Tree-based ensemble methods for sensitivity analysis of environmental models
A performance comparison with Sobol and Morris techniques

Jaxa-Rozen, Marc; Kwakkel, Jan

DOI
10.1016/j.envsoft.2018.06.011

Publication date
2018

Document Version
Accepted author manuscript

Published in
Environmental Modelling and Software

Tree-based ensemble methods for sensitivity analysis of environmental models: a performance comparison with Sobol and Morris techniques

Marc Jaxa-Rozen[a,*], Jan Kwakkel[a]

[a] *Faculty of Technology, Policy and Management, Delft University of Technology, Jaffalaan 5, Delft 2628 BX, The Netherlands*

Abstract

Complex environmental models typically require global sensitivity analysis (GSA) to account for non-linearities and parametric interactions. However, variance-based GSA is highly computationally expensive. While different screening methods can estimate GSA results, these techniques typically impose restrictions on sampling methods and input types. As an alternative, this work evaluates two decision tree-based methods to approximate GSA results: random forests, and Extra-Trees. These techniques are applicable with common sampling methods, and continuous or categorical inputs. The tree-based methods are compared to reference Sobol GSA and Morris screening techniques, for three cases: an Ishigami-Homma function, a H1N1 pandemic model, and the CDICE integrated assessment model. The Extra-Trees algorithm performs favorably compared to Morris elementary effects, accurately approximating the relative importance of Sobol total effect indices. Furthermore, Extra-Trees can estimate variable interaction importances using a pairwise permutation measure. As such, this approach could offer a user-friendly option for screening in models with inputs of mixed types.

Keywords: Global sensitivity analysis, factor screening, decision tree methods, ensemble learning methods

1. Introduction

Sensitivity analysis (SA) is recognized as a key step for analyses which involve the assessment and propagation of uncertainty in mathematical models (Frey and Patil, 2002; Helton and Oberkampf, 2004). In particular, techniques for global sensitivity analysis (GSA) have become an accepted standard for the evaluation of the impact and interactions of uncertain inputs in complex environmental models (as described in this journal by e.g. Saltelli and Annoni 2010; Nossent et al. 2011; Pianosi and Wagener 2015). These techniques consider the output behaviour of the model over the full domain of uncertain inputs; specifically, this implies that the full distribution of each input parameter should be evaluated, and that the importance of each input should be evaluated across the domain of all other parameters (Liu and Homma, 2009). This is in contrast to "one-at-a-time" (OAT) sensitivity analysis – which focuses on response to changes in individual inputs around an initial baseline value, and which for instance inadequately captures non-additive responses caused by interactions between input parameters. These properties make GSA particularly relevant for applications such as integrated assessment models, which frequently combine a large number of highly uncertain inputs with a non-linear, non-additive structure. In these conditions, a OAT analysis can lead to an incomplete or misleading interpretation of model uncertainty. As such, GSA can help analysts and decision-makers better understand and communicate the results of complex models, and ultimately make these models more credible in a decision support context. However, the computational cost of existing GSA methods can quickly become prohibitive with complex simulation models.

This paper therefore draws on the statistical learning literature to evaluate the performance of decision tree-based ensemble methods, when applied to typical sensitivity analysis problems. These methods rely on ensembles of decision trees which match partitions of the input space with a predicted output, and are commonly implemented using the random forests and Extra-Trees algorithms (Breiman, 2001; Geurts et al., 2006). These techniques perform well at relatively small sample sizes for non-linear regression or classification problems in which input interactions are significant; they are also able to handle both numerical and categorical inputs (Louppe, 2014). Building on previous investigations of decision tree methods for sensitivity analysis (e.g. Harper et al., 2011), this paper will show that these methods can replicate some of the key insights of GSA by estimating relative variable importances and interactions, at a much smaller computational cost.

In the context of GSA, Saltelli et al. (2008) summarize four analysis objectives, or "settings": i) factor prioritiza-

*Corresponding author

Email addresses: m.jaxa-rozen@tudelft.nl (Marc Jaxa-Rozen), j.kwakkel@tudelft.nl (Jan Kwakkel)

tion, which identifies inputs (or groups of inputs) which contribute the most towards output uncertainty; ii) factor fixing, which conversely identifies inputs which have a negligible contribution to output uncertainty and may thus be fixed at a given value; iii) variance cutting, which investigates the assumptions on input values under which output uncertainty can be reduced below a given threshold; and iv) factor mapping, which identifies regions of the input space associated with a given output space. Factor prioritization is especially valuable for identifying uncertain inputs on which additional data collection and modelling efforts should be focused, while factor fixing can make models easier to test and interpret by discarding non-influential inputs. These two settings are arguably the most common for sensitivity analysis in environmental modelling. Variance cutting can be applied in risk and reliability analysis, in which analysts may need to meet a certain tolerance (e.g. Plischke et al., 2013; Saltelli and Tarantola, 2002), while factor mapping can be related to techniques for scenario discovery (e.g. Bryant and Lempert, 2010; Kwakkel and Jaxa-Rozen, 2016; Guivarch et al., 2016).

GSA results are typically interpreted through quantitative importance indices, which can be used to compare the uncertain inputs in the context of the desired setting (e.g. factor prioritization or factor fixing). Liu and Homma (2009) and Saltelli (2002b) describe several features of an ideal uncertainty importance index. Notably, the measure should be i) unconditional, in the sense of the index being independent of assumptions about the input value (so that the sensitivity metric of an input is not conditional on a given baseline value); ii) easy to interpret, for instance by representing an input's proportional contribution to output uncertainty; iii) easy to compute numerically; iv) stable across different samples (e.g. robust to bootstrapped resamples); and v) model-free, so that the indices are independent from structural properties of the model such as linearity and additivity. Borgonovo (2007) and Pianosi and Wagener (2015) further propose vi) moment independence as a criteria, so that the influence of the entire input distribution can be assessed on the output distribution independently of the shape of the latter, without being conditional on a specific moment of the output distribution.

In practice, the estimation of these indices often presents analysts with a trade-off between computational cost, and the information gained from the sensitivity analysis. Variance-based GSA (Sobol, 2001; Saltelli, 2002b) is arguably the most prominent approach in the literature. This technique can be used under factor prioritization or factor fixing settings to directly assess the contribution of uncertain inputs to unconditional output variance. A typical application of the Sobol technique provides first-order and total indices, which respectively describe the fraction of output variance contributed by each factor on its own, and by the sum of first-order and all higher-order interactions for each factor. Additional terms which decompose these higher-order interactions, such as pairwise second-order interactions between variables, can be computed at an additional computational cost. These indices satisfy the above requirements except for moment independence (by relying on variance as a proxy for output uncertainty – which may cause issues with multimodal or skewed distributions, e.g. Pianosi and Wagener, 2015). Given their clear mathematical interpretability and straightforward computation, Sobol indices have for example been increasingly applied for hydrological and integrated assessment models (Tang et al., 2006; Pappenberger et al., 2008; Nossent et al., 2011; Herman et al., 2013; Butler et al., 2014). The indices can also be extended to cover non-scalar inputs – e.g. "switches" for structural model uncertainties – in addition to scalar input ranges (Baroni and Tarantola, 2014). However, the use of variance-based GSA can be difficult for models with a large number of input parameters. In principle, the model evaluations N required to calculate Sobol indices grow linearly with the number of input parameters p, so that $N = n(p + 2)$ for the calculation of first-order and total indices (where n is a baseline sample size). In practice, this baseline sample size also tends to increase significantly for complex models with multiple parameters, and may vary from 100 to 10,000 or more (e.g. Butler et al., 2014, in which $n > 130,000$ was needed for a simulation model with 30 inputs). The computational cost of variance-based GSA may therefore prevent its use for models with a significant runtime.

The literature presents a variety of alternative methods which can be used under such circumstances to reproduce some of the insights of variance-based GSA, at a smaller number of model evaluations. These are often used in a factor fixing setting to screen non-influential variables (see e.g. Kleijnen, 2009 for a review of screening techniques). The elementary effects method (Morris, 1991; Campolongo et al., 2007) is commonly applied to estimate sensitivity measures, using an efficient sampling design to cover the domain of uncertain inputs with a set of sampling trajectories. However, while elementary effects indices can be used to rank inputs based on their influence on model output, the interpretation of the indices is essentially qualitative rather than quantitative, as their relative values may not match the relative importances estimated by variance-based methods. In addition, the sampling trajectories assume uniformly distributed continuous inputs, so that these indices are unsuitable for models with categorical or non-scalar inputs; they also do not provide information about specific interactions between variables. The specific input sampling required for elementary effects is also a drawback: for instance, this prevents the use of model datasets which may have been generated from a typical uncertainty analysis, and which could be reused for SA under a "given data" approach (Borgonovo et al., 2017; Plischke et al., 2013). A generic input sampling can otherwise support a multi-method framework which covers complementary aspects of model sensitivity at the same computational cost (such as Pianosi et al. (2017)'s framework for the estimation of first-order indices, density-based

indices, and interactions using a Latin Hypercube sample). Under Liu and Homma (2009)'s criteria, the elementary effects indices would therefore be suboptimal in terms of interpretability and ease of computation.

These sensitivity analysis methods have largely been developed and applied in the context of model-based risk analysis and environmental science. However, a parallel domain of research has also focused on the problem of feature selection in statistical learning, which offers some useful analogies to the factor fixing setting in sensitivity analysis. As described by Guyon and Elisseeff (2003), feature selection aims to reduce the dimensionality of the input data used in a learning problem by selecting a subset of the original variables, and eliminating variables which are not relevant. This process offers several advantages, such as making output data easier to analyze, making the prediction model more understandable, or improving the accuracy of the prediction model by avoiding overfitting. Several definitions of variable relevance (described more extensively in e.g. Blum and Langley, 1997; Kohavi and John, 1997) can be followed, leading to different paths for feature selection. For instance, the feature selection literature describes *wrapper* methods, in which variables are assessed based on their relevance for a given predictor (Kohavi and John, 1997). In this application, feature selection aims to select a subset of variables which maximizes the accuracy of a predictor, which is considered as a "black box". When combined with a suitable predictor, this approach enables a more flexible analysis, for instance by relaxing assumptions on input types or distributions (Lazar et al., 2012). Decision trees are a popular example of such a predictor, which combine several desirable properties for statistical learning in general, and for feature selection in particular. As such, these predictors can represent arbitrary relations between inputs and outputs, without prior assumptions about inputs or structural relationships (Louppe, 2014). They can also be used for non-linear problems with heterogeneous input data (such as continuous or categorical parameters), and implicitly account for variable interactions. Decision trees are therefore a popular option for feature selection (Guyon and Elisseeff, 2003); they are commonly used within ensemble methods which combine multiple decision trees to improve performance, such as random forests and Extra-Trees (see e.g. Hapfelmeier and Ulm, 2013 for a review of random forests in a feature selection context).

