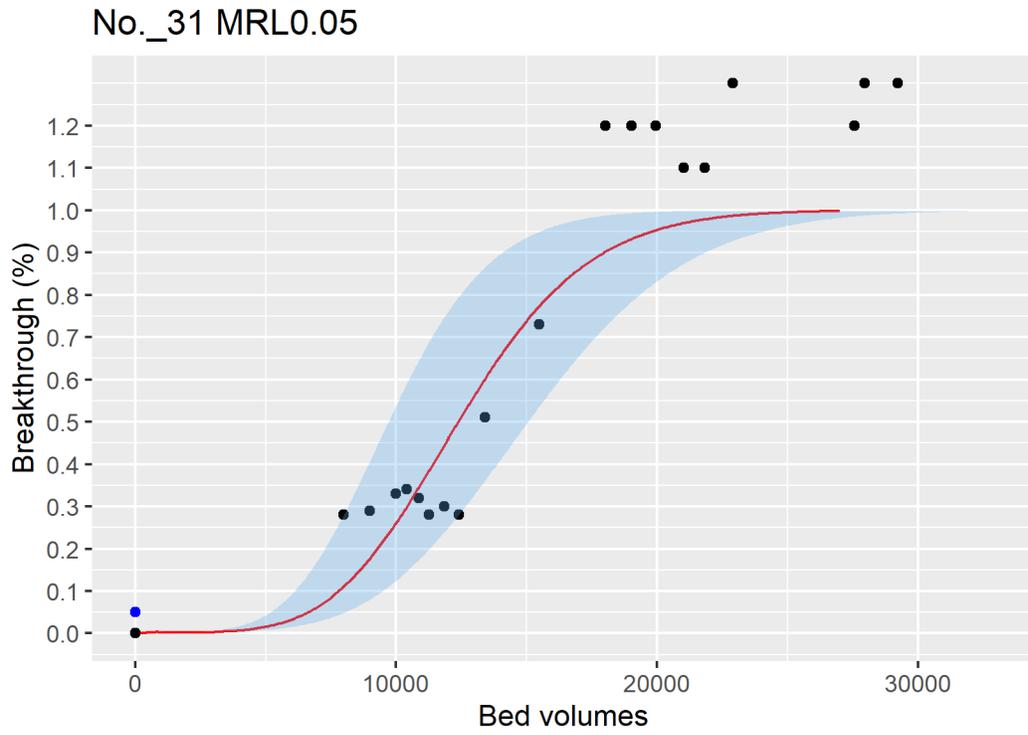


Figure ID 19- 30

No._32 MRL0.04

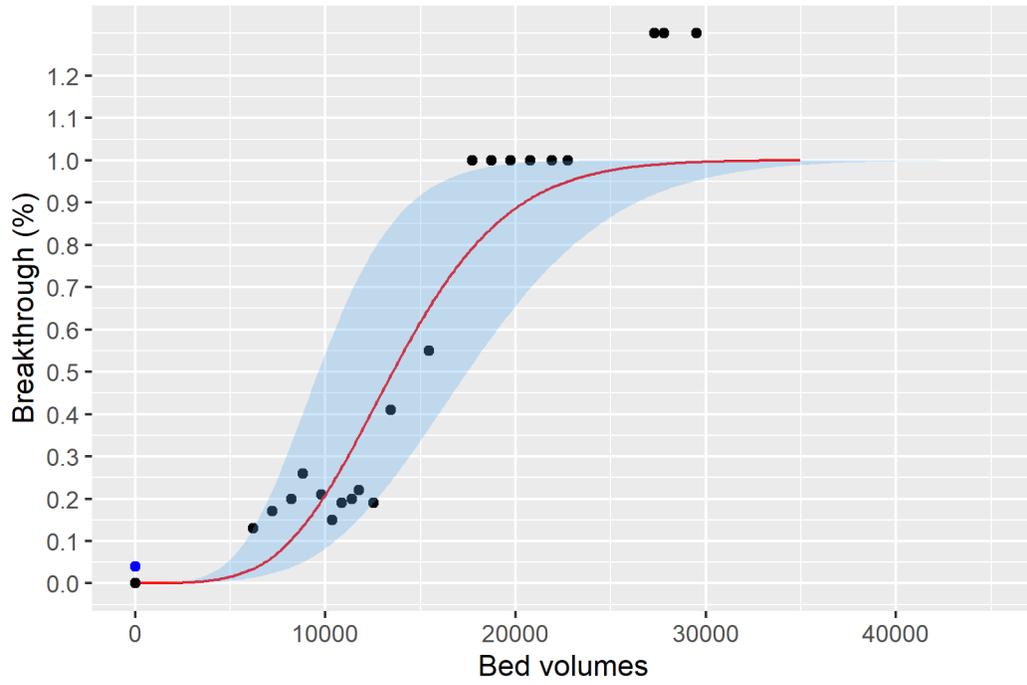


Figure ID 19- 31

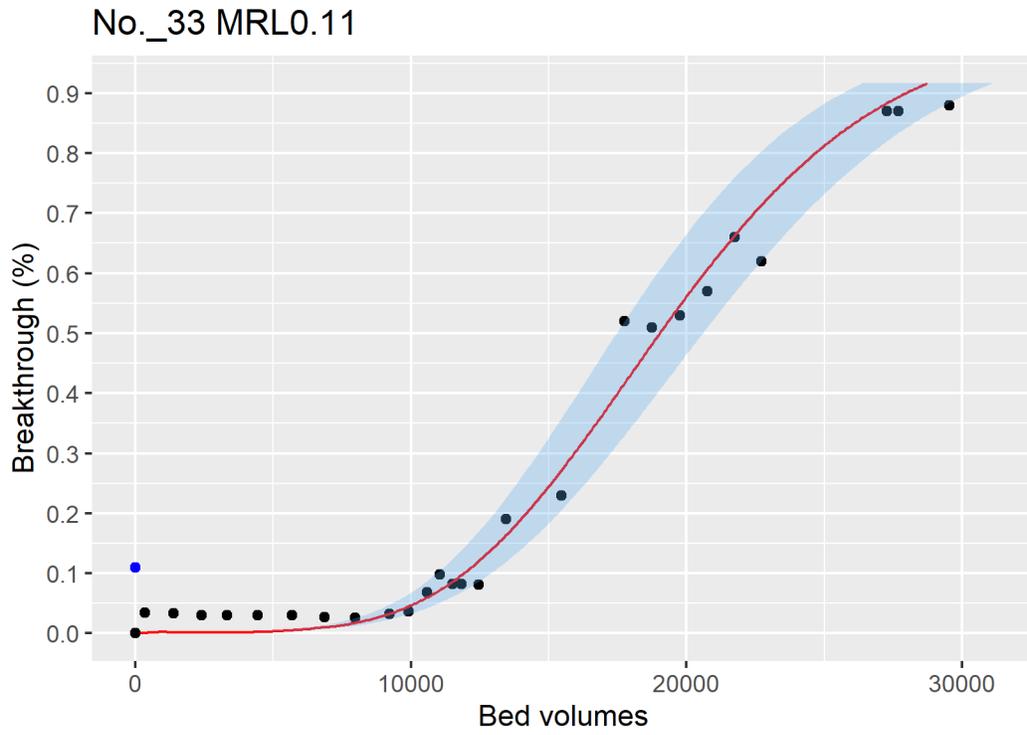


Figure ID 19- 32

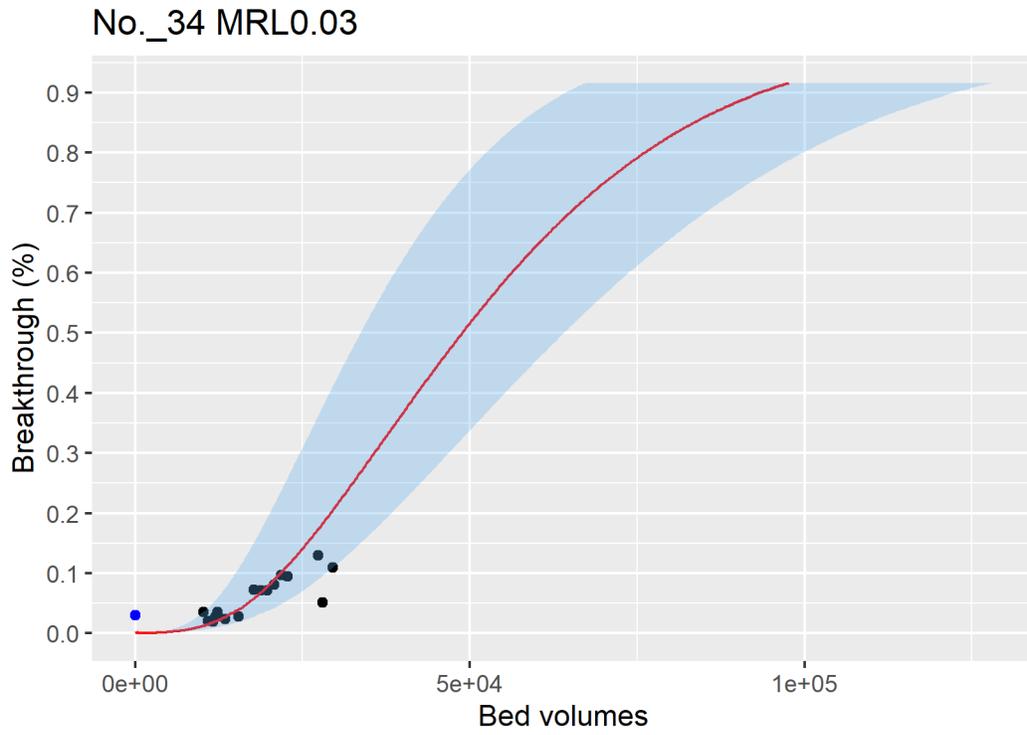


Figure ID 19- 33

No._35 MRL0.22

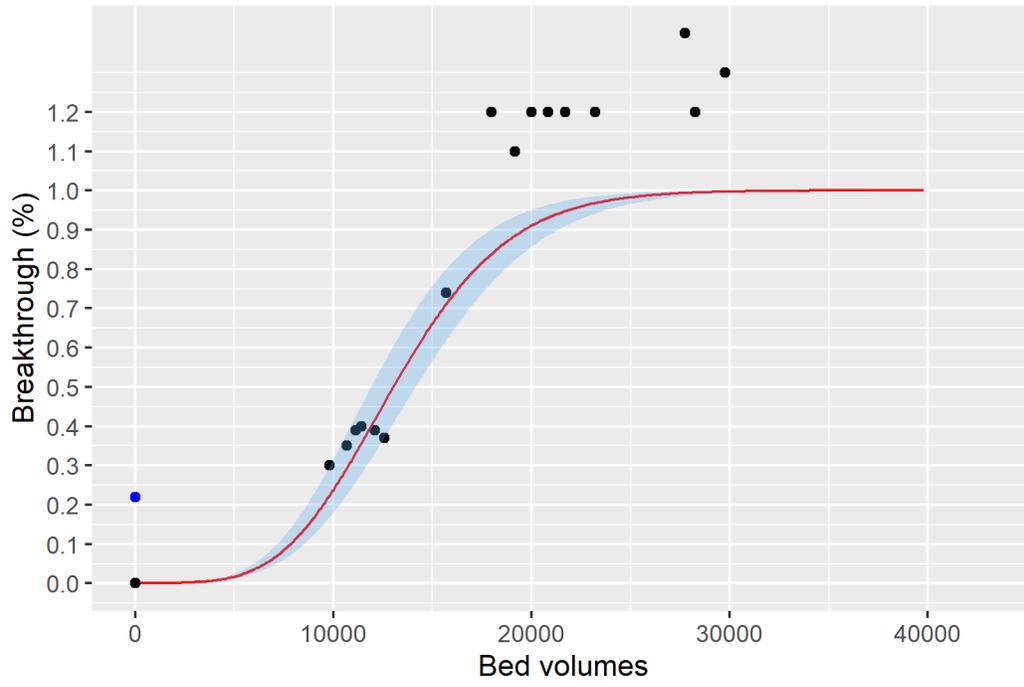


Figure ID 19- 34

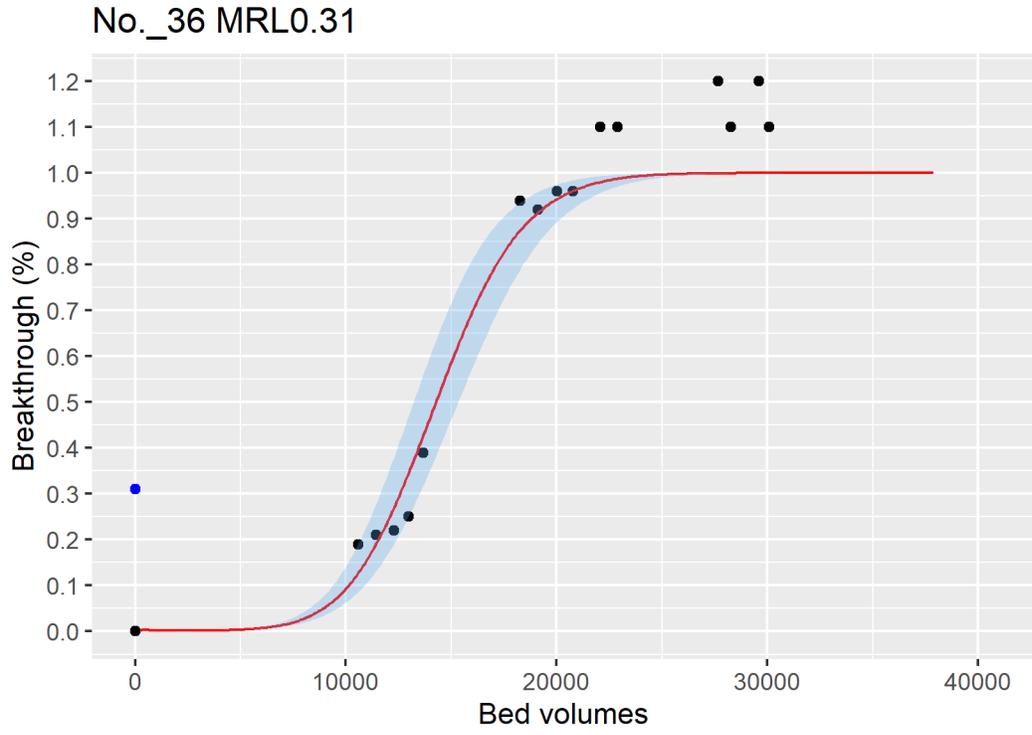


Figure ID 19- 35

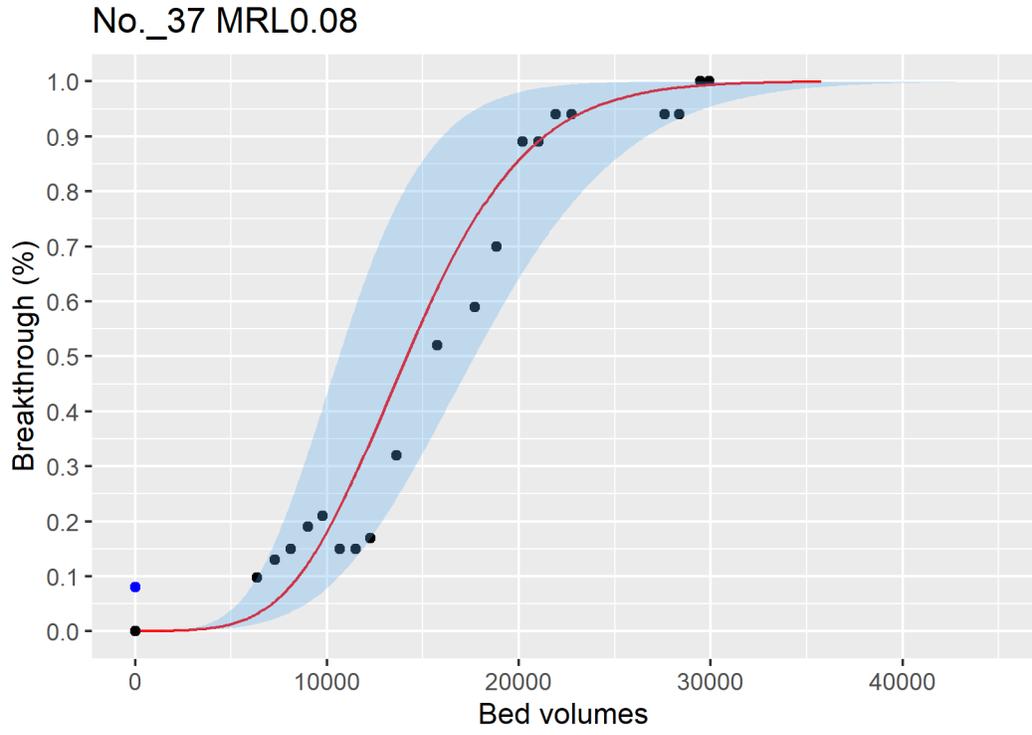


Figure ID 19- 36

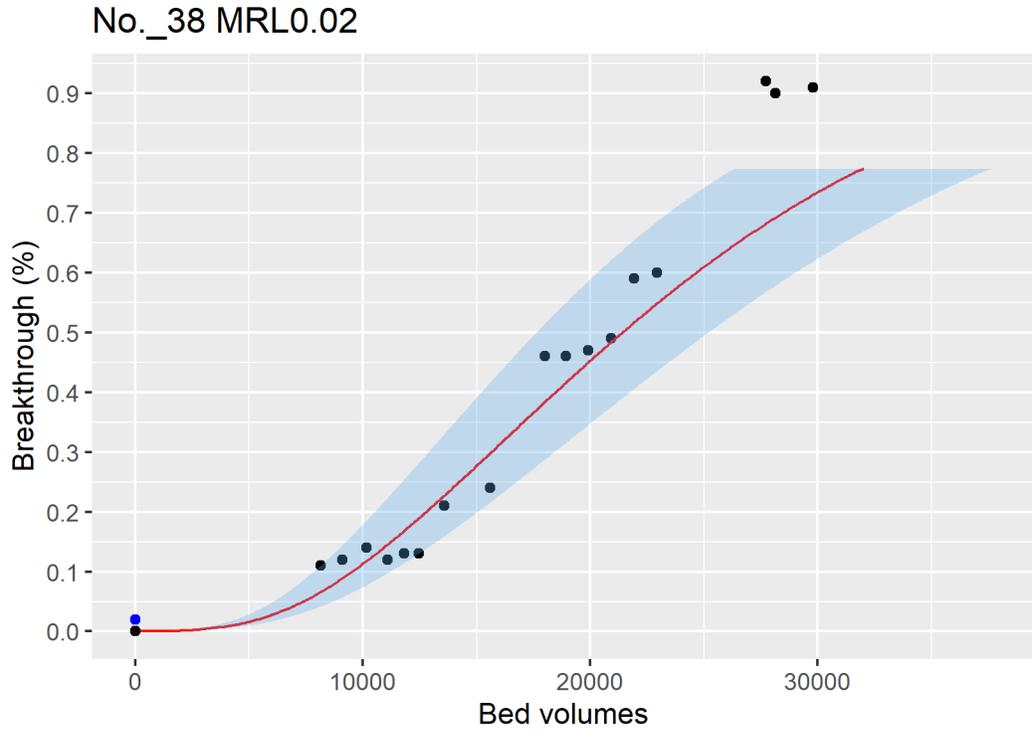


Figure ID 19- 37

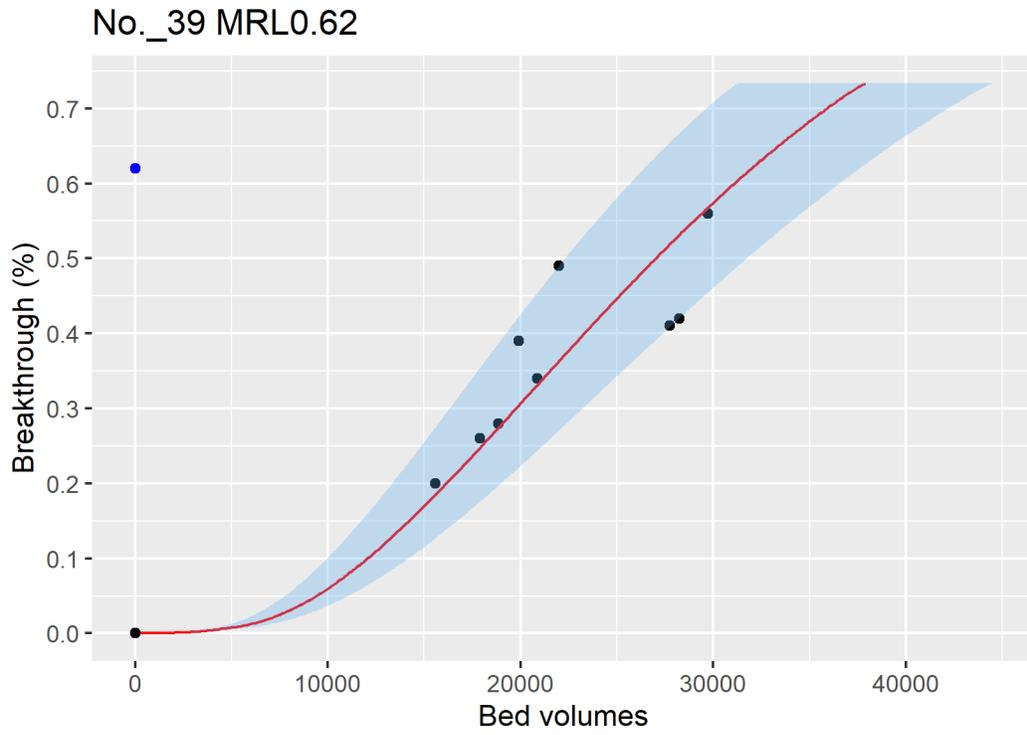


Figure ID 19- 38

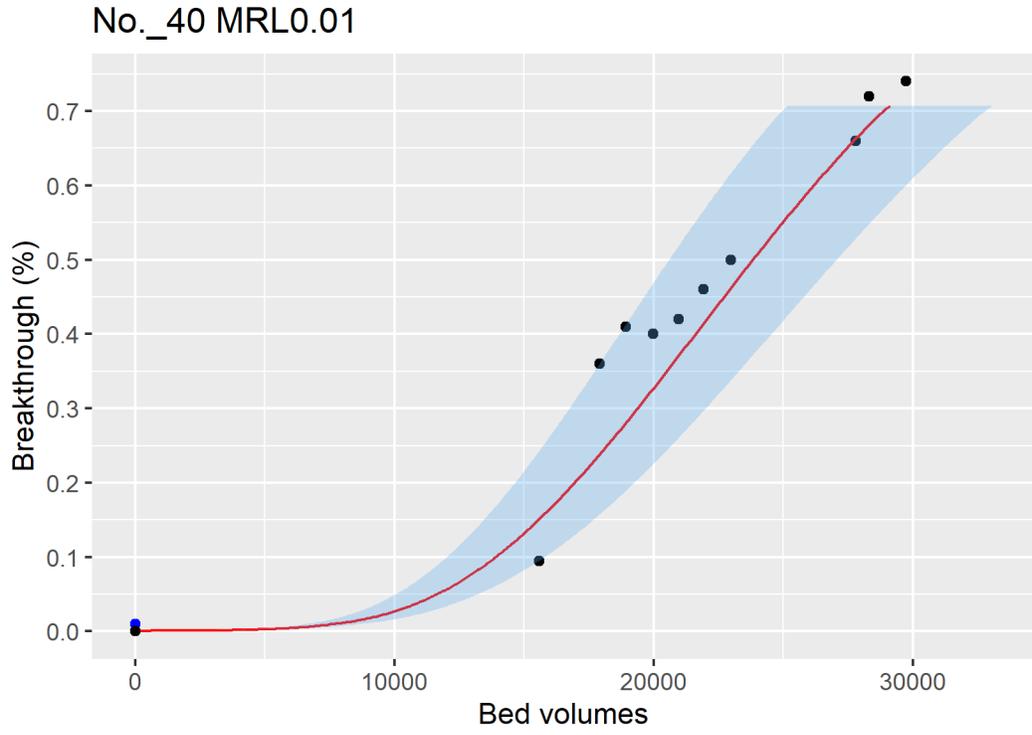


Figure ID 19- 39

Breakthrough data of Ref ID 35

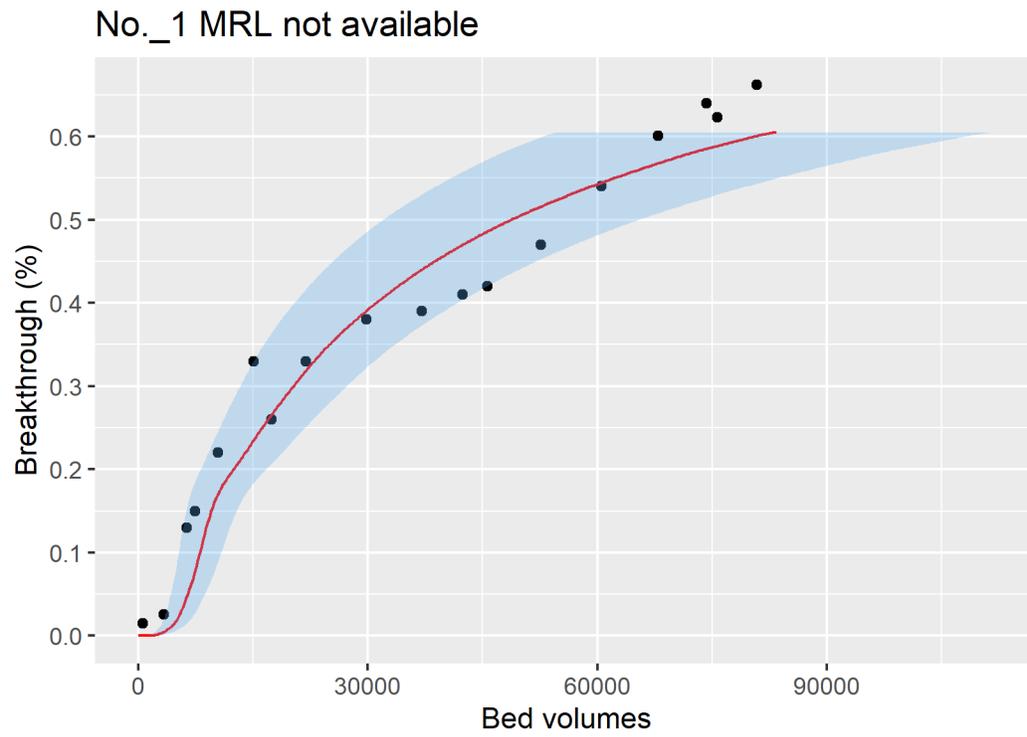


Figure ID 35- 1

No._2 MRL not available

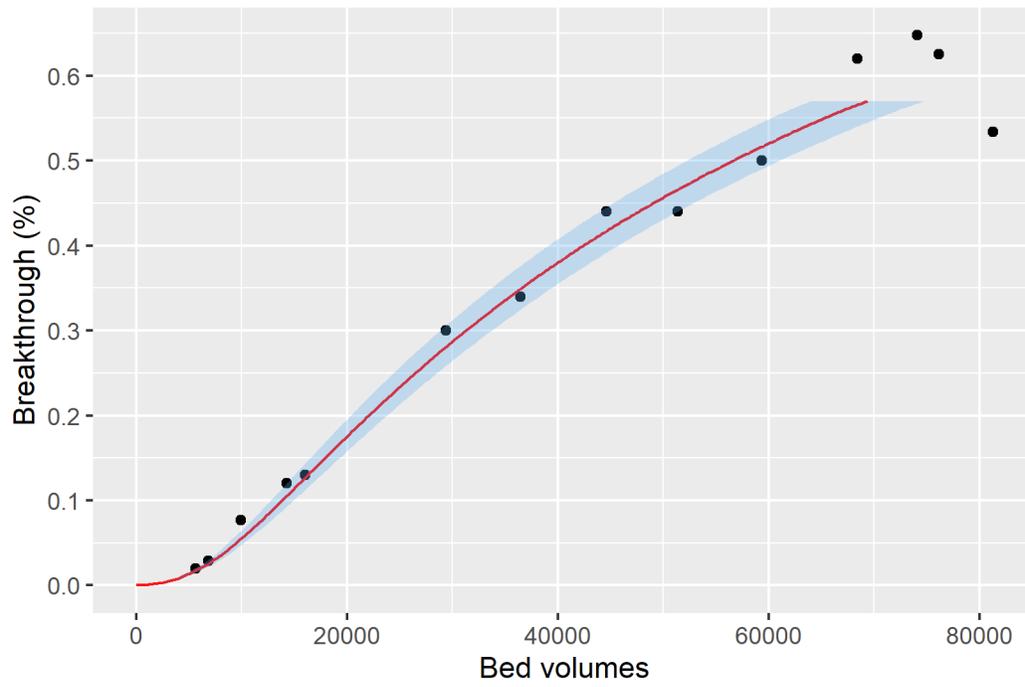


Figure ID 35- 2

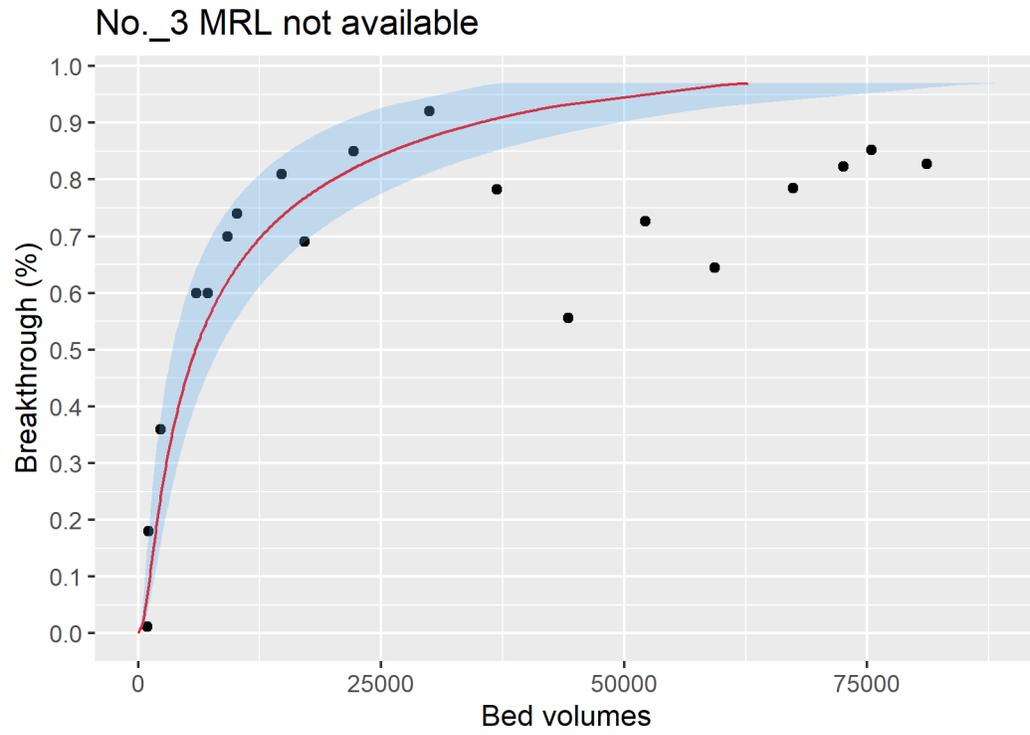


Figure ID 35- 3

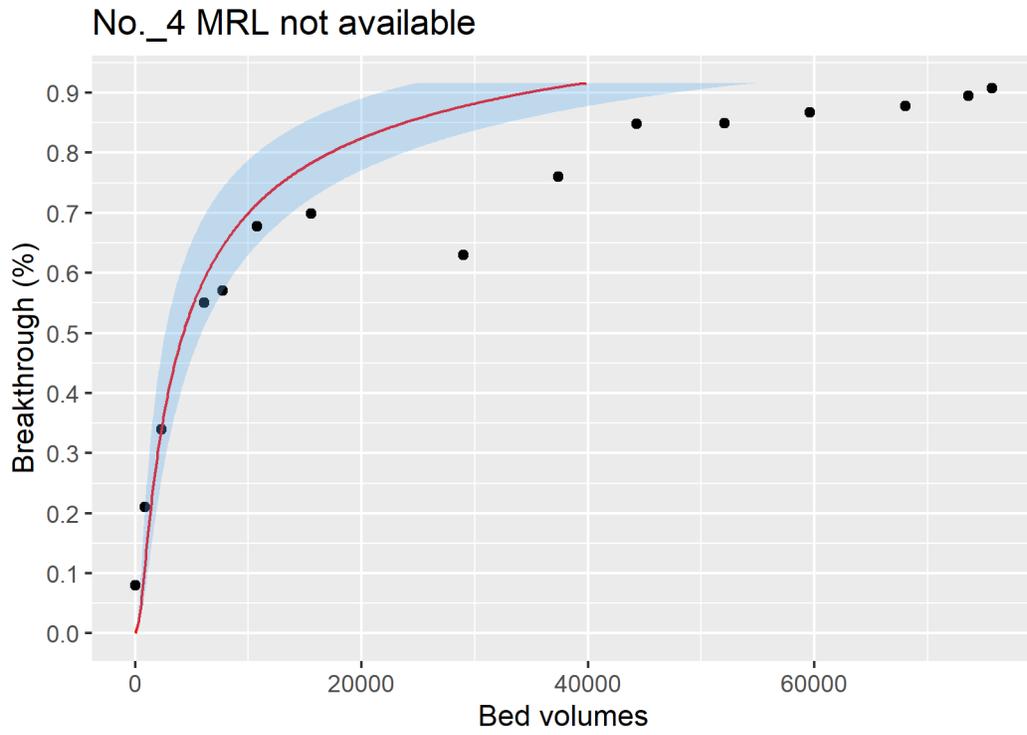


Figure ID 35- 4

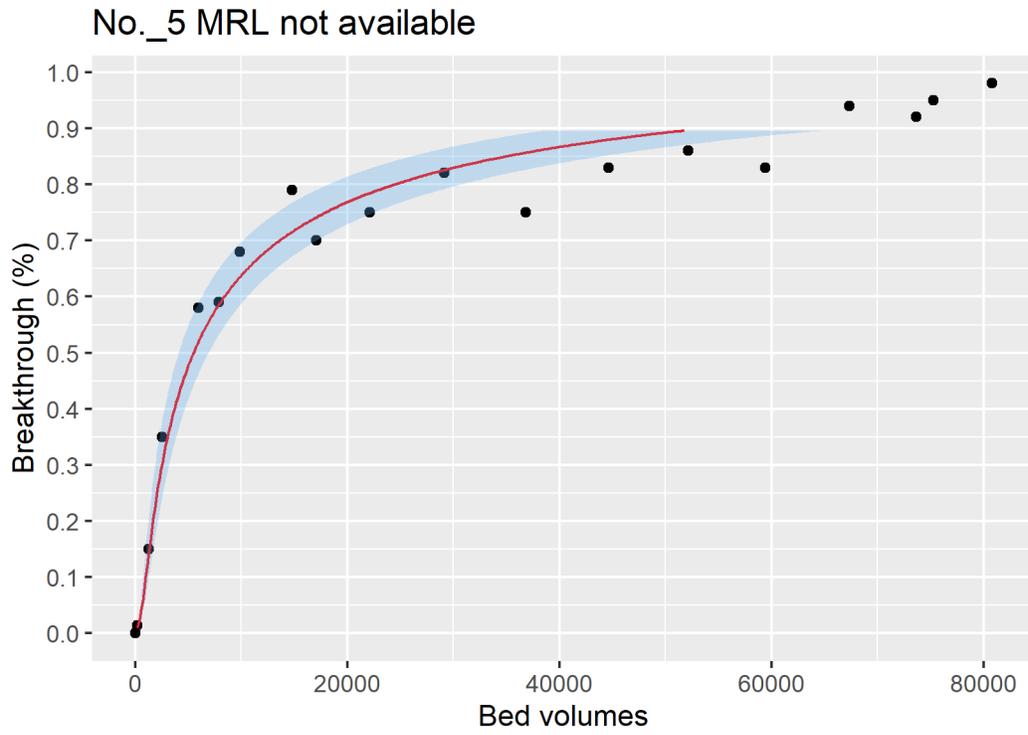


Figure ID 35- 5

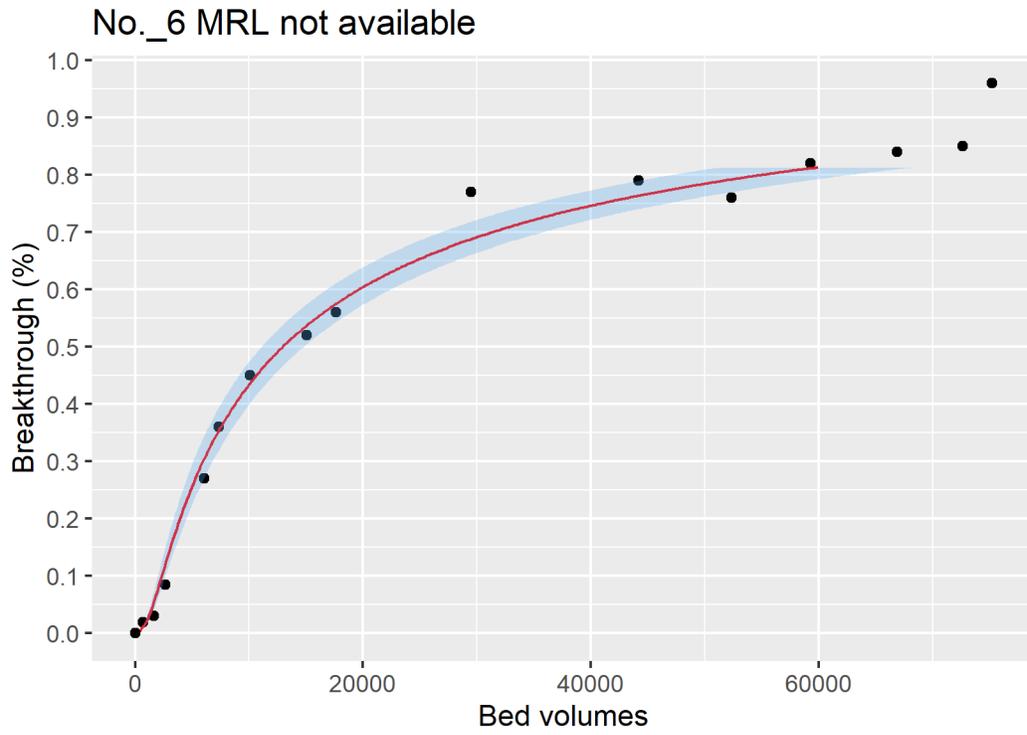


Figure ID 35- 6

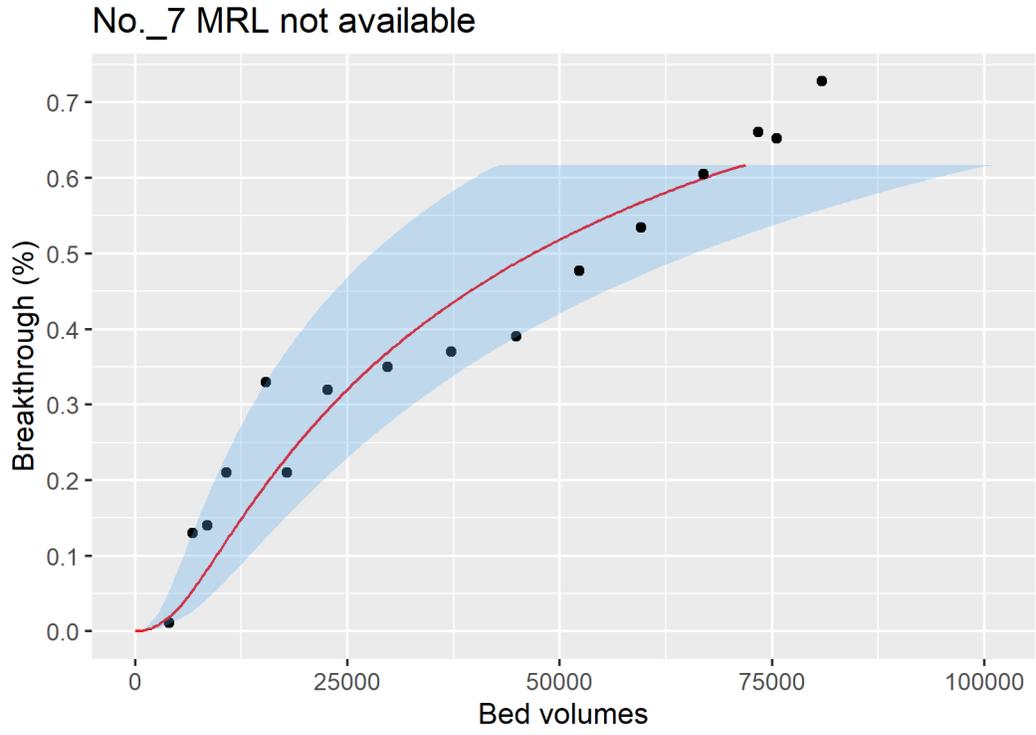


Figure ID 35- 7

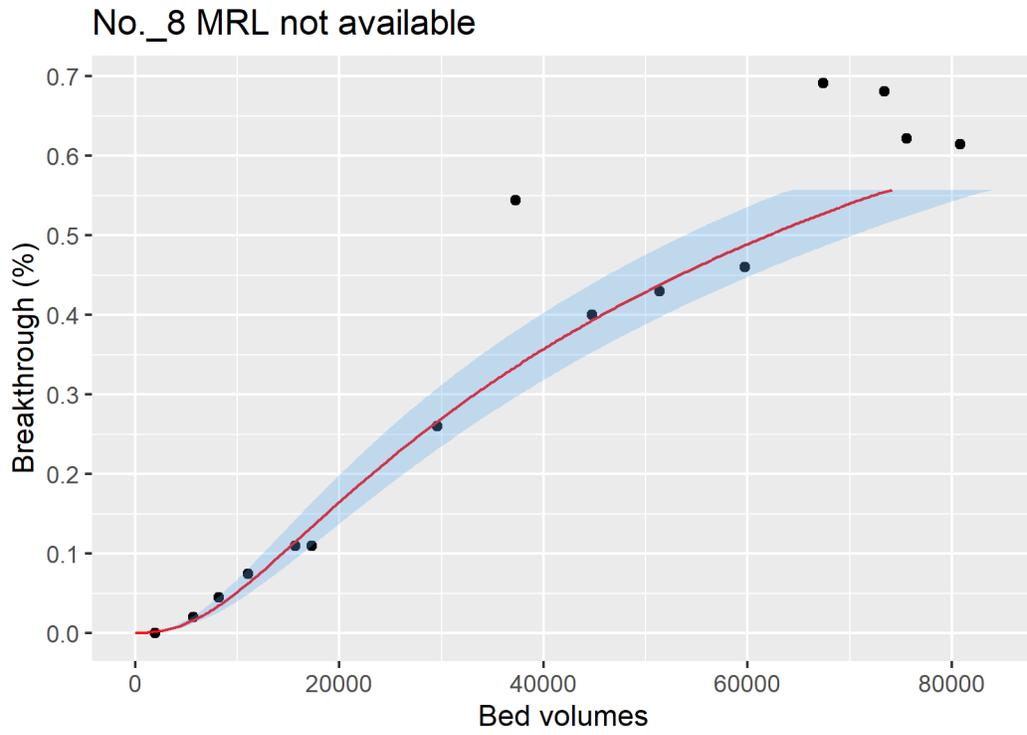


Figure ID 35- 8

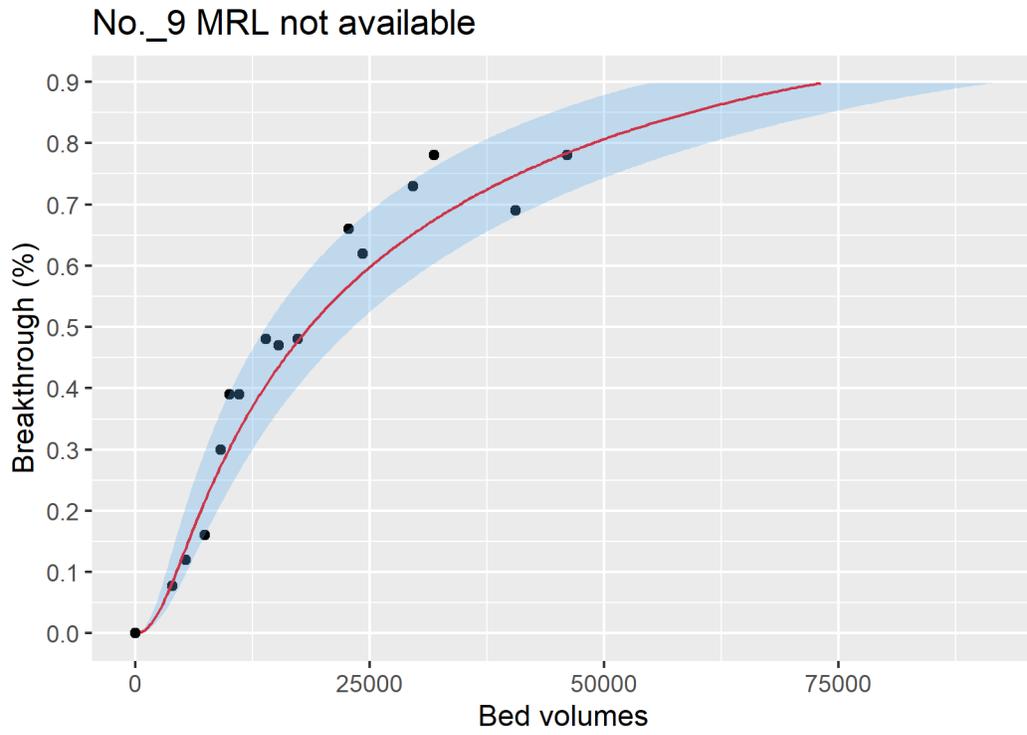


Figure ID 35- 9

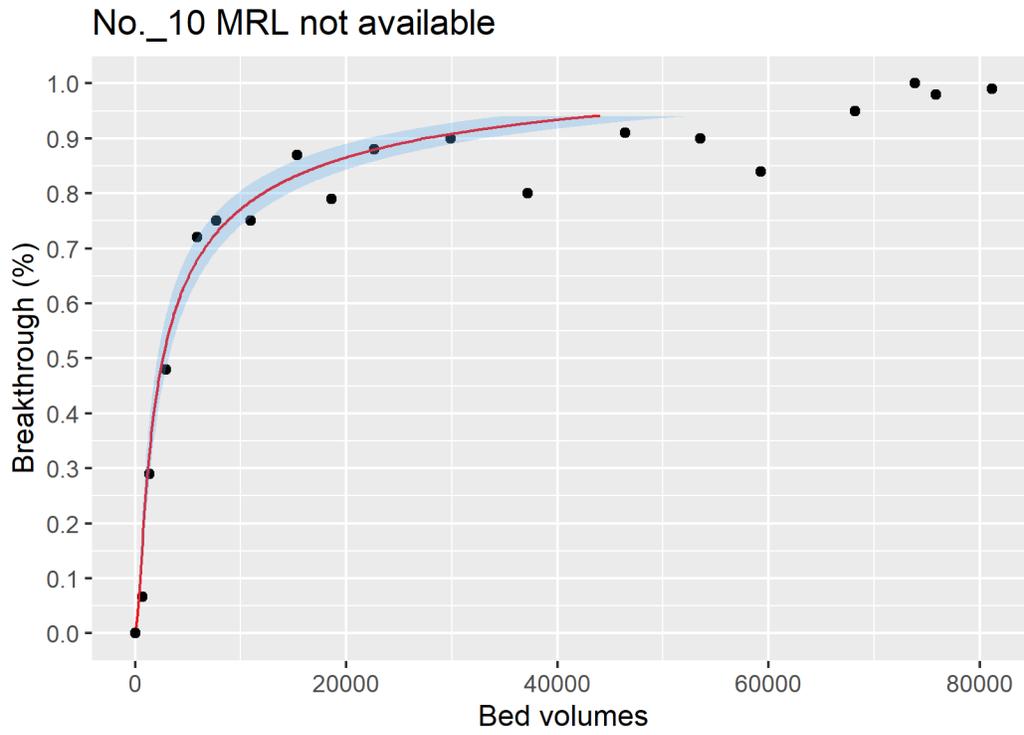


Figure ID 35- 10

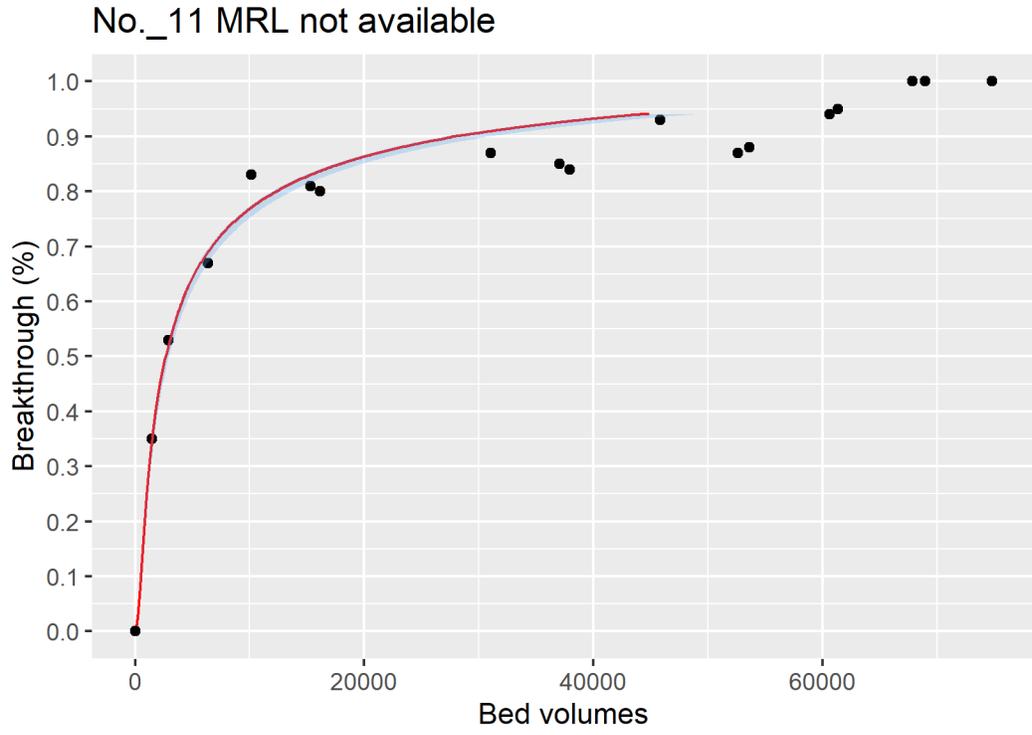


Figure ID 35- 11

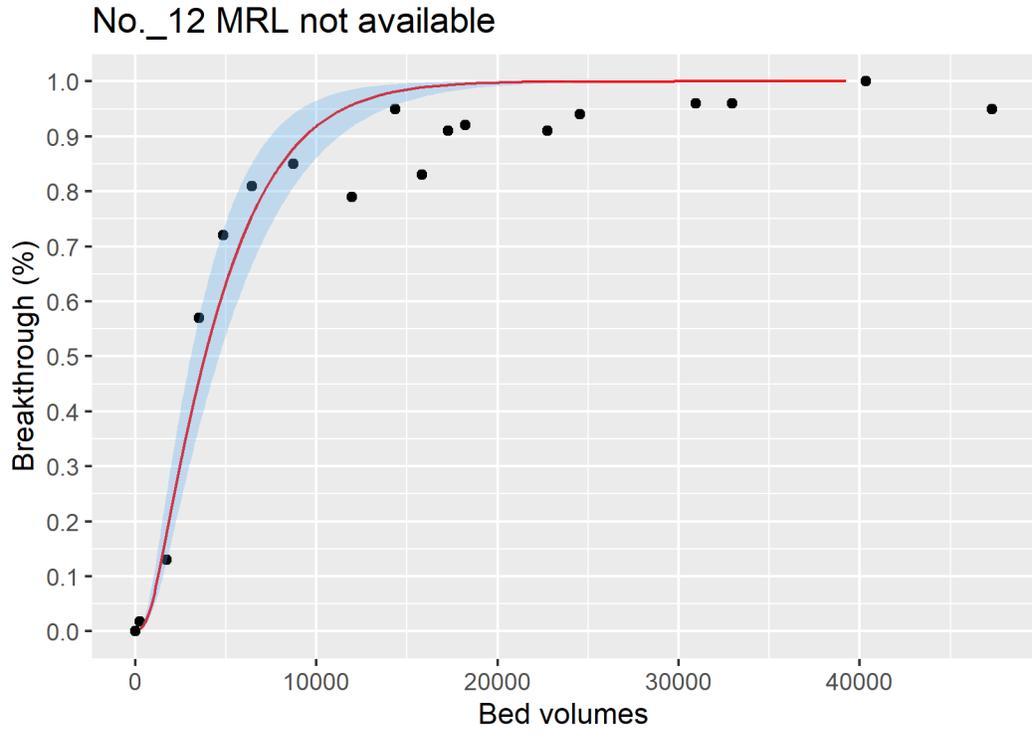


Figure ID 35- 12

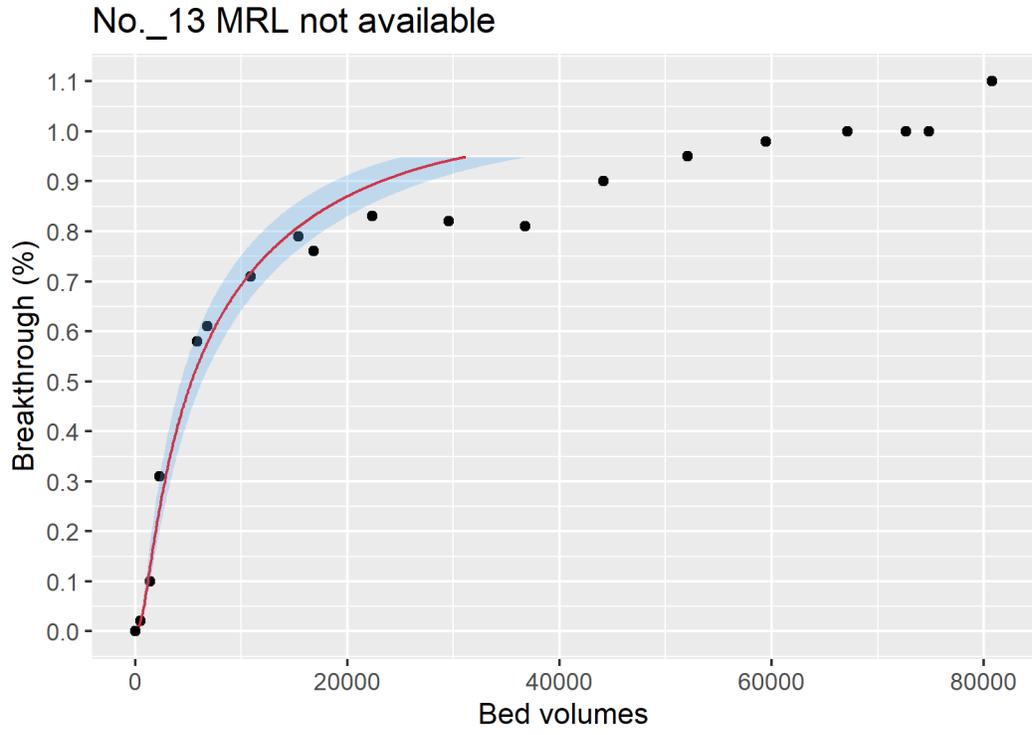


Figure ID 35- 13

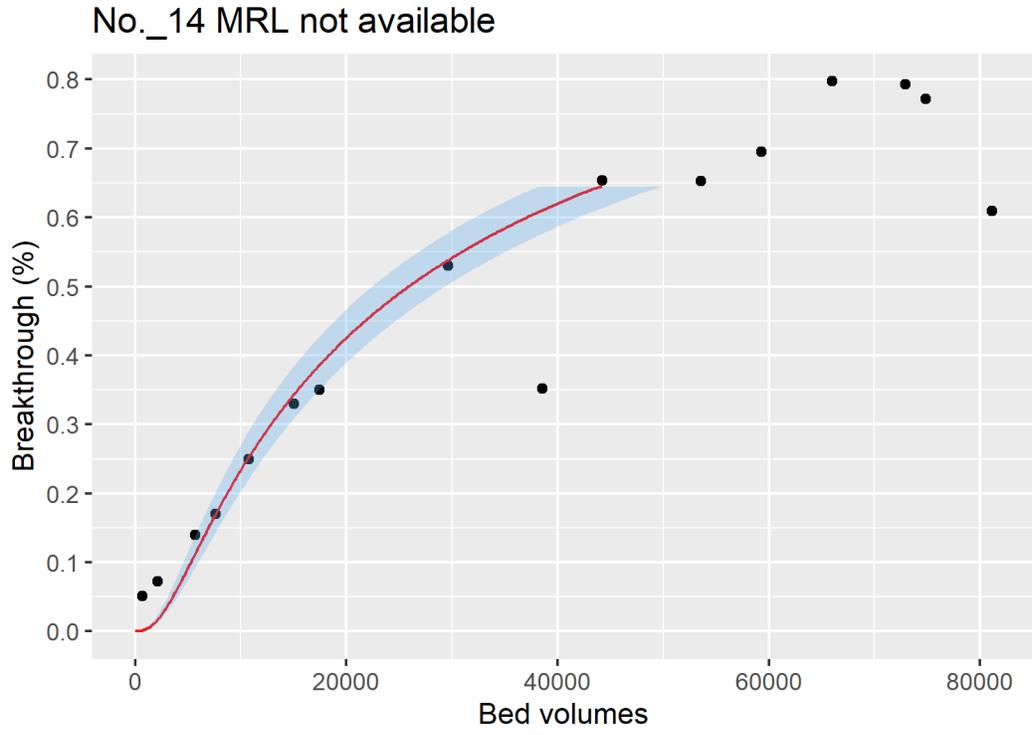


Figure ID 35- 14

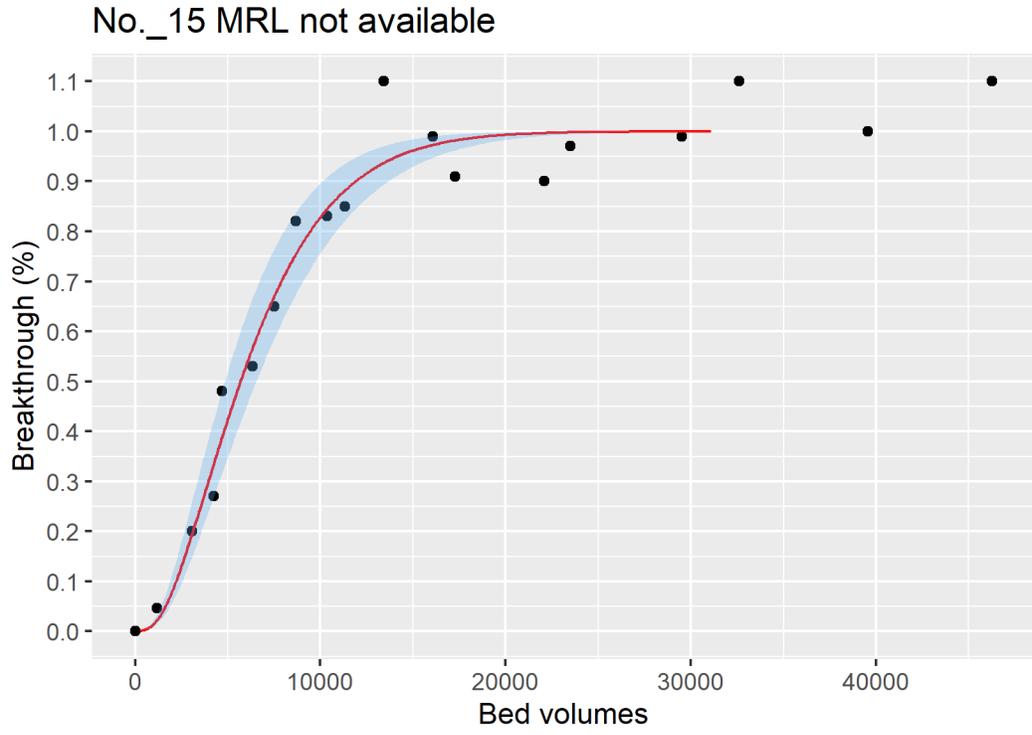


