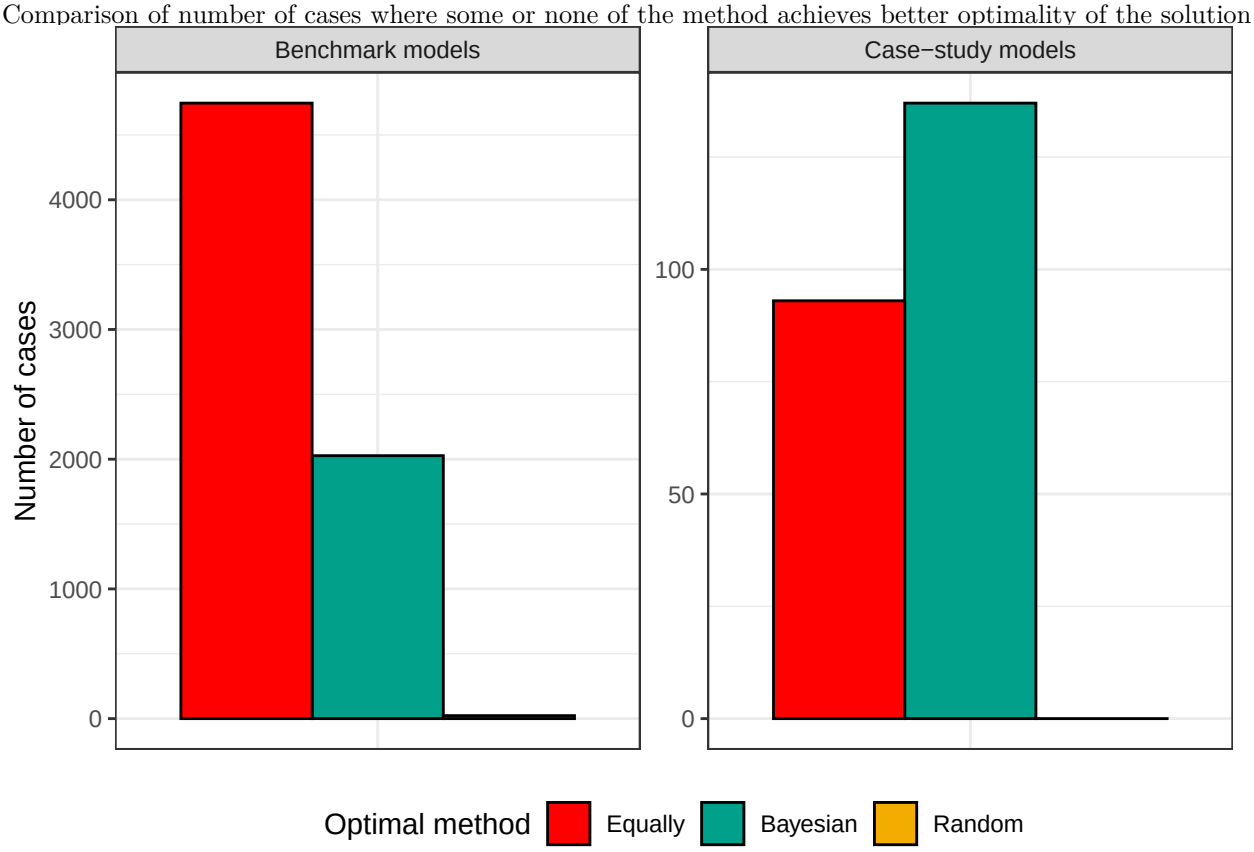
Supplementary Materials: Analysis of stochastic methods for options generation

The analysis includes stochastic methods for option generation from DEX models, and in particular a comparison of Bayesian Optimisation-based (below referred to as Bayesian and abbreviated with BO) method with a random search (below referred to as Random and abbreviated with RS). Both methods are evaluated on benchmark (mock) DEX models and a model from a case-study on modeling primary productivity as a soil function. Performance of the applied method is expressed as quantities related to a time of first optimal discovery, and number of optimal solutions found (optimal set size), for which a statistical analysis is performed.