It is therefore interesting to assess whether insights from the literature on feature selection can be transferred to the sensitivity analysis of complex environmental models. In particular, decision tree-based predictors may mitigate some of the drawbacks of common screening techniques, as they can be applied with generic input sampling designs and categorical uncertainties, while supporting the study of variable interactions. This work builds on past applications of decision tree-based predictors in the environmental modelling literature, such as Harper et al. (2011); this study combined random forests and individual trees to evaluate variable importances and interactions in a model of cottonwood dynamics. Similarly, Almeida et al. (2017) and Singh et al. (2014) used individual classification trees to study critical thresholds in a factor mapping setting, for a hydro-climactic watershed modelling framework and for a model of slope stability, respectively. Given the demonstrated performance and widespread availability of the random forests and Extra-Trees ensemble predictors, this paper will focus on comparing both of these methods with the reference Sobol and elementary effects techniques, using typical model cases.

Section 2 of the paper provides more background about the Sobol and elementary effects methods for global sensitivity analysis, and describes the selected decision tree-based methods. Section 3 then compares the performance of the tree-based ensemble methods against reference GSA results, for three cases: an Ishigami test function, a H1N1 flu pandemic model, and the CDICE integrated assessment model. Section 4 discusses the results and describes potential avenues for future work.

2. Methods

2.1. Reference methods for global sensitivity analysis

2.1.1. Variance-based Sobol indices

The Sobol technique for global sensitivity analysis uses variance decomposition to establish the contribution of each uncertain input to the unconditional output variance of a model, which can be non-linear and non-additive (e.g. Sobol (2001); Homma and Saltelli (1996)). Given a model output Y and a set $X = (x_1, ..., x_p)$ of independent parameters, the corresponding function $f(X)$ can be decomposed into terms of increasing order:

$$Y = f(X) = f(x_1, ..., x_p) \tag{1}$$

$$\begin{aligned} f(x_1, ..., x_p) =& f_0 + \sum_{j=1}^{p} f_j(x_j) + \sum_{j=1}^{p} \sum_{k=j+1}^{p} f_{jk}(x_j, x_k) \\ &+ ... + f_{1,...,p}(x_1, ..., x_p) \end{aligned} \tag{2}$$

The unconditional variance $V(Y)$ can correspondingly be decomposed into partial variances, where e.g. V_j and V_{jk} represent the variances of f_j and f_{jk}, respectively:

$$V(Y) = \int_{\Omega} f^2(X)\, dX - f_0^2 \tag{3}$$

$$V(Y) = \sum_{j=1}^{p} V_j + \sum_{j=1}^{p-1} \sum_{k=j+1}^{p} V_{jk} + ... + V_{1,...,p} \tag{4}$$

Using these partial variances, the first-order, second-order and total Sobol sensitivity indices can then be defined in relation to the total variance:

$$S_j = \frac{V_j}{V(Y)} = \frac{V_{x_j}\left[E_{X\sim j}(Y \mid x_j)\right]}{V(Y)} \tag{5}$$

$$S_{jk} = \frac{V_{jk}}{V(Y)} = \frac{E_{X\sim x_j,x_k}\left[V_{x_j,x_k}(Y \mid X_{\sim x_j,x_k})\right]}{V(Y)} - S_j - S_k \tag{6}$$

$$ST_j = 1 - \frac{V_{\sim j}}{V(Y)} = \frac{E_{X\sim x_j}\left[V_{x_j}(Y \mid X_{\sim x_j})\right]}{V(Y)} \tag{7}$$

The first-order index S_j , or main effect, represents the fraction by which the output variance would be reduced on average by fixing x_j within its range. The second-order index S_{jk} then represents the fraction of output variance linked to inputs x_j and x_k which is not captured by the superposition of each input's first-order index, and thus corresponds to interaction effects in a non-additive model. Finally, the total effect ST_j includes the contribution of the first-order effect and the sum of all higher-order interaction effects. For a non-additive model, the difference $ST_j - S_j$ thus indicates the importance of interaction effects, which can be directly assessed for pairwise interactions using the second-order S_{jk} index. These indices can be used for factor prioritization, in which the input parameters with the highest main effect S would be assessed as the most influential. Conversely, for factor fixing, input parameters with $ST \approx 0$ can be judged to be non-influential and discarded from the analysis, given that they do not contribute to output variance either through their main effect or through interactions (Saltelli et al., 2008). As shown by Baroni and Tarantola (2014), these indices can similarly be applied to assess the contribution of non-scalar inputs (such as structural model "switches") to output variance.

In practice, the unconditional variance $V(Y)$ typically needs to be estimated using Monte Carlo integrals rather than an analytical form. Saltelli (2002a) for instance presents an input sampling strategy which can be used to estimate the first-order, second-order and total indices at a cost of $N = n(2p + 2)$ evaluations. This sampling design has been implemented in a variety of software packages; for the purposes of this work, the Python SALib library (Herman and Usher, 2017) is used to generate input samples, and to calculate the resulting Sobol indices with bootstrapped confidence intervals.

2.1.2. Morris elementary effects

For models with a large number of uncertain inputs and/or a high computational cost, the elementary effects method is used as a standard screening technique for factor fixing (Morris, 1991; Campolongo et al., 2007). The method relies on a systematic sampling of the input space to generate a randomized ensemble of "one-at-a-time" experiments. Taking a set $X = (x_1, ..., x_p)$ of independent input parameters transformed so as to be uniformly distributed in the interval [0,1], a certain number r of sampling "trajectories" of $(p + 1)$ points are then constructed to vary one input at a time, across k levels of the [0,1] input domain. Starting from a given value of X and taking $\Delta \in \{1/(k-1), ..., 1 - 1/(k-1)\}$, the elementary effect of x_j is given by:

$$EE_j(X) = \left(\frac{f(x_1, ..., x_{j-1}, x_j + \Delta, x_{j+1}, ..., x_p) - f(X)}{\Delta}\right) \tag{8}$$

The distribution F_j of this elementary effect can then be obtained by sampling multiple initial values of X. Morris (1991) originally proposed using the mean μ and standard deviation σ of this distribution to respectively assess the overall influence of the variable on output, and the magnitude of higher-order effects due to non-linearities and interactions. However, the μ measure was shown to be vulnerable to type II error (i.e. potentially ignoring influential variables) in the case of non-monotonic models, as elementary effects may cancel each other out at different points of the input set X. Campolongo et al. (2007) thus introduced a measure $\mu*$, which takes the mean of the distribution of the absolute values of the elementary effects. This index was shown to acceptably estimate the ST indices obtained from a variance-based global sensitivity analysis. The elementary effects technique can thus reliably be used to identify factors which have a negligible influence, and which may be discarded from the analysis. However, the index σ is more difficult to interpret; it combines the effect of interactions as well as non-linearities, so that specific interactions between pairs of variables cannot be evaluated. The assumption of uniformly distributed scalar inputs X also makes the indices unsuitable for non-scalar inputs.

As with the Sobol technique, the SALib library will be used to sample input trajectories (with the efficient trajectories introduced by Campolongo et al. (2007)) and to estimate the elementary effect indices.

2.2. Decision tree-based ensemble methods

Decision trees are a simple and well-established general approach for statistical learning; such trees aim to identify the splitting criteria which describe the relationship between a set of input combinations, and regions of the output space (graphically illustrated for an idealized case in Figure 1). Decision trees can be fitted through several specific algorithms, such as classification and regression trees (CART; Breiman et al., 1984). The right panel of Figure 1 presents a simple example of a regression tree for a test case $\boldsymbol{y} = f(\boldsymbol{X})$. The tree is fitted to an output vector $\boldsymbol{y} = (y_1, ..., y_i)^T$, with vectors of predictor values $\boldsymbol{x}_j = (x_{1,j}, ..., x_{i,j})^T$ for $j \in \{1, 2\}$, forming the matrix $\boldsymbol{X} = (\boldsymbol{x_1}, \boldsymbol{x_2})$. The depth of the tree is here artificially constrained to create a small number of nodes t; each of

the terminal nodes t_3, t_4, t_5, t_6 corresponds to a rectangular region of the input space shown in the left panel of the figure. The predicted value $\hat{y}_t$ at each node (i.e. for each corresponding combination of ranges for the predictor values) is then the mean of the output values in each node, $y_i \in t$.

Starting from the root node t_0, the tree is "grown" using an optimization procedure to search over all possible binary splits $s = (\boldsymbol{x}_j \leq c)$, and identify the splitting point c across the values of variable $\boldsymbol{x}_j$ which leads to the greatest reduction in the impurity of the resulting "child" nodes (typically using Gini impurity for classification, or mean square error for regression). We let t_L and t_R represent the left and right child nodes obtained when partitioning node t with a binary split. The reduction in impurity from split s at node t is then:

$$\Delta i(s,t) = i(t) - \frac{N_{tL}}{N_t} i(t_L) - \frac{N_{tR}}{N_t} i(t_R) \tag{9}$$

where N_t, N_{tL}, N_{tR} are the number of samples in the parent node and the left and right child nodes. For regression, we use the mean square error as a measure of impurity:

$$i_R(t) = \frac{1}{N_t} \sum_{i=1}^{N_t} (y_i - \hat{y}_t)^2 \tag{10}$$

In the example shown in Figure 1, this leads to the selection of a splitting value $c_0 = 2.572$ on $\boldsymbol{x}_2$ in the root node. This splitting procedure is repeated until a stopping criterion is reached, which can be e.g. the depth of the tree, or the maximum number of samples to be found in a terminal node.