Figure ID 35- 15

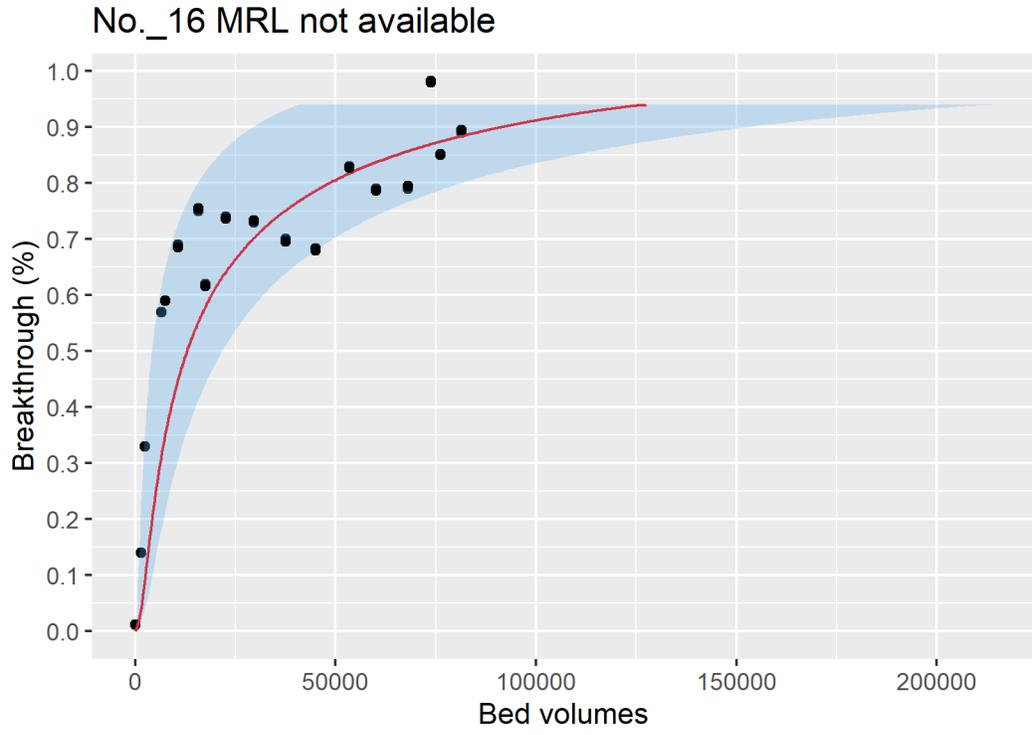


Figure ID 35- 16

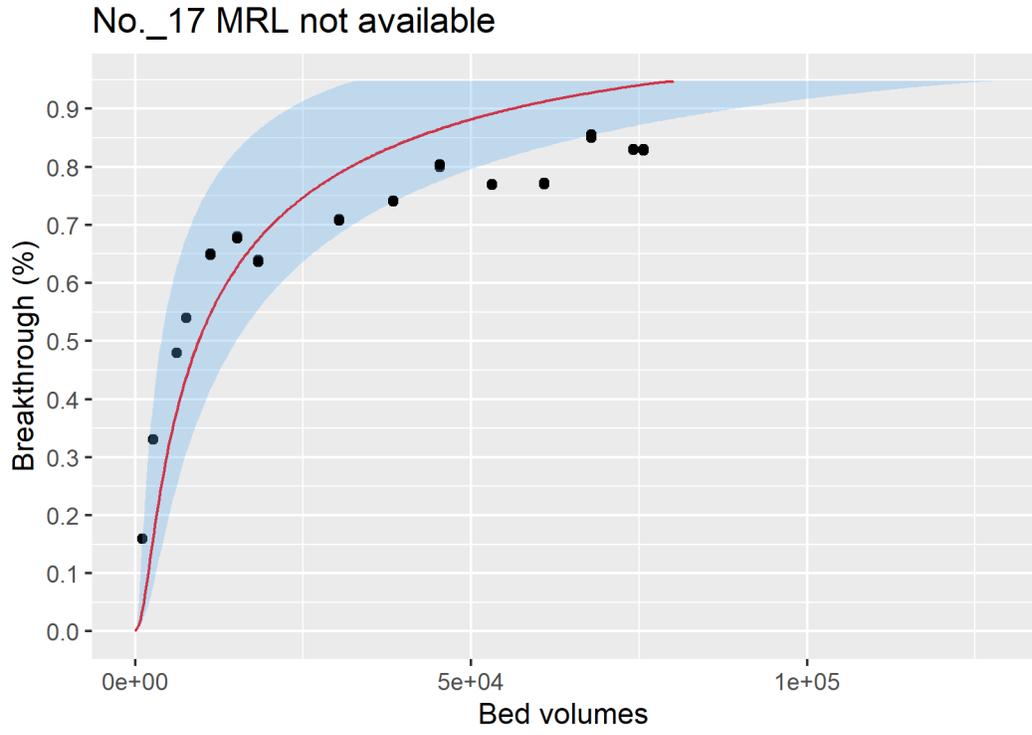


Figure ID 35- 17

Breakthrough data of Ref ID 40

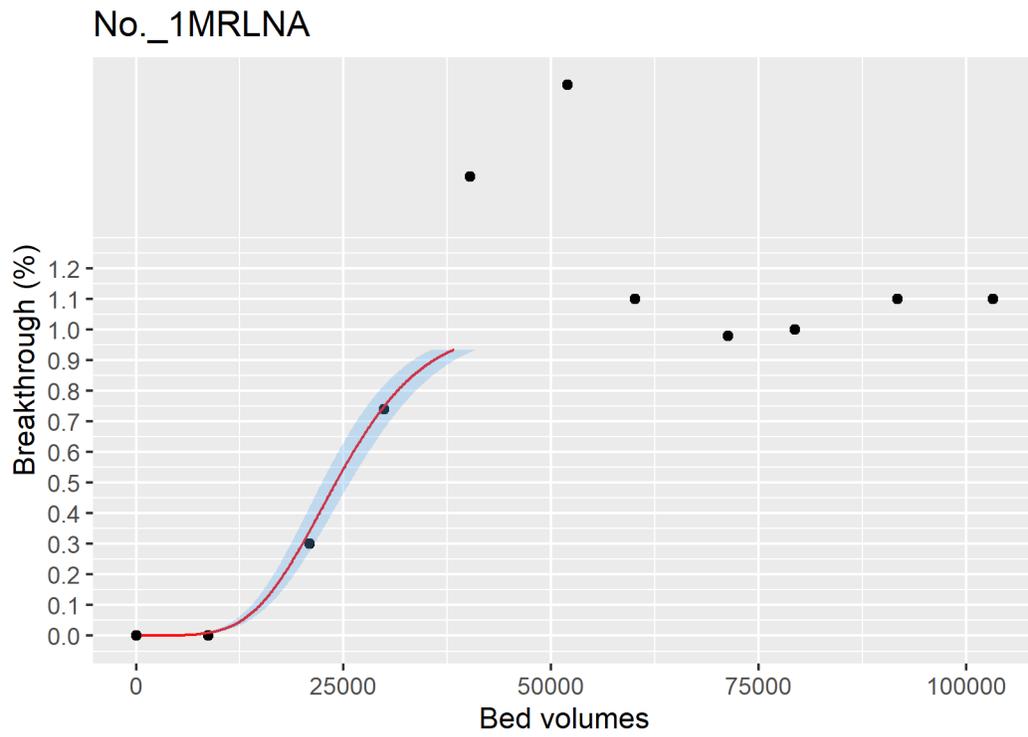


Figure ID 40- 1

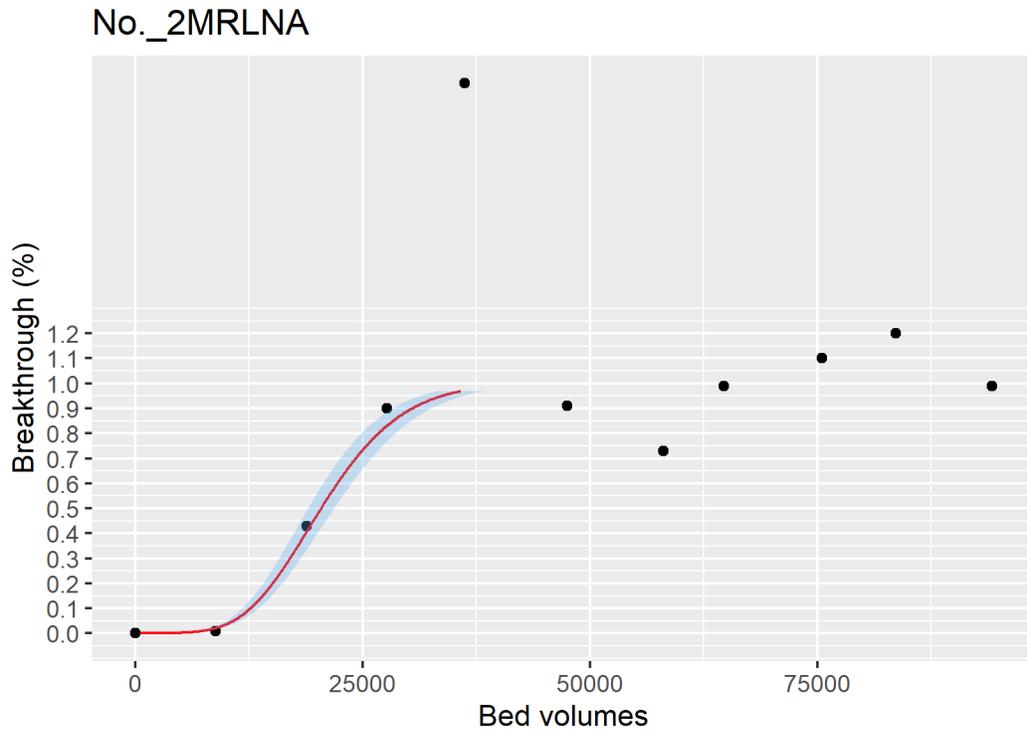


Figure ID 40- 2

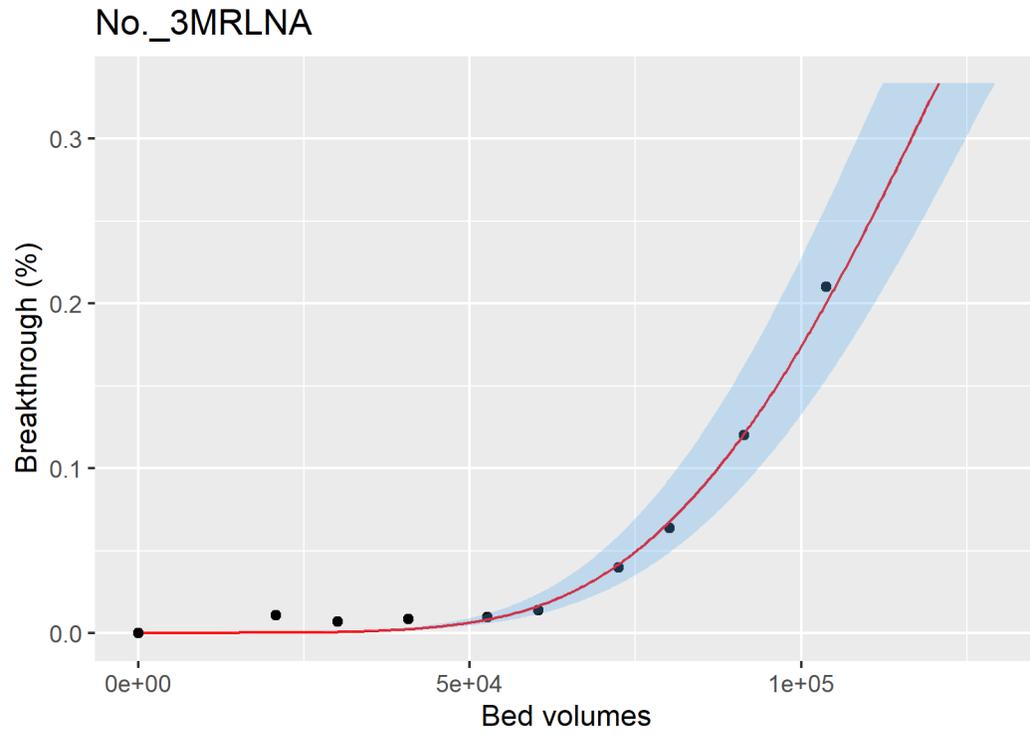


Figure ID 40- 3

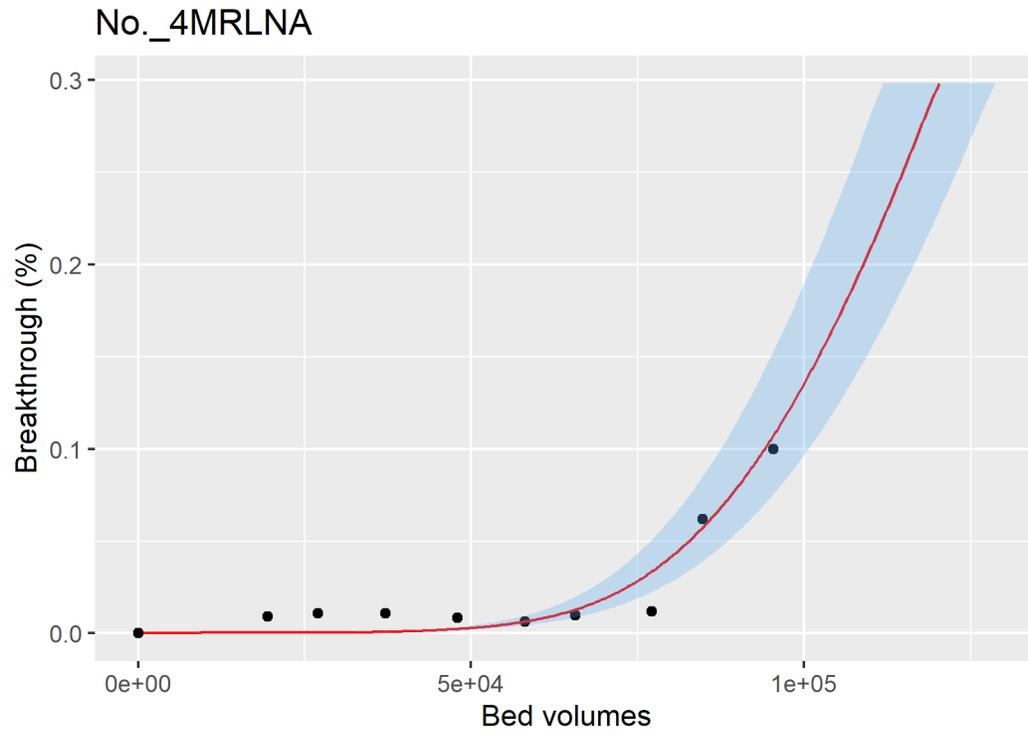


Figure ID 40- 4

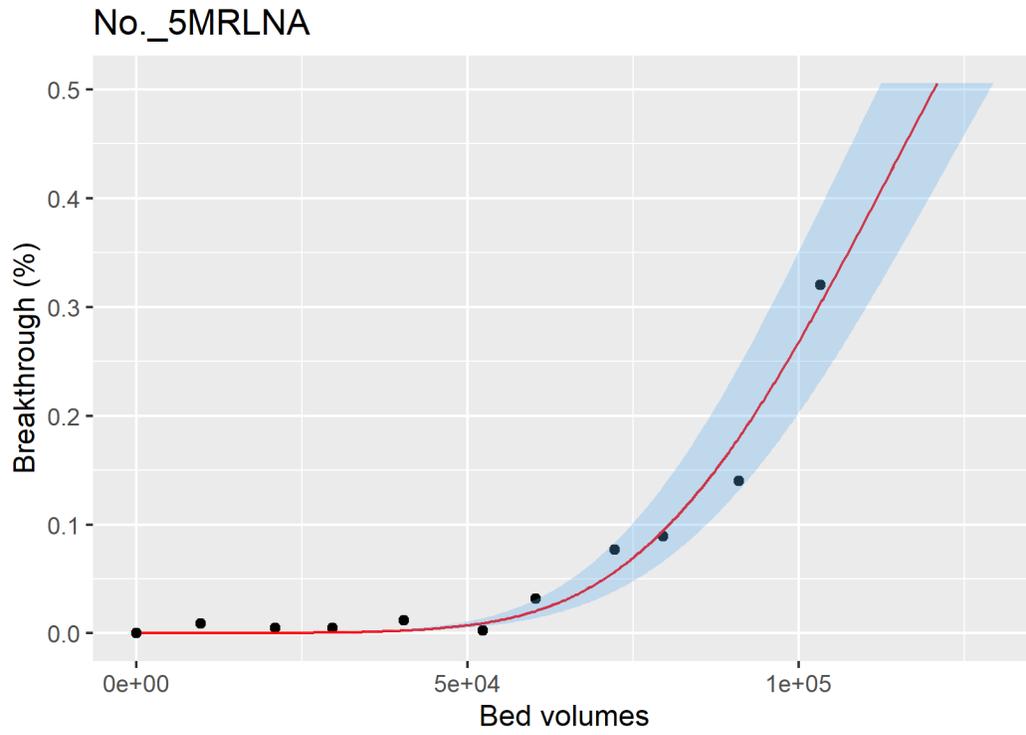


Figure ID 40- 5

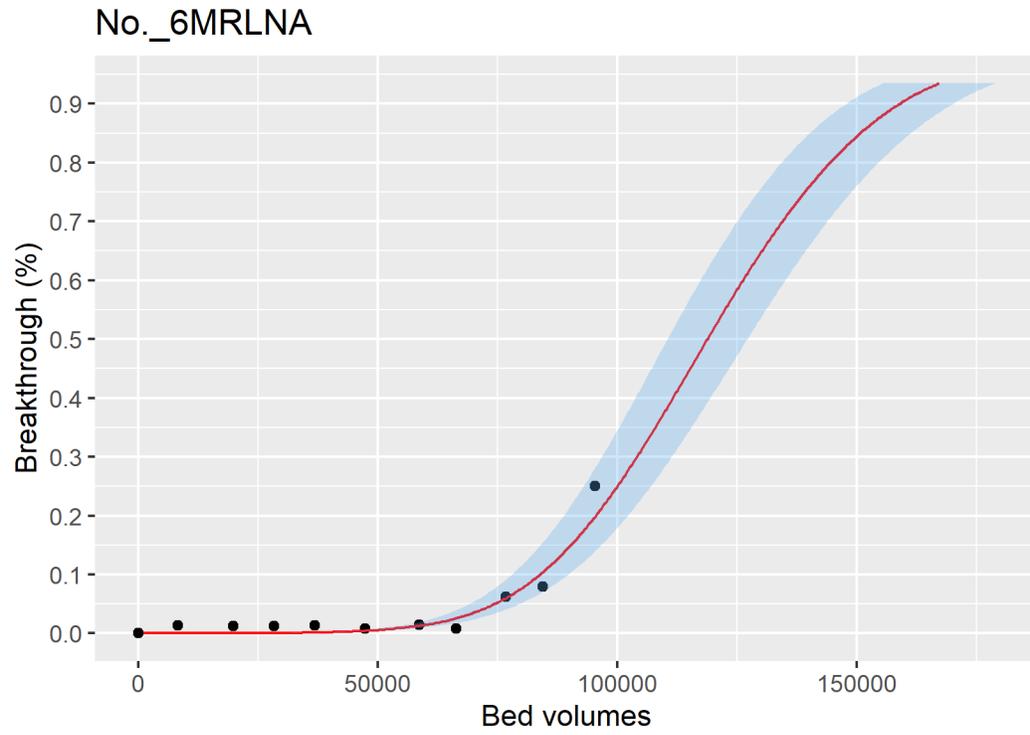


Figure ID 40- 6

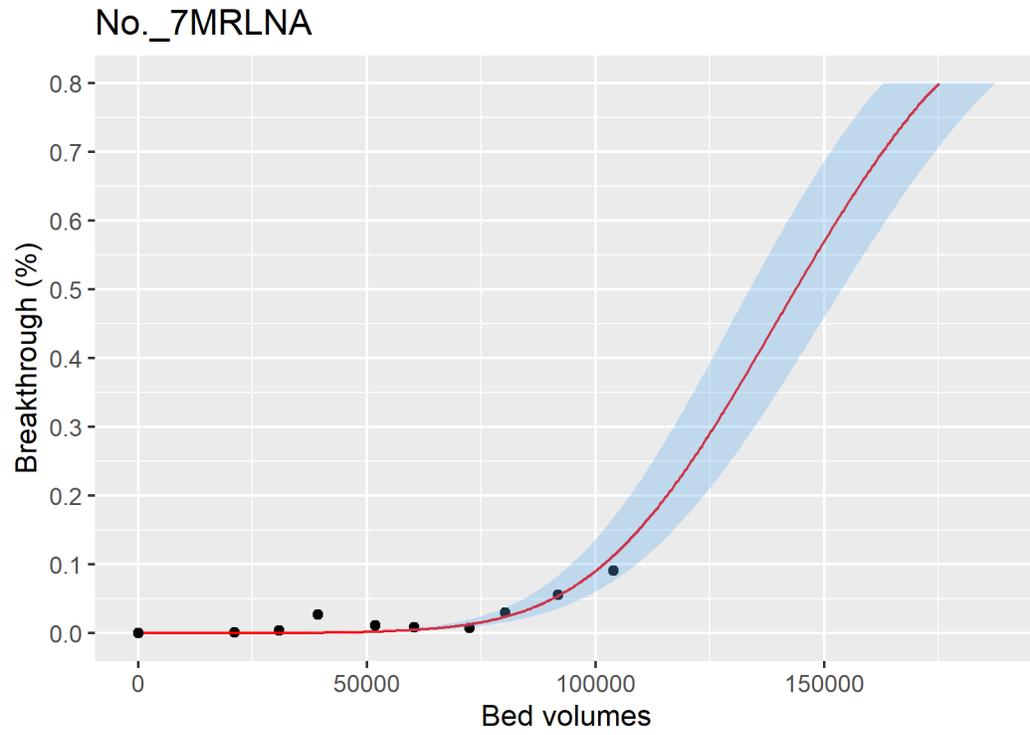


Figure ID 40- 7

Breakthrough data of Ref ID 47

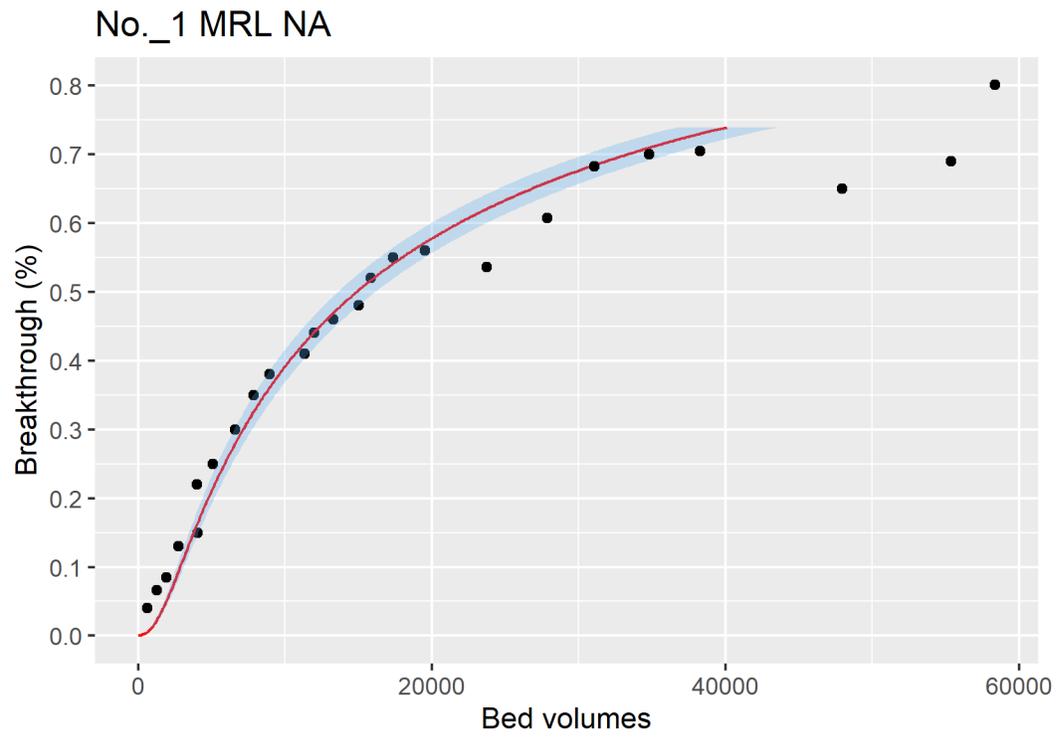


Figure ID 47- 1

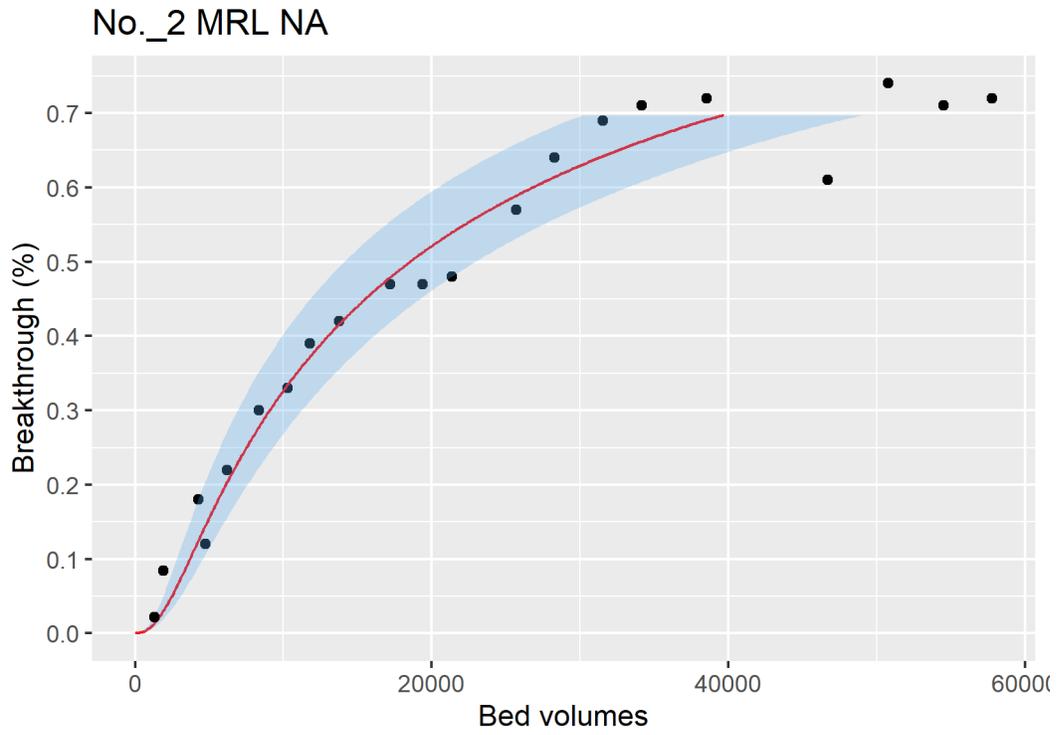


Figure ID 47- 2

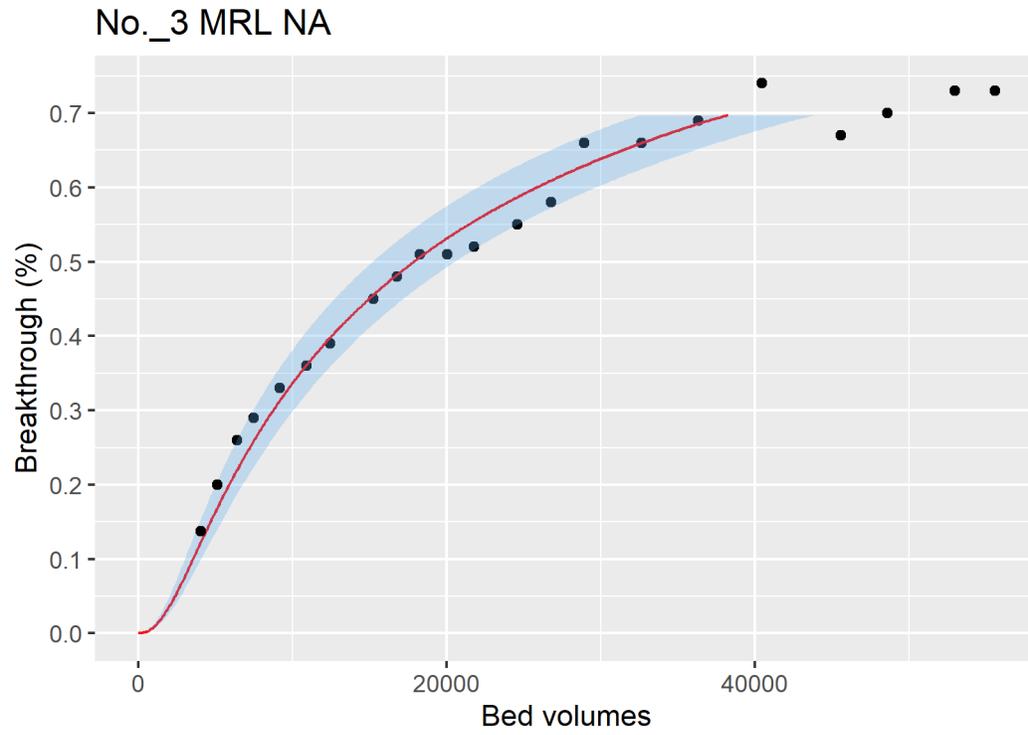


Figure ID 47- 3

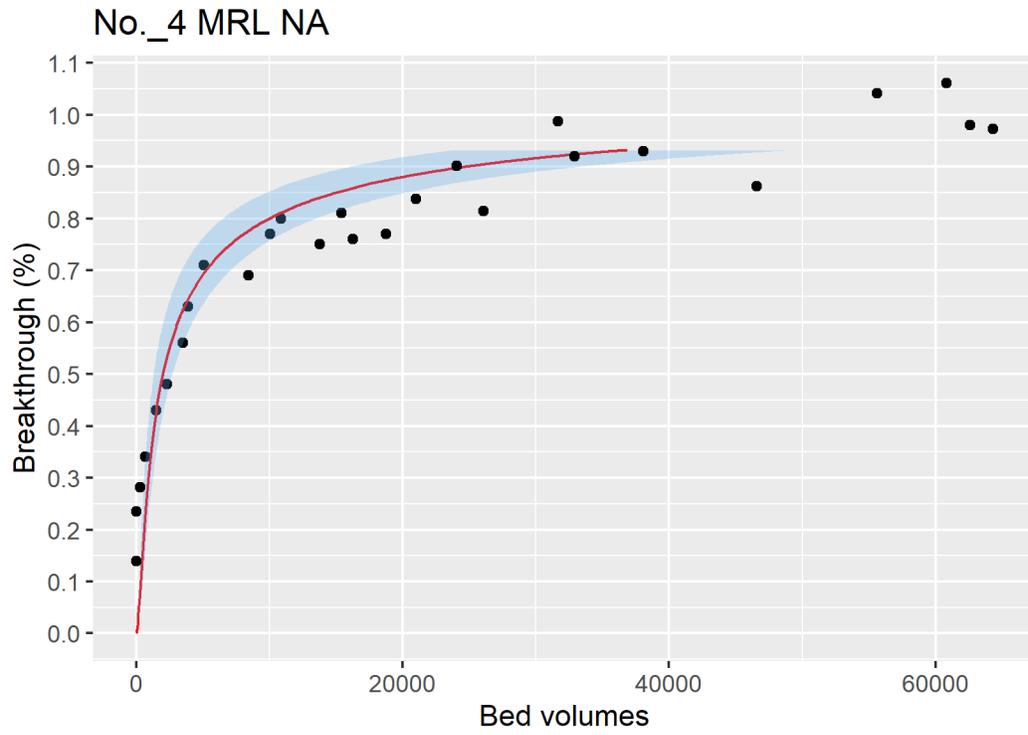


Figure ID 47- 4

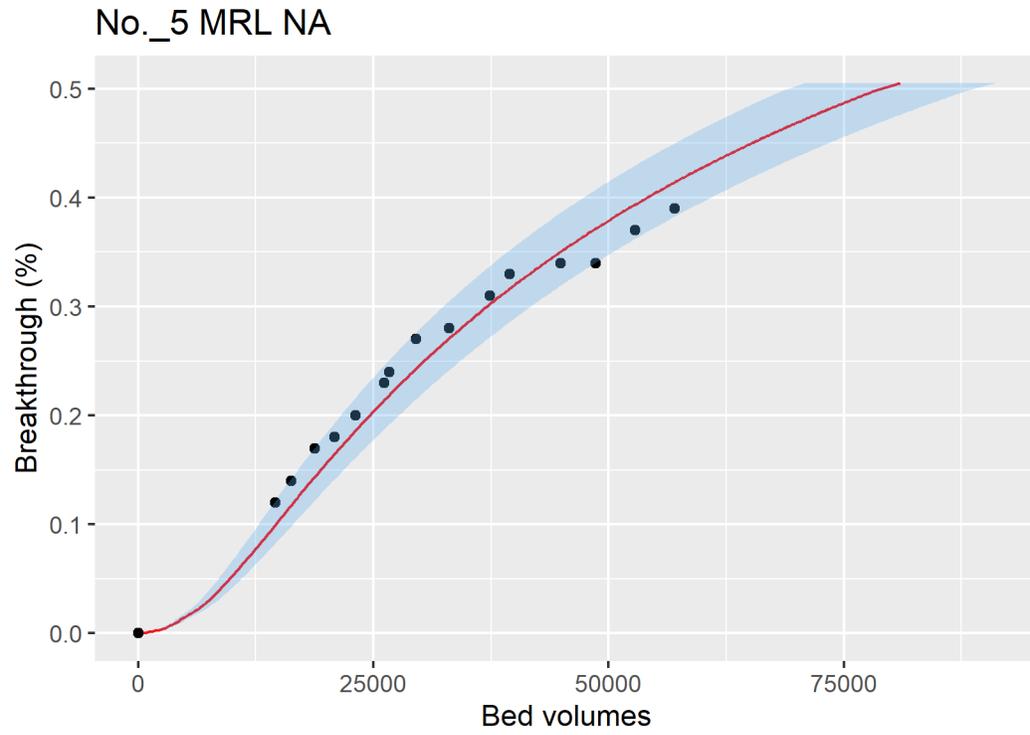


Figure ID 47- 5

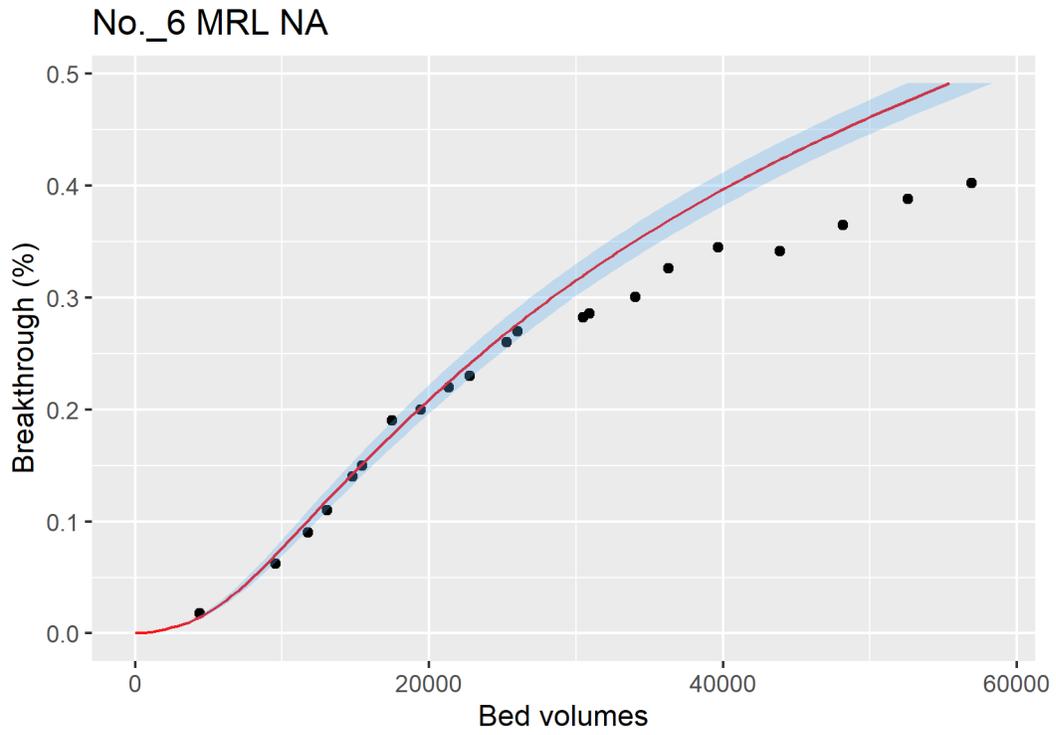


Figure ID 47- 6

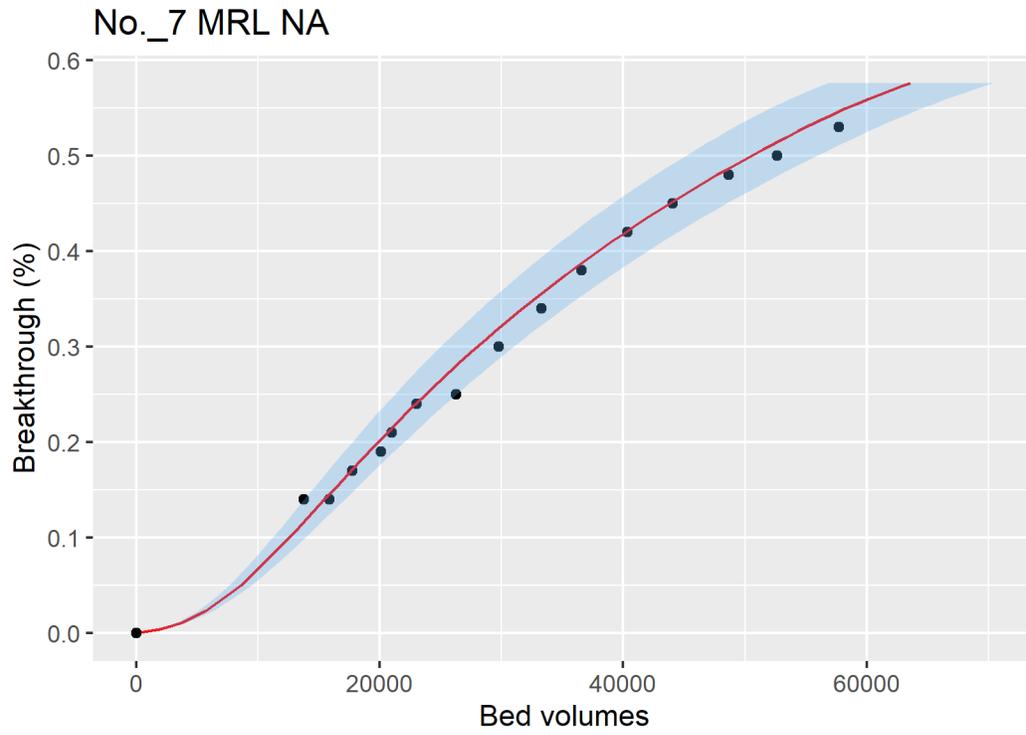


Figure ID 47- 7

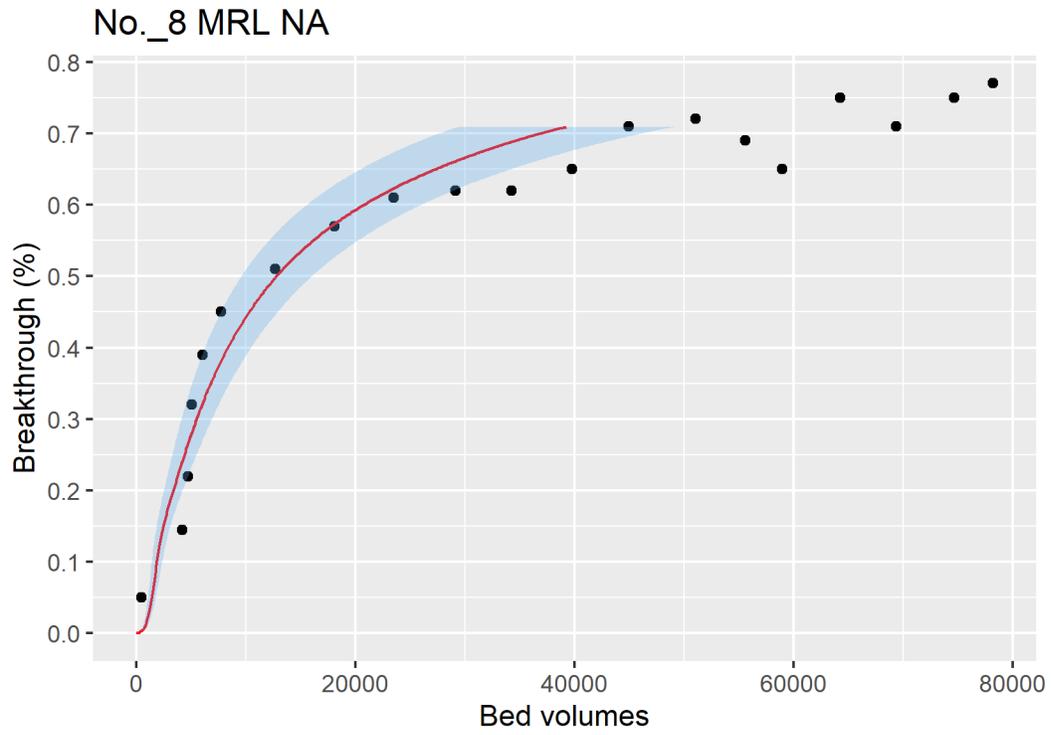


Figure ID 47- 8

Breakthrough data of Ref ID 50

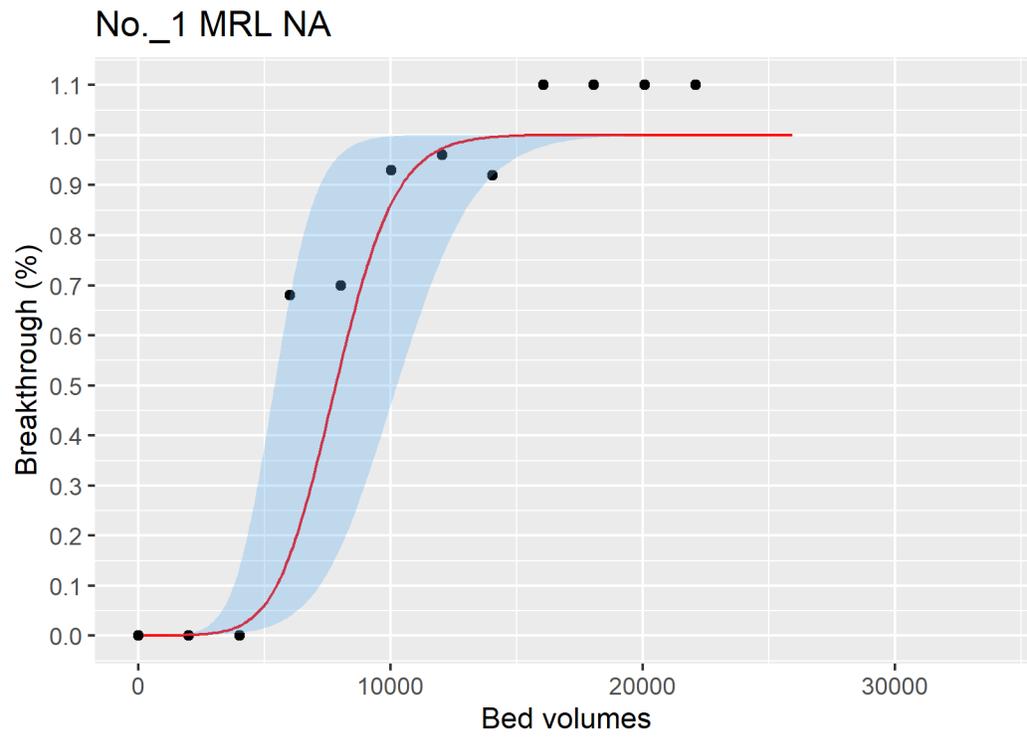


Figure ID 50- 1

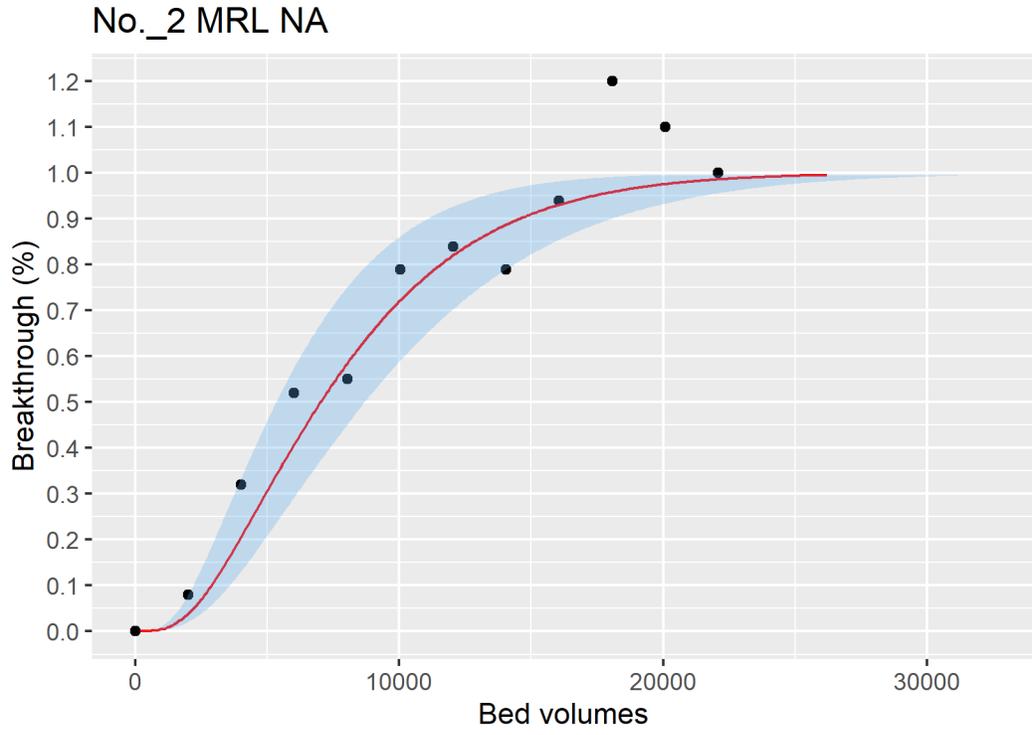


Figure ID 50- 2

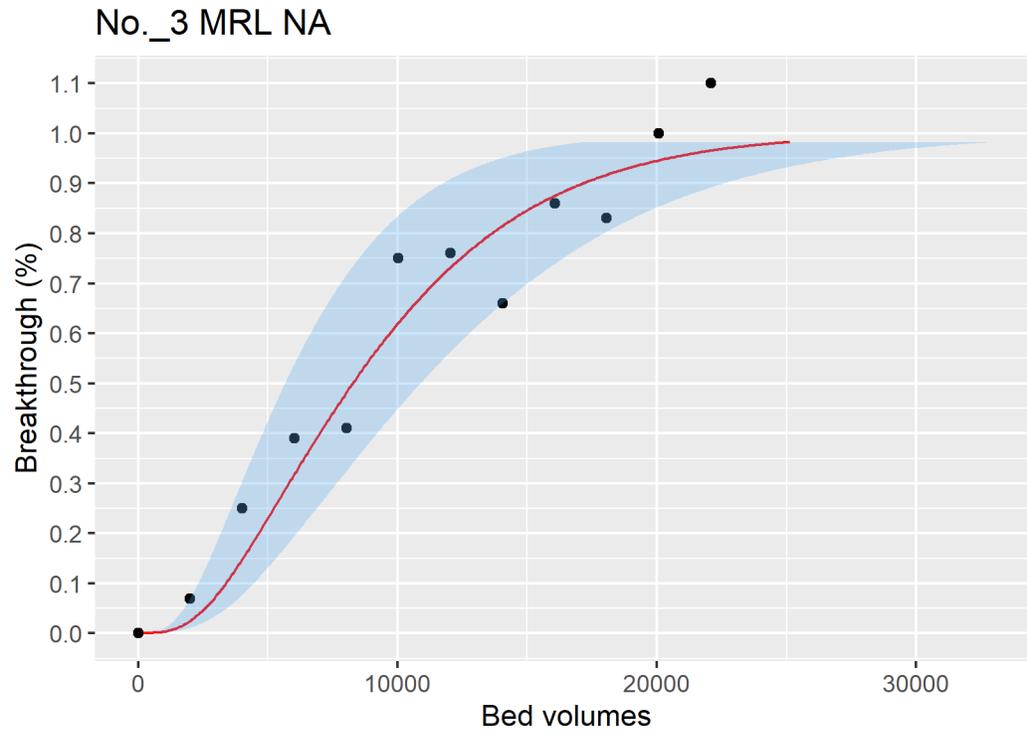


Figure ID 50- 3

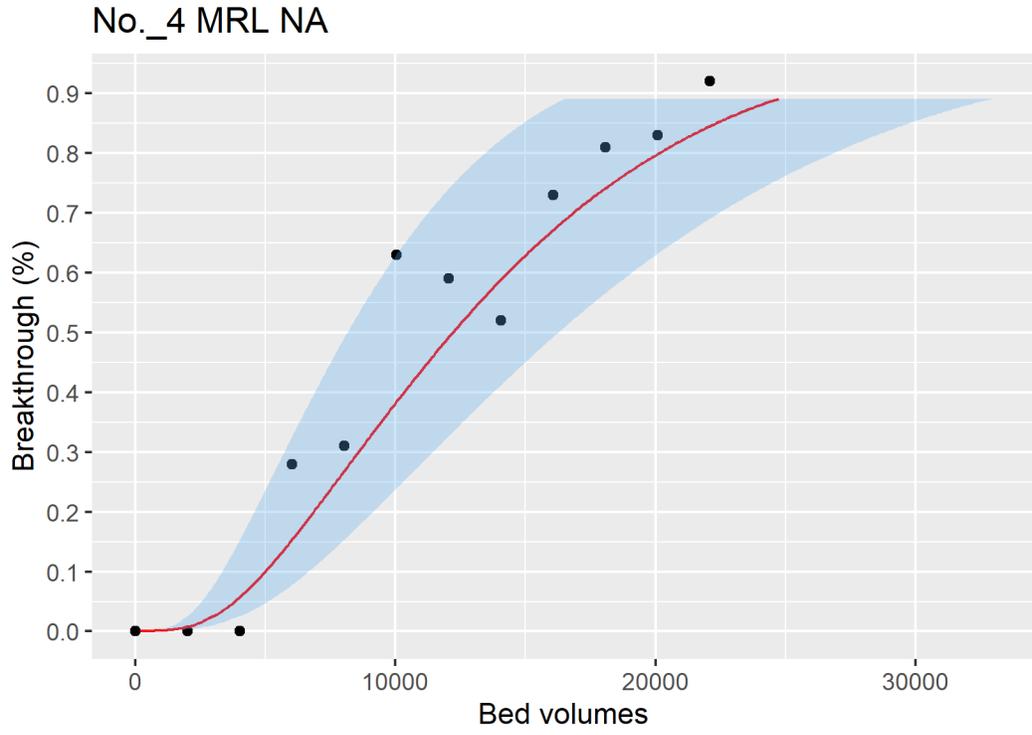


Figure ID 50- 4

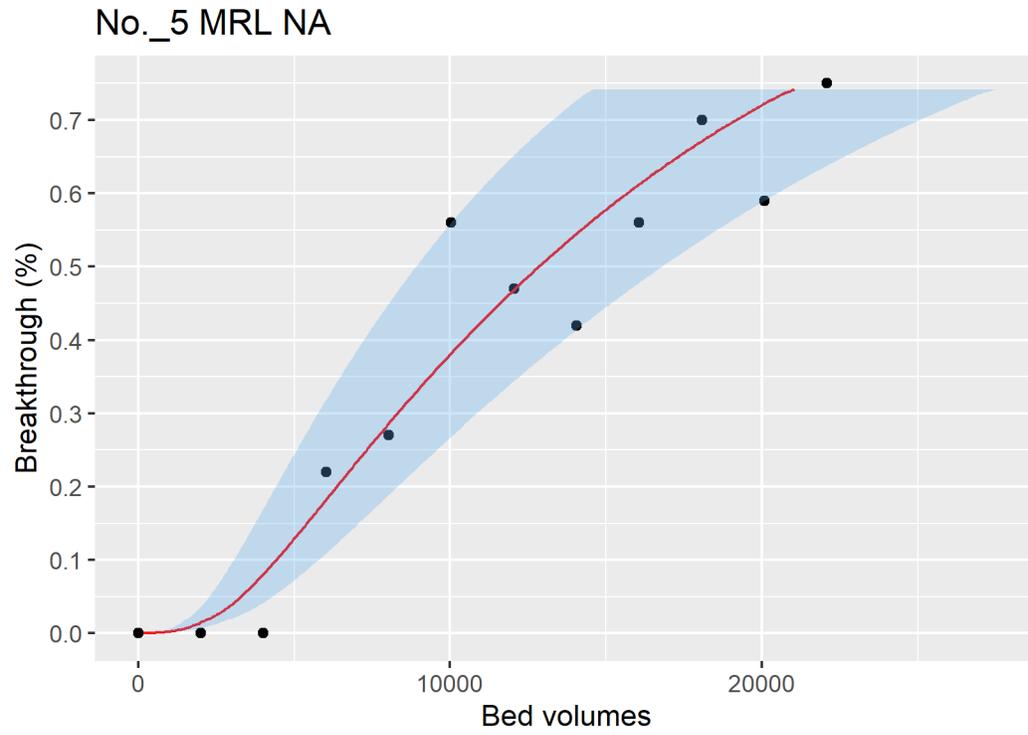


Figure ID 50- 5

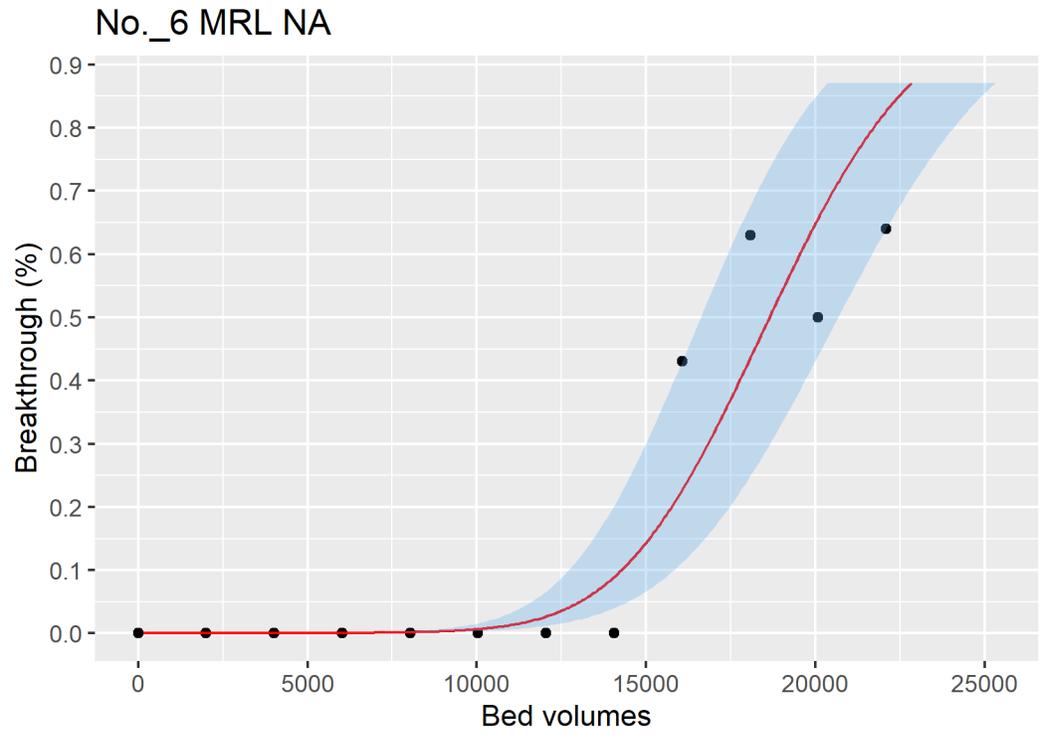


Figure ID 50- 6

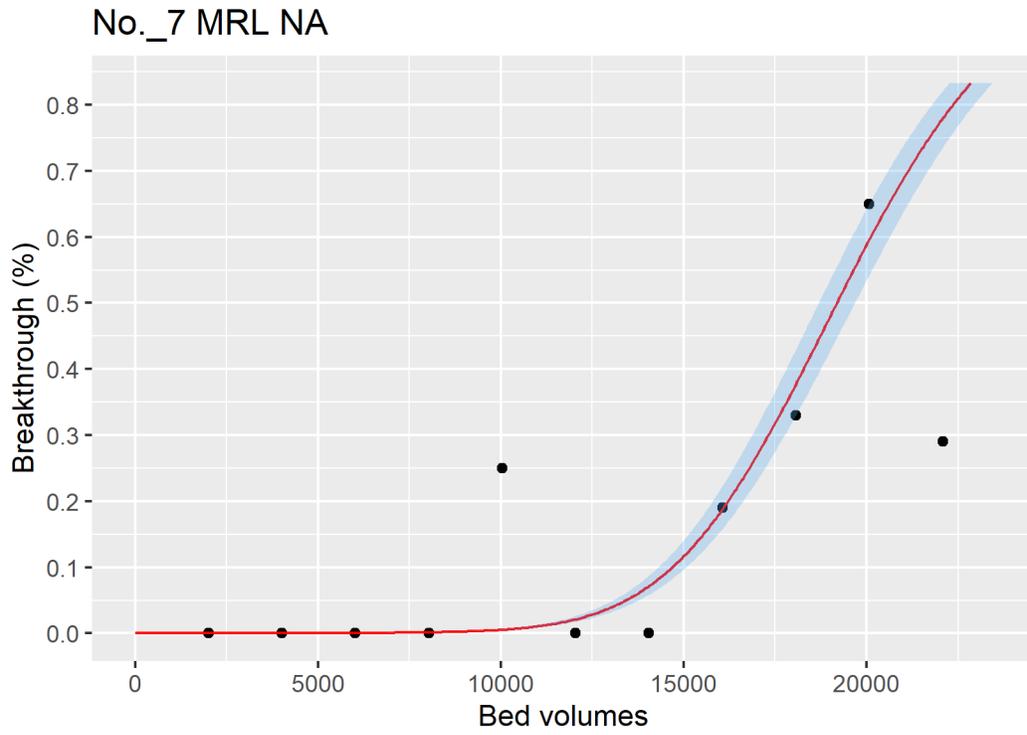


Figure ID 50- 7

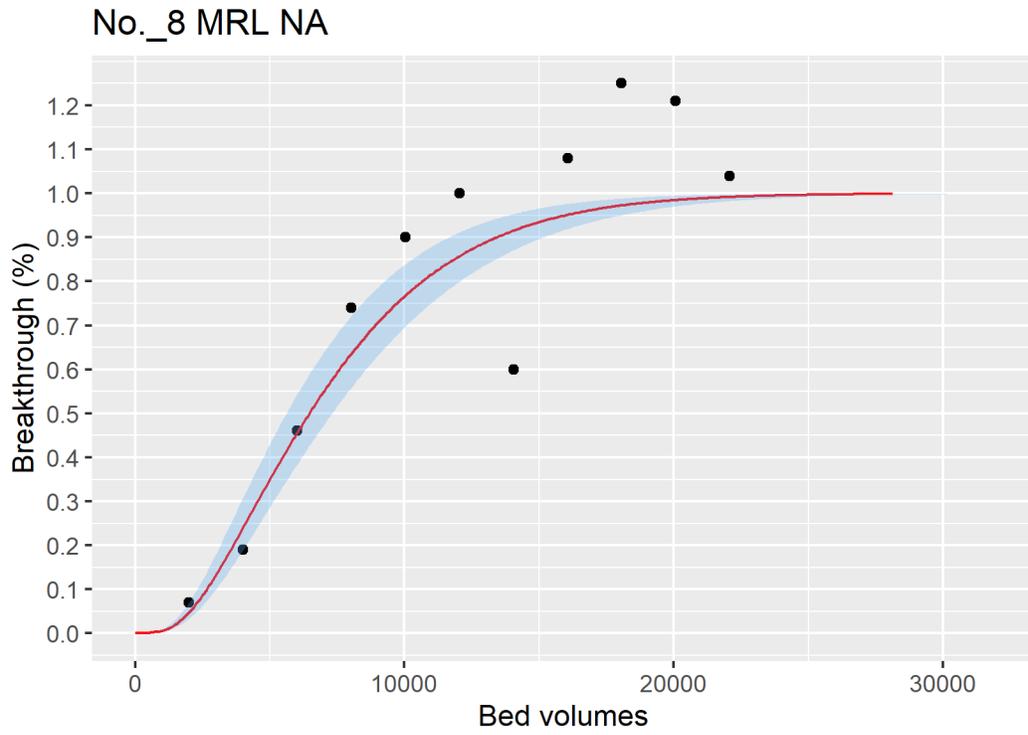