1 Compare the optimality of solution sets

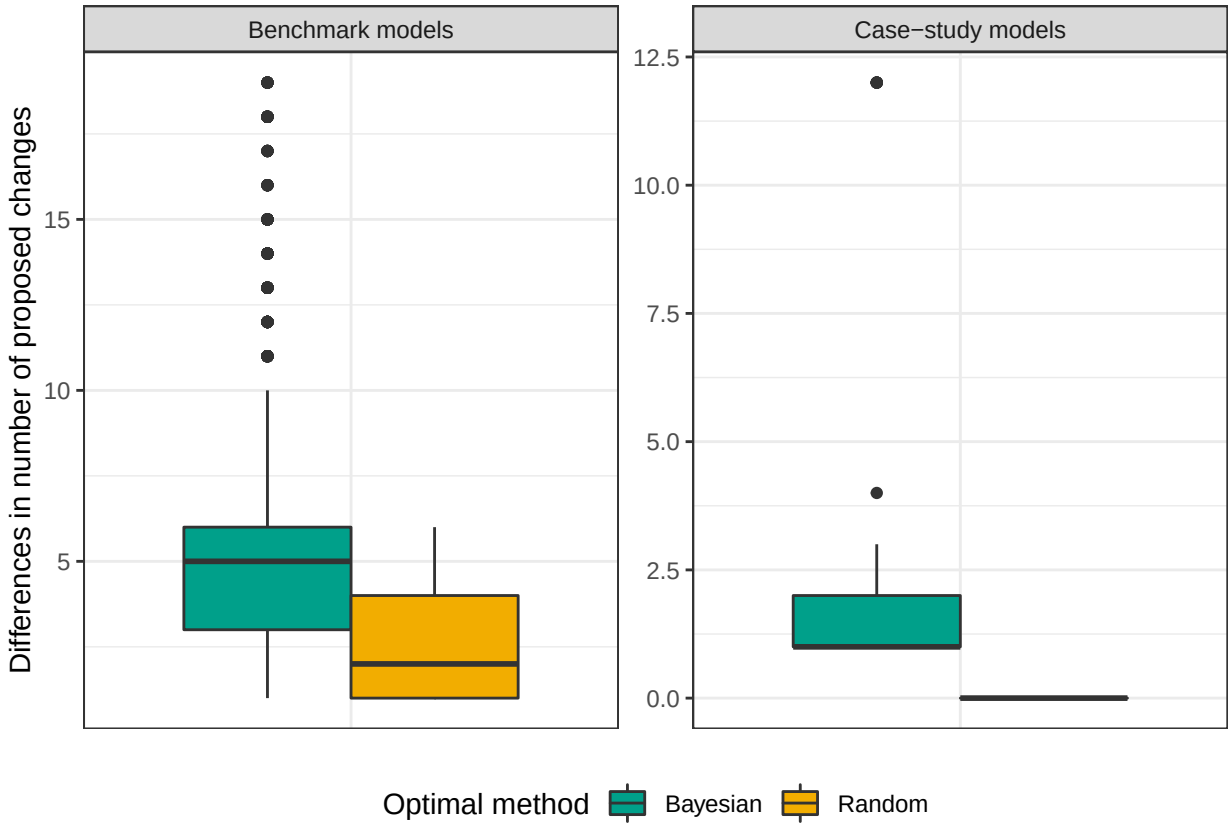
Comparison of number of cases where some or none of the method achieves better optimality of the solution



The above figure shows that for the benchmark models, in 4745 cases (out of 6795) both Bayesian and Random methods returned solutions with equal number of proposed changes. In 2027 cases, the method using Bayesian Optimisation had solutions with less number of changes, while in only 23 cases the Random method had better solutions. Unlike the benchmark models, in case of the case-study models, the method using Bayesian Optimisation dominates (137/230) by the number of cases, in which it resulted in better

solution, i.e. solutions with less number of changes compared to the Random method. In the rest 93 cases the proposed solutions demanded equal number of changes, for both methods.

