SIMPLIFIED PREDICTIVE MODELS FOR CO_2 SEQUESTRATION PERFORMANCE ASSESSMENT

RESEARCH TOPICAL REPORT ON TASK #3
STATISTICAL LEARNING BASED MODELS

Reporting Period: July 1, 2014 through September 30, 2014

Principal Investigator: Dr. Srikanta Mishra
mishras@battelle.org 614-424-5712
Principal Authors: Jared Schuetter and Srikanta Mishra

Date Report Issued: October 2014

U.S. Department of Energy National Energy Technology Laboratory
DOE Award No. DE-FE0009051, Task #3

Submitting Organization:

Battelle Memorial Institute
505 King Avenue
Columbus, OH 43201
DUNS Number: 00 790 1598

This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency thereof, nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

Abstract

We compare two approaches for building a statistical proxy model (metamodel) for CO_2 geologic sequestration from the results of full-physics compositional simulations. The first approach involves a classical Box-Behnken or Augmented Pairs experimental design with a quadratic polynomial response surface. The second approach used a space-filling maxmin Latin Hypercube sampling or maximum entropy design with the choice of five different meta-modeling techniques: quadratic polynomial, kriging with constant and quadratic trend terms, multivariate adaptive regression spline (MARS) and additivity and variance stabilization (AVAS). Simulations results for CO_2 injection into a reservoir-caprock system with 9 design variables (and 97 samples) were used to generate the data for developing the proxy models. The fitted models were validated with using an independent data set and a cross-validation approach for three different performance metrics: total storage efficiency, CO_2 plume radius and average reservoir pressure. The Box-Behnken–quadratic polynomial metamodel performed the best, followed closely by the maximin LHS–kriging metamodel.

Table of Contents

Page

List of Figures

Page

List of Tables

Executive Summary

The objective of this research project is to develop and validate a portfolio of simplified modeling approaches for CO_2 sequestration in deep saline formations – based on simplified physics, statistical learning, and/or mathematical approximations – for predicting: (a) injection well and formation pressure buildup, (b) lateral and vertical CO_2 plume migration, and (c) brine displacement to overlying formations and the far-field. Such computationally-efficient alternatives to conventional numerical simulators can be valuable assets during preliminary CO_2 injection project screening, serve as a key element of probabilistic system assessment modeling tools, and assist regulators in quickly evaluating geological storage projects. The project team includes Battelle and Stanford University. Support for the project is provided by U.S. DOE National Energy Technology Laboratory and the Ohio Development Service Agency Office of Coal Development (ODSA).

This topical report presents results from Task 3 of the research, which focuses on statistical learning based models, with the objective of identifying and comparing several different ways of creating such predictive models. These are commonly called “proxy models” or “metamodels” in the geoscience literature. In applications related to subsurface flow, response variables of interest are often simulated with full physics mathematical models that are based on a large number of predictor variables. When a deep understanding of the relationship between the predictors and response is required, e.g., for optimization, many runs of the predictors at different combinations of settings may be necessary. Due to time and cost, running such a model for a large number of runs may not be feasible. The idea of a proxy model is to first acquire a small number of simulation runs at prescribed combinations of predictors, called a design matrix. These combinations are specially chosen to be representative of all possible predictor settings, called the input space. The runs are also chosen to allow estimation of large scale effects in the response. Using the observed runs, a statistical model is then developed. This model describes a specific mathematical relationship between the predictor variables and the response.

A good metamodel needs to have two characteristics. First, it must provide an accurate approximation of the full physics simulation. That is, for any combination of predictor settings, the metamodel should predict a value of the response that is close to the value one would get by running the full simulation at the same settings. Second, the metamodel must run orders of magnitude faster than the full physics simulation. If these two requirements are met, then the metamodel may be used as a proxy for the full physics simulation, and since it can produce responses quickly, it can be used to explore the input space for optimal predictor combinations.

After conducting a survey of geoscience literature, several designs and models were selected for the comparison study. Regarding designs, both experimental and sampling design approaches were considered. From the former group, Box-Behnken (BB) and augmented pairs (AP) designs were selected. BB designs are the industry standard, and AP is a competitor of the BB that uses fewer runs. From the latter group, maximum entropy (ME) and maximin Latin hypercube sampling (LHS) designs were selected. LHS designs are also popular in the geoscience literature, and ME designs are a leading competitor.

Regarding modeling techniques, five different approaches were considered. These include quadratic polynomial regression, which is common in oil and gas applications; kriging, which is a popular choice often used with LHS designs; MARS, which is another method often cited in the literature; and AVAS, which is a non-parametric modeling option. In addition, a version of quadratic modeling that uses LASSO variable selection was also considered as a more refined alternative to traditional quadratic regression modeling.

All 20 combinations of designs and models were used to predict each of three responses in a 9-input full-physics simulation of CO_2 injection into a closed reservoir using the compositional simulator, GEM. The performance of each metamodel was evaluated by fitting to this data set using three criteria: root mean squared error (RMSE), scaled RMSE, and pseudo-R^2. Evaluation was performed both for 5-fold cross-validated predictions on the training set as well as predictions on an independent test set.

In this latter case, the traditional approach of a BB design with a quadratic regression model came out as the top performer in terms of general performance scores and robustness to different responses. In particular, it beat out the other models in the validation study, and was competitive with the top performer in the cross-validation study. Of the other models, the maximin LHS with either kriging or quadratic regression models also showed good performance and robustness to different responses.

The poorest performing design was augmented pairs (AP), which was not competitive with the other three designs. This could be due to the fact that the AP design has fewer runs and is designed to work best with linear modeling approaches like quadratic regression. It does not have the kind of space-filling characteristics that one would expect for good performance using the other types of models. The worst performing modeling approaches were MARS and AVAS, which showed decent performance on some responses, but poor results on others.

1 Introduction

1.1 Background

To understand the behavior of a response function with respect to multiple predictor values, one typically needs a large number of observations to adequately cover the input space. An inefficient approach is to compute the response for all combinations of predictor values chosen on a suitably fine grid. Usually, this is not feasible. In physical experiments, some combinations of predictors may not be available to the experimenter, or may produce responses that are beyond the capability of the instrumentation to measure. In simulated experiments (e.g., finite element computer models), a large amount of computation may be required to collect each response. Therefore, computing responses over a grid of predictor values may take too long, or be too expensive to complete.

The standard method for avoiding costly data collection is to only observe the response at a subset of predictor values, and then fit a metamodel (also called a “proxy model” or “response surface model” or “reduced-order model”) to those points. Metamodels approximate the response at unobserved combinations of predictor values using the available sampled data, and are typically designed for rapid prediction. In this way, an approximate response surface can be generated for the entire input space in a short amount of time, and it can subsequently be used to meet project-specific research goals.

1.2 Previous Work

In the oil and gas literature, metamodels are often used as proxies for the underlying simulation models, especially for optimization and uncertainty quantification studies. Osterloh (2008) [1], Ekeoma and Appah (2009) [2], and Zubarev (2009) [3] provide overall guidance on sampling and metamodeling strategy for reservoir simulations. In particular, Osterloh (2008) [1] examines Latin hypercube sampling (LHS) designs and compares polynomial and kriging metamodels, Ekeoma and Appah (2009) [2] focuses specifically on LHS designs, and Zubarev (2009) [3] compares polynomial, kriging, thin plate spline, and artificial neural network metamodels.

There are also examples of specific case studies in which metamodeling was used. Kalla and White (2005) [4] compared a second order polynomial model and kriging model using an orthogonal array (OA) sample design in a gas coning case study. In this case, the second order polynomial outperformed kriging with a 36-run design in 14 variables. Anbar (2010) [5] settled on first order polynomial models for fitting outputs of a CMG STARS simulation for CO_2 sequestration in deep saline carbonate aquifers. The models were fit using LHS designs of size 100 over 16 variables. Finally, Wriedt *et al*. (2014) [6] used a Box-Behnken design and a stepwise quadratic regression model to develop probability distributions for responses related to CO_2 injection into deep saline reservoirs.

1.3 Scope and Organization

The aim of this research project is to evaluate and compare several combinations of study designs and metamodeling techniques in the context of injection of CO_2 into a closed-volume reservoir. In conducting this research, there are two outcomes of interest. First of all, as in the studies mentioned above, this study provides another piece of information regarding which designs and metamodels are most effective in this application area. Such information may be of use to other investigators who work in carbon sequestration. Secondly, the hope is that this task produces one or more metamodels that can predict the simulated responses with an acceptable degree of accuracy, and can be compared to some of the other simplified approaches being investigated under this project.

The remainder of this document contains a more detailed description of this study and the results. Sections 1 and 1 contain background information on experimental designs and metamodels, respectively, that are commonly used in response surface modeling. Section 1 identifies several frameworks for evaluating such metamodels. Section 4.5 describes a case study that was carried out using pre-existing data from the Arches province in the American Midwest. This work was the subject of a presentation at the 8th International Congress on Environmental Modelling and Software (iEMSs), and subsequent publication in the conference proceedings. Finally, Section 5 describes the application of the methodology from Section 4 using results of a reservoir simulation study of CO_2 injection into a bounded saline formation.

2 Experimental Designs

2.1 Factorial Designs

Factorial designs are typically used for variable screening or response surface optimization. These designs set each of the predictor variables at one of several levels, usually a “low” and “high”, or a “low”, “center”, and “high”. Typically, “low”, “center”, and “high” levels are denoted -1, 0, and +1, respectively. When the number of inputs is small, factorial designs can use a relatively small number of runs to explore the predictor space and allow estimation of simple linear or quadratic models, which can in turn be used to identify the regions of the space corresponding to optimal response values. As long as the response surface can be adequately modeled with simple functions, factorial designs are sufficient; however, other designs may be necessary for understanding the behavior of more complex functions (see Section 2.2). As the number of inputs increases, full factorial designs can get quite large due to exponential growth in the number of runs. In that case, smaller factorial designs can be used to understand the response surface. A description of several of those designs is given below.

2.1.1 Plackett-Burman

Plackett-Burman designs [7] are a class of designs that are chosen to provide the best possible estimates of the main effects of the predictors on the response. Main effect estimates for Plackett-Burman designs have the minimum variance possible for a limited number of runs. The designs themselves are chosen so that each unique combination of levels for every pair of predictors appears the same number of times throughout the design. Typically, there are only two levels (+1 and -1) assigned for each input. While main effects are estimable, interaction effects between predictors are typically confounded with the main effects and cannot be separated without additional runs. Plackett-Burman designs for k inputs can have a number of unique runs anywhere between the nearest multiple of 4 from k (not any larger than $k + 4$) and 2^k runs, where they become full 2^k factorial designs. One example of a Plackett-Burman design is shown in Figure 1. In this case, the design has 12 runs over 3 inputs, although there are only $2^3 = 8$ unique runs; the other runs are duplicates.

X_1	X_2	X_3
1	1	1
-1	1	-1
-1	-1	1
-1	-1	-1
1	-1	-1
1	1	-1
1	1	1
-1	1	1
1	-1	1
-1	1	-1
-1	-1	1
1	-1	-1

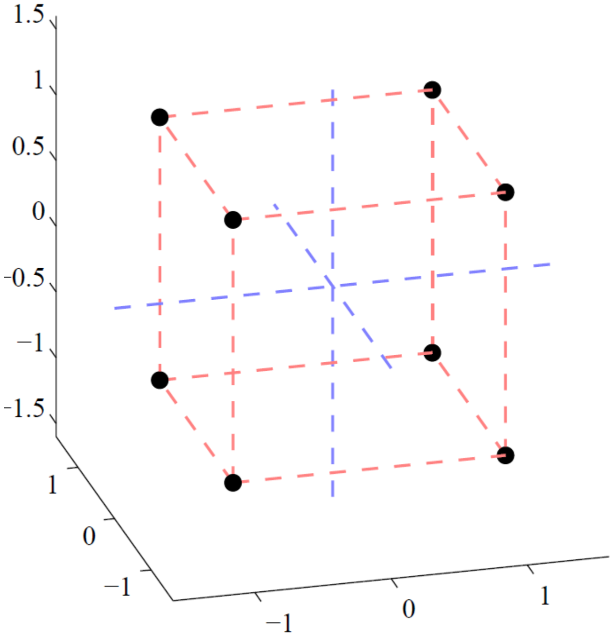


Figure 1. An example of a Plackett-Burman design for three inputs (left) and its representation in the predictor space (right).

2.1.2 Central Composite and Box-Behnken

Central Composite (CC) and Box-Behnken (BB) [8] designs are related methods that use three levels for each predictor. Both designs make judicious use of observations and allow estimation of linear and quadratic terms in a polynomial surface model. The CC design samples points at the corners of a hypercube in the input space and at points at the centers of the faces, as shown in Figure 2. In contrast, the BB design samples points along the edges of the hypercube, as shown in Figure 3**.** One commonly cited disadvantage to the CC design is that combinations where multiple predictors have simultaneous extreme values (i.e., at the corners of the hypercube) are typically unrealistic. The BB design places observations at less extreme predictor combinations to provide a better model fit over the center of the space.

X_1	X_2	X_3
-1	-1	-1
-1	-1	+1
-1	+1	-1
-1	+1	+1
-1.68	0	0
0	-1.68	0
0	0	-1.68
0	0	0
0	0	1.68
0	1.68	0
1.68	0	0
+1	-1	-1
+1	-1	+1
+1	+1	-1
+1	+1	+1

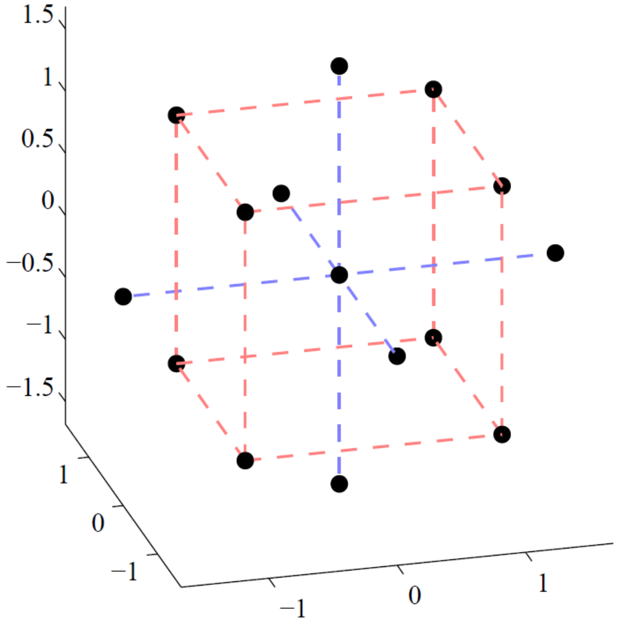


Figure 2. Central Composite design for three inputs (left) and its representation in the input space (right).

X_1	X_2	X_3
-1	-1	0
-1	0	-1
-1	0	+1
-1	+1	0
0	-1	-1
0	-1	+1
0	0	0
0	+1	-1
0	+1	+1
+1	-1	0
+1	0	-1
+1	0	+1
+1	+1	0

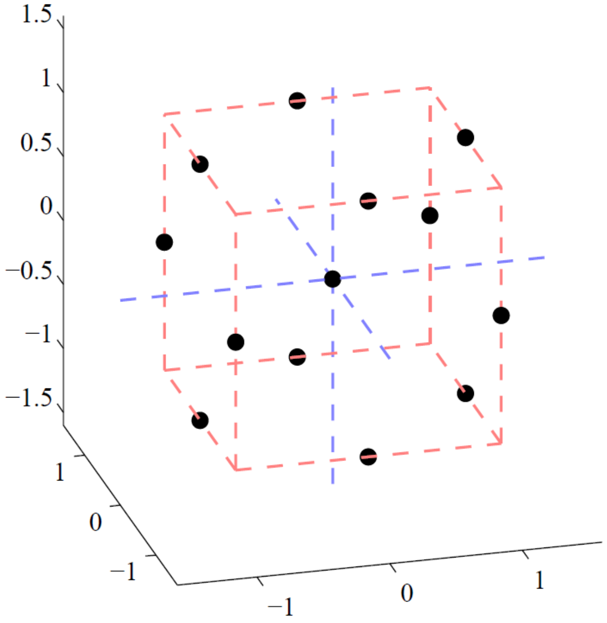


Figure 3. Box-Behnken design for three inputs (left) and its representation in the predictor space (right).

2.1.3 Augmented Pairs

The augmented pairs (AP) design described by Morris (2000) [9] is an alternative to Central Composite and Box-Behnken designs, and is made to work well with sequential response surface search and optimization procedures. The strength of the AP design is that it builds the latter 3-level targeted design by augmenting the 2-level design used in the initial exploration phase. In this way, none of the runs are wasted. To construct an AP design, one begins with a 2-level (preferably orthogonal) design, with observations at various combinations of {-1, +1} for the different factors. An example of such a design is the Plackett-Burman design. To augment the design, first n_0 center-point replicates are added (e.g., repeated runs with level 0 for all factors). Next, each pair of runs in the 2-level design are used to construct a new single run, where the levels of the factors in the new run are $L_{\text{new}} = -0.5*(L_1 + L_2)$. Here, L_1 and L_2 are the factor levels in the two parent runs, so that the new level of the factor will be 0 if the original runs were at +1 and -1, -1 if both original runs were at +1, or +1 if both original runs were at -1. The resulting design is smaller in size than a CC or BB design, but still retains many of their advantages.

X_1	X_2	X_3
0	0	0
1	1	1
-1	1	-1
1	-1	-1
-1	-1	1
0	1	0
1	0	0
0	0	1
0	0	-1
-1	0	0
0	-1	0

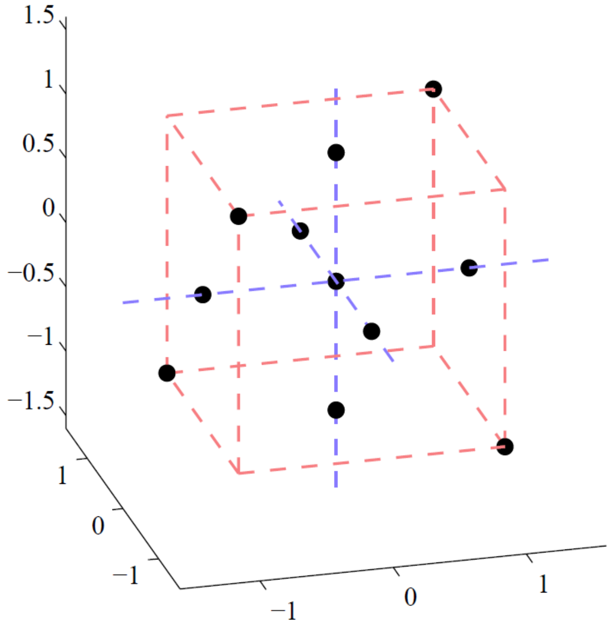


Figure 4. Augmented pairs design for three inputs (left) and its representation in the predictor space (right).

2.1.4 Run Comparison

Figure 5 shows a comparison of the number of unique runs required by each type of factorial design described above. The most expensive design is a full 2-level factorial design, which has 2^k runs for k inputs (see the curve indicated in magenta). Such designs are a special case of Plackett-Burman design, but Plackett-Burman designs can have as low as $k + 1$ runs. The minimum number of runs for a Plackett-Burman design is shown in Figure 5 in cyan. Note, however, that such designs do not allow estimation of much more than the main effects of the inputs, and are not good in general for response surface modeling. Of the 3-level designs, the Box-Behnken and Central Composite designs (red and green, respectively) have comparable numbers of unique runs, while the augmented pairs design typically has fewer runs. The maximum number of 3-level runs possible is 3^k (not shown).

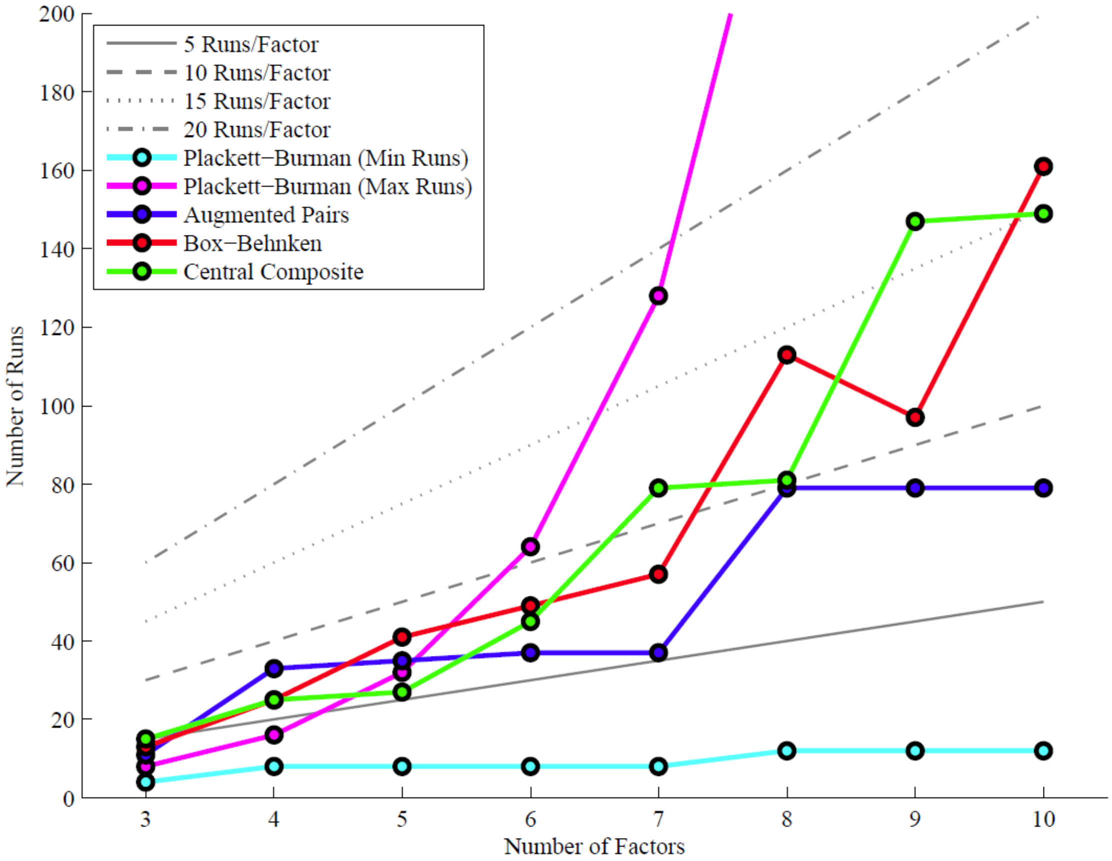


Figure 5. A comparison of the number of unique runs needed for the different factorial designs described in this section.

2.2 Sampling Designs

For smooth, well-behaved responses, factorial designs provide a means of fitting polynomial surfaces (e.g., linear for two-level designs, quadratic for three-level designs) to the data to guide further exploration in the predictor space. Because they were developed in the tradition of modeling physical experiments, predictors in these designs are only set to one of a few levels in each run; this allows the estimation of predictor effects (i.e., through an ANOVA decomposition) and the magnitude of the random variability present in the system.

In this case, the goal is to fit a metamodel to the output of deterministic simulation code. That is to say, the variability in the system is zero. There is less of a need to sample predictors at one of a small set of values from run to run, since estimating variability is no longer required. Furthermore, it is possible that the simulation surface is not smooth and well-behaved. There could be local discontinuities present that cannot be easily observed from a factorial design that only examines behavior at the low, center, and high end of the ranges for each predictor.

An alternative approach is a sampling design, which has runs that are not restricted to low, center, and high values of each predictor. Instead, the samples are randomly chosen across the ranges of values for each predictor. Generally, the goal is to spread observations across the predictor space with as few “holes” or “gaps” as possible.

2.2.1 Purely Random Design

The most basic sampling design is a simple random sample over the input space. Observations are chosen by drawing independent random samples of size n over the range of possible values for each input. The result is a design with n runs. Variations on this approach could use different marginal distributions in the sampling of the inputs, or possibly include draws from a joint distribution over subsets of inputs. Random designs are easy and straightforward to produce. However, they could also suffer from poor “space-filling” characteristics. That is, multiple observations frequently end up clustered in one part of the space and provide largely redundant information about the behavior of the response surface in that region. Other parts of the space may be sparsely populated, and the redundant observations could be put to better use filling in those gaps.

2.2.2 Latin Hypercube Sampling

A Latin hypercube sample (LHS) design described by McKay *et al*. (1997) [10] is intended to fill the predictor space by randomly selecting observations in equal probability bins across the range of the inputs. These designs sample values in [0, 1] for each of the inputs at each design point.

The sampling is done in such a way that for a sample of size *n*, there will be exactly one observation in each of the intervals [0, 1/*n*), [1/*n*, 2/*n*), …, [(*n*-1)/*n*, 1] for each of the inputs.

In practice, the [0, 1] bounds on the values in LHS samples are interpreted to be a probability, and the design points are transformed through some probability distribution on the inputs. This has the effect of spreading the sampled points across equal regions of probability for each input, according to the chosen distribution. Several examples of LHS designs are shown in Figure 6 for two predictors.

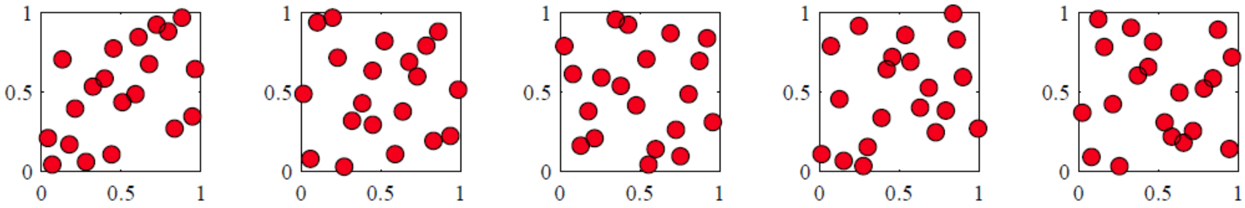


Figure 6. Examples of LHS designs using 20 observations for two predictors.

2.2.3 Maximin LHS

A maximin LHS design described by Johnson *et al*. (1990) [11] is created by generating a large number (e.g., thousands) of LHS designs and selecting the design that has the largest value of the function

$$M(\mathbf{x}^1, \mathbf{x}^2, \dots, \mathbf{x}^n) = min_{i,j} \left\|\mathbf{x}^i - \mathbf{x}^j\right\|,$$

where $\mathbf{x}^1, \mathbf{x}^2, \dots, \mathbf{x}^n$ are the *n* sampled observations and $\left\|\mathbf{x}^i - \mathbf{x}^j\right\|$ is the Euclidean distance between observations *i* and *j*. In other words, the maximin LHS design is the one that maximizes the minimum distance between any pair of observations in the sample. Examples of maximin LHS designs are shown in Figure 7.

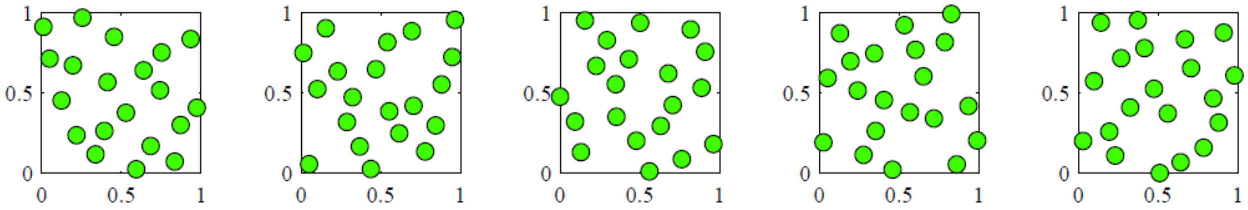


Figure 7. Examples of maximin LHS designs using 20 observations for two predictors.

Maximizing the minimum distance between any pair of points has the effect of spreading the observations out as much as possible across the input space, under the constraint that the design is still based on a Latin hypercube. Maximin LHS designs, therefore, tend to have better space-filling characteristics. With a generic LHS design, there is a rare chance that, for example, all of the runs could be drawn from bins along the diagonal of the hypercube. This would result in a poor design for response surface modeling. Since maximin designs are selected from hundreds or thousands of candidate models, the chance of such a diagonal model is infinitesimally small. In general, for any location in the input space, the distance to the closest observation will be on average less in a maximin LHS design than in a generic LHS design.