Figure ID 50- 8

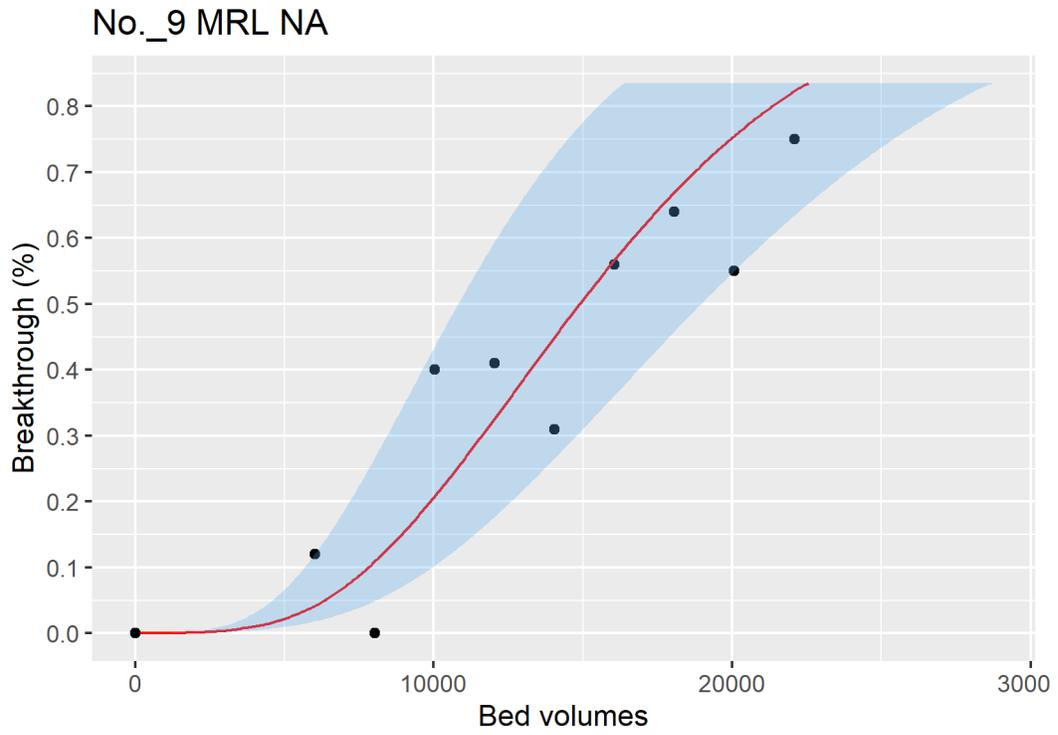


Figure ID 50- 9

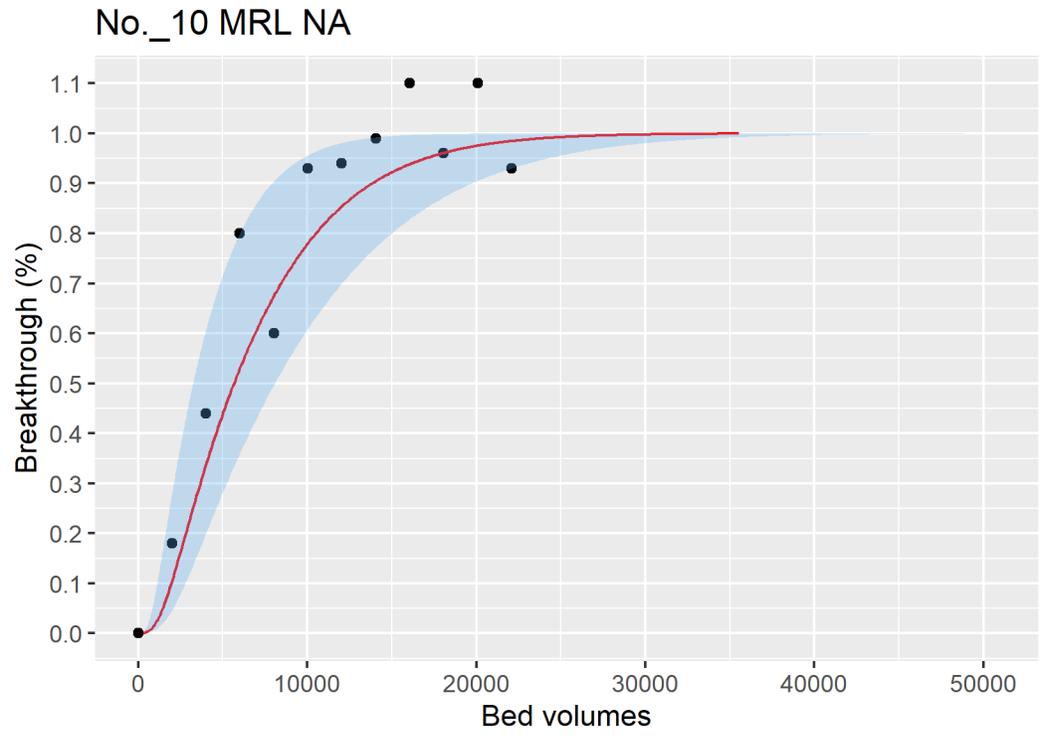


Figure ID 50- 10

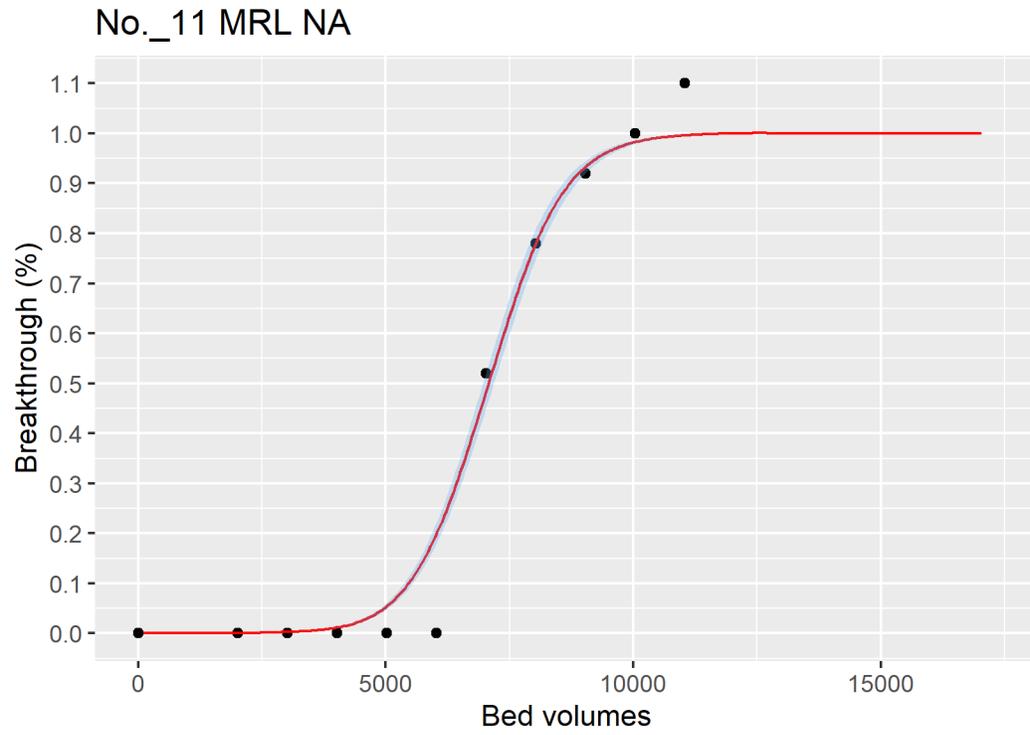


Figure ID 50- 11

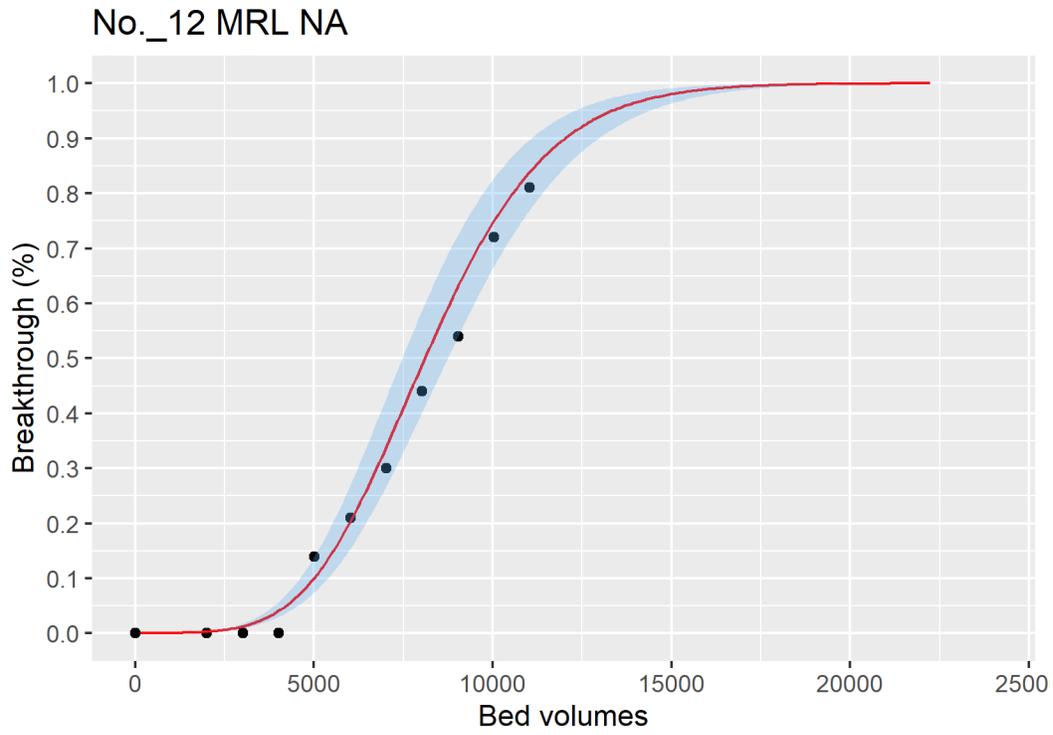


Figure ID 50- 12

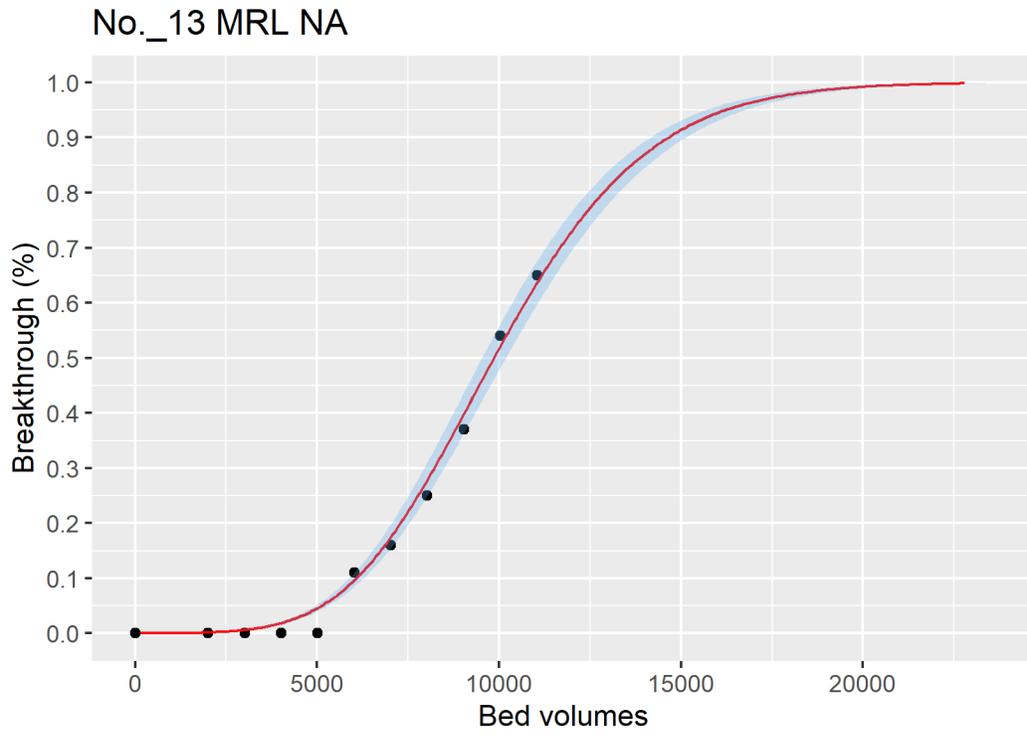


Figure ID 50- 13

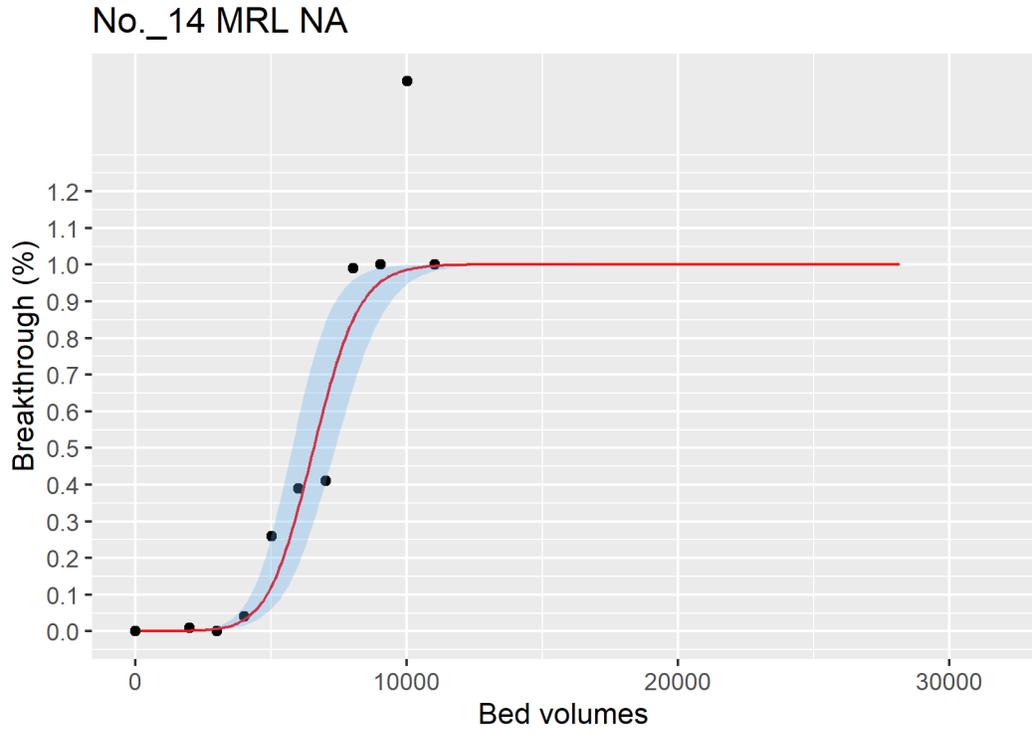


Figure ID 50- 14

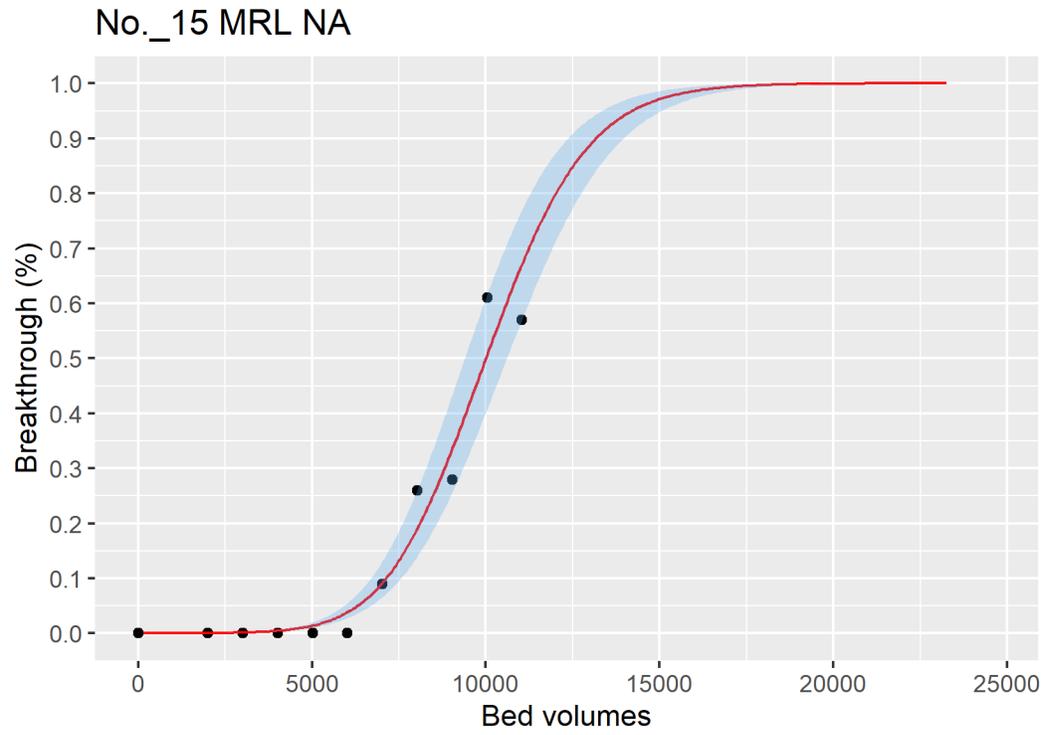


Figure ID 50- 15

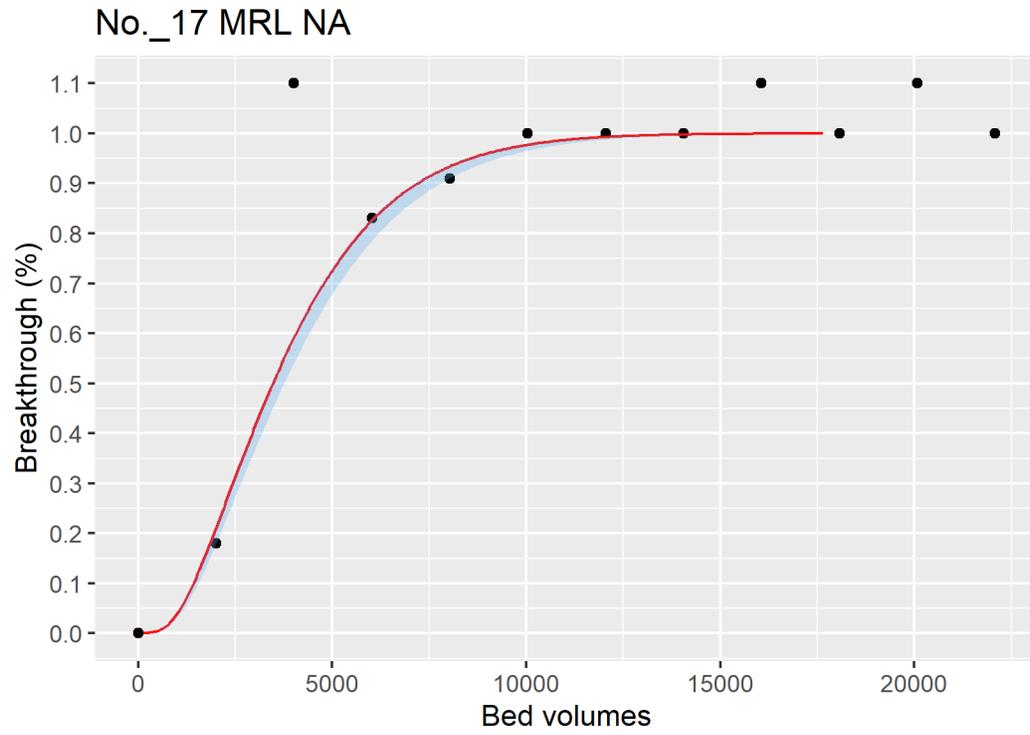


Figure ID 50- 16

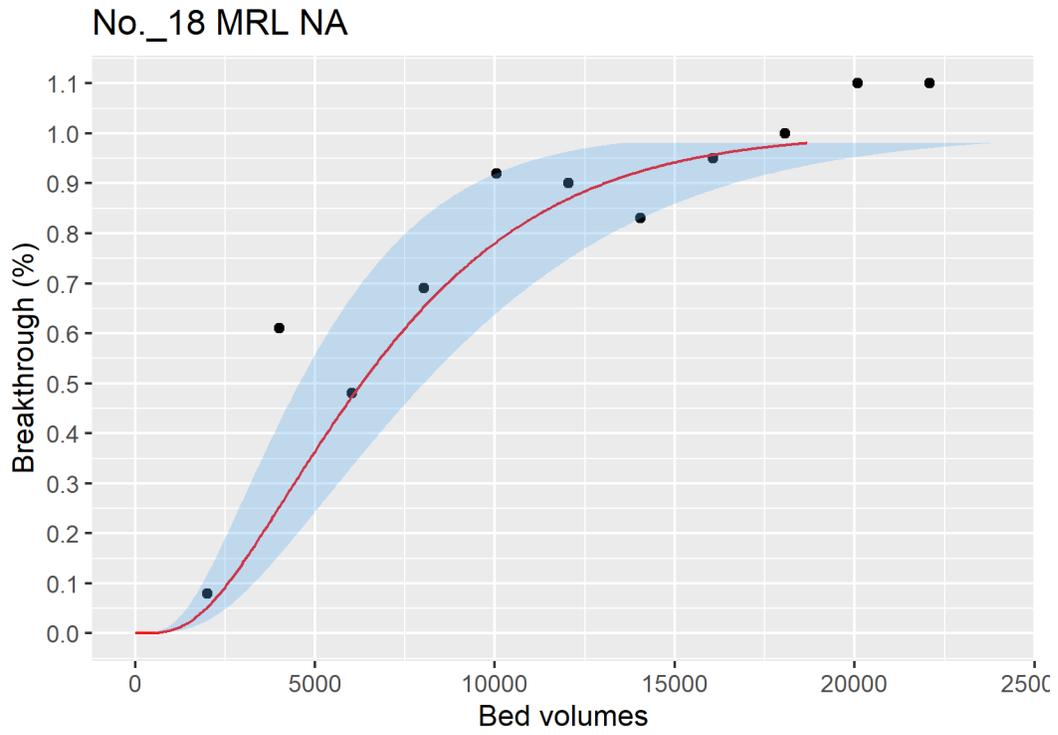


Figure ID 50- 17

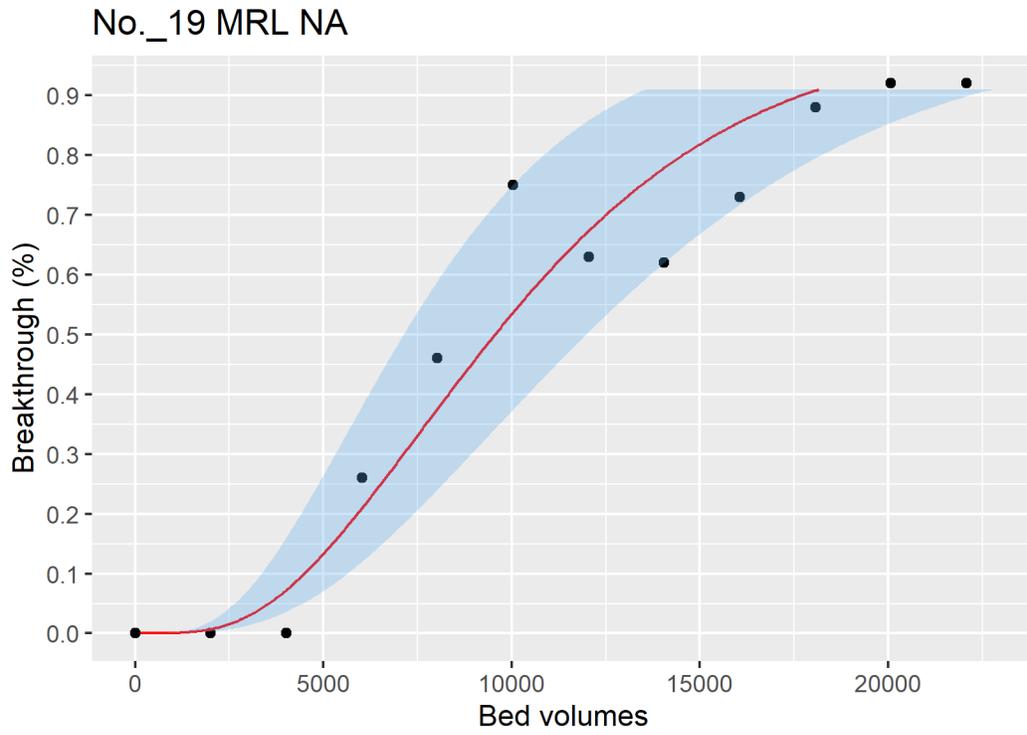


Figure ID 50- 18

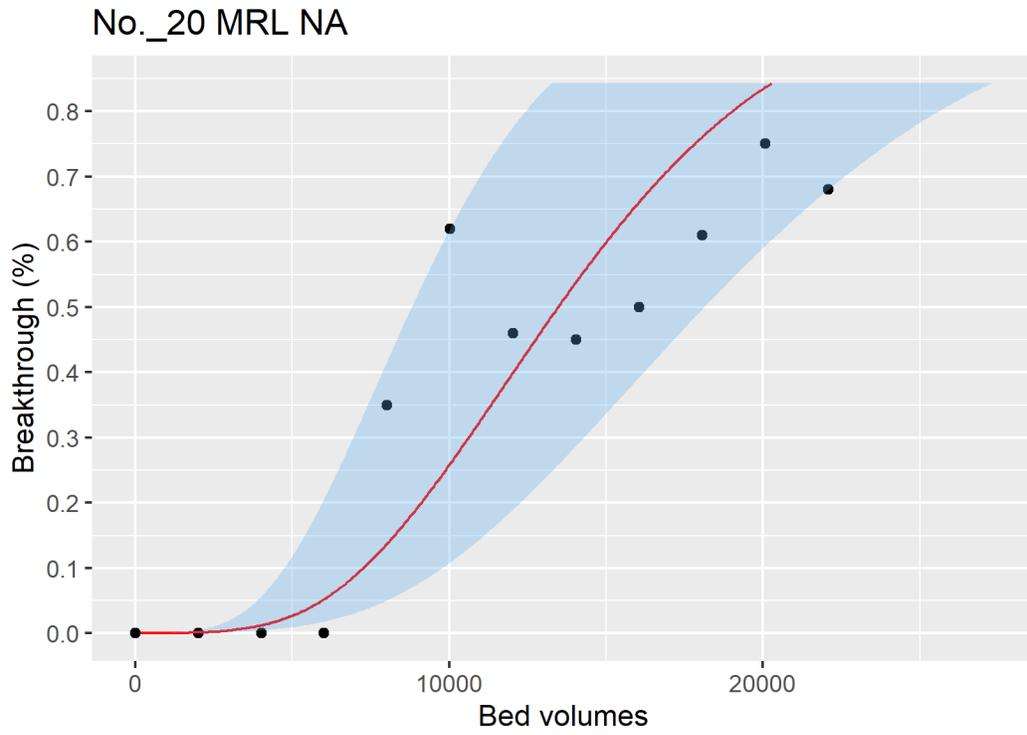


Figure ID 50- 19

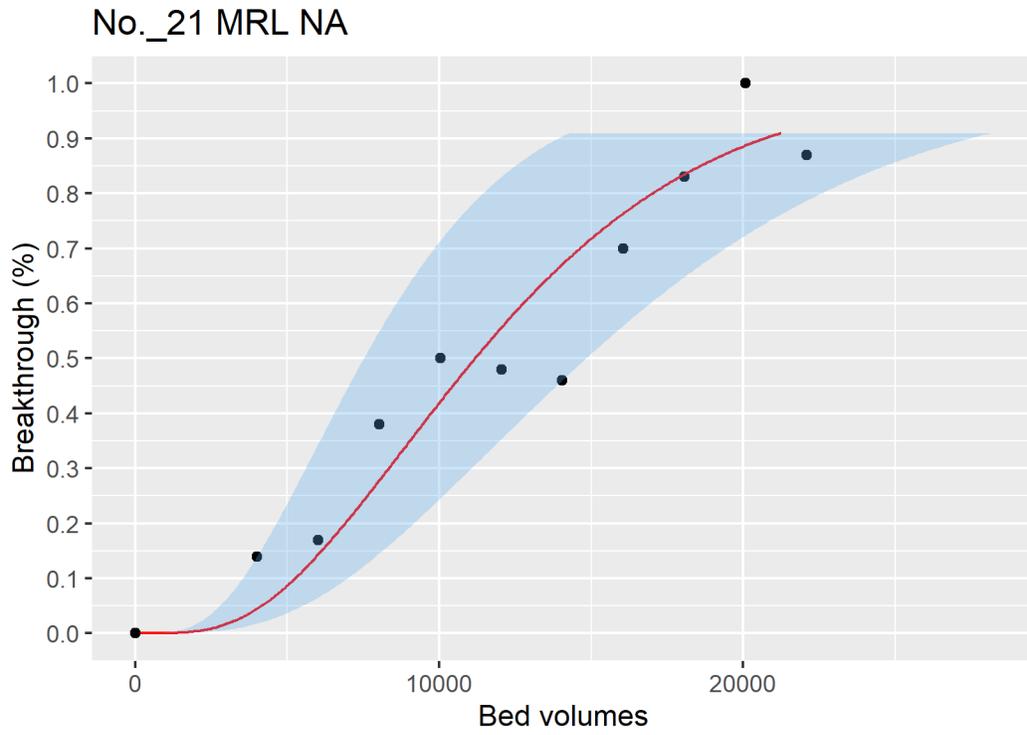


Figure ID 50- 20

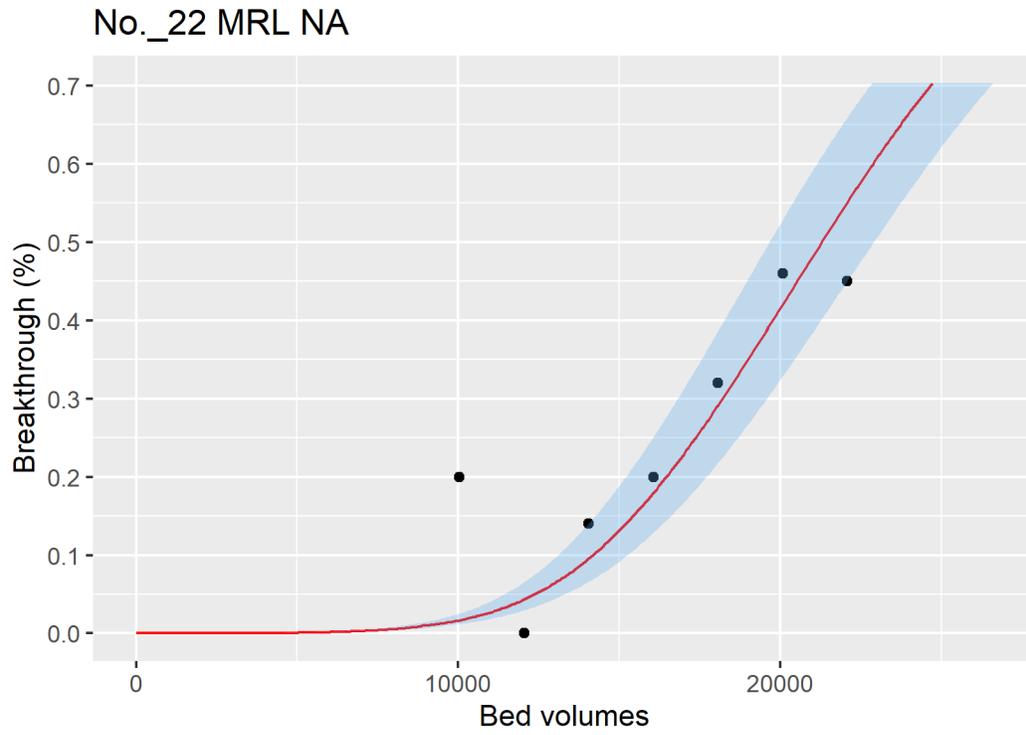


Figure ID 50- 21

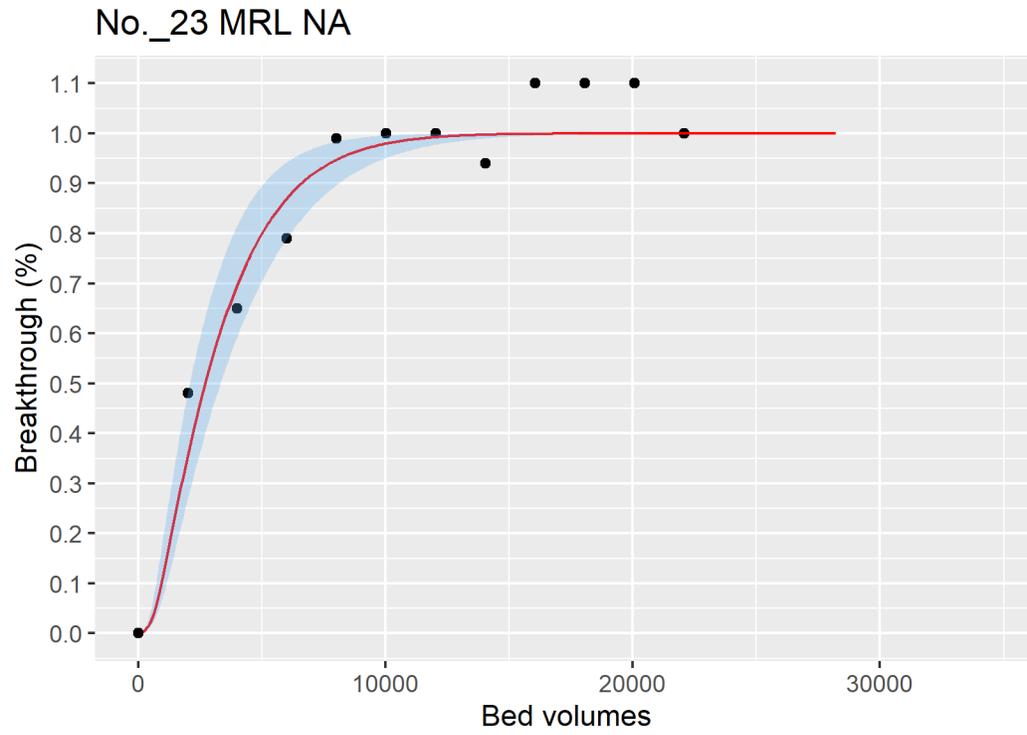


Figure ID 50- 22

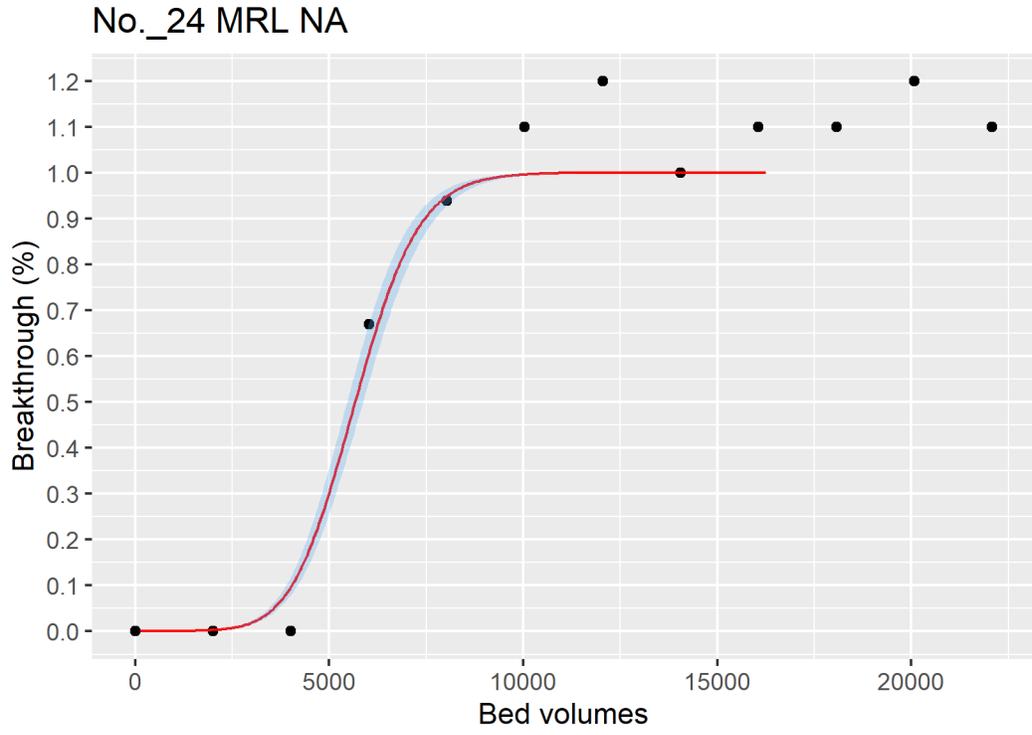


Figure ID 50- 23

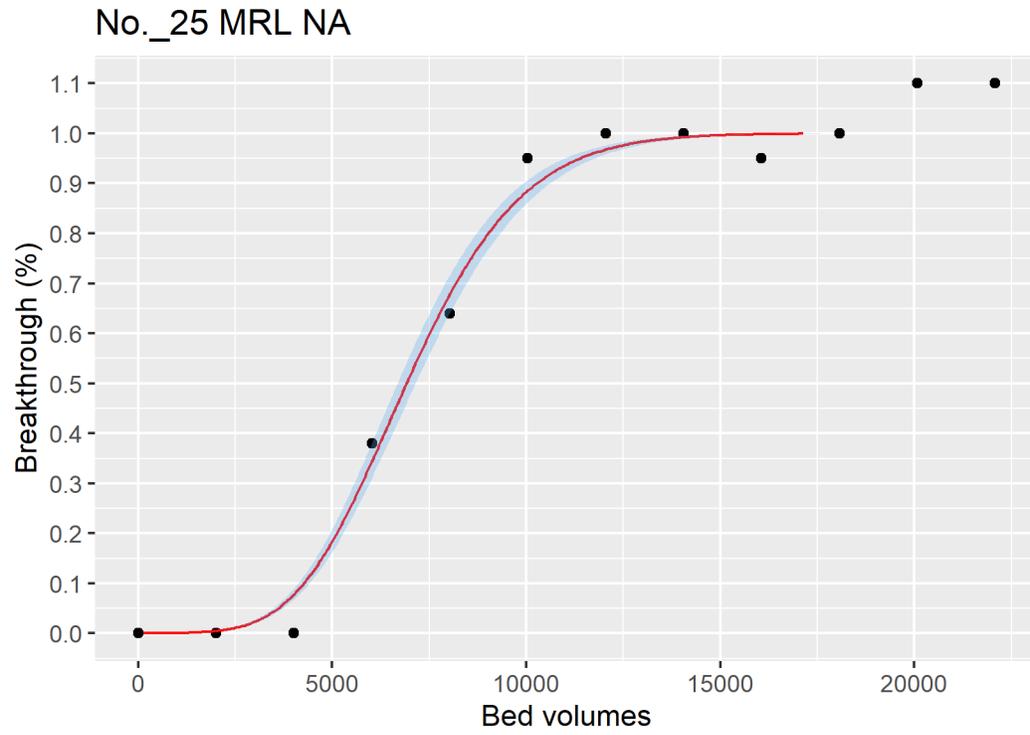


Figure ID 50- 24

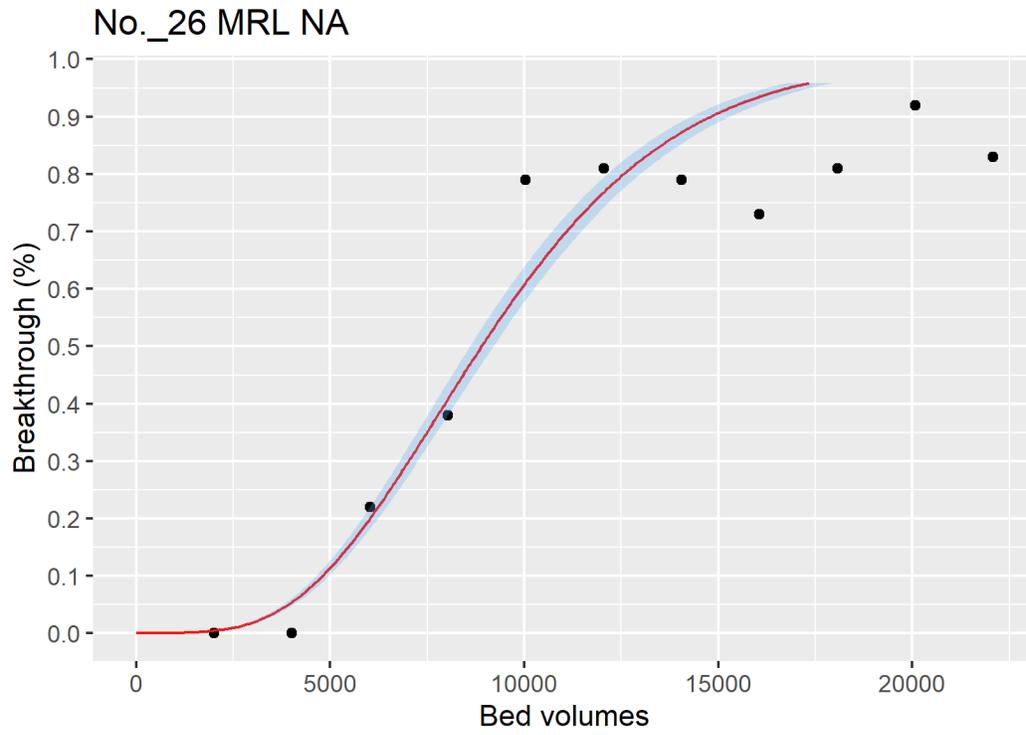


Figure ID 50- 25

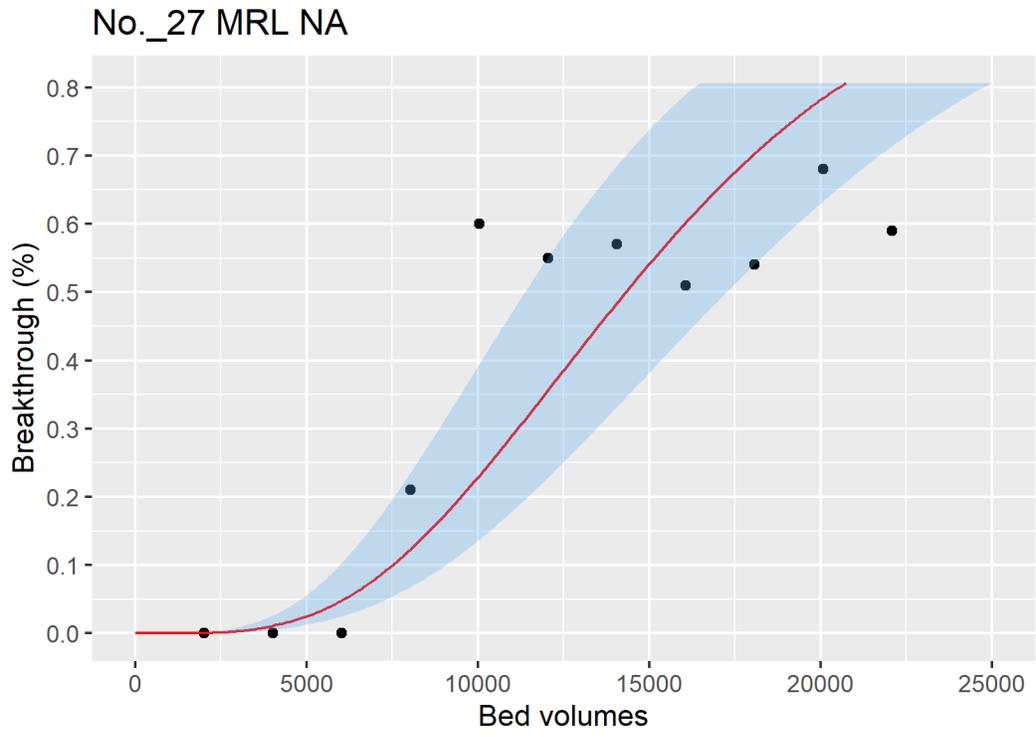


Figure ID 50- 26

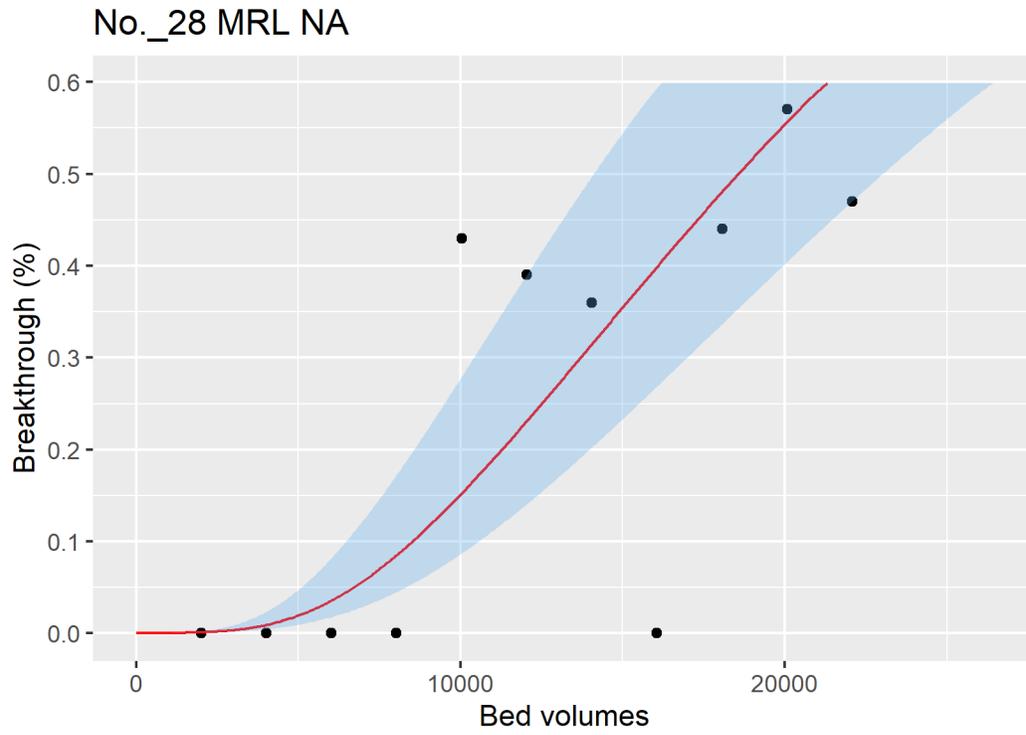


Figure ID 50- 27

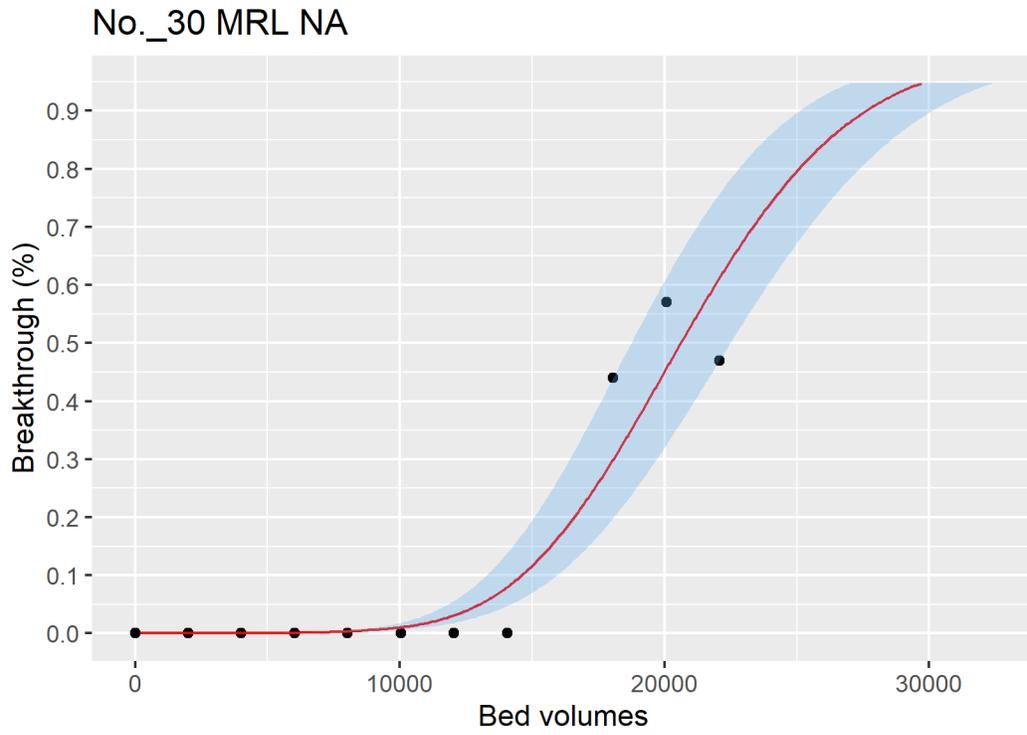


Figure ID 50- 28

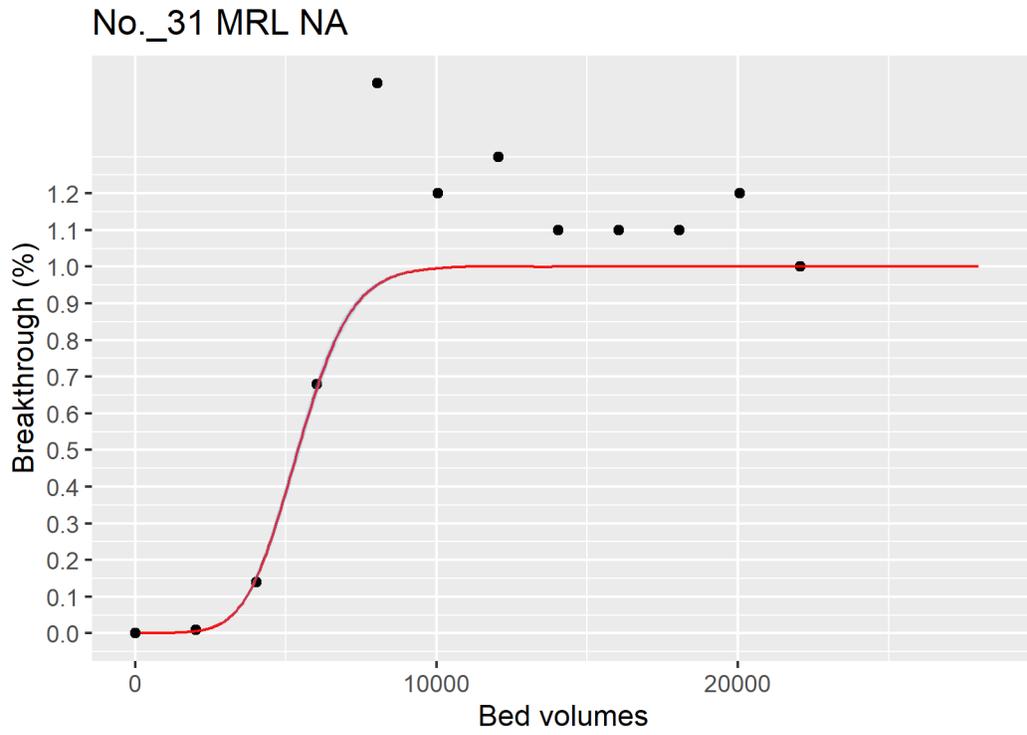


Figure ID 50- 29

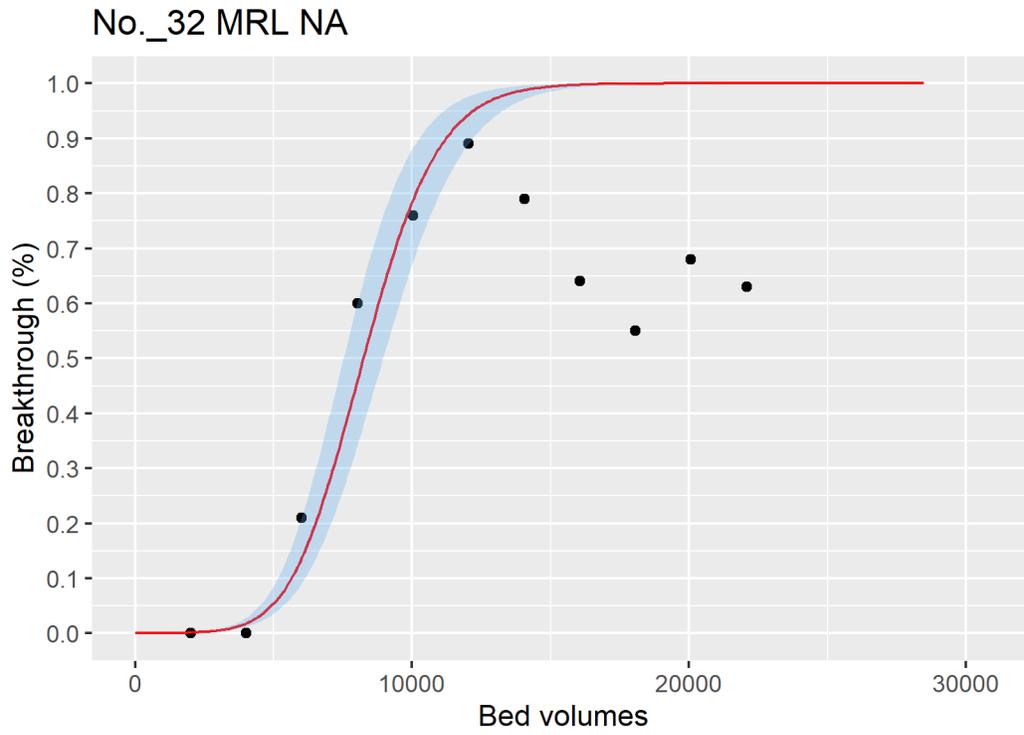


Figure ID 50- 30

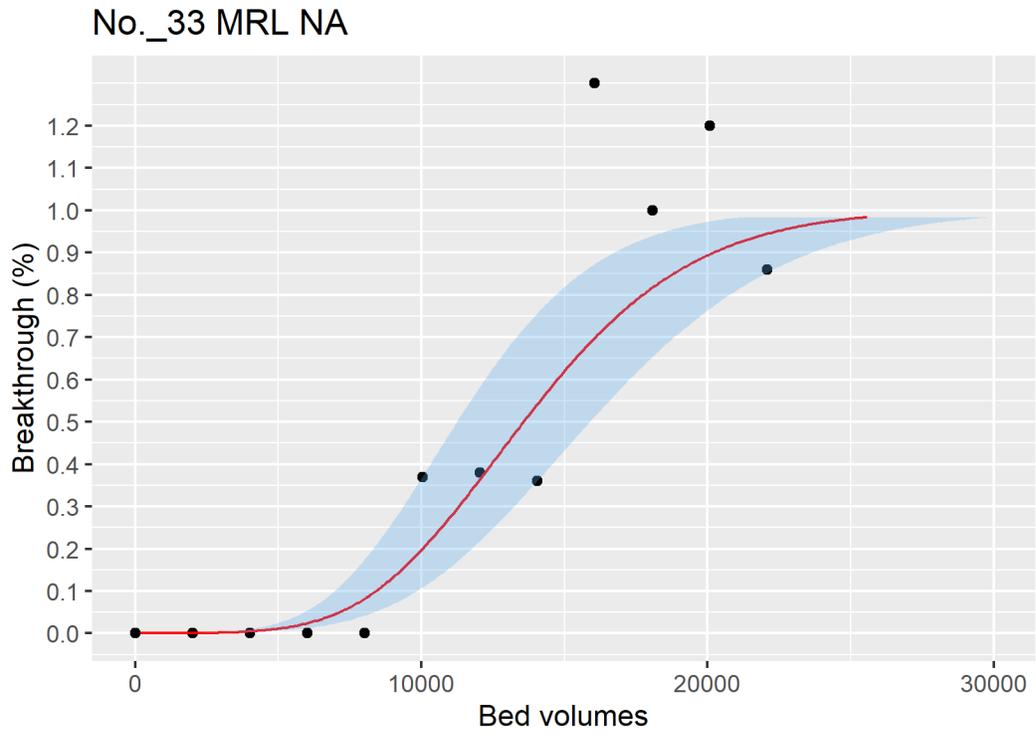


Figure ID 50- 31

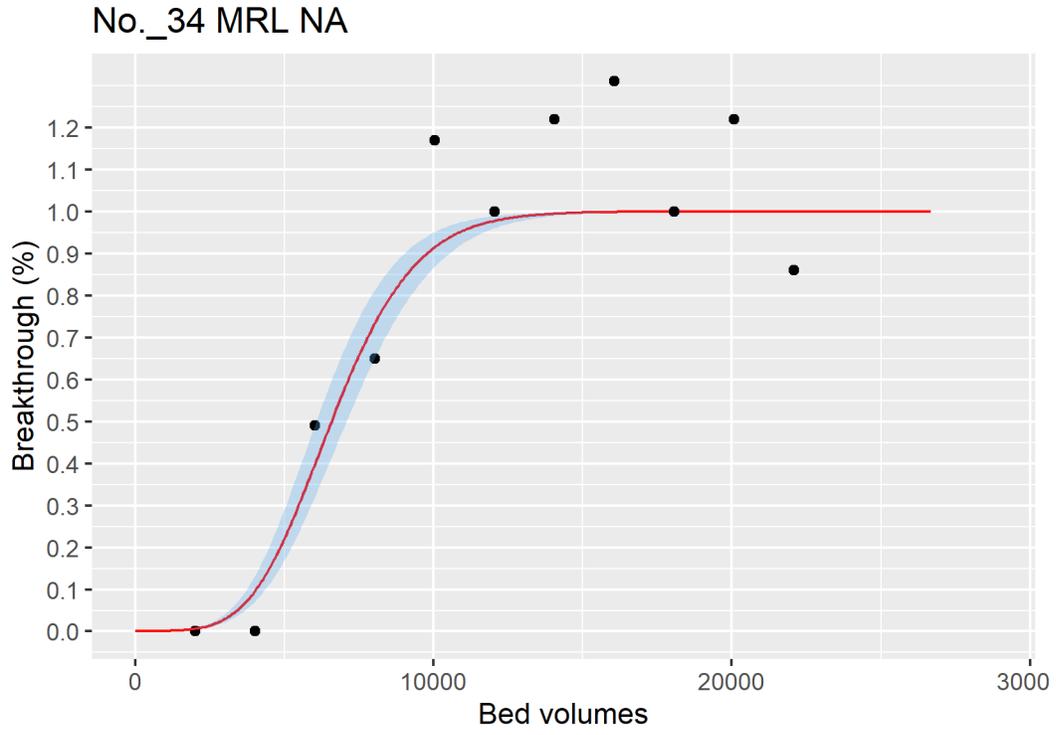


Figure ID 50- 32

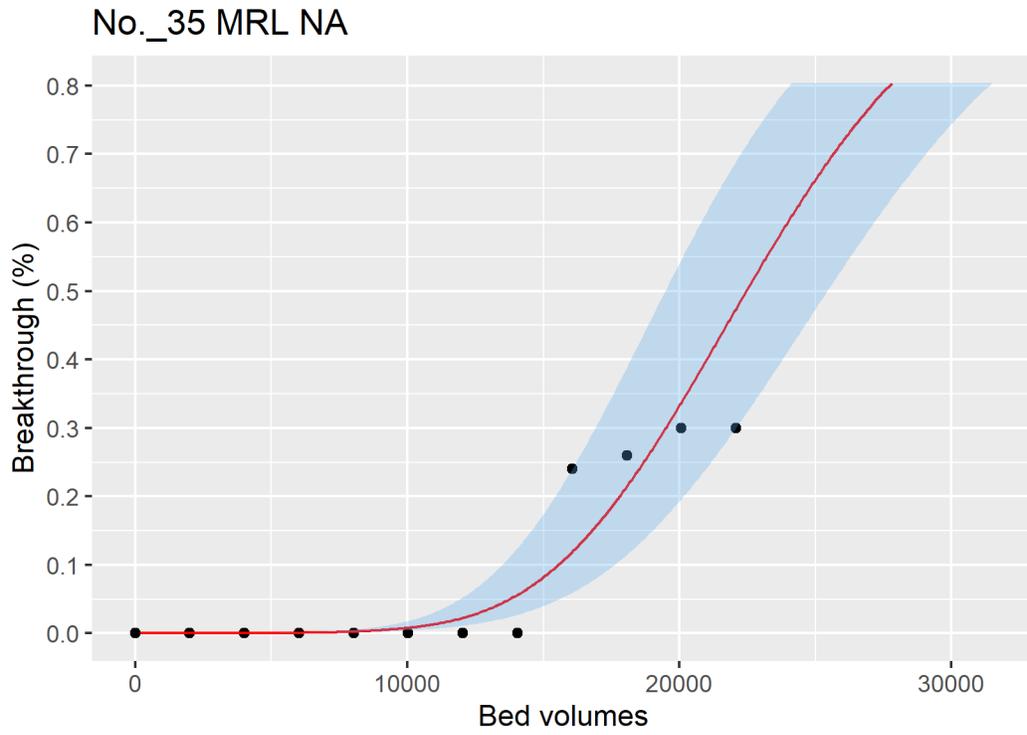


Figure ID 50- 33

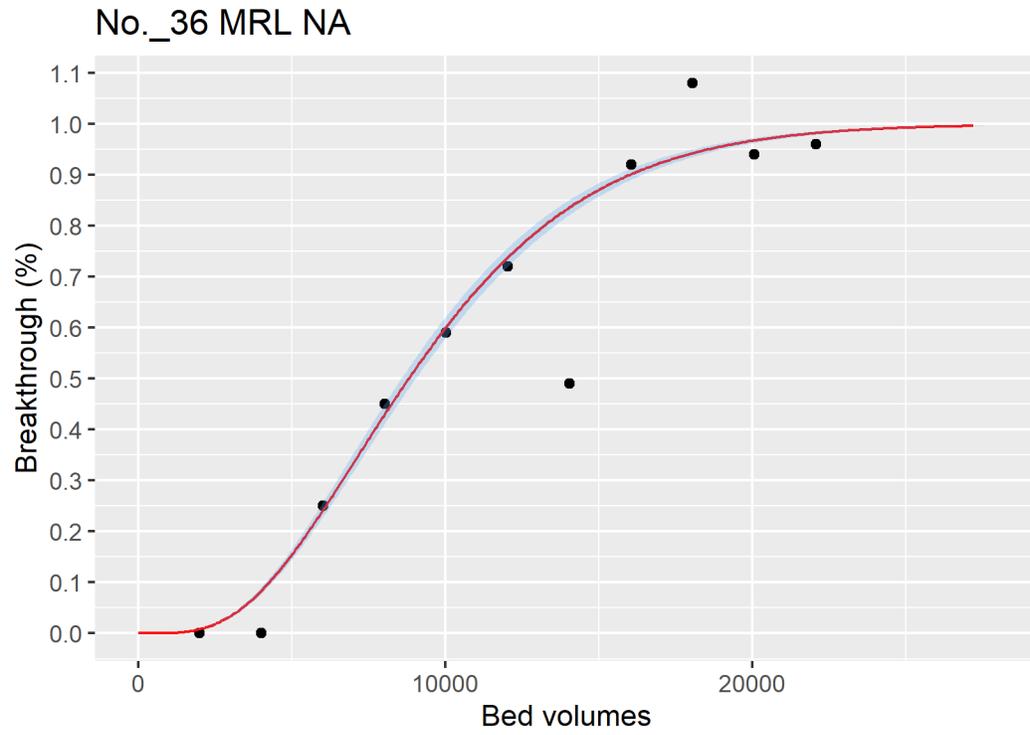