Dominance of the method using Bayesian Optimisation can be observed in terms of absulte difference of required chnages in the proposed solutions, when they differ. As can be seen below, when Bayesian Optimisation method has better solutions, it usually has drastically lower number of required changes, as compared to the Random method. This is in particularly visible among benchmark models.



2 Statistical analysis

Statistical analyses are performed to test significance of achieved performance per model characteristics. For benchmark models, the models are investigated in terms of the model' weights and model's depth, while for the case-study model, overall test of statistical significance is performed. For that purpose a set of statistical test are used, as follows:

- *Shapiro-Wilk test* - for testing the normality of value distributions
- *Brown–Forsythe Levene's* or *Kruskal test* - for testing the samples' homoscedasticity (the selection depends on the test of normality)
- *T-test* or *Mann-Whitney U Test* - for testing the significance of differences (T-test is used in case of observed normality and homoscedasticity, otherwise Mann-Whitney U Test)

The significance level is set to 0.05 for all tests.

The performance is expressed in terms of **Time to first optimal solution** and **Size of the optimal solution/option set**. During the statistical analysis, quantities from both measures were transformed using normalization (maximum - minimum) and log transformation.

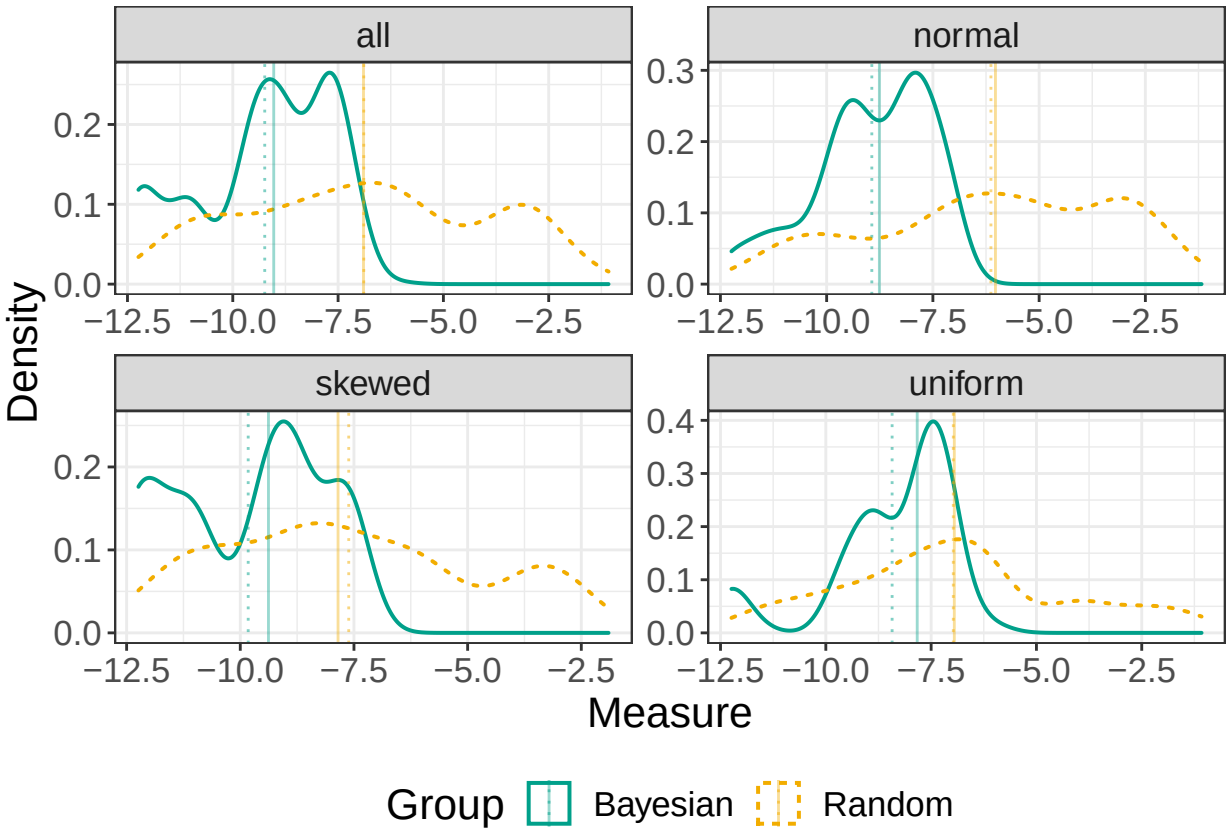
2.1 Benchmark (mock) models

2.1.1 Time to first optimal solution