Individual decision trees will typically display high levels of variance, so that small changes in the selected input data may cause significant changes in the structure of the fitted tree. As such, ensemble methods – in which multiple, randomly generated instances of an estimator are aggregated – can increase the accuracy of decision trees for classification and regression. The most popular of these has been the random forests algorithm (RF; Breiman, 2001), in which multiple CARTs are fitted to bootstrap samples of the data and aggregated (or "bagged"). The trees are randomized by selecting a subset of the input variables as candidates for splitting at each node. Their predictions are then simply averaged for a regression problem, or taken as a majority vote for classification. Geurts et al. (2006) add an additional randomization step for the construction of "extremely randomized trees" (or Extra-Trees), in which the random selection of variables for splitting is combined with randomized cutting points at each node (typically using the full input set, rather than bootstrap samples). This step can improve accuracy as well as computational performance. This paper will thus focus on the RF and Extra-Trees (ET) algorithms, due to the demonstrated accuracy and versatility of these techniques for non-linear regression problems with heterogeneous inputs (Hastie et al., 2009; Louppe, 2014).

The performance of random forests and Extra-Trees can be tuned with parameters which control the construction of the ensemble. The most significant of these are i) the number of trees T used in the ensemble, ii) the size of the candidate subset m of the input variables p which is assessed for each split of the individual trees, and iii) the depth to which the trees are grown (which can be controlled with the same criteria described above for individual trees, such as the minimum number of samples N_{leaf} to be left in the nodes created after a split).

Increasing the number of trees T used in the ensemble will in principle reduce prediction error, with the methods being robust to overfitting (Geurts et al., 2006). In practice, the size of the ensemble is likely to be driven by computational constraints, with a trade-off between accuracy and time. The size of the subset of variables m will affect correlation between the trees within the ensemble, with a smaller value increasing randomness; in the extreme case of $m = 1$, each split is determined by a single random input, and the trees are said to be totally randomized. The choice of this parameter depends on the problem, with $m = p/3$ as a starting point for regression (Hastie et al., 2009). Finally, the depth of the trees will affect generalization error: fully developed trees may overfit the data, while smaller trees will typically have larger bias. The empirical results presented by Geurts et al. (2006) suggest a value of $N_{min} = 5$ as a robust starting point for regression, for the minimum number of samples required to split a node.

Variable importance metrics

Different measures can be used to assess the importance (or predictive strength) of input variables in random forests and Extra-Trees. The most common metrics are *Mean Decrease Impurity* (MDI) and *Mean Decrease Accuracy* (MDA) (Breiman, 2001; detailed in Louppe, 2014). MDI relies on the criterion used to select an optimal split in CART (defined in eq. 9), extending it across the ensemble of trees. The MDI importance of a variable x_j can thus be computed from the total decrease in node impurity (across the trees in the ensemble) which is obtained when x_j is used for splitting. A variable associated with a large decrease in impurity is then influential. We use the definition given by Louppe (2014), with an ensemble of T trees:

$$MDI(x_j) = \frac{1}{T} \sum_{\tau=1}^{T} \sum_{t \in \varphi_\tau} 1(j_t = x_j) \left[\frac{N_t}{N} \Delta i(s_t, t) \right] \tag{11}$$

where the change in impurity $\Delta i(s_t, t)$ is summed in tree φ_τ over all nodes t in which x_j is used for splitting, weighted by the fraction of total samples present in the node (N_t/N); j_t is the variable used for splitting at node t. This value is averaged over all trees φ_τ in the ensemble.

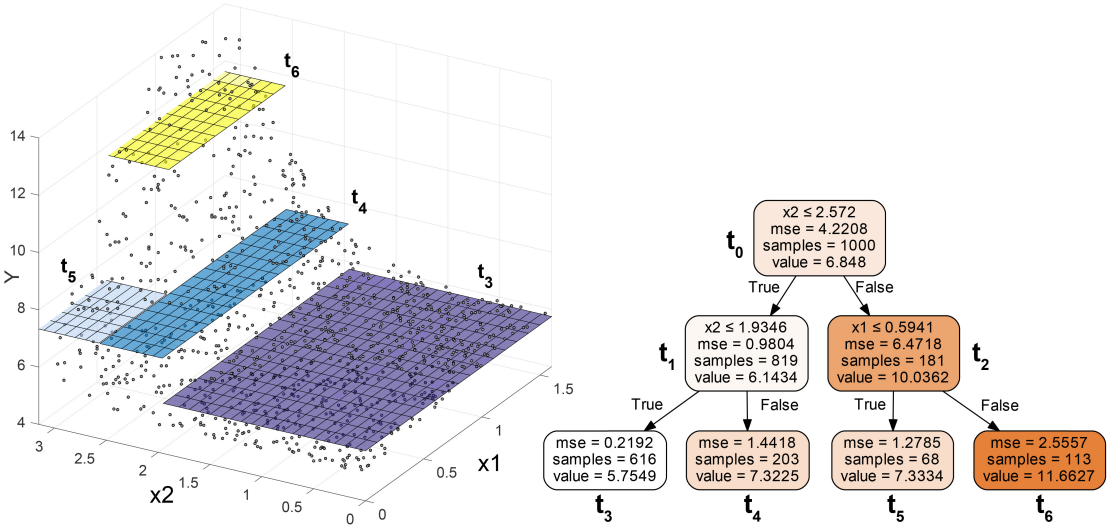


Figure 1: Graphical representation of a regression tree. Left panel: two-dimensional partition of a feature space; right panel: decision tree corresponding to the partition.

An alternate measure is given by the MDA (or permutation) importance, in which the change in prediction accuracy of the ensemble is assessed after randomly permuting the input values for variable x_j. When using bootstrapping, MDA can be estimated on the out-of-bag (OOB) samples at each tree, i.e. the samples which were not part of the bootstrapped training set for each tree. Following Strobl et al. (2008), we compare prediction accuracy on the OOB samples for the original vector of input values $\boldsymbol{x}_i = (x_{i,1}, ..., x_{i,p})$, and for a vector $\boldsymbol{x^{\pi j}}_i$ in which the values of x_j are permuted across the observations i: $\boldsymbol{x^{\pi j}}_i = (x_{i,1}, ..., x_{\pi j(i),j}, x_{i,j+1}, ..., x_{i,p})$. An influential variable will cause a large decrease in prediction accuracy, while a non-influential variable would not significantly change the performance of the ensemble. The mean square error is typically used as a measure of prediction accuracy for regression. Taking $\hat{y}_{\tau,i} = \varphi_\tau(\boldsymbol{x}_i)$ as the prediction given by tree φ_τ for observation i, averaging over each observation in the set of OOB samples B^τ , then averaging over the ensemble of trees, we obtain:

$$MDA(x_j) = \frac{1}{T}\sum_{\tau=1}^{T} \frac{\sum_{i \in B^\tau} (y_i - \varphi_\tau(\boldsymbol{x^{\pi j}}_i))^2 - (y_i - \varphi_\tau(\boldsymbol{x}_i))^2}{|B^\tau|} \tag{12}$$

These variable importance measures have been extensively studied and refined in the context of RF and feature selection (e.g. Ishwaran, 2007; Strobl et al., 2007, 2008; Wright et al., 2016; Bureau et al., 2005). An advantage of the measures is their implicit consideration of interactions across variables, which follows from the tree induction process. This makes RF importance measures a potential candidate for approximating the total effect indices obtained through global sensitivity analysis. In feature selection, Qi et al. (2006) for instance found that RF outperformed five other classifier methods for the detection of interactions in large datasets. However, the MDI metric tends to be biased towards especially salient variables, due to the underlying bias of the splitting criterion. In the case of categorical variables, MDI also tends to be biased towards inputs with a larger number of categories (Strobl et al., 2007). The bias of the MDA measure was less obvious in the results discussed by Strobl et al. (2007) but can nonetheless affect the reliability of the measures, particularly in the case of correlated predictors. Strobl et al. (2009) and Altmann et al. (2010) thus introduced revised metrics to address these characteristics. For the purposes of this work, the relative values of the ST and $\mu*$ indices obtained from the Sobol and elementary effects techniques will be compared to the standard MDI importance index, which offers better computational performance than MDA on large datasets. The revised metrics of e.g. Strobl et al. (2009) are less relevant for this application, due to the typical assumptions on uncorrelated parameters which are used when sampling inputs for sensitivity analysis.

In addition to MDI, we use a variant of the MDA metric (Bureau et al., 2005) to directly estimate the effect of pairwise interactions between variables, by permuting both of the corresponding input samples across observations in a vector $\boldsymbol{x^{\pi j,k}}_i$, and subtracting individual MDA importances. For variables x_j, x_k, the pairwise MDA is then given by:

$$MDA(x_j, x_k) = \left[\frac{1}{T}\sum_{\tau=1}^{T}\frac{\sum_{i\in B^\tau}(y_i - \varphi_\tau(\boldsymbol{x}^{\boldsymbol{\pi j,k}}{}_i))^2 - (y_i - \varphi_\tau(\boldsymbol{x}_i))^2}{|B^\tau|}\right] - MDA(x_j) - MDA(x_k) \quad (13)$$

To assess the stability of the MDI indices, we use a convergence criterion presented by Touzani and Busby (2014) (eq. 14), where $\boldsymbol{V}_N = (v_1, \ldots, v_p)$ is the vector of estimated variable importance indices at a sample size of N observations. The criterion considers the Euclidean norm $\|\cdot\|$ of the vector rather than individual indices, so that more influential indices have a greater effect on measured convergence. We compute the indices sequentially over an increasing sample size at intervals of ΔN total samples. The convergence criterion κ_N is then computed backwards from N over t intervals, with t and ΔN being specified for each case study. This criterion will also be used to ensure the stability of the reference vector of Sobol ST indices, $\boldsymbol{ST}$.

$$\kappa_N = \frac{1/t\sum_{s=1}^{t}\|\boldsymbol{V}_N - \boldsymbol{V}_{N-s\Delta N}\|}{\|\boldsymbol{V}_N\|} \quad (14)$$

Finally, the accuracy of the proportional estimated variable importances is assessed with the root mean square error and mean bias error of $\boldsymbol{V}_N$, relative to $\boldsymbol{ST}$. As the indices measure different quantities (e.g. the decrease in mean square error for MDI, and fraction of output variance for ST), the values are not directly comparable; however, by first rescaling each vector relative to its maximum value across all p variables, we can compare the proportional importances estimated by each method. We avoid normalizing the estimated importances over $[0, 1]$ to preserve negative values which may indicate numerical artifacts. RMSE is used as an overall indicator of accuracy, while MBE provides information about the average overestimation or under-estimation of variable importances.

$$RMSE = \frac{\|\boldsymbol{ST}/\max(\boldsymbol{ST}) - \boldsymbol{V}_N/\max(\boldsymbol{V}_N)\|}{\sqrt{p}} \quad (15)$$

$$MBE = \frac{\sum(\boldsymbol{ST}/\max(\boldsymbol{ST}) - \boldsymbol{V}_N/\max(\boldsymbol{V}_N))}{p} \quad (16)$$

2.3. Software availability

The model cases are tested in the Python environment using the Exploratory Modeling Workbench (Kwakkel, 2017). This library provides an interface for sensitivity analysis using the *scikit-learn* implementation of the random forests and Extra-Trees algorithms (Pedregosa et al., 2011), as well as the Sobol and Morris techniques through the *SALib* library (Herman and Usher, 2017). These libraries are available through the *pip* package manager for Python. Alternative implementations of the tree-based methods can be found in the R environment, with the *party* and *extraTrees* packages (Hothorn et al., 2017; Simm and de Abril, 2015).