Figure ID 50- 34

Breakthrough data of Ref ID 57

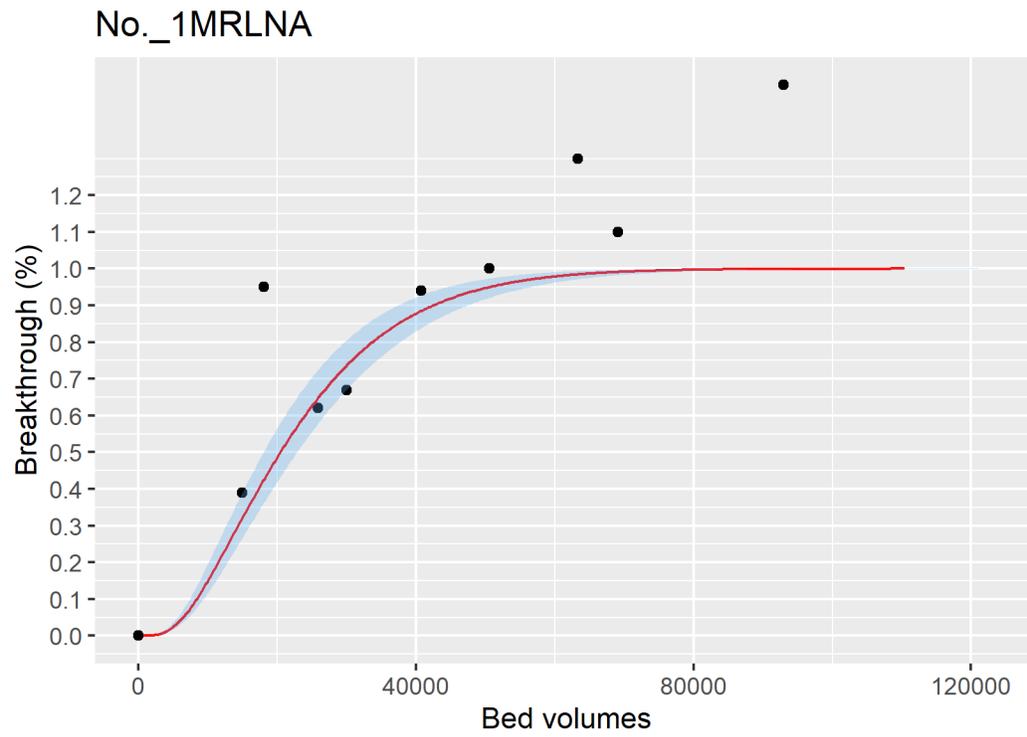


Figure ID 57- 1

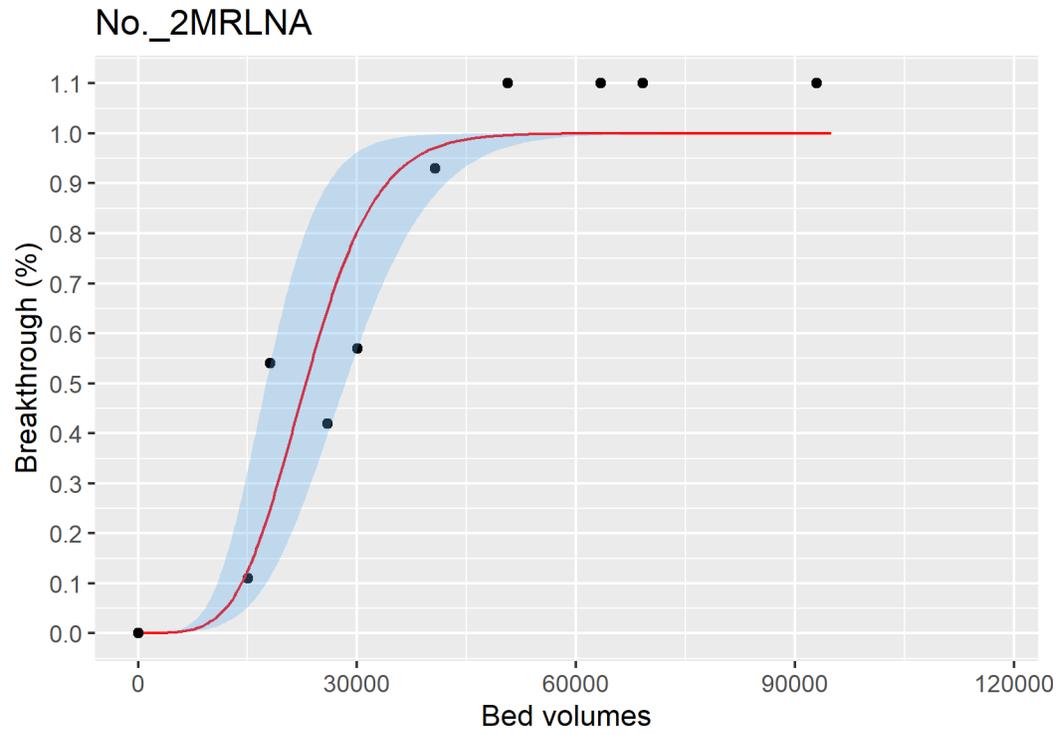


Figure ID 57- 2

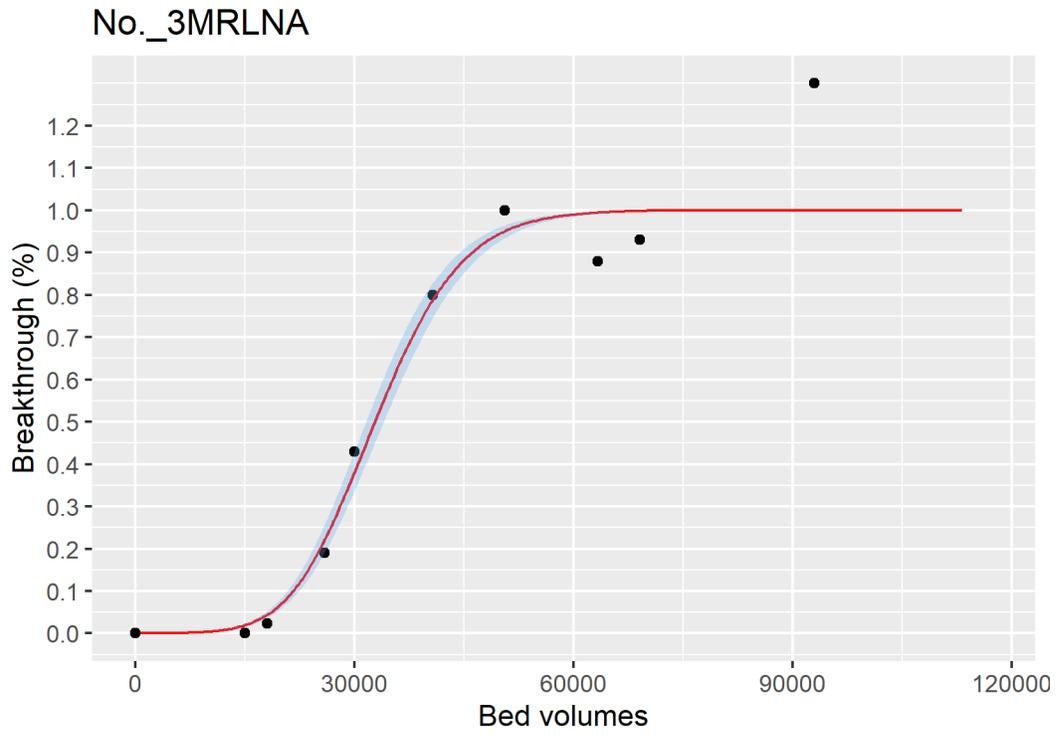


Figure ID 57- 3

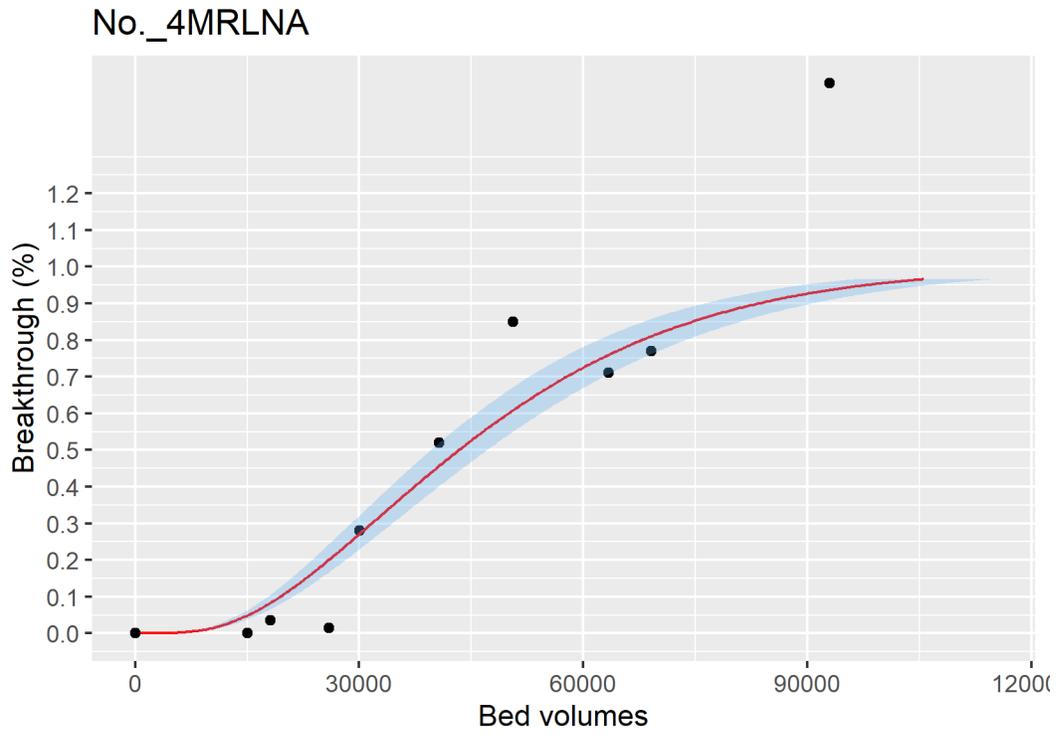


Figure ID 57- 4

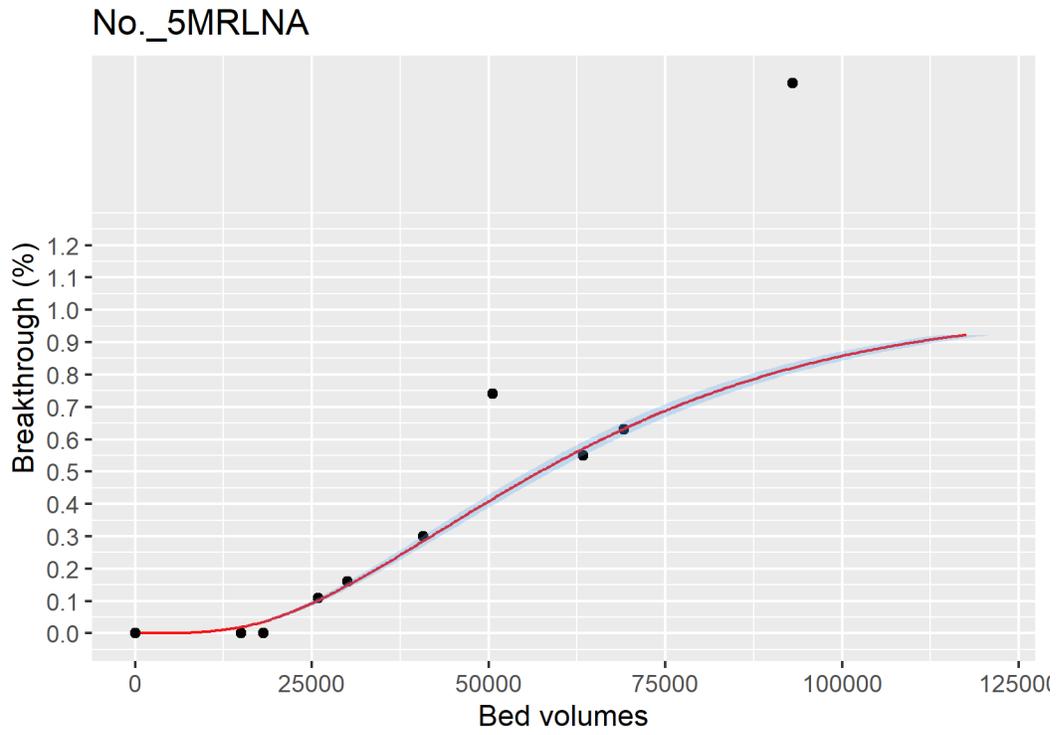


Figure ID 57- 5

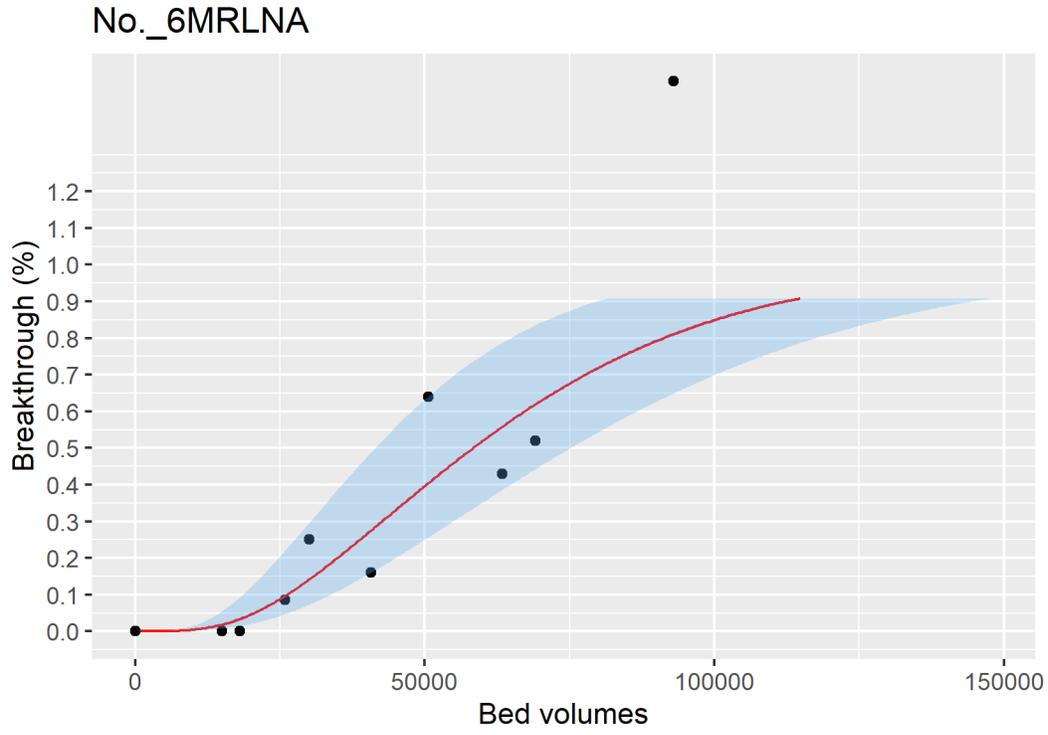


Figure ID 57- 6

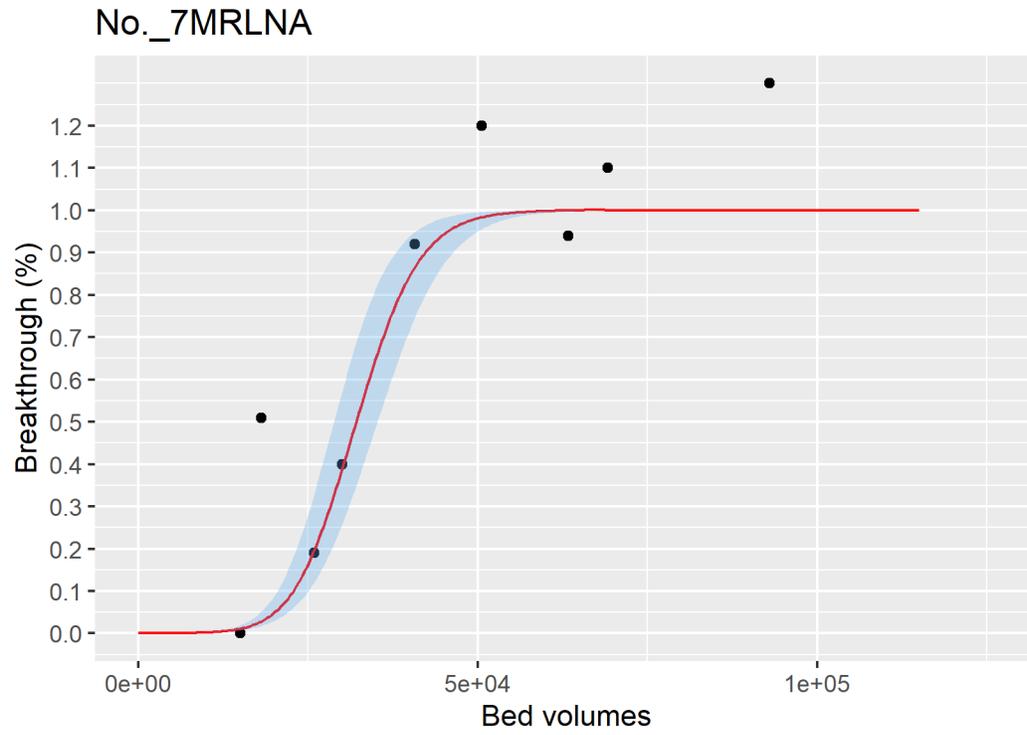


Figure ID 57- 7

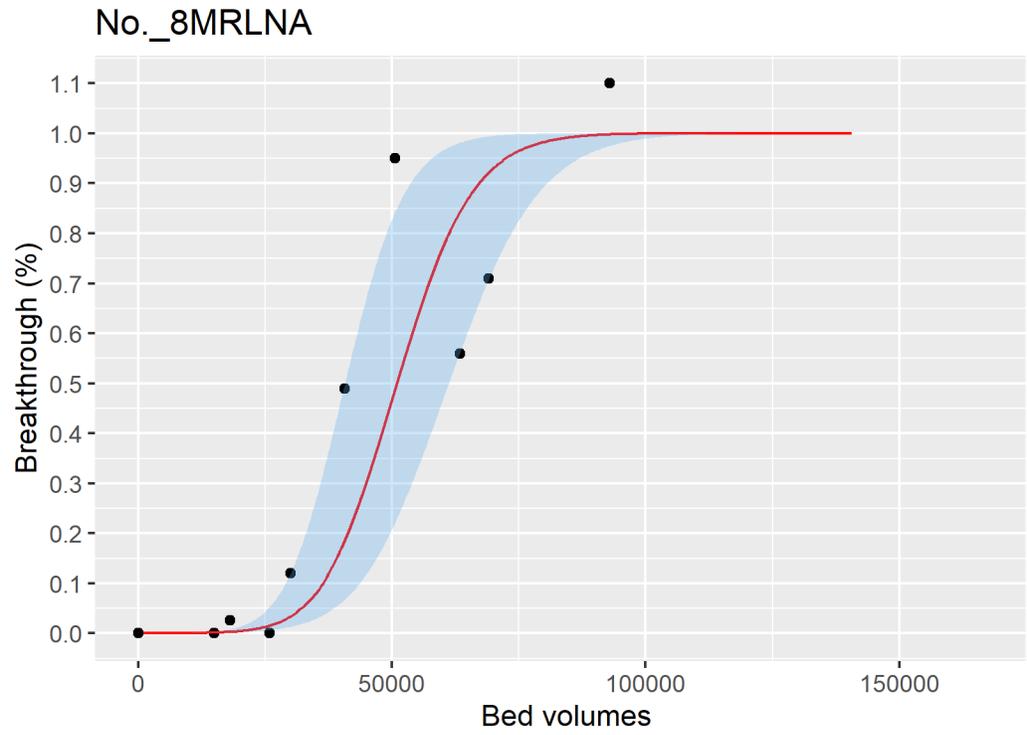


Figure ID 57- 8

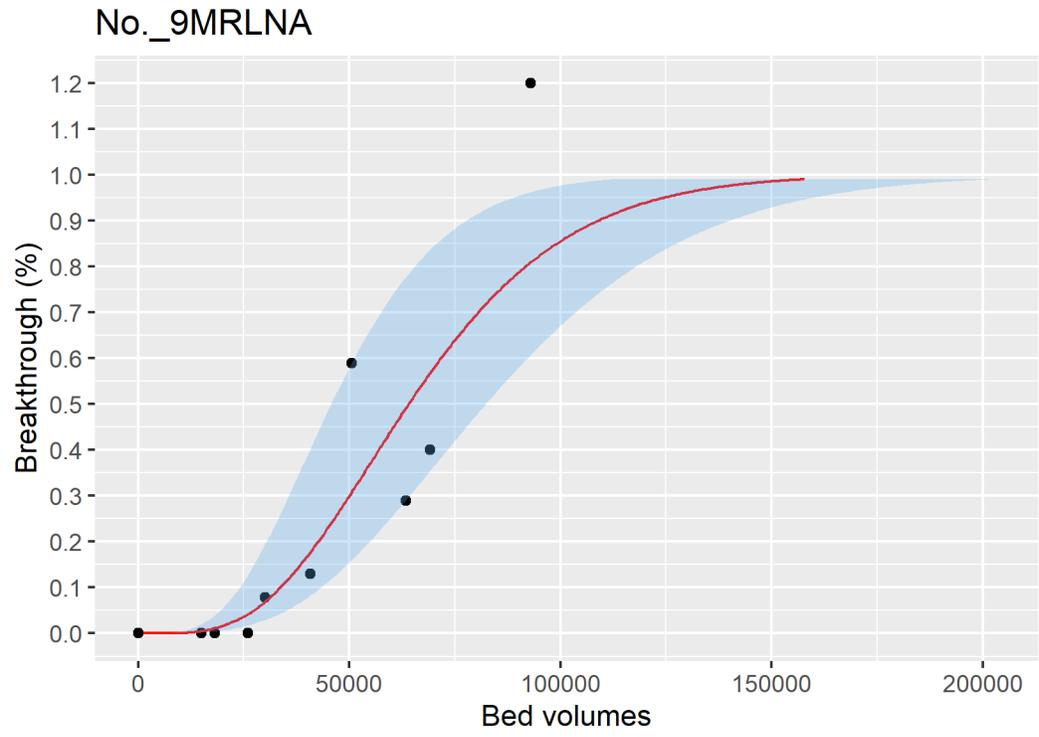


Figure ID 57-9

Breakthrough data of Ref ID 60

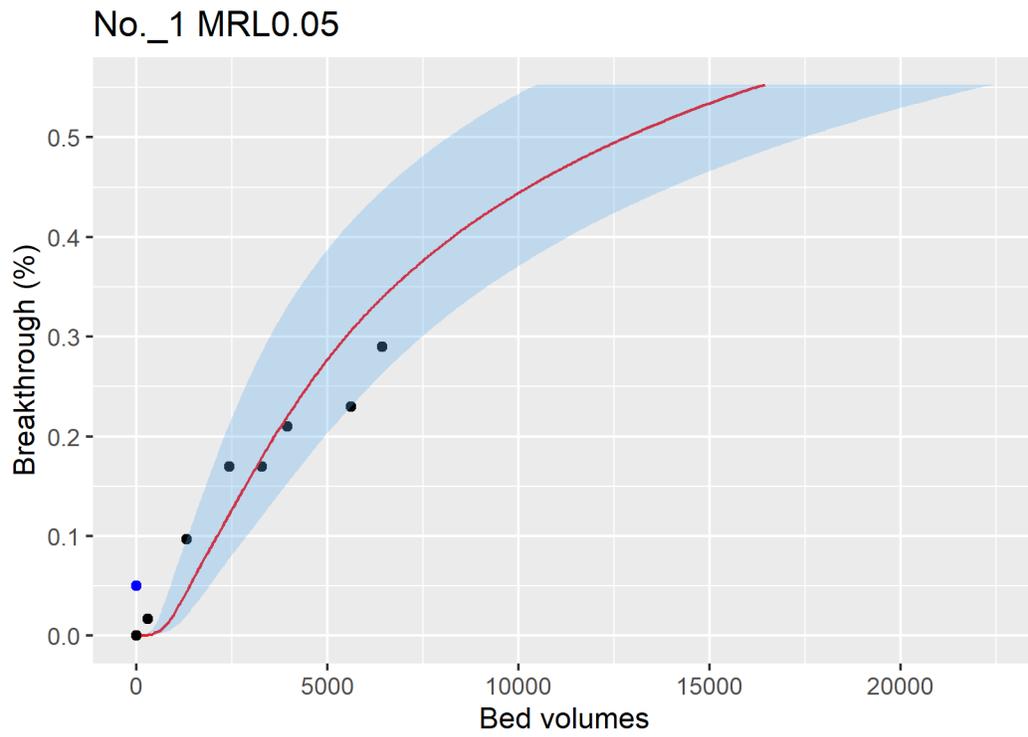


Figure ID 60- 1

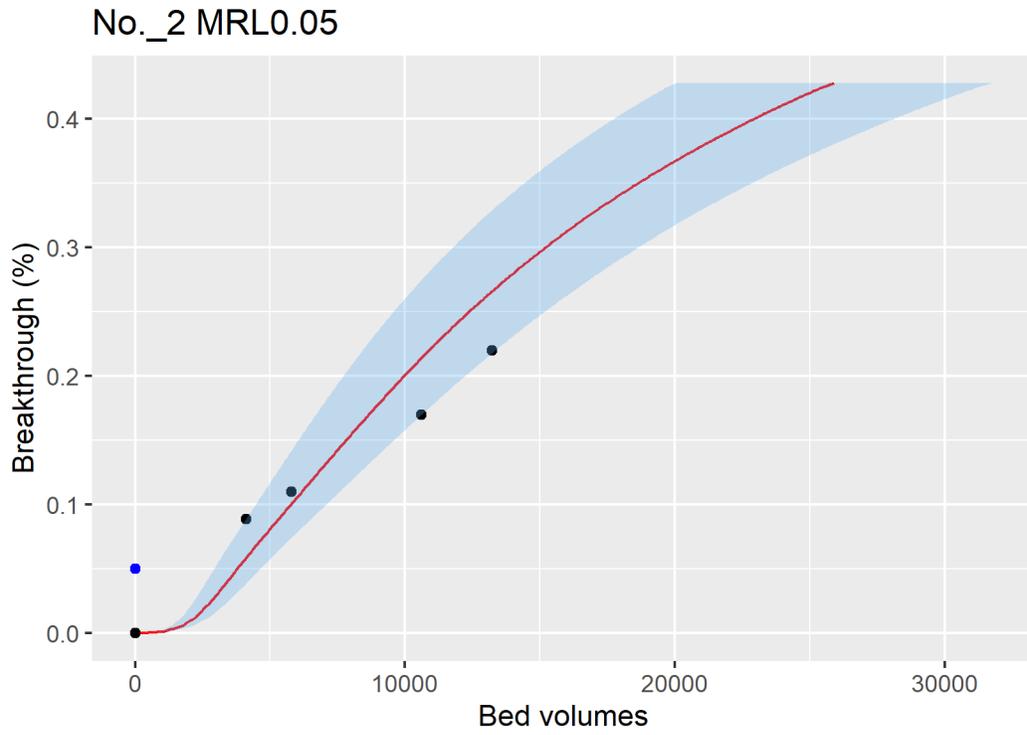


Figure ID 60- 2

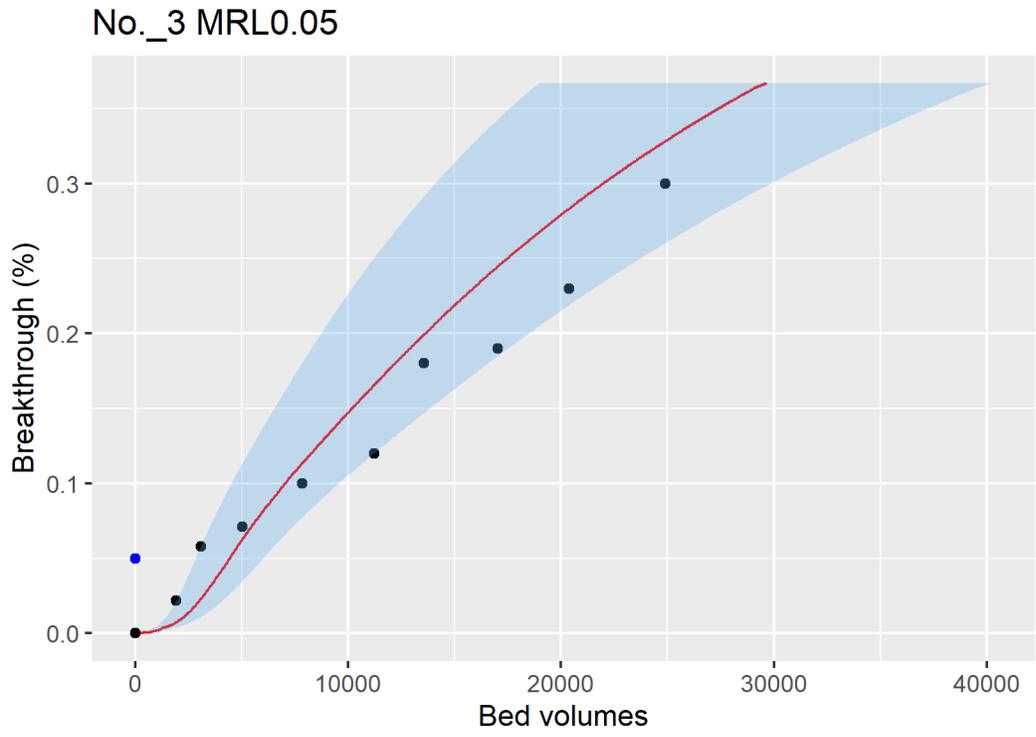


Figure ID 60- 3

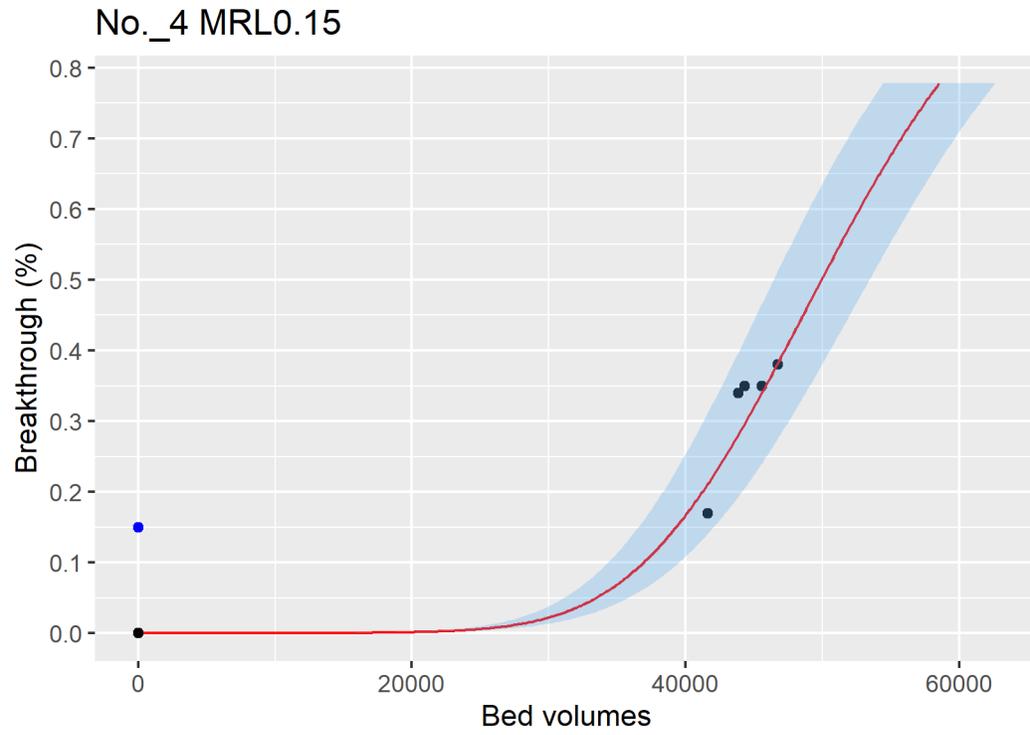


Figure ID 60- 1

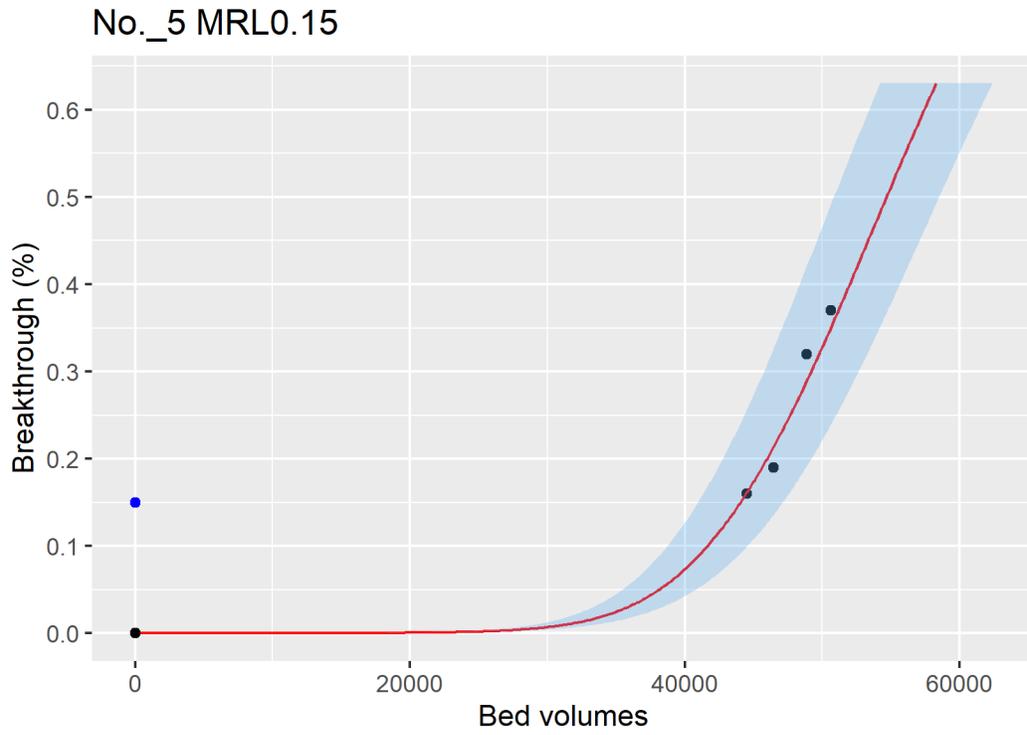


Figure ID 60- 2

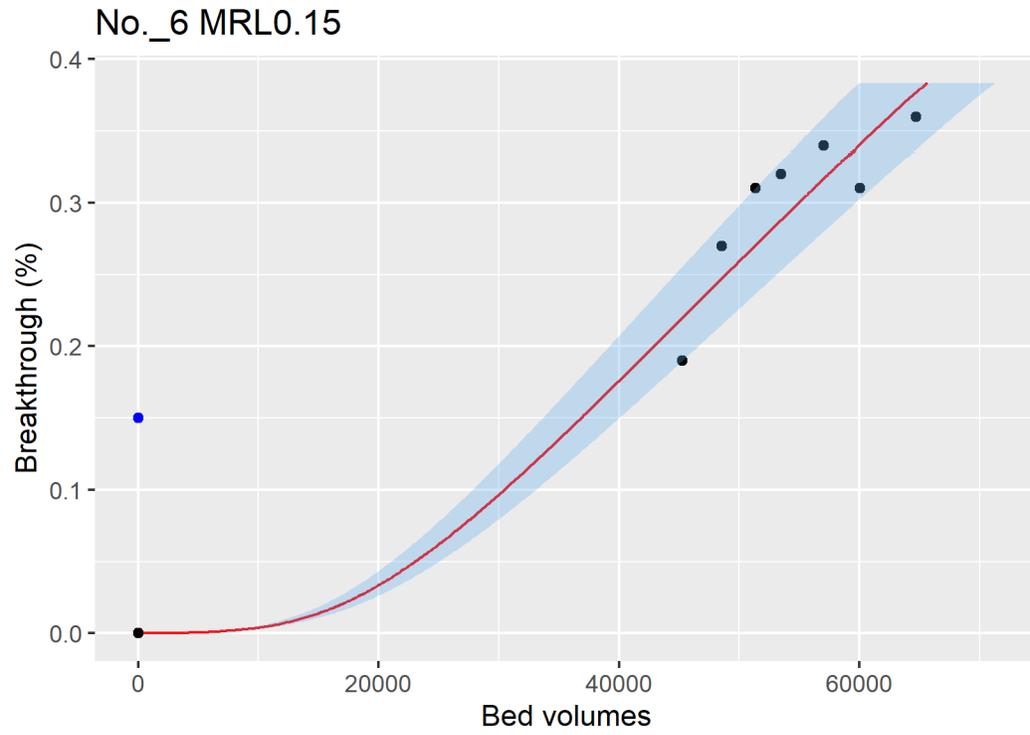


Figure ID 60- 3

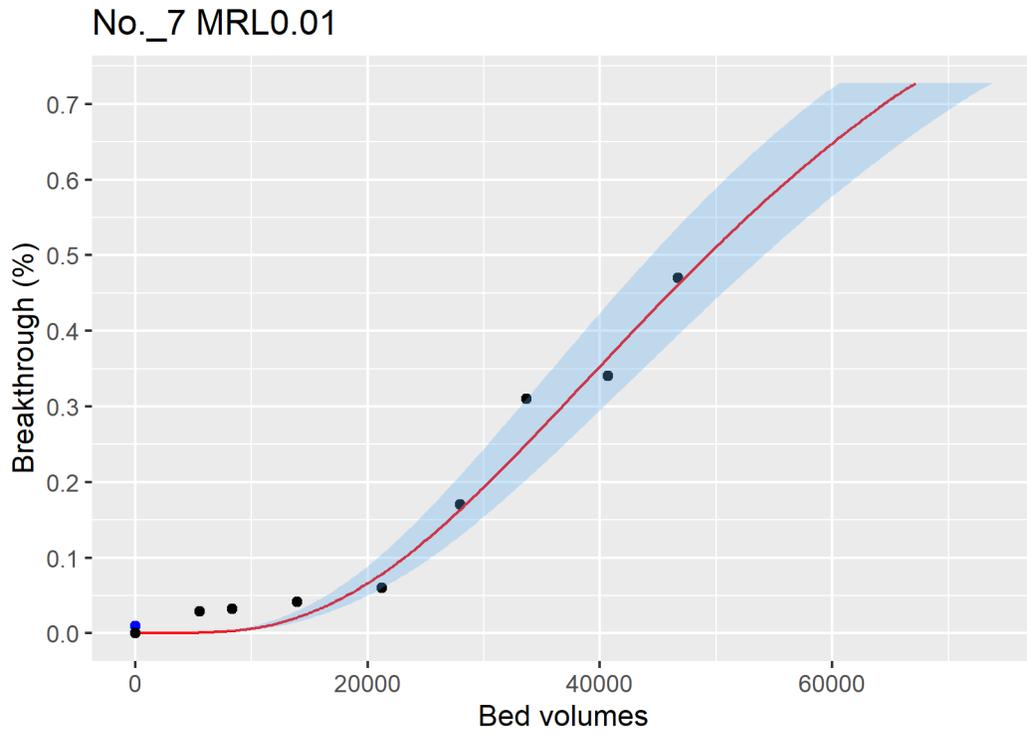


Figure ID 60- 4

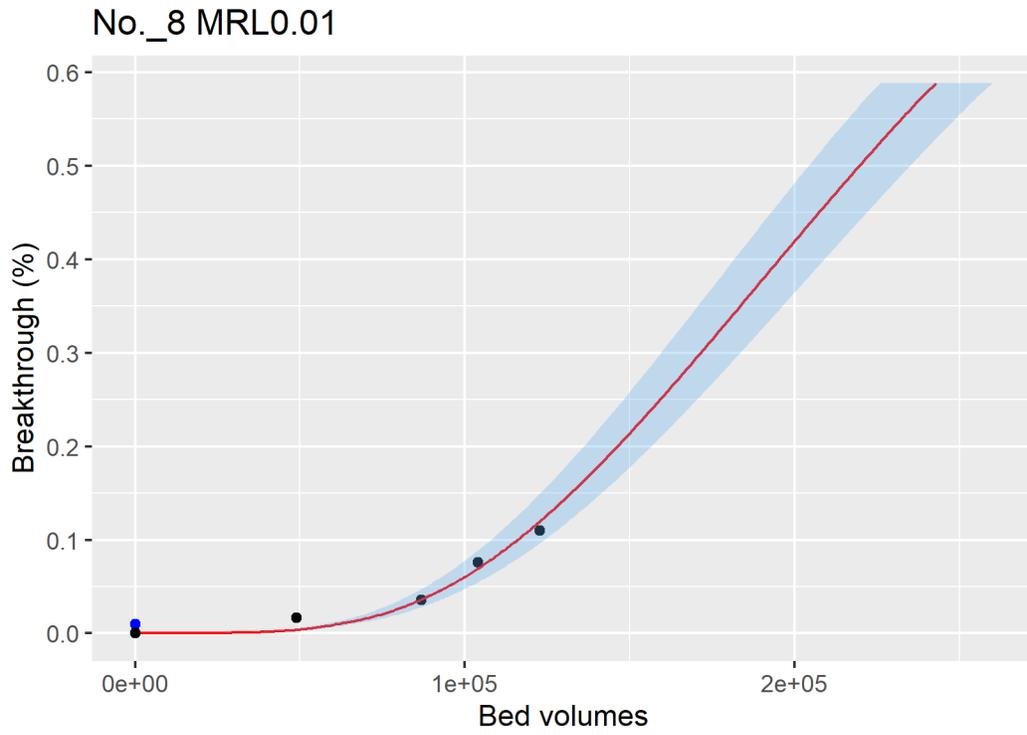


Figure ID 60- 5

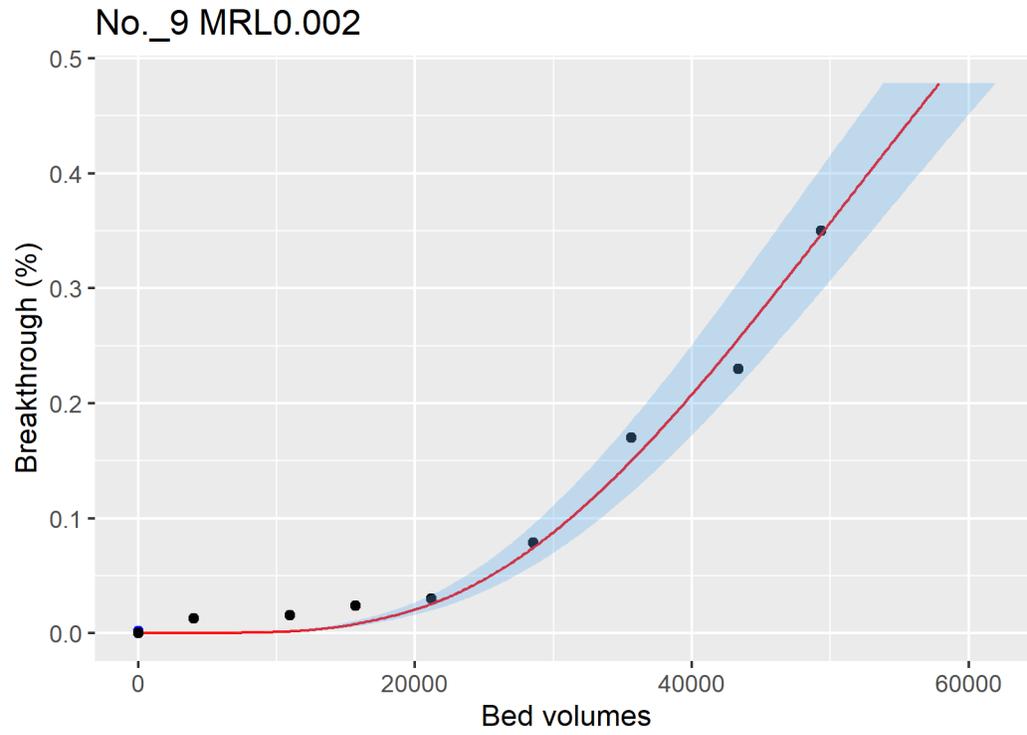


Figure ID 60- 6

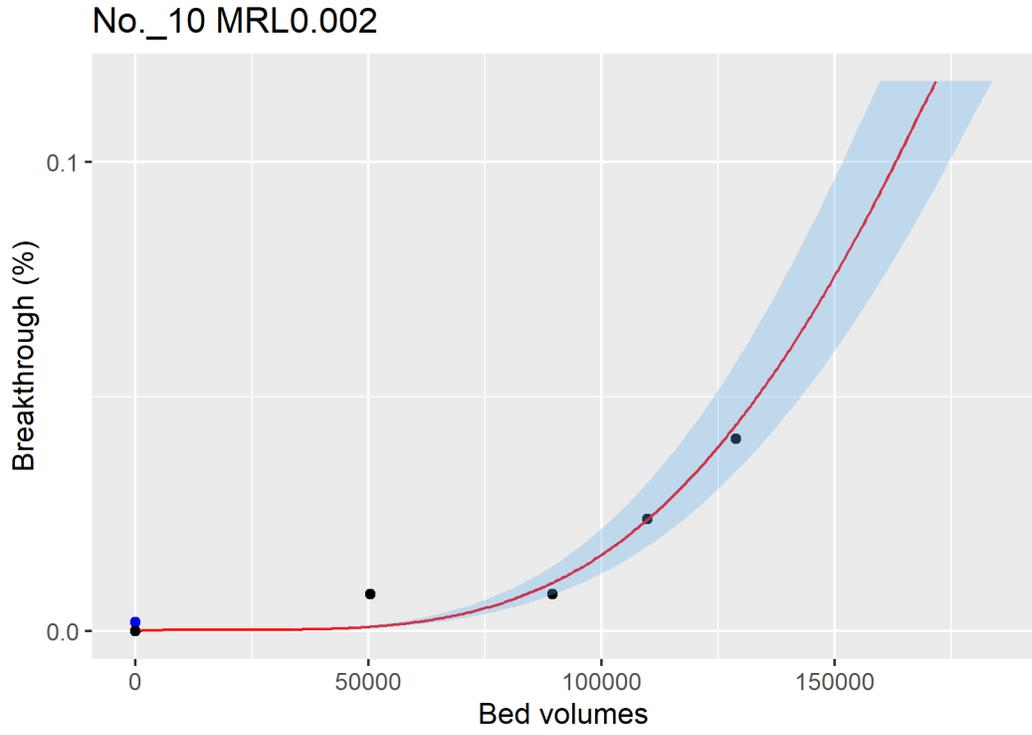


Figure ID 60- 7

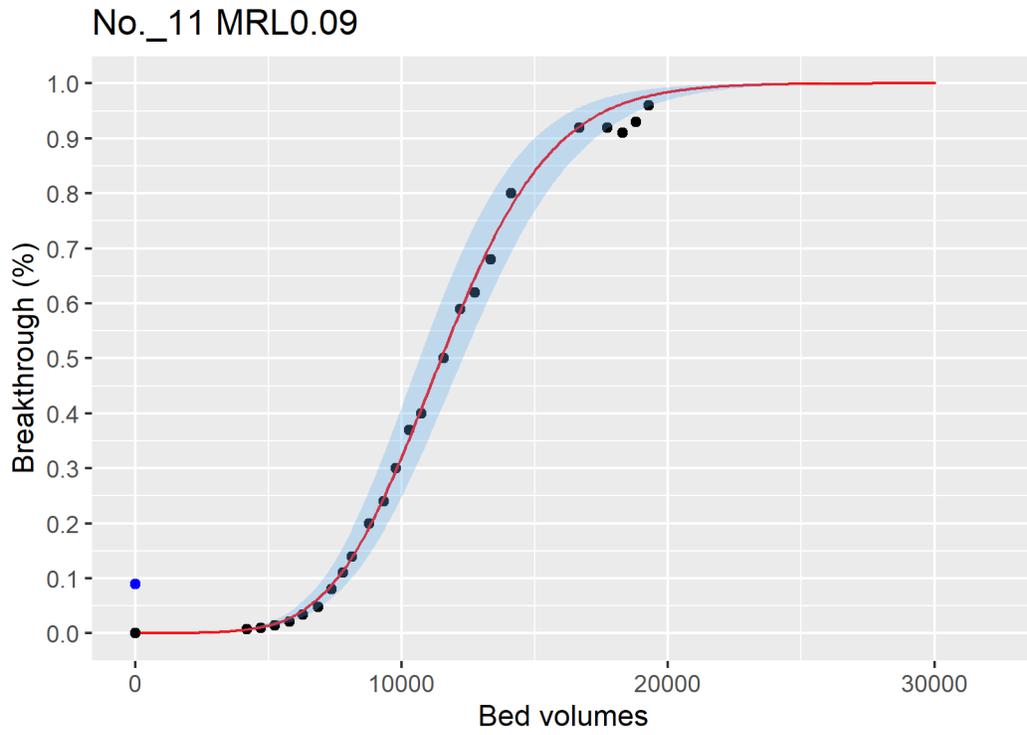


Figure ID 60- 8

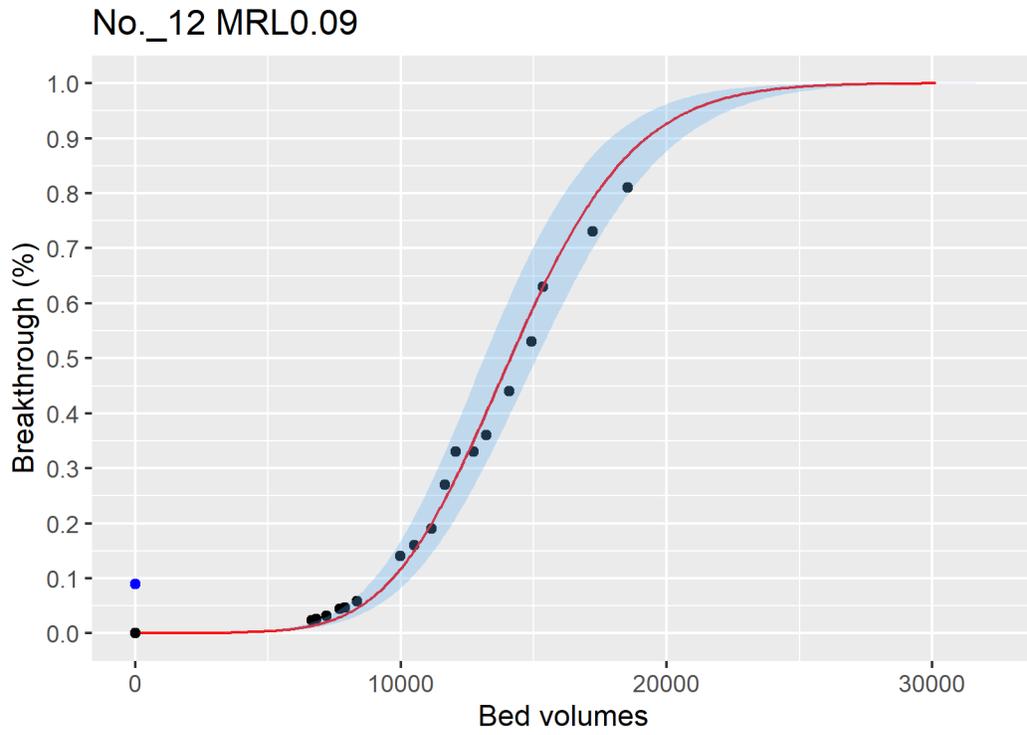


Figure ID 60-9

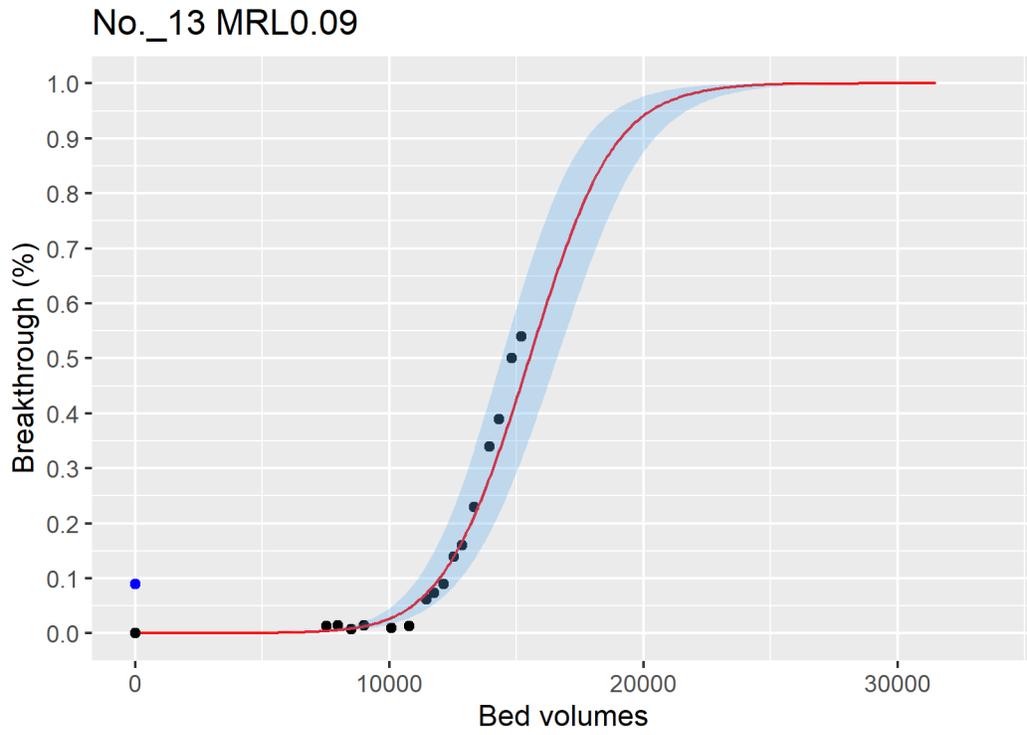


Figure ID 60- 10

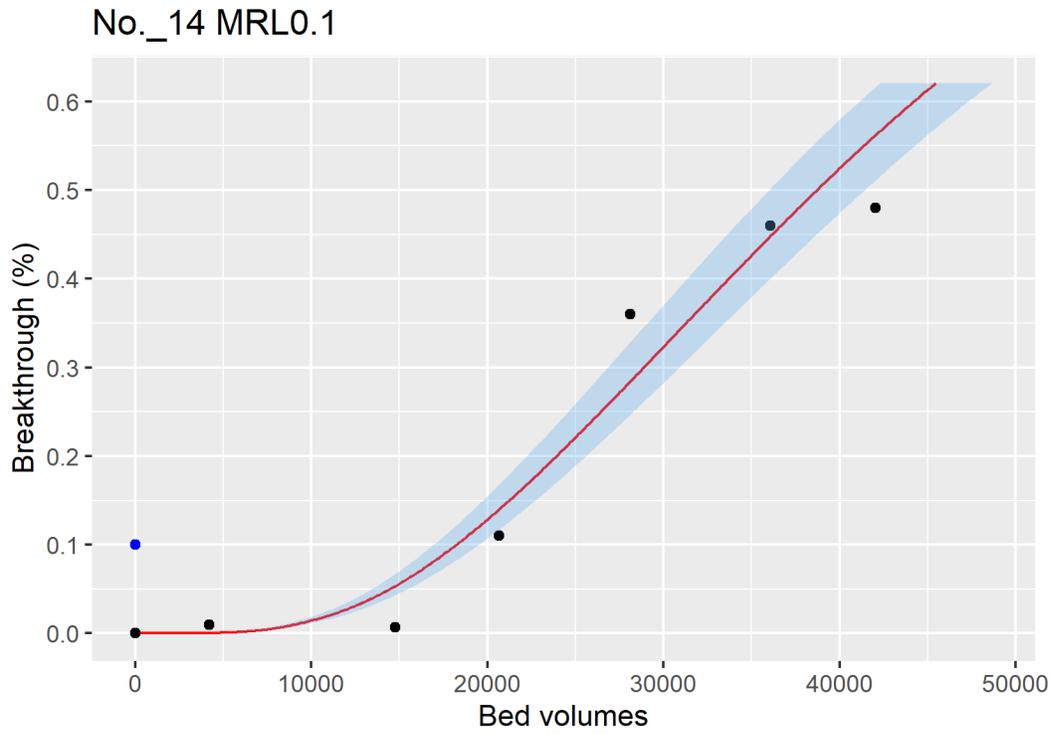


Figure ID 60- 11

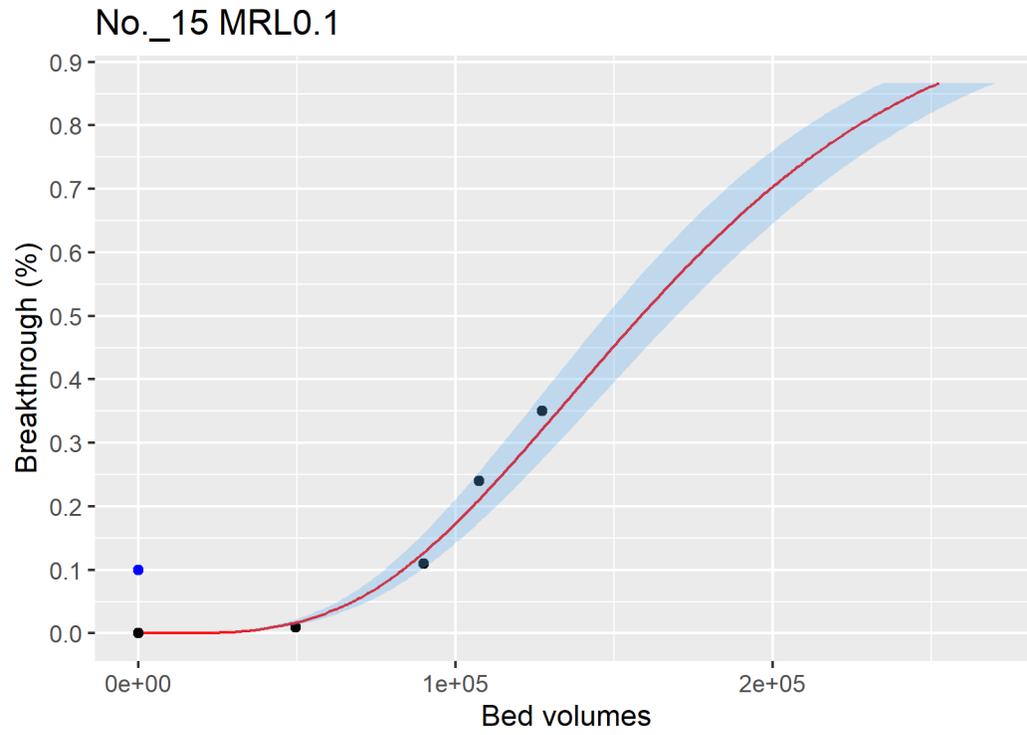


Figure ID 60- 12

Breakthrough data of Ref ID 61

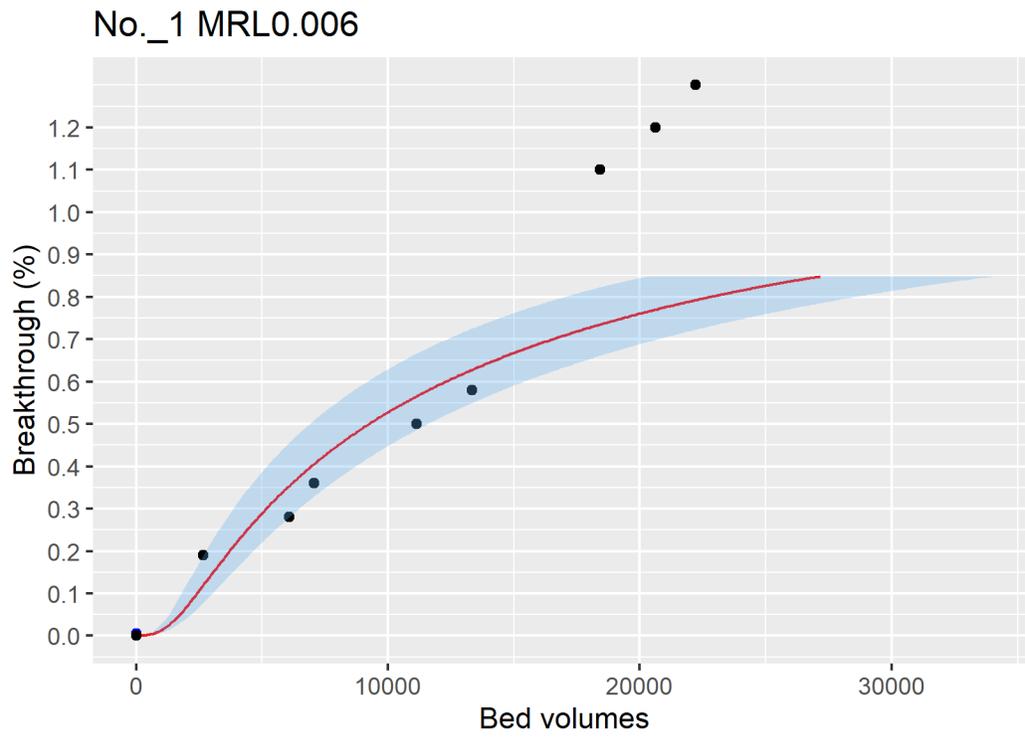