Statistical analysis of methods' performance achieved in terms of the time to first optimal solution reveals that the method based on Bayesian Optimisation stochastically dominates the random search, i.e., significantly dominates across the whole range of possible values. This is observable in general, across all models, but also if individual type of models is considered (e.g., models with normal, skewed or uniform distribution). Comparatively, the Bayesian Optimisation-based method has less dominance in the problems described with uniform distribution of models' weights, compared to the other two. This is explained if the solution space (set of possible solutions) is considered, which is largest for the uniformly distributed weights, and smallest in the case of skewed distributions of the models' weights.

Table S1: Statistical test of normality of, homoscedasticity and difference between values for time to first solutions of both Bayesian and Random methods. Reported values are p-values, with significance reported by the asterisk character (*). The difference is tested with the test given in brackets and presented with a null hypothesis stating that of both samples are sampled from same distribution. Vertical solid and dotted lines represent the sample's median and mean, respectively.

Model weight	Normality (BO)	Normality (RS)	Homoscedasticity (BO vs. RS)	Sample distribution (BO vs. RS)
all	0 *	0 *	(Kruskal-Wallis) 0 *	(Mann-Whitney U) 0 *
normal	0 *	0 *	(Kruskal-Wallis) 0 *	(Mann-Whitney U) 0 *
skewed	0 *	0 *	(Kruskal-Wallis) 0 *	(Mann-Whitney U) 0 *
uniform	0 *	0.0497 *	(Kruskal-Wallis) 0 *	(Mann-Whitney U) 1e-04 *