3. Model cases

This section will present model case studies in increasing order of complexity, using the benchmark Ishigami-Homma function (Ishigami and Homma, 1990), an exploratory SIR model of the A(H1N1)v swine flu epidemic (Pruyt and Hamarat, 2010), and the CDICE simulation version of the DICE-2007 integrated assessment model (Butler et al., 2014; Nordhaus, 2007). Each case will first present reference sensitivity analysis results with the Sobol and Morris techniques. These results will then be compared with the MDI and pairwise MDA variable importances, as estimated from the random forests and Extra-Trees ensemble techniques.

3.1. Ishigami test function

The first test case is the Ishigami-Homma function (Ishigami and Homma, 1990), which is a common test case for sensitivity analysis due to its analytical tractability and non-additive properties:

$$Y = sin(x_1) + \alpha sin(x_2)^2 + \beta x_3^4 sin(x_1) \quad (17)$$

where x_1, x_2, x_3 are uniformly distributed in $[-\pi, \pi]$, with $\alpha = 7$ and $\beta = 0.1$. Using a Latin Hypercube sample with $N = 1500$ yields the output distribution shown in Figure 2.

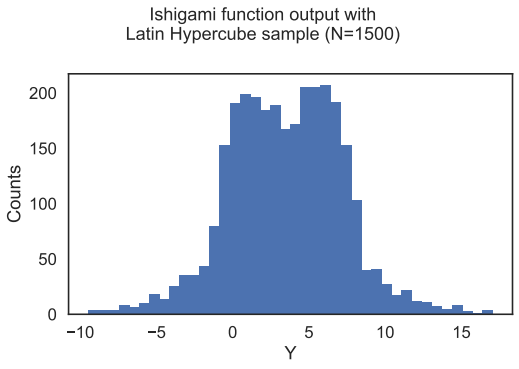


Figure 2: Output distribution for Ishigami function.

Figure 3 presents the convergence of the Sobol (top panel, left) and elementary effects (middle panel, left) indices as a function of the total number of input samples, and the relationships between the key indices provided by each technique (right panels). The Sobol sample size of $N = 15000$ is chosen to achieve a convergence criterion of $\kappa_N < 0.01$ (using intervals of $\Delta N = 400$ samples and $t = 4$ intervals); the shaded envelopes present 95% confidence bounds for the indices. The relationship between the

first-order and total Sobol indices indicates higher-order interactions for x_1 and x_3, as expected from the structure of the function, while S and ST are identical for x_2.

Using $k = 8$ levels, with $\Delta = k/[2(k-1)]$ as recommended by Campolongo et al. (2007), the variable ranking obtained from the $\mu*$ elementary effects converges at a relatively small number of samples. However, the ranking does not match the order of the ST indices, underestimating the relative importance of x_1. This is illustrated in the bottom panel of the figure by plotting the proportional values of the $\mu*$ and ST indices against each other; the values of each group of indices are scaled relatively to the maximum value in each group, for $N = 5000$ and $N = 15000$ respectively. x_1 and x_3 show relatively higher values on the σ index, compared to their values for the $\mu*$ index. This could potentially be related to their interaction effects (which, in this case, can be inferred from the structure of the model), but the contribution of interactions towards the value of σ cannot be distinguished from the contribution of non-linearities (Saltelli and Annoni, 2010). This is indeed highlighted by x_2, which has approximately the same value on σ as x_3; although it has a non-linear impact, it does not interact with other variables in the model structure.

Figure 4 shows the convergence of the mean MDI importance indices for the random forests (top panel, left) and Extra-Trees (bottom panel, left) techniques over a Latin Hypercube sample, using 50 bootstrap resamples to estimate confidence bounds (shown by shaded envelopes which contain the full range of estimated values). Appendix A presents detailed convergence results, indicating that both algorithms stabilize below $\kappa_N < 0.02$ around $N = 3000$ samples, similarly to the Morris indices. Both algorithms are parameterized with $T = 100$ trees, $m \approx p/3 = 1$ (so that the trees are totally randomized), and a stopping criterion of $N_{leaf} = 2$. The right panels compare the mean estimated MDI importances (scaled relative to the highest MDI value at $N = 5000$), against the scaled reference ST indices.

For both techniques, Appendix A shows the root mean square error (RMSE) and mean bias error (MBE) estimated over all scaled MDI values, compared to scaled ST values (where positive bias is linked with an underestimation of relative variable importances compared to ST; eq. 15). Compared to the Morris $\mu*$ results, both ensemble techniques correctly rank the input variables; compared to random forests, Extra-Trees show quicker convergence, and a lower error compared to the relative ST values.

A potential drawback of the ensemble techniques is the requirement of choosing suitable tuning parameters. Focusing on the Extra-Trees technique due to its favorable performance, Figure 5 shows the RMSE (relative to scaled ST values) for scaled estimated importances, bootstrapped confidence interval on RMSE across 50 resamples, and MBE. These metrics are presented across a range of values for the number of trees T, the number of splitting features m (subplot rows) and the minimum number of samples per node N_{leaf} (subplot columns).

RMSE appears robust to the number of trees T. The combination of m and N_{leaf} has a significant influence on RMSE; the starting point of $m \approx p/3$ suggested by Hastie et al. (2009) provides good results on RMSE, when combined with a small value for N_{leaf} (which controls the depth of the trees). N_{leaf} has a significant influence on MBE at a given value of m, which is particularly relevant for sensitivity analysis: a positive value indicates that relative variable importances are underestimated compared to ST, which could lead to a type II error in a screening setting (i.e. discarding potentially influential variables). Smaller trees appear more vulnerable to this error, which emphasizes more salient variables (x_1 and x_2). This can be compensated by increasing m to decrease the randomness of the trees; however, at smaller values of N_{leaf} (e.g. $m = 3$ and $N_{leaf} = 1$), this increases RMSE.

As indicated by the relative values of ST and S for x_1 and x_3, the interaction between these variables contributes significantly to the output behavior. The left panel of Figure 6 shows the pairwise interaction importances estimated by the second-order Sobol S2 indices; the right panel presents MDA interaction importances estimated with Extra-Trees (averaged over 50 bootstrap resamples, on a 1500 sample set). The analytical relationship between x_1 and x_3 is therefore identified by both techniques, with other pairwise importances being negligible.

3.2. H1N1 swine flu epidemic model

Pruyt and Hamarat (2010) present a simple exploratory system dynamics model of the 2009 swine flu epidemic, based on a two-region SIR model. This provides a more complex test case for sensitivity analysis due to the larger number of input variables (with 17 continuous inputs and two structural switches), and a broad output distribution. Table 1 shows the input variables and their bounds, assuming uniform distributions for all continuous variables. Figure 7 presents the resulting output distribution on the outcome of interest (defined as the number of fatalities in region 1 of the model) with a Latin Hypercube sample.

For this example, the Sobol technique requires $N > 150,000$ for a stable estimation of variable rankings, as shown in the top panel of Figure 8. A reference value of $N = 800,000$ was chosen by setting the convergence criterion to $\kappa_N < 0.01$ (using intervals of $\Delta N = 40,000$ samples and $t = 4$ intervals). The relationship between ST and S indicates that higher-order interactions are present for most of the variables, with a group of 7 variables contributing significantly to output behavior.

The middle panel shows Morris results with $k = 8$ levels and $\Delta = k/[2(k-1)]$. While the same group of 7 variables is identified by the $\mu*$ indices, they require a relatively large sample size for a stable estimation. Appendix A presents a convergence analysis with $\Delta N = 10,000$ samples and $t = 4$ intervals, which requires approximately 190,000 samples for a stable convergence at $\kappa_N < 0.02$. Although both of the structural "switch" uncertainties