2.2.4 Maximum Entropy

Maximum entropy designs described by Shewry and Wynn (1987) [12] are also designed to have space-filling characteristics. The design is chosen to maximize the amount of "information" given by the sample, which in this case is captured by the entropy measure as defined in Shannon's information theory [13]. One way to do this is to maximize the determinant of the correlation matrix $\mathbf{C} = (r[i,j])$, where

$$r[i,j] = \begin{cases} 1 - \Gamma(h_{ij}) & if\ h_{ij} \leq a \\ 0 & if\ h_{ij} > a \end{cases}.$$

Here, h_{ij} is the distance between two observations x^i and x^j and $\Gamma\left(h_{ij}\right)$ is a spherical variogram with range a, defined by

$$\Gamma(h) = \frac{3h}{2a} - \frac{1}{2}\left(\frac{h}{a}\right)^3.$$

Maximum entropy designs are not restricted to equal probability bins, as LHS designs are. Several examples of these designs are shown in Figure 8.

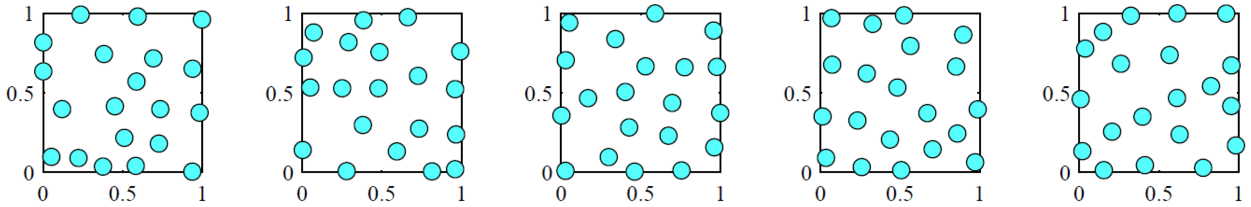


Figure 8. Examples of maximum entropy designs using 20 observations for two predictors.

2.2.5 Design Comparison

The figures below show comparisons of the various types of sampling designs with respect to several space-filling criteria. The wrap-around L_2 discrepancy described by Hickernell (1998)

[14], *WL2*, measures the difference between the number of design points per sub-volume compared to the same count for a uniform distribution of points across the input space. It is computed with the formula shown below, where *p* is the number of inputs, and $\mathbf{x}^1, \mathbf{x}^2, \dots, \mathbf{x}^n$ are the *n* observations (i.e., design runs).

$$WL2 = -\left(\frac{4}{3}\right)^p + \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\prod_{k=1}^{p}\left(\frac{3}{2} - \left|x_i^k - x_j^k\right|\left(1 - \left|x_i^k - x_j^k\right|\right)\right)$$

The second criterion is the maximin criterion described in Section 2.2.3:

$$M = min_{i,j}\left\|\mathbf{x}^i - \mathbf{x}^j\right\|$$

The final criterion is the entropy measure, defined as $E = \det(\mathbf{C})$, where the matrix $\mathbf{C} = (r[i,j])$ as described in Section 2.2.4.

To compare the space-filling characteristics of each of the sampling designs, 100 designs of each type were sampled over *n* = 20 runs and *d* = 2 inputs. Each of the three criteria were then computed for each design. Comparisons of the designs are shown in Figure 9, Figure 10, and Figure 11, corresponding to wrap-around L_2 discrepancy, maximin, and entropy measures, respectively. Maximum entropy is the top performer for two of the metrics, including the maximin measure. It is able to outperform the maximin LHS design in the latter case because it is not bound by the restriction to be a Latin hypercube design. In terms of wrap-around L_2 discrepancy, maximin LHS is the clear winner.

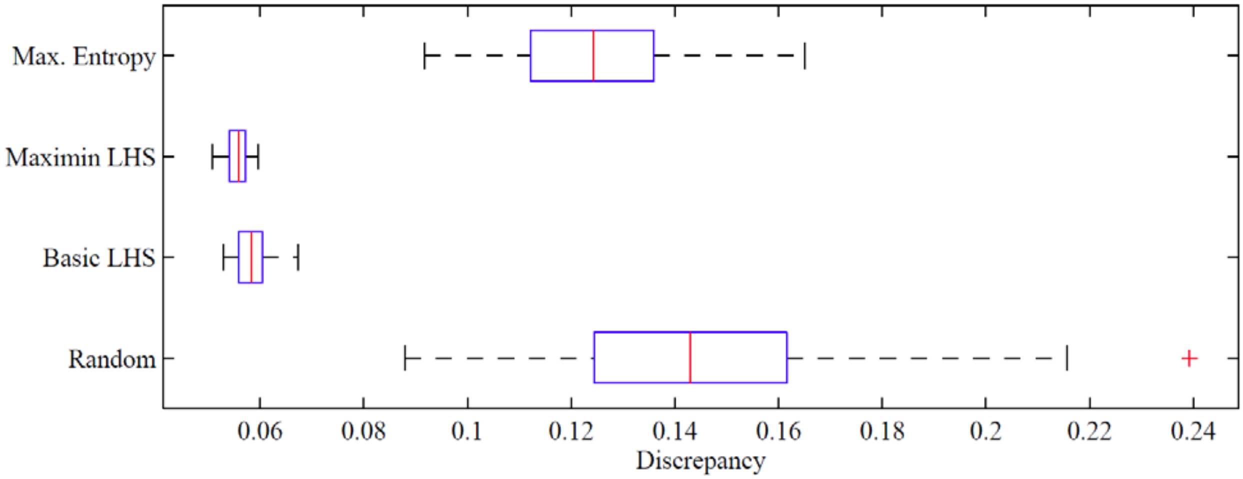


Figure 9. Comparison of the sampling designs with respect to the wrap-around L_2 discrepancy measure.

Smaller values indicate better space-filling characteristics.

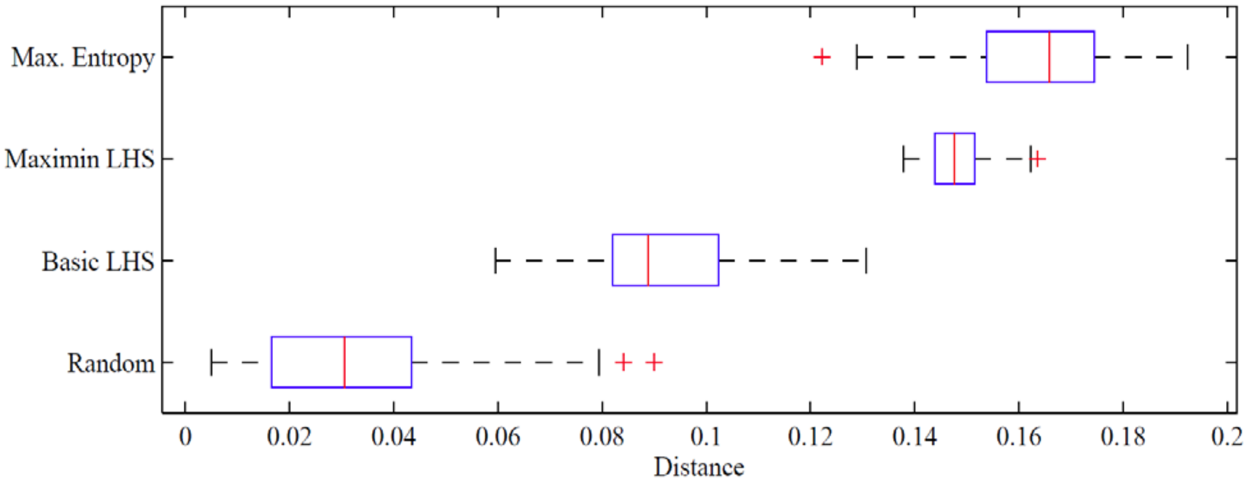


Figure 10. Comparison of the sampling designs with respect to the maximin distance measure.

Larger values indicate better space-filling characteristics.

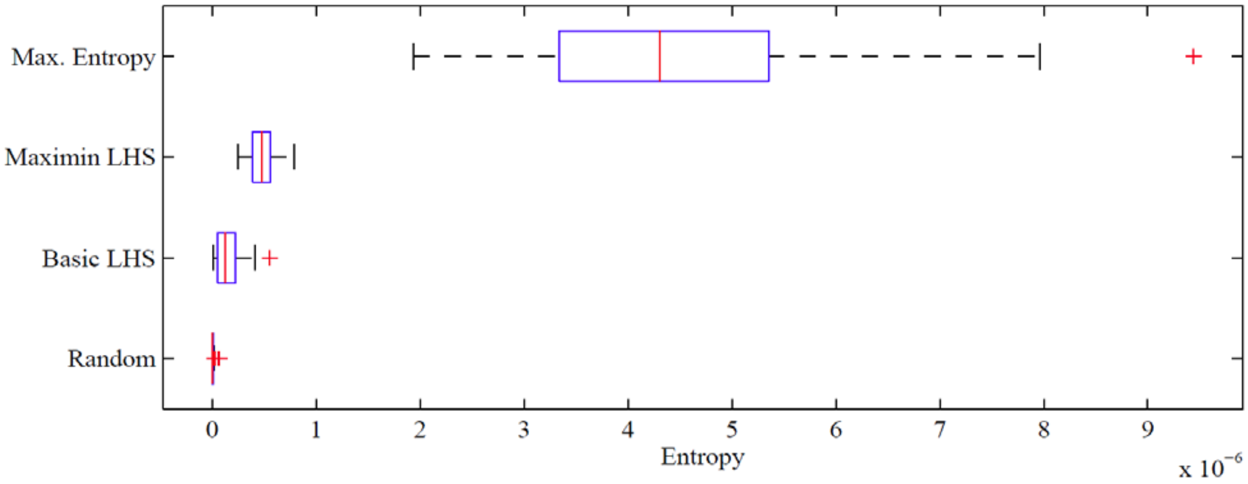


Figure 11. Comparison of the sampling designs with respect to the entropy measure.

Larger values indicate better space-filling characteristics.

3 Metamodeling

3.1 Introduction

After deciding on an experimental design, the experiment can be run at each of the prescribed predictor settings, and the responses can be observed. Using the design and the observed response, a response surface model may be used to predict what the response would have been at an unobserved combination of predictor values. In the context of computer experiments, where the response being modeled is the result of deterministic computer code, the response surface model is also referred to as a proxy model or a metamodel. Both terms capture the fact that one is using a model (i.e., the metamodel) to predict the output of another model (i.e., the deterministic computer code).

There are many variations of metamodels, but the goal is generally the same for all of them. Some assumptions are made about either the shape of the response surface, its smoothness, and/or the correlation in responses between points that are close in the space. The parameters for these assumptions are estimated with the sampled observations, and a criterion is optimized. Typically, that criterion balances the smoothness and simplicity of the surface with its ability to match available data.

3.2 Quadratic Model

The quadratic polynomial model fits a parametric model to the response that is the analogue of the parabola in p dimensions. It is defined as a sum of all linear, quadratic, and pair-wise cross-product terms between predictors. That is, the approximating function $\hat{f}(\mathbf{x})$ is defined by:

$$\hat{f}(\mathbf{x}) = \hat{y} = b_0 + \sum_{i=1}^{p} b_i x_i + \sum_{i=1}^{p} b_{ii}(x_i)^2 + \sum_{i=1}^{p} \sum_{j>i} b_{ij} x_i x_j$$

The coefficients in the quadratic polynomial model are estimated by solving the linear model **Y** = **XB**, where

$$\mathbf{Y} = \begin{pmatrix} f(\mathbf{x}^1) \\ f(\mathbf{x}^2) \\ \vdots \\ f(\mathbf{x}^n) \end{pmatrix}, \mathbf{X} = \begin{pmatrix} 1 & x_1^1 & \dots & x_p^1 & (x_1^1)^2 & \dots & \left(x_p^1\right)^2 & x_1^1 x_2^1 & x_1^1 x_3^1 & \dots & x_{p-1}^1 x_p^1 \\ 1 & x_1^2 & \dots & x_p^2 & (x_1^2)^2 & \dots & \left(x_p^2\right)^2 & x_1^2 x_2^2 & x_1^2 x_3^2 & \dots & x_{p-1}^2 x_p^2 \\ \vdots & \vdots & & \vdots & \vdots & & \vdots & \vdots & \vdots & & \vdots \\ 1 & x_1^n & \dots & x_p^n & (x_1^n)^2 & \dots & \left(x_p^n\right)^2 & x_1^n x_2^n & x_1^n x_3^n & \dots & x_{p-1}^n x_p^n \end{pmatrix},$$

and

$$\mathbf{B} = \left(b_0, b_1, \dots, b_p, b_{11}, \dots, b_{pp}, b_{12}, b_{13}, \dots, b_{p-1,p}\right)^T.$$

The solution is given by $\widehat{\mathbf{B}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$. This is a special case of multivariate linear regression.

3.3 Quadratic Model with LASSO Variable Selection

Typically, in industry, the analyst will perform a variable selection technique before proceeding with a quadratic fit. This could be done, for example, using exploratory analysis, stepwise regression, or comparison of candidate models using information criteria like AIC or BIC. Ultimately, the final model fit will only use a subset of the main effects, interactions, and squared effects. This results in a parsimonious model and can often lead to better predictions because noisy, less relevant covariates have been removed from consideration.

One way of performing variable selection is through an automatic procedure based on LASSO regression. LASSO (Least Absolute Shrinkage and Selection Operator) regression described by Tibshirani (1966) [15] is a technique for fitting a basic multiple linear regression model while shrinking the coefficients toward zero. Mathematically, this is done by adding a penalty term to the least squares term in the objective function for linear regression. LASSO regression has the interesting property that some of the fitted coefficients will be exactly zero. In these cases, LASSO serves as a variable selection algorithm where variables whose coefficients are zero are removed from the model.

The full procedure for the LASSO variable selection and quadratic fit is as follows:

1. Determine an appropriate value of λ using 10-fold cross-validation on the root mean squared error (RMSE) of the regression fit.

2. Fit a LASSO model using the quadratic regression model:

$$\hat{f}(\mathbf{x}) = \hat{y} = b_0 + \sum_{i=1}^{p} b_i x_i + \sum_{i=1}^{p} b_{ii}(x_i)^2 + \sum_{i=1}^{p}\sum_{j>i} b_{ij} x_i x_j$$

3. Identify which coefficients (b0, bi, bij, bii) are non-zero in the LASSO model. Remove all main effects, interactions, and squared terms that are associated with the zero coefficients.

4. Refit an OLS regression model using only the remaining terms from the LASSO model.

3.4 Kriging Model

The kriging model described by Simpson *et al*. (1998), Cressie (1993), and Krige (1951) [16-18] has an approximation function that is composed of a trend term and an autocorrelation term. That is,

$$\hat{f}(\mathbf{x}) = \mu(\mathbf{x}) + Z(\mathbf{x}),$$

where $\mu(\mathbf{x})$ is the overall trend and $Z(\mathbf{x})$ is the autocorrelation term. $Z(\mathbf{x})$ is treated as the realization of a mean zero stochastic process with a covariance structure given by $Cov\big(Z(\mathbf{x})\big) = \sigma^2\mathbf{R}$, where $\mathbf{R}$ is an *n*×*n* matrix whose $(i, j)^{\text{th}}$ element is the correlation function $R\big(\mathbf{x}^i, \mathbf{x}^j\big)$ between any two of the sampled observations $\mathbf{x}^i$ and $\mathbf{x}^j$. *Ordinary kriging* assumes a scalar trend $(\mathbf{x}) = \mu_0$, whereas *universal kriging* uses a parametric trend term.

In this study, the Matérn(5/2, θ) correlation below is used, where, $d_k = \big(x_k^i - x_k^j\big)$. The Matérn correlation is often favored for kriging models because it tends to produce estimates that are smoother on a local level than other common alternatives structures, like the exponential. However, it is also more flexible than Gaussian correlation, which can be overly smooth.

$$R\big(\mathbf{x}^i, \mathbf{x}^j\big) = \prod_{k=1}^{p}\left[1 + \frac{d_k\sqrt{5}}{\theta_k} + \frac{5d_k^2}{\theta_k^2}\right] exp\left(-\frac{d_k\sqrt{5}}{\theta_k}\right)$$

In the universal kriging model, the quadratic polynomial trend term below was used.

$$\mu(\mathbf{x}) = b_0 + \sum_{i=1}^{p} b_i x_i + \sum_{i=1}^{p} b_{ii}(x_i)^2 + \sum_{i=1}^{p}\sum_{j>i} b_{ij} x_i x_j$$

3.5 Multivariate Adaptive Regression Splines (MARS)

MARS models described by Friedman (1991) [19] approximate the response surface using a collection of simple step and hinge functions (described below). Each function is only defined over a particular region of the input space, and all of the functions collectively form a single piecewise function over the full input space. The model is defined by:

$$\hat{f}(\mathbf{x}) = \sum_{i=1}^{k} c_i B_i(\mathbf{x}),$$

where each B_i is a basis function that is constant, a hinge function, or a product of two or more hinge functions. Hinge functions are flat at zero over a portion of the space and linear elsewhere. In the case of a single variable x, a hinge function takes the form $\max(0, x - q)$ or $\max(0, q - x)$ for a constant q. Here, q is the location of the hinge, also called a knot.

3.6 Additivity and Variance Stabilization (AVAS)

The AVAS model described by Breiman and Friedman (1985) and Tibshirani (1988) [20, 21] uses a non-parametric, iterative procedure to find some transformation of the response that can be represented as a sum of transformed predictors. That is, it finds functions g_0, g_1, …, g_p such that:

$$g_0\big(f(\mathbf{x})\big) = \sum_{i=1}^{p} g_i(\mathbf{x}_i)$$

3.7 Thin Plate Splines (TPS)

TPS, as described by Duchon (1977), [22] are a generalization of splines in multiple dimensions. The name refers to the modeling of the response using a surface analogous to a thin semi-rigid sheet of metal. This surface can be deformed to fit the response, but at the expense of a penalty applied for non-smoothness.

3.8 Support Vector Regression

Support vector regression (SVR) is closely related to support vector machines (SVMs), which are widely used in classification tasks. In the case of SVR as described by Drucker (1977) [23], the p-dimensional input vectors $\mathbf{x}$ are represented by a d-dimensional set of features $\mathbf{z}$, where each element of the vector $\mathbf{z}$ is some function of the values in $\mathbf{x}$. For example, $\mathbf{z}$ could contain each of the terms in a polynomial formulation of the inputs. Given the features **z**, the support vector regression model is given by

$$\hat{f}(\mathbf{x}) = \mathbf{z}'\mathbf{w},$$

where **w** is a vector of real-valued linear model parameters. The estimation of the parameter vector **w** is done with two goals in mind: (i) allow errors no larger than ε when comparing actual and predicted responses, and (ii) make the function as “flat” as possible. These requirements are formally stated as:

$$\text{minimize} \qquad \tfrac{1}{2}\|\mathbf{w}\|^2 + C\textstyle\sum_{i=1}^{l}(\xi_i + \xi_i^*)$$

$$\text{subject to} \qquad \begin{cases} f(\mathbf{x}) - \hat{f}(\mathbf{x}) \le \varepsilon + \xi_i \\ \hat{f}(\mathbf{x}) - f(\mathbf{x}) \le \varepsilon + \xi_i^* \\ \xi_i, \xi_i^* \ge 0 \end{cases}$$

Here, ξ_i and ξ_i^* are “slack” variables that, together with the tuning parameter $C > 0$ specify the tradeoff between the flatness and accuracy of the predictive function.

The model is fit using a quadratic programming approach that depends only on knowing the dot product between observations. For this reason, the “kernel trick” can be used to specify non-linear SVR models, in which case the feature vector **z** and model parameters **we** are not explicitly given.

3.9 Radial Basis Functions (RBF)

Radial basis functions described by Chen *et al.* (1991) [24] are any functions that depend solely on the distance of an observation to some fixed location **c**. That is, an RBF $\phi(\cdot)$ satisfies$\phi(\mathbf{x}) = \phi(\|\mathbf{x} - \mathbf{c}\|)$. An RBF regression model takes the following form.

$$\hat{f}(\mathbf{x}) = b_0 + \sum_{i=1}^{p} b_i \phi_i(\|\mathbf{x} - \mathbf{x}_i\|)$$

That is, the response surface is approximated by a weighted sum of radial basis functions, each of which depends on the distance from the location of interest, **x**, and one of the sampled observations, $\mathbf{x}_i$. The regression weights b_i are then trained using an ordinary least squares approach. Other variations on this theme may be used to improve model fit. One way to provide a smoother fit is to include a smaller number of basis functions that involve alternative centers $\mathbf{c}_1$, $\mathbf{c}_2$, …, $\mathbf{c}_{p'}$ instead of $\mathbf{x}_1$, $\mathbf{x}_2$, …, $\mathbf{x}_p$, where *p'* << *p*. Another alternative is to allow the parameters of the $\phi_i(\cdot)$ functions to vary by location.

3.10 Projection Pursuit Regression (PPR)

PPR is an iterative procedure that estimates the response surface in a nonparametric fashion using linear combinations of the predictors.

$$\hat{f}(\mathbf{x}) = \sum_{k=1}^{K} g_k(\mathbf{z}_k) = \sum_{k=1}^{K} g_k(\boldsymbol{\alpha}'_k\mathbf{x}) = \sum_{k=1}^{K} g_k\left(\sum_{i=1}^{n} \alpha_{ki}x_i\right)$$

Here, the weights $\boldsymbol{\alpha}_k$ define a projection of the variables in observation $\mathbf{x}$ to a new variable $\mathbf{z}_k$. The functions g_k are smooth univariate functions (e.g., a linear approximation or spline fit). To fit this model, the response is centered and the residuals are initialized to the response values. At each iteration in the model fitting, the projection $\boldsymbol{\alpha}_k$ is chosen to maximize the amount of variability in the residuals r_1, r_2, …, r_n that can be explained. That is, $\boldsymbol{\alpha}_k$ is chosen as $\boldsymbol{\alpha}_k = \arg max_{\boldsymbol{\alpha}}\ I(\boldsymbol{\alpha})$, where

$$I(\boldsymbol{\alpha}) = 1 - \frac{\sum_{i=1}^{n}\left(r_i - g(\boldsymbol{\alpha}'\mathbf{x})\right)}{\sum_{i=1}^{n} r_i^2}.$$

In this sense, projections are chosen in pursuit of explaining the variability in the residuals. After selecting the projection $\boldsymbol{\alpha}_k$, the residuals are updated for the next iteration. The model fitting process ends when residuals are sufficiently small to trigger a stopping threshold.

4 Metamodel Evaluation

4.1 Performance Evaluation Metrics

The most desirable property of a metamodel is that it will provide the closest match between the prediction and the truth for future independent test data. When comparing different metamodels, it is useful to be able to capture the quality of the metamodel fit in a single statistic. There are many ways to do this, but two of the most common are root mean squared error (RMSE) and R^2. RMSE is defined as the square root of the average squared difference between predictions $\hat{y}_i = \hat{f}(\mathbf{x}^i)$ and true response values $y_i = f(\mathbf{x}^i)$ for a set of observations$\{\mathbf{x}^1, \mathbf{x}^2, \dots, \mathbf{x}^n\}$.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

The RMSE can also be normalized by, for example, dividing it by the median observed response. This puts it on a similar scale regardless of the response, allowing for comparison of metamodel fits to different response surfaces.

$$\text{"Scaled" } RMSE = SRMSE = \frac{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}}{median\{y_1, y_2, \dots, y_n\}}$$

Another metamodel accuracy measure is R^2, which is defined as the amount of variation in the response that is explained by the predictors. In a simple linear regression model, the R^2 statistic is the square of the correlation between the actual and predicted response values. For other models, a pseudo-R^2 statistic is typically used.

$$\text{Pseudo-}R^2 = R_p^2 = 1 - \frac{SS_{model}}{SS_{error}} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

Note that while R^2 in simple linear regression is always in [0, 1], the pseudo-R^2 is in [-∞, 1]. A negative pseudo-R^2 statistic means the model predicts the response worse than a flat model that predicts the mean observed response value everywhere in the predictor space.

4.2 Independent Validation

The most straightforward way to quantify the accuracy of a metamodel is to collect new, independent test data that were not involved in training the model. The metamodel is then used to produce predictions of the response at those test locations, and an RMSE or pseudo-R^2 statistic can capture the quality of the fit. If these statistics have favorable values, then the model can be considered accurate over the region of the predictor space "covered" by the test observations. For this reason, it is important to select independent test data that span the region of the predictor space of interest to the investigator, and that they are selected in a sufficient density to capture variation in the response surface at a resolution fine enough to meet the needs of the study.

4.3 Cross-Validation

In many cases, independent test data are not available at the time of model fitting. In this case, one must use the training data to measure model accuracy. One approach is to fit a metamodel using all available training data, then calculate model accuracy statistics based on those training observations. In this case, the statistics will be biased optimistically, since the metamodel first and foremost is designed to fit those particular observations well. An overtrained model will fit the training data very well, but perform poorly on independent test data.

4.3.1 k-Fold

A better approach is to use a *k*-fold cross-validation approach (see, e.g., Hastie *et al*. (2008) [25], Chapter 7). Under this paradigm, the dataset is randomly partitioned into *k* folds, which are mutually exclusive and exhaustive subsets of the observations. Each fold is then systematically held out and the metamodel is fit to a dataset consisting of only the remaining $k - 1$ folds. This model is then used to make a prediction on the fold that was left out. After repeating this process on all *k* folds, there are a total of *k* models that are constructed, each of which are used to predict the value of the single fold that was left out of the training set.

While the cross-validation approach does not specifically test the unique model that is created by using all *n* training observations together, it does test the algorithm that is used to construct the model. When each fold is held out of the training set, it will behave like independent test data as far as that particular model is concerned. Therefore, the error magnitudes from the cross-validation more accurately reflect error rates in the model fit over parts of the response surface that have not been sampled.

In general, selecting the number of folds *k* is a balance between the bias in estimating the model accuracy and the variance in model fits. If *k* is small (i.e., each fold contains many observations), the model fits will be made using a smaller subset of the dataset, with less overlap in training data between models. This can result in large variations in the models as successive folds are held out, and thus create highly variable estimates of model accuracy from fold to fold. If *k* is large (i.e. each fold contains only a few observations), the model fits will be less variable, since most of the training data will be in common between fits. However, estimation of model accuracy will tend to be more optimistically biased, since all of the models will begin to look similar to the full model trained using all n observations. The rule of thumb is to use somewhere between $k = 5$ and $k = 10$ (see discussion in Hastie *et al*. (2008) [25], Ch.7).

4.3.2 Leave-One-Out

Leave-one-out cross-validation is the extreme case of *k*-fold cross-validation where $k = n$. That is, each fold is a single observation. In this case, each cross-validation model is fit with $n – 1$ observations and then used to predict the response at the single observation that was left out of the training data. Although this is a popular approach to measuring model accuracy, it can lead to optimistic estimates of model quality due to the low variance of model fits from fold to fold. It can also come with a high computational cost, as $n – 1$ different models need to be trained.

4.4 Variable Importance

Often, it is of interest to identify which of the inputs are the drivers of the model response. This can aid in narrowing down the set of input values that produces an optimum response, or can allow for a more parsimonious explanation of the interactions between the inputs and the response. Determination of a variable’s importance can be approached in a number of ways.

For example, in classical experimental design, the coefficients of each input in an ANOVA model, along with the associated standard error and significance, can be used to rank the variables and remove the insignificant ones from the model. Such a procedure has even been automated in stepwise model-fitting approaches. Some other models have their own specific methods for determining variable importance, as well.

One more general method for assigning variable importance is to compute the R^2 loss. The reasoning behind this approach is that removing important inputs from the model will result in models that have much worse predictive ability. In contrast, removing inputs that are not important should have a minimal effect on the quality of the model fit. The pseudo-R^2 value (see Section 4.1) is one way of capturing how well the inputs are able to explain the variability in the response. Therefore, an input's importance can be measured by how large a reduction in the pseudo-R^2 occurs as a result of removing that input from the model. A summary of the procedure is given below.