Figure ID 61- 1

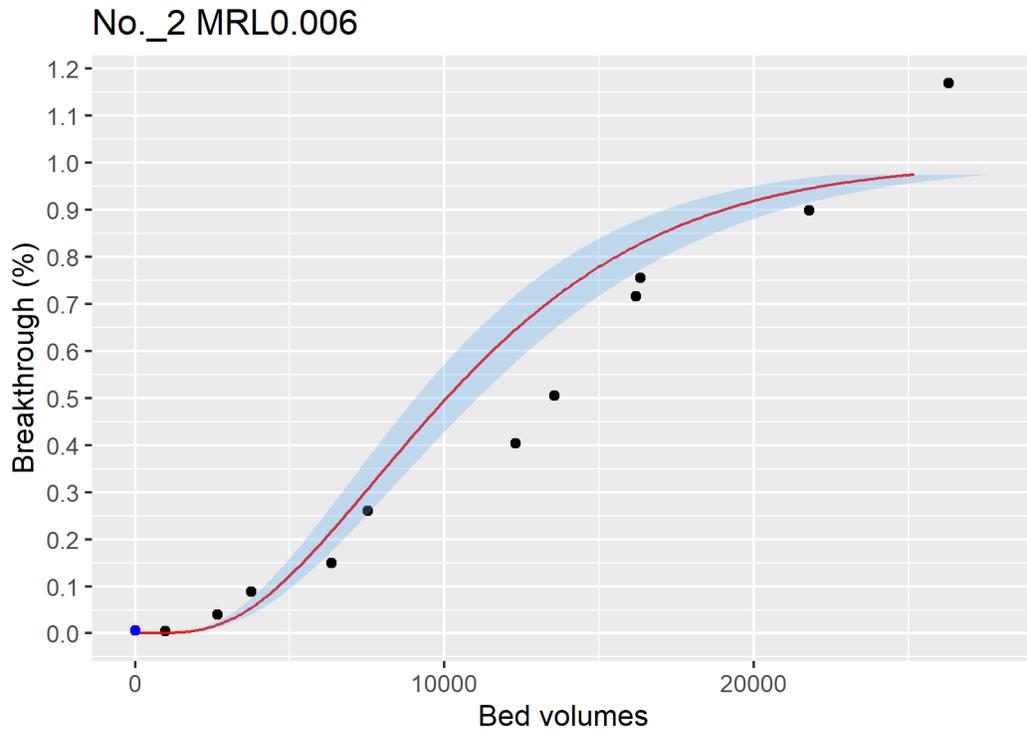


Figure ID 61- 2

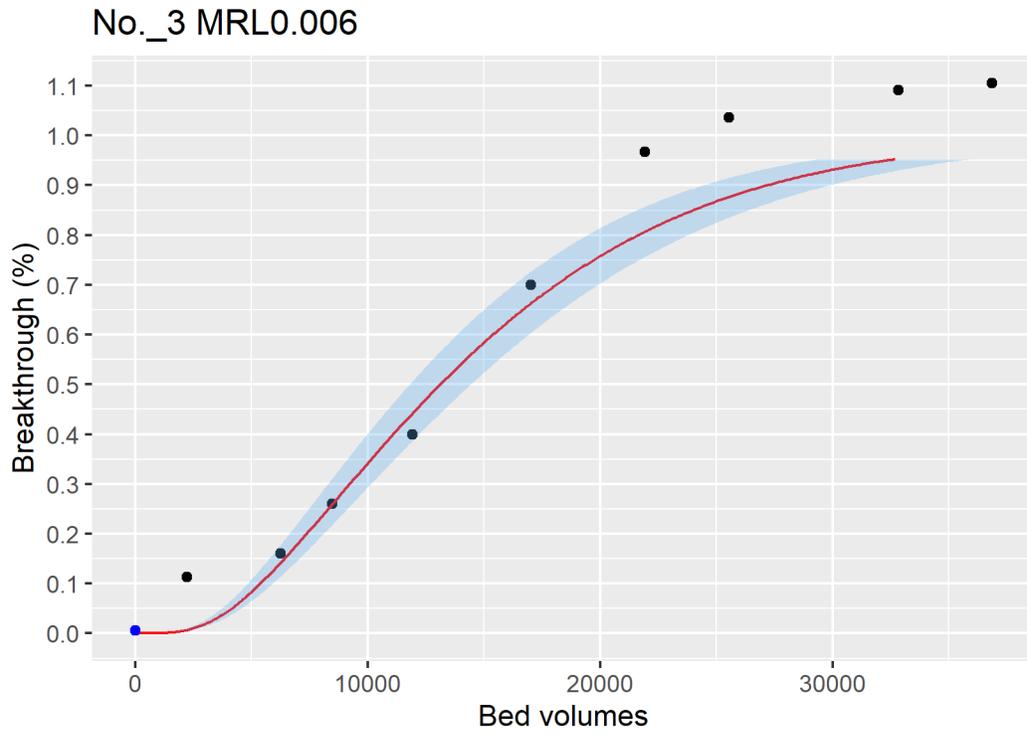


Figure ID 61- 3

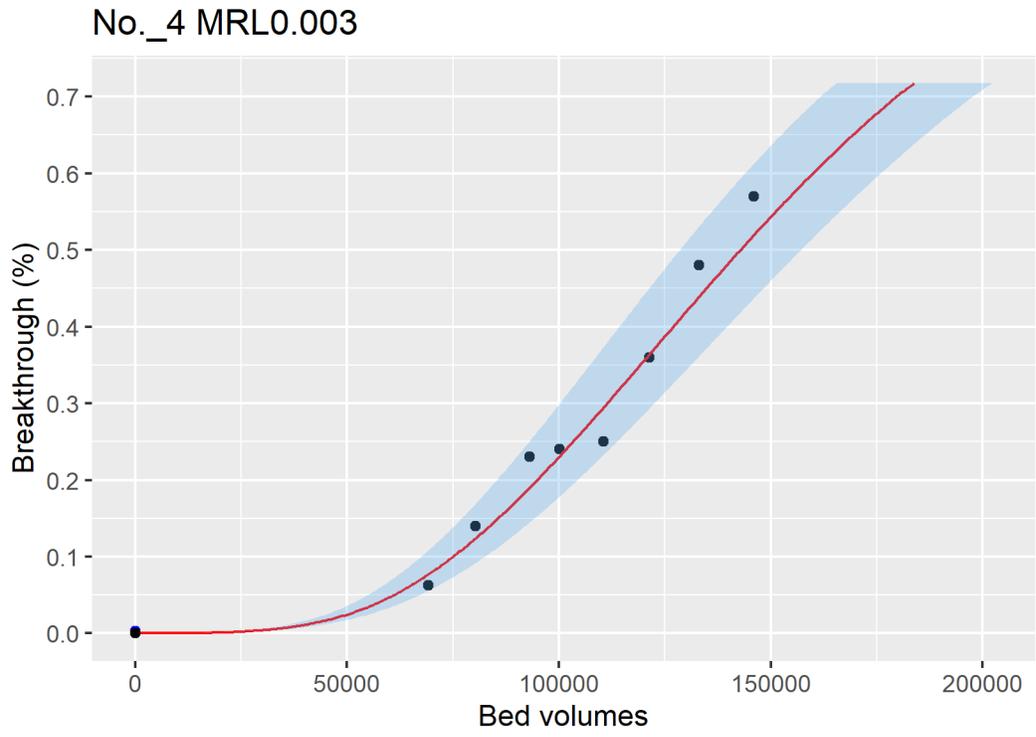


Figure ID 61- 4

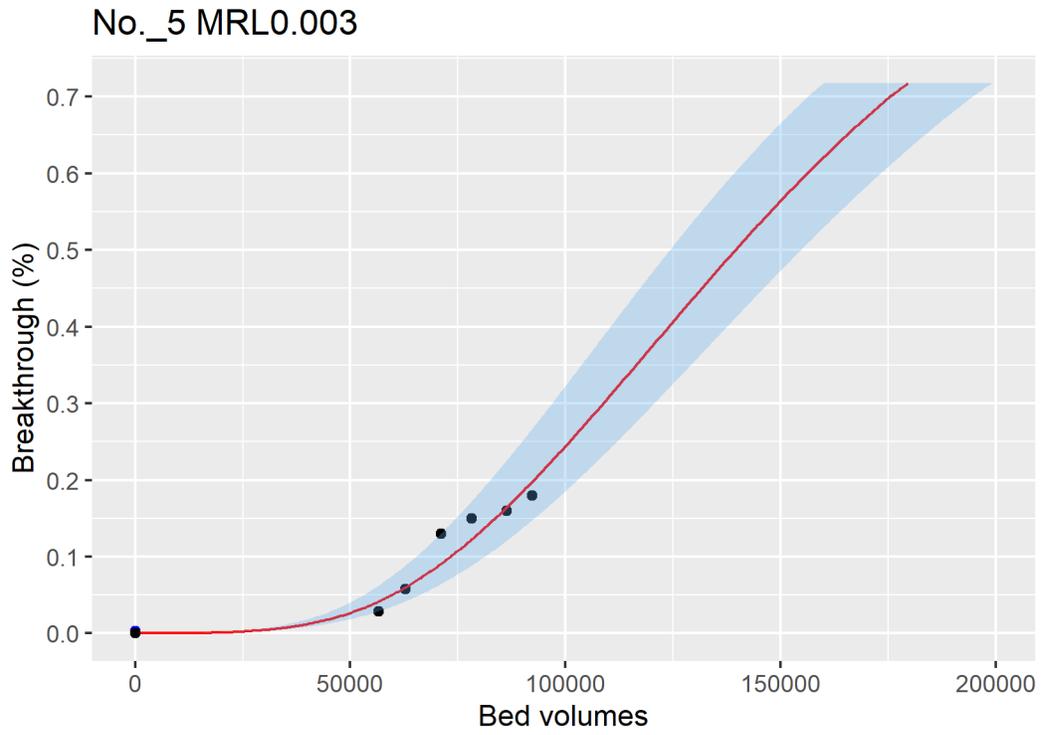


Figure ID 61- 5

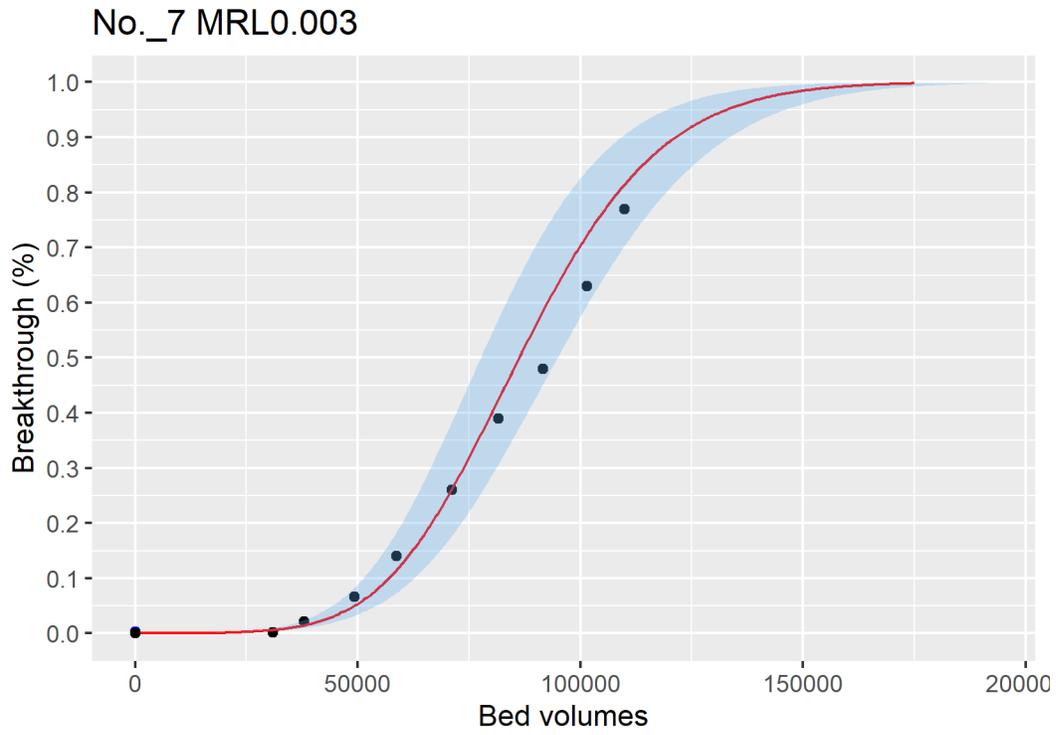


Figure ID 61- 6

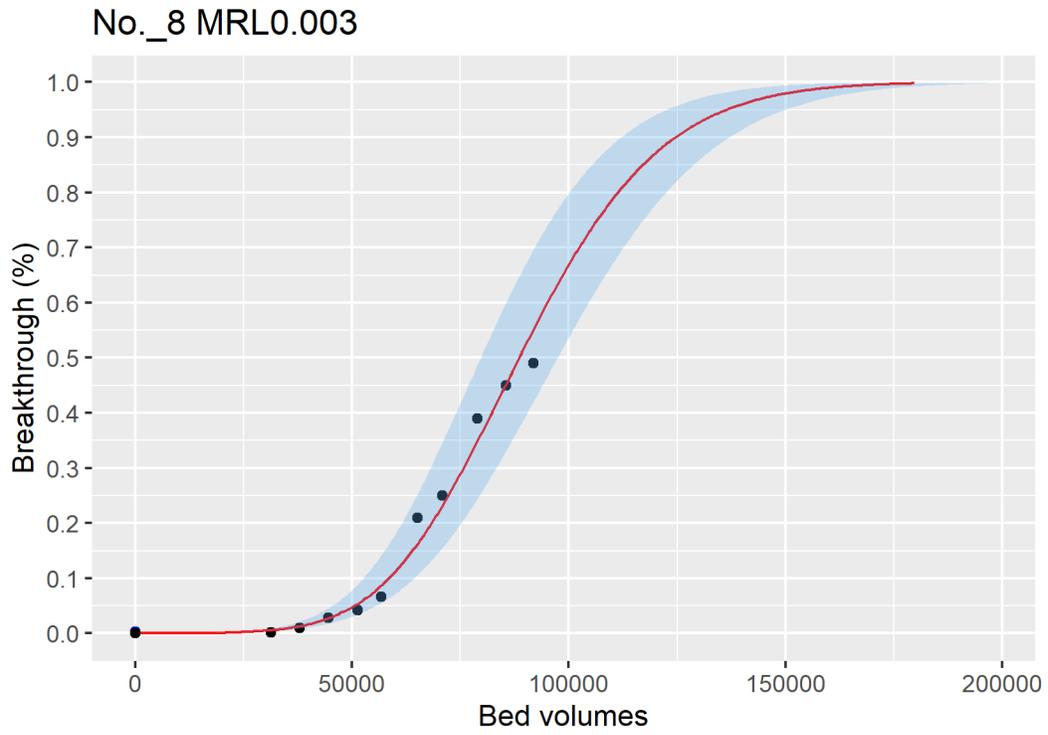


Figure ID 61- 7

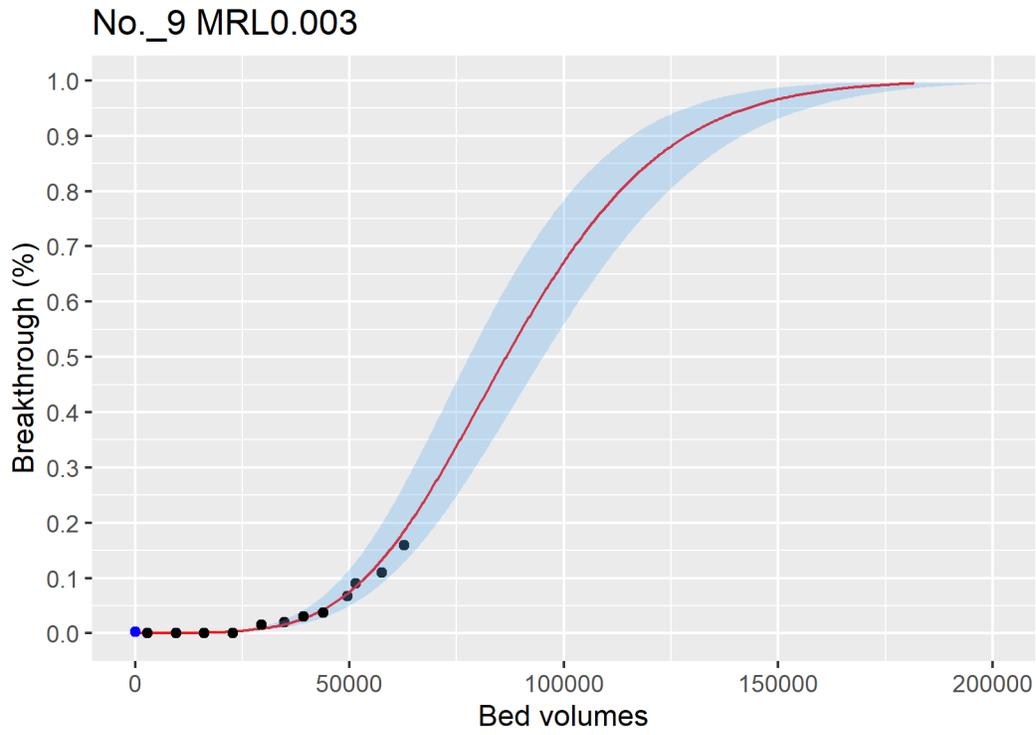


Figure ID 61- 8

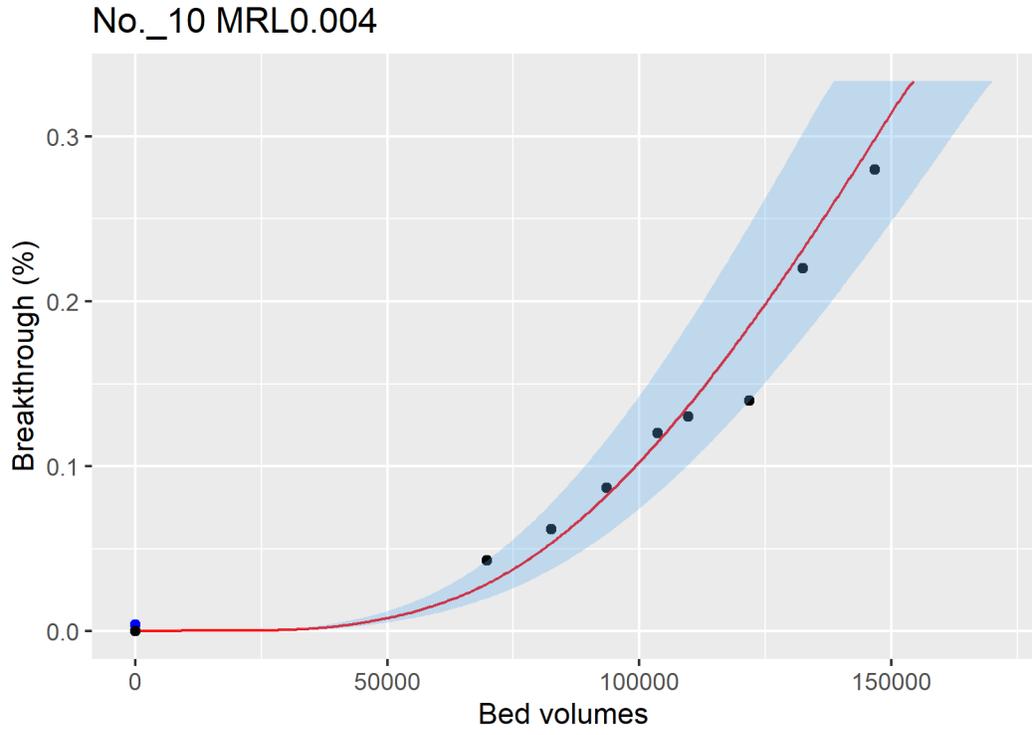


Figure ID 61- 9

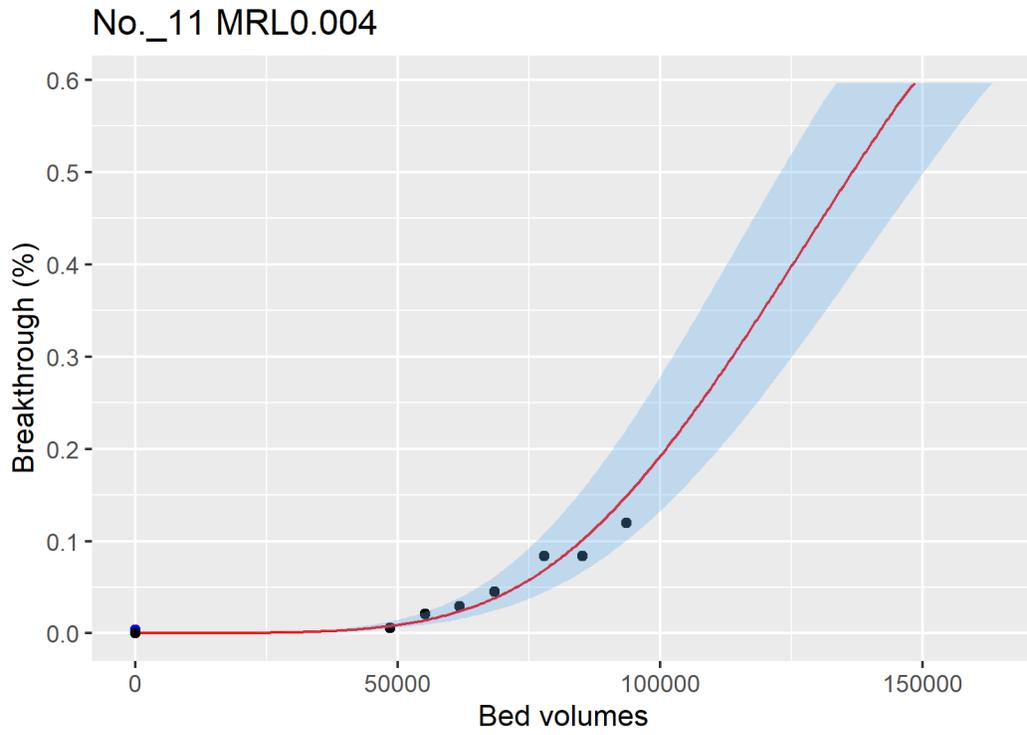


Figure ID 61- 10

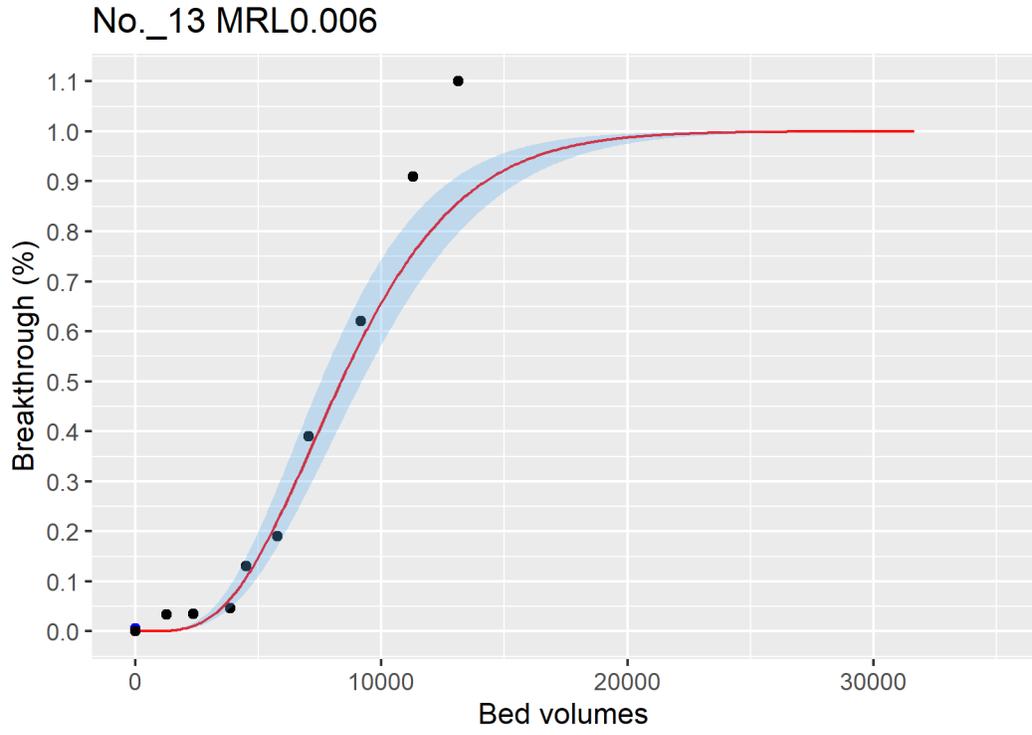


Figure ID 61- 11

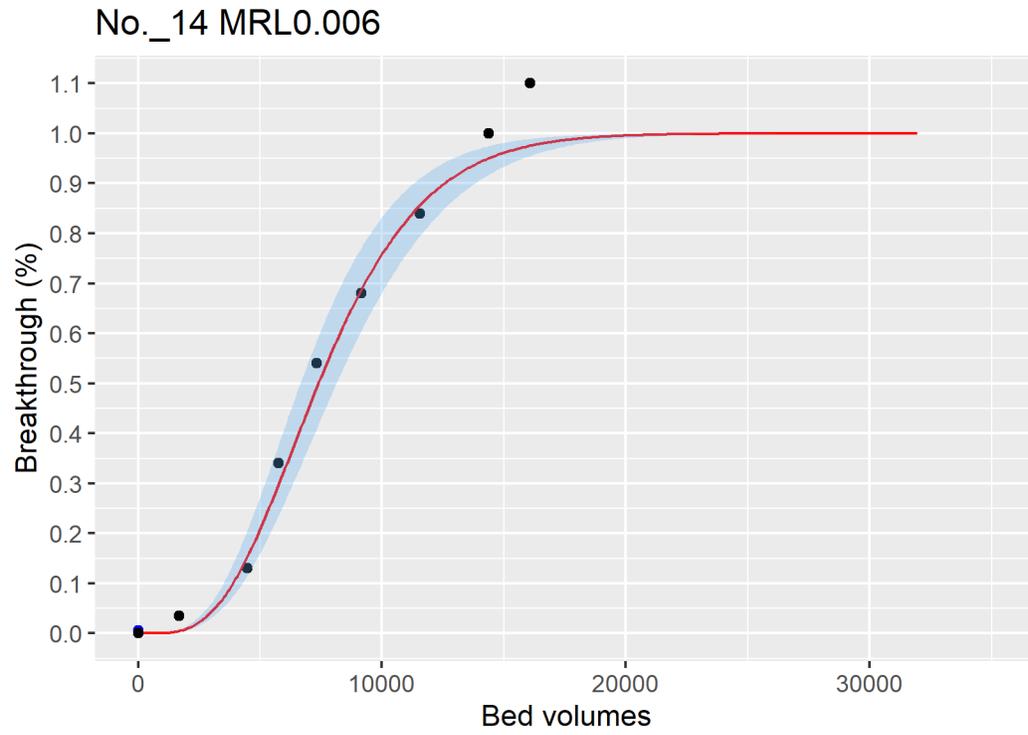


Figure ID 61- 12

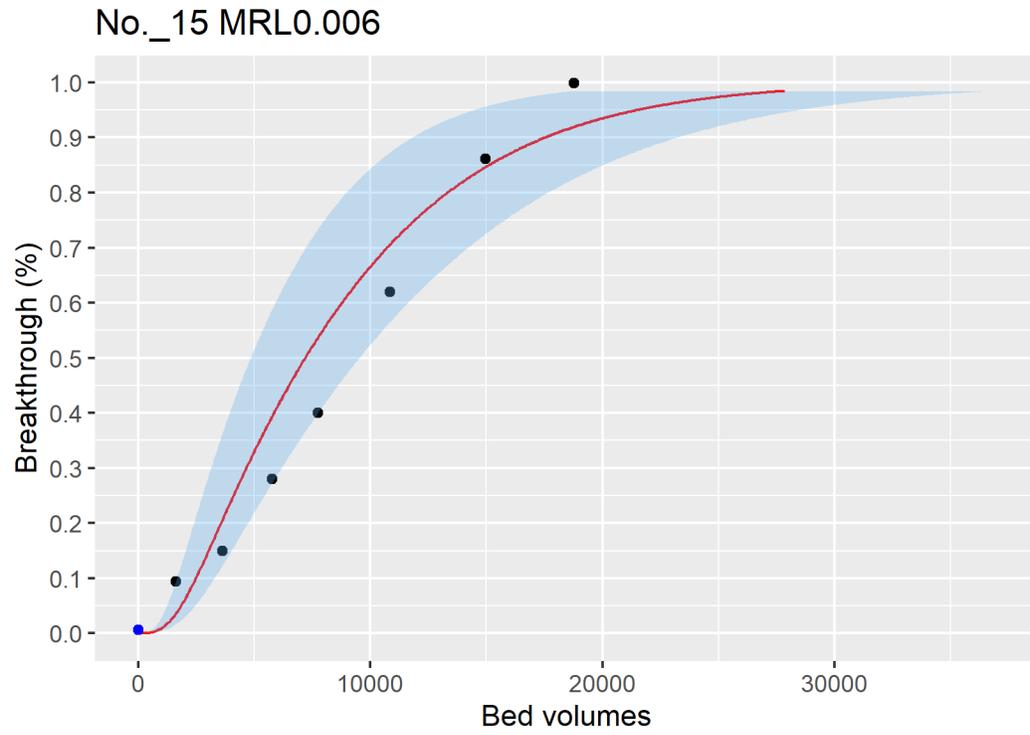


Figure ID 61- 13

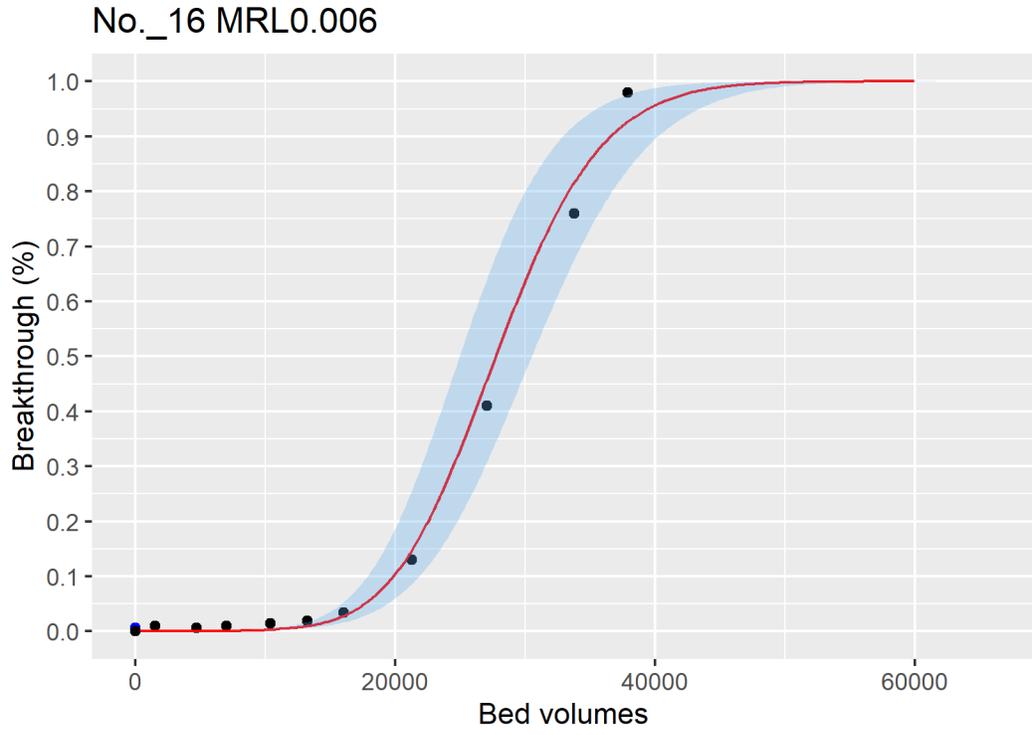


Figure ID 61- 14

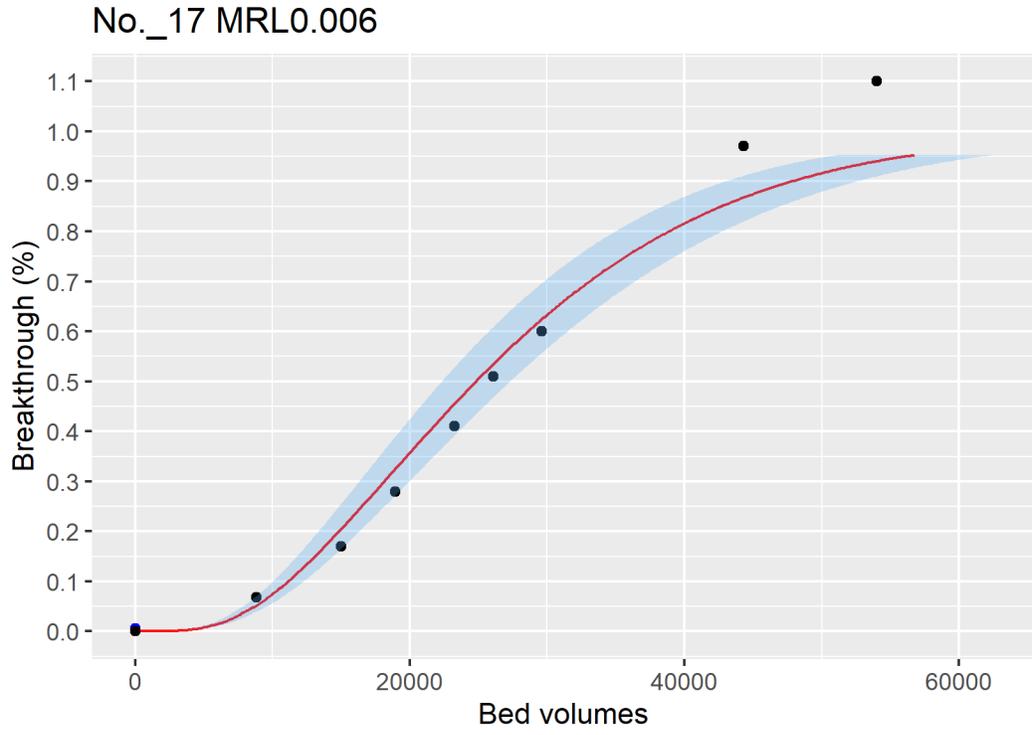


Figure ID 61- 15

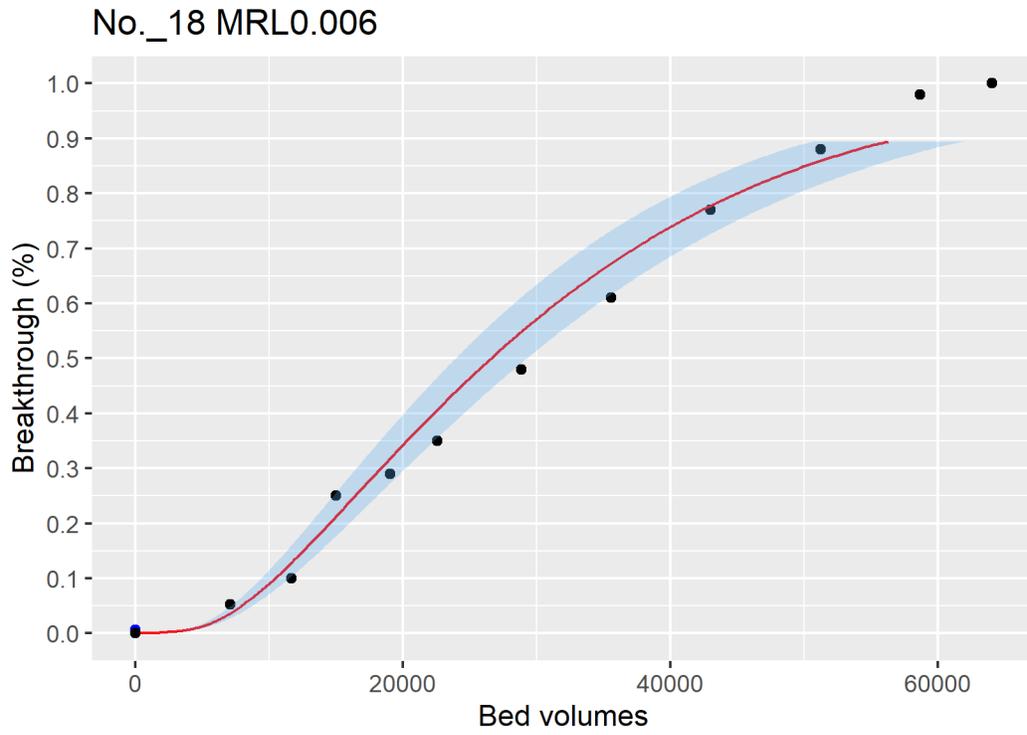


Figure ID 61- 16

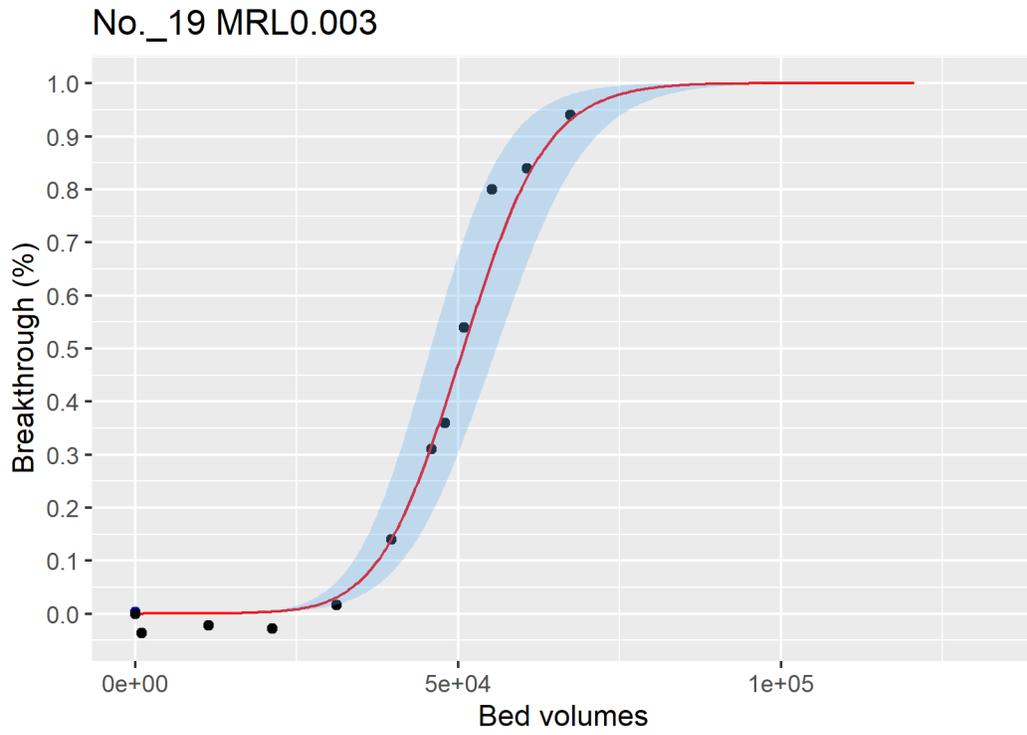


Figure ID 61- 17

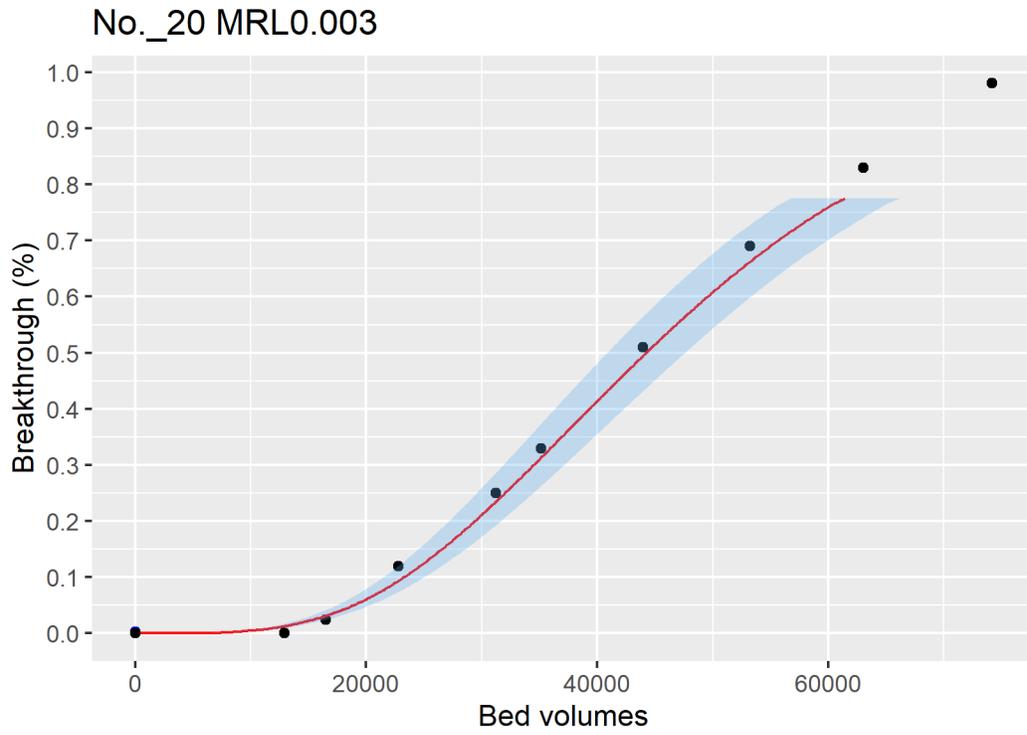


Figure ID 61- 18

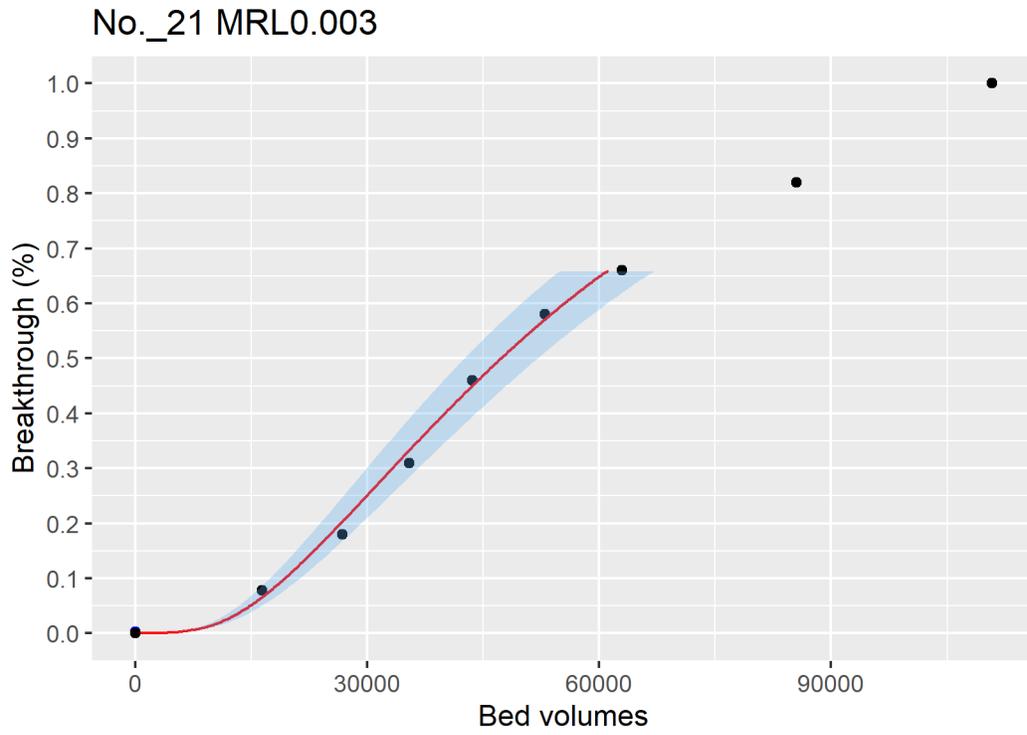


Figure ID 61- 19

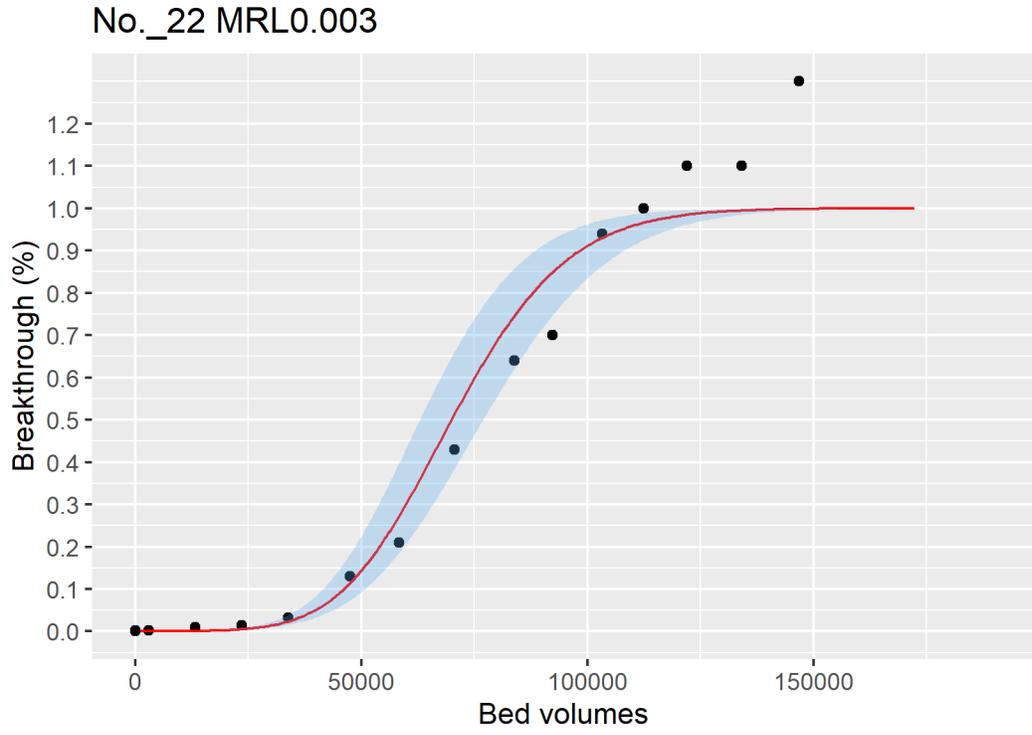


Figure ID 61- 20

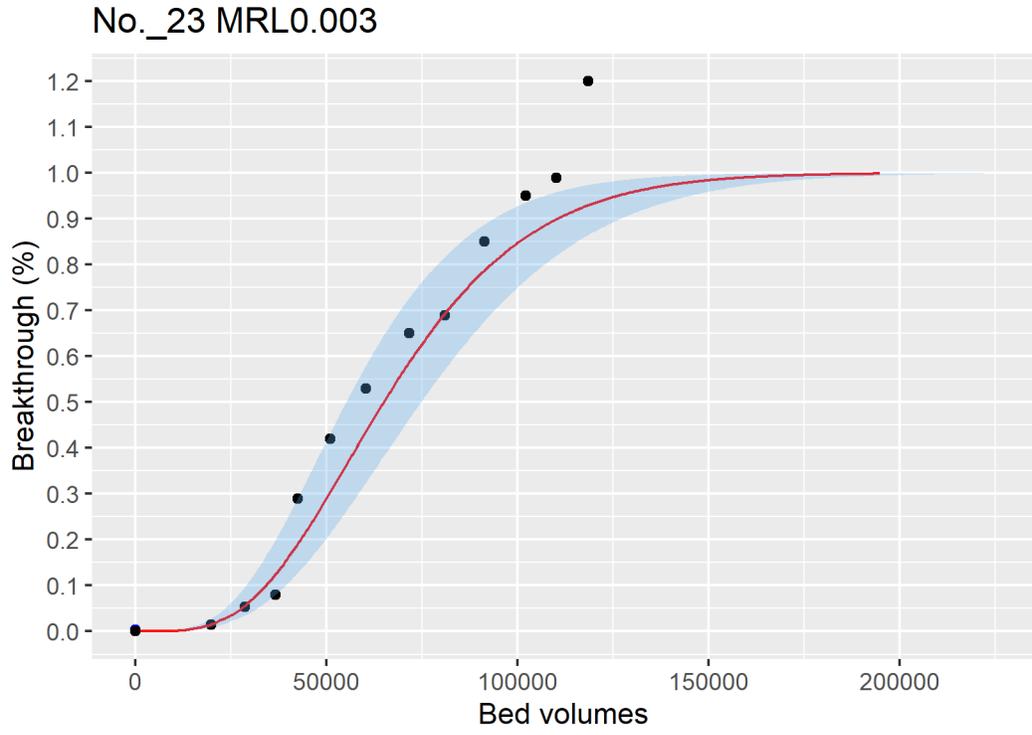


Figure ID 61- 21

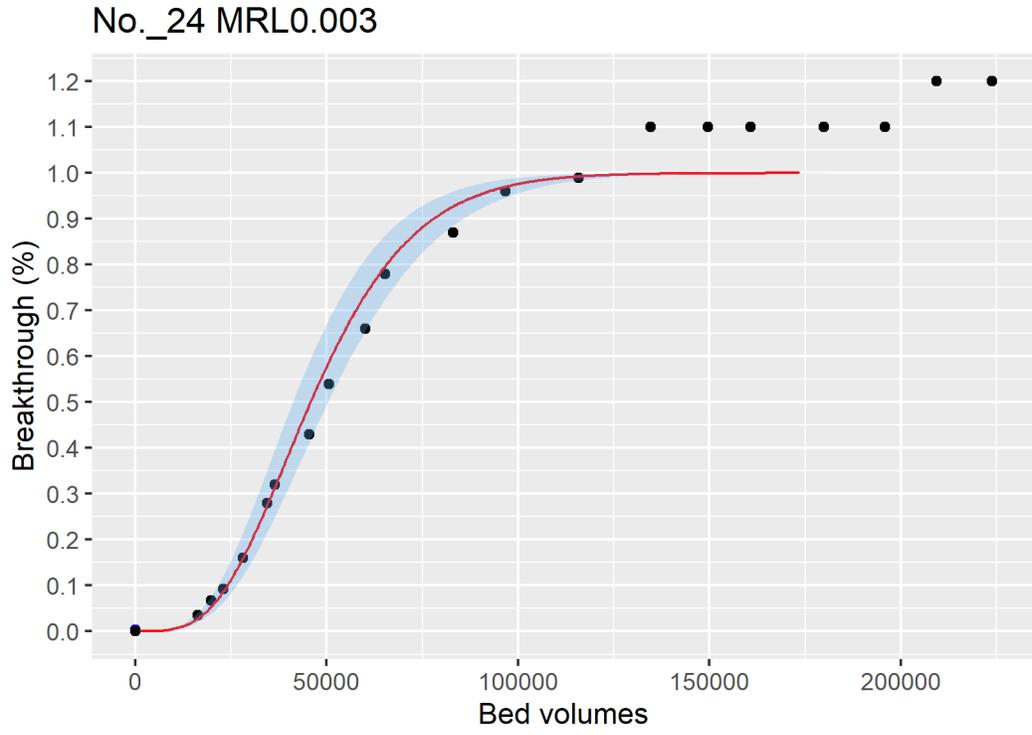


Figure ID 61- 22

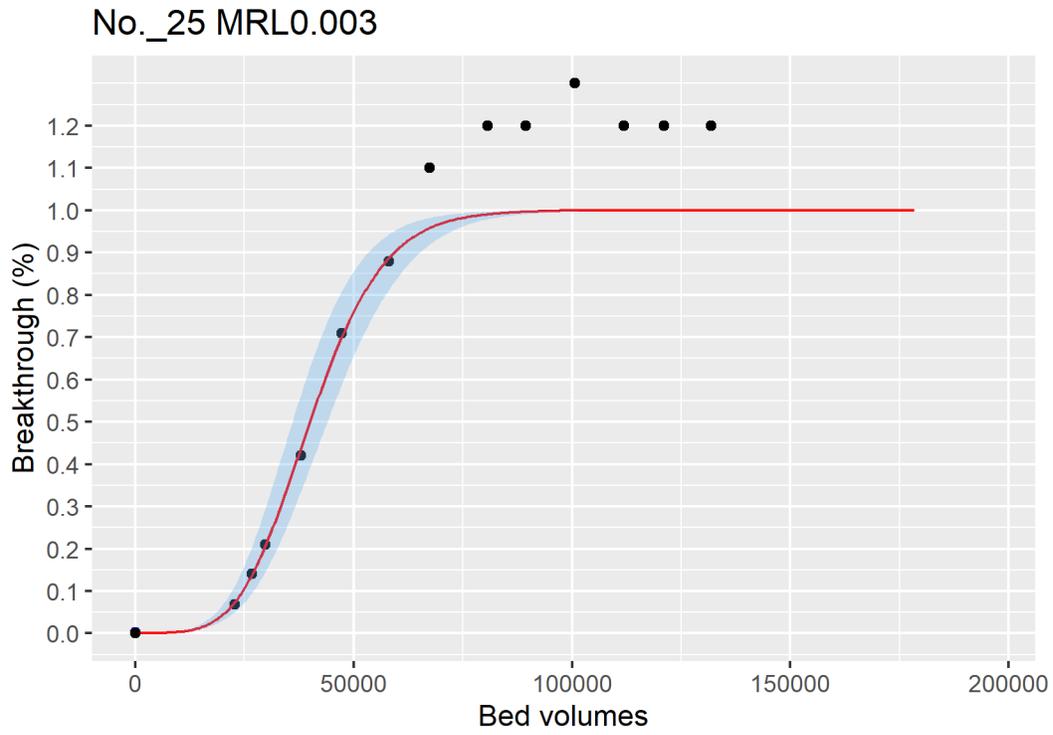


Figure ID 61- 23

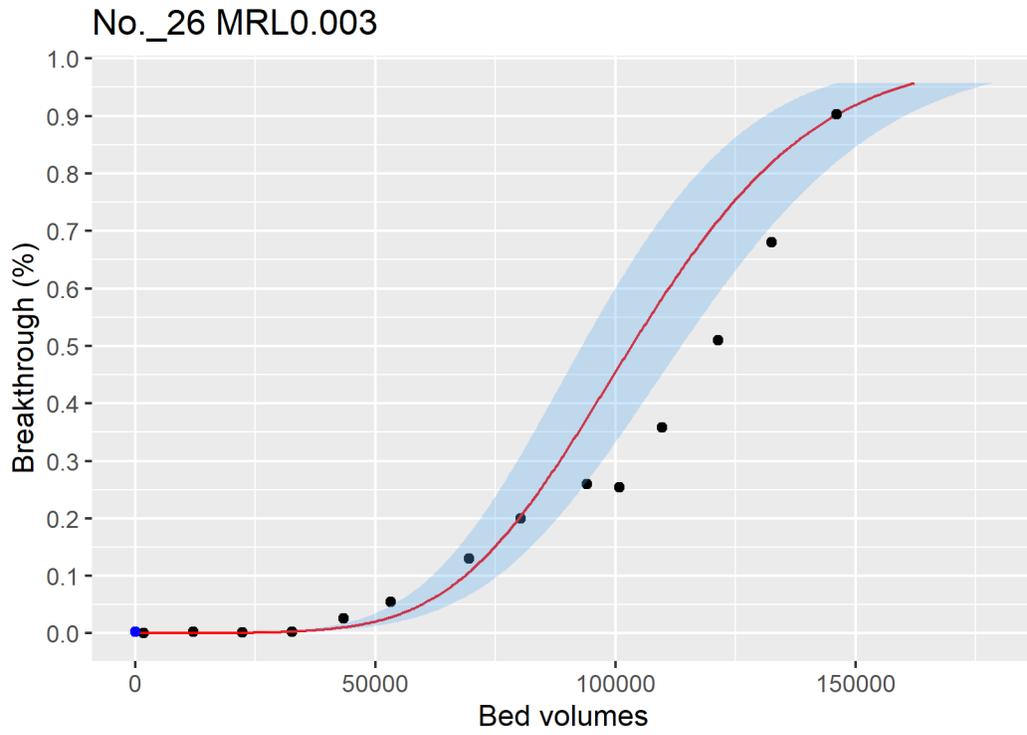


Figure ID 61- 24

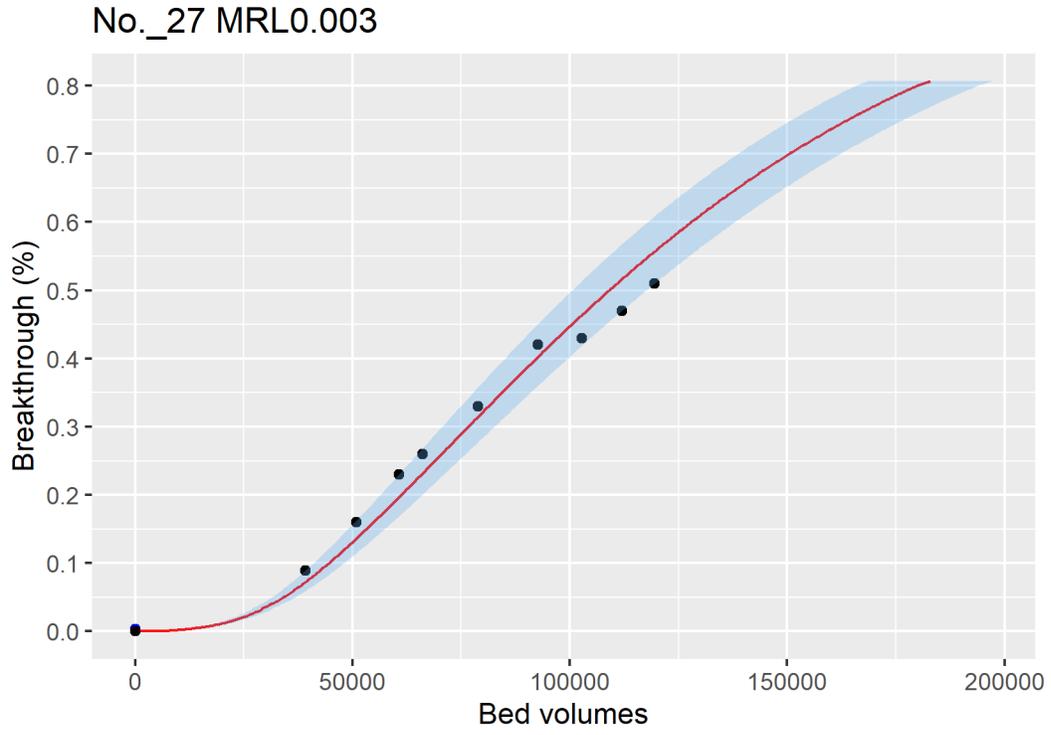


Figure ID 61- 25

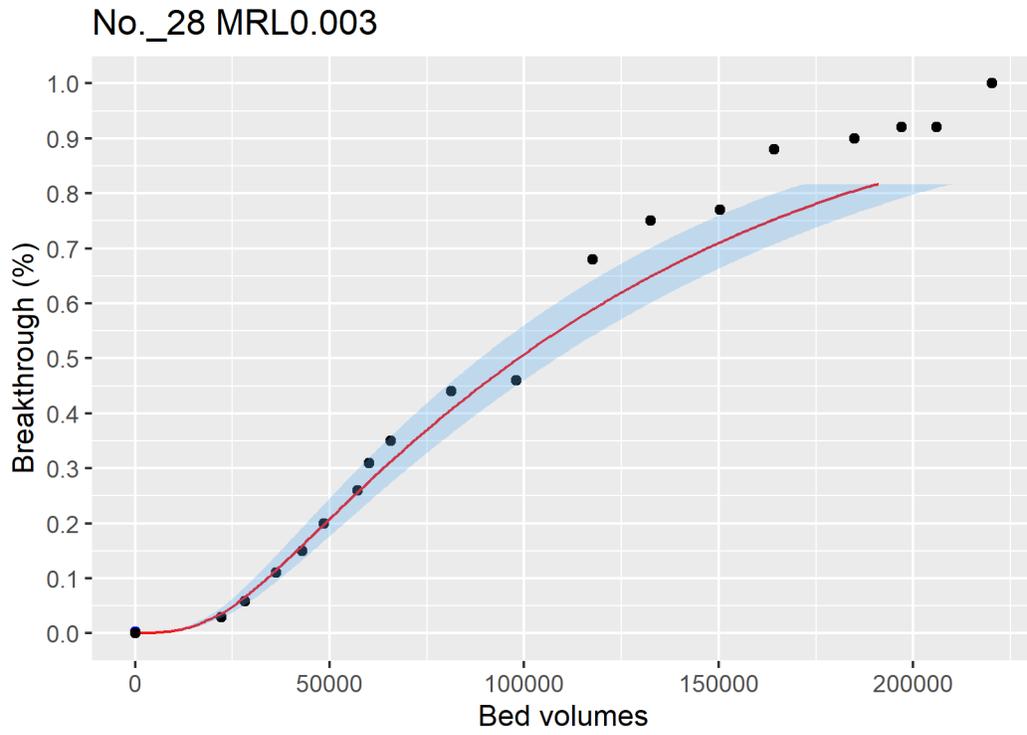


Figure ID 61- 26

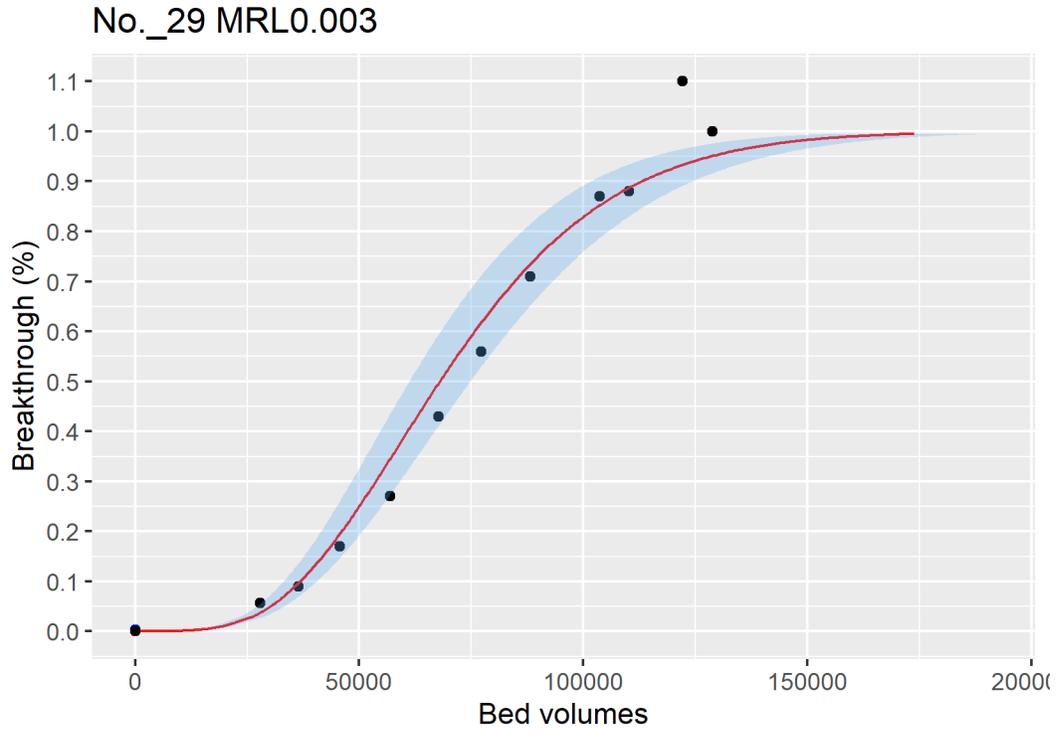