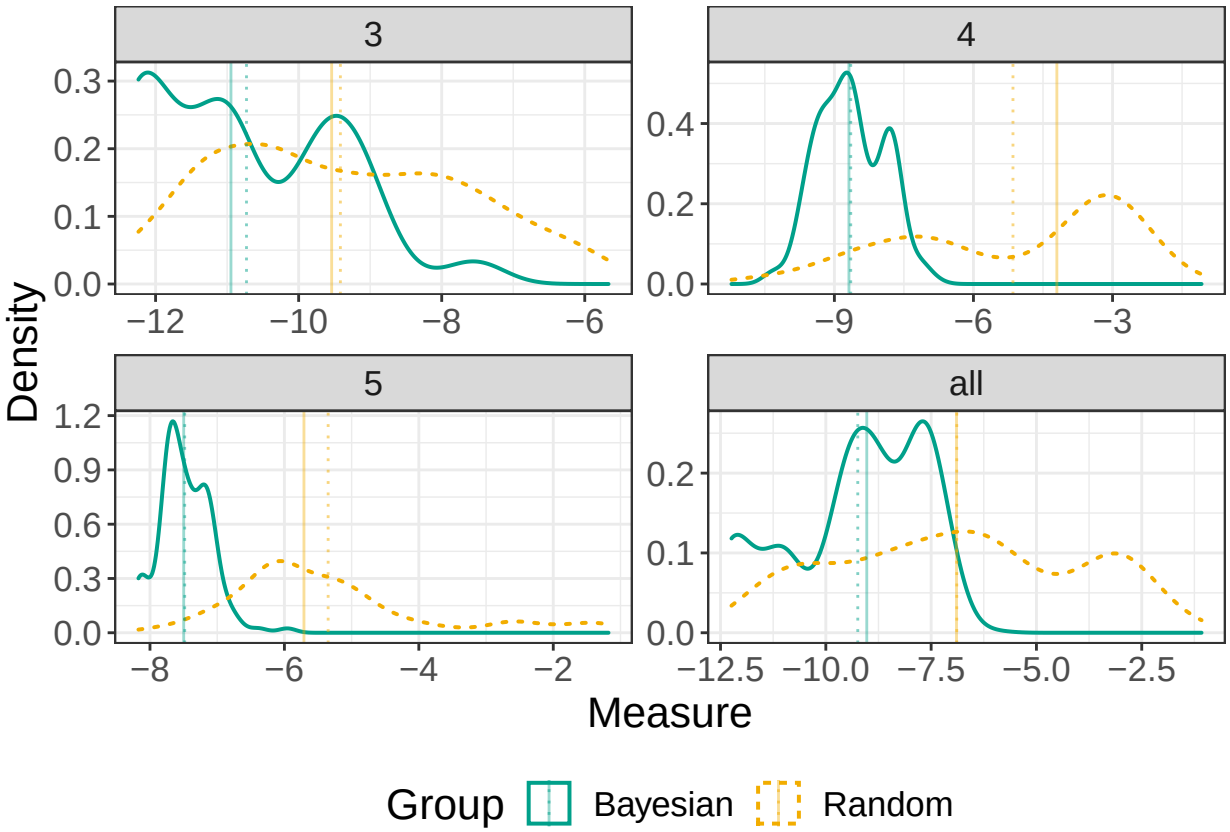


Similarly, significant stochastic dominance is observed if the performance is compared in terms of other property of the models, i.e., models' depth. The visual analysis reveals that as the depth increases, the

dominances become more significant, which is confirmed by the statistical analisys.

Table S2: Statistical test of normality of, homoscedasticity and difference between values for time to first solutions of both Bayesian and Random methods. Reported values are p-values, with significance reported by the asterisk character (*). The difference is tested with the test given in brackets and presented with a null hypothesis stating that of both samples are sampled from same distribution. Vertical solid and dotted lines represent the sample's median and mean, respectively.

Model weight	Normality (BO)	Normality (RS)	Homoscedasticity (BO vs. RS)	Sample distribution (BO vs. RS)
3	0 *	0 *	(Kruskal-Wallis) 0 *	(Mann-Whitney U) 0 *
4	0.0028 *	0 *	(Kruskal-Wallis) 0 *	(Mann-Whitney U) 0 *
5	2e-04 *	0 *	(Kruskal-Wallis) 0 *	(Mann-Whitney U) 0 *
all	0 *	0 *	(Kruskal-Wallis) 0 *	(Mann-Whitney U) 0 *



2.1.2 Optimal set size

Unlike the analysis of performance in accordance to time to a first solution, the achieved advances of the Bayesian are less significant compared to Random, when the size of the optimal set is compared. Namely, an overall significant stochastic dominance is observed, as shown below in Table 3 and Table 4, but such dominance is not observed over models with uniformly distributed weights and models with depth of 3. The former is explained by the overall size of the candidate set, which is largest in case of uniformly distributed weifghts, and the fact that with the given number of iterations, Random search is capable of finding many options/solutions. This is in particularly true for models of depth 3 that are characterised with relatively small alternative spaces, and hence with those models, both methods achieve comparable results.

Table S3: Statistical test of normality of, homoscedasticity and difference between values for the size of solution sets of both Bayesian and Random methods. Reported values are p-values, with significance reported by the asterisk character (*). The difference is tested with the test given in brackets and presented with a null hypothesis stating that of both samples are sampled from same distribution. Vertical solid and dotted lines represent the sample's median and mean, respectively.

Model weight	Normality (BO)	Normality (RS)	Homoscedasticity (BO vs. RS)	Sample distribution (BO vs. RS)
all	0 *	0 *	(Kruskal-Wallis) 0 *	(Mann-Whitney U) 0 *
normal	0 *	0 *	(Kruskal-Wallis) 0 *	(Mann-Whitney U) 0 *
skewed	0 *	0 *	(Kruskal-Wallis) 0.27	(Mann-Whitney U) 0 *
uniform	0 *	0 *	(Kruskal-Wallis) 0.678	(Mann-Whitney U) 0.0689