1. Fit a model of the response using all of the inputs. Compute the pseudo-R^2 for this full model, and call it R^2_{full}.

2. Fit a reduced model of the response using all inputs except input k. Compute the pseudo-R^2 for the reduced model, and call it R^2_{-k}.

3. Define the importance of input k to be $I_k = R^2_{full} - R^2_{-k}$

4.5 Case Study – Arches Metamodeling

During this year, a preliminary study for Task 3 was conducted using an existing dataset from the Arches province in the American Midwest, which is described by Mishra *et al*. (2014) [26]. In that study, single-well simulations of CO_2 injection into a closed volume (as would be the case in a network of wells employed for regional scale CO_2 storage) were carried out using the STOMP-CO_2 simulator described in White *et al*. (2012) [27]. Stratigraphic columns corresponding to three different ratios of reservoir (Mount Simon sandstone, MS) and caprock (Eau Claire shale, EC) thickness were considered, with different depths to injection zone in each case.

For each synthetic site case, simulations with 4 different well patterns (3×3, 4×4, 5×5, and 6×6 well arrays), and 3 different permeability group variations (High *k*, Medium *k*, and Low *k*) were run. Each permeability group consists of a set of correlated variables: permeability of MS, permeability of EC, and the capillary entry pressure. This brings the total number of simulations to 3×4×3 = 36. Simulations of pressure-constrained injection (at 90% of the fracture pressure) were carried out using a 2-D *r-z* model with 20 vertical rows and 100 radial columns. The predictors and response variables extracted from the simulations are described in Table 1.

Since the Arches dataset contains only 36 observations and 3 predictors, there is clearly the potential for highly variable model fits if k is too small. For this study, k was chosen to be 12, which places 3 observations within each fold. Since model fits are being made with 33 observations, they are unlikely to vary wildly from fold to fold. Also, with such a small sample size, bias is unlikely to be a problem since all of the observations carry a lot of weight in the model fitting process. This guarantees that a model created with one fold removed will be different than the model trained using all of the data.

Table 1. Predictors and Responses in the Arches Dataset

Predictor	Description
D	Depth to injection (m), which affects the fracture pressure, and hence, the maximum pressure differential under which injection can be carried out
L	Well spacing (m), which determines the volume of the closed system into which CO_2 is injected
kh_MS	Permeability-thickness product (md-ft) for the injection reservoir (Mount Simon sandstone), which controls the amount of CO_2 that can be injected for a given pressure differential

Response	Description
Cum_CO2	Cumulative volume of CO_2 injected (millions of metric tons, MMT)
CO2_R	Radius of CO_2 plume (m)
PCT_CO2	% Mass flux entering the caprock

The 12-fold cross-validation procedure was repeated 100 times for each metamodel, with fold memberships being randomly selected each time. For each response, this produced 100 cross-validated predictions at every set of sampled predictor values. The overall measure of metamodel accuracy was the RMSE. In this case, let $\hat{f}_j(\mathbf{x}^i)$ be the prediction of the response in the jth cross-validation replicate for the ith observation. In similar fashion, let $f_j\left(\mathbf{x}^i\right)$ be the true response. Then define the metamodel accuracy for that response to be the following.

$$RMSE = \sqrt{\frac{1}{36}\sum_{i=1}^{36}\frac{1}{100}\sum_{j=1}^{100}\left(f_j(\mathbf{x}^i) - \hat{f}_j(\mathbf{x}^i)\right)^2}$$

Table 2 shows a comparison of model accuracy across all metamodels and responses. Note that the raw RMSE values have different magnitudes from response to response. This is due to different original scales for the observed responses. To allow better comparison of the RMSE values, they may be scaled by the average response across the n = 36 observations, which is denoted here as SRMSE. Both the RMSE and the SRMSE are given in Table 2 for each combination of metamodel and response.

Table 2. Metamodel Performance, 12-Fold Cross-Validation

Model Name	RMSE (SRMSE)		
	CUM_C02	CO2_R	PCT_CO2
Quadratic	2.376 (-0.108)	900.504 (-0.095)	0.466 (-0.216)
Ordinary Kriging	0.640 (-0.029)	784.904 (-0.083)	0.087 (-0.04)
Universal Kriging	0.766 (-0.035)	536.496 (-0.057)	0.080 (-0.037)
MARS	9.811 (-0.445)	1322.644 (-0.14)	1.092 (-0.507)
AVAS	6.996 (-0.317)	1393.492 (-0.147)	0.443 (-0.206)
TPS	3.924 (-0.178)	1035.832 (-0.109)	0.694 (-0.322)

For all three responses, the kriging models outperform the other four types of metamodels. In particular, the universal kriging model with a quadratic trend seems best overall. This can be seen, for example, in Figure 12, which shows plots of cross-validated metamodel performance on one of the responses, CO2_R. Note that the universal kriging model provides relatively stable results across the range of values for this response. The residuals appear to be largely uncorrelated as well, which indicates that there are not large systematic components to the response that are not being accounted for.

The entire cross-validation procedure was also repeated using 6 folds and 100 replicate runs. In this case, there are 6 observations in each fold, and the successive metamodel fits are made using fewer observations (30 instead of the 33 in the 12-fold case). Using a smaller number of folds should have the effect of reducing overfitting effects on the predictions and increasing model variability from run to run. Results for the 6-fold cross-validation are shown in Table 3. The kriging models are still the best performers for all responses, although now the universal kriging model is the top performer for the first response as well.

Table 3. Metamodel Performance, 6-Fold Cross-Validation

Model Name	RMSE (SRMSE)		
	CUM_C02	CO2_R	PCT_CO2
Quadratic	2.455 (-0.111)	933.624 (-0.098)	0.480 (-0.223)
Ordinary Kriging	1.305 (-0.059)	743.243 (-0.078)	0.137 (-0.064)
Universal Kriging	1.186 (-0.054)	611.898 (-0.065)	0.117 (-0.054)
MARS	10.061 (-0.456)	1574.248 (-0.166)	1.157 (-0.537)
AVAS	7.083 (-0.321)	1253.068 (-0.132)	0.476 (-0.221)
TPS	4.214 (-0.191)	1141.412 (-0.12)	0.763 (-0.354)

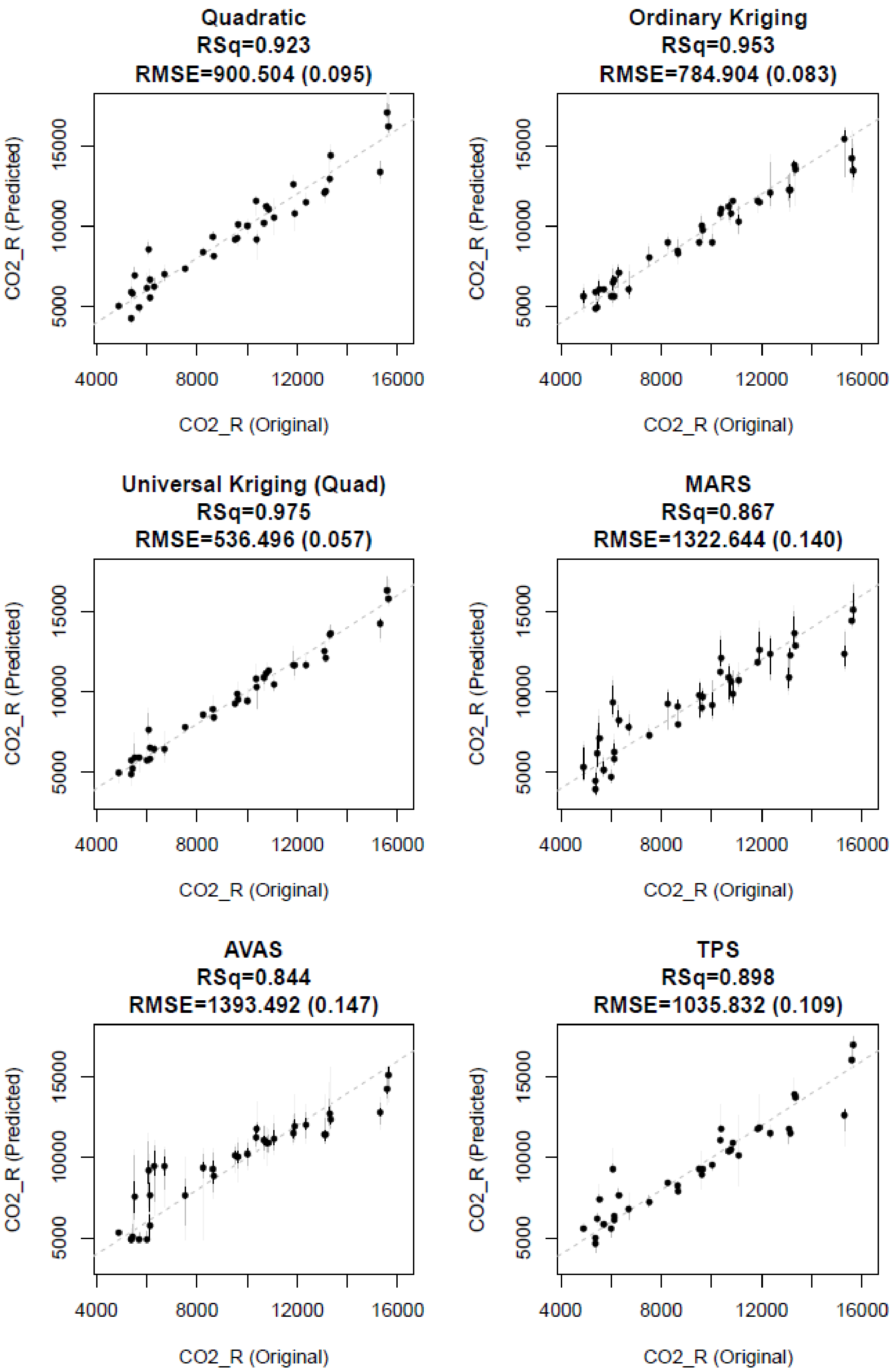


Figure 12. CO2_R metamodel performance in 12-fold cross-validation.

Circles represent the median prediction over the 100 replicate runs. Vertical lines indicate various percentile ranges over the 100 replicate runs: Black = 25th – 75th, Dark Grey = 5th – 95th, Light Grey = Min – Max.

In this study, the cross-validation performance of five different metamodeling approaches was measured in a closed volume injection case study of the Arches dataset. The dataset contained three CO_2 predictors, three responses, and 36 observations. Results showed that the kriging metamodels outperformed the others for all three responses – in particular, the universal kriging model with a quadratic trend had the best overall cross-validated RMSE. The quadratic polynomial model was second best, followed by TPS, AVAS, and MARS. These rankings held true for both the 6- and 12-fold cross-validation.

Clearly, a single case study with 36 observations is not sufficient to make a robust comparison between metamodels. Performance is dependent not only on the number of observations, but also on what sampling design was used and how many predictors there are. However, in this particular case, the kriging metamodel was better at uncovering the structure of the response surface (at least at the 36 observed sample points) than the other models.

5 Comparison of Metamodeling Approaches

5.1 Problem Description

For the full combined experimental design and metamodel comparison study, sampling methods and metamodels were developed for a reservoir simulator called GEM described in Computer Modeling Group (2014) [28]. A simulation run requires nine input parameters, and results in a host of responses over a 30 year period. Of these responses, three were chosen for the proxy model comparison. The first is the average pressure in the reservoir, the second is the radius of the CO_2 plume, and the third is the total storage efficiency of the reservoir. All responses were selected at the end of the 30 year period.

The designs and metamodels under consideration in the study are listed in Table 4. The selection of these options was made with two objectives in mind. The first goal was to provide a slice through a continuum of possible approaches that would likely be used by others in the field. The second goal was to allow for clear interpretation by keeping the number of models from being needlessly large. For analysts opting for an experimental design route, the Box-Behnken design seems to be the industry standard, more so than Central Composite, Plackett-Burman, and factorial designs. Augmented pairs was also selected as an alternative to the Box-Behnken that needs fewer runs; such a design may be useful in cases where design runs are expensive to obtain in terms of time or money.

Table 4. Designs and Metamodels (Size *n*) Used in the Study

Experimental Designs	Metamodels
Box-Behnken (n = 97)	Ordinary Kriging
Augmented Pairs (n = 79)	Universal Kriging
Maximum Entropy (n = 97)	Quadratic Polynomial
Maximin LHS (n = 97)	Quadratic Polynomial + LASSO Variable Selection
	MARS
	AVAS

For those opting for a sampling approach, the expected advantage of such designs over experimental designs would be their space-filling nature. LHS designs are popular choices in this area, and maximin LHS designs are the typical choice of space-filling LHS design. Occasionally, maximum entropy designs are used in literature, and they show better space-filling characteristics than maximin LHS designs according to two of the three space-filling metrics described in Section 2.2.5. For that reason, maximum entropy designs are considered in this study as well.

Each of the sampling designs contained $n = 97$ runs with different values for the nine input variables. The selection of 97 runs was made because the Box-Behnken design for $p = 9$ input variables has $n = 97$ unique observations. In physical experiments, the Box-Behnken design contains a number of duplicate points that are intended to be used to capture unknown sources of variability. However, since the GEM simulation is deterministic, there is no need to measure the response at a particular set of inputs more than once. This reduces the Box-Behnken design from 130 runs to 97. To avoid any bias that could be attributed to unequal sample sizes, all of the maximin LHS designs were restricted to the same number of runs as the Box-Behnken design. The exception to this rule was the augmented pairs design, which for 9 predictors is defined by $n = 79$ runs after removing duplicate observations.

The maximum entropy and maximin LHS designs were sampled over the 9-dimensional unit hypercube $[0, 1]^9$ and then converted back to the original predictor scale using the distributions shown in Table 5. Here, T(l, m, h) is a triangular distribution extending from l to h, with its peak at m.

Table 5. Input Distributions used with LHS Sampling

Input	Description	Distribution
$\mathbf{H_R}$	Thickness of the reservoir	T(50,150,250)
$\mathbf{H_{CR}}$	Thickness of the caprock	T(100,150,200)
$\boldsymbol{\mu_{LNKR}}$ $\mathbf{V_{DP}}$	Log-mean reservoir permeability, Dykstra-Parson's coefficient (perfectly correlated)	μ_{ln_KR} ~ T(2.45, 3.56, 4.67) VDP ~ T(0.35, 0.55, 0.75)
$\mathbf{K_{CR}}$	Average horizontal permeability of the caprock	lnT(0.002,0.02,0.2)
$\mathbf{K_V/K_H}$	Anisotropy ratio	lnT(0.01,0.1,1)
Q	CO_2 injection rate	discrete with equal probability – {0.33, 0.83, 1.33}
$\boldsymbol{\varphi_R}$	Porosity of the reservoir	T(0.08,0.12,0.18)
$\boldsymbol{\varphi_{CR}}$	Porosity of the caprock	T(0.05,0.07,0.10)
$\mathbf{I_V}$	Order of permeability layering	Discrete w/equal probability, $I_V \in$ {"random", " increasing", "decreasing"}
$\mathbf{P_{C,CR}}$	Capillary pressure model of caprock	Fixed value: Decrease P_C by 3x
$\mathbf{k_{r,R}}$	Relative permeability model of reservoir	Fixed value: Linear for k_{rw}

Each of the metamodels was trained on the dataset using the R statistical computing language described by R Development Core Team (2011) [29]. Specifically, kriging metamodels were trained using the *km* method of the **DiceKriging** package described by Roustant *et al*. (2011) [30], the quadratic polynomial fit was performed using the *lm* method in the **stats** package (included in base R), LASSO was performed using the **caret** package described by Kuhn *et al*. (2012) [31], MARS was performed using the *mars* method of the **mda** package described by Hastie *et al*. (2011) [32], and AVAS was performed using the *avas* method of the **acepack** package described by Spector *et al*. (2010) [33].

5.2 Model Fit Results

In this study, there were four candidate designs (Box-Behnken, augmented pairs, maximum entropy, and maximin Latin hypercube sampling (LHS)) and six metamodels (ordinary kriging, universal kriging, quadratic, LASSO + quadratic, MARS, and AVAS). This yields a total of 4 x 6 = 24 combinations of designs and metamodels. For each of these combinations, a metamodel was trained using the design inputs.

The model performance was measured using root mean squared error (RMSE), scaled root mean squared error (SRMSE), and pseudo-R^2, which are described in Section 4.1. Table 6 and

Table 7 show the SRMSE and pseudo-R^2 for each of the metamodels, respectively, measured over the training data. Here, “BB” stands for Box-Behnken, “AP” for augmented pairs, “ME” for maximum entropy, and “MM” for maximin LHS. Note that this is a biased view of model performance, since the models are being evaluated over the same dataset used to train them. For example, since kriging models are interpolators (i.e., they pass through each observation by design), they always achieve zero error over the training set. However, one obviously could not expect them to perfectly model the response at other points in the input space.

Table 6. Full Model Fit results (Scaled RMSE shown for each combination)

Total Storage Efficiency

	Ordinary Kriging	Universal Kriging	Quadratic Poly	Quadratic w/LASSO	MARS	AVAS
BB	0.000	0.000	0.063	0.070	0.103	0.097
AP	0.000	0.000	0.042	0.059	0.125	0.107
ME	0.000	0.000	0.054	0.063	0.090	0.080
MM	0.000	0.000	0.048	0.058	0.082	0.072

Plume Radius

	Ordinary Kriging	Universal Kriging	Quadratic Poly	Quadratic w/LASSO	MARS	AVAS

BB	0.000	0.000	0.046	0.047	0.123	0.068
AP	0.000	0.000	0.032	0.032	0.138	0.094
ME	0.000	0.000	0.033	0.033	0.068	0.044
MM	0.000	0.000	0.024	0.024	0.059	0.074

Average Pressure

	Ordinary Kriging	Universal Kriging	Quadratic Poly	Quadratic w/LASSO	MARS	AVAS
BB	0.000	0.000	0.027	0.037	0.107	0.033
AP	0.000	0.000	0.044	0.058	0.187	0.313
ME	0.000	0.000	0.017	0.028	0.097	0.025
MM	0.000	0.000	0.014	0.024	0.088	0.015

Table 7. Full Model Fit results (Pseudo-R^2 shown for each combination)

Total Storage Efficiency

	Ordinary Kriging	Universal Kriging	Quadratic Poly	Quadratic w/LASSO	MARS	AVAS
BB	1.000	1.000	0.926	0.910	0.802	0.825
AP	1.000	1.000	0.980	0.960	0.819	0.869
ME	1.000	1.000	0.921	0.892	0.786	0.830
MM	1.000	1.000	0.938	0.908	0.819	0.858

Plume Radius

	Ordinary Kriging	Universal Kriging	Quadratic Poly	Quadratic w/LASSO	MARS	AVAS
BB	1.000	1.000	0.977	0.976	0.834	0.949
AP	1.000	1.000	0.992	0.992	0.851	0.931
ME	1.000	1.000	0.984	0.984	0.933	0.972
MM	1.000	1.000	0.990	0.989	0.937	0.902

Average Pressure

	Ordinary Kriging	Universal Kriging	Quadratic Poly	Quadratic w/LASSO	MARS	AVAS
BB	1.000	1.000	0.964	0.932	0.435	0.947
AP	1.000	1.000	0.975	0.957	0.544	-0.283
ME	1.000	1.000	0.987	0.965	0.570	0.971
MM	1.000	1.000	0.990	0.970	0.579	0.988

5.3 Independent Validation Results

The validation study is a comparison of various combinations of designs and metamodels in terms of prediction performance on a common independent test design. This avoids the issue of bias in evaluating model performance, as discussed in the previous section. These models were then used to predict the value of the response on an independent design, which in this case was a basic Latin hypercube sample.

The results of the validation study are shown in Table 8 and Table 9. Cells shaded green indicate better performance, and those in red indicate worse performance. For the first response (Total Storage Efficiency), the Box-Behnken design was clearly the best performer, and the quadratic fit (with and without LASSO) seemed best. The LASSO quadratic fit seemed more robust across the designs than the other metamodels, giving consistently good performance. However, all of the first four model types (kriging and quadratic fits) seemed to have similar quality.

Table 8. Validation Study Results (Scaled RMSE shown for each combination)

Total Storage Efficiency

	Ordinary Kriging	Universal Kriging	Quadratic Poly	Quadratic w/LASSO	MARS	AVAS
BB	0.100	0.090	0.090	0.089	0.103	0.111
AP	0.100	0.110	0.110	0.106	0.109	0.107
ME	0.114	0.114	0.114	0.103	0.109	0.100
MM	0.110	0.118	0.118	0.093	0.109	0.113

Plume Radius

	Ordinary Kriging	Universal Kriging	Quadratic Poly	Quadratic w/LASSO	MARS	AVAS
BB	0.077	0.068	0.068	0.069	0.099	0.066
AP	0.081	0.087	0.087	0.087	0.117	0.093
ME	0.072	0.062	0.062	0.069	0.081	0.063
MM	0.064	0.065	0.065	0.064	0.089	0.099

Average Pressure

	Ordinary Kriging	Universal Kriging	Quadratic Poly	Quadratic w/LASSO	MARS	AVAS
BB	0.035	0.040	0.040	0.040	0.118	0.047
AP	0.076	0.108	0.108	0.104	0.151	0.410
ME	0.036	0.040	0.041	0.045	0.119	0.048
MM	0.044	0.042	0.042	0.060	0.134	0.050

Table 9. Validation Study Results (Pseudo-R^2 shown for each combination)

Total Storage Efficiency

	Ordinary Kriging	Universal Kriging	Quadratic Poly	Quadratic w/LASSO	MARS	AVAS
BB	0.765	0.808	0.808	0.811	0.748	0.706
AP	0.763	0.715	0.715	0.734	0.720	0.731
ME	0.693	0.693	0.693	0.746	0.720	0.763
MM	0.712	0.672	0.672	0.793	0.718	0.699

Plume Radius

	Ordinary Kriging	Universal Kriging	Quadratic Poly	Quadratic w/LASSO	MARS	AVAS
BB	0.921	0.938	0.938	0.937	0.870	0.943
AP	0.913	0.900	0.900	0.901	0.821	0.886
ME	0.932	0.950	0.950	0.936	0.913	0.947
MM	0.946	0.945	0.945	0.946	0.895	0.871

Average Pressure

	Ordinary Kriging	Universal Kriging	Quadratic Poly	Quadratic w/LASSO	MARS	AVAS
BB	0.952	0.940	0.940	0.939	0.464	0.916
AP	0.779	0.555	0.554	0.588	0.128	-5.428
ME	0.950	0.938	0.936	0.924	0.458	0.913
MM	0.925	0.933	0.932	0.862	0.318	0.903

For the second response (Plume Radius), the maximin LHS is the best performing design. As was the case in the first response, the first four model types perform similarly. This pattern continues for the third response (Average Pressure), where ordinary kriging is the best model. In terms of sampling designs, the Box-Behnken, maximum entropy, and maximin LHS all have similar performance when predicting Average Pressure. A summary of the findings from this study appear in Table 10.

A visualization of the model fit qualities may also be found in the Appendix. Each row of plots shows a comparison of the predicted response to the actual response on different datasets. The left plot shows the prediction of the metamodel on the same data used to train it. These results are expected to be optimistically biased. The center plot shows the prediction of the metamodel on the independent validation set, which was a Latin hypercube sample. The right plot contains cross-validated results, which are described in more detail in Section **5.4**.

Table 10. Summary of the Validation Study Findings

Response	Total Storage Efficiency	Plume Radius	Average Pressure
Best Design	Box-Behnken	Maximin LHS	Box-Behnken / Maximum Entropy
Best Metamodel	Quadratic w/LASSO	Tie Between All Kriging and Quadratic Variants	Ordinary Kriging

5.4 Cross-Validation Results

Another comparison of interest in this study was between cross-validation and validation. Had validation data not been available, would cross-validation have given similar results in terms of which metamodels had the best performance on each of the responses? To investigate this question, a 5-fold cross-validation procedure was implemented 100 times for each of the metamodels. For each response, this produced 100 cross-validated predictions at every set of sampled predictor inputs. The metamodels were compared using the average scaled RMSE (SMRSE) over the 100 sets of predictions. The SRMSE is given in the formula below, where y_{ij} is the prediction for the i^{th} response in the j^{th} cross-validation.

$$SRMSE = \frac{\sqrt{\frac{1}{n}\sum_{i=1}^{n}\frac{1}{100}\sum_{j=1}^{100}\left(y_{ij}-\hat{y}_{ij}\right)^2}}{median_i(y_i)} = \frac{\sqrt{\frac{1}{100n}\sum_{i=1}^{n}\sum_{j=1}^{100}\left(y_{ij}-\hat{y}_{ij}\right)^2}}{median_i(y_i)}$$

Results of the cross-validation runs are given in Table 11 and Table 12. There are several interesting things to note in comparing these results to the validation results. First of all, the cross-validation error rates seem higher than the validation error rates across the board. The effect is most prominent for the Box-Behnken and augmented pairs designs. This is likely because predictions by cross-validated models can only be made at sampled locations in the response surface. In the case of the BB and AP designs, the only samples points were on the boundaries of the predictor space, where models are not as likely to fit well, especially when those points are left out of the training process.

Due to the disproportionate bias in measuring error in the factorial designs, the maximum entropy and maximin LHS designs look much more favorable in the cross-validation study. Maximin LHS is the top performing design in all cases. The ordinary kriging and quadratic + LASSO metamodels seem best overall as well.

Table 11. 5-fold Cross-Validation Study Results (Scaled RMSE shown for each combination)

Total Storage Efficiency

	Ordinary Kriging	Universal Kriging	Quadratic Poly	Quadratic w/LASSO	MARS	AVAS
BB	0.119	0.181	0.182	0.170	0.119	0.119
AP	0.136	0.327	0.333	0.367	0.141	0.144
ME	0.115	0.177	0.175	0.125	0.117	0.117
MM	0.104	0.148	0.151	0.106	0.107	0.109

Plume Radius

	Ordinary Kriging	Universal Kriging	Quadratic Poly	Quadratic w/LASSO	MARS	AVAS
BB	0.100	0.130	0.128	0.216	0.134	0.104
AP	0.098	0.186	0.187	0.347	0.162	0.135
ME	0.092	0.111	0.110	0.071	0.087	0.086
MM	0.067	0.075	0.076	0.057	0.079	0.082

Average Pressure

	Ordinary Kriging	Universal Kriging	Quadratic Poly	Quadratic w/LASSO	MARS	AVAS
BB	0.054	0.078	0.078	0.107	0.118	0.052
AP	0.173	0.289	0.299	0.303	0.233	0.320
ME	0.040	0.055	0.056	0.039	0.114	0.033
MM	0.031	0.046	0.045	0.036	0.106	0.028

Table 12. 5-fold Cross-Validation Study Results (Pseudo-R^2 shown for each combination)

Total Storage Efficiency

	Ordinary Kriging	Universal Kriging	Quadratic Poly	Quadratic w/LASSO	MARS	AVAS
BB	0.760	0.640	0.632	0.625	0.749	0.761
AP	0.824	0.741	0.741	0.709	0.792	0.786
ME	0.691	0.462	0.468	0.678	0.681	0.684
MM	0.748	0.623	0.614	0.765	0.728	0.724

Plume Radius

	Ordinary Kriging	Universal Kriging	Quadratic Poly	Quadratic w/LASSO	MARS	AVAS
BB	0.906	0.888	0.887	0.721	0.808	0.907
AP	0.944	0.909	0.907	0.712	0.836	0.910
ME	0.896	0.882	0.883	0.938	0.904	0.923
MM	0.936	0.935	0.934	0.956	0.906	0.925

Average Pressure

	Ordinary Kriging	Universal Kriging	Quadratic Poly	Quadratic w/LASSO	MARS	AVAS
BB	0.874	0.827	0.826	0.709	0.347	0.904
AP	0.656	0.699	0.693	0.533	0.427	0.200
ME	0.940	0.913	0.909	0.946	0.535	0.971
MM	0.959	0.928	0.929	0.945	0.525	0.973

Table 13. Summary of the Cross-Validation Study Findings

Response	Total Storage Efficiency	Plume Radius	Average Pressure
Best Design	Maximin LHS	Maximin LHS	Maximin LHS
Best Metamodel	Ordinary Kriging	Quadratic w/LASSO	Ordinary Kriging

Scatterplots summarizing the cross-validation results can be found in the Appendix in the right-most column. Each plot compares the cross-validated predicted responses to the actual response. For each observation, the circle represents the median predicted response over the 100 cross-validation runs. The black line shows the range of the 25th to 75th percentile predictions, the dark gray represents the 5th to 95th percentiles, and the light gray shows the full range of predictions, minimum to maximum.