Figure ID 61- 27

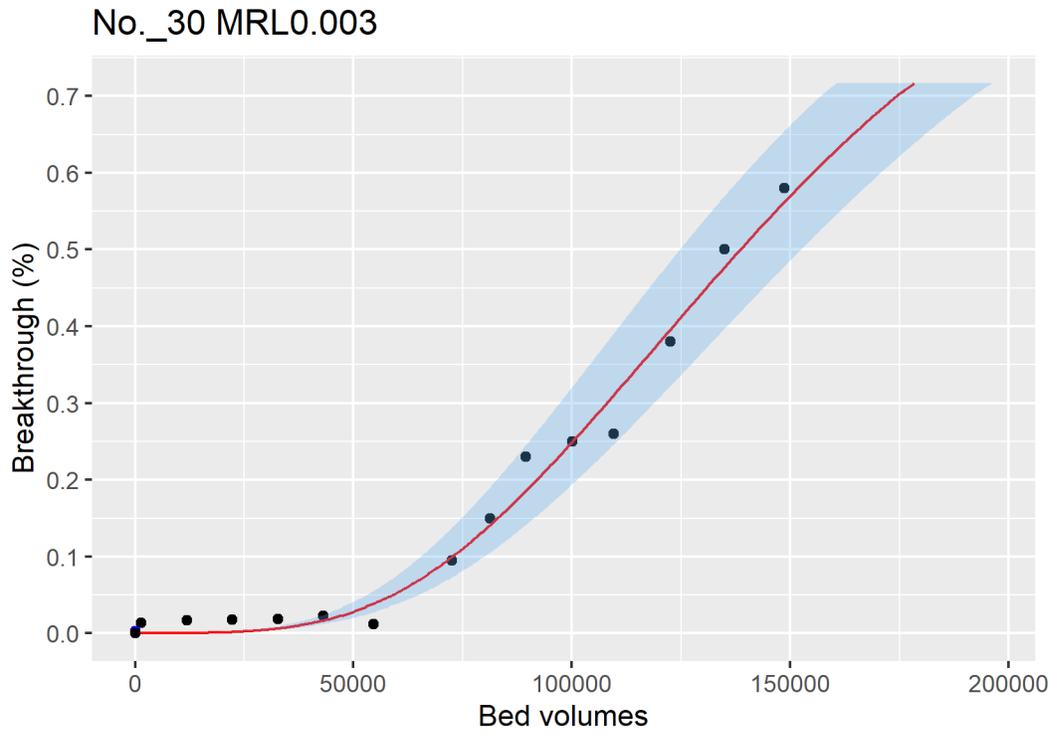


Figure ID 61- 28

No._31 MRL0.003

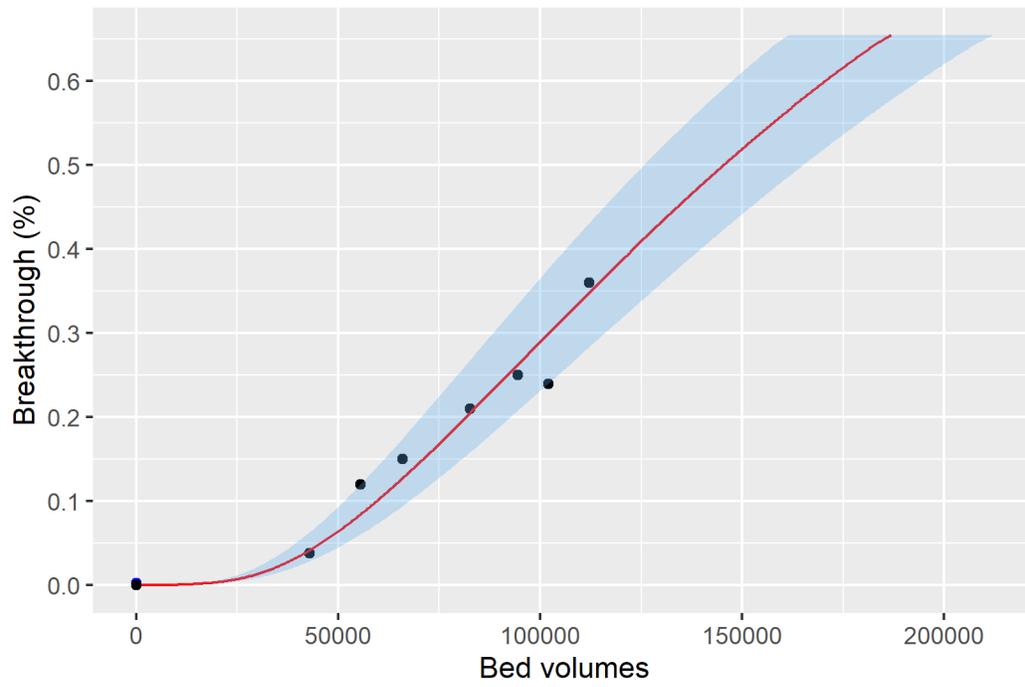


Figure ID 61- 29

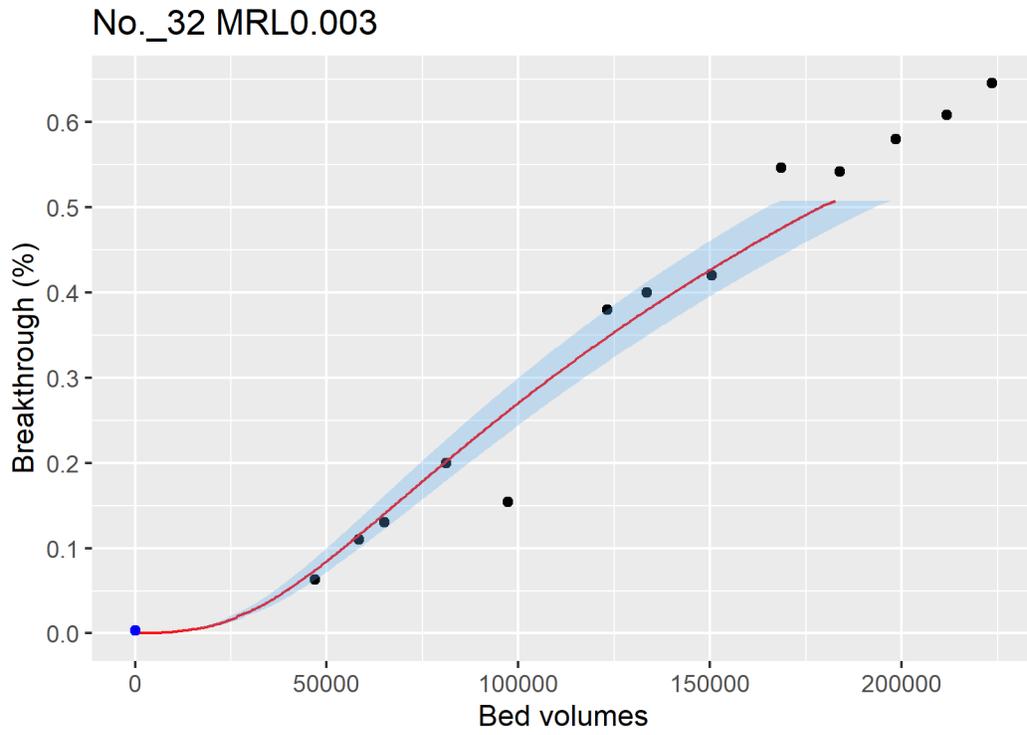


Figure ID 61- 30

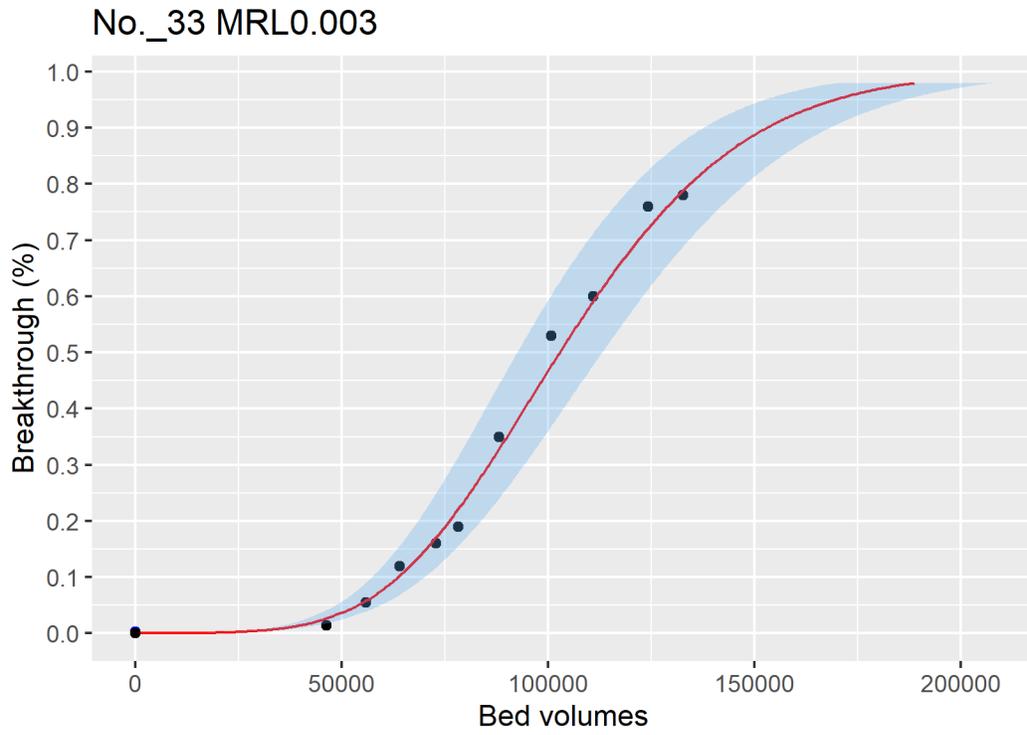


Figure ID 61- 31

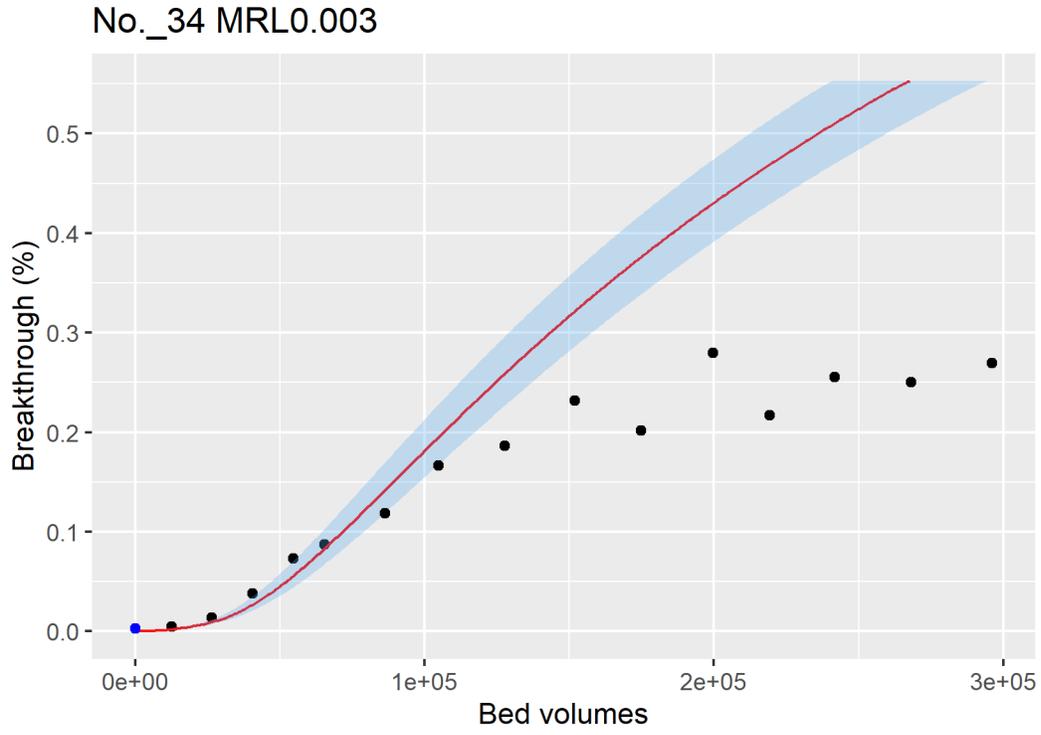


Figure ID 61- 32

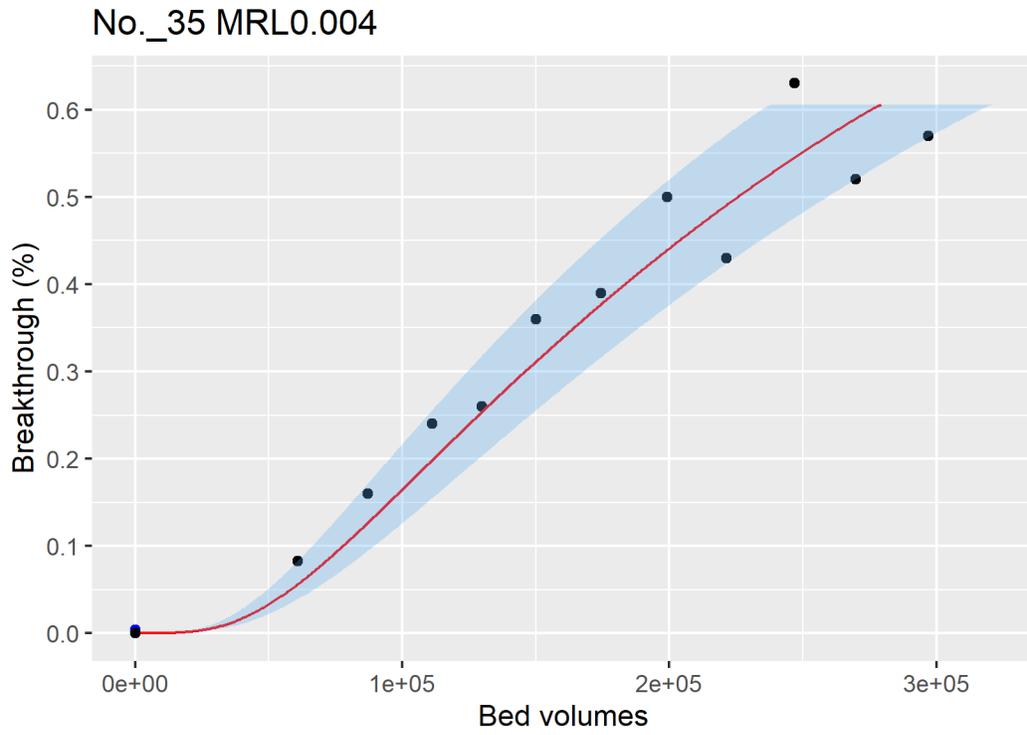


Figure ID 61- 33

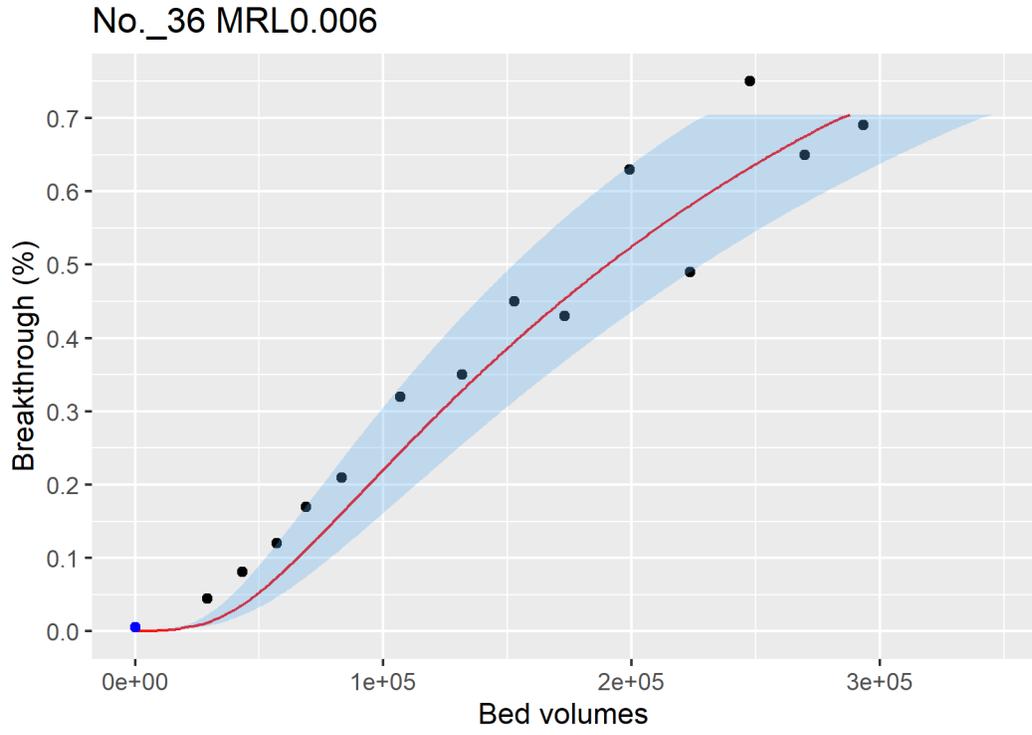


Figure ID 61- 34

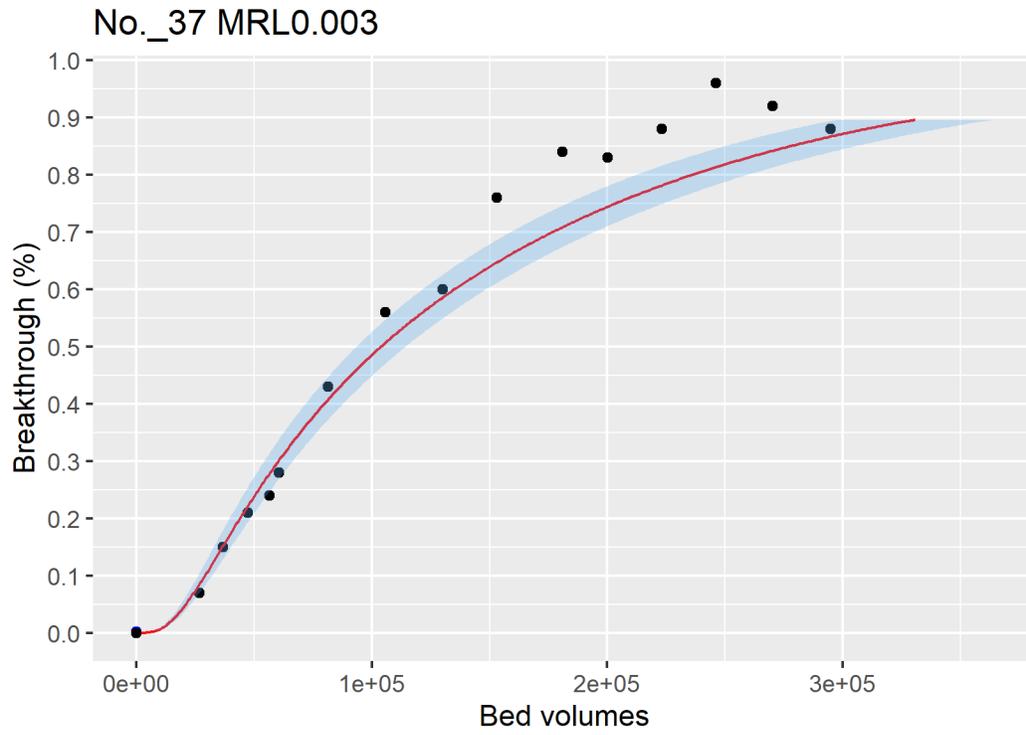


Figure ID 61- 35

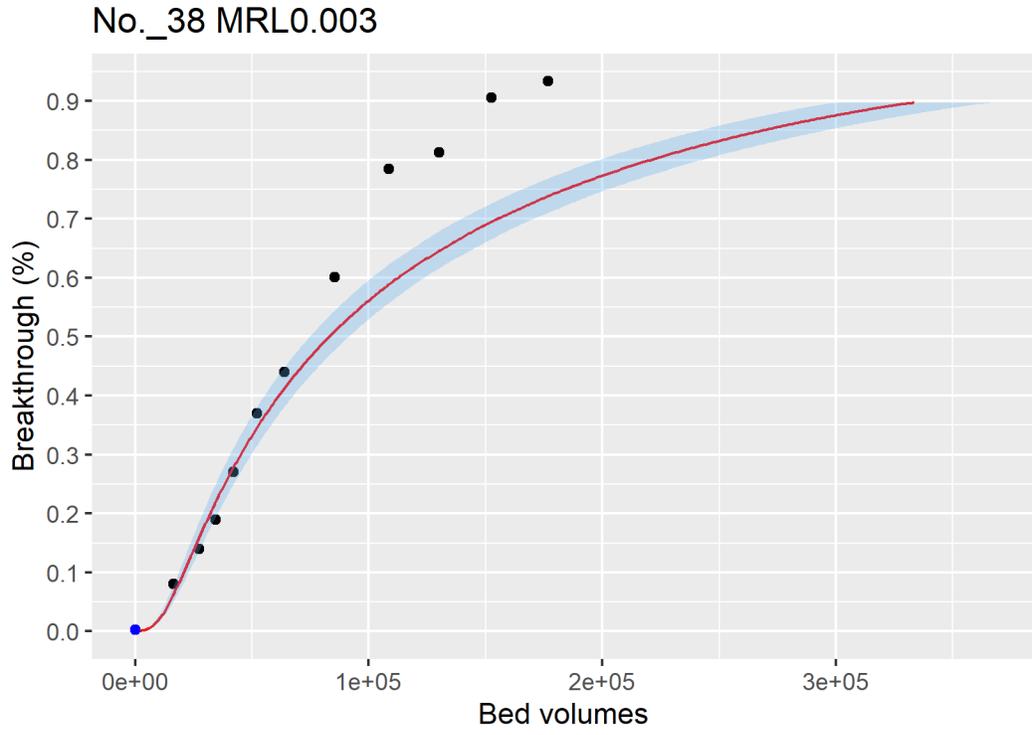


Figure ID 61- 36

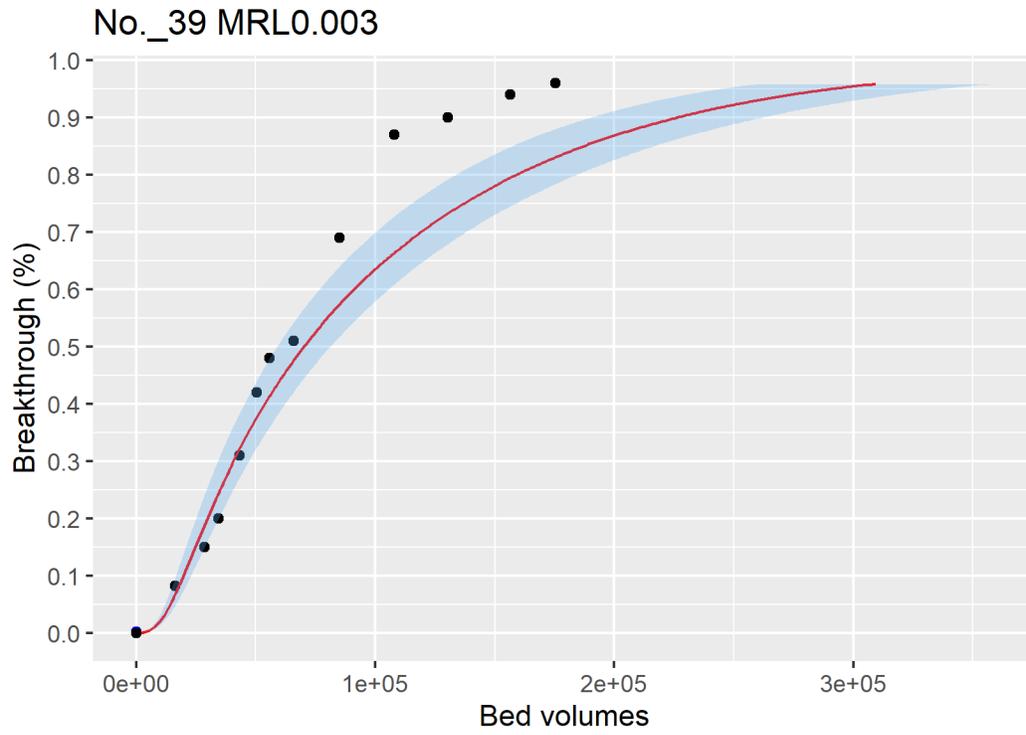


Figure ID 61- 37

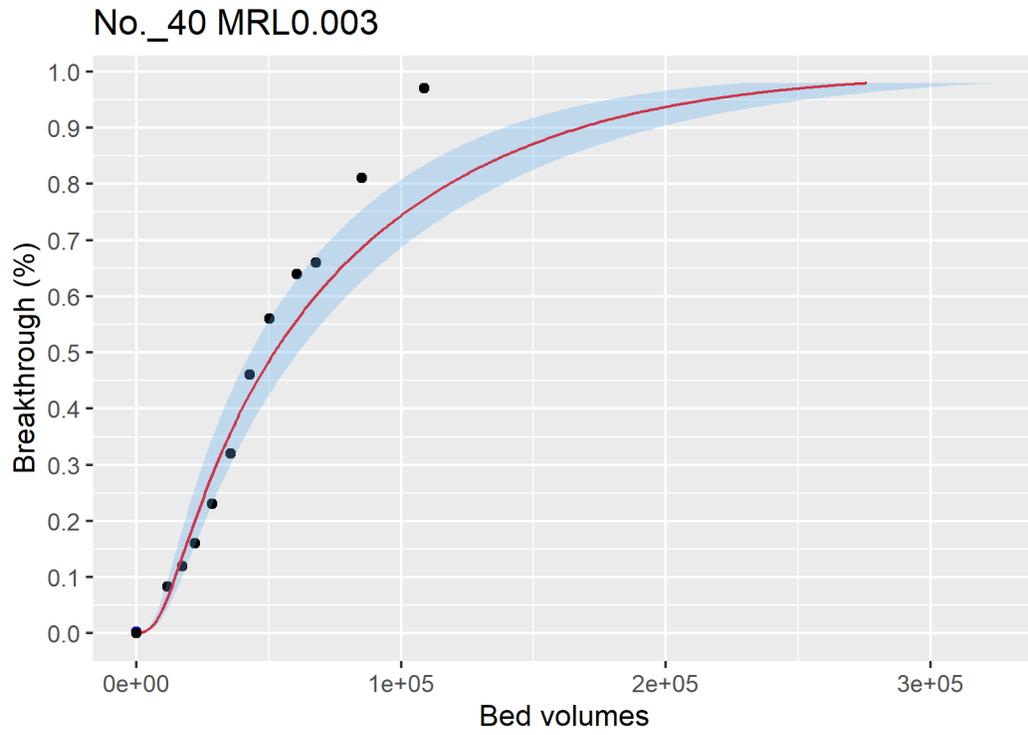


Figure ID 61- 38

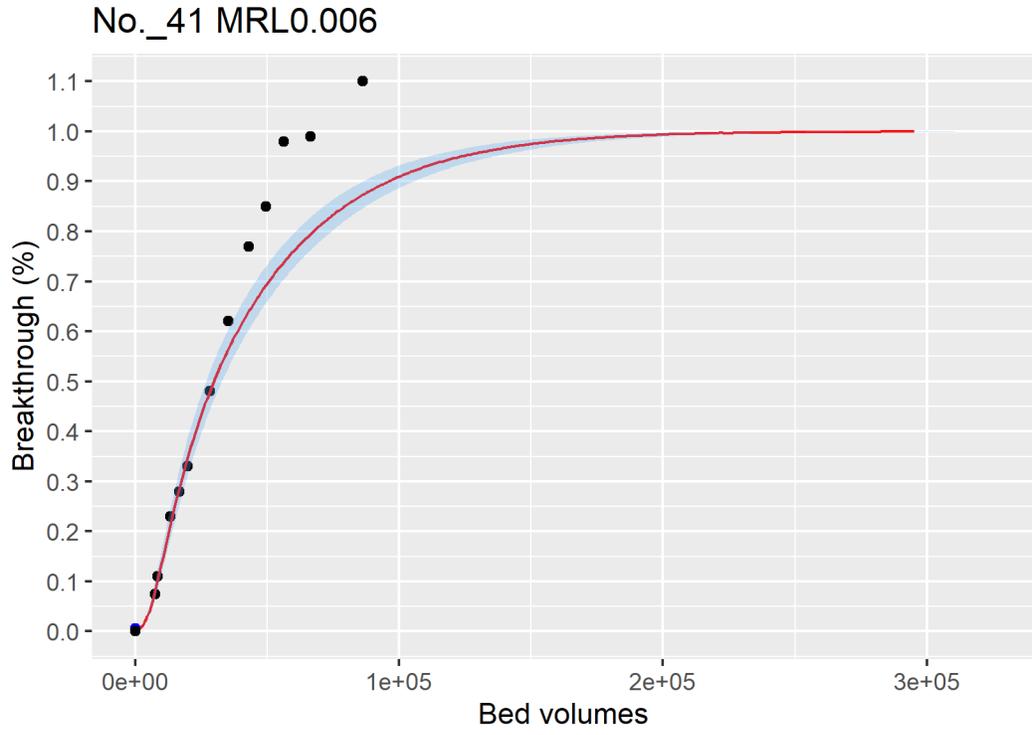


Figure ID 61- 39

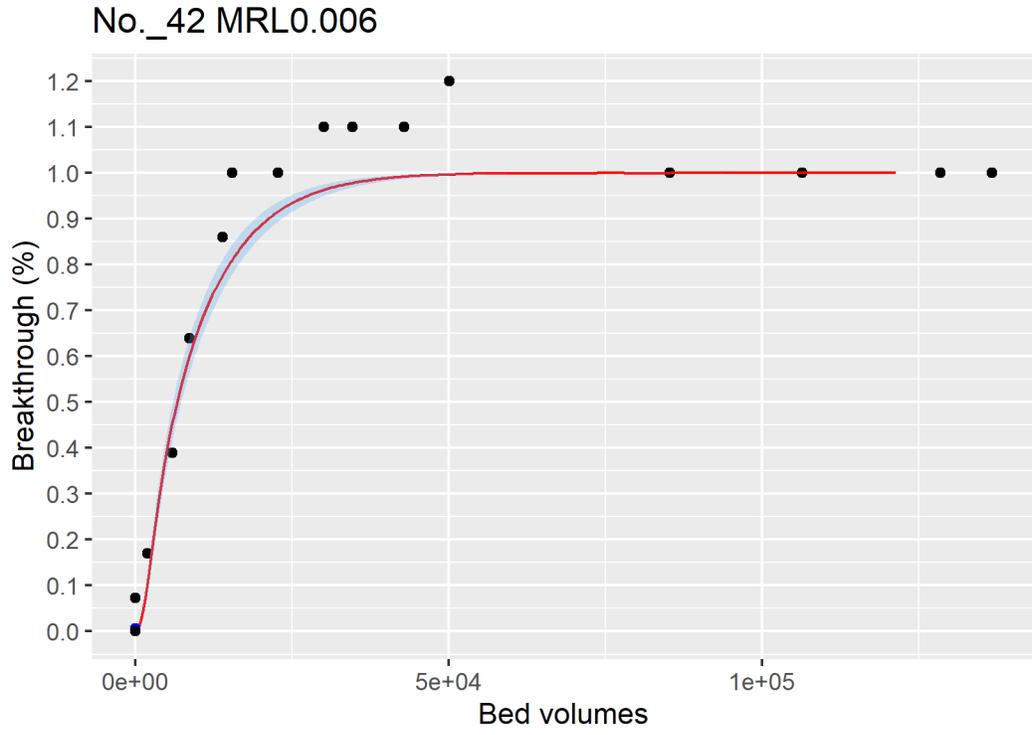


Figure ID 61- 40

Breakthrough data of Ref ID 62

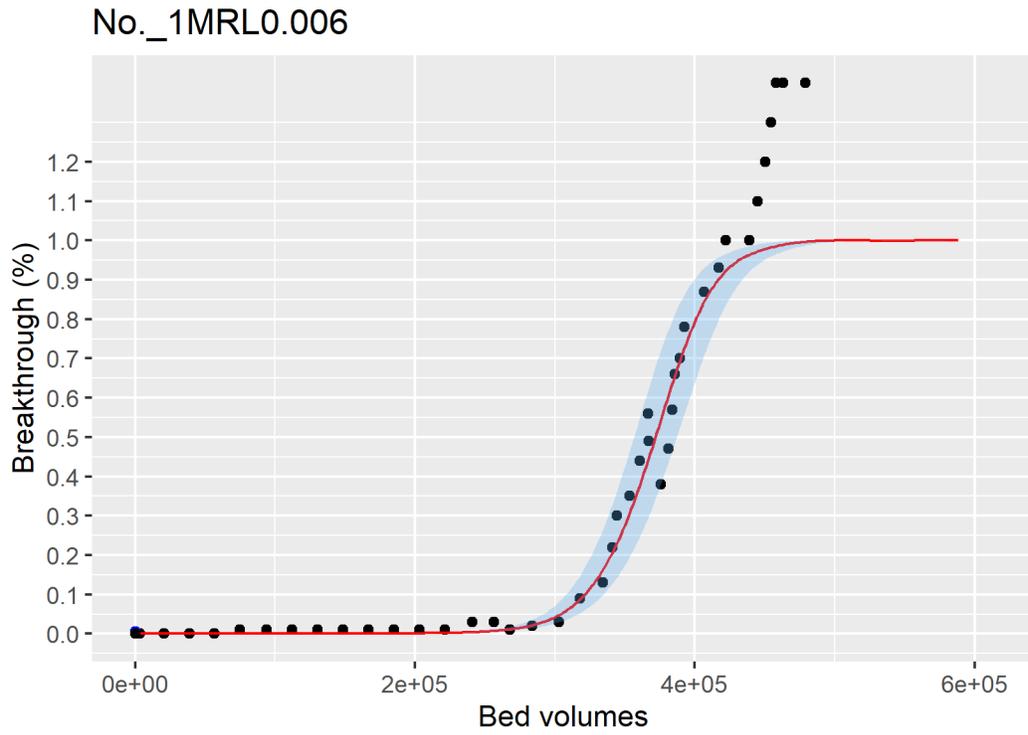


Figure ID 62- 1

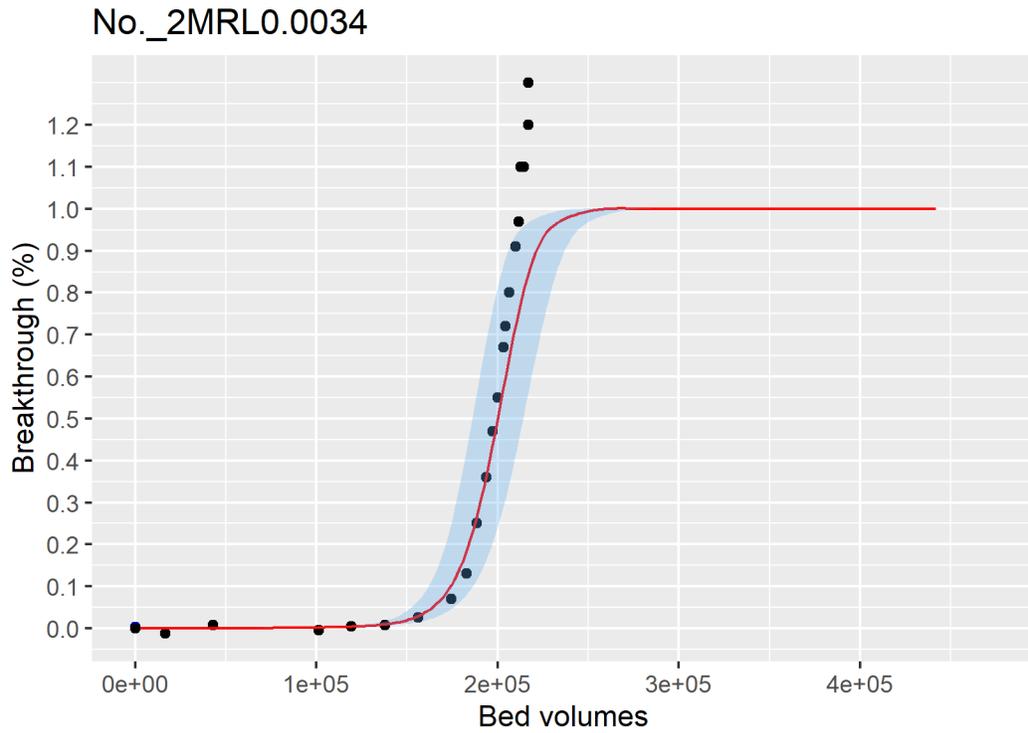


Figure ID 62- 2

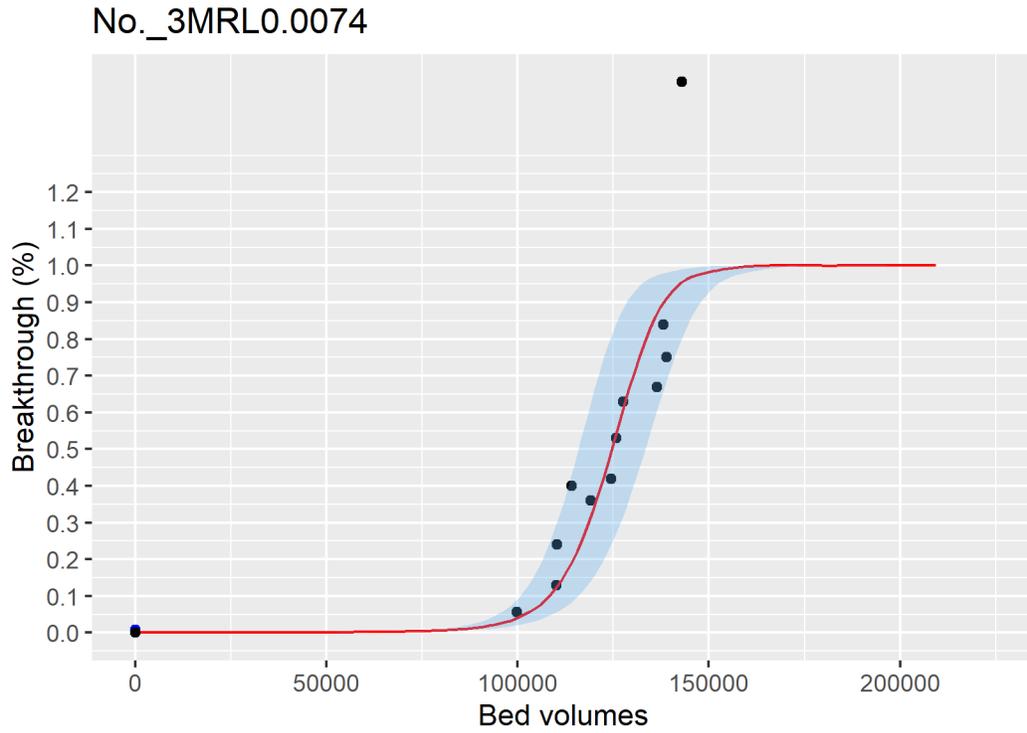


Figure ID 62- 3

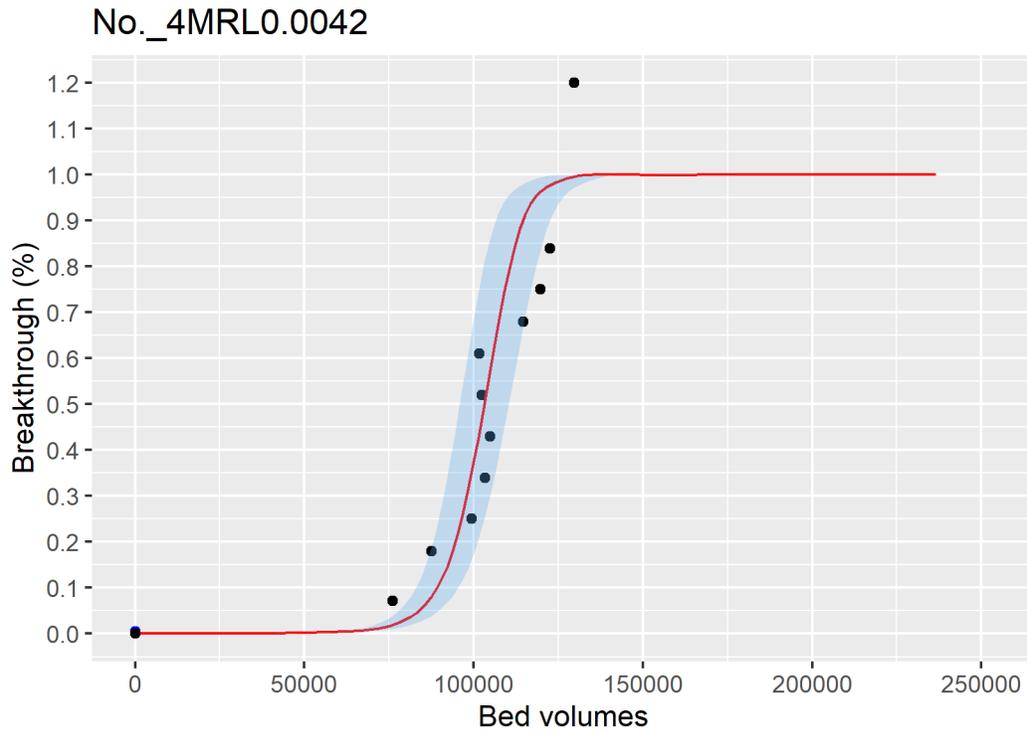


Figure ID 62- 4

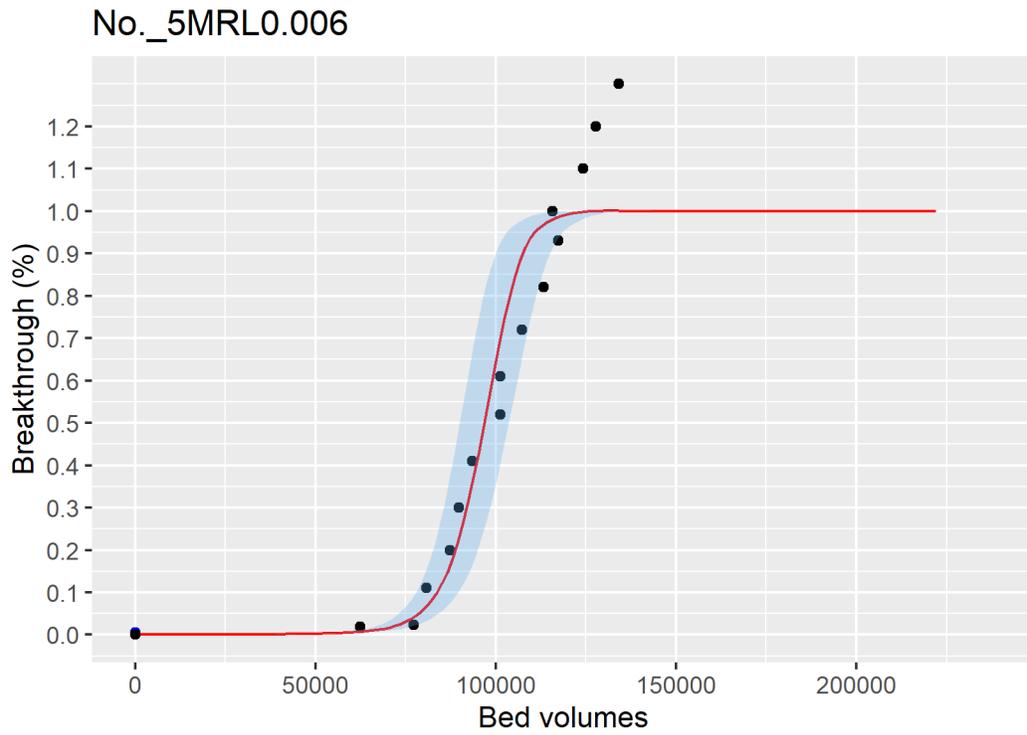


Figure ID 62- 5

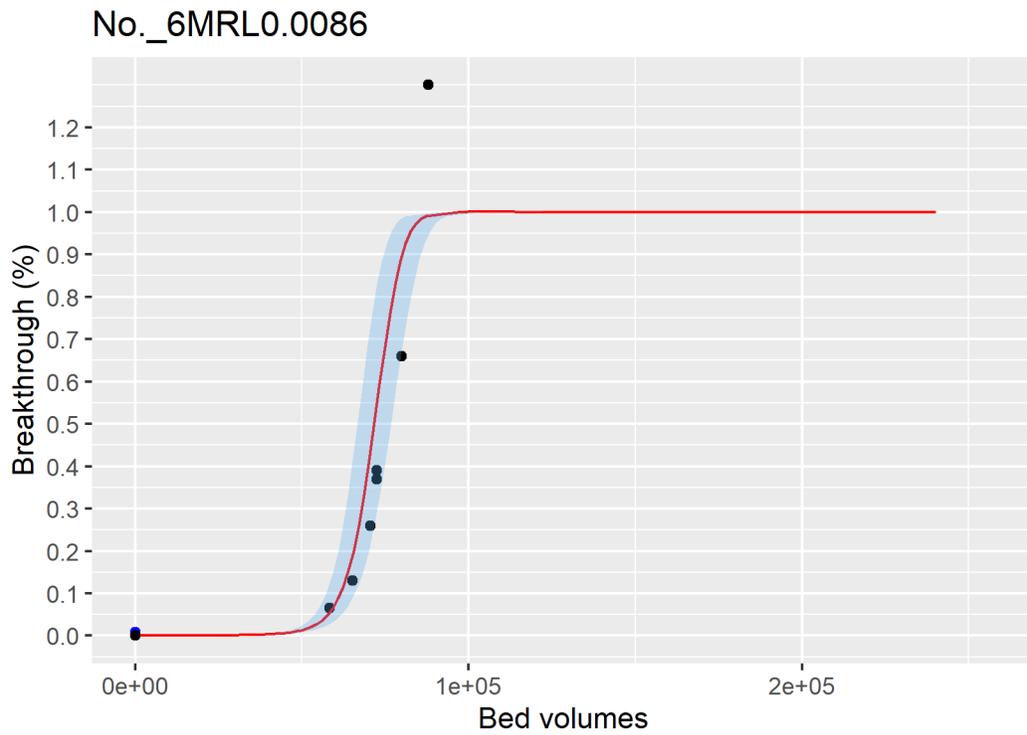


Figure ID 62- 6

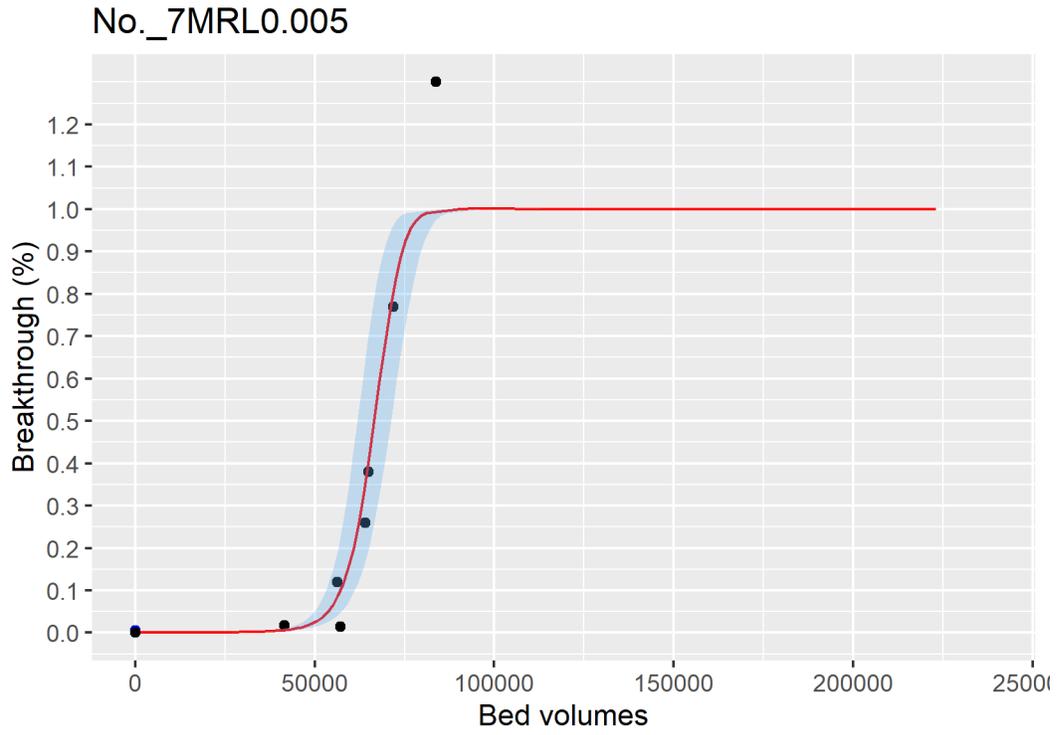


Figure ID 62- 7

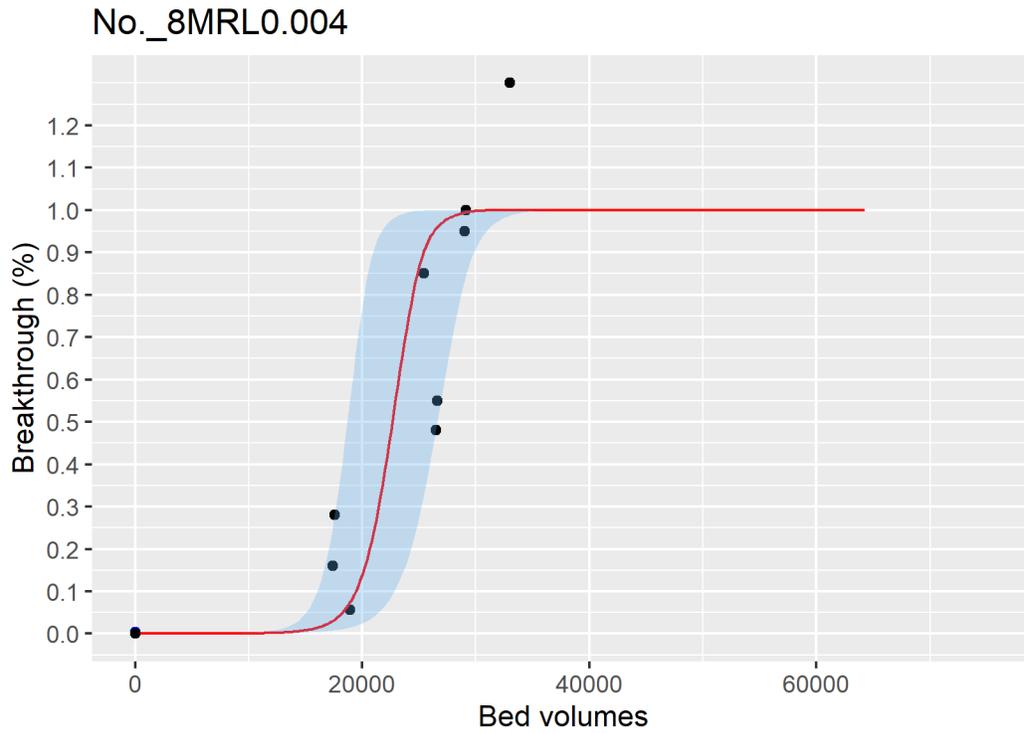


Figure ID 62- 8

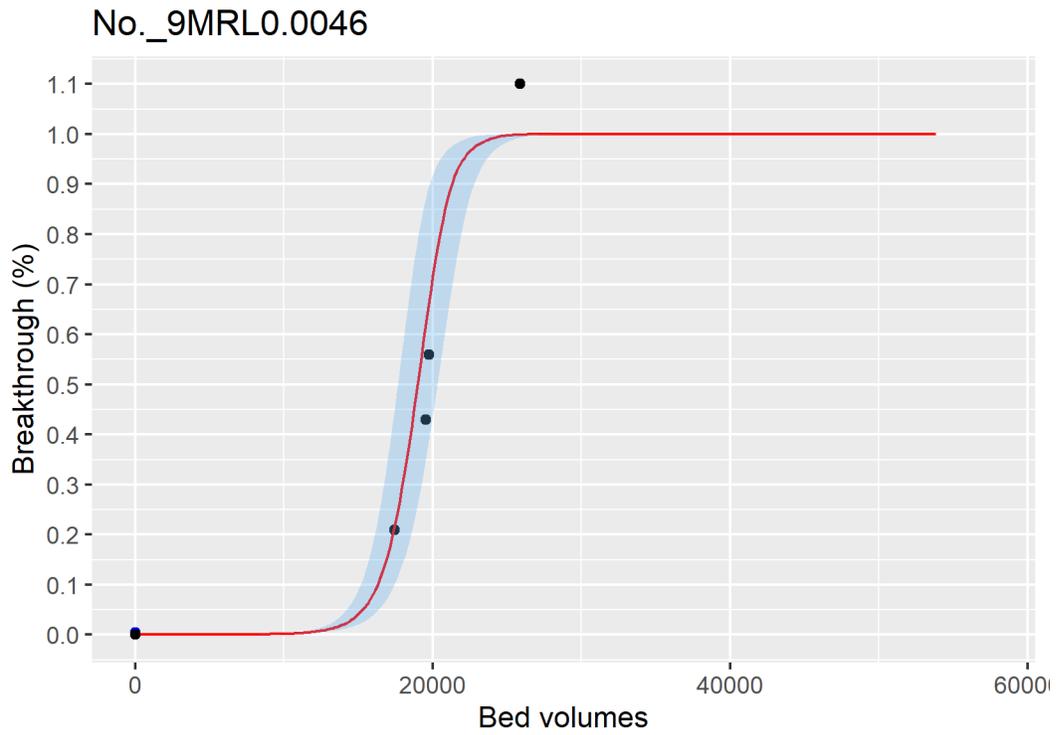


Figure ID 62- 9

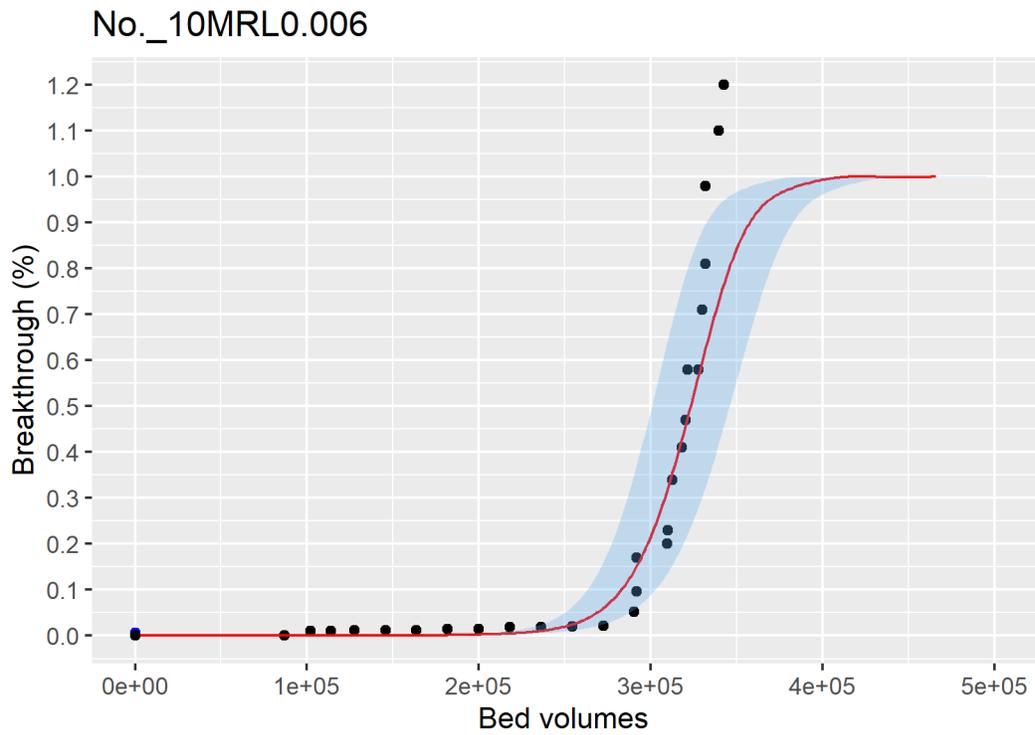


Figure ID 62- 10

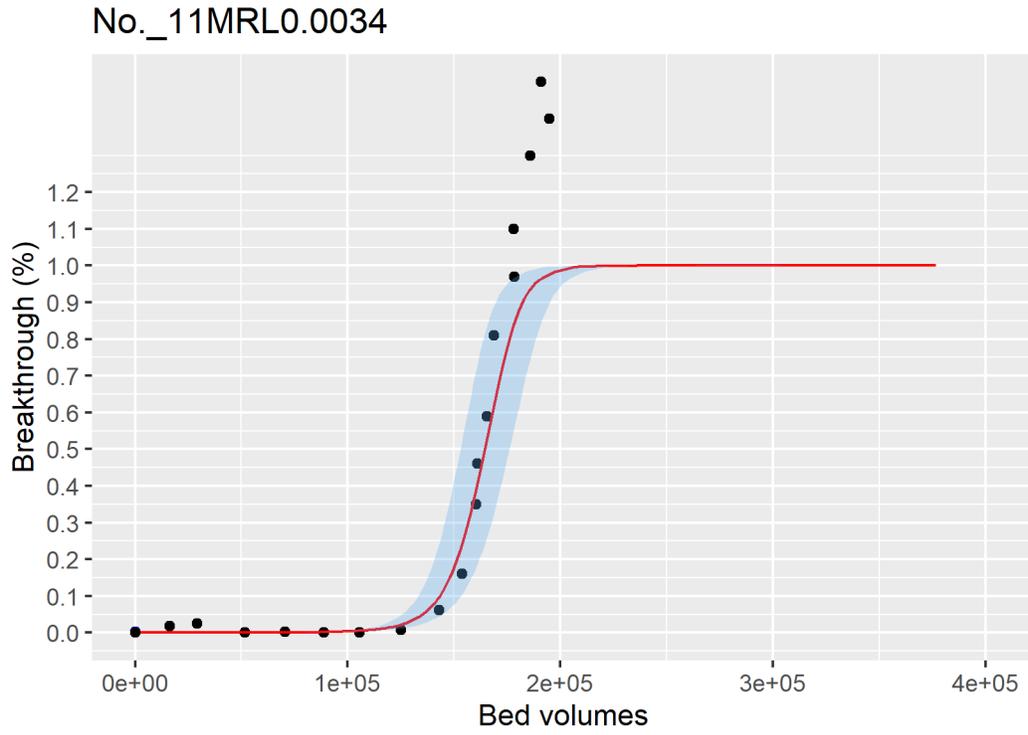


Figure ID 62- 11

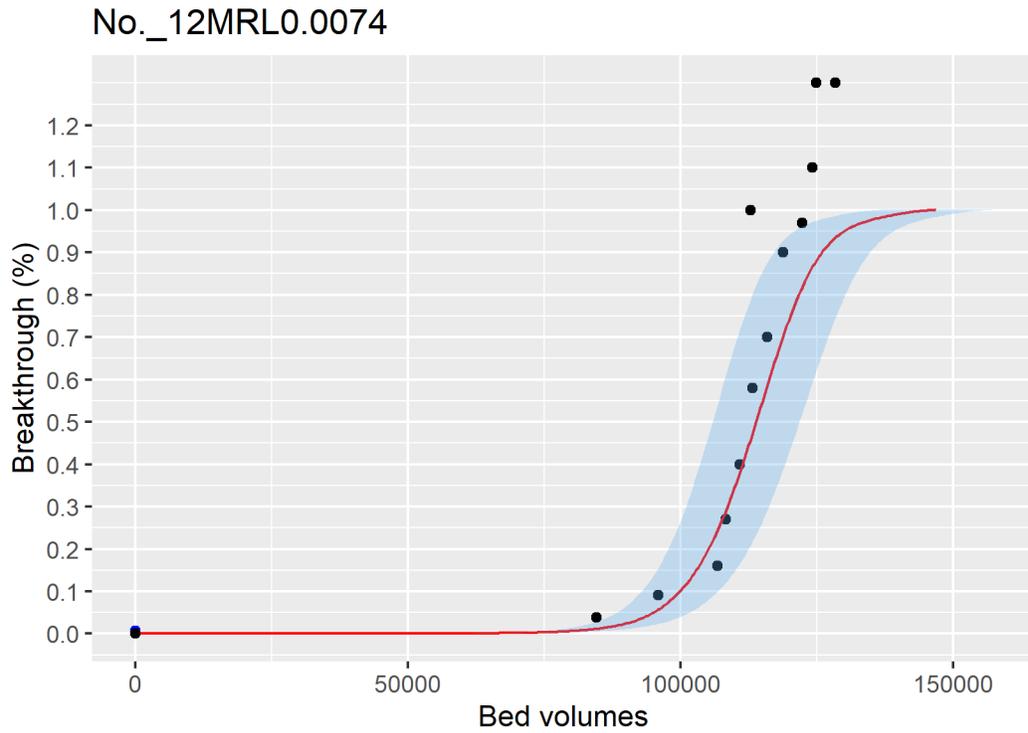


Figure ID 62- 12