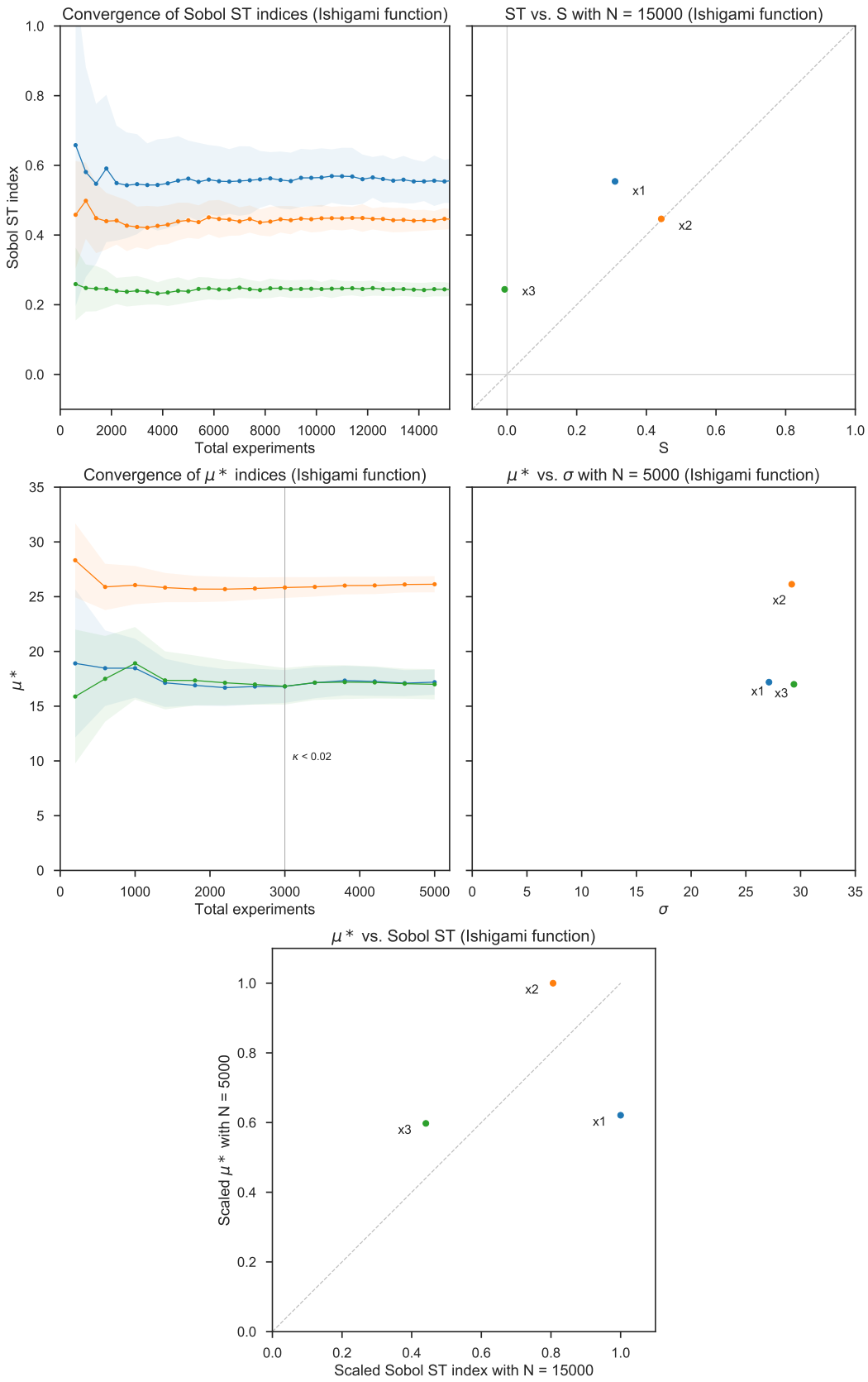


Figure 3: Results of Sobol (top panel) and elementary effects (middle panel) methods for the Ishigami test function. The vertical line indicates the $\kappa < 0.02$ convergence criterion for the $\mu*$ indices.

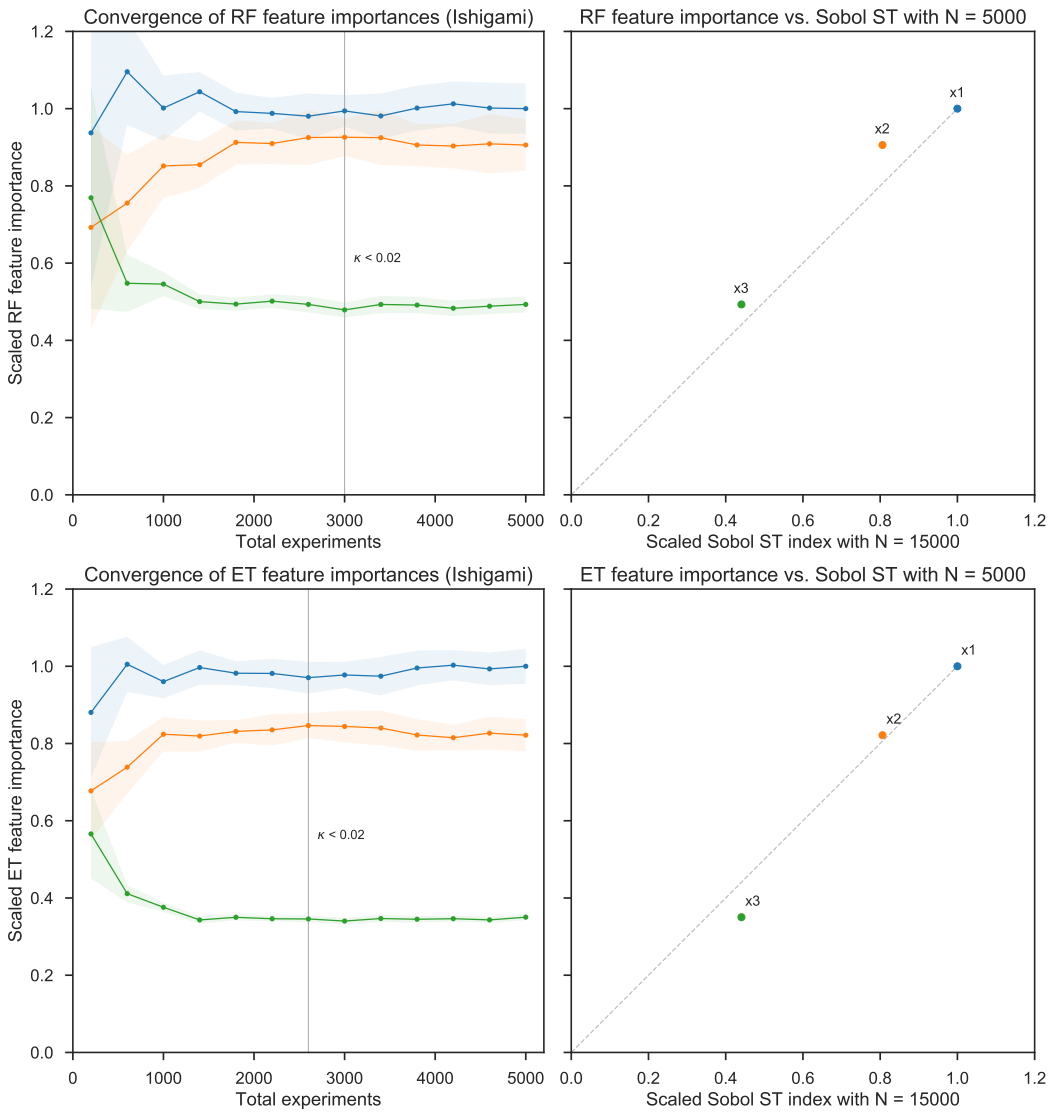


Figure 4: Estimation of MDI variable importances with the random forests (top panel) and Extra-Trees (bottom panel) techniques for the Ishigami test function. Vertical lines indicate the $\kappa < 0.02$ convergence criterion.

are correctly ranked by $\mu*$, their relative estimated importance is less stable than the continuous uncertainties across sample sizes.

Figure 9 shows the convergence of the MDI variable importances for the random forests (top panel, left) and Extra-Trees (bottom panel, left) techniques over a Latin Hypercube sample, using 30 bootstrap resamples to estimate confidence bounds. Both algorithms are parameterized with $T = 100$ trees, $m \approx p/3 = 6$, and a stopping criterion of $N_{leaf} = 6$. The right panels compare the mean estimated importances (scaled relative to the highest MDI at $N = 150,000$) with the scaled reference ST indices.

As with the Ishigami function, Extra-Trees are more accurate than random forests and the Morris $\mu*$ indices for approximating ST; under the parameterization used, random forests present a higher error relative to ST than the Morris $\mu*$ indices. ET and random forests converge more quickly than the $\mu*$ indices (in particular for the "switch" uncertainties), with a largely stable variable ranking for $N > 10,000$, and a convergence criterion $\kappa_N < 0.02$ above 80,000 samples.

Figure 10 shows the influence of the tuning parameters on RMSE and MBE, compared to the reference scaled ST values. In this application, totally randomized ($m = 1$), fully grown ($N_{leaf} = 1$) trees perform significantly worse. The assumption of $m \approx p/3 = 6$ provides consistent performance, and the mean bias can be tuned by adjusting the value of N_{leaf} in a range of approximately 1 to 16 without introducing a larger error. N_{leaf} has a similar effect as in the Ishigami-Homma test case, with relatively smaller trees having a smaller negative bias.

As evidenced by the large difference between the ST and S indices, higher-order interactions are influential for output behavior. The left panel of Figure 11 shows second-

Table 1: Input variables for H1N1 flu model

Name	ID	Min.	Max.
Structural switch on immunity	immunity_switch	{0,1}	
Structural switch on contact rate lookup function	lookup_switch	{0,1,2,3}	
Additional seasonal immune population fraction - region 1	x11	0.1	0.5
Additional seasonal immune population fraction - region 2	x12	0.1	0.5
Fatality rate - region 1	x21	0.01	0.1
Fatality rate - region 2	x22	0.01	0.1
Initial immune fraction of the population - region 1	x31	0.1	0.5
Initial immune fraction of the population - region 2	x32	0.1	0.5
Normal interregional contact rate	x41	0.1	0.9
Permanent immune population fraction - region 1	x51	0.1	0.5
Permanent immune population fraction - region 2	x52	0.1	0.5
Recovery time - region 1	x61	0.1	0.8
Recovery time - region 2	x62	0.1	0.8
Root contact rate - region 1	x81	1	10
Root contact rate - region 2	x82	1	10
Infection rate - region 1	x91	0.01	0.1
Infection rate - region 2	x92	0.01	0.1
Normal contact rate - region 1	x101	10	100
Normal contact rate - region 2	x102	10	70

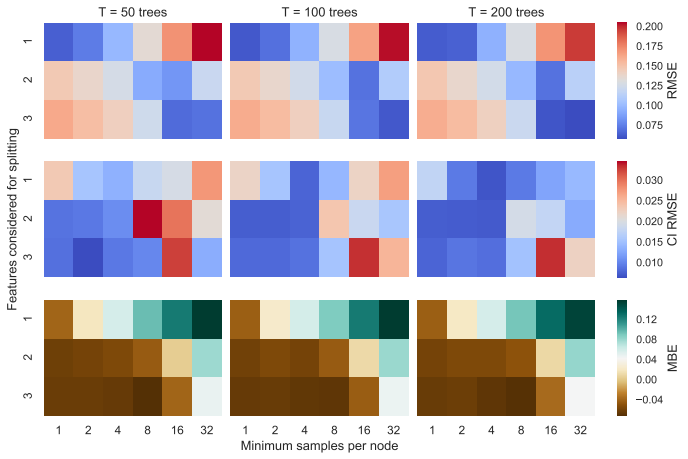


Figure 5: Extra-Trees performance relative to ST across key tuning parameters, for Ishigami function (N=3000). The figure shows RMSE (top three panels), bootstrapped confidence interval on RMSE (middle), and MBE (bottom panels) across a range of values for the number of trees T, the number of features considered for splitting m (subplot rows), and the minimum number of samples per node N_{leaf} (subplot columns).