These figures give particular insights into why some models fare better than others. For example, the universal kriging and quadratic polynomial models tend to underperform compared to ordinary kriging and the quadratic w/LASSO model. The plots show that those models perform worse because they are more variable in the predicted response with respect to the choice of cross-validation grouping. That is, those models seem more sensitive to the data that are available, whereas the ordinary kriging and quadratic w/LASSO models predict similar responses regardless of how the training set is divided into cross-validation groups.

The fact that these results do not entirely agree with the validation results is somewhat surprising. Cross-validation is intended to approximate the error rates that would be obtained on an independent test set without having to actually collect that independent test data. However, cross-validated predictions can only be made at training data points. Since the designs under consideration in this study have fundamentally different paradigms (e.g., sampling over input ranges vs. assigning levels), it is difficult to compare model performance on the basis of cross-validation. A clear example of this is the augmented pairs models, which have the worst cross-validation results. This is at odds with the validation results, which showed that AP models were in the middling range of performance. The fact that the cross-validation metrics for AP models were so much worse could be related to the fact that the AP design had fewer runs to begin with, and those it did have were specifically chosen to not contain any redundant information. Holding some of those crucial observations, therefore, could remove too much information for the AP model to be able to adequately describe the surface.

5.5 Variable Importance

Using the procedure described in Section 4.4, the importance of the 9 GEM inputs was assessed for each combination of response, design, and metamodel. For a given combination, the metamodel was trained using the selected design and used to make predictions on the validation set described in Section 5.3. The pseudo-R^2 was recorded over the validation set using this full model. Next, each input was held out of the model training process to obtain a reduced model. This reduced model was also used to predict over the validation set, with the input removed from the test set as well. The pseudo-R^2 was then recorded for the reduced model. The difference between the two R^2 values is the variable importance measure for the input that was held out.

Results are shown in Figure 13, Figure 14, and Figure 15 for the total storage efficiency, plume radius, and average pressure responses, respectively. In each set of plots, the length of the bars represents the variable importance for the 9 predictors arranged in the rows. Different color bars correspond to the different designs, while plot columns correspond to different types of metamodels. One thing to note is that the pseudo-R^2 is not bound between 0 and 1. As a result, in some cases the R^2 loss can be negative or larger than 1.

Results for total storage efficiency (Figure 13) show that the most influential inputs are anisotropy ratio, mean reservoir permeability, permeability layer arrangement, and reservoir thickness. In general, this pattern holds true over all combinations of design and metamodel. There is evidence of a slight contribution from the injection rate.

The plume radius response (Figure 14) appears to be almost entirely driven by the injection rate and reservoir thickness, with all combinations of designs and metamodels in broad agreement. The same inputs are influential for average pressure as well (Figure 15). In this case, the instability observed in the augmented pairs/AVAS combination produces some very extreme R^2 loss values. Note that in the validation results in Table 9, the AVAS model for the augmented pairs design had a pseudo-R^2 value of -5.428, which indicates that the model fit is much worse than that obtained by always predicting the mean response no matter what the input values are. With a full model that is predicting so poorly, removal of any inputs is going to make predictions worse.

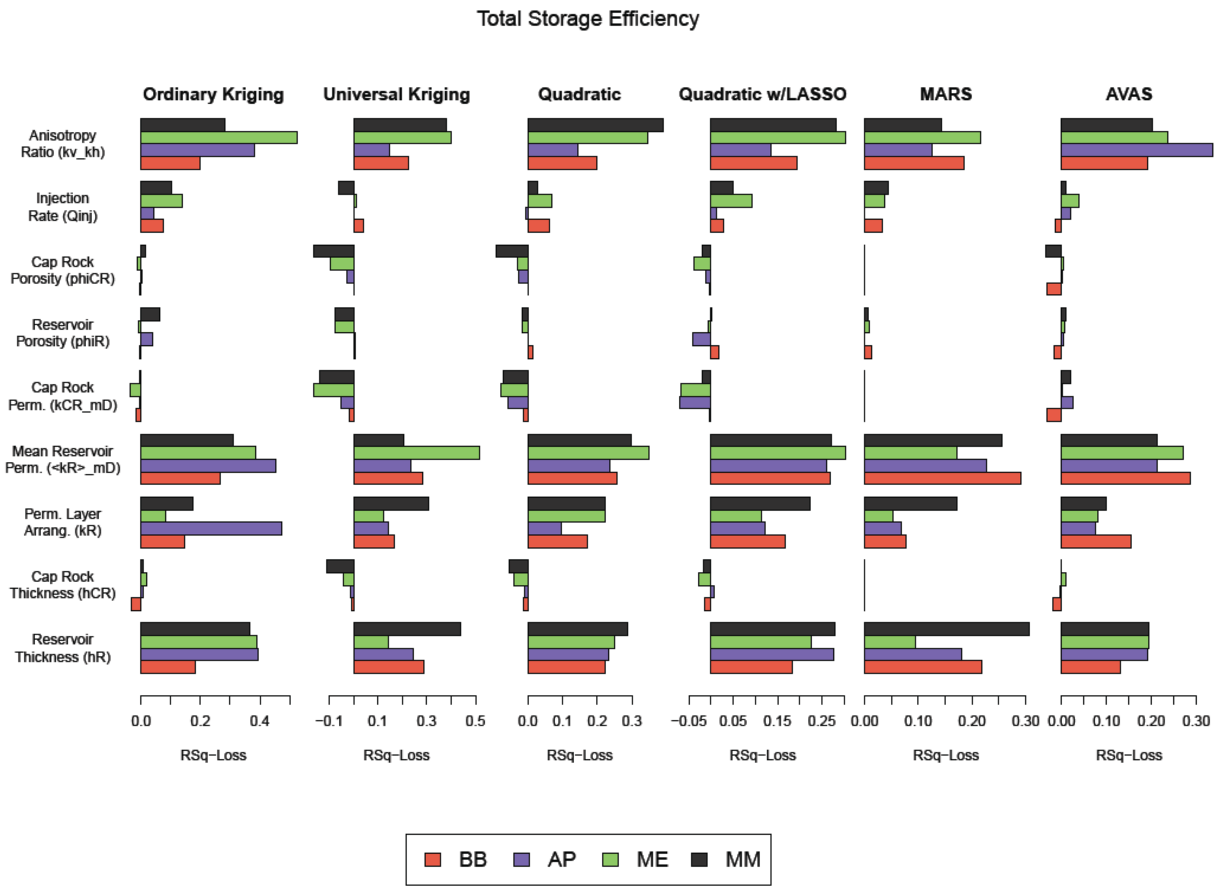


Figure 13. Variable importance results for the "Total Storage Efficiency" response.

Influential variables include anisotropy ratio, mean reservoir permeability, permeability layer arrangement, and reservoir thickness.

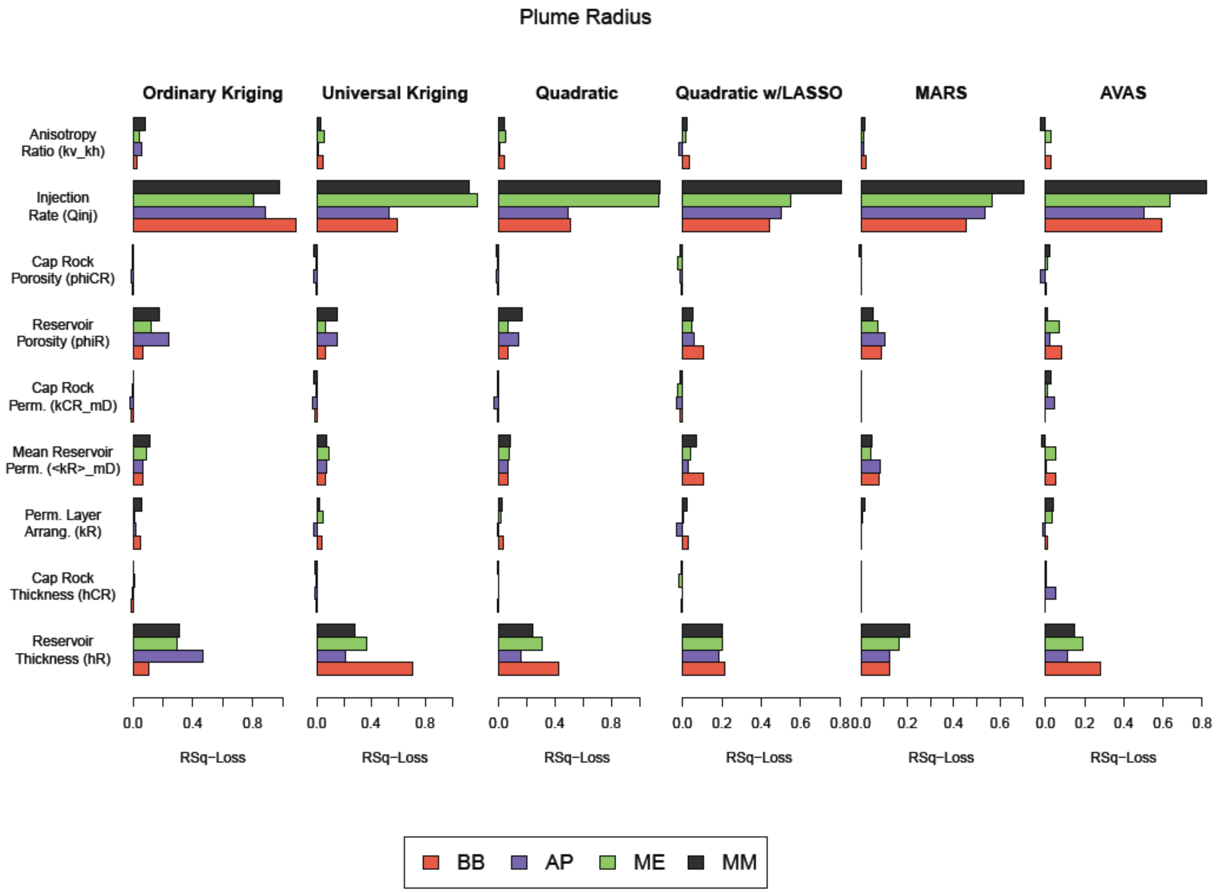


Figure 14. Variable importance results for the "Plume Radius" response.

Influential variables are injection rate and reservoir thickness.

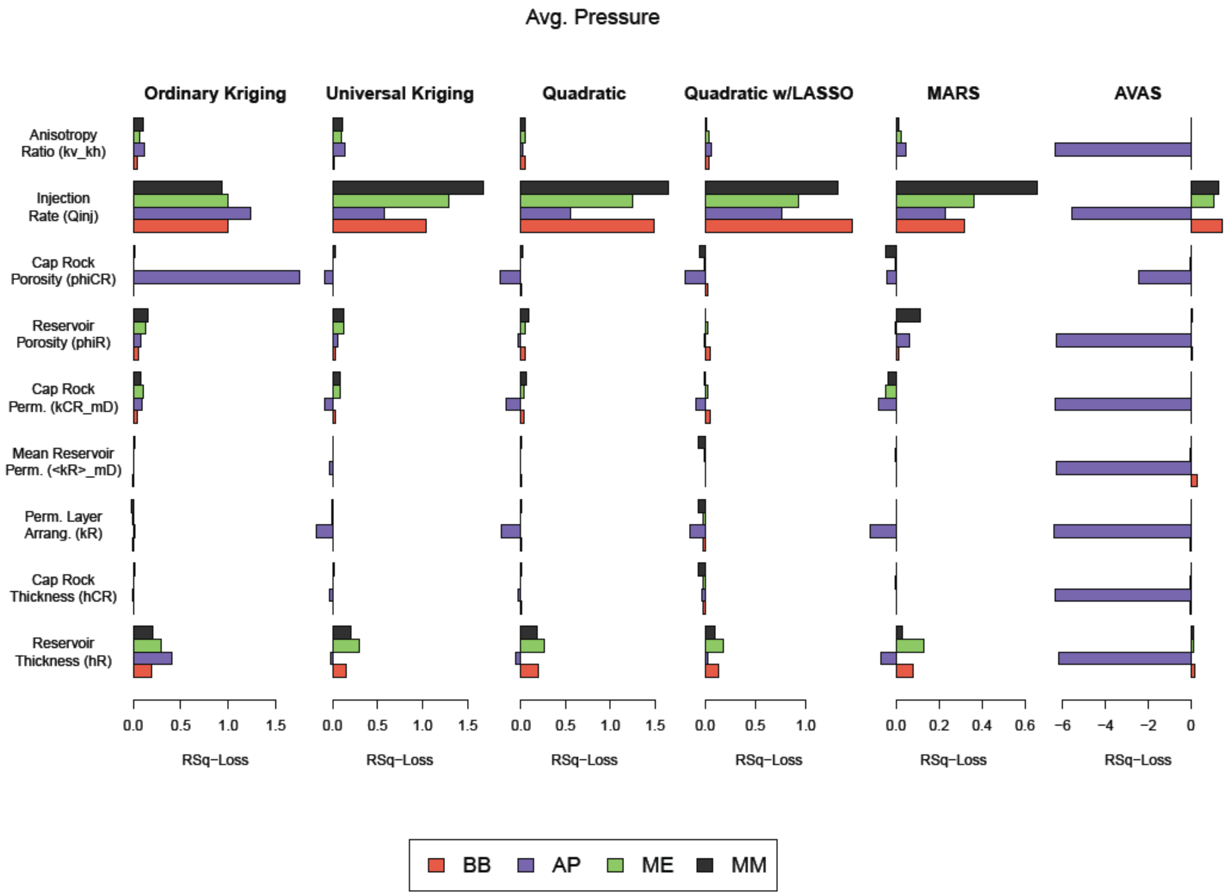


Figure 15. Variable importance results for the "Average Pressure" response.

Influential variables are injection rate and, to a small extend, reservoir thickness. Note that the augmented pairs design was very unstable for the AVAS procedure, resulting in abnormal R^2 loss values.

5.6 Discussion of Results

In this study, four design strategies and six metamodels were tested using a GEM simulation. Each combination of design and metamodel was evaluated over the 9-predictor simulation using both an independent validation on a LHS design and 5-fold cross-validation. The validation results indicated that the best designs were Box-Behnken and maximin LHS, and the best metamodels were ordinary kriging and the quadratic polynomial with LASSO variable selection. MARS and AVAS models did not work as well as the other models in most cases, and the AP design was generally the worst performer, possibly due to it having a smaller number of runs than the other designs.

A comparison of the validation and cross-validation results showed that the cross-validation error rates were larger than the validation error rates, and that the effect was disproportionately greater on the Box-Behnken and augmented pairs designs (see Figure 16 and Figure 17). This resulted in the maximin LHS appearing far more favorable in the cross-validation study. Again, the ordinary kriging and quadratic w/LASSO metamodels were the best performers.

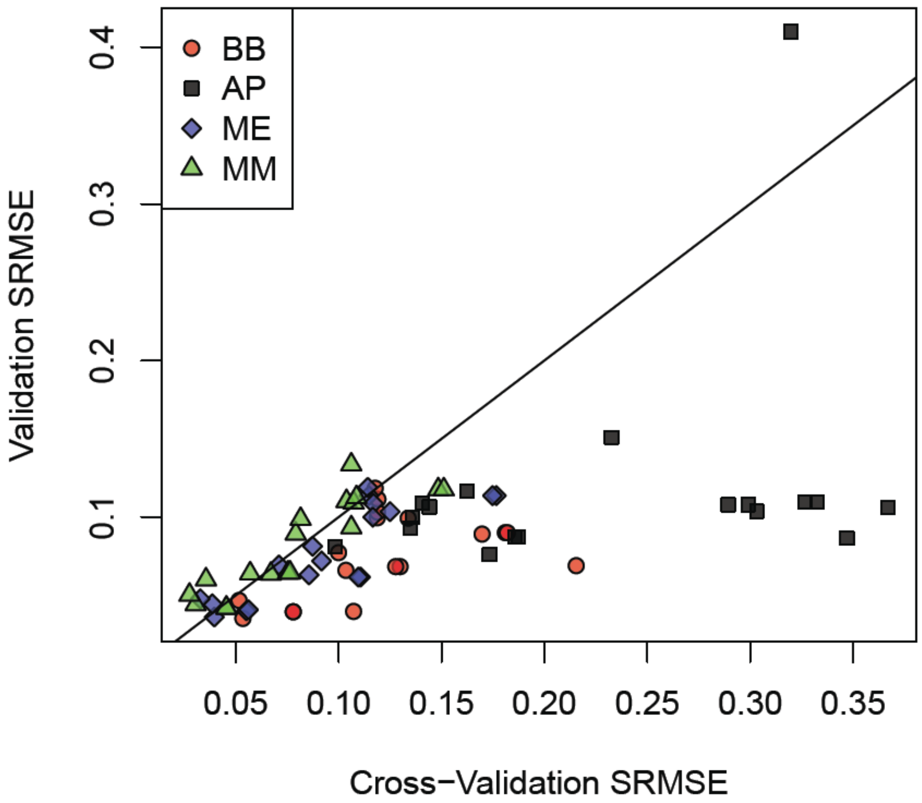


Figure 16. Comparison of validation and cross-validation scaled RMSE values for the metamodels, colored by design type.

The diagonal line indicates where error rates would be the same between validation and cross-validation.

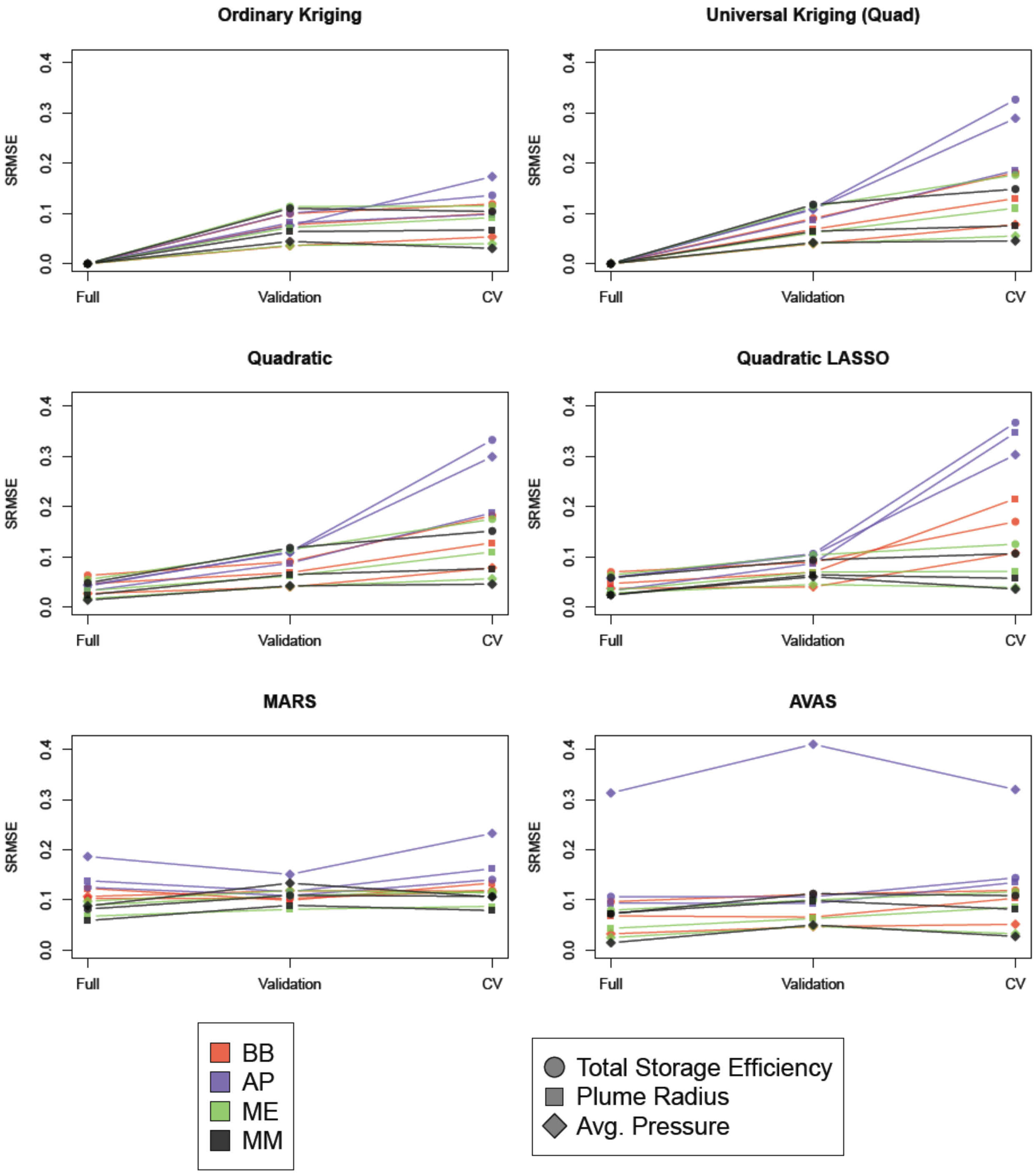


Figure 17. Progression through full model fit SRMSE, validation SRMSE, and cross-validation SRMSE, by design, metamodel, and response.

From the discussion above, there are two major conclusions to be drawn from the study. First, the Box-Behnken design with a quadratic metamodel seems to work quite well with this simulation, especially if a variable selection method like LASSO is used first. The maximin LHS design and ordinary kriging metamodel also performed well, but not consistently better than the traditional approach.

Second, the results highlight the fact that factorial designs like the Box-Behnken design may be disproportionately affected by higher error rates when a cross-validation approach is used. One must be careful in using cross-validation as a technique with factorial design data, since the error rates paint a much different picture when compared to the validation approach, which is more straightforward and assessed model fit over the entire input space.

6 Summary and Conclusions

There were two major objectives to be accomplished in Simplified Modeling Task 3. The first goal was to identify several candidate approaches that predict simulation output using an approximation model (i.e., a metamodel). These approaches include both a design of inputs at which to observe the true simulation output as well as a modeling paradigm whose parameters are fit using those observed outputs. The second goal was to compare those approaches in order to determine which methods work well in different situations, and to make recommendations for others who may be attempting to model simulation outputs in this field.

Regarding the first objective, a collection of representative experimental designs were selected from two broad categories. The Box-Behnken and augmented pairs designs were chosen to represent the more traditional factorial design approach. The former was chosen because it appears to be the most commonly used design of this type in the gas and petroleum literature; the latter was chosen as an alternative to the Box-Behnken that requires fewer simulation runs. Maximin Latin hypercube sampling (LHS) and maximum entropy designs were selected to represent sampling designs, which form the second category. Sampling designs have grown in popularity in recent years, and LHS designs are the most common of these designs seen in the literature. The maximin LHS is a special type of LHS design that optimizes a desirable space-filling criterion. Maximum entropy designs are a popular alternative to LHS designs, and also have excellent space-filling characteristics.

In terms of metamodeling approaches, several common competing methods were selected for comparison in this task. These included quadratic polynomials, which are typically used in conjunction with factorial designs, kriging models, which are often used with sampling design approaches, and two other competing methods called MARS and AVAS. In addition, a version of quadratic modeling that uses LASSO variable selection was also considered as a more refined alternative to traditional quadratic regression modeling.

Regarding the second objective, all combinations of designs and metamodels were used to predict three different responses from a 9-input CMG-GEM simulation of gas injection into a closed reservoir. Models were evaluated using three criteria: root mean squared error (RMSE), scaled RMSE, and pseudo-R^2. Evaluation was performed both for 5-fold cross-validated predictions on the training set as well as predictions on an independent test set. Results showed that using the traditional approach of a Box-Behnken design with a quadratic metamodel appears to work just as well, if not better, than using newer sampling designs and more complex modeling strategies. In the validation study, this traditional approach was the top performer, and it was competitive in the cross-validation study as well. The maximin LHS with either ordinary kriging or the quadratic polynomial model with LASSO variable selection was another competitive method that showed generally robust performance across the three responses. The worst performing design in general was the augmented pairs design, which may be attributed to the fact that it has fewer observations than the other designs. The worst performing metamodels were MARS and AVAS.

7 Acknowledgments

This work was supported by U.S. Department of Energy National Energy Technology Laboratory award DE-FE0009051 and Ohio Development Services Agency grant D-13-02. We thank our colleagues Dr. Doug Mooney and Dr. Rodney Osborne for their technical review of our report.