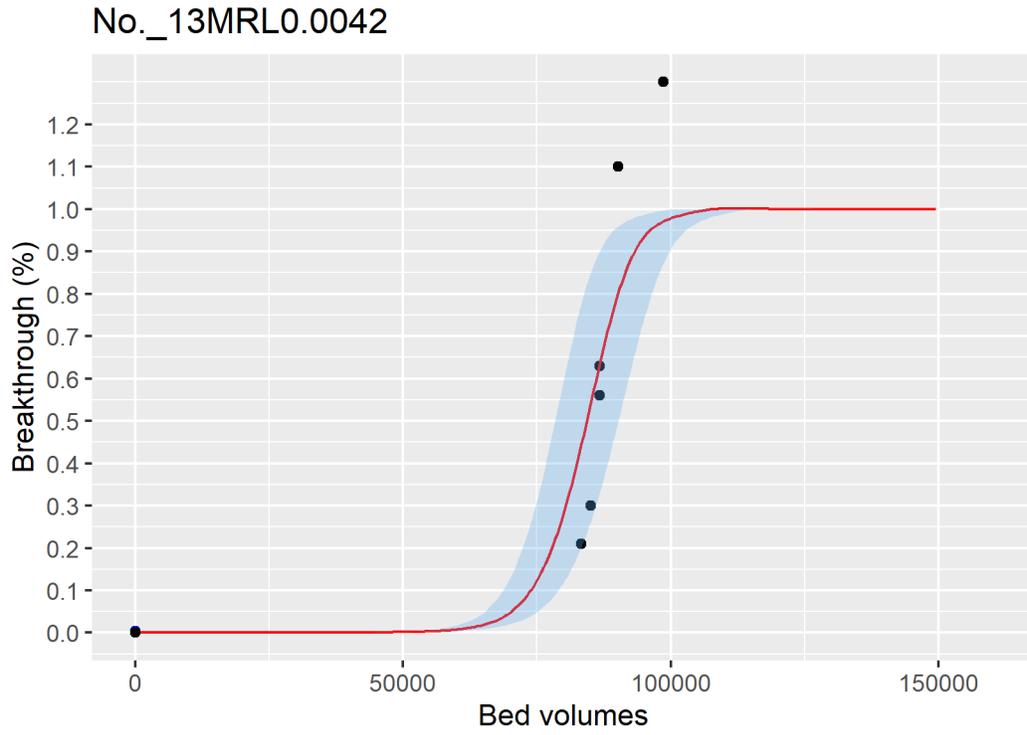


Figure ID 62- 13

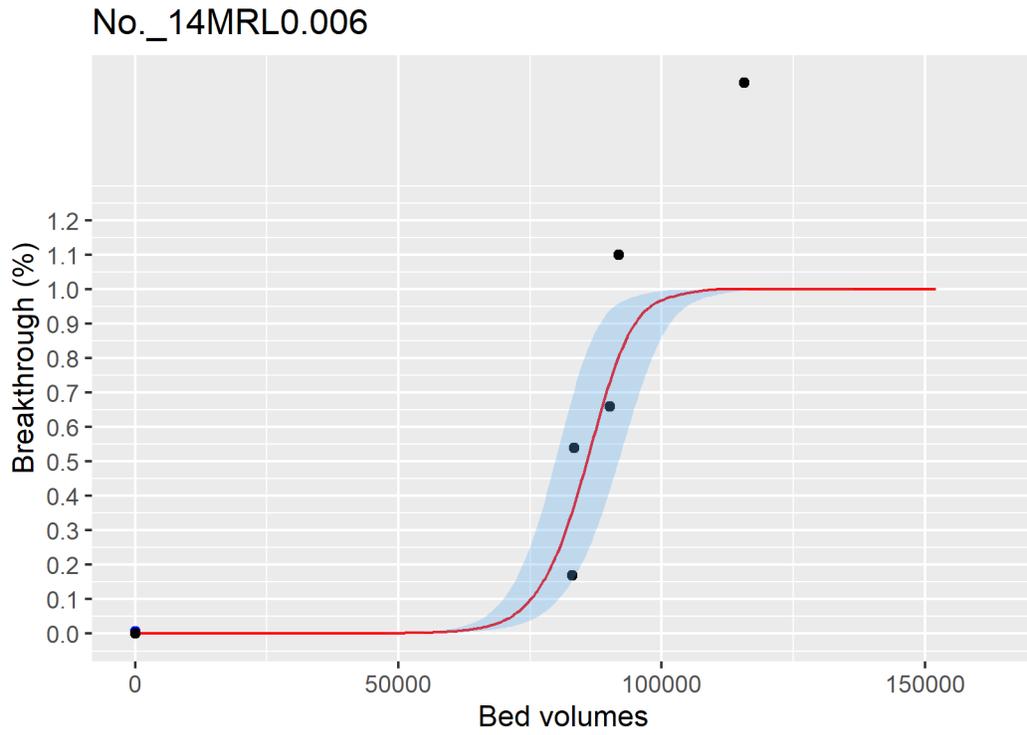


Figure ID 62- 14

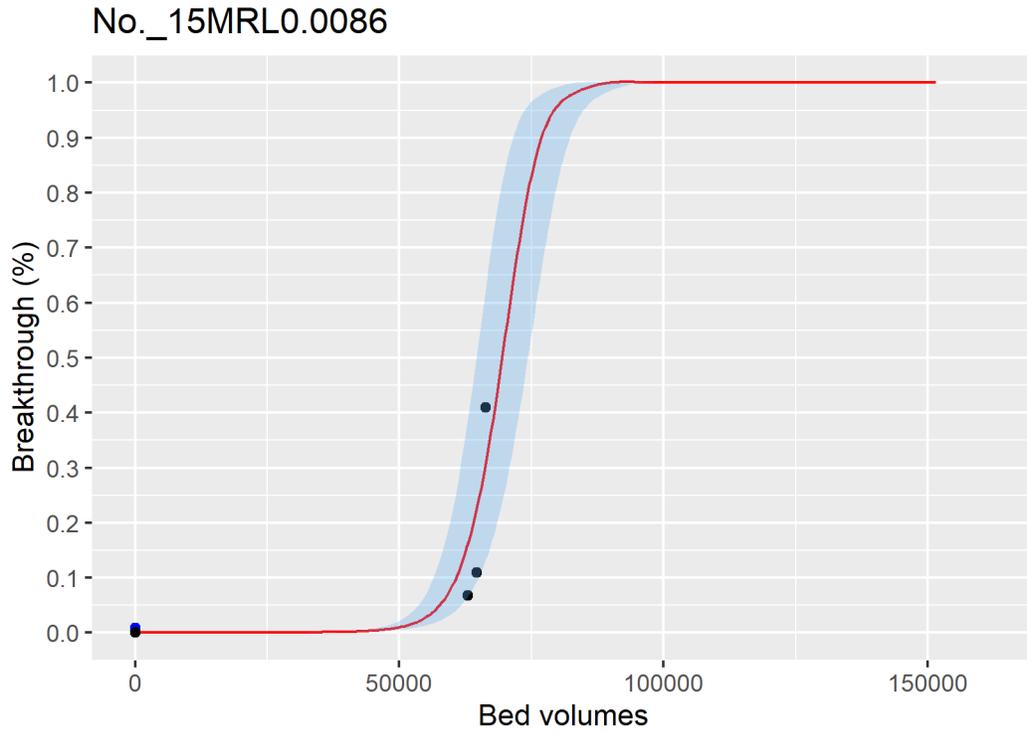


Figure ID 62- 15

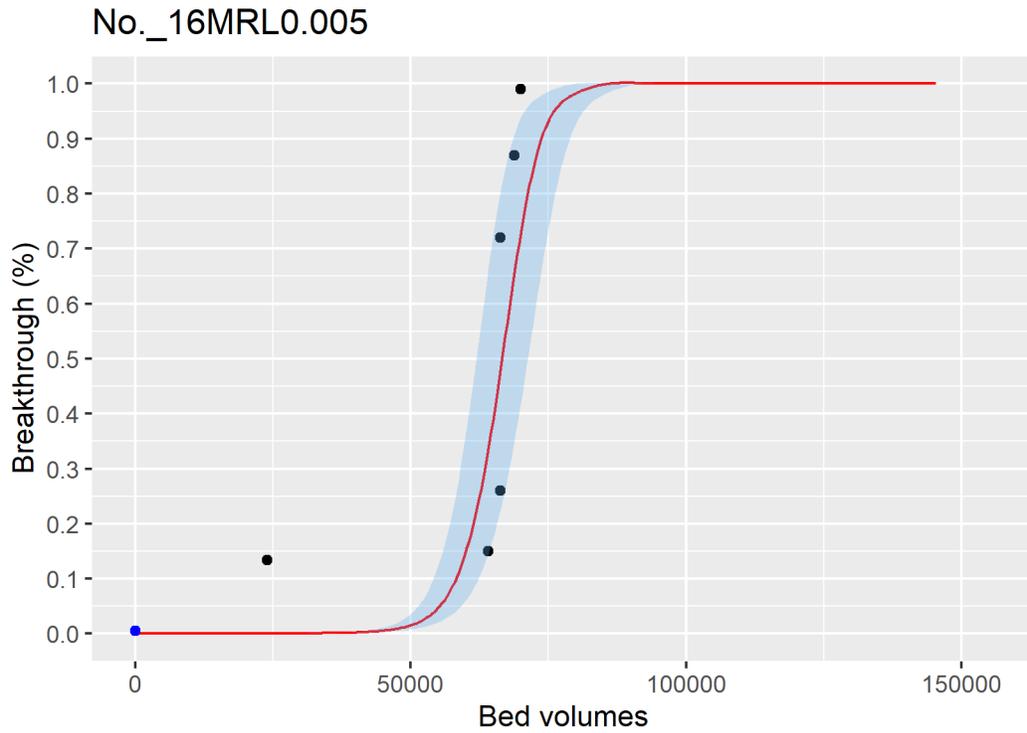


Figure ID 62- 16

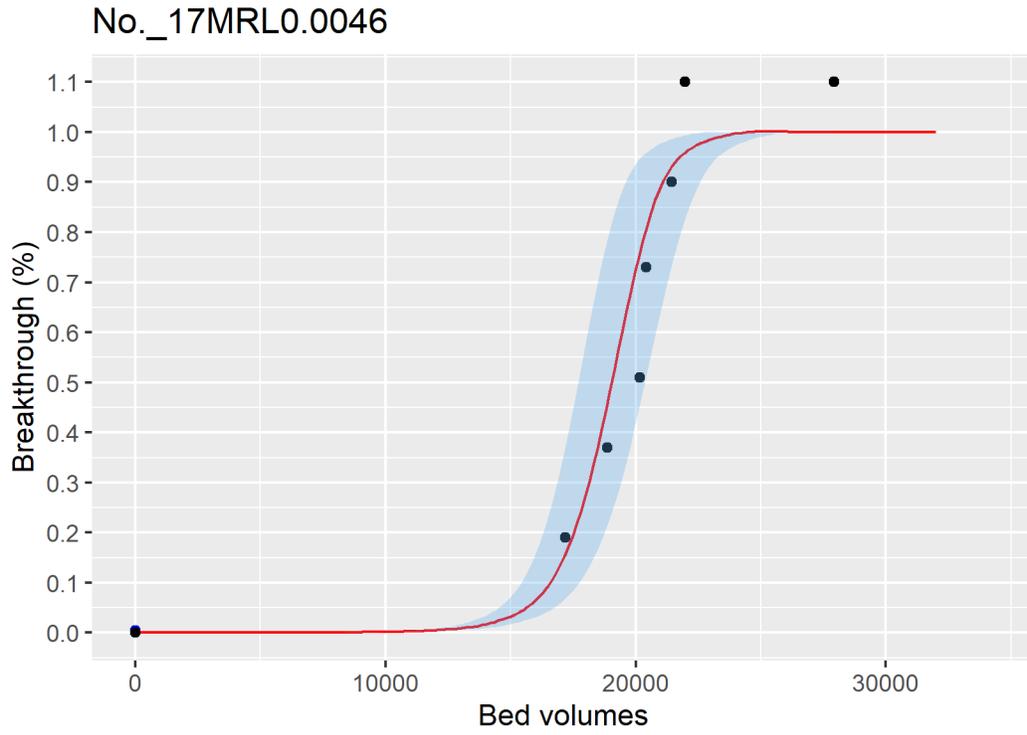


Figure ID 62- 17

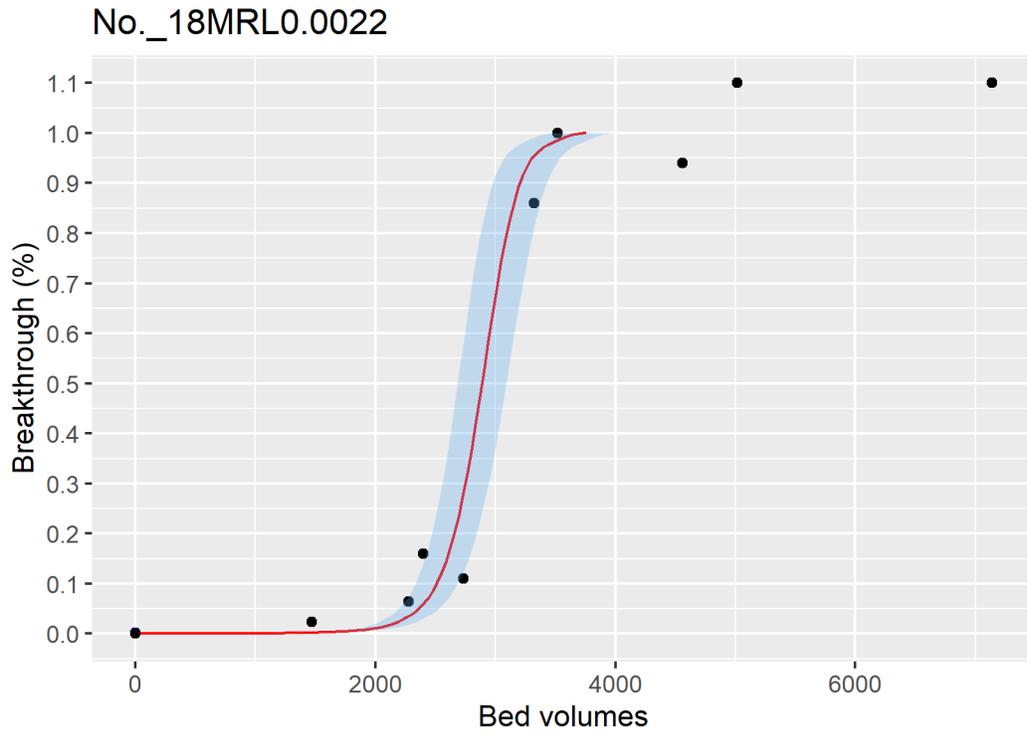


Figure ID 62- 18

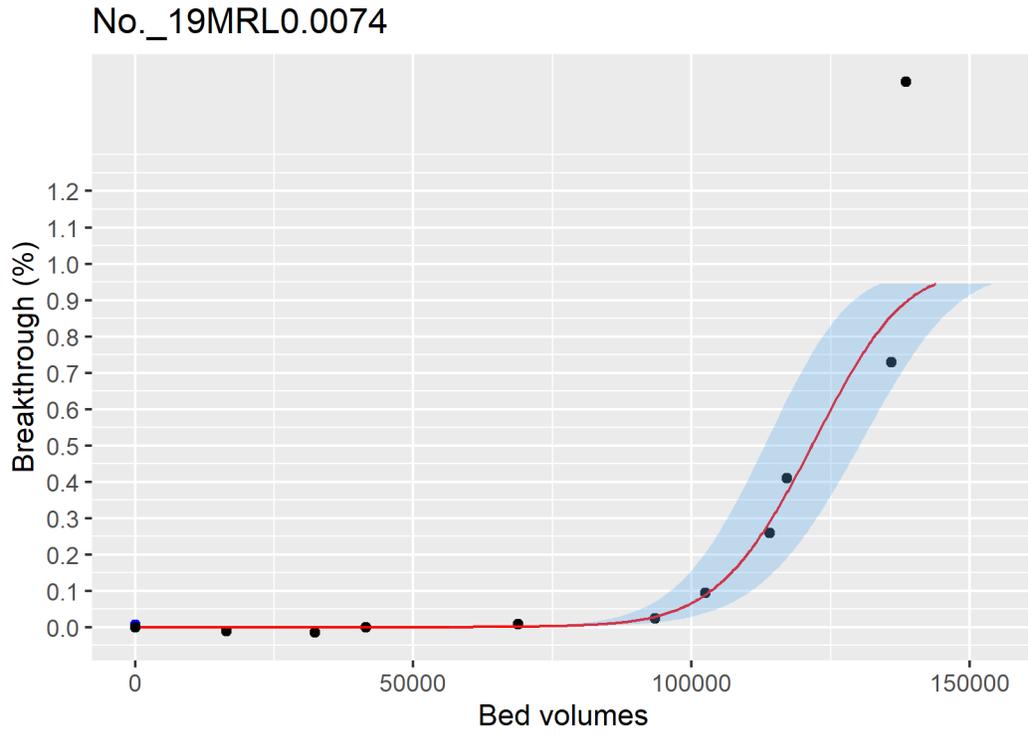


Figure ID 62- 19

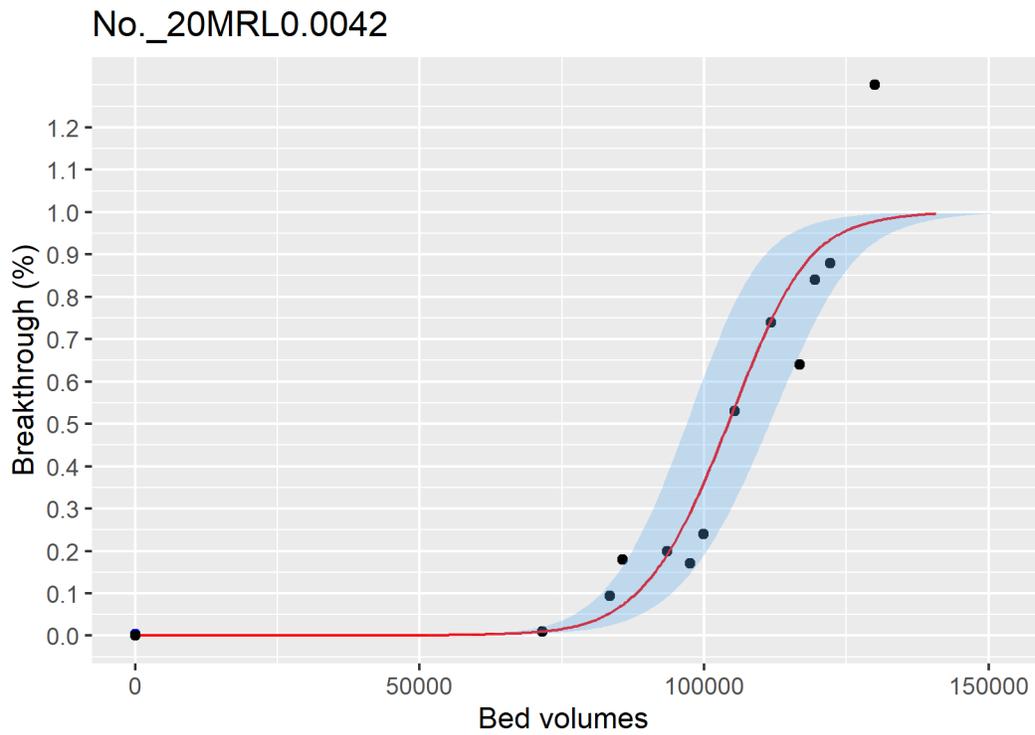


Figure ID 62- 20

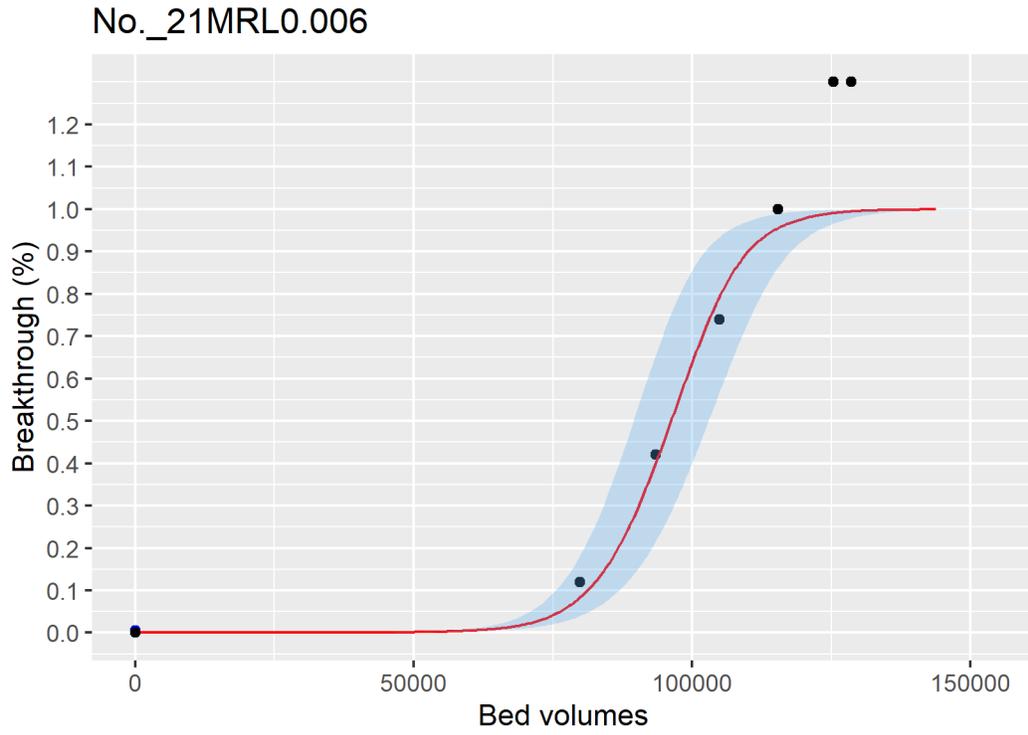


Figure ID 62- 21

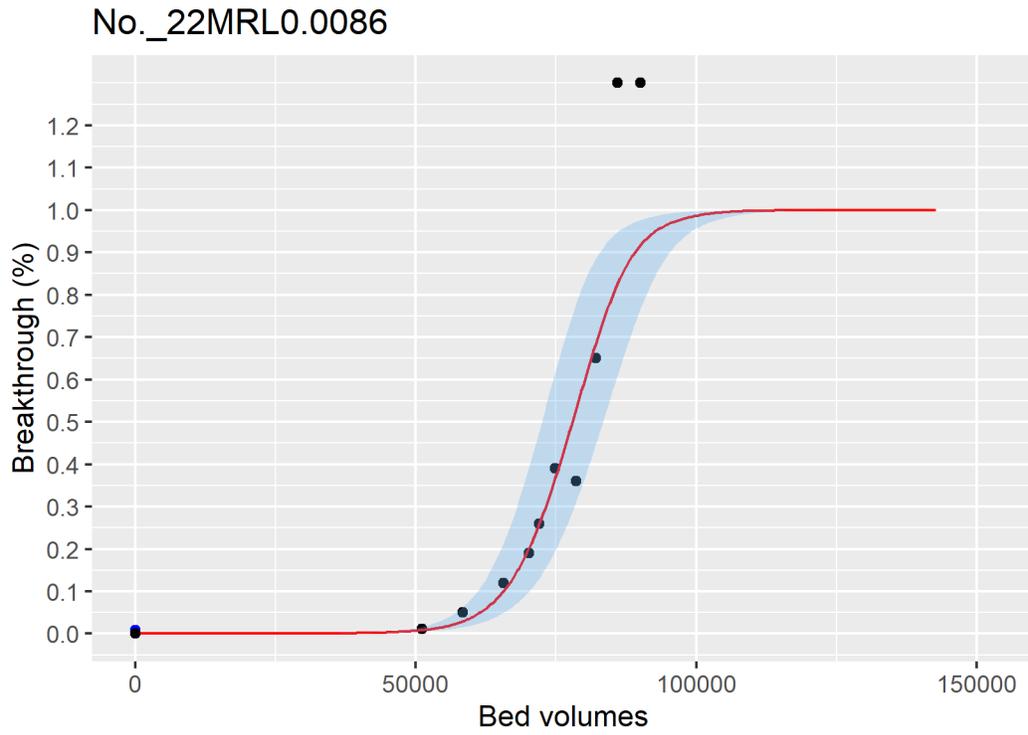


Figure ID 62- 22

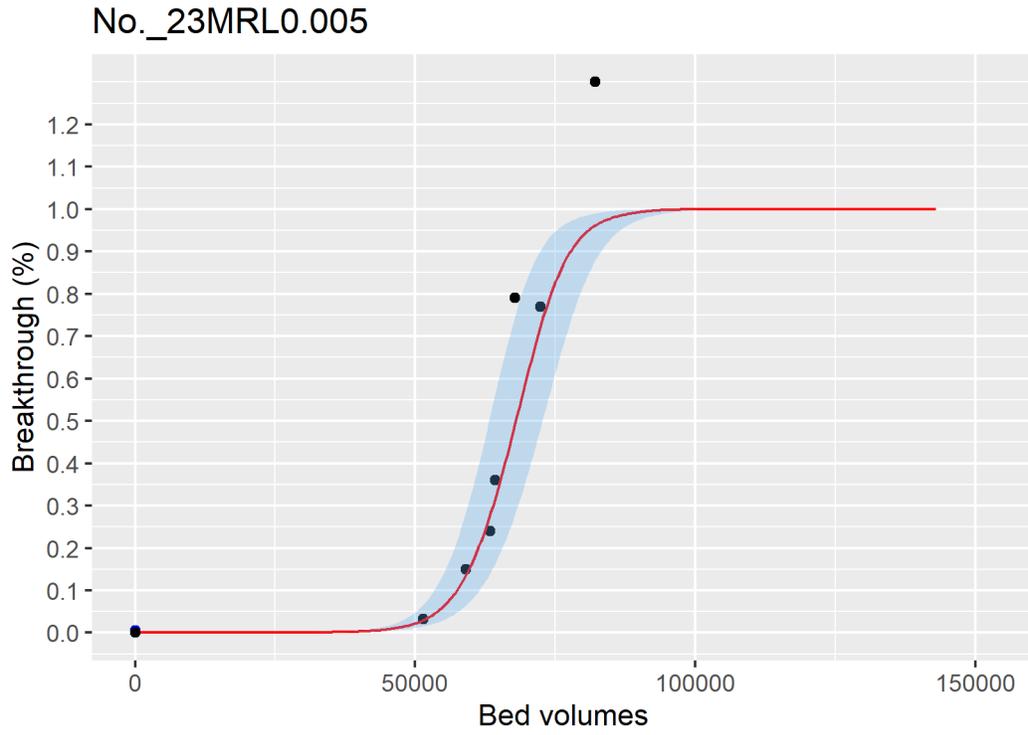


Figure ID 62- 23

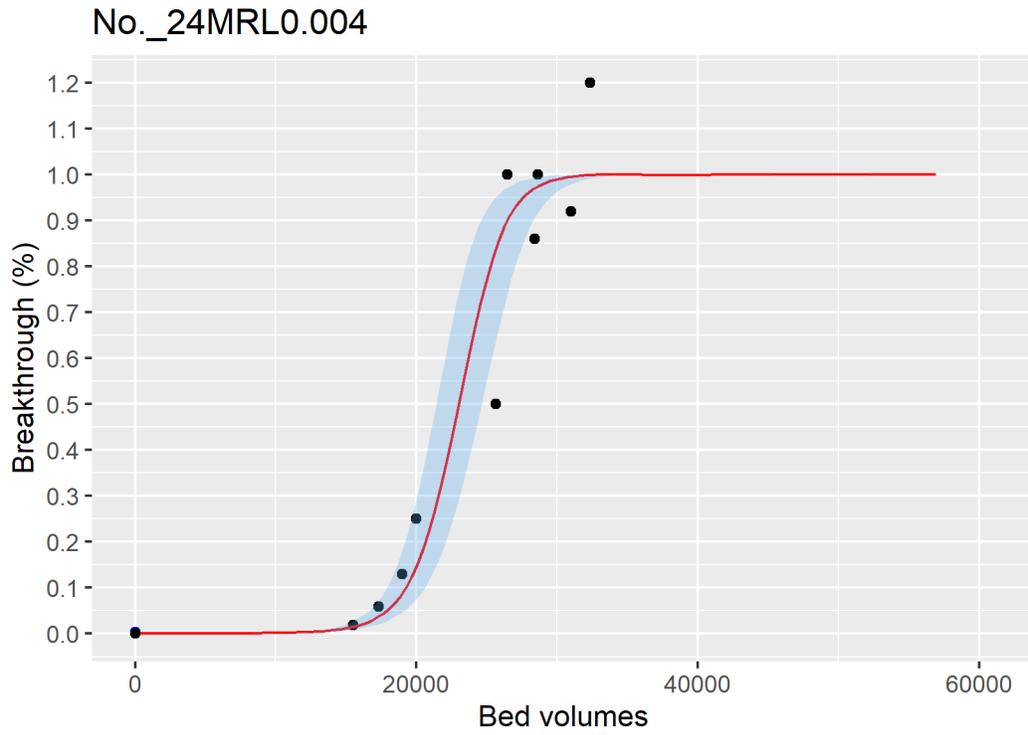


Figure ID 62- 24

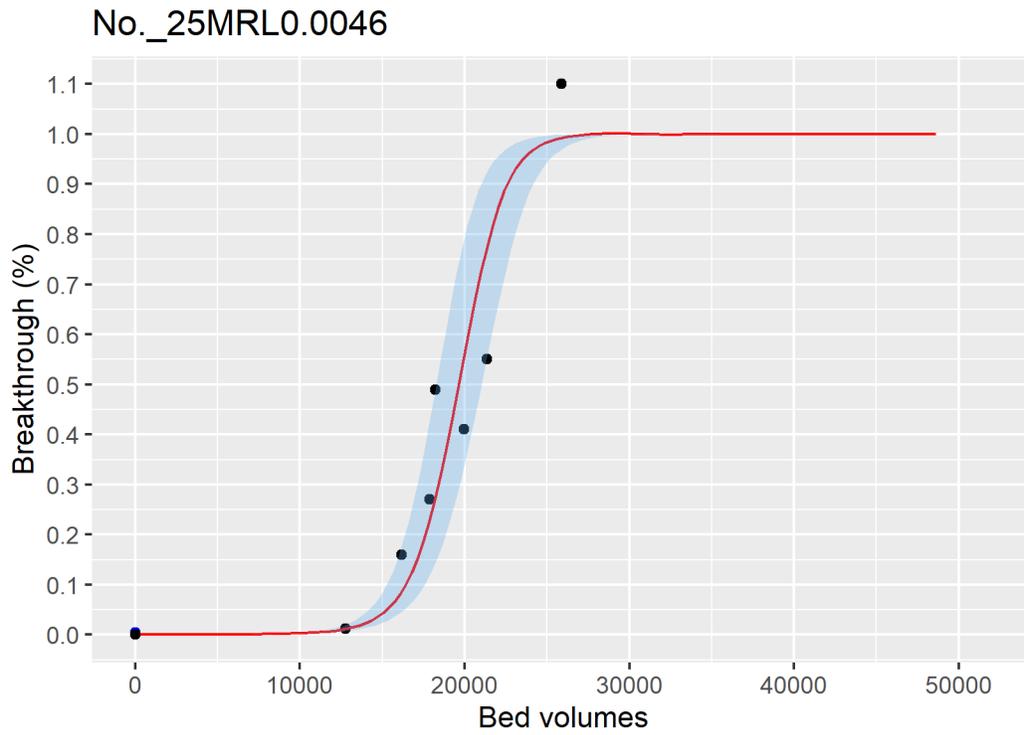


Figure ID 62- 25

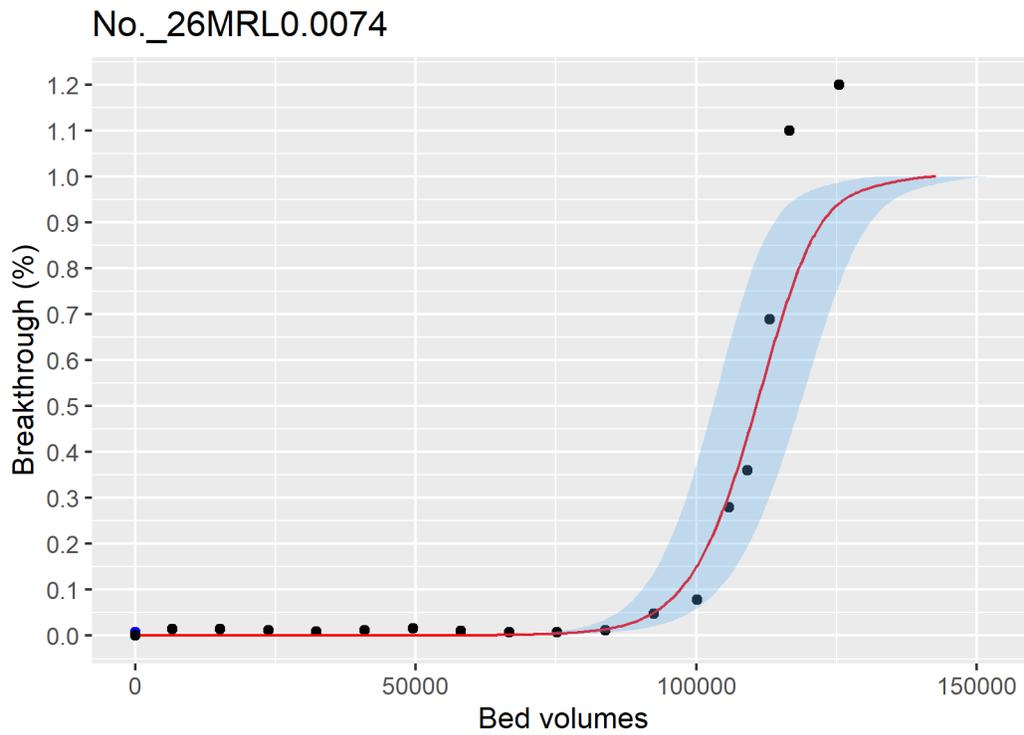


Figure ID 62- 26

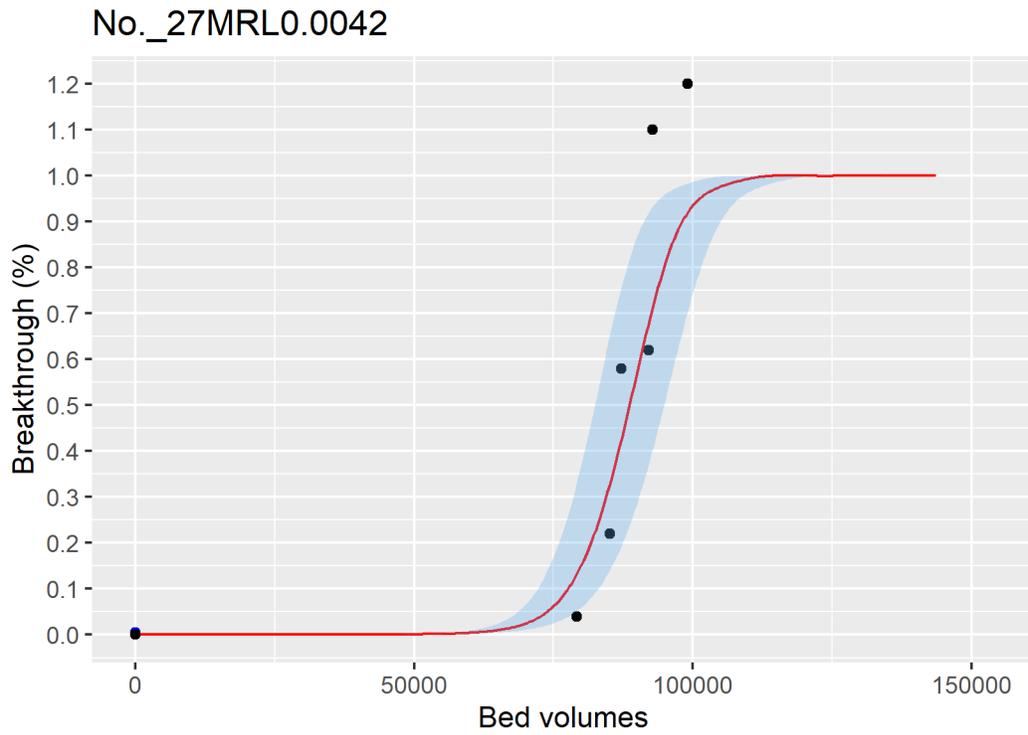


Figure ID 62- 27

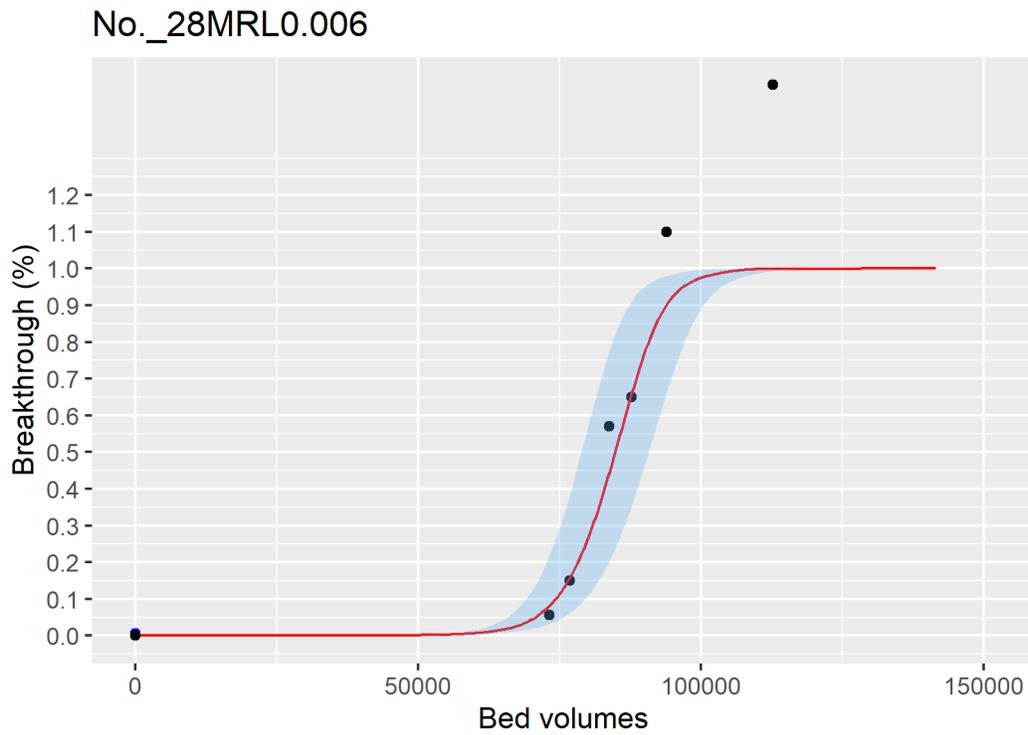


Figure ID 62- 28

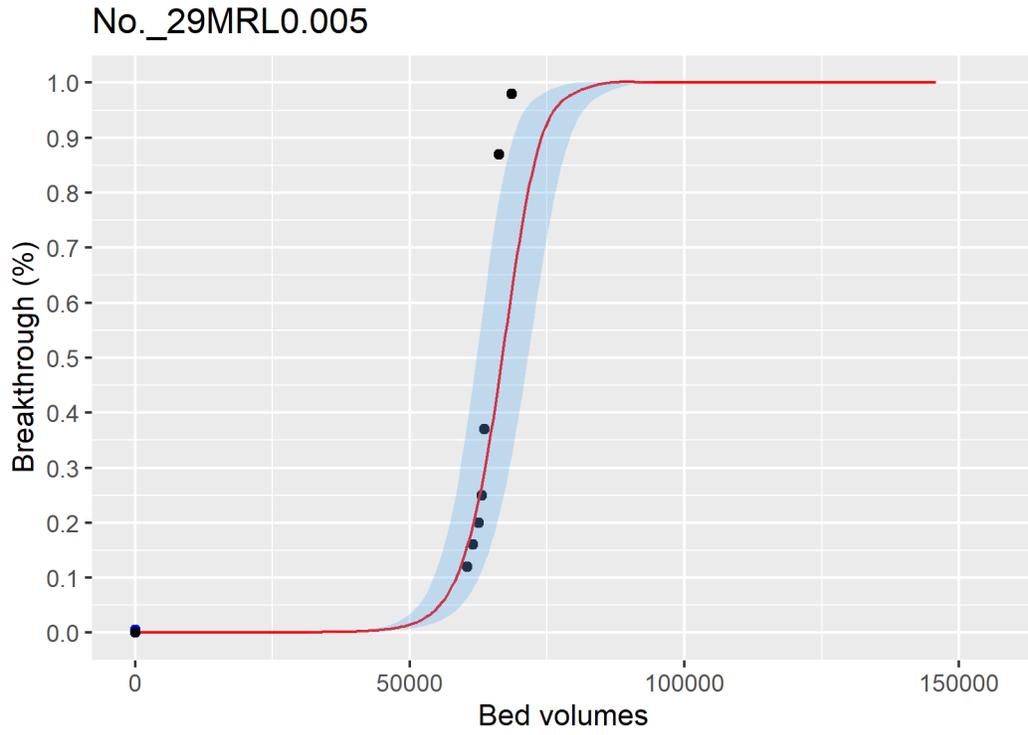


Figure ID 62- 29

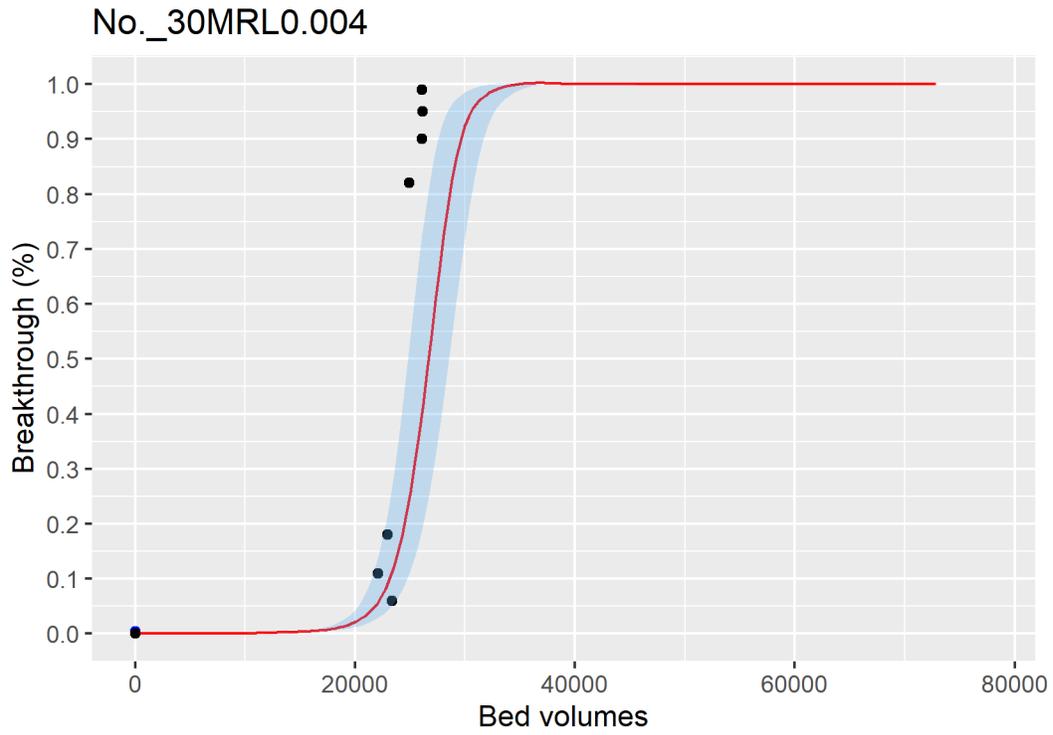


Figure ID 62- 30

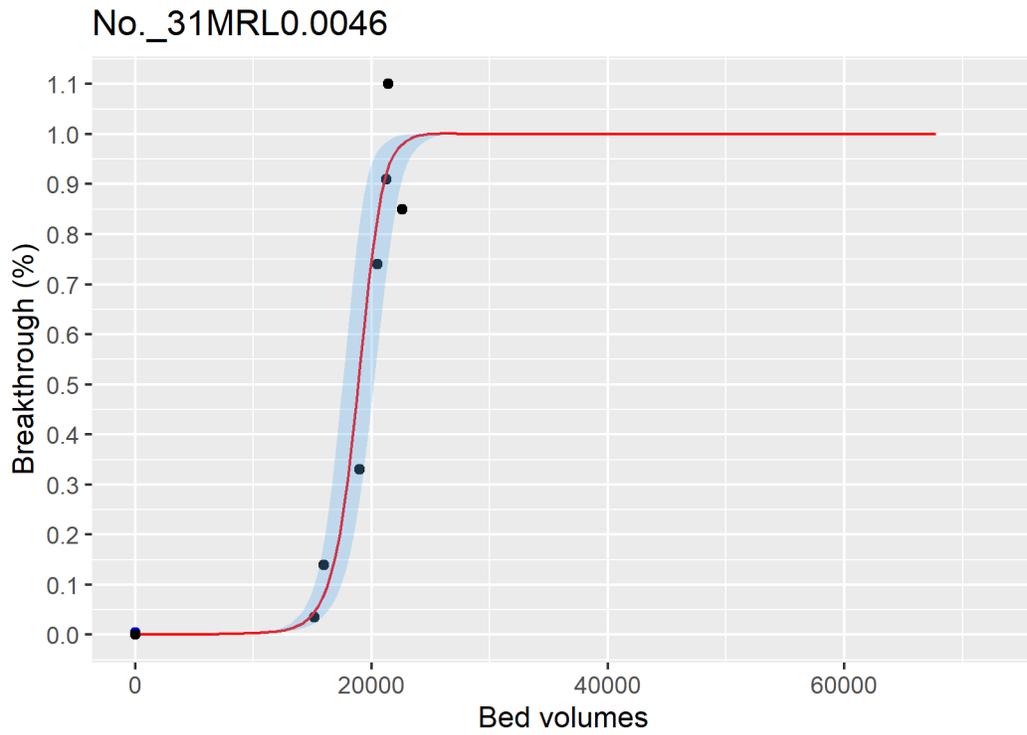


Figure ID 62- 31

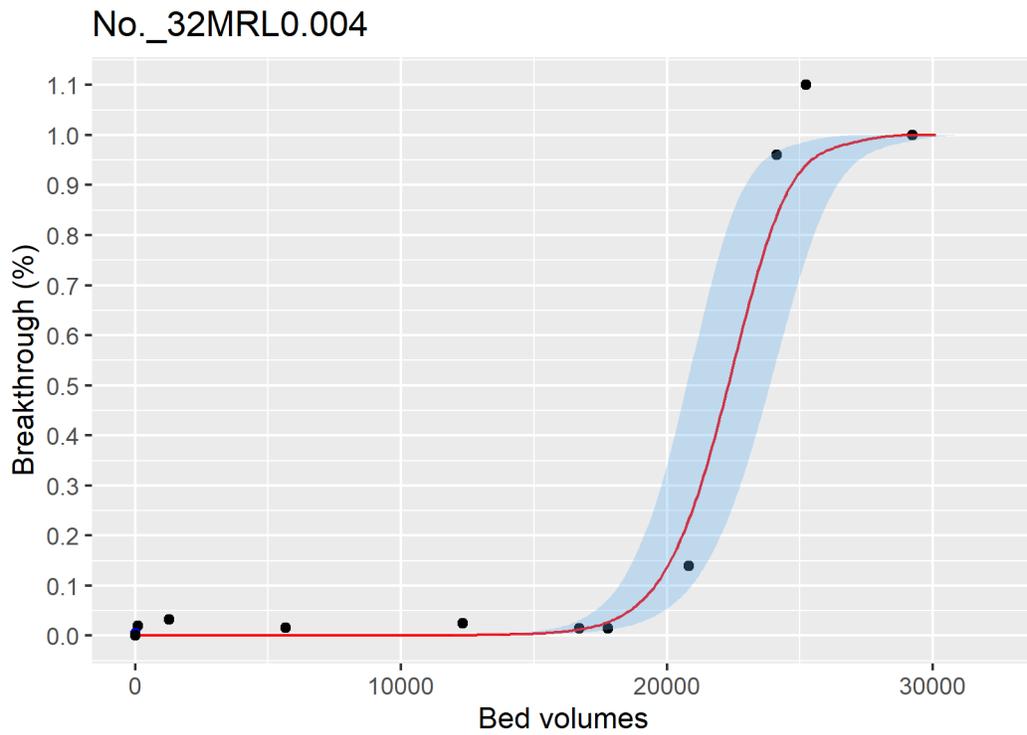


Figure ID 62- 32

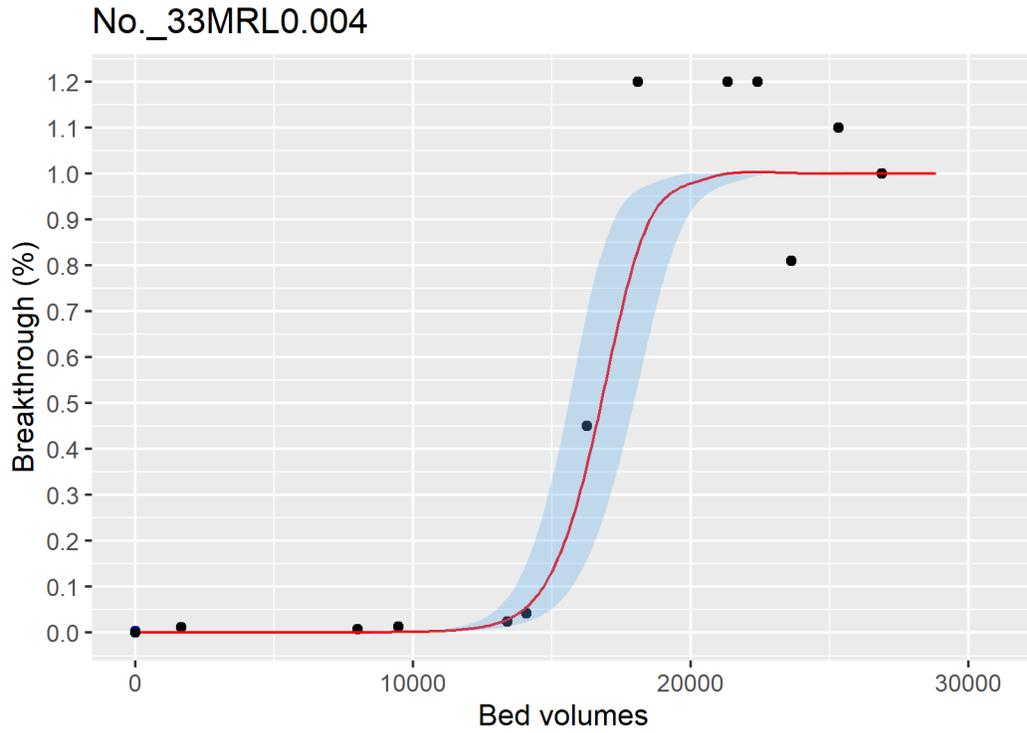


Figure ID 62- 33

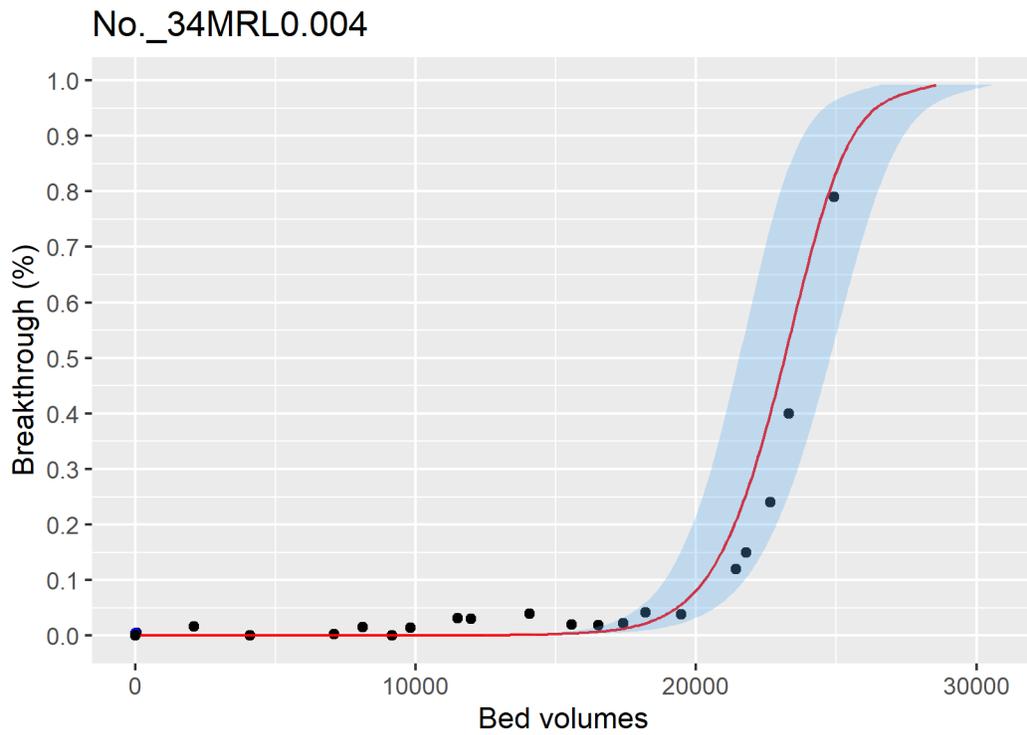


Figure ID 62- 34

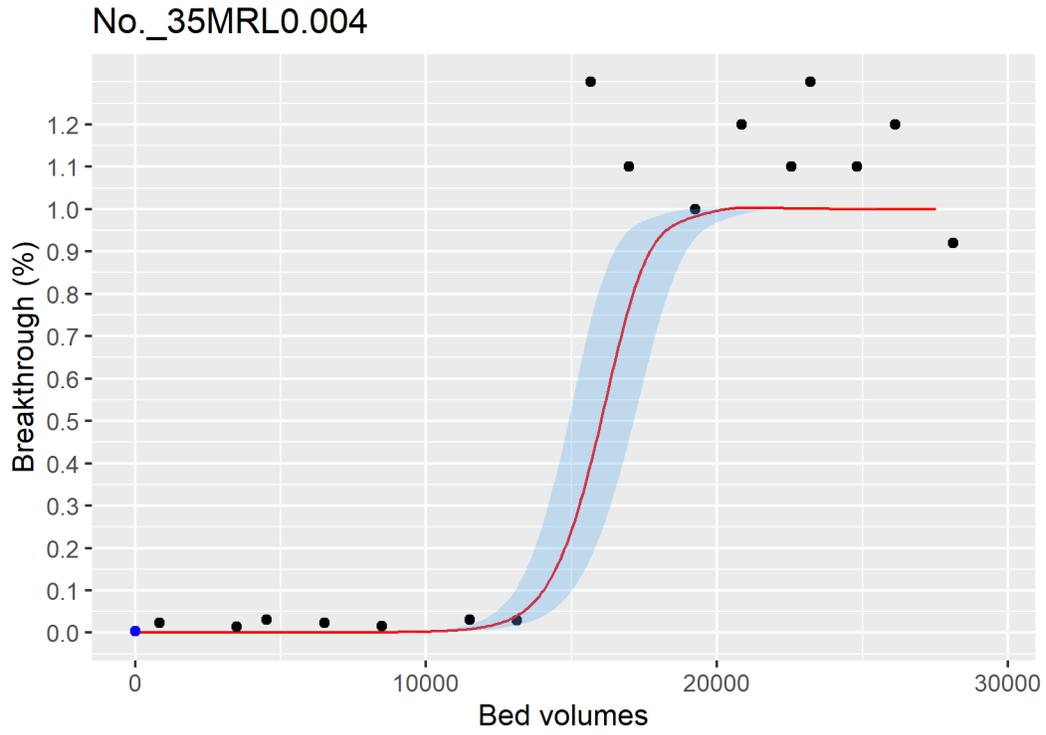


Figure ID 62- 35

Breakthrough data of Ref ID 63

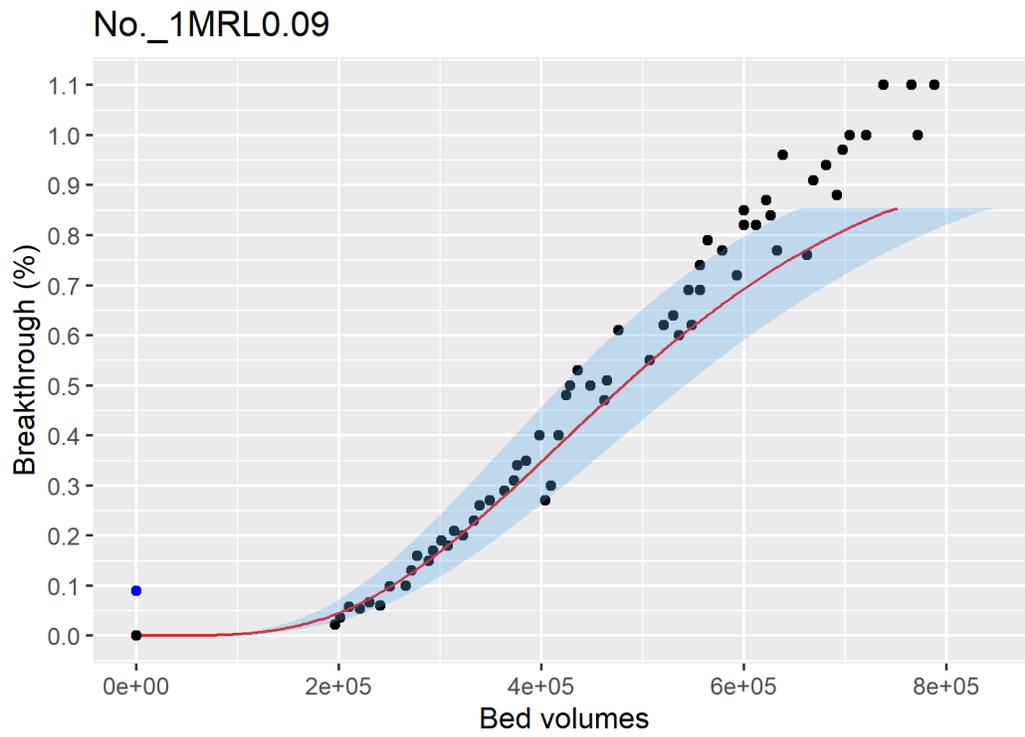


Figure ID 63- 1

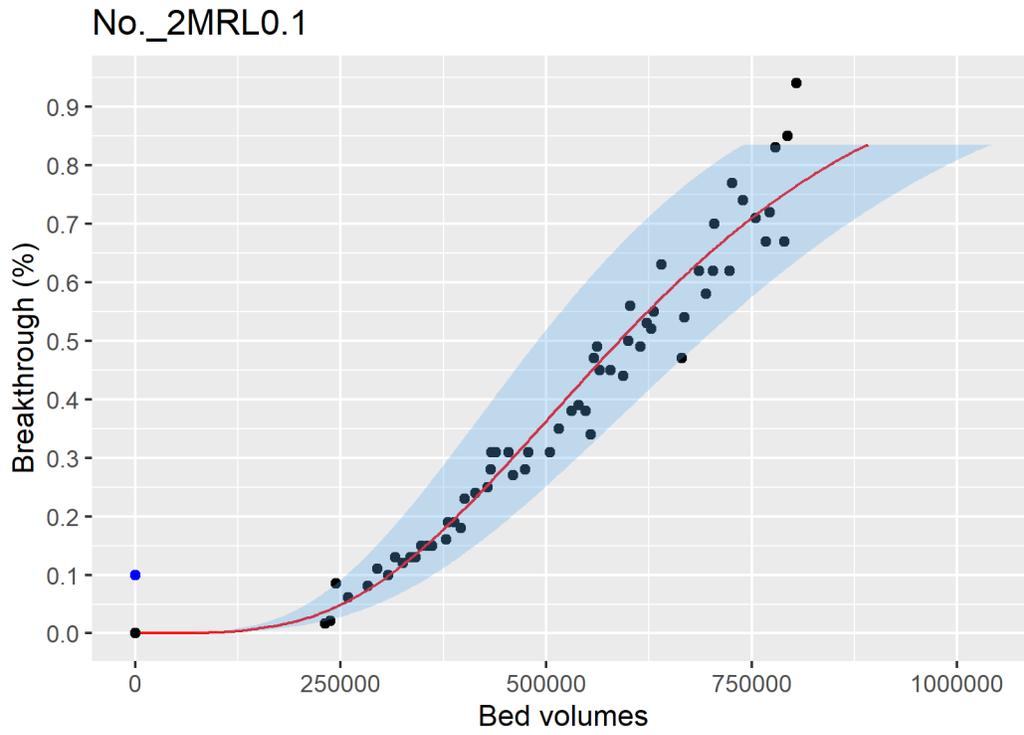
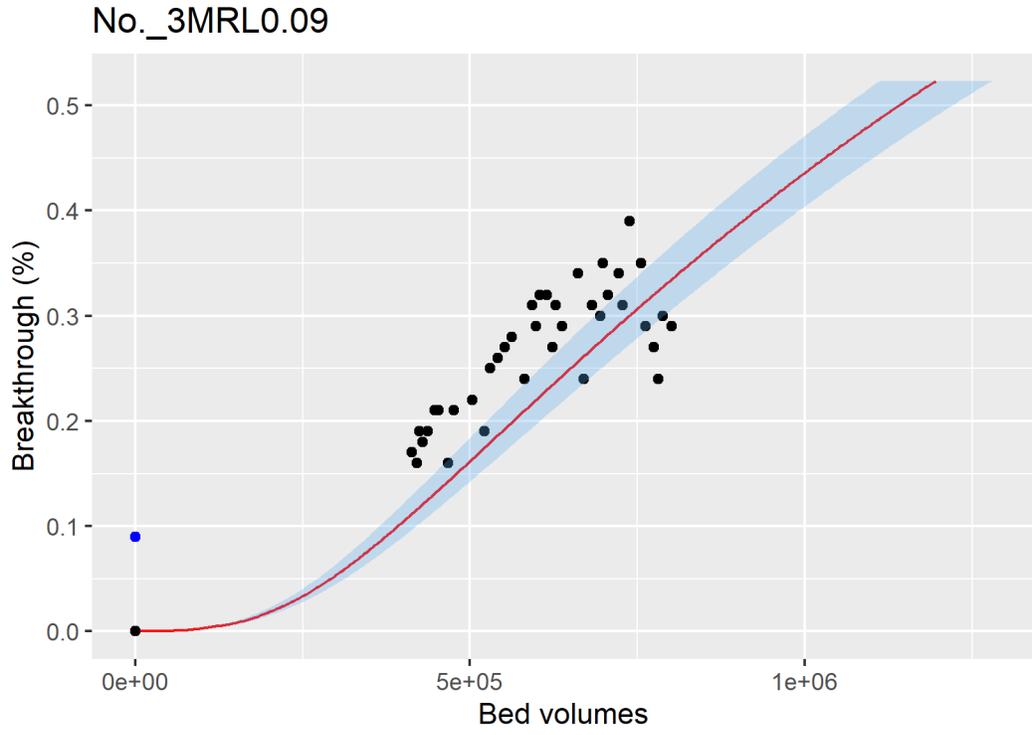