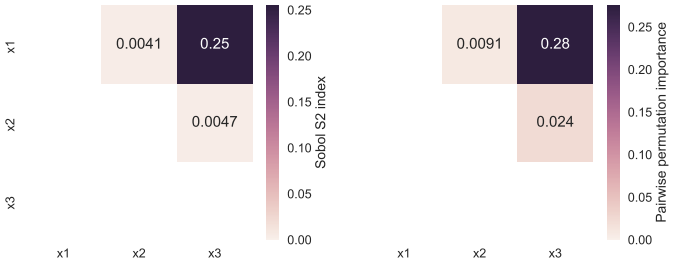


Figure 6: Comparison of pairwise variable interactions in the Ishigami function, using Sobol S2 indices (left) and Extra-Trees MDA pairwise permutation importances (right).

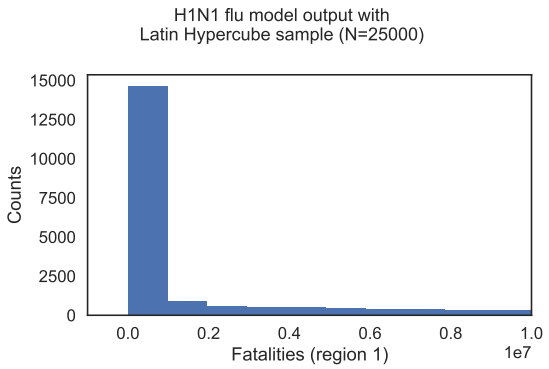


Figure 7: Output distribution for H1N1 model.

order interaction importances as estimated by the Sobol S2 indices, for the same sample size of $N = 800,000$. The right panel presents the mean pairwise MDA interaction importances estimated with Extra-Trees (with 30 bootstrap resamples on a 50,000 sample set). These estimated importances for each interacting pair are plotted in the bottom panel, after scaling relatively to the highest value in each set (S2 and MDA).

The interpretation of these results should take into account the numerical sensitivity of the reference S2 results. As shown on the left panel, each of the second-order inter-

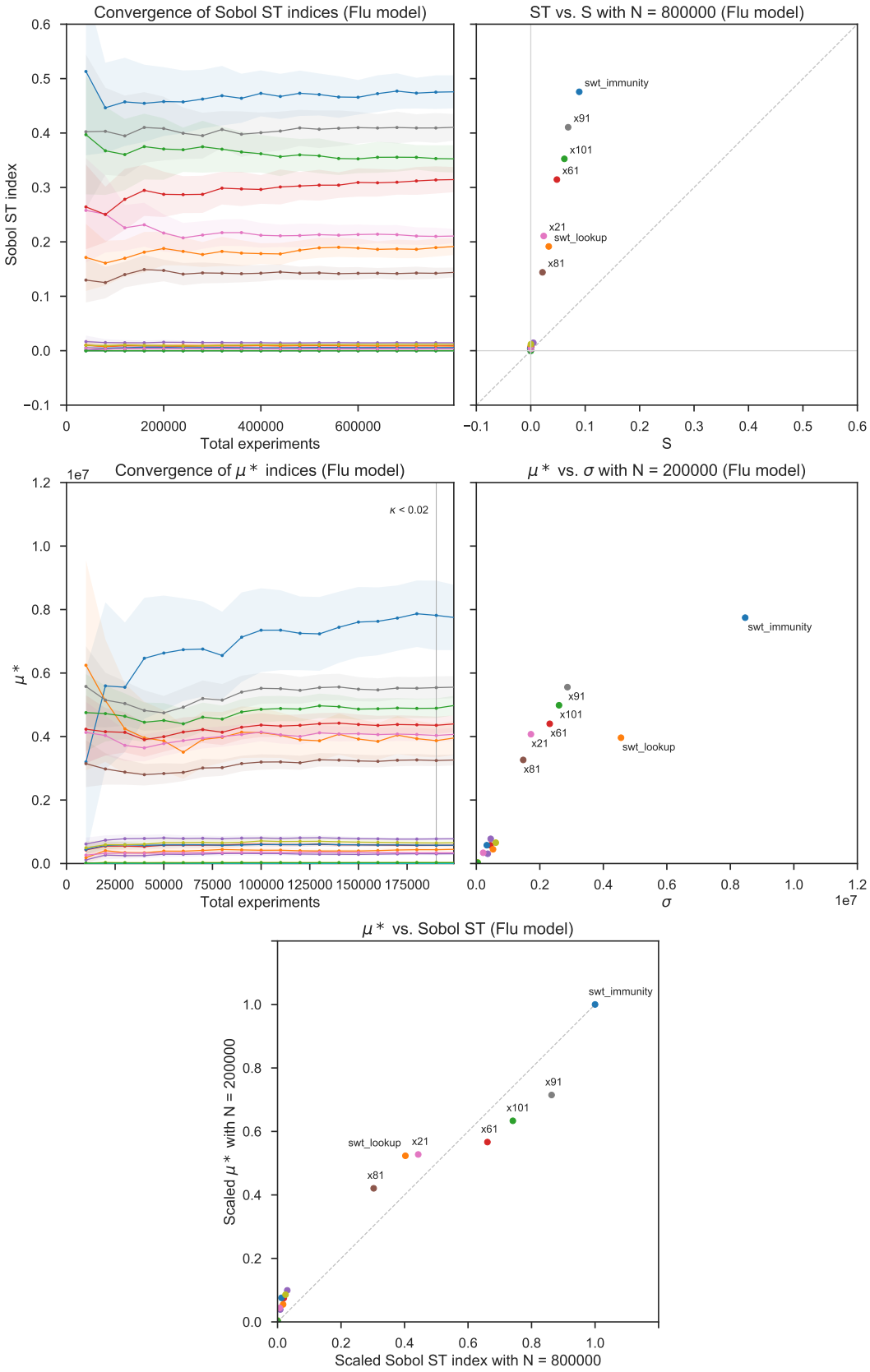


Figure 8: Results of Sobol (top panel) and elementary effects (middle panel) methods for H1N1 flu model. The vertical line indicates the $\kappa < 0.02$ convergence criterion for the $\mu*$ indices.

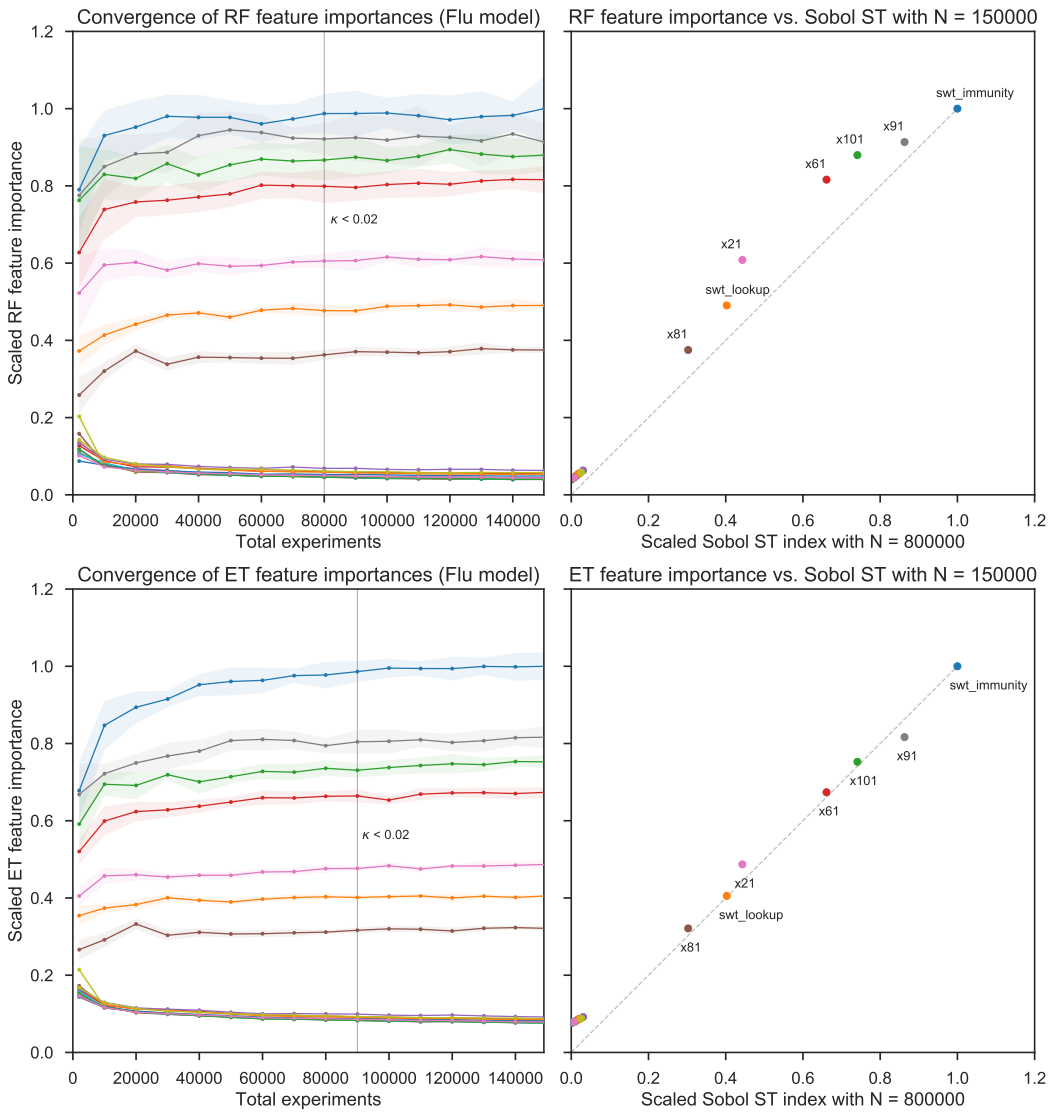


Figure 9: Estimation of MDI variable importances with the random forests (top panel) and Extra-Trees (bottom panel) techniques for H1N1 flu model. Vertical lines indicate the $\kappa < 0.02$ convergence criterion.

action terms only contributes a small portion of variance, which is typically smaller than the 95% confidence interval provided by SALib. This remains the case at significantly larger sample sizes ($N > 1e6$). The bottom panel illustrates this result with light gray markers for values of the S2 indices which are smaller than the estimated confidence interval, and are therefore likely to be unreliable. Nonetheless, the pairwise permutation generally performs well for identifying more significant interactions, for which the S2 index is outside the confidence interval (e.g. between the infection rate *x91* and other parameters to which it is structurally related in the model, such as the normal contact rate *x101* and the structural switch on immunity).