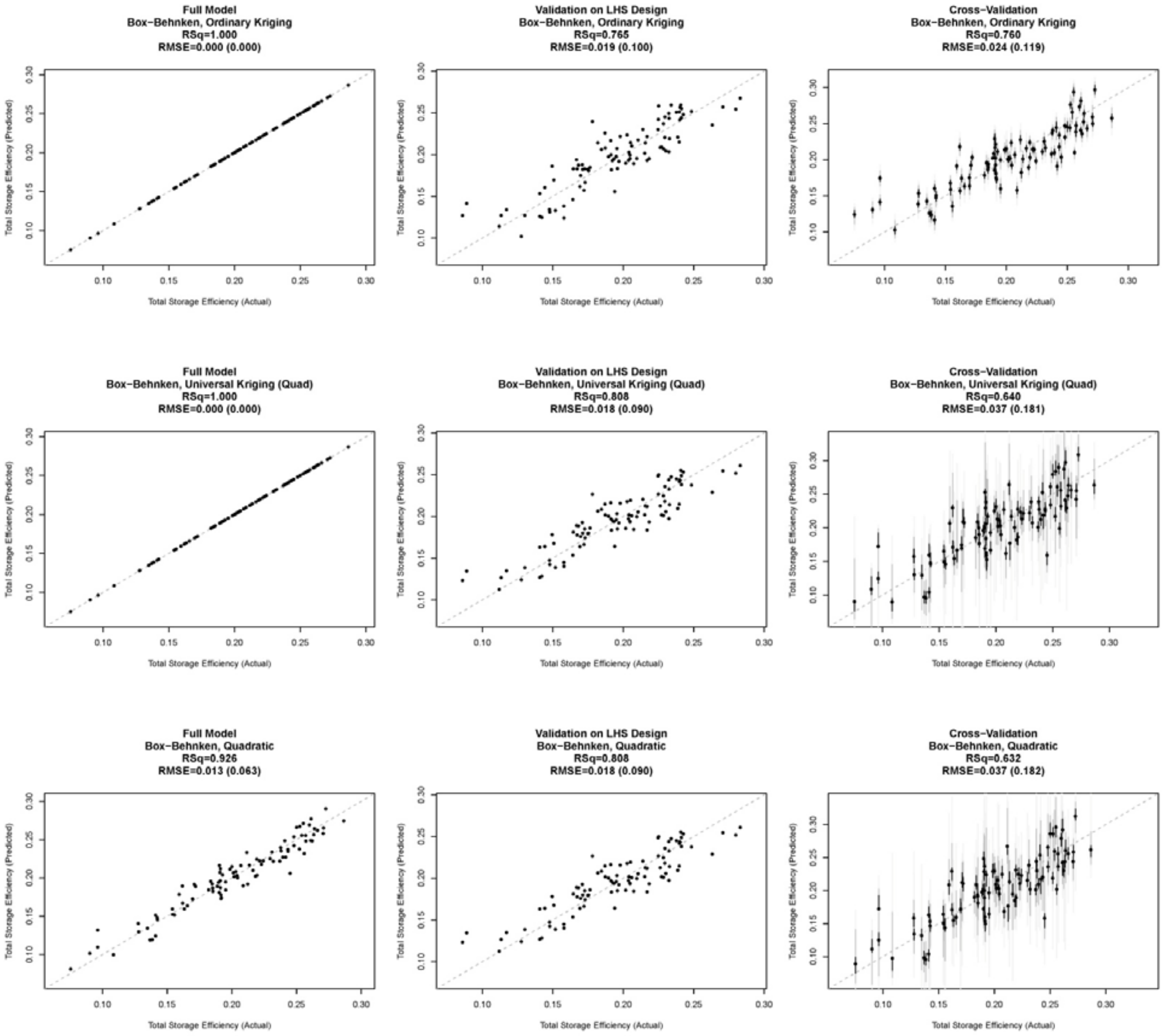
8 References

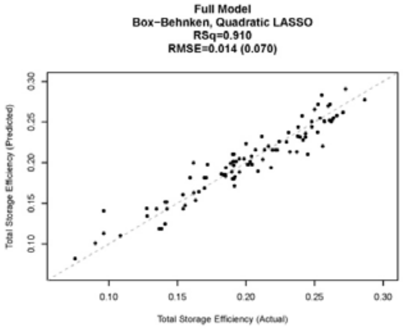
1. Osterloh, W. *Use of Multiple-Response Optimization to Assist Reservoir Simulation Probabilistic Forecasting and History Matching*. in *SPE Annual Technical Conference and Exhibition*. 2008.
2. Ekeoma, E. and D. Appah. *Latin Hypercube Sampling (LHS) for Gas Reserves*. in *Nigeria Annual International Conference and Exhibition*. 2009.
3. Zubarev, D. *Pros and cons of applying proxy-models as a substitute for full reservoir simulations*. in *SPE Annual Technical Conference and Exhibition*. 2009.
4. Kalla, S. and C.D. White. *Efficient design of reservoir simulation studies for development and optimization*. in *SPE Annual Technical Conference and Exhibition*. 2005.
5. Anbar, S. *Development of a predictive model for carbon dioxide sequestration in deep saline carbonate aquifers*. in *SPE Annual Technical Conference and Exhibition*. 2010.
6. Wriedt, J., *et al.*, *A methodology for quantifying risk and likelihood of failure for carbon dioxide injection into deep saline reservoirs.* International Journal of Greenhouse Gas Control, 2014. **20**: p. 196–211.
7. Plackett, R.L. and J.P. Burman, *The design of optimum multifactorial experiments.* Biometrika, 1946: p. 305–325.
8. Box, G.E. and D. Behnken, *Some new three level designs for the study of quantitative variables.* Technometrics, 1960. **2**(4): p. 455–475.
9. Morris, M.D., *A class of three-level experimental designs for response surface modeling.* Technometrics, 2000. **42**(2): p. 111–121.
10. McKay, M.D., R.J. Beckman, and W.J. Conover, *Comparison of three methods for selecting values of input variables in the analysis of output from a computer code.* Technometrics, 1979. **21**(2): p. 239–245.
11. Johnson, M.E., L.M. Moore, and D. Ylvisaker, *Minimax and maximin distance designs.* Journal of statistical planning and inference, 1990. **26**(2): p. 131–148.
12. Shewry, M.C. and H.P. Wynn, *Maximum entropy sampling.* Journal of Applied Statistics, 1987. **14**(2): p. 165–170.
13. Shannon, C.E., *A mathematical theory of communication.* ACM SIGMOBILE Mobile Computing and Communications Review, 2001. **5**(1): p. 3–55.
14. Hickernell, F., *Lattice rules: how well do they measure up*, in *Random and Quasi-Random Point Sets*, P. Hellekalek and G. Larcher, Editors. 1998, Springer. p. 109–166.
15. Tibshirani, R., *Regression shrinkage and selection via the lasso.* Journal of the Royal Statistical Society. Series B (Methodological), 1996: p. 267–288.
16. Simpson, T.W., *et al.*, *Comparison of response surface and kriging models for multidisciplinary design optimization.* AIAA paper 98, 1998. **4758**(7).
17. Cressie, N., *Statistics for spatial data (wiley series in probability and statistics).* 1993.
18. Krige, D.G., *A statistical approach to some mine valuation and allied problems on the Witwatersrand*, 1951, University of the Witwatersrand.
19. Friedman, J.H., *Multivariate adaptive regression splines.* The Annals of Statistics, 1991: p. 1–67.
20. Breiman, L. and J.H. Friedman, *Estimating optimal transformations for multiple regression and correlation.* Journal of the American Statistical Association, 1985. **80**(391): p. 580–598.

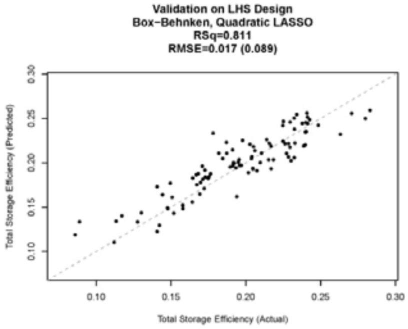
21. Tibshirani, R., *Estimating transformations for regression via additivity and variance stabilization.* Journal of the American Statistical Association, 1988. **83**(402): p. 394–405.
22. Duchon, J., *Splines minimizing rotation-invariant semi-norms in Sobolev spaces*, in *Constructive theory of functions of several variables.* 1977, Springer. p. 85–100.
23. Drucker, H., *et al.*, *Support vector regression machines.* Advances in neural information processing systems, 1997. **9**: p. 155–161.
24. Chen, S., C.F. Cowan, and P.M. Grant, *Orthogonal least squares learning algorithm for radial basis function networks.* Neural Networks, IEEE Transactions on, 1991. **2**(2): p. 302–309.
25. Hastie, T., R. Tibshirani, and J.H. Friedman, *The elements of statistical learning: data mining, inference, and prediction*. Second Edition. 2008: Springer.
26. Mishra, S., Y.D. Oruganti, and J. Sminchak, *Parametric analysis of CO_2 geologic sequestration in closed volumes.* Environmental Geosciences, 2014. **21**(2): p. 59–74.
27. White, M.D., *et al.*, *STOMP Subsurface Transport Over Multiple Phases: STOMP-CO2 and STOMP-CO2e Guide: Version 1.0*, 2012, Pacific Northwest National Laboratory (PNNL), Richland, WA (US).
28. Computer Modelling Group LTD., *GEM –- Compositional and Unconventional Reservoir Simulator*, 2014.
29. R Development Core Team, *R: A Language and Environment for Statistical Computing*, 2011: Vienna, Austria.
30. Roustant, O., D. Ginsbourger, and Y. Deville, *DiceKriging: Kriging methods for computer experiments, R package version 1.3.2*, 2011.
31. Kuhn, M., *et al.*, *caret: Classification and regression training. R package version 5.15-044*, 2012.
32. Hastie, T., *et al.*, *mda: Mixture and flexible discriminant analysis, R package version 0.4-2*, 2011.
33. Spector, P., *et al.*, *acepack: ace() and avas() for selecting regression transformations. R package version 1.3-3.0.* 2010.

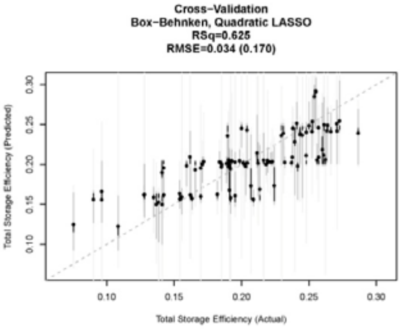
Appendix

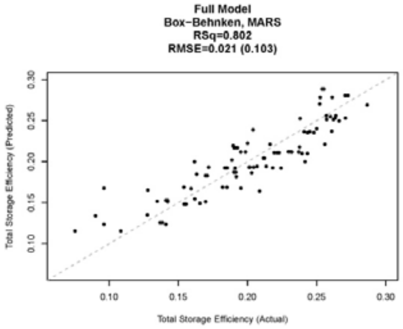
Total Storage Efficiency

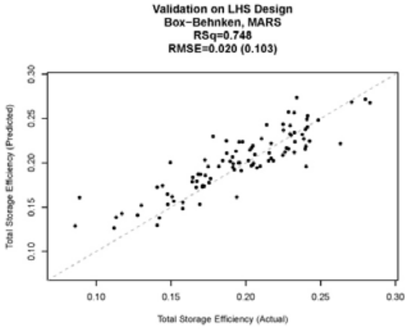


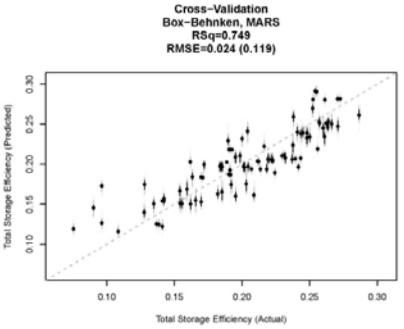

Full Model
Box−Behnken, Quadratic LASSO
RSq=0.910
RMSE=0.014 (0.070)

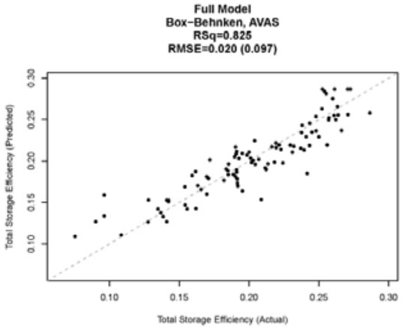

Validation on LHS Design
Box−Behnken, Quadratic LASSO
RSq=0.811
RMSE=0.017 (0.089)

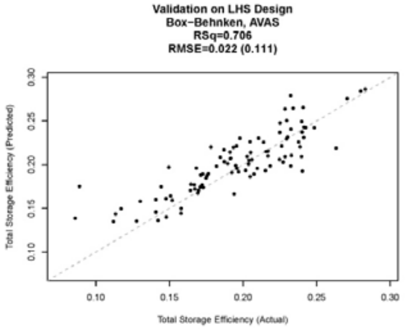

Cross−Validation
Box−Behnken, Quadratic LASSO
RSq=0.625
RMSE=0.034 (0.170)

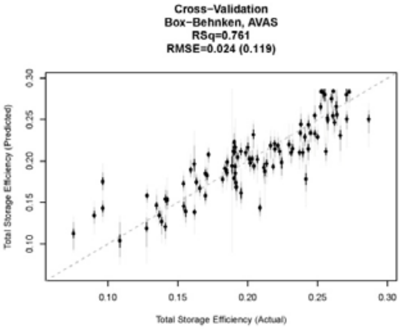

Full Model
Box−Behnken, MARS
RSq=0.802
RMSE=0.021 (0.103)

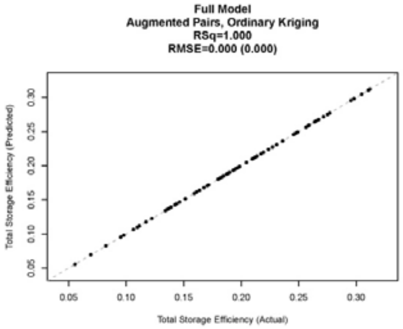

Validation on LHS Design
Box−Behnken, MARS
RSq=0.748
RMSE=0.020 (0.103)

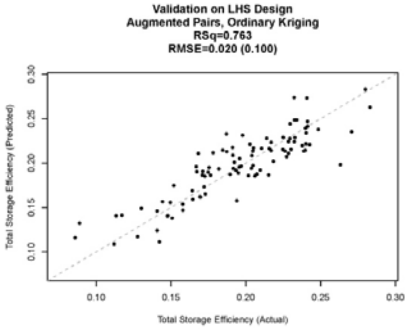

Cross−Validation
Box−Behnken, MARS
RSq=0.749
RMSE=0.024 (0.119)

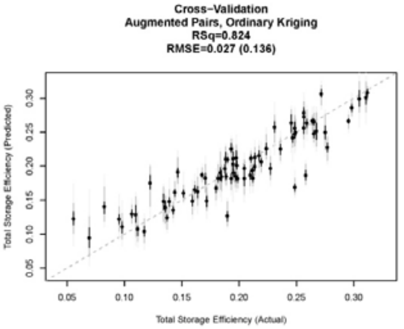

Full Model
Box−Behnken, AVAS
RSq=0.825
RMSE=0.020 (0.097)

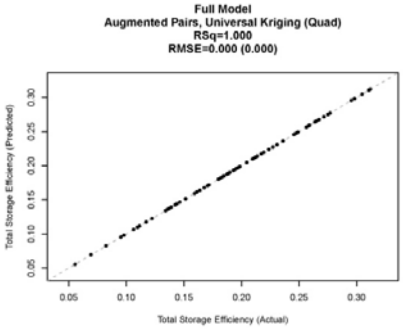

Validation on LHS Design
Box−Behnken, AVAS
RSq=0.706
RMSE=0.022 (0.111)

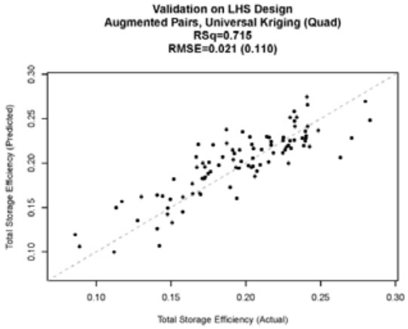

Cross−Validation
Box−Behnken, AVAS
RSq=0.761
RMSE=0.024 (0.119)

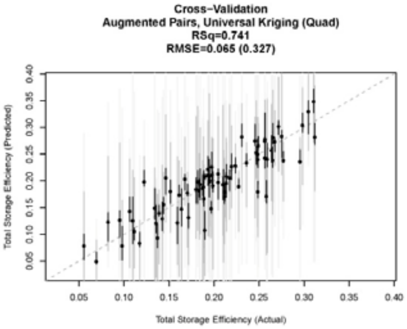

Full Model
Augmented Pairs, Ordinary Kriging
RSq=1.000
RMSE=0.000 (0.000)

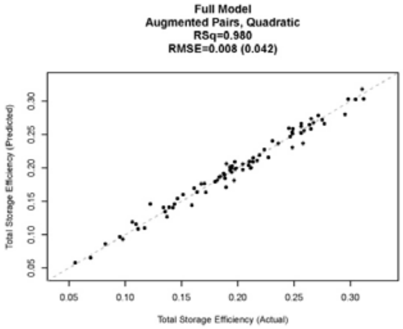

Validation on LHS Design
Augmented Pairs, Ordinary Kriging
RSq=0.763
RMSE=0.020 (0.100)

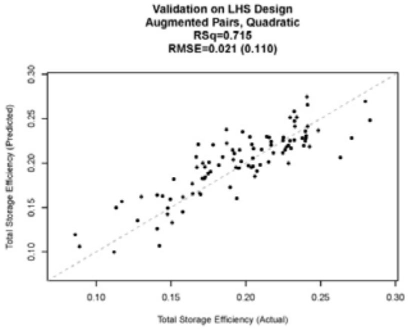

Cross−Validation
Augmented Pairs, Ordinary Kriging
RSq=0.824
RMSE=0.027 (0.136)

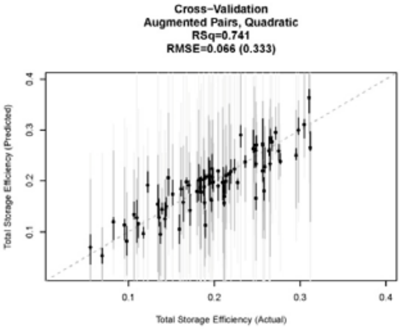

Full Model
Augmented Pairs, Universal Kriging (Quad)
RSq=1.000
RMSE=0.000 (0.000)
Total Storage Efficiency (Predicted)
Total Storage Efficiency (Actual)
0.05
0.10
0.15
0.20
0.25
0.30

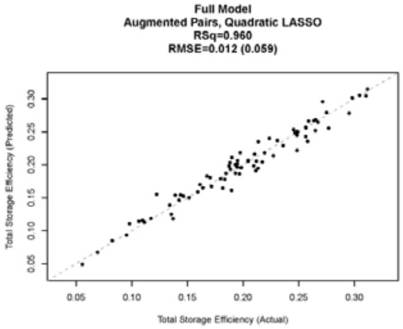

Validation on LHS Design
Augmented Pairs, Universal Kriging (Quad)
RSq=0.715
RMSE=0.021 (0.110)
Total Storage Efficiency (Predicted)
Total Storage Efficiency (Actual)
0.10
0.15
0.20
0.25
0.30

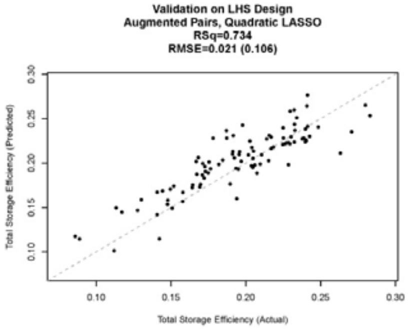

Cross−Validation
Augmented Pairs, Universal Kriging (Quad)
RSq=0.741
RMSE=0.065 (0.327)
Total Storage Efficiency (Predicted)
Total Storage Efficiency (Actual)
0.05
0.10
0.15
0.20
0.25
0.30
0.35
0.40

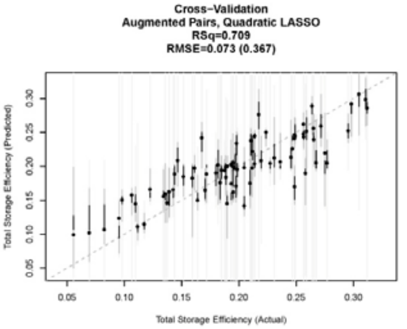

Full Model
Augmented Pairs, Quadratic
RSq=0.980
RMSE=0.008 (0.042)
Total Storage Efficiency (Predicted)
Total Storage Efficiency (Actual)
0.05
0.10
0.15
0.20
0.25
0.30

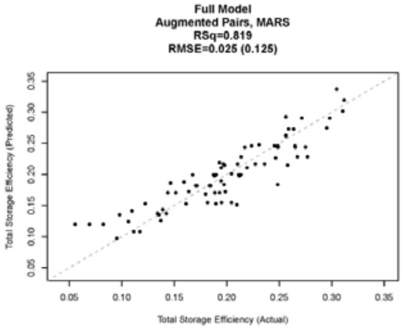

Validation on LHS Design
Augmented Pairs, Quadratic
RSq=0.715
RMSE=0.021 (0.110)
Total Storage Efficiency (Predicted)
Total Storage Efficiency (Actual)
0.10
0.15
0.20
0.25
0.30

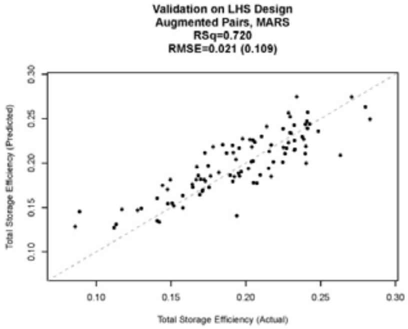

Cross−Validation
Augmented Pairs, Quadratic
RSq=0.741
RMSE=0.066 (0.333)
Total Storage Efficiency (Predicted)
Total Storage Efficiency (Actual)
0.1
0.2
0.3
0.4

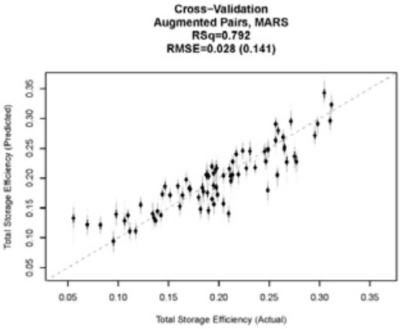

Full Model
Augmented Pairs, Quadratic LASSO
RSq=0.960
RMSE=0.012 (0.059)
Total Storage Efficiency (Predicted)
Total Storage Efficiency (Actual)
0.05
0.10
0.15
0.20
0.25
0.30

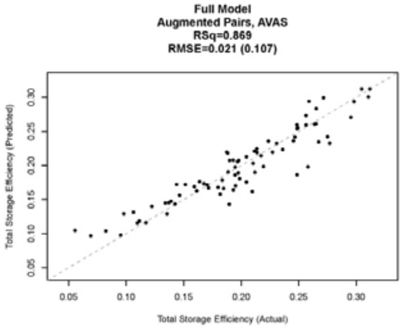

Validation on LHS Design
Augmented Pairs, Quadratic LASSO
RSq=0.734
RMSE=0.021 (0.106)
Total Storage Efficiency (Predicted)
Total Storage Efficiency (Actual)
0.10
0.15
0.20
0.25
0.30

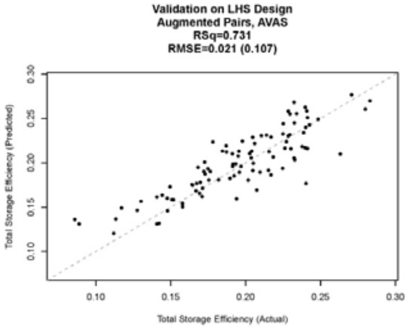

Cross−Validation
Augmented Pairs, Quadratic LASSO
RSq=0.709
RMSE=0.073 (0.367)
Total Storage Efficiency (Predicted)
Total Storage Efficiency (Actual)
0.05
0.10
0.15
0.20
0.25
0.30

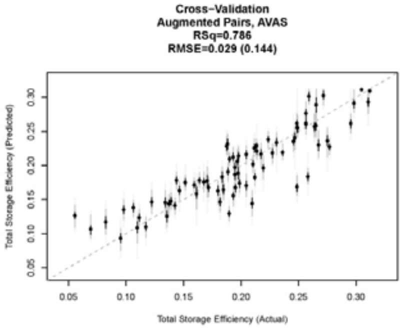

Full Model
Augmented Pairs, MARS
RSq=0.819
RMSE=0.025 (0.125)
Total Storage Efficiency (Predicted)
Total Storage Efficiency (Actual)
0.05
0.10
0.15
0.20
0.25
0.30
0.35

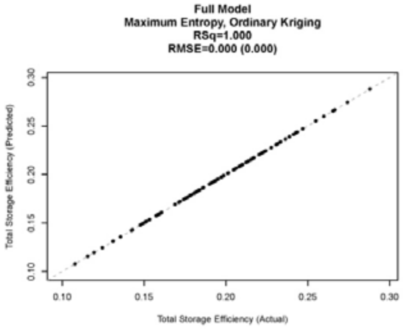

Validation on LHS Design
Augmented Pairs, MARS
RSq=0.720
RMSE=0.021 (0.109)
Total Storage Efficiency (Predicted)
Total Storage Efficiency (Actual)
0.10
0.15
0.20
0.25
0.30

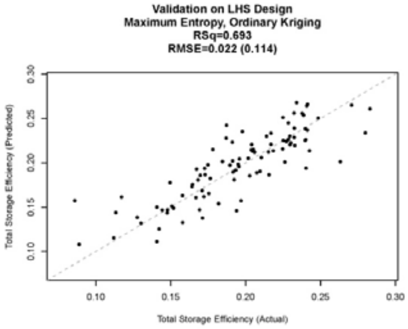

Cross−Validation
Augmented Pairs, MARS
RSq=0.792
RMSE=0.028 (0.141)
Total Storage Efficiency (Predicted)
Total Storage Efficiency (Actual)
0.05
0.10
0.15
0.20
0.25
0.30
0.35

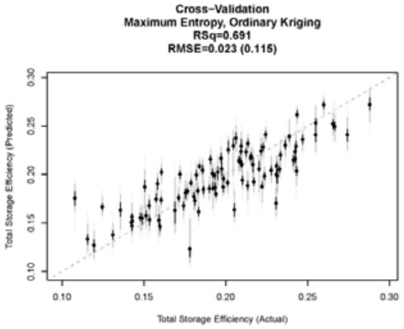

Full Model
Augmented Pairs, AVAS
RSq=0.869
RMSE=0.021 (0.107)

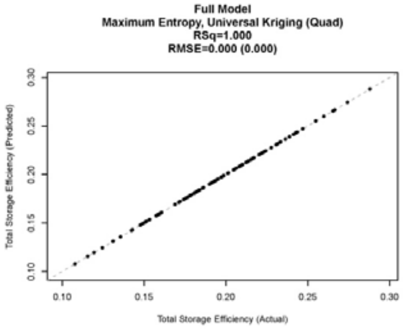

Validation on LHS Design
Augmented Pairs, AVAS
RSq=0.731
RMSE=0.021 (0.107)

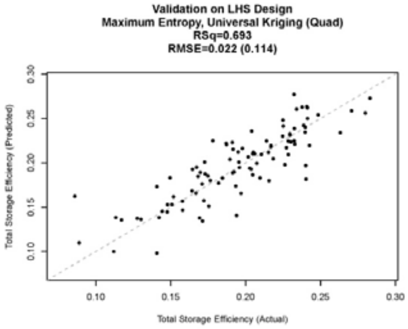

Cross−Validation
Augmented Pairs, AVAS
RSq=0.786
RMSE=0.029 (0.144)

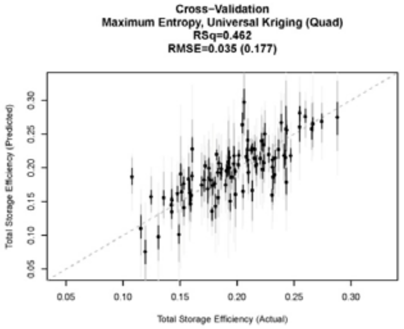

Full Model
Maximum Entropy, Ordinary Kriging
RSq=1.000
RMSE=0.000 (0.000)

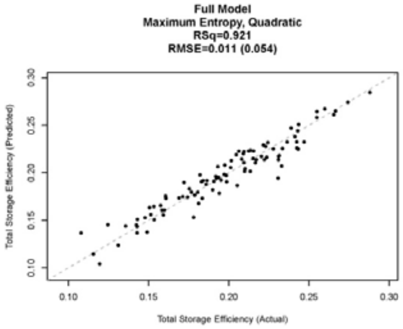

Validation on LHS Design
Maximum Entropy, Ordinary Kriging
RSq=0.693
RMSE=0.022 (0.114)

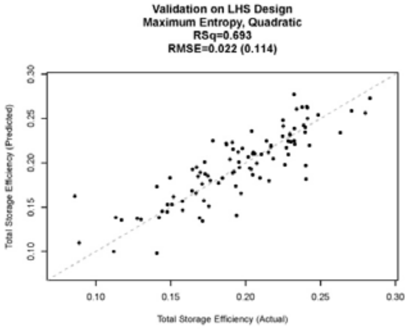

Cross−Validation
Maximum Entropy, Ordinary Kriging
RSq=0.691
RMSE=0.023 (0.115)

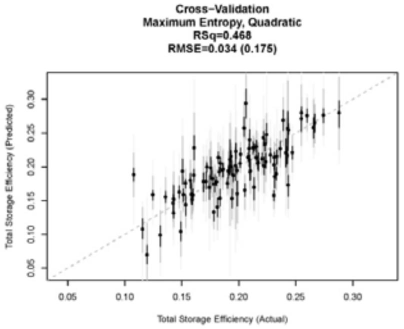

Full Model
Maximum Entropy, Universal Kriging (Quad)
RSq=1.000
RMSE=0.000 (0.000)