Figure ID 63- 2



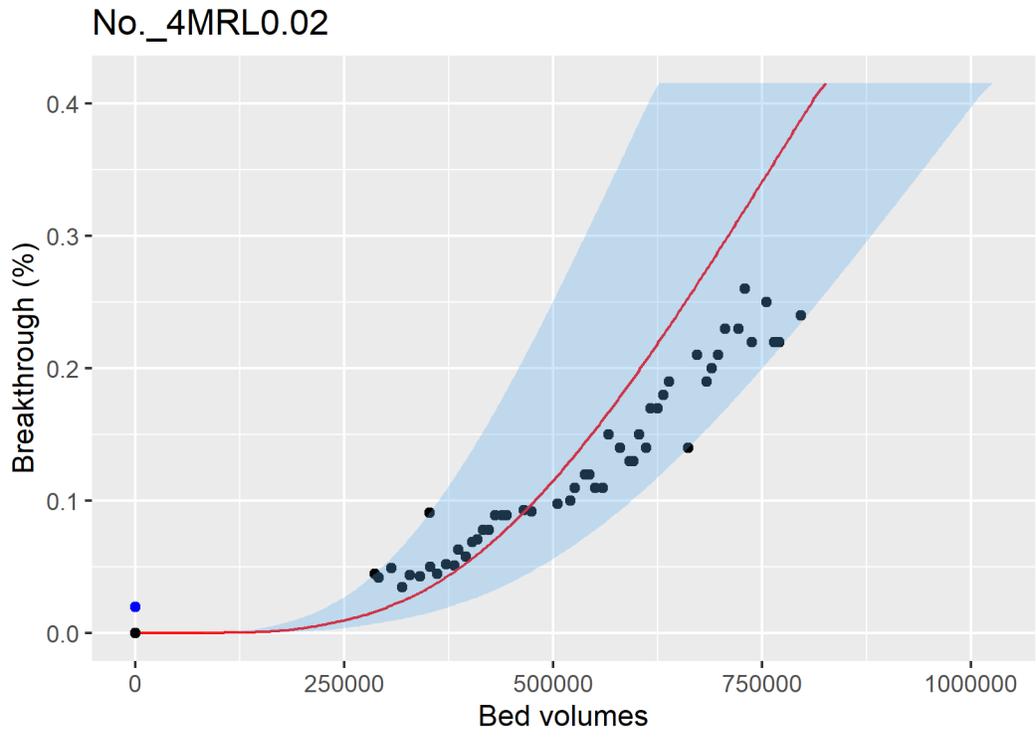


Figure ID 63- 4



Figure ID 63- 5

No._6MRL0.03

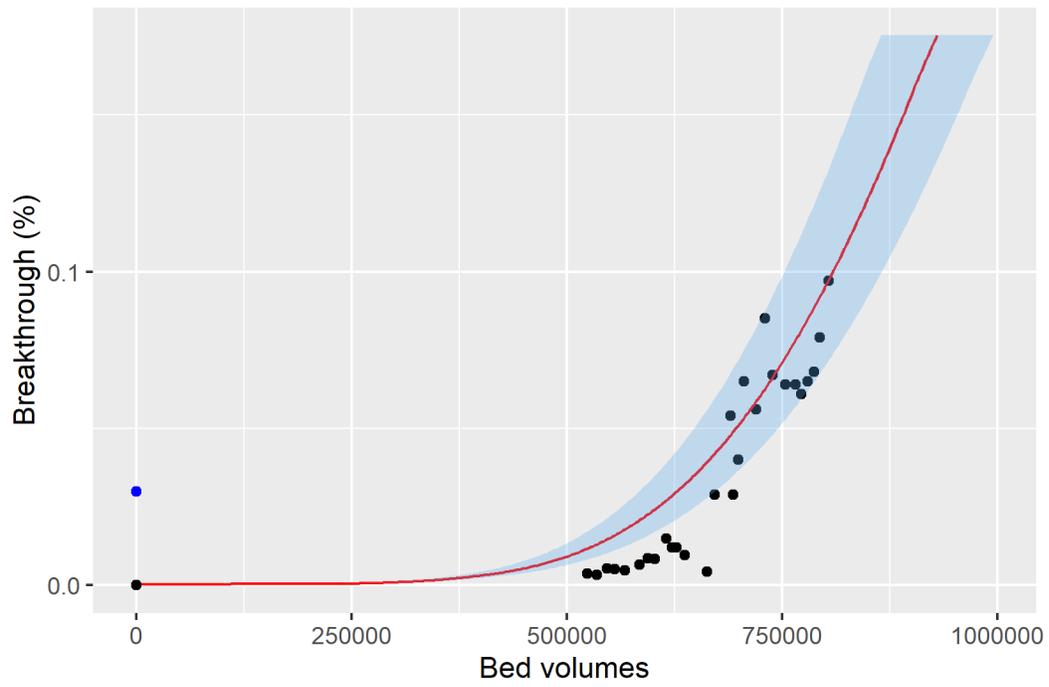


Figure ID 63- 6

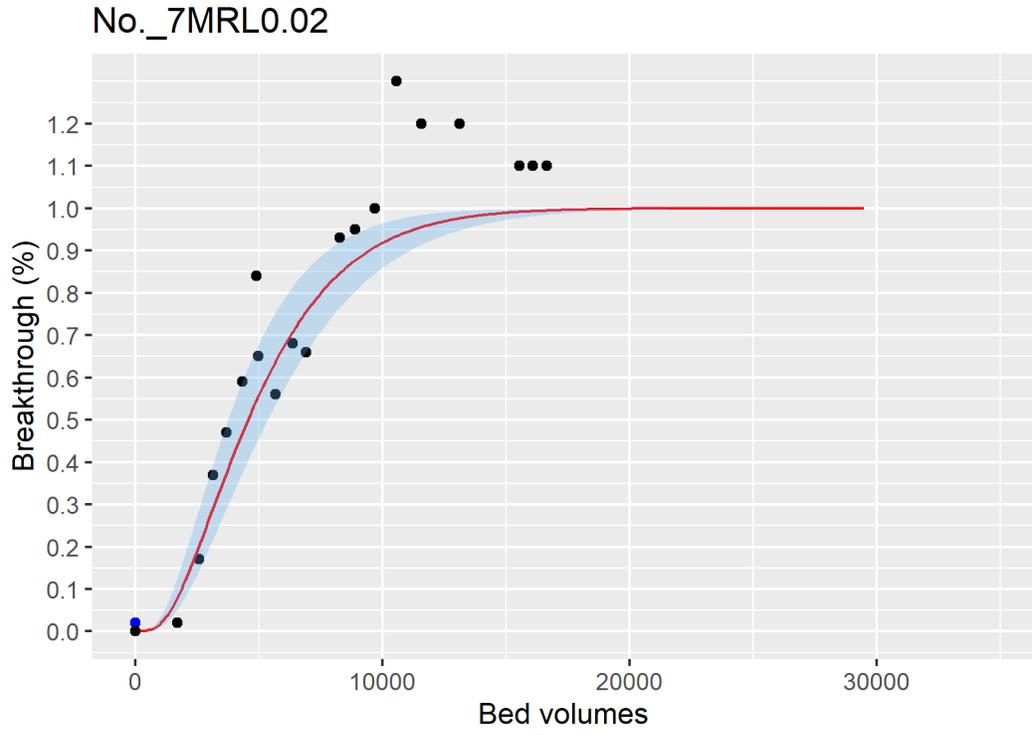


Figure ID 63- 7

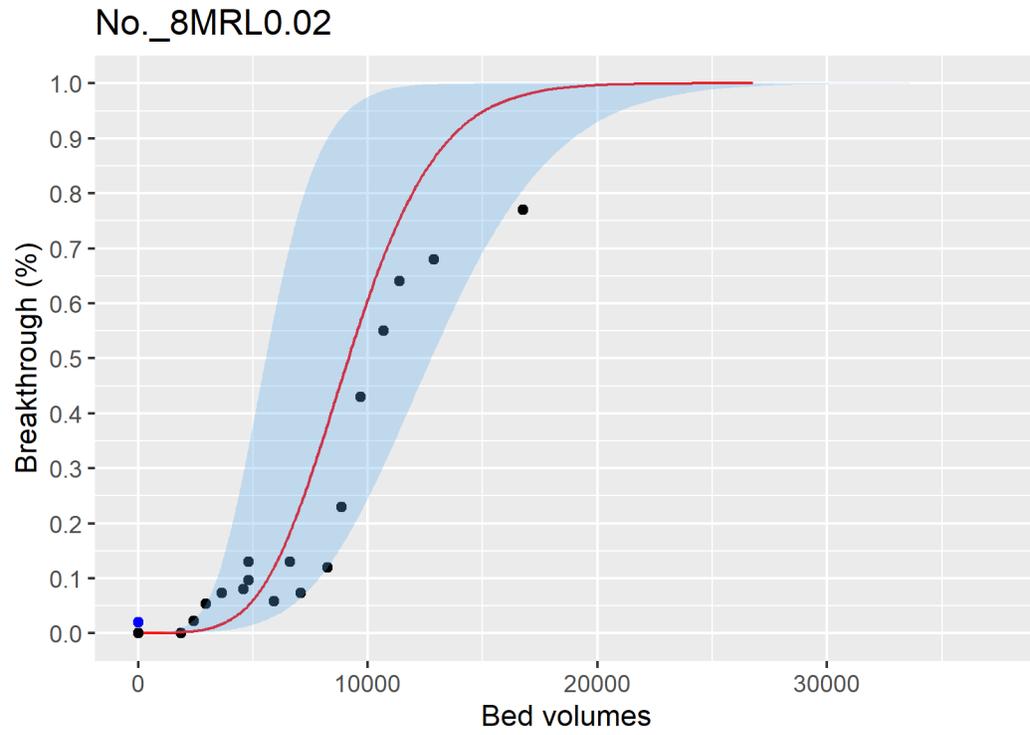


Figure ID 63- 8

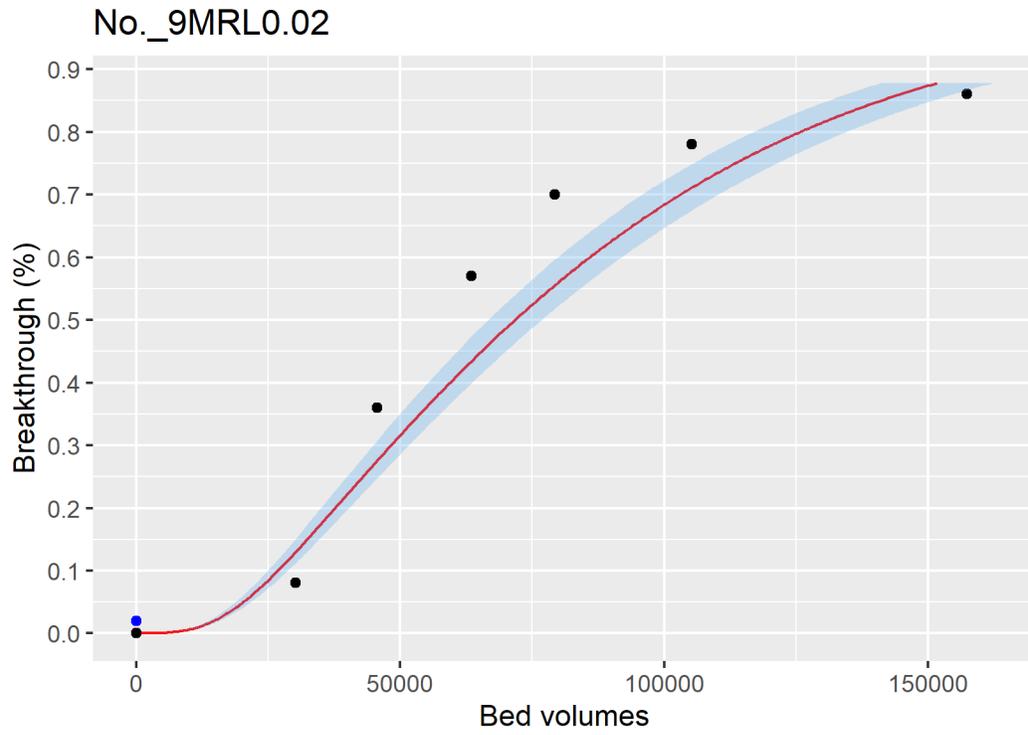


Figure ID 63- 9

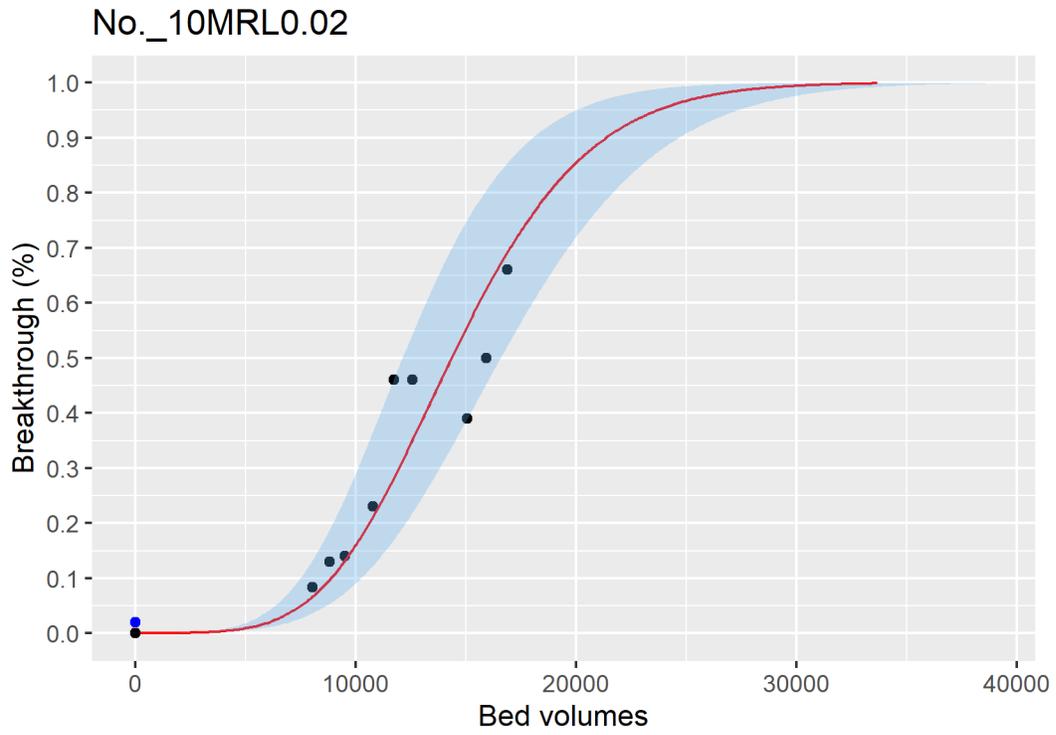


Figure ID 63- 10

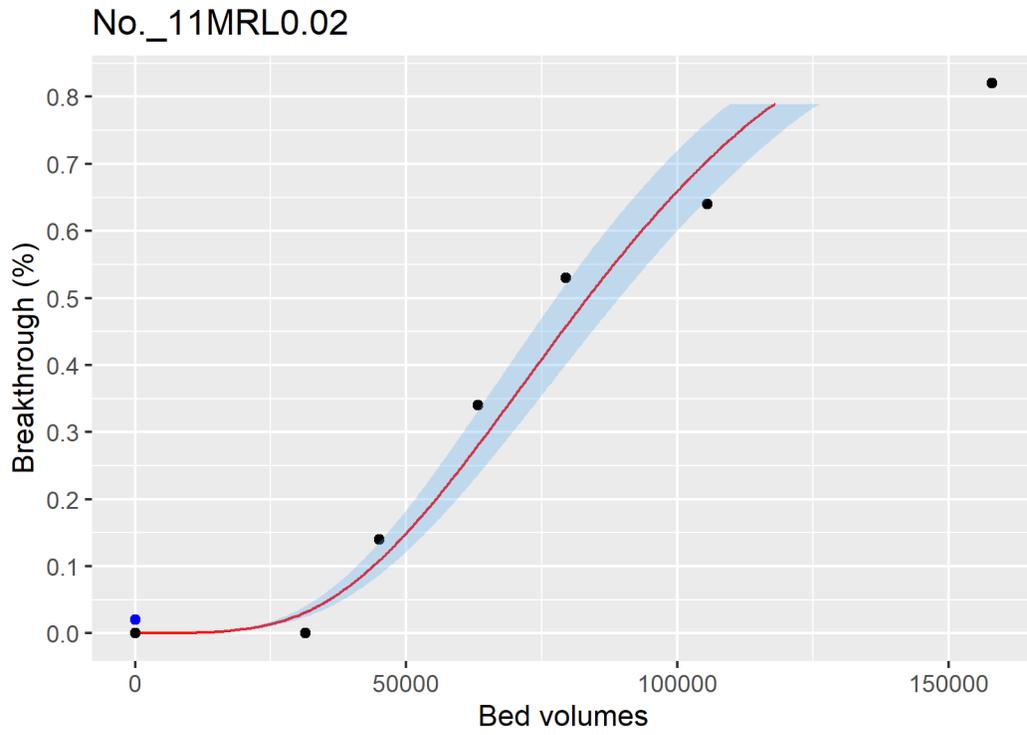


Figure ID 63- 11

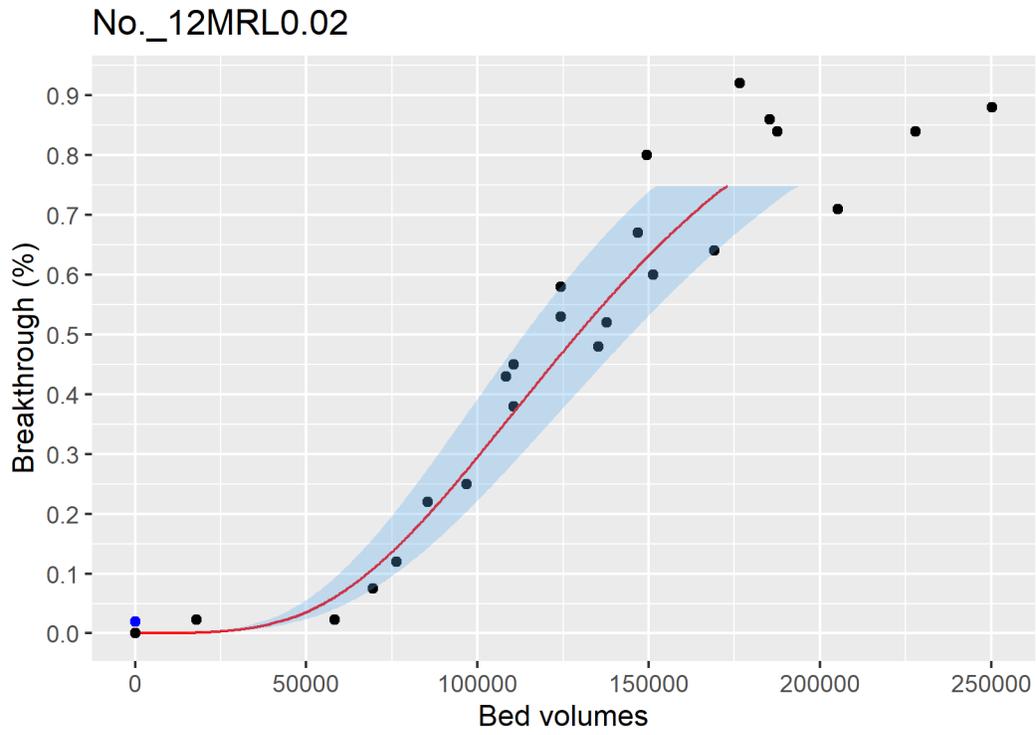


Figure ID 63- 12

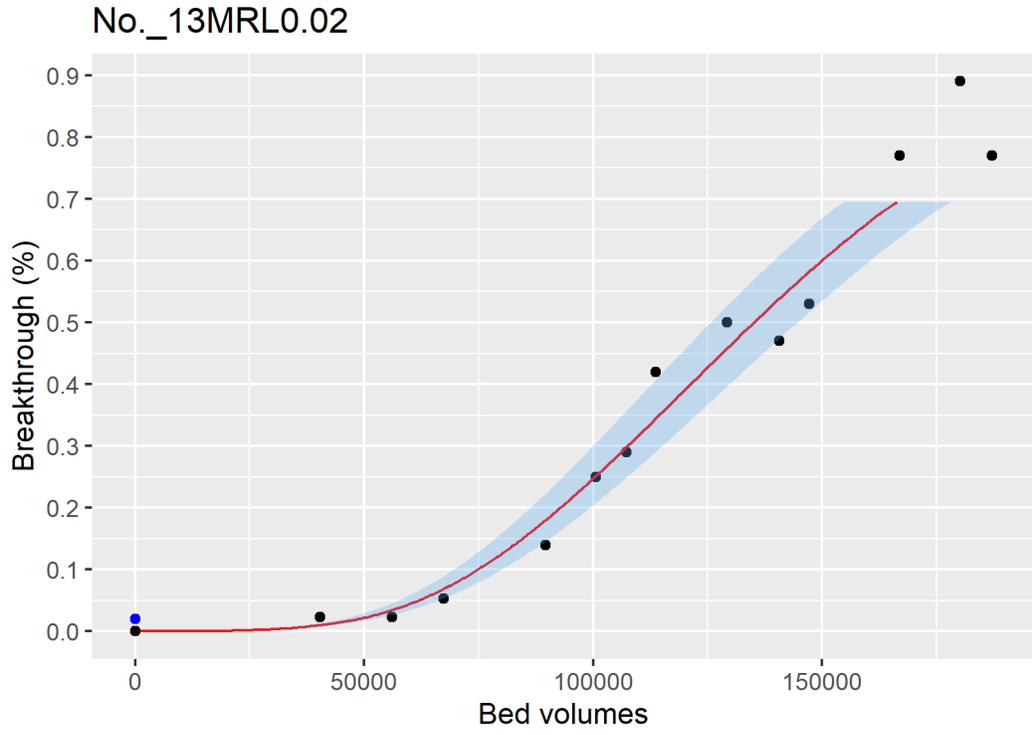


Figure ID 63- 13

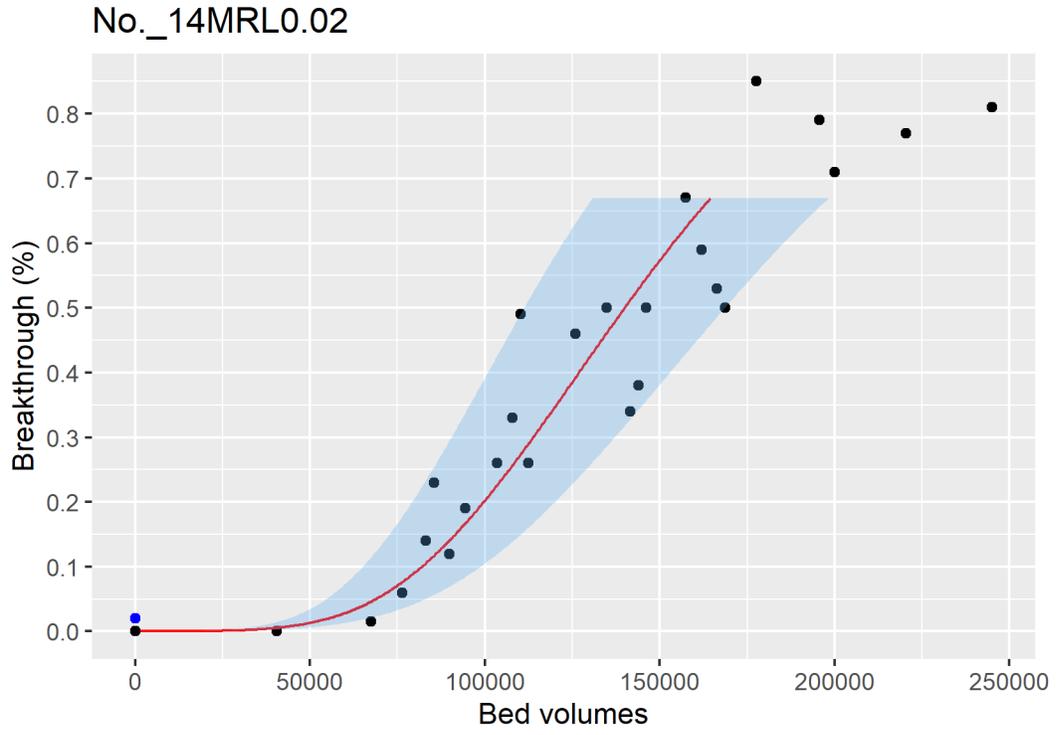


Figure ID 63- 14

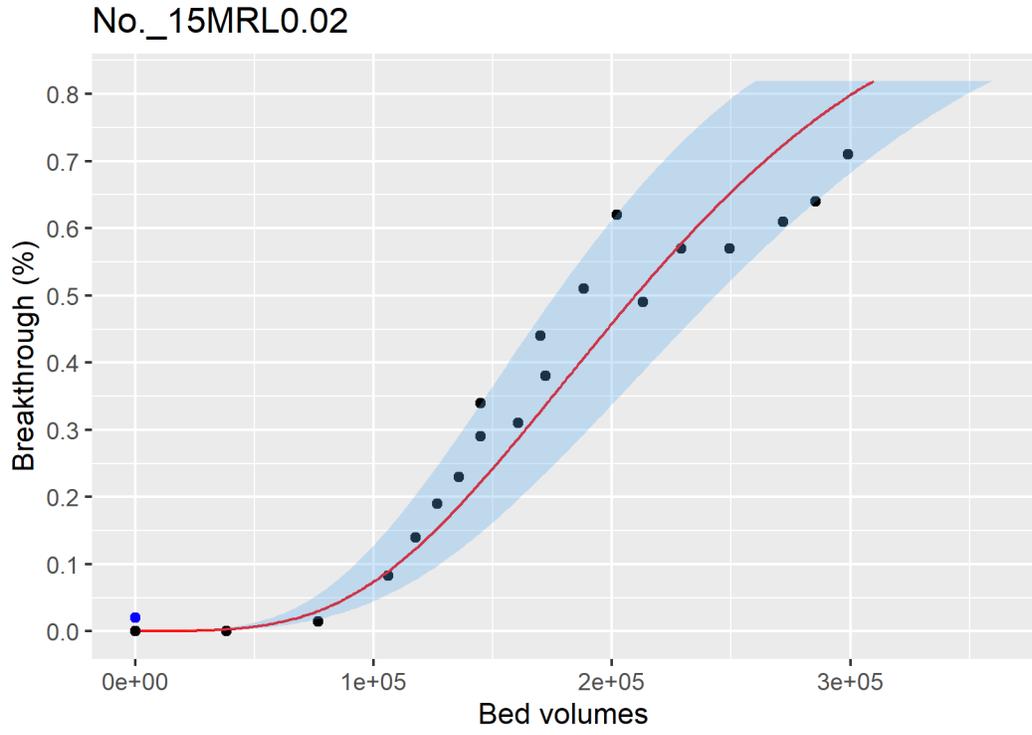


Figure ID 63- 15

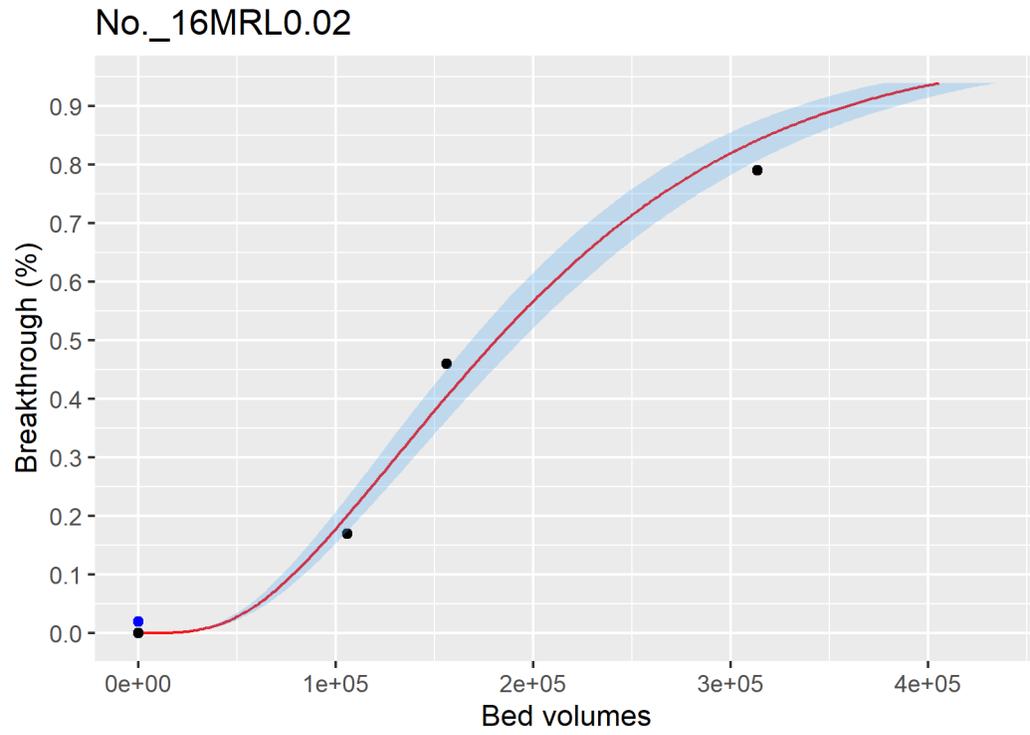


Figure ID 63- 16

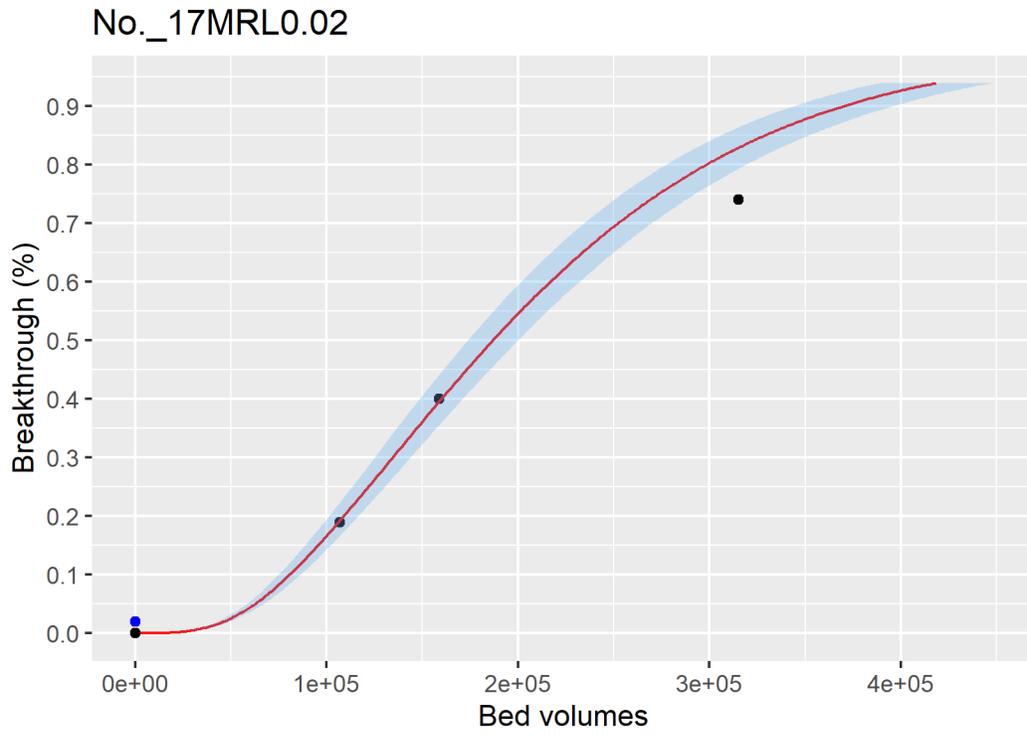


Figure ID 63- 17

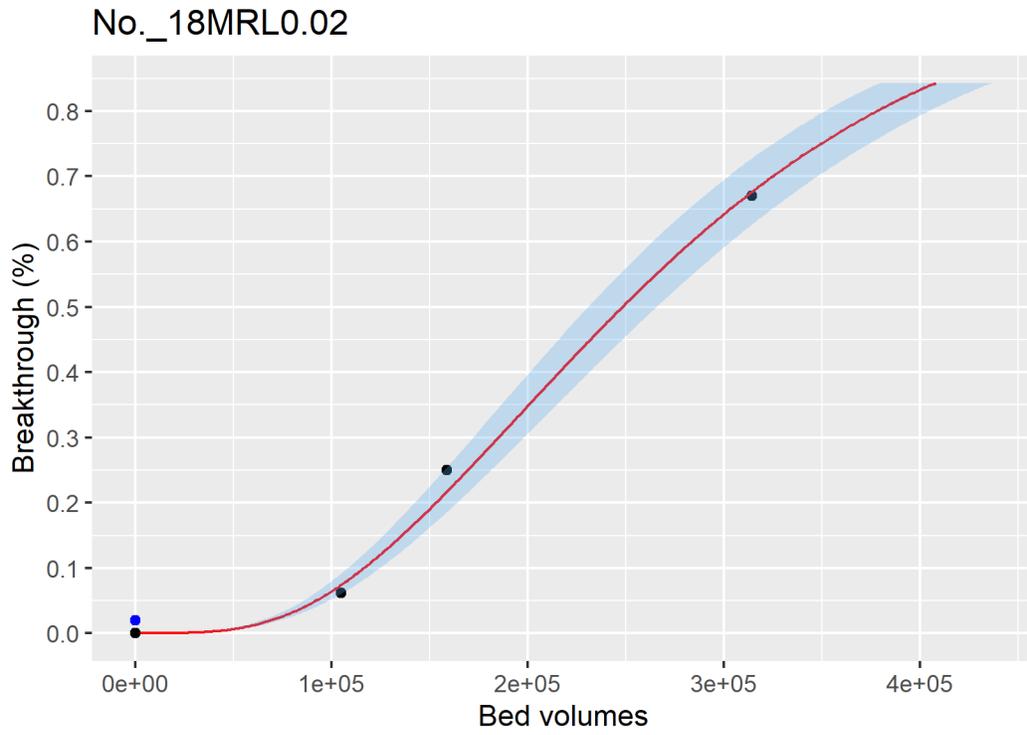


Figure ID 63- 18

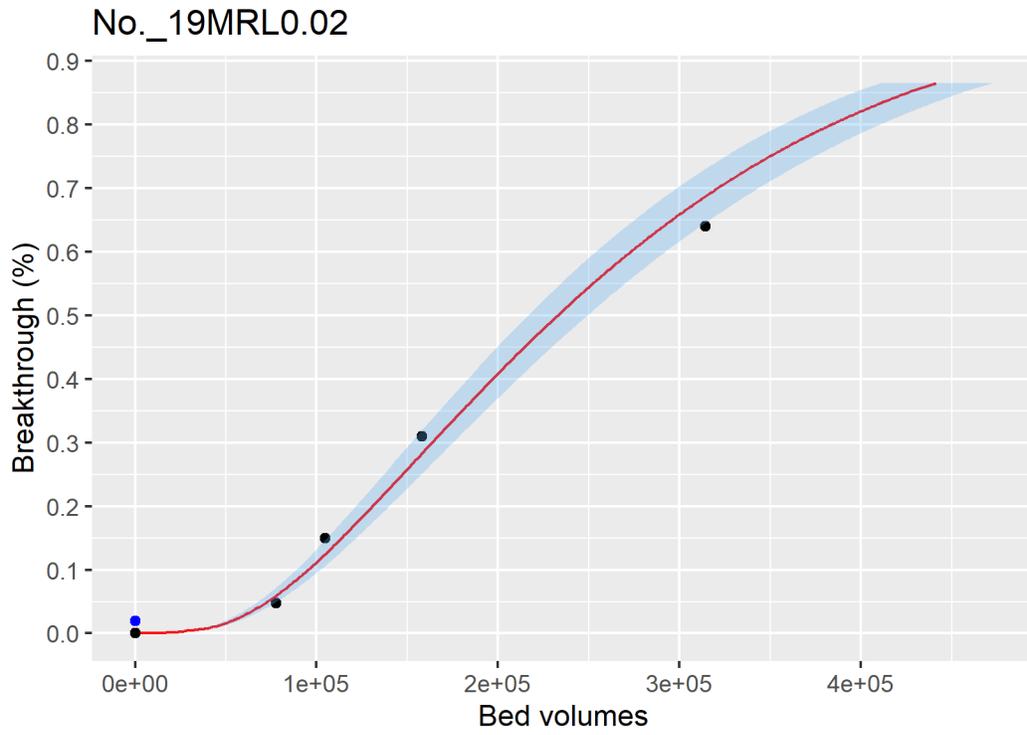


Figure ID 63- 19

Breakthrough data of Ref ID 69

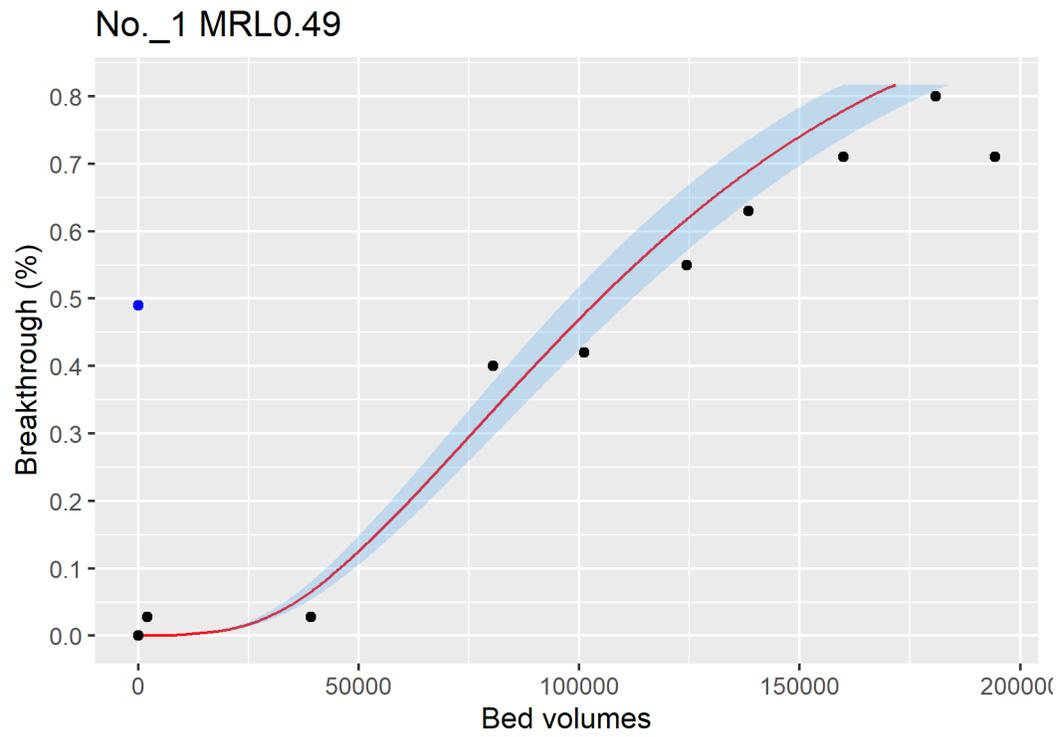


Figure ID 69- 1

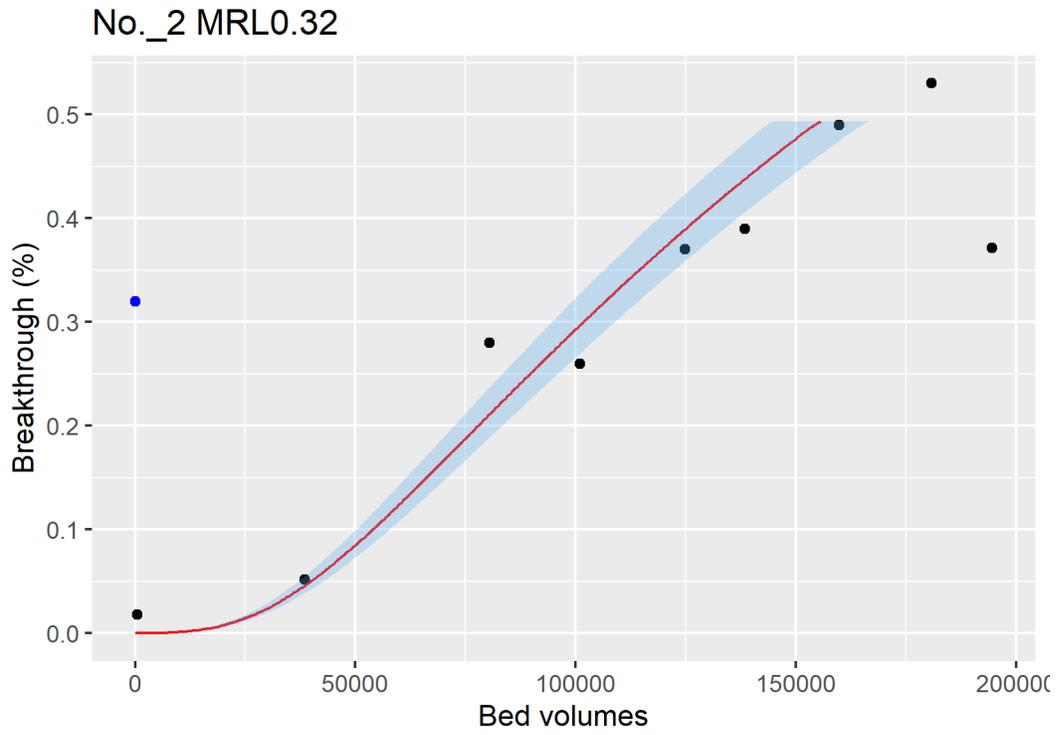


Figure ID 69- 2

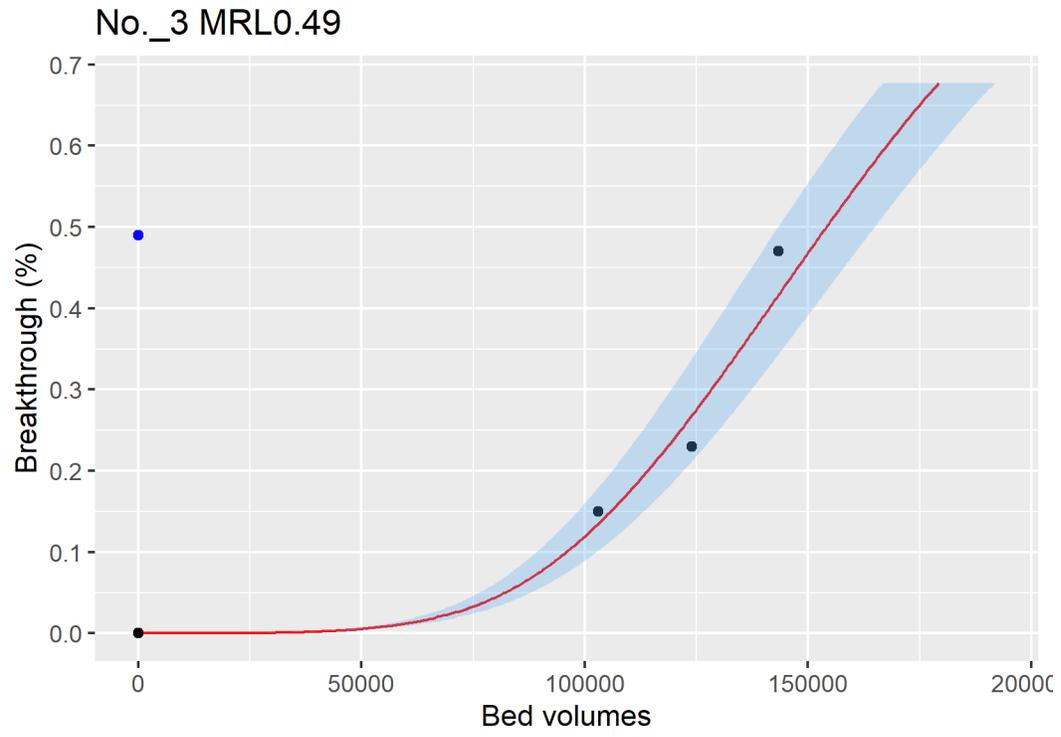


Figure ID 69- 3

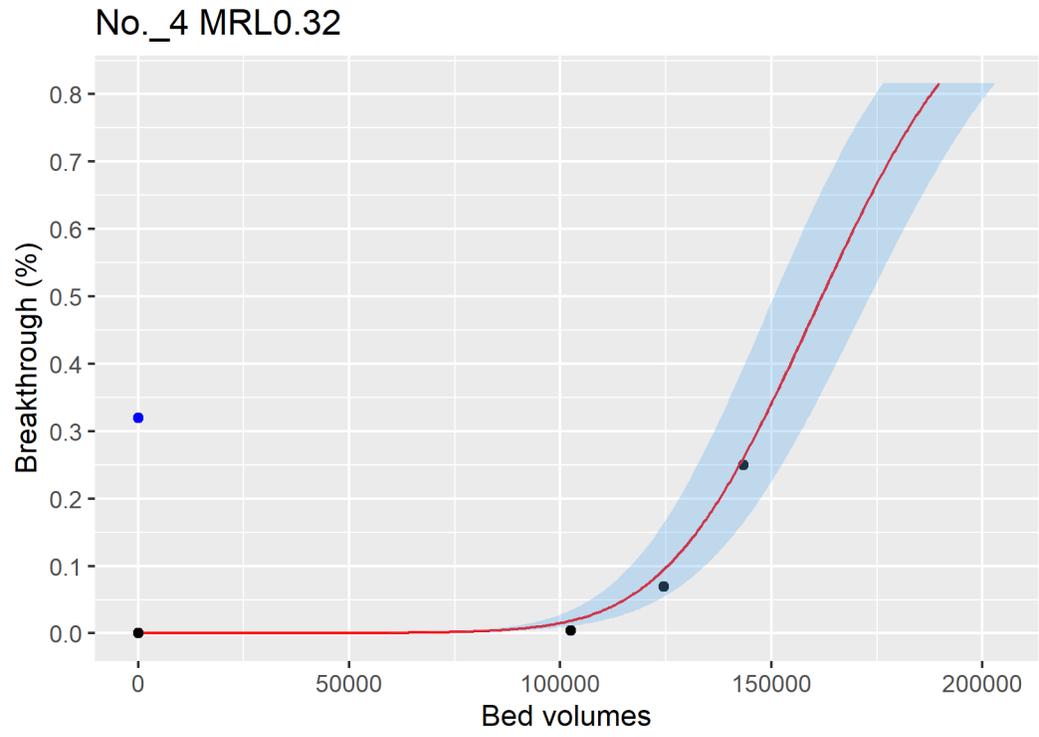


Figure ID 69- 4

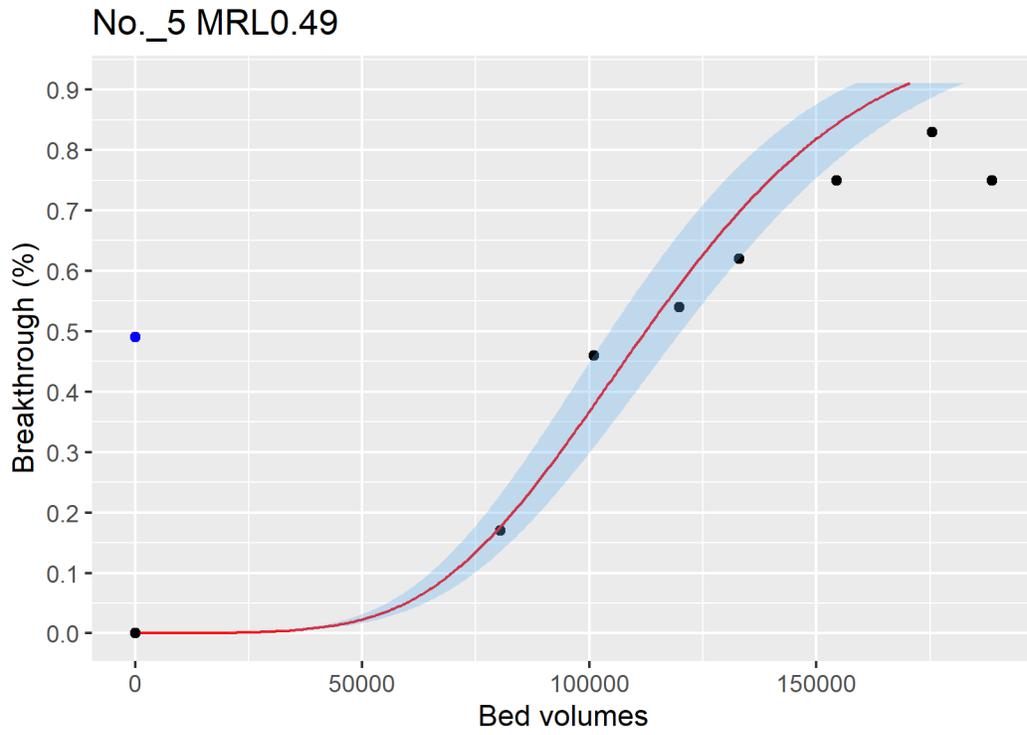


Figure ID 69- 5

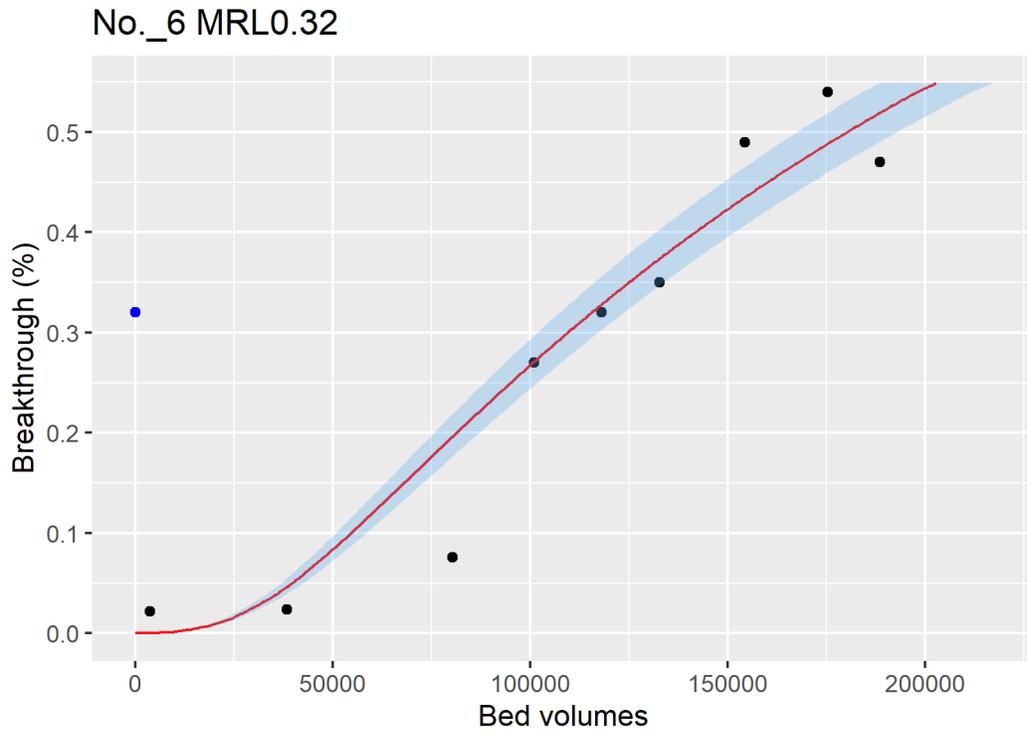


Figure ID 69- 6

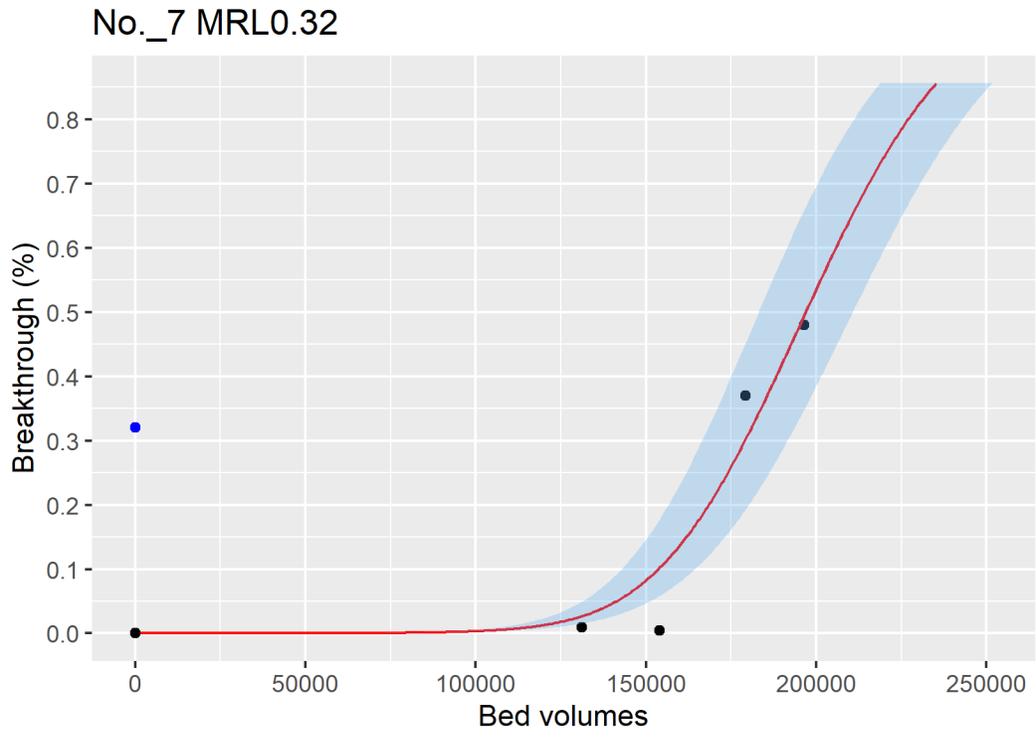


Figure ID 69- 7

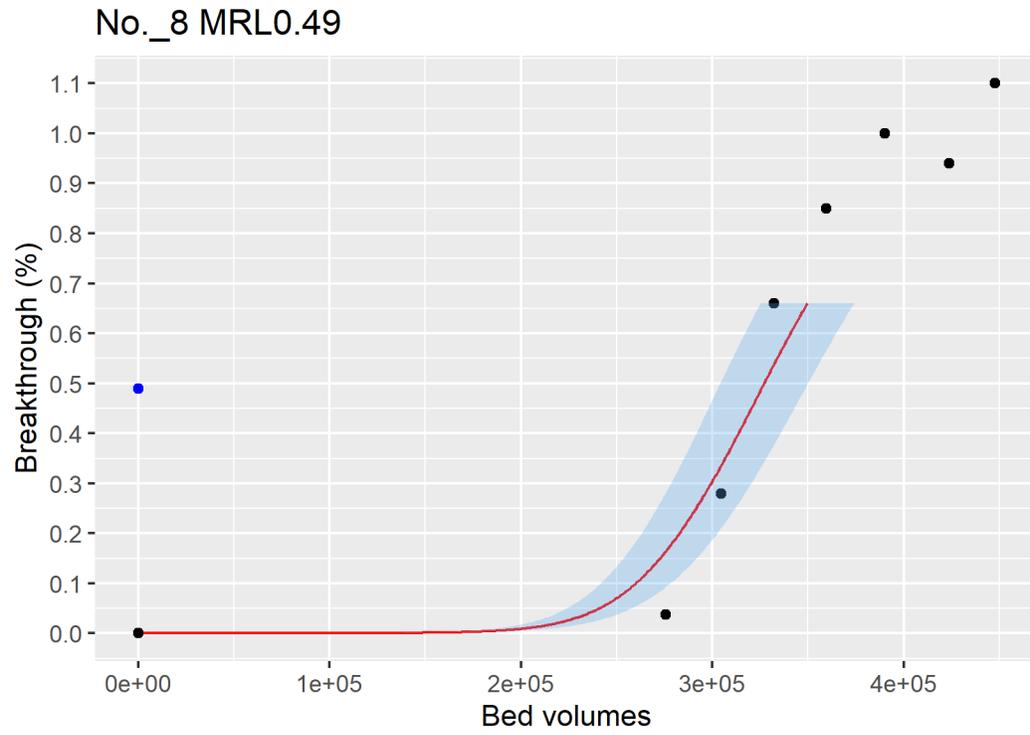


Figure ID 69- 8

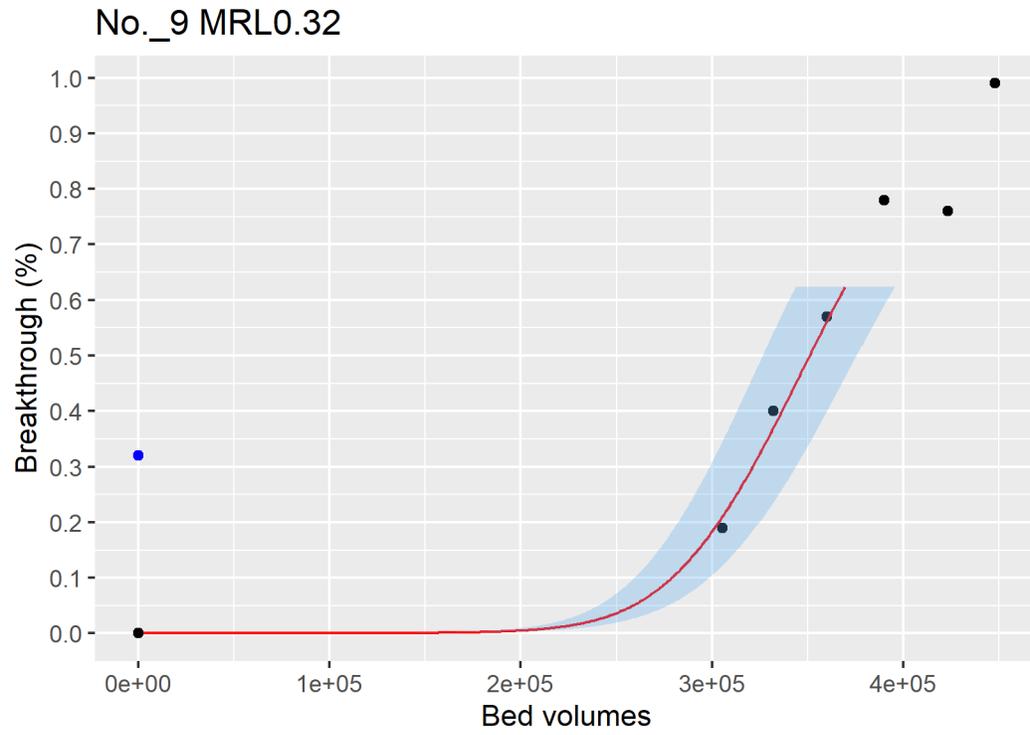


Figure ID 69- 9

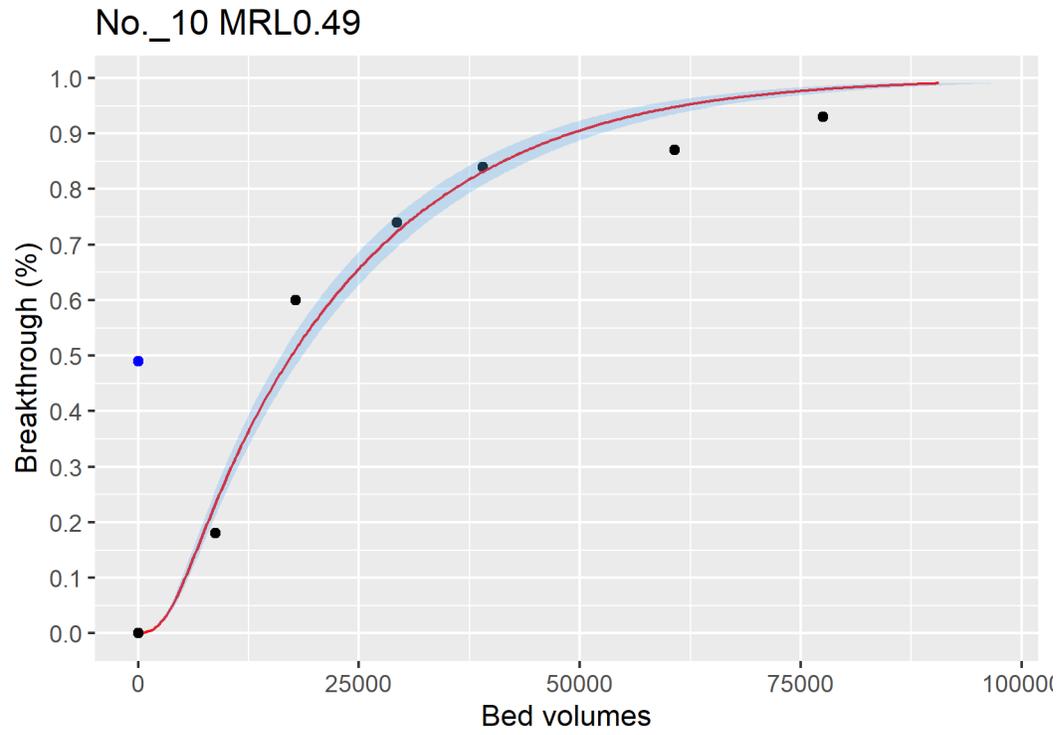


Figure ID 69- 10

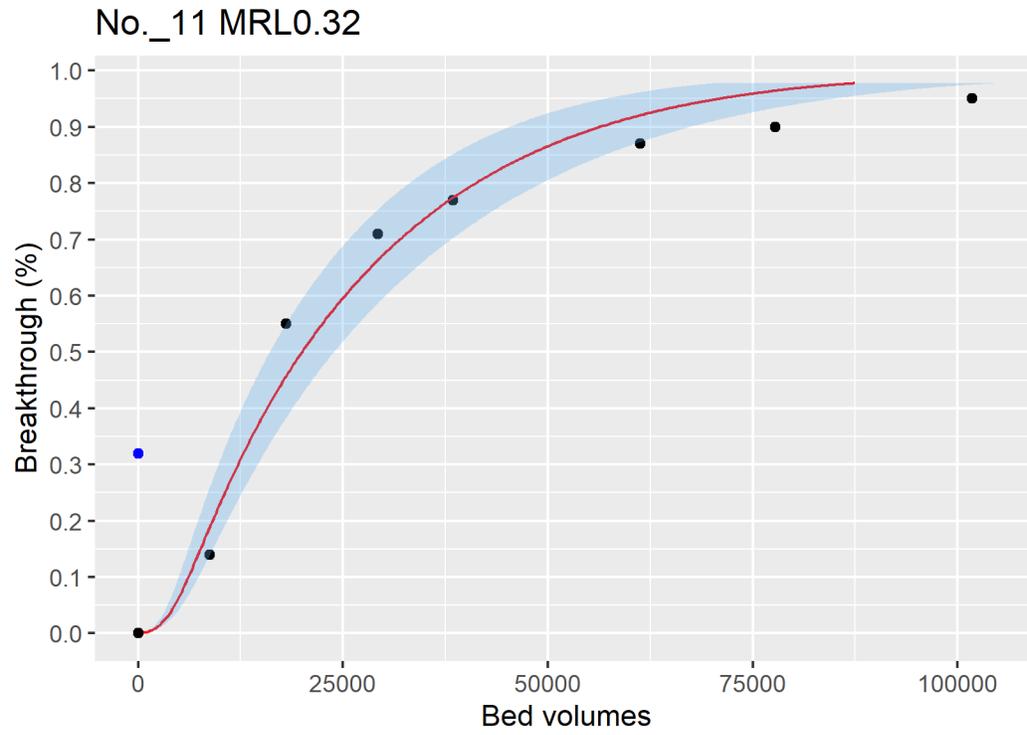


Figure ID 69- 11

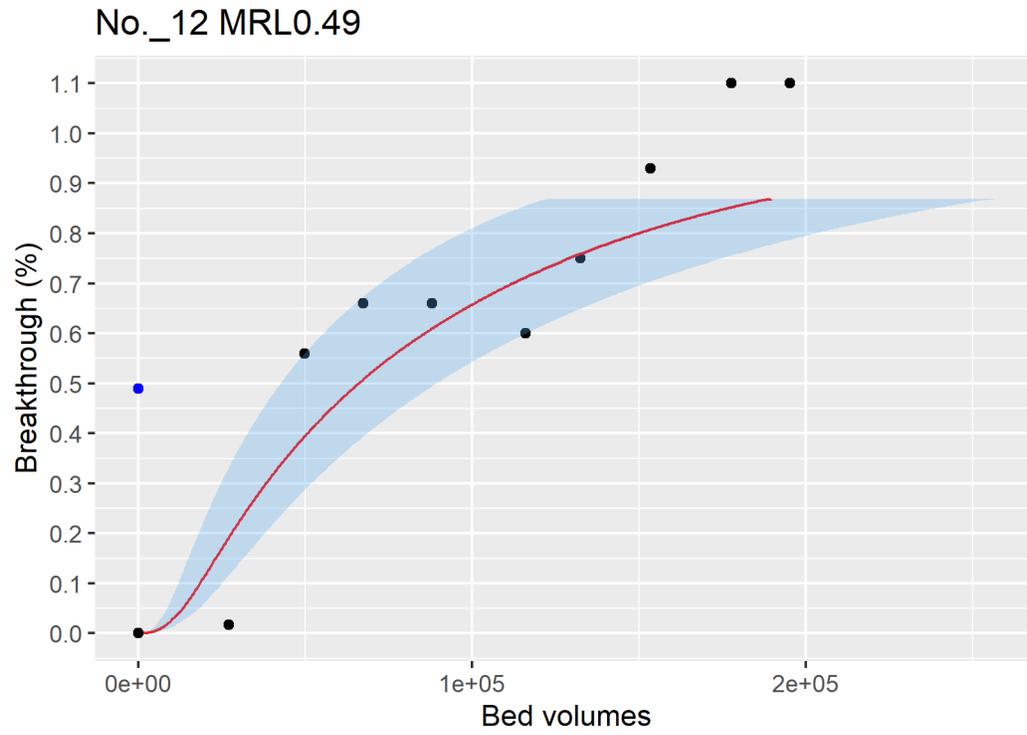


Figure ID 69- 12

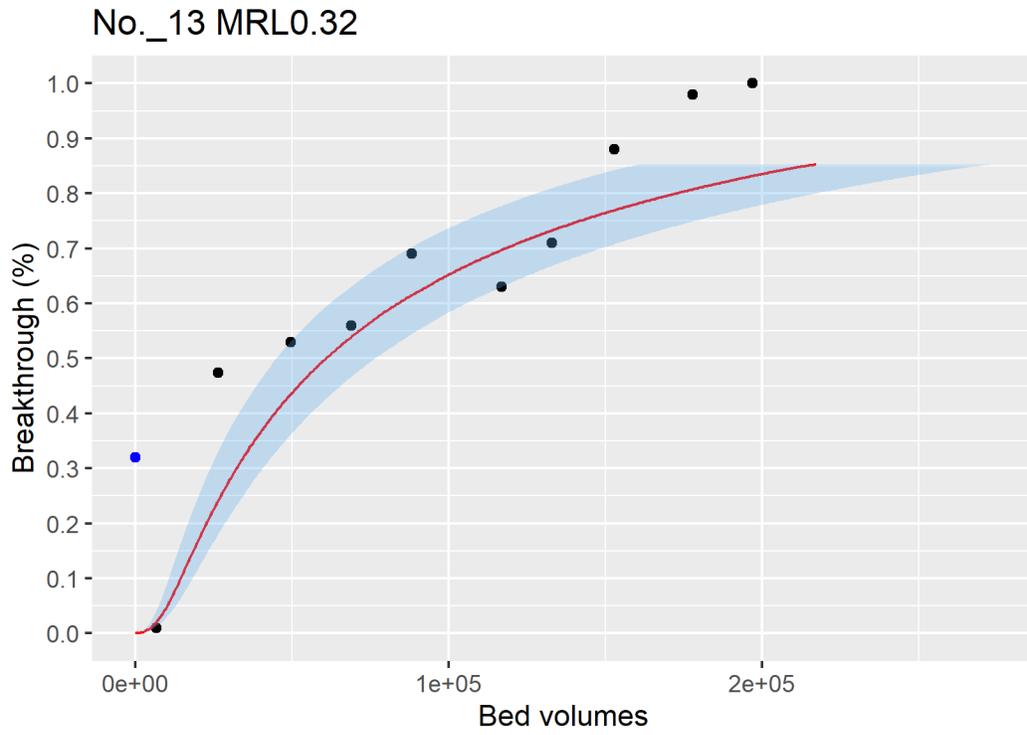


Figure ID 69- 13

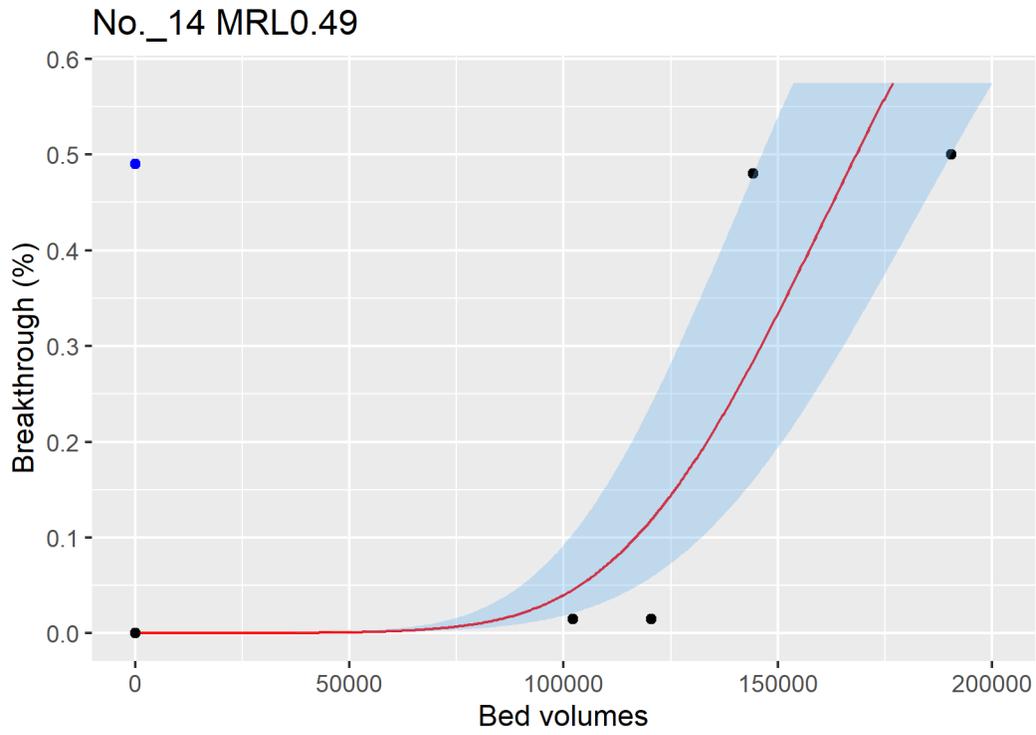


Figure ID 69- 14

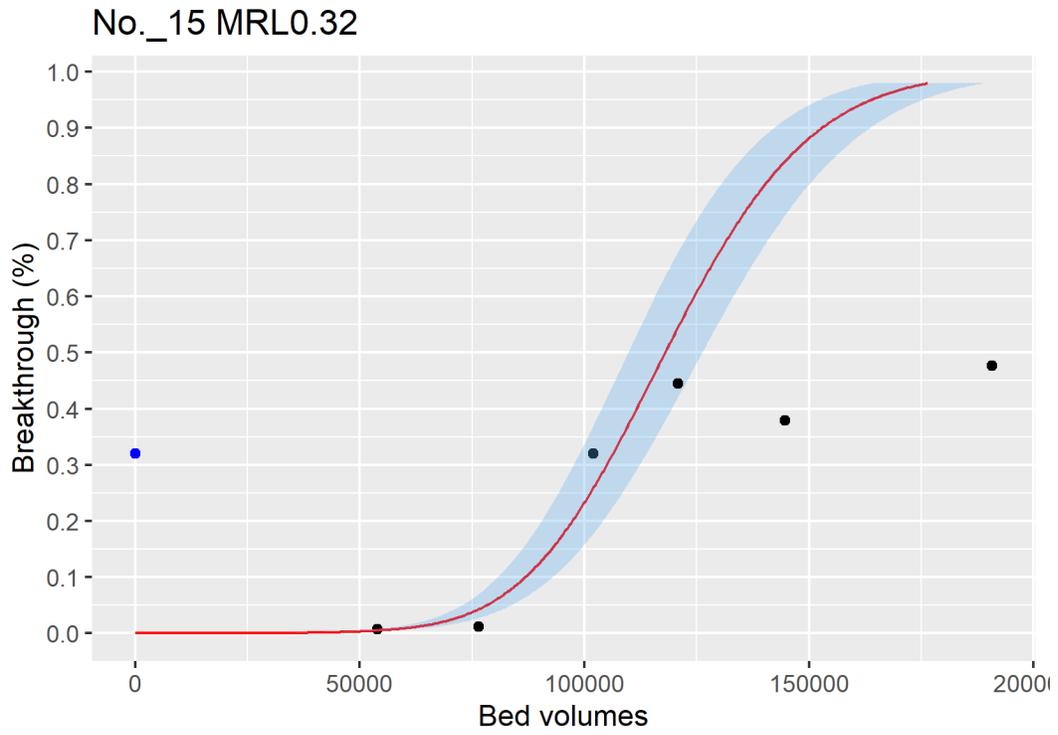


Figure ID 69- 15

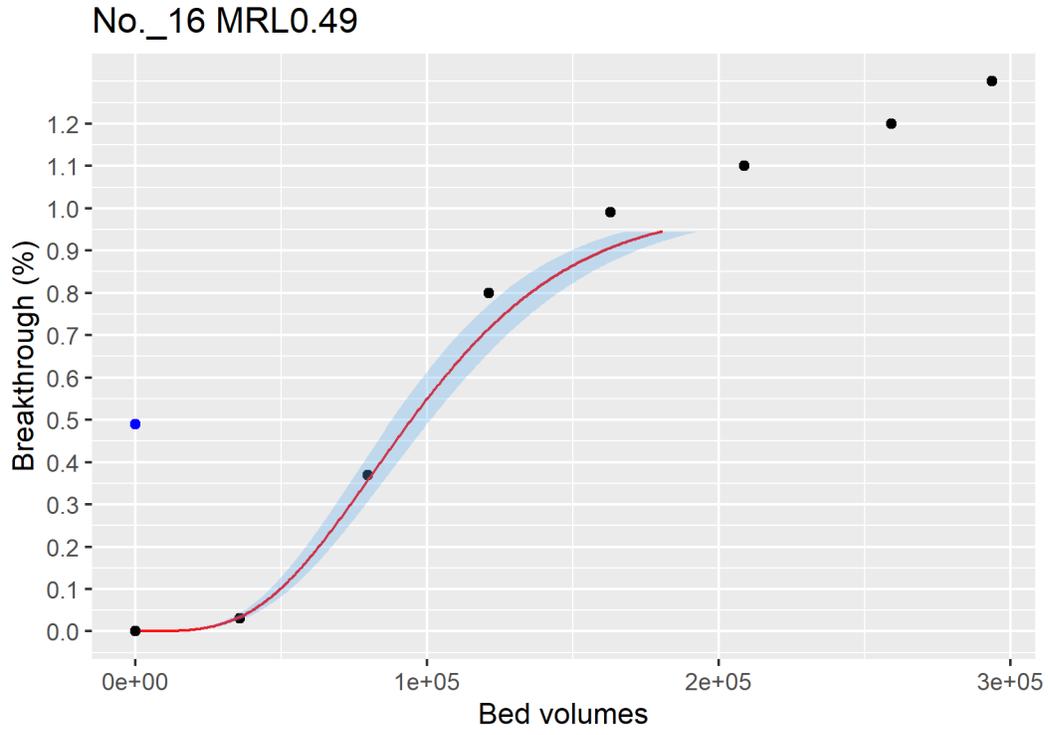


Figure ID 69- 16

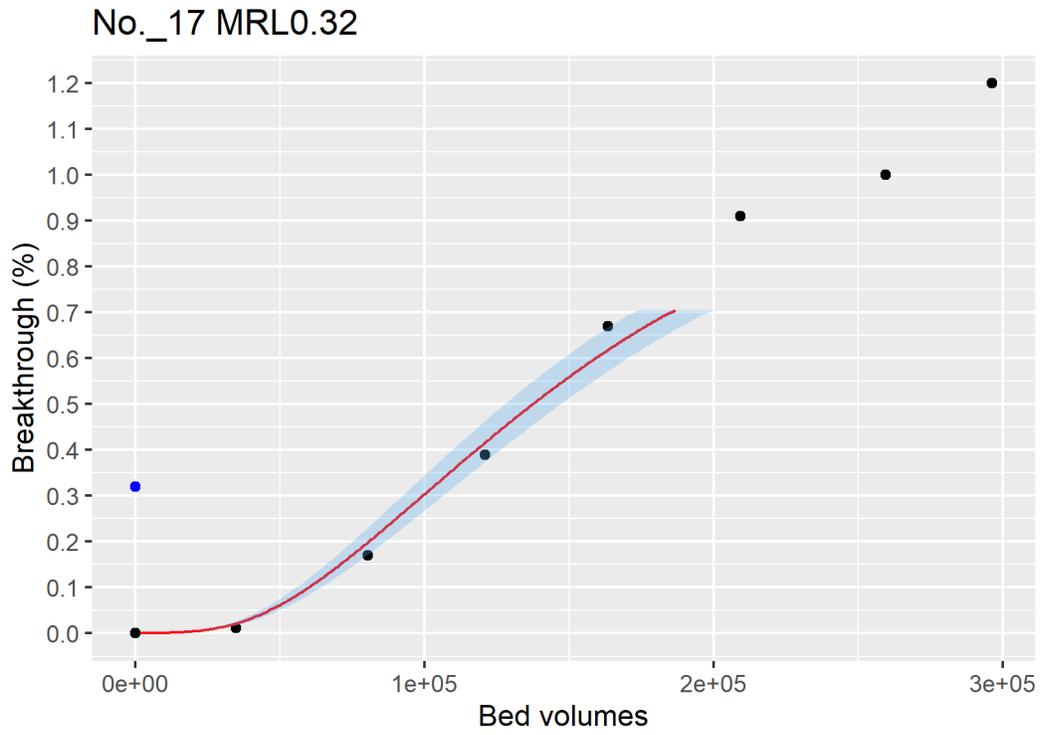


Figure ID 69- 17

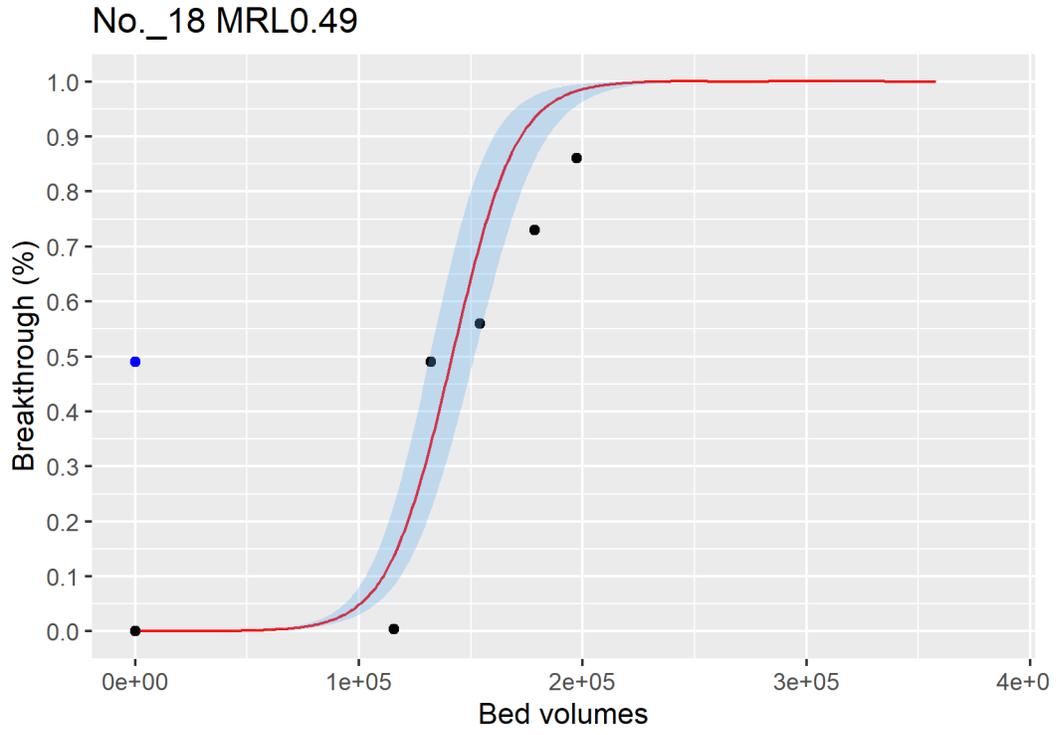


Figure ID 69- 18