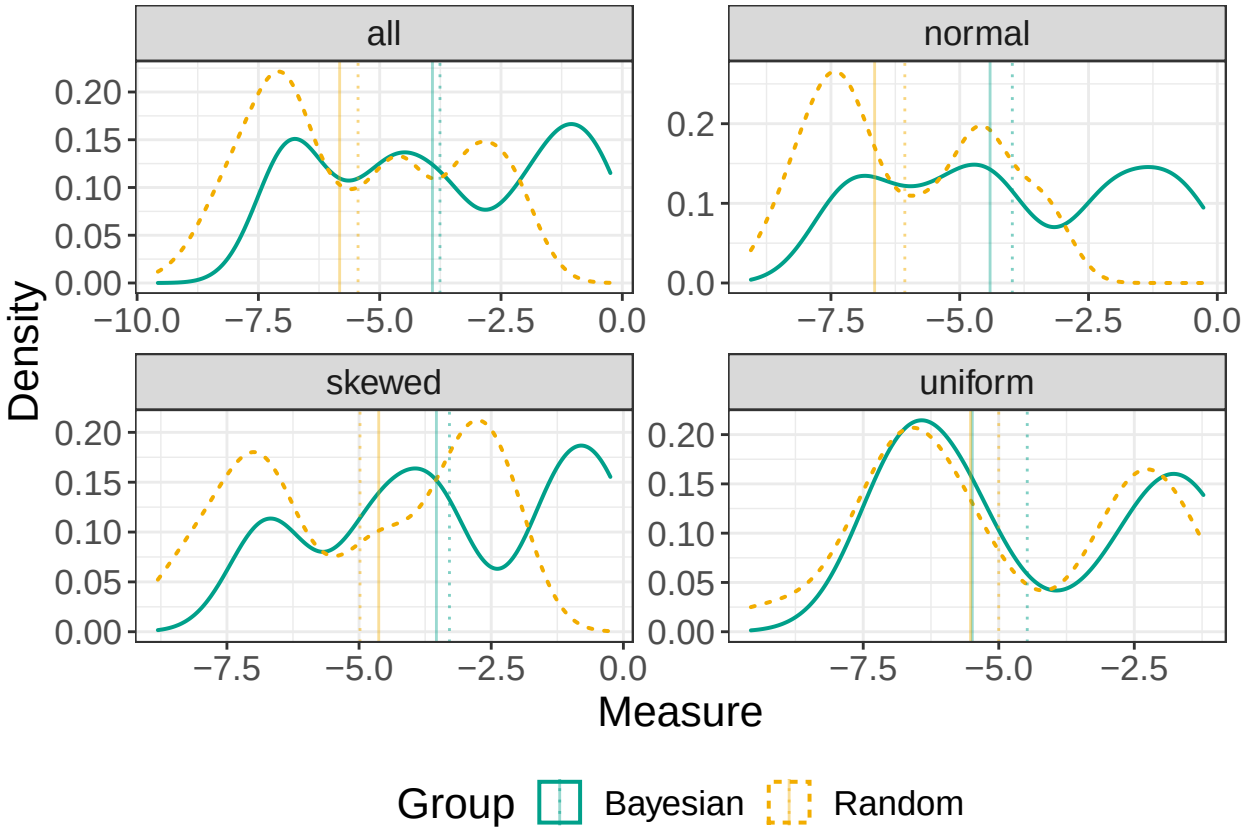