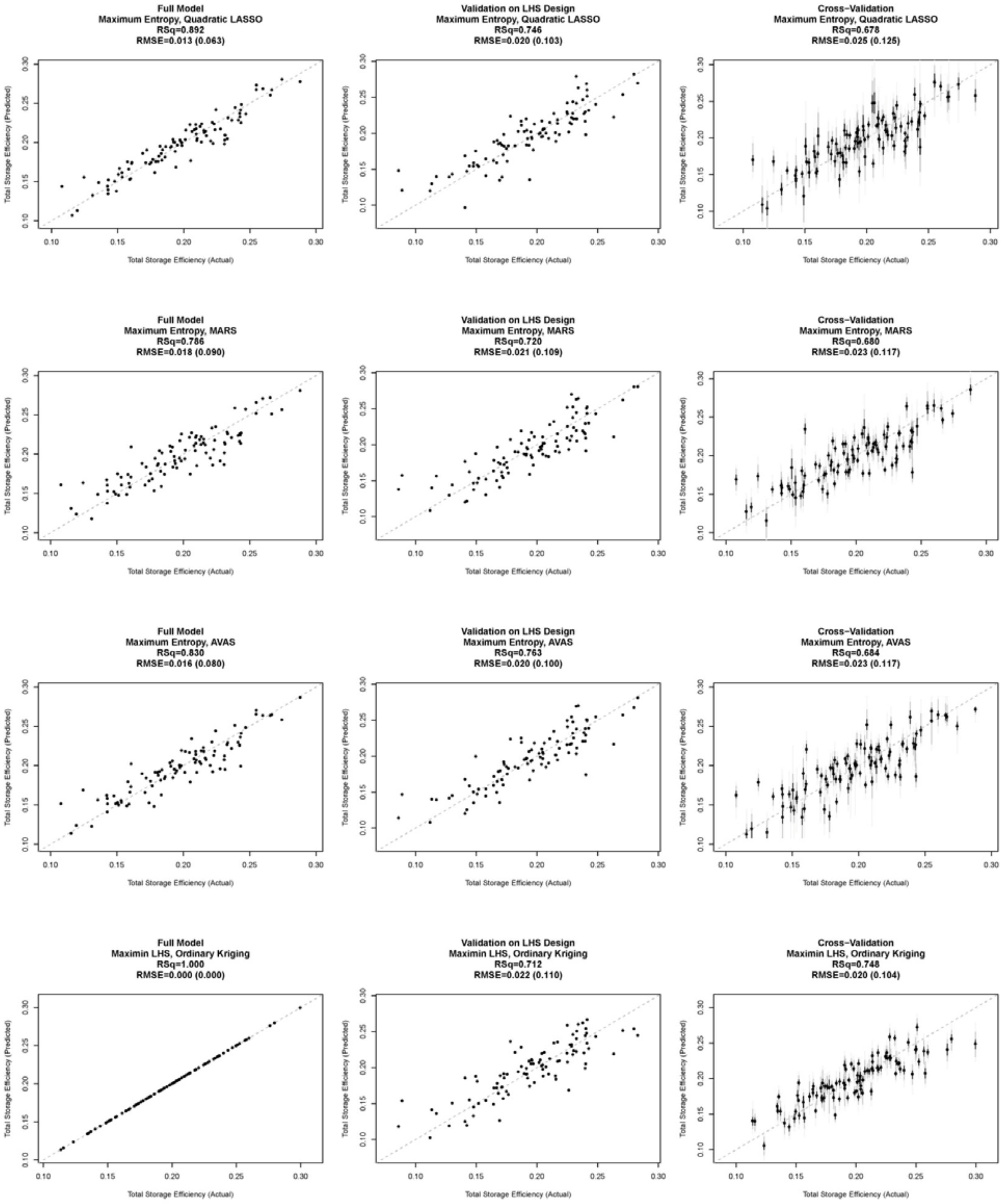

Validation on LHS Design
Maximum Entropy, Universal Kriging (Quad)
RSq=0.693
RMSE=0.022 (0.114)


Cross−Validation
Maximum Entropy, Universal Kriging (Quad)
RSq=0.462
RMSE=0.035 (0.177)


Full Model
Maximum Entropy, Quadratic
RSq=0.921
RMSE=0.011 (0.054)


Validation on LHS Design
Maximum Entropy, Quadratic
RSq=0.693
RMSE=0.022 (0.114)


Cross−Validation
Maximum Entropy, Quadratic
RSq=0.468
RMSE=0.034 (0.175)



Full Model
Maximum Entropy, Quadratic LASSO
RSq=0.892
RMSE=0.013 (0.063)

Validation on LHS Design
Maximum Entropy, Quadratic LASSO
RSq=0.746
RMSE=0.020 (0.103)

Cross−Validation
Maximum Entropy, Quadratic LASSO
RSq=0.678
RMSE=0.025 (0.125)

Full Model
Maximum Entropy, MARS
RSq=0.786
RMSE=0.018 (0.090)

Validation on LHS Design
Maximum Entropy, MARS
RSq=0.720
RMSE=0.021 (0.109)

Cross−Validation
Maximum Entropy, MARS
RSq=0.680
RMSE=0.023 (0.117)

Full Model
Maximum Entropy, AVAS
RSq=0.830
RMSE=0.016 (0.080)

Validation on LHS Design
Maximum Entropy, AVAS
RSq=0.763
RMSE=0.020 (0.100)

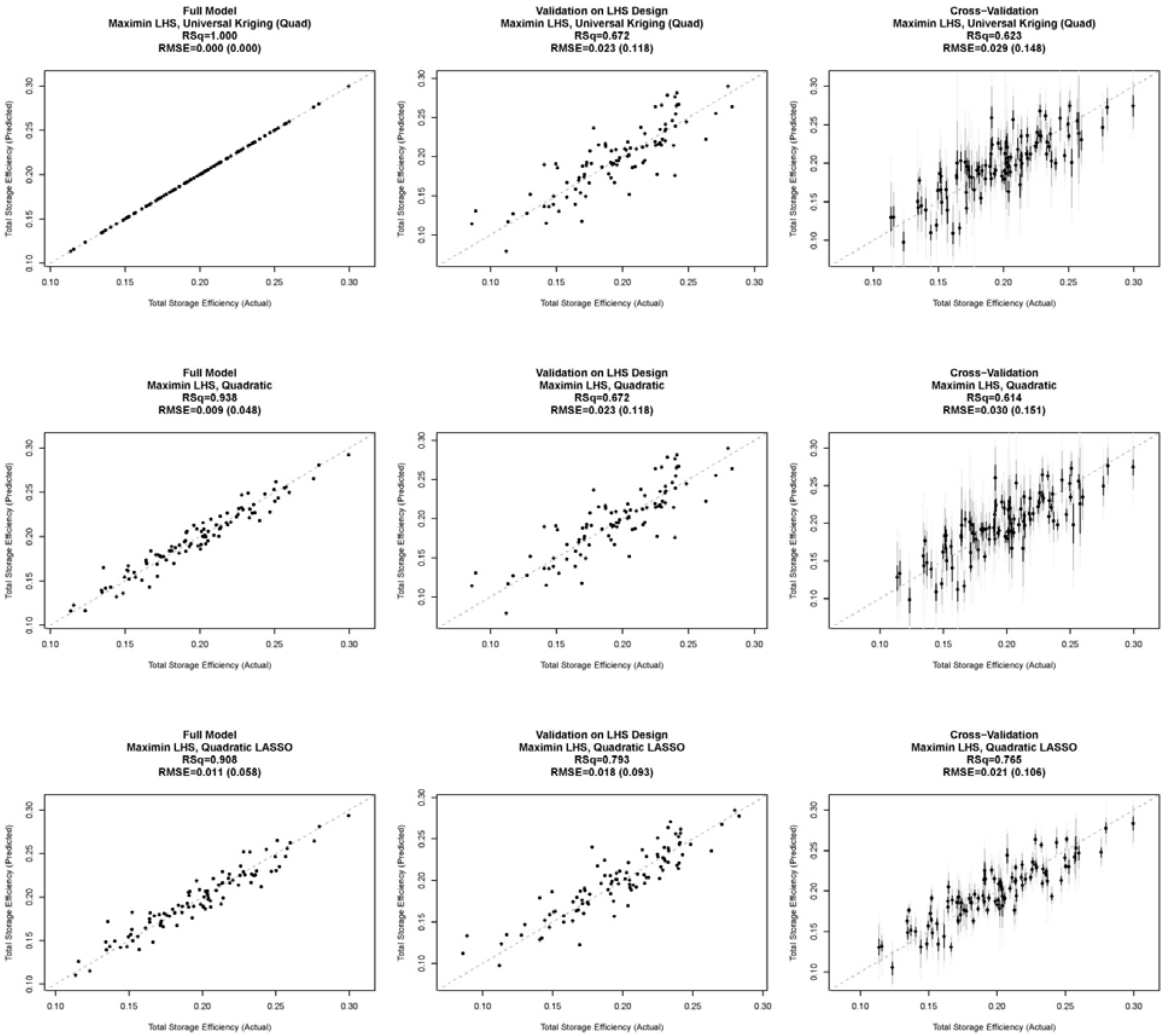
Cross−Validation
Maximum Entropy, AVAS
RSq=0.684
RMSE=0.023 (0.117)

Full Model
Maximin LHS, Ordinary Kriging
RSq=1.000
RMSE=0.000 (0.000)

Validation on LHS Design
Maximin LHS, Ordinary Kriging
RSq=0.712
RMSE=0.022 (0.110)

Cross−Validation
Maximin LHS, Ordinary Kriging
RSq=0.748
RMSE=0.020 (0.104)

Total Storage Efficiency (Predicted)
Total Storage Efficiency (Actual)
0.10
0.15
0.20
0.25
0.30



Full Model
Maximin LHS, Universal Kriging (Quad)
RSq=1.000
RMSE=0.000 (0.000)

Validation on LHS Design
Maximin LHS, Universal Kriging (Quad)
RSq=0.672
RMSE=0.023 (0.118)

Cross−Validation
Maximin LHS, Universal Kriging (Quad)
RSq=0.623
RMSE=0.029 (0.148)

Full Model
Maximin LHS, Quadratic
RSq=0.938
RMSE=0.009 (0.048)

Validation on LHS Design
Maximin LHS, Quadratic
RSq=0.672
RMSE=0.023 (0.118)

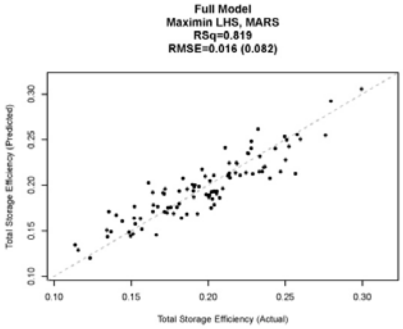
Cross−Validation
Maximin LHS, Quadratic
RSq=0.614
RMSE=0.030 (0.151)

Full Model
Maximin LHS, Quadratic LASSO
RSq=0.908
RMSE=0.011 (0.058)

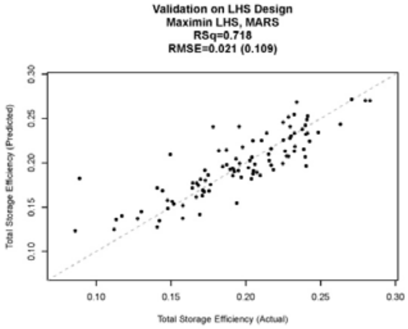
Validation on LHS Design
Maximin LHS, Quadratic LASSO
RSq=0.793
RMSE=0.018 (0.093)

Cross−Validation
Maximin LHS, Quadratic LASSO
RSq=0.765
RMSE=0.021 (0.106)

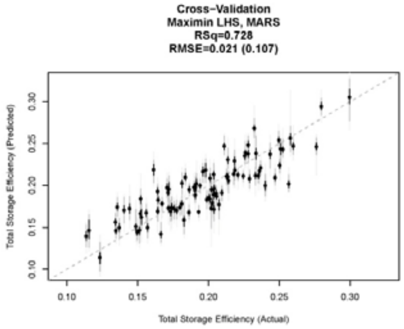
Total Storage Efficiency (Predicted)
0.10
0.15
0.20
0.25
0.30
Total Storage Efficiency (Actual)
0.10
0.15
0.20
0.25
0.30



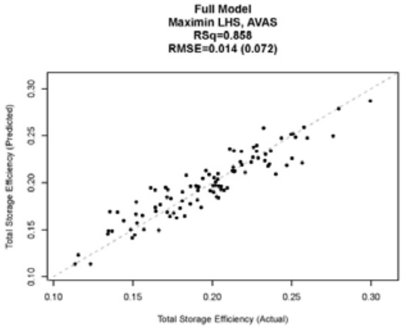
Full Model
Maximin LHS, MARS
RSq=0.819
RMSE=0.016 (0.082)



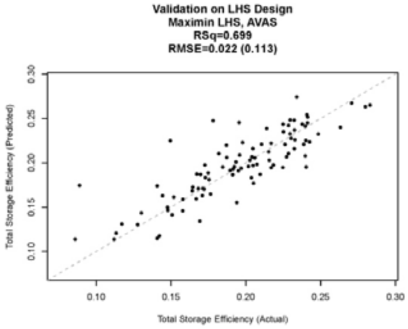
Validation on LHS Design
Maximin LHS, MARS
RSq=0.718
RMSE=0.021 (0.109)



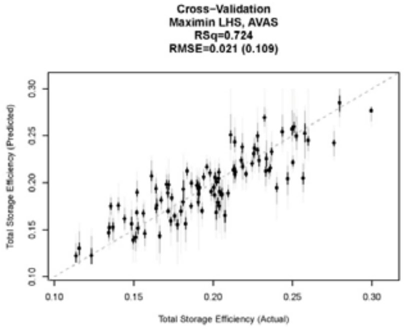
Cross-Validation
Maximin LHS, MARS
RSq=0.728
RMSE=0.021 (0.107)



Full Model
Maximin LHS, AVAS
RSq=0.858
RMSE=0.014 (0.072)

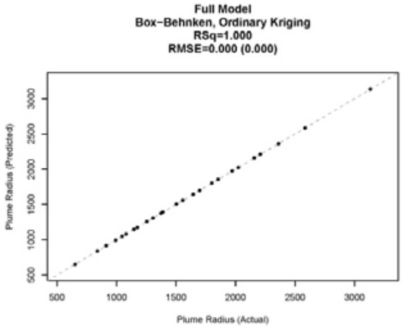


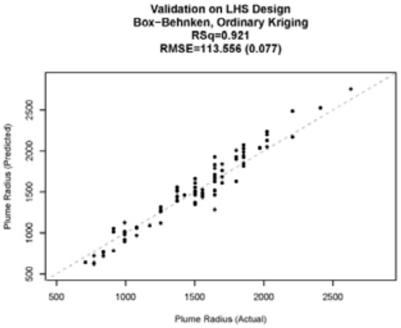
Validation on LHS Design
Maximin LHS, AVAS
RSq=0.699
RMSE=0.022 (0.113)

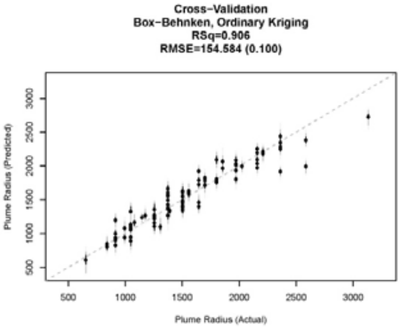


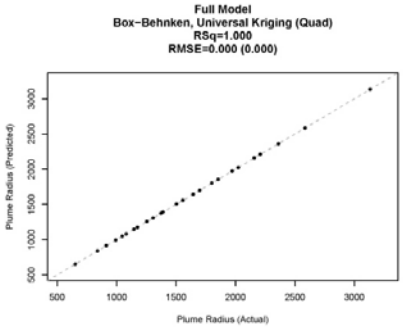
Cross-Validation
Maximin LHS, AVAS
RSq=0.724
RMSE=0.021 (0.109)

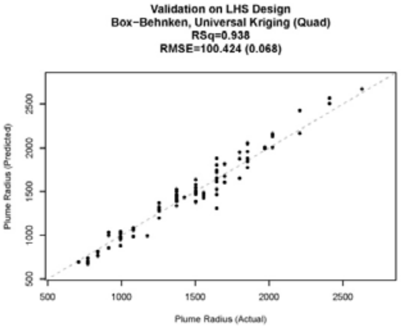
Plume Radius

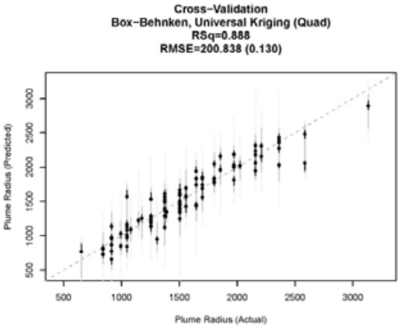


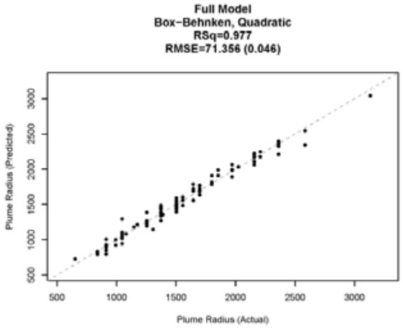


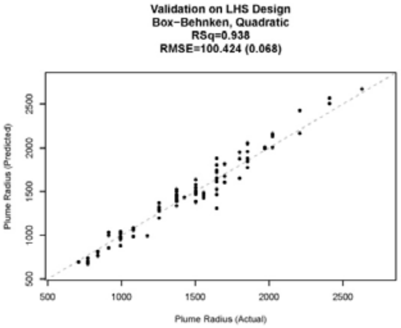


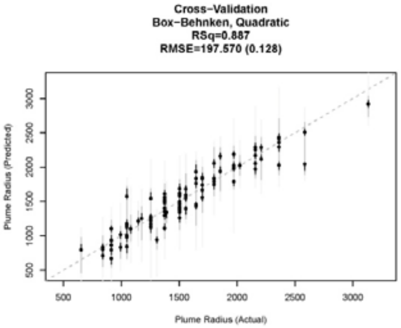


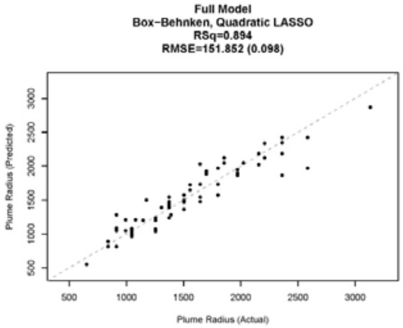


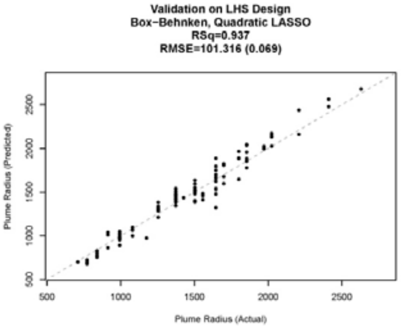


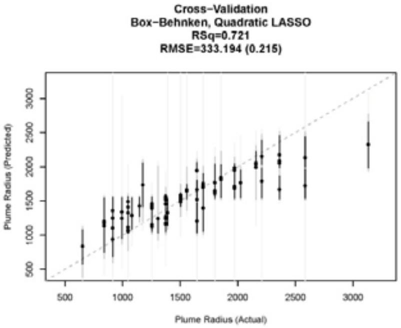


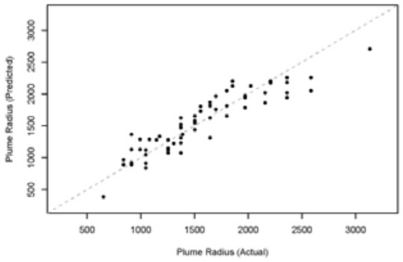


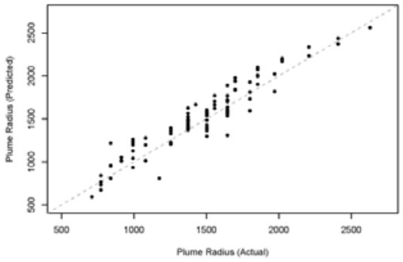


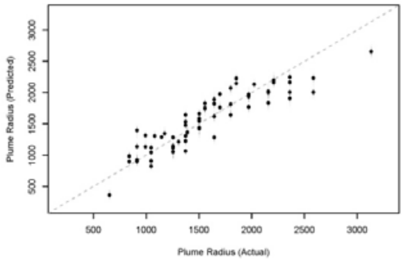


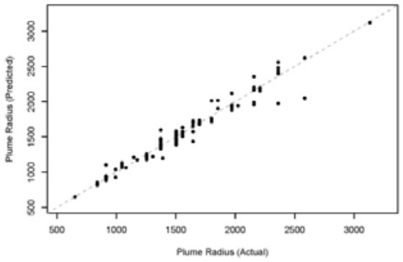


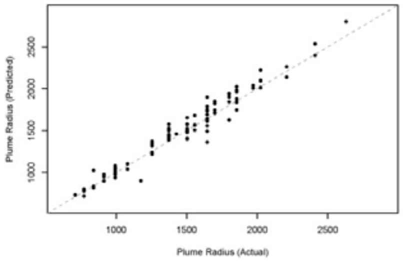


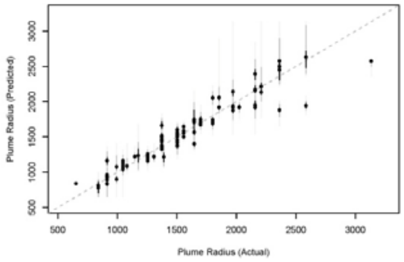

Full Model
Box−Behnken, MARS
RSq=0.834
RMSE=190.409 (0.123)

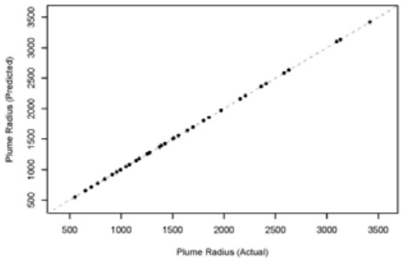

Validation on LHS Design
Box−Behnken, MARS
RSq=0.870
RMSE=145.793 (0.099)

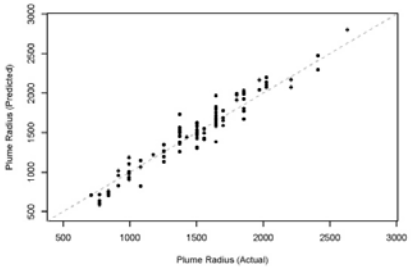

Cross−Validation
Box−Behnken, MARS
RSq=0.808
RMSE=206.926 (0.134)

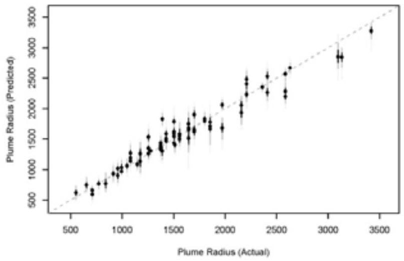

Full Model
Box−Behnken, AVAS
RSq=0.949
RMSE=105.538 (0.068)

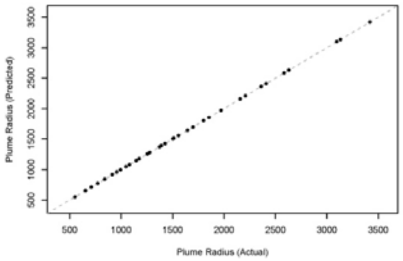

Validation on LHS Design
Box−Behnken, AVAS
RSq=0.943
RMSE=96.952 (0.066)

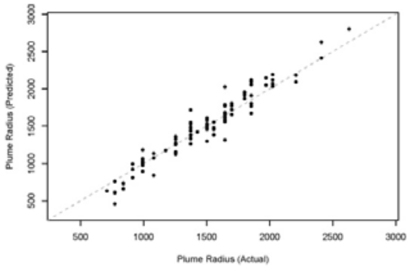

Cross−Validation
Box−Behnken, AVAS
RSq=0.907
RMSE=160.153 (0.104)

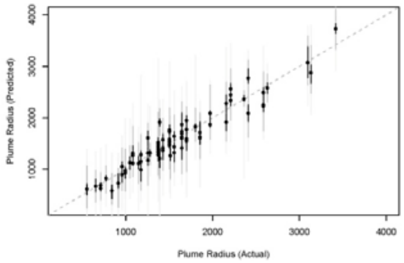

Full Model
Augmented Pairs, Ordinary Kriging
RSq=1.000
RMSE=0.000 (0.000)


Validation on LHS Design
Augmented Pairs, Ordinary Kriging
RSq=0.913
RMSE=119.444 (0.081)

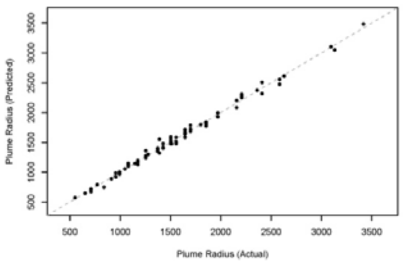

Cross−Validation
Augmented Pairs, Ordinary Kriging
RSq=0.944
RMSE=156.976 (0.098)


Full Model
Augmented Pairs, Universal Kriging (Quad)
RSq=1.000
RMSE=0.000 (0.000)

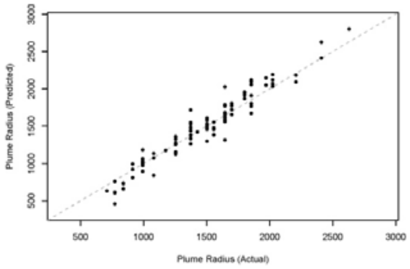

Validation on LHS Design
Augmented Pairs, Universal Kriging (Quad)
RSq=0.900
RMSE=128.198 (0.087)


Cross−Validation
Augmented Pairs, Universal Kriging (Quad)
RSq=0.909
RMSE=296.331 (0.186)

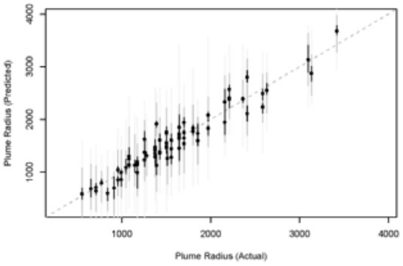
Full Model
Augmented Pairs, Quadratic
RSq=0.992
RMSE=50.874 (0.032)



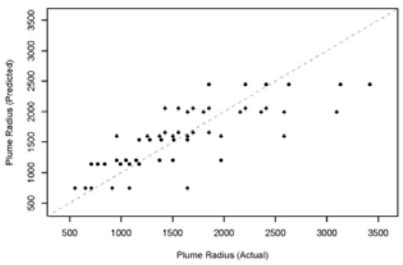
Validation on LHS Design
Augmented Pairs, Quadratic
RSq=0.900
RMSE=128.196 (0.087)



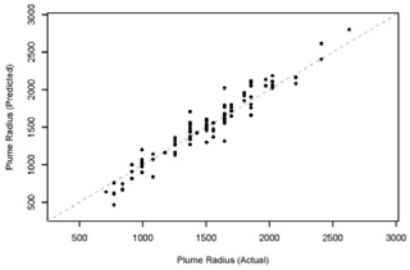
Cross–Validation
Augmented Pairs, Quadratic
RSq=0.907
RMSE=298.865 (0.187)



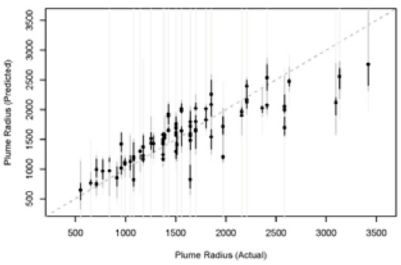
Full Model
Augmented Pairs, Quadratic LASSO
RSq=0.592
RMSE=365.888 (0.229)



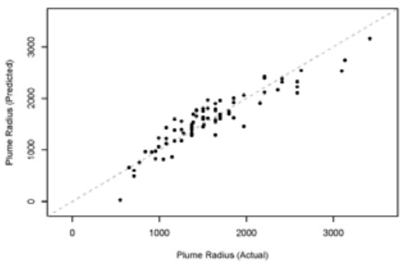
Validation on LHS Design
Augmented Pairs, Quadratic LASSO
RSq=0.901
RMSE=127.049 (0.087)



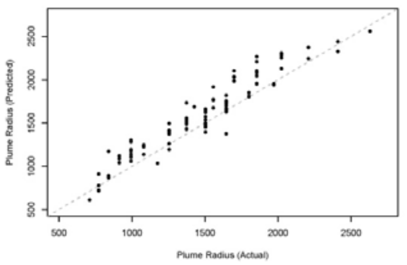
Cross–Validation
Augmented Pairs, Quadratic LASSO
RSq=0.712
RMSE=553.673 (0.347)



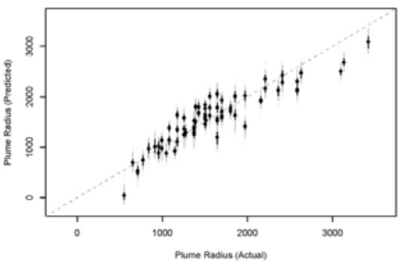
Full Model
Augmented Pairs, MARS
RSq=0.851
RMSE=220.762 (0.138)