3.3. CDICE integrated assessment model

The last case study uses the CDICE model (Butler et al., 2014), which replicates the outcomes of the globally-aggregated DICE-2007 integrated assessment model (Nordhaus, 2007) under given policy scenarios. This model represents a simplified global economy, coupled with a 3-reservoir carbon cycle model and a 2-reservoir climate model; the feedbacks between these components lead to highly non-linear outputs. When used in an optimization setting, DICE yields an optimal policy for the time series of GHG emission control rates and investments that maximize the discounted utility of consumption over the modelled time frame. Conversely, the CDICE simulation version introduced by Butler et al. (2014) can be used to evaluate the impact of exogenous uncertainties on the performance of policy scenarios.

The full version of this model uses 31 exogenous input variables, shown with their input ranges in Table 2. With uniform input distributions, these assumptions yield the output distribution shown in Figure 12 for the NPV of

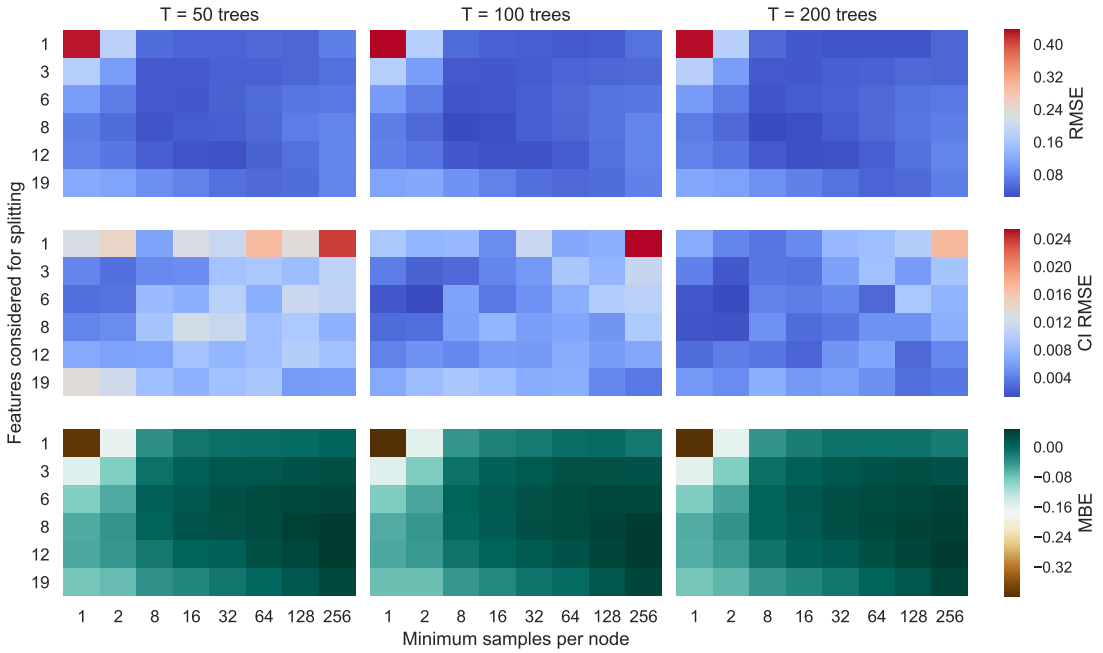


Figure 10: Extra-Trees performance relative to ST across key tuning parameters, for H1N1 flu model (N=50,000). The figure shows RMSE (top three panels), bootstrapped confidence interval on RMSE (middle), and MBE (bottom panels) across a range of values for the number of trees T, the number of features considered for splitting m (subplot rows), and the minimum number of samples per node N_{leaf} (subplot columns).

abatement costs. This outcome will be used for the analysis due to its relatively quicker convergence with Sobol measures.

Figure 13 shows the convergence of ST (top panel) and Morris $\mu*$ (middle panel) indices, for the NPV of abatement costs. Due to the large number of parameters, the Sobol indices require $N > 9e6$ for a stable ranking. As shown by the low values of the first-order S indices relative to ST, higher-order interactions are significant for the behavior of this outcome.

The Morris results use a sampling of $k = 10$ levels and $\Delta = k/[2(k-1)]$, yielding a mostly stable estimation of variable rankings above $N > 150,000$. However, the bottom panel of Figure 13 shows several inconsistencies in the variable rankings given by $\mu*$ compared to scaled ST values.

Figure 14 shows the convergence of the MDI variable importances for the random forests (top panel, left) and Extra-Trees (bottom panel, left) techniques over 30 resamples on a Latin Hypercube sample, for the same outcome. The algorithms are parameterized with $T = 100$ trees, $m \approx p/3 = 10$, and $N_{leaf} = 8$. The right panels compare the scaled mean estimated importances with the scaled reference ST indices.

The Extra-Trees variable rankings mostly stabilize for $N > 100,000$, with a better approximation of relative ST values than the Morris $\mu*$ indices. For random forests, however, the variable ranking shows some discrepancies with the ST results. Appendix A presents a convergence analysis with $\Delta N = 16,000$ samples and $t = 4$ intervals;

Table 2: Input variables for CDICE model (baseline scenario)

ID	Min.	Max.
popasym	5000	13000
gpop0	0.2	0.35
ga0	0.092	0.2
dela	0.001	0.016
sig0	0.13364	0.15273
gsigma	-0.16	-0.07
dsig	0.001	0.003
dsig2	0	0.0002
eland0	9	15
dtree	0.05	0.2
b12	0.155288	0.223288
b23	0.025	0.1
fex0	-0.3	0
fex1	-0.2	0.5
t2xco2	1	8
fco22x	3.6	3.9
c1	0.2	0.24
c3	0.27	0.33
c4	0.045	0.055
a1	0	0.001
a2	0.002255	0.003123
a3	1.5	3
pback0	0.6	3
theta2	2.6	3
backrat	1.5	2.5
gback	0.045	0.055
partfrac1	0.1	1
partfrac2	0.25372	1
partfracn	0.5	1
dpartfrac	0	0.25
saverate0	0.2	0.24

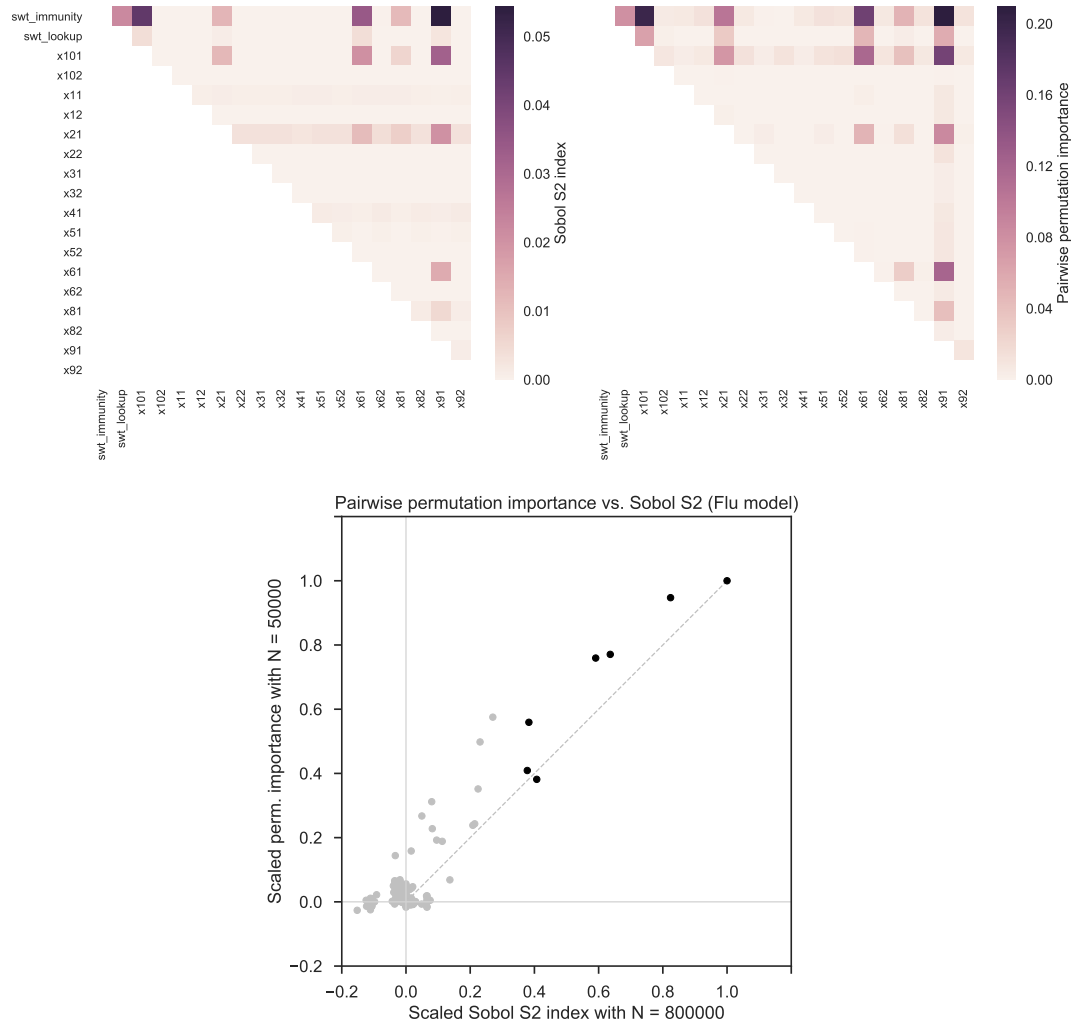


Figure 11: Comparison of pairwise variable interactions in the H1N1 flu model, using Sobol S2 indices (left) and Extra-Trees MDA pairwise permutation importances (right). The bottom panel plots scaled Sobol S2 and Extra-Trees interaction importances against each other, with light gray markers corresponding to S2 values which are within the confidence bounds estimated by SALib.

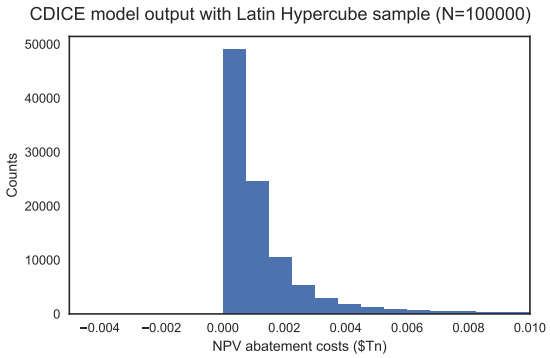


Figure 12: Output distribution for CDICE model (NPV of abatement costs).

both ensemble methods, as well as Morris indices, reach a criterion of $\kappa_N < 0.02$ with approximately 150,000 samples.