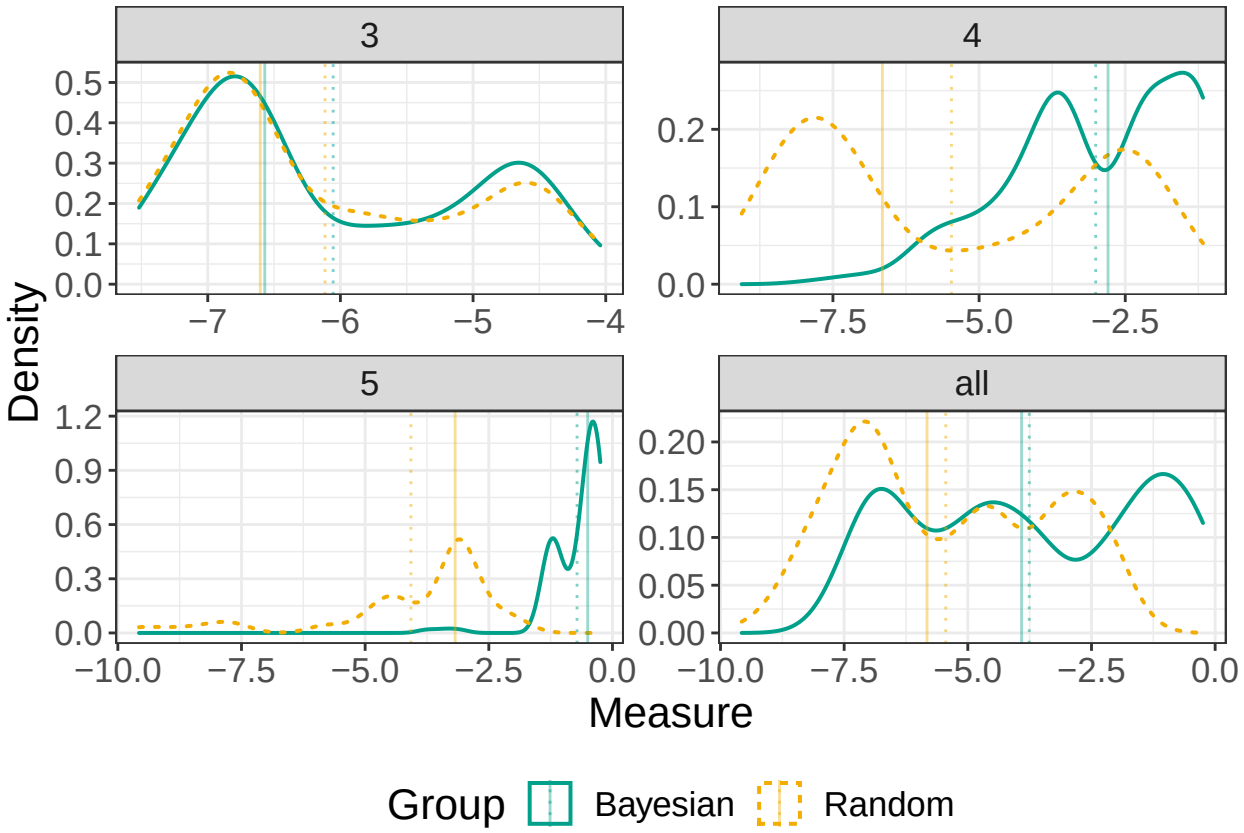