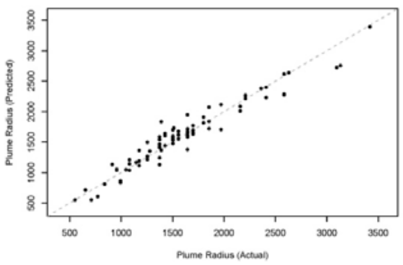
Validation on LHS Design
Augmented Pairs, MARS
RSq=0.821
RMSE=171.105 (0.117)



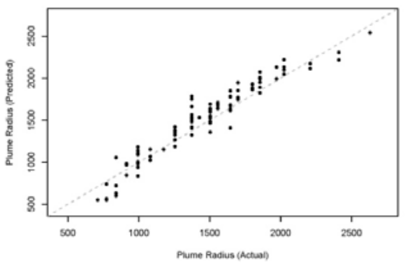
Cross–Validation
Augmented Pairs, MARS
RSq=0.836
RMSE=259.016 (0.162)



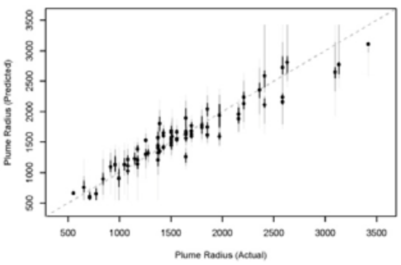
Full Model
Augmented Pairs, AVAS
RSq=0.931
RMSE=149.974 (0.094)

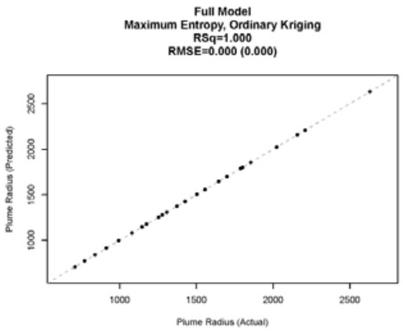


Validation on LHS Design
Augmented Pairs, AVAS
RSq=0.886
RMSE=136.832 (0.093)

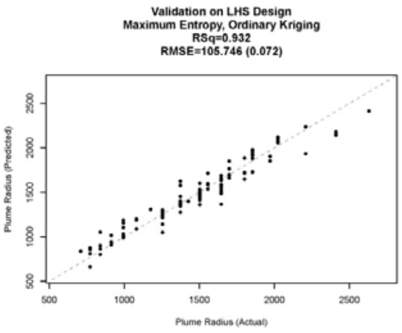


Cross–Validation
Augmented Pairs, AVAS
RSq=0.910
RMSE=215.312 (0.135)

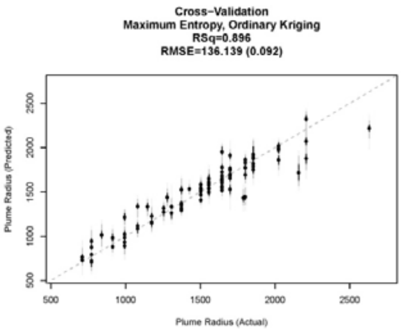




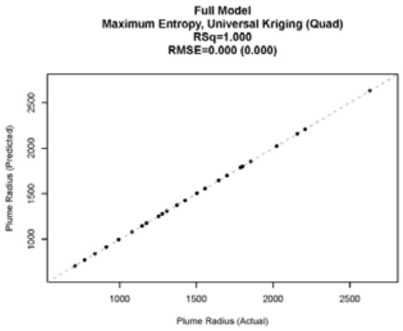
Full Model
Maximum Entropy, Ordinary Kriging
RSq=1.000
RMSE=0.000 (0.000)



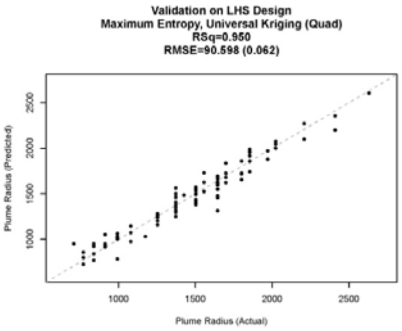
Validation on LHS Design
Maximum Entropy, Ordinary Kriging
RSq=0.932
RMSE=105.746 (0.072)



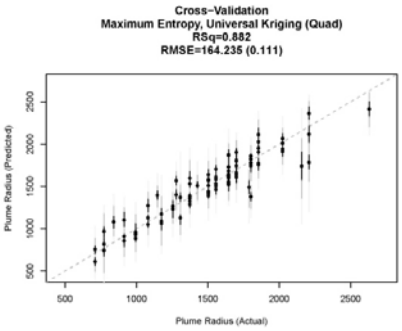
Cross−Validation
Maximum Entropy, Ordinary Kriging
RSq=0.896
RMSE=136.139 (0.092)



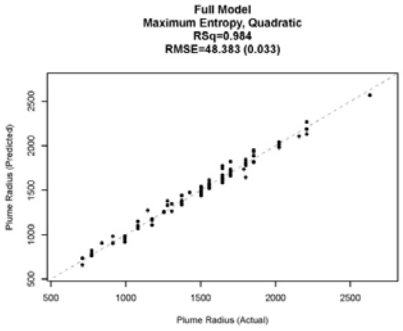
Full Model
Maximum Entropy, Universal Kriging (Quad)
RSq=1.000
RMSE=0.000 (0.000)



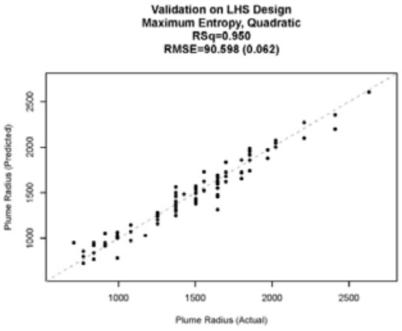
Validation on LHS Design
Maximum Entropy, Universal Kriging (Quad)
RSq=0.950
RMSE=90.598 (0.062)



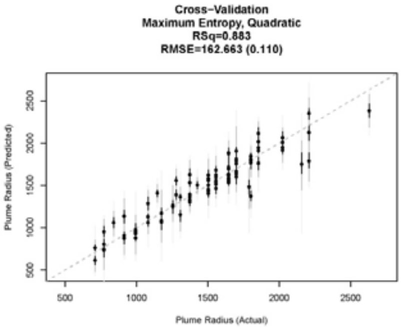
Cross−Validation
Maximum Entropy, Universal Kriging (Quad)
RSq=0.882
RMSE=164.235 (0.111)



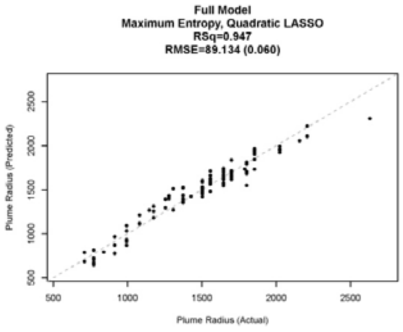
Full Model
Maximum Entropy, Quadratic
RSq=0.984
RMSE=48.383 (0.033)



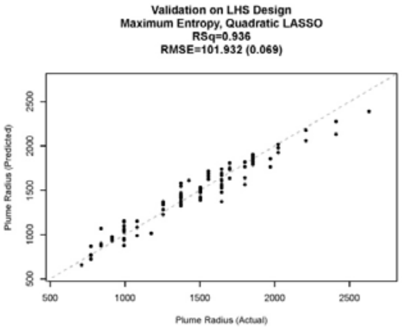
Validation on LHS Design
Maximum Entropy, Quadratic
RSq=0.950
RMSE=90.598 (0.062)



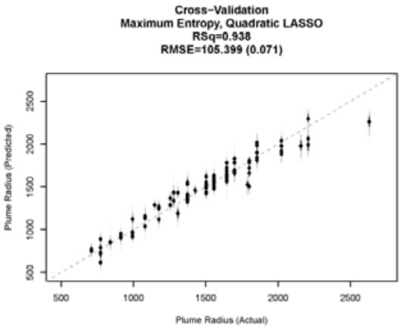
Cross−Validation
Maximum Entropy, Quadratic
RSq=0.883
RMSE=162.663 (0.110)



Full Model
Maximum Entropy, Quadratic LASSO
RSq=0.947
RMSE=89.134 (0.060)

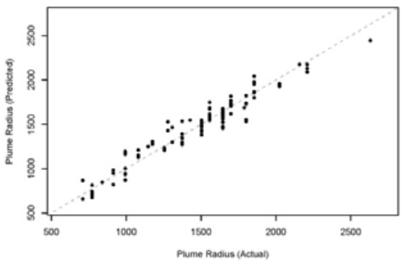


Validation on LHS Design
Maximum Entropy, Quadratic LASSO
RSq=0.936
RMSE=101.932 (0.069)

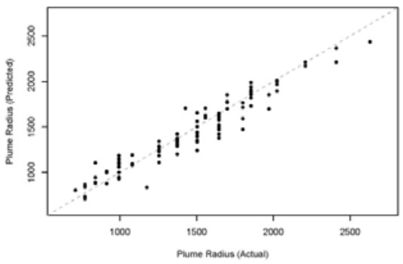


Cross−Validation
Maximum Entropy, Quadratic LASSO
RSq=0.938
RMSE=105.399 (0.071)

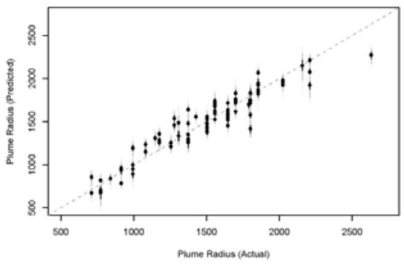
Full Model
Maximum Entropy, MARS
RSq=0.933
RMSE=100.441 (0.068)



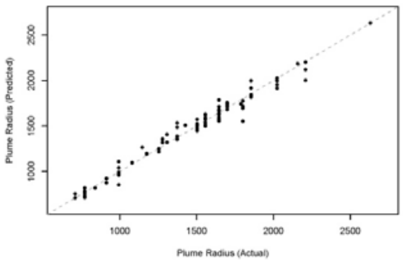
Validation on LHS Design
Maximum Entropy, MARS
RSq=0.913
RMSE=119.516 (0.081)



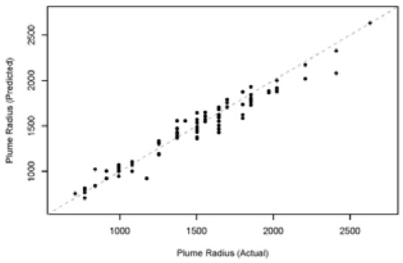
Cross−Validation
Maximum Entropy, MARS
RSq=0.904
RMSE=129.700 (0.087)



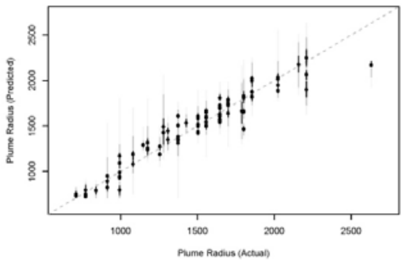
Full Model
Maximum Entropy, AVAS
RSq=0.972
RMSE=64.673 (0.044)



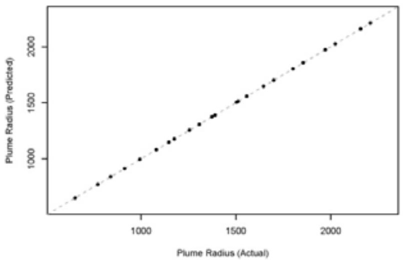
Validation on LHS Design
Maximum Entropy, AVAS
RSq=0.947
RMSE=92.741 (0.063)



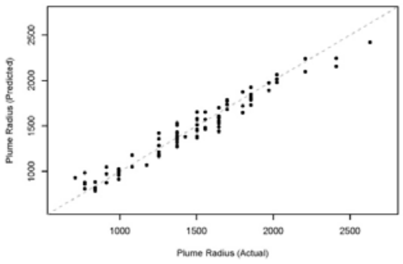
Cross−Validation
Maximum Entropy, AVAS
RSq=0.923
RMSE=127.067 (0.086)



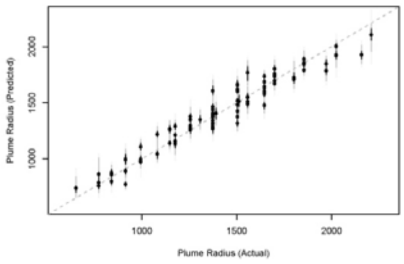
Full Model
Maximin LHS, Ordinary Kriging
RSq=1.000
RMSE=0.000 (0.000)



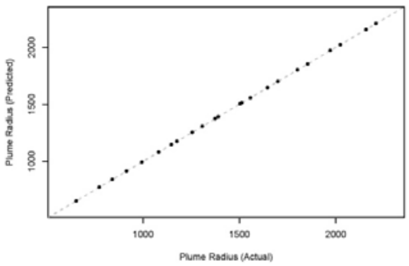
Validation on LHS Design
Maximin LHS, Ordinary Kriging
RSq=0.946
RMSE=93.654 (0.064)



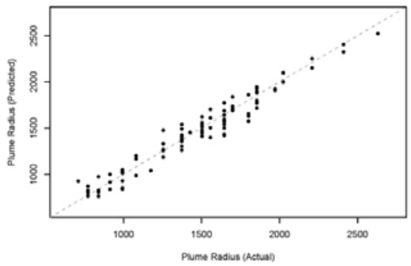
Cross−Validation
Maximin LHS, Ordinary Kriging
RSq=0.936
RMSE=96.615 (0.067)



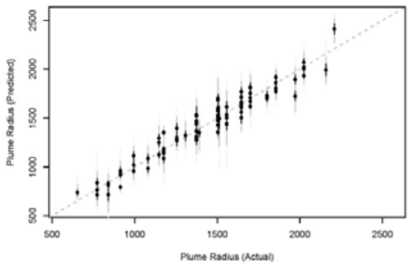
Full Model
Maximin LHS, Universal Kriging (Quad)
RSq=1.000
RMSE=0.000 (0.000)



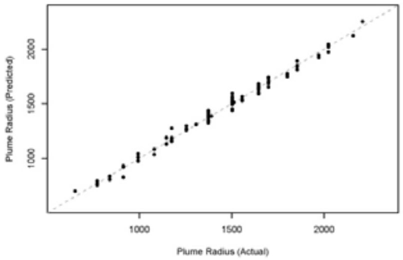
Validation on LHS Design
Maximin LHS, Universal Kriging (Quad)
RSq=0.945
RMSE=94.805 (0.065)



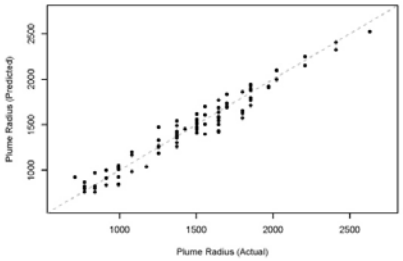
Cross−Validation
Maximin LHS, Universal Kriging (Quad)
RSq=0.935
RMSE=108.583 (0.075)



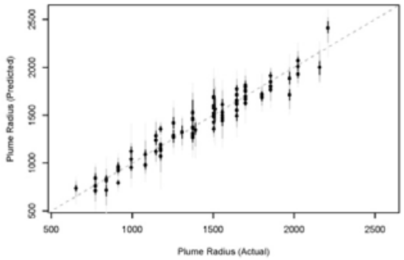
Full Model
Maximin LHS, Quadratic
RSq=0.990
RMSE=34.735 (0.024)



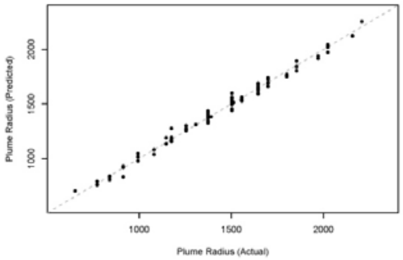
Validation on LHS Design
Maximin LHS, Quadratic
RSq=0.945
RMSE=94.805 (0.065)



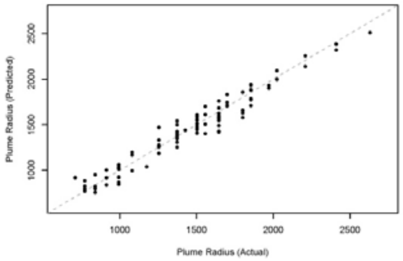
Cross−Validation
Maximin LHS, Quadratic
RSq=0.934
RMSE=110.047 (0.076)



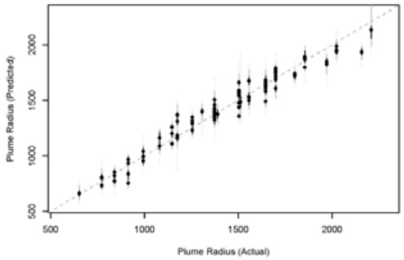
Full Model
Maximin LHS, Quadratic LASSO
RSq=0.989
RMSE=34.888 (0.024)



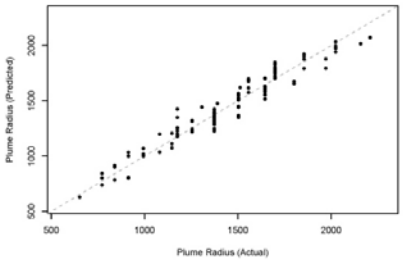
Validation on LHS Design
Maximin LHS, Quadratic LASSO
RSq=0.946
RMSE=93.990 (0.064)



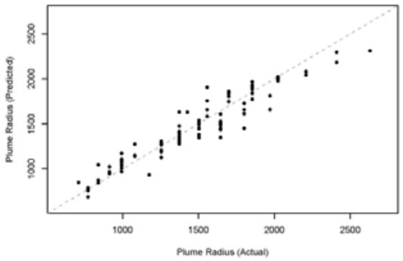
Cross−Validation
Maximin LHS, Quadratic LASSO
RSq=0.956
RMSE=82.135 (0.057)



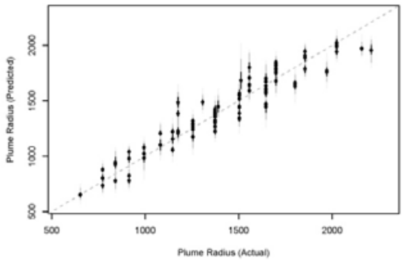
Full Model
Maximin LHS, MARS
RSq=0.937
RMSE=84.988 (0.059)



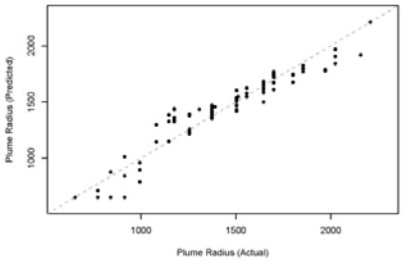
Validation on LHS Design
Maximin LHS, MARS
RSq=0.895
RMSE=131.260 (0.089)



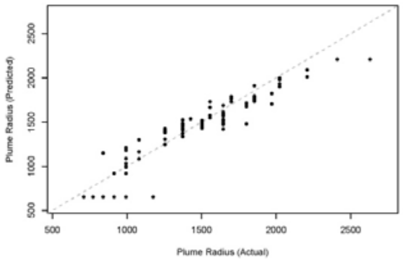
Cross−Validation
Maximin LHS, MARS
RSq=0.906
RMSE=114.009 (0.079)



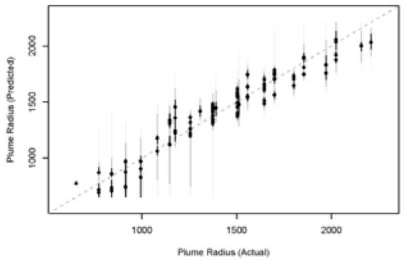
Full Model
Maximin LHS, AVAS
RSq=0.902
RMSE=106.427 (0.074)



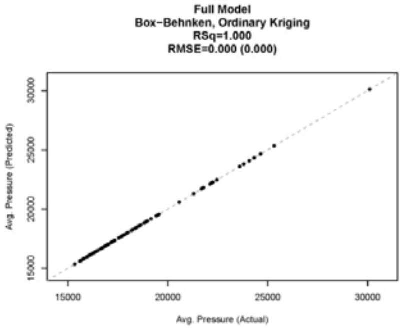
Validation on LHS Design
Maximin LHS, AVAS
RSq=0.871
RMSE=145.009 (0.099)

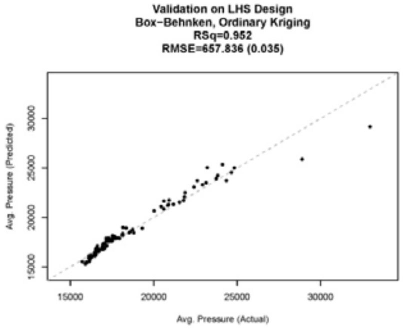


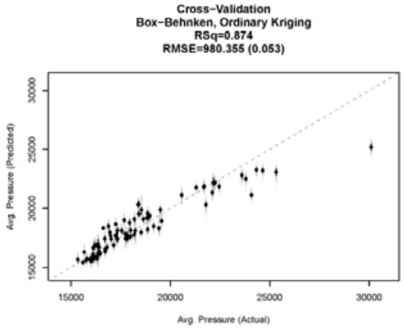
Cross−Validation
Maximin LHS, AVAS
RSq=0.925
RMSE=117.447 (0.082)

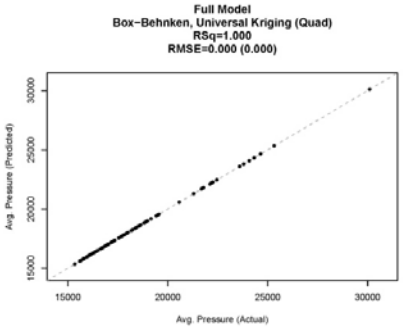


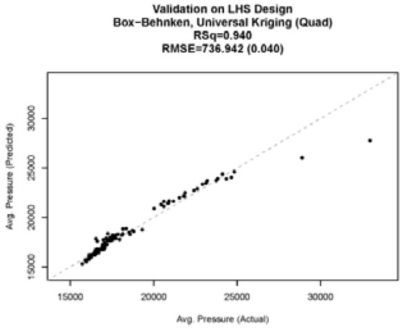
Average Pressure

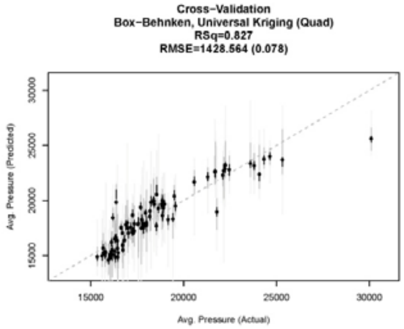


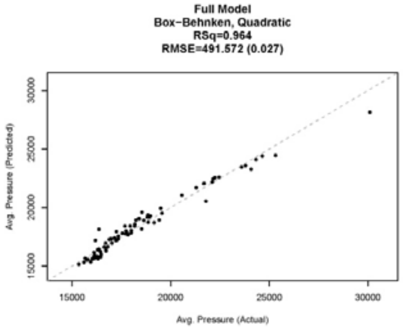


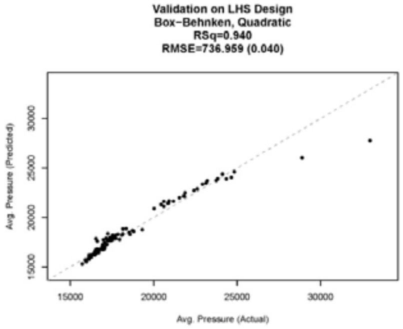


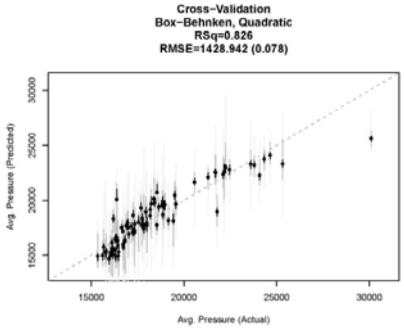


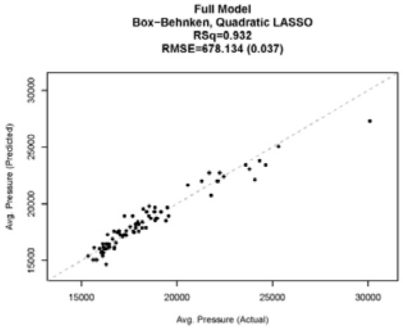


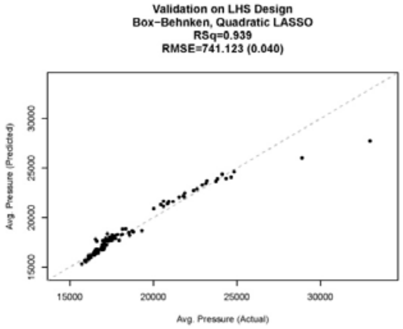


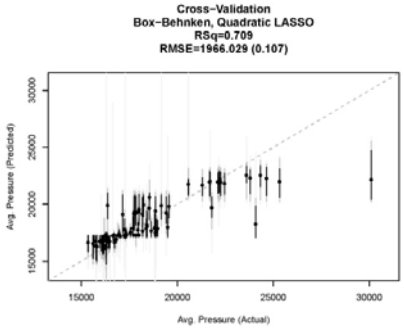




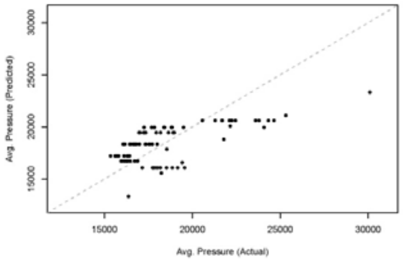




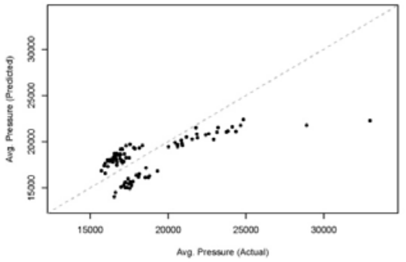




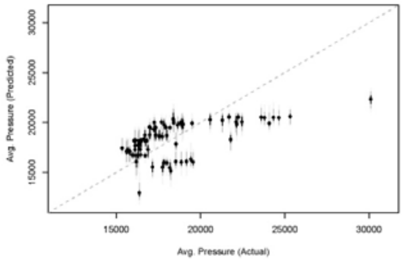
Full Model
Box−Behnken, MARS
RSq=0.435
RMSE=1960.424 (0.107)



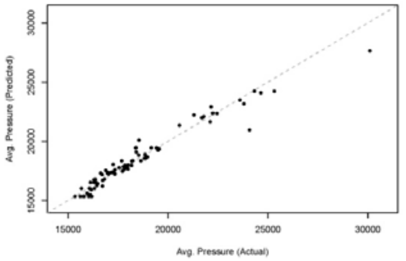
Validation on LHS Design
Box−Behnken, MARS
RSq=0.464
RMSE=2204.270 (0.118)



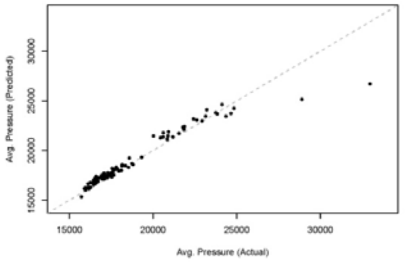
Cross−Validation
Box−Behnken, MARS
RSq=0.347
RMSE=2158.447 (0.118)



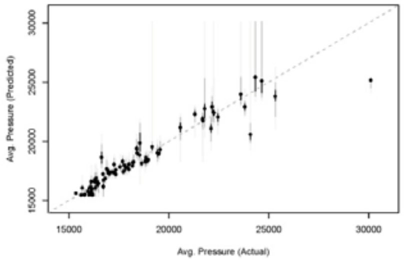
Full Model
Box−Behnken, AVAS
RSq=0.947
RMSE=598.093 (0.033)