Table S4: Statistical test of normality of, homoscedasticity and difference between values for the size of solution sets of both Bayesian and Random methods. Reported values are p-values, with significance reported by the asterisk character (*). The difference is tested with the test given in brackets and presented with a null hypothesis stating that of both samples are sampled from same distribution. Vertical solid and dotted lines represent the sample's median and mean, respectively.

Model weight	Normality (BO)	Normality (RS)	Homoscedasticity (BO vs. RS)	Sample distribution (BO vs. RS)
3	0 *	0 *	(Kruskal-Wallis) 0.893	(Mann-Whitney U) 0.6011
4	0 *	0 *	(Kruskal-Wallis) 0 *	(Mann-Whitney U) 0 *
5	0 *	0 *	(Kruskal-Wallis) 0.026 *	(Mann-Whitney U) 0 *
all	0 *	0 *	(Kruskal-Wallis) 0 *	(Mann-Whitney U) 0 *



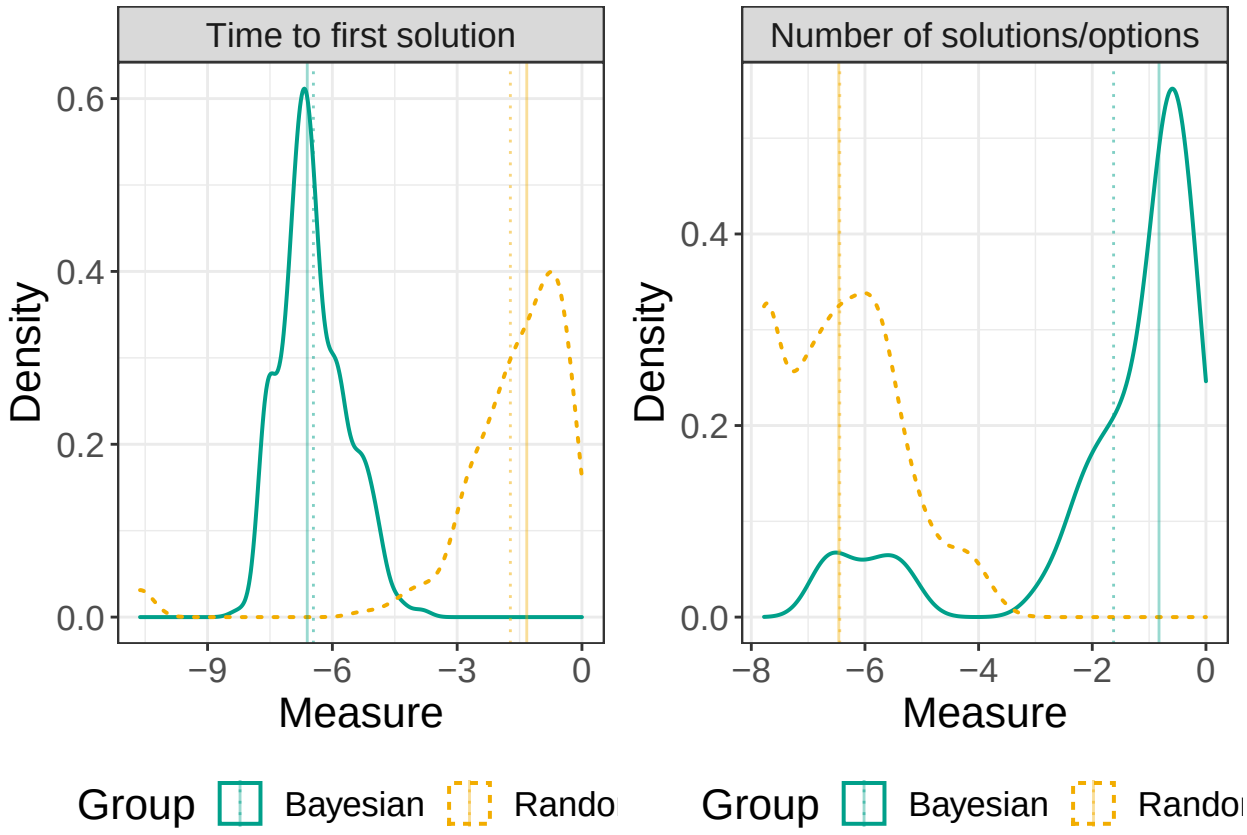
2.2 Case-study models on primary productivity

Unlike the comparison over benchmark models, the case-study model is evaluated in general terms, i.e., overall performances, because the case-study model lacks characteristics observed among benchmark models. Instead, generally the model is described with depth of 5 and attributes' weights that closely resemble the normal distribution.

The statistical analysis clearly shows the performance-wise dominance of the Bayesian method in terms of both measures: the time to first solution and the size of solutions (Table 5).

Table S5: Statistical test of normality of, homoscedasticity and difference between values for both performance measures (time to first solutions and size of solution set) of both Bayesian and Random methods. Reported values are p-values, with significance reported by the asterisk character (*). The difference is tested with the test given in brackets and presented with a null hypothesis stating that of both samples are sampled from same distribution. Vertical solid and dotted lines represent the sample's median and mean, respectively.

Performance	Normality (BO)	Normality (RS)	Homoscedasticity (BO vs. RS)	Sample distribution (BS vs. RO)
# Solutions/options	0 *	0 *	(Kruskal-Wallis) 0.012 *	(Mann-Whitney U) 0 *
Time to first solution	0 *	0 *	(Kruskal-Wallis) 0 *	(Mann-Whitney U) 0 *


Time to first solution
Number of solutions/options
Density
Measure
0.0
0.2
0.4
0.6
−9
−6
−3
0
−8
−4
−2
Group
Bayesian
Rando