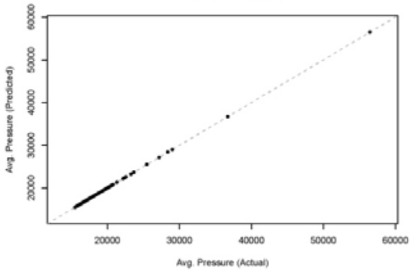
Validation on LHS Design
Box−Behnken, AVAS
RSq=0.916
RMSE=871.533 (0.047)



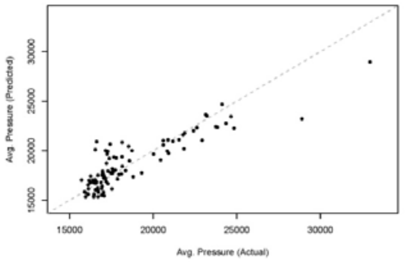
Cross−Validation
Box−Behnken, AVAS
RSq=0.904
RMSE=946.230 (0.052)



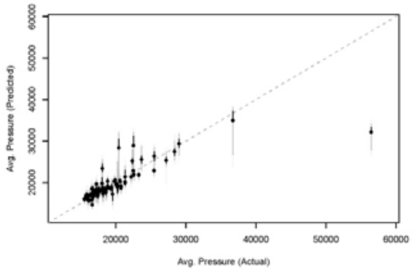
Full Model
Augmented Pairs, Ordinary Kriging
RSq=1.000
RMSE=0.000 (0.000)



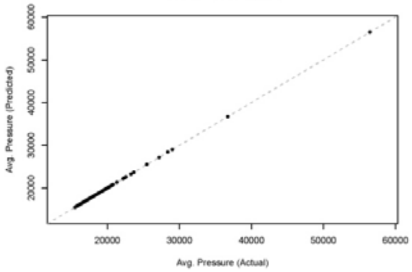
Validation on LHS Design
Augmented Pairs, Ordinary Kriging
RSq=0.779
RMSE=1414.982 (0.076)



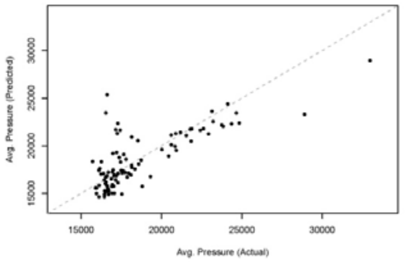
Cross−Validation
Augmented Pairs, Ordinary Kriging
RSq=0.655
RMSE=3419.664 (0.173)



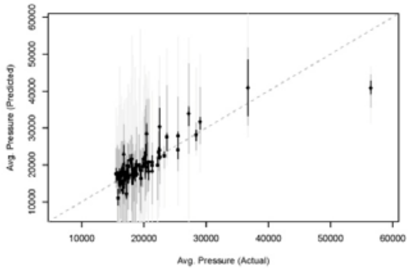
Full Model
Augmented Pairs, Universal Kriging (Quad)
RSq=1.000
RMSE=0.000 (0.000)

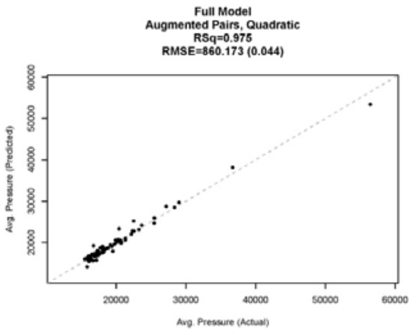


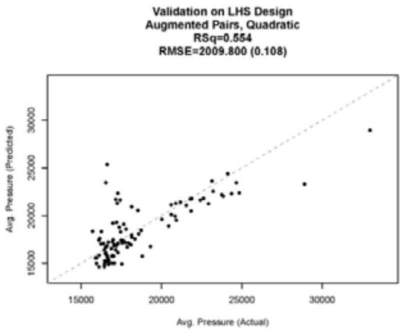
Validation on LHS Design
Augmented Pairs, Universal Kriging (Quad)
RSq=0.555
RMSE=2008.437 (0.108)

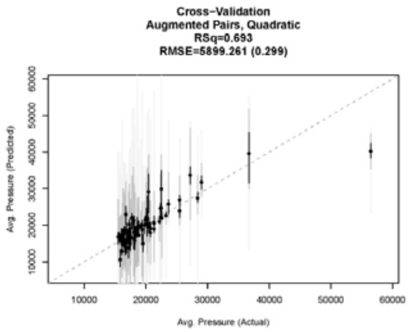


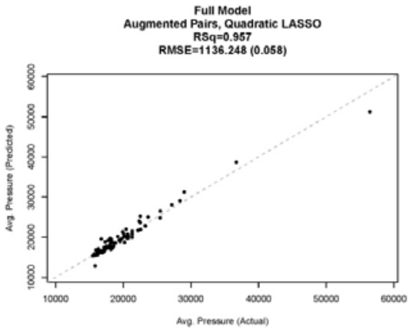
Cross−Validation
Augmented Pairs, Universal Kriging (Quad)
RSq=0.699
RMSE=5708.371 (0.289)

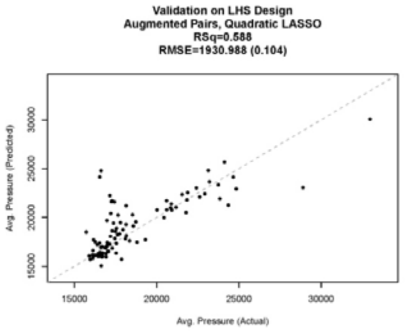


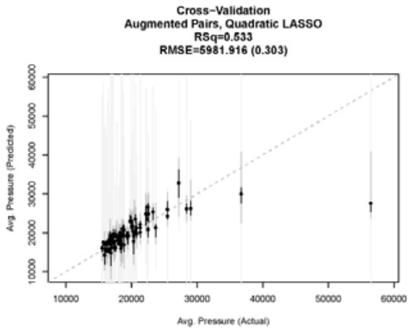

Full Model
Augmented Pairs, Quadratic
RSq=0.975
RMSE=860.173 (0.044)

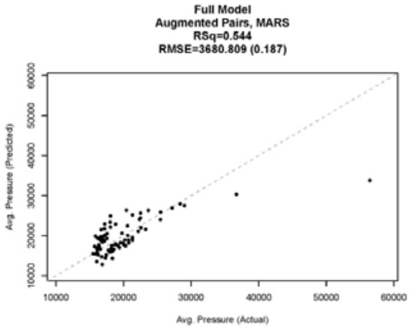

Validation on LHS Design
Augmented Pairs, Quadratic
RSq=0.554
RMSE=2009.800 (0.108)

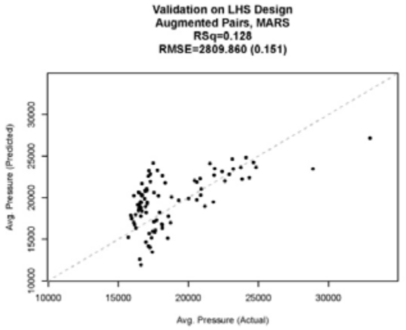

Cross−Validation
Augmented Pairs, Quadratic
RSq=0.693
RMSE=5899.261 (0.299)

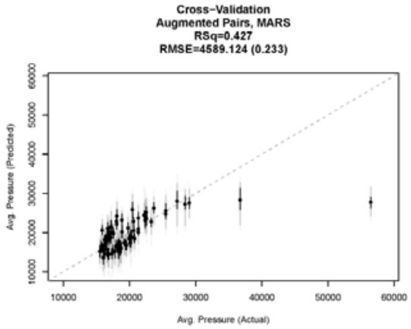

Full Model
Augmented Pairs, Quadratic LASSO
RSq=0.957
RMSE=1136.248 (0.058)

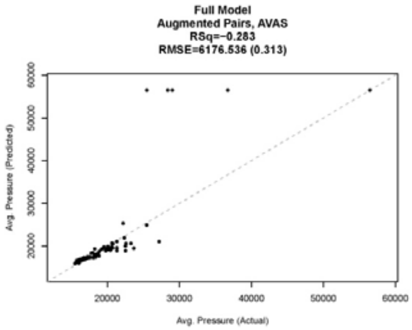

Validation on LHS Design
Augmented Pairs, Quadratic LASSO
RSq=0.588
RMSE=1930.988 (0.104)

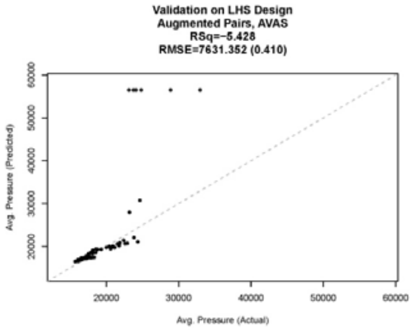

Cross−Validation
Augmented Pairs, Quadratic LASSO
RSq=0.533
RMSE=5981.916 (0.303)

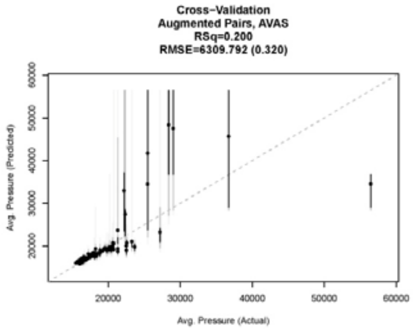

Full Model
Augmented Pairs, MARS
RSq=0.544
RMSE=3680.809 (0.187)

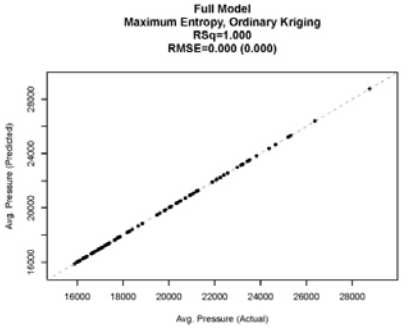

Validation on LHS Design
Augmented Pairs, MARS
RSq=0.128
RMSE=2809.860 (0.151)


Cross−Validation
Augmented Pairs, MARS
RSq=0.427
RMSE=4589.124 (0.233)

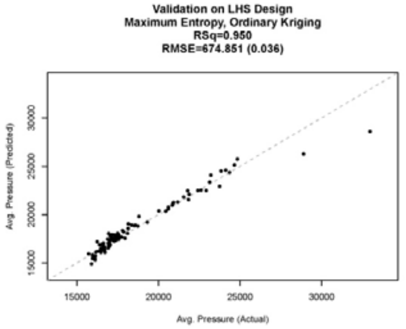

Full Model
Augmented Pairs, AVAS
RSq=−0.283
RMSE=6176.536 (0.313)


Validation on LHS Design
Augmented Pairs, AVAS
RSq=−5.428
RMSE=7631.352 (0.410)

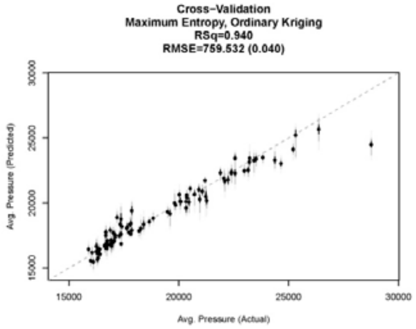

Cross−Validation
Augmented Pairs, AVAS
RSq=0.200
RMSE=6309.792 (0.320)



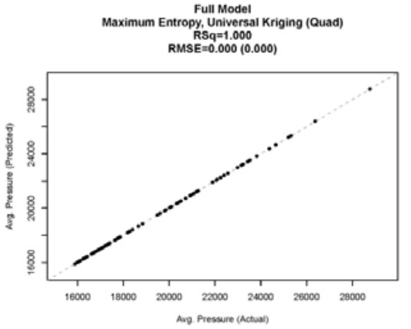
Full Model
Maximum Entropy, Ordinary Kriging
RSq=1.000
RMSE=0.000 (0.000)



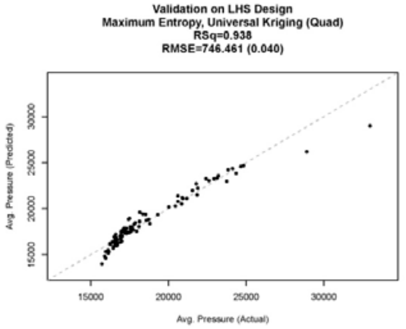
Validation on LHS Design
Maximum Entropy, Ordinary Kriging
RSq=0.950
RMSE=674.851 (0.036)



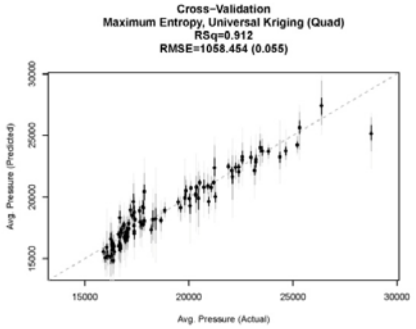
Cross−Validation
Maximum Entropy, Ordinary Kriging
RSq=0.940
RMSE=759.532 (0.040)



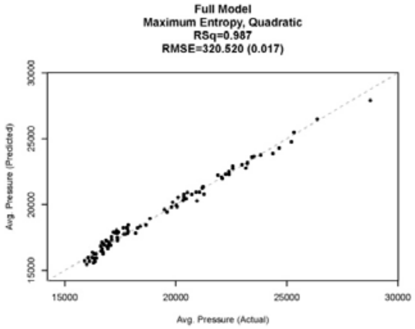
Full Model
Maximum Entropy, Universal Kriging (Quad)
RSq=1.000
RMSE=0.000 (0.000)



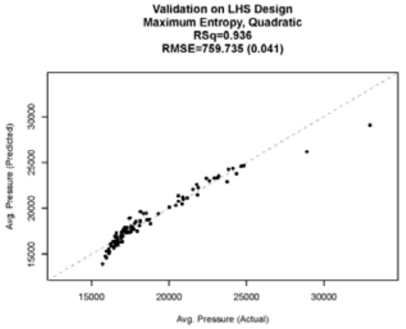
Validation on LHS Design
Maximum Entropy, Universal Kriging (Quad)
RSq=0.938
RMSE=746.461 (0.040)



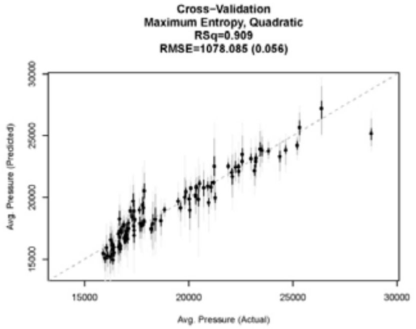
Cross−Validation
Maximum Entropy, Universal Kriging (Quad)
RSq=0.912
RMSE=1058.454 (0.055)



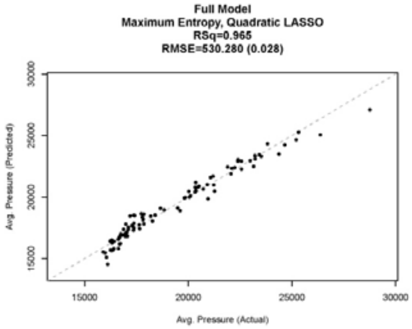
Full Model
Maximum Entropy, Quadratic
RSq=0.987
RMSE=320.520 (0.017)



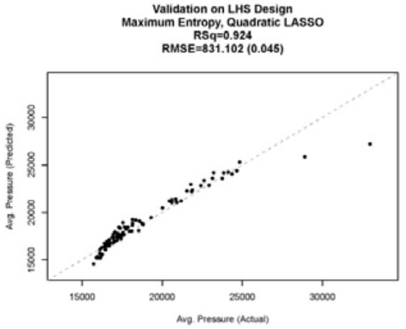
Validation on LHS Design
Maximum Entropy, Quadratic
RSq=0.936
RMSE=759.735 (0.041)



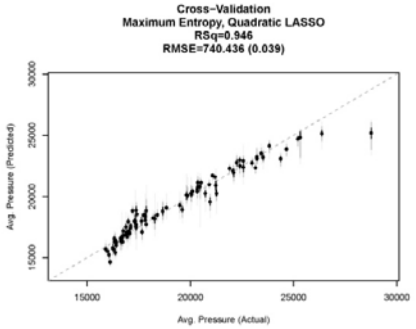
Cross−Validation
Maximum Entropy, Quadratic
RSq=0.909
RMSE=1078.085 (0.056)



Full Model
Maximum Entropy, Quadratic LASSO
RSq=0.965
RMSE=530.280 (0.028)

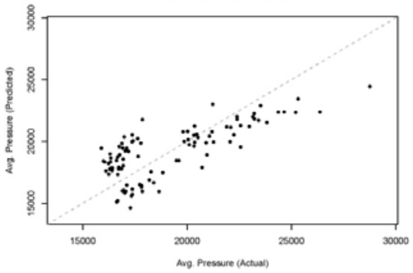


Validation on LHS Design
Maximum Entropy, Quadratic LASSO
RSq=0.924
RMSE=831.102 (0.045)

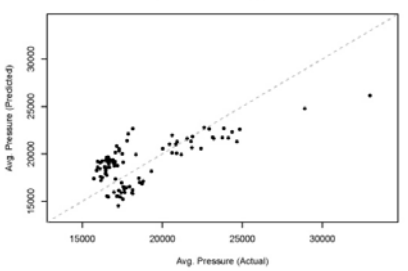


Cross−Validation
Maximum Entropy, Quadratic LASSO
RSq=0.946
RMSE=740.436 (0.039)

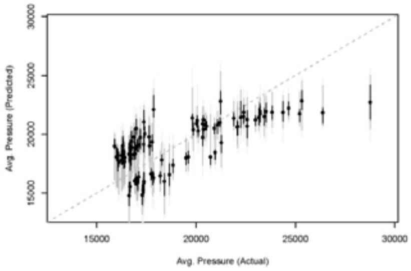
Full Model
Maximum Entropy, MARS
RSq=0.570
RMSE=1869.291 (0.097)



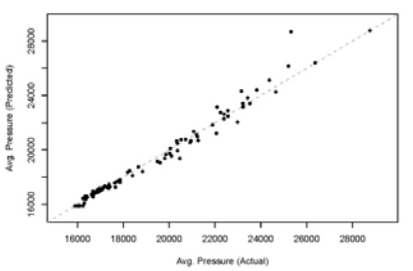
Validation on LHS Design
Maximum Entropy, MARS
RSq=0.458
RMSE=2215.020 (0.119)



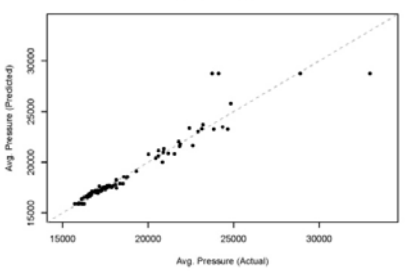
Cross-Validation
Maximum Entropy, MARS
RSq=0.535
RMSE=2190.138 (0.114)



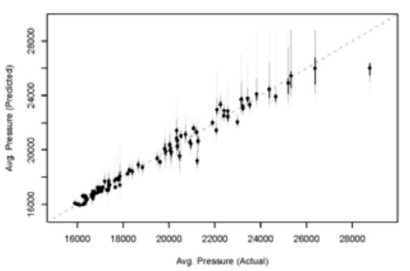
Full Model
Maximum Entropy, AVAS
RSq=0.971
RMSE=485.311 (0.025)



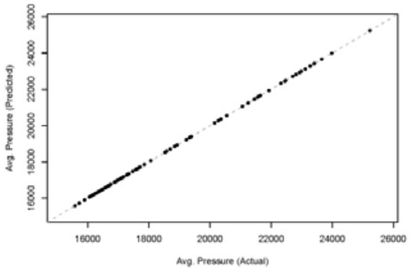
Validation on LHS Design
Maximum Entropy, AVAS
RSq=0.913
RMSE=890.089 (0.048)



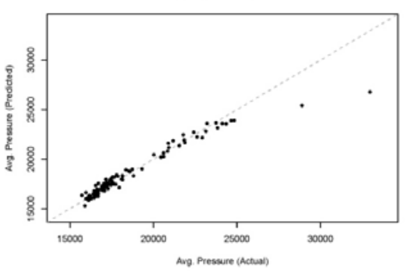
Cross-Validation
Maximum Entropy, AVAS
RSq=0.971
RMSE=629.513 (0.033)



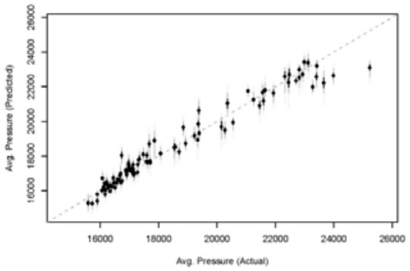
Full Model
Maximin LHS, Ordinary Kriging
RSq=1.000
RMSE=0.000 (0.000)



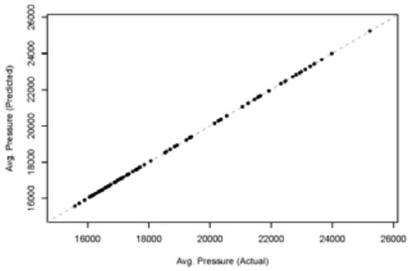
Validation on LHS Design
Maximin LHS, Ordinary Kriging
RSq=0.925
RMSE=824.285 (0.044)



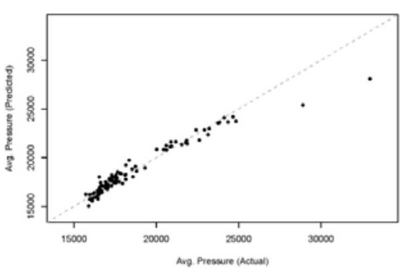
Cross-Validation
Maximin LHS, Ordinary Kriging
RSq=0.959
RMSE=564.617 (0.031)



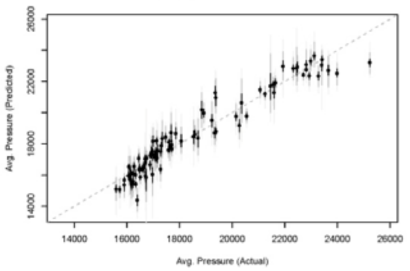
Full Model
Maximin LHS, Universal Kriging (Quad)
RSq=1.000
RMSE=0.000 (0.000)



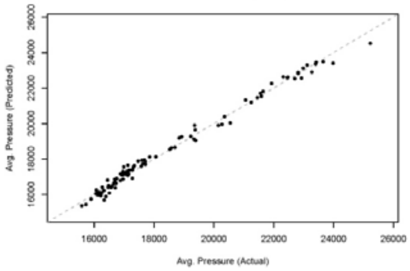
Validation on LHS Design
Maximin LHS, Universal Kriging (Quad)
RSq=0.933
RMSE=780.808 (0.042)



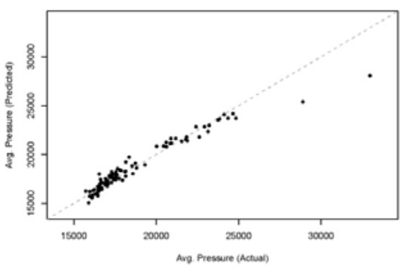
Cross-Validation
Maximin LHS, Universal Kriging (Quad)
RSq=0.928
RMSE=841.891 (0.046)



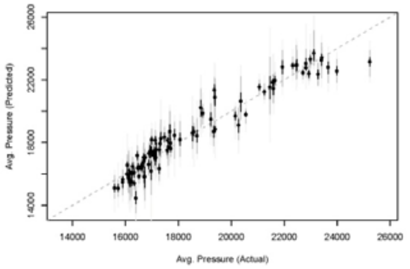
Full Model
Maximin LHS, Quadratic
RSq=0.990
RMSE=254.650 (0.014)



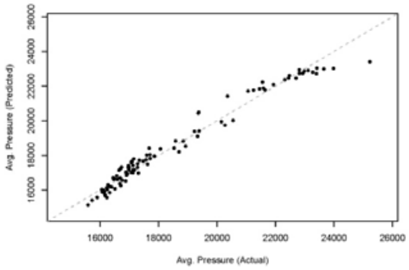
Validation on LHS Design
Maximin LHS, Quadratic
RSq=0.932
RMSE=783.426 (0.042)



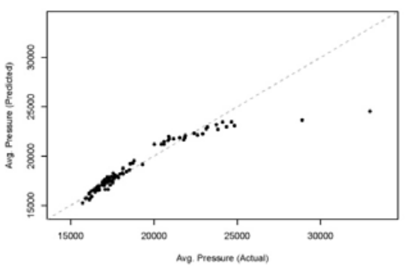
Cross−Validation
Maximin LHS, Quadratic
RSq=0.929
RMSE=840.182 (0.045)



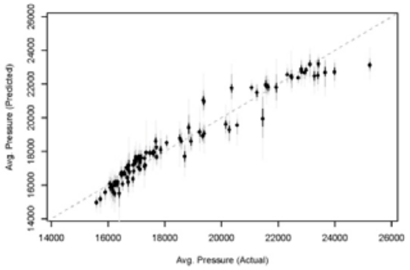
Full Model
Maximin LHS, Quadratic LASSO
RSq=0.970
RMSE=438.557 (0.024)



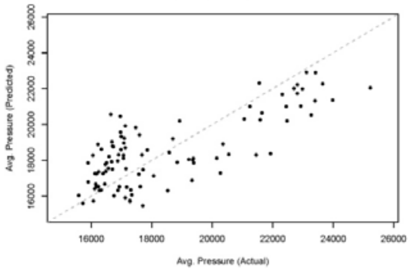
Validation on LHS Design
Maximin LHS, Quadratic LASSO
RSq=0.862
RMSE=1119.999 (0.060)



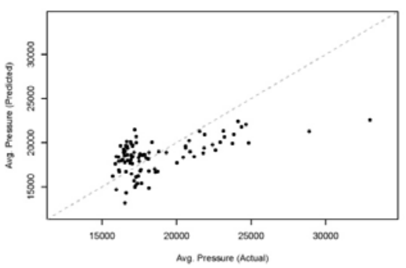
Cross−Validation
Maximin LHS, Quadratic LASSO
RSq=0.945
RMSE=657.602 (0.036)



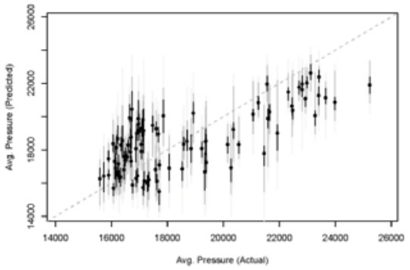
Full Model
Maximin LHS, MARS
RSq=0.579
RMSE=1631.922 (0.088)



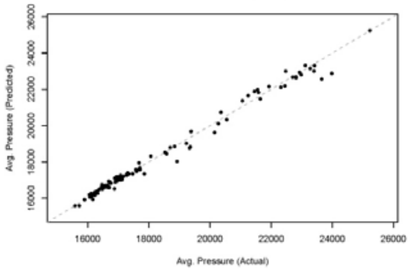
Validation on LHS Design
Maximin LHS, MARS
RSq=0.318
RMSE=2485.536 (0.134)



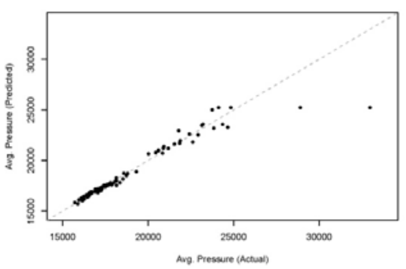
Cross−Validation
Maximin LHS, MARS
RSq=0.525
RMSE=1963.522 (0.106)



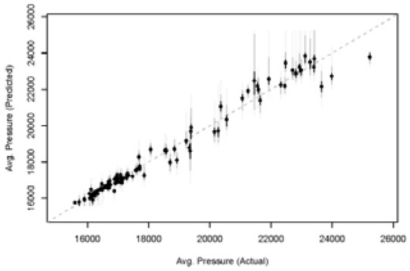
Full Model
Maximin LHS, AVAS
RSq=0.988
RMSE=274.782 (0.015)



Validation on LHS Design
Maximin LHS, AVAS
RSq=0.903
RMSE=937.737 (0.050)



Cross−Validation
Maximin LHS, AVAS
RSq=0.973
RMSE=509.487 (0.028)


Avg. Pressure (Predicted)
Avg. Pressure (Actual)