Figure 15 shows the performance of the Extra-Trees estimation across the tuning parameters, compared to the scaled relative ST values. As with the H1N1 flu model, highly randomized and fully developed trees do not perform adequately, but the estimated importances are robust in a range of m/p of 0.3 to 0.6 ($m = 9$ to $m = 18$). The MBE metric also presents a comparable pattern to the H1N1 model results, with larger values of m/p leading to a negative bias unless compensated by a larger stopping criterion N_{leaf}.

The left panel of Figure 16 shows second-order interaction importances estimated by the Sobol S2 indices, with the same sample of $N = 9.22e6$. The right panel presents the mean pairwise MDA interaction importances estimated with Extra-Trees (with 30 bootstrap resamples, on a 100,000 sample set).

The scaled estimated importances for each interacting pair are plotted against each other in the bottom panel.

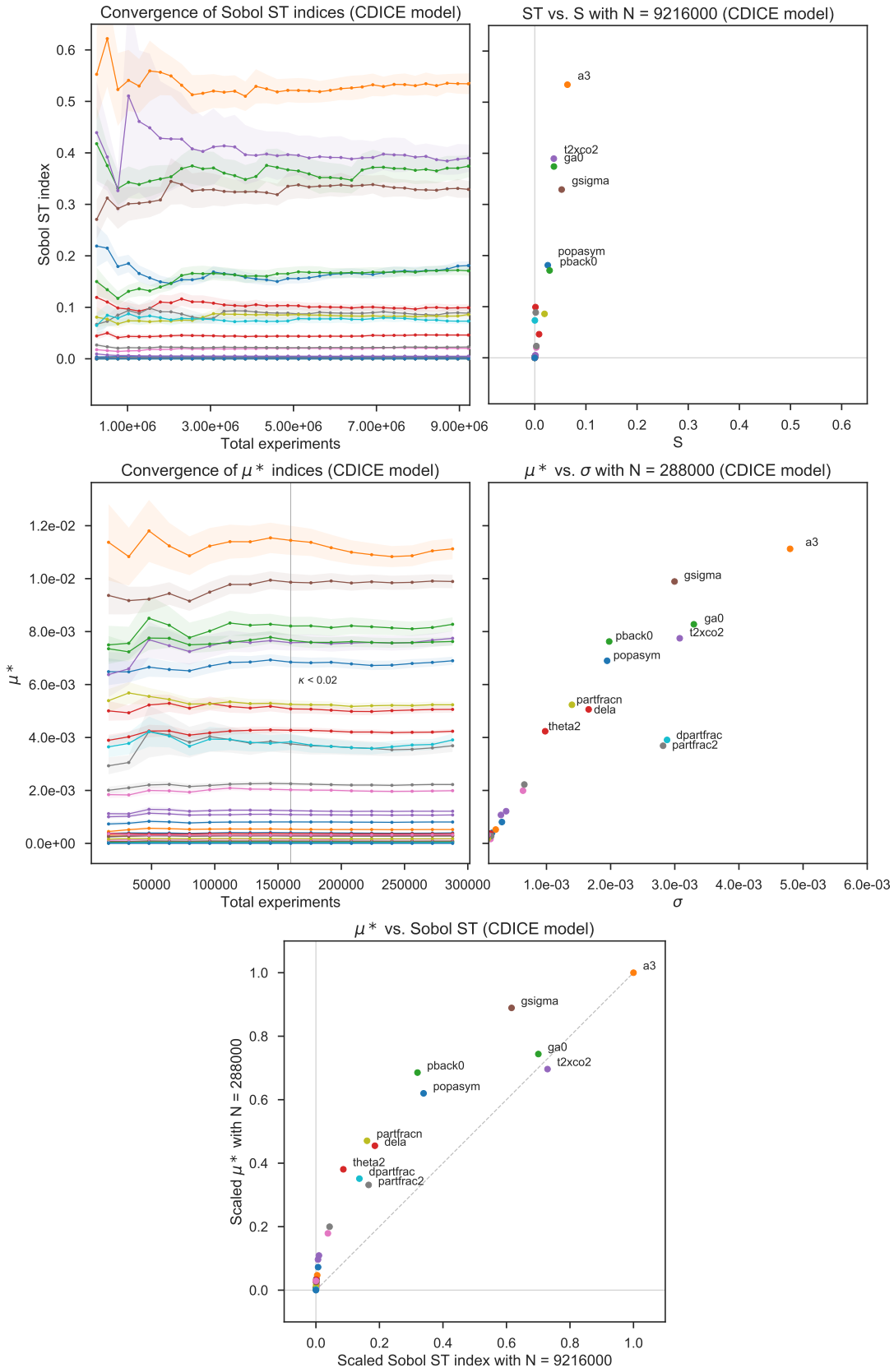


Figure 13: Results of Sobol (top panel) and elementary effects (middle panel) methods for the CDICE model (NPV of abatement costs). The vertical line indicates the $\kappa < 0.02$ convergence criterion for the $\mu*$ indices.

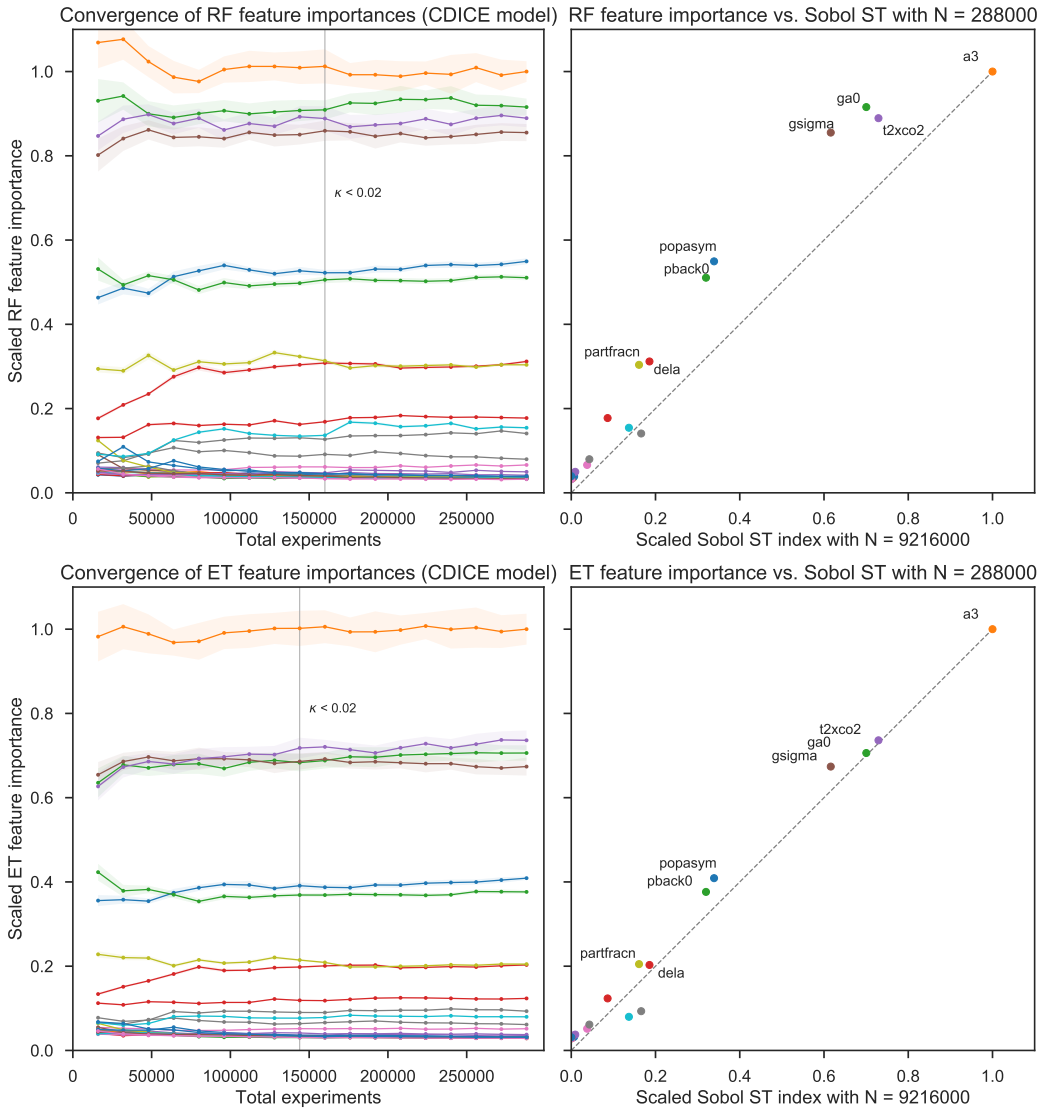


Figure 14: Estimation of MDI variable importances with random forests (top panel) and Extra-Trees (bottom panel) for the CDICE model (NPV of abatement costs). Vertical lines indicate the $\kappa < 0.02$ convergence criterion.

The most significant pairwise interactions appear to be identified by the permutation measure, such as the interactions involving the *a3* exponent of the model's climate damage function. As with the H1N1 flu model, however, the S2 indices may be numerically unreliable due to relatively large confidence bounds. It can be noted that some of the S2 indices present negative values, which is clearly a numerical artifact. The analysis was in this case limited by the computational costs of the larger input samples which would be required for a stable estimation of S2 indices.

4. Discussion and conclusions

This paper assessed the performance of decision tree-based ensemble methods for the estimation of global sensitivity analysis measures, focusing on the random forests and Extra-Trees algorithms. Compared to the Morris elementary effects method which is commonly used for screening non-influential variables, the Extra-Trees technique in particular performed well to estimate relative Sobol ST total effect indices, using the Mean Decrease Impurity (MDI) metric for variable importance. Across the three case studies presented in the paper, Extra-Trees therefore outperformed the Morris $\mu*$ indices on measures of RMSE and variable ranking error, compared to the proportional values of ST indices. For the more complex H1N1 and CDICE cases, a sample size of less than 10% of the Sobol sample size was sufficient for a stable estimation of variable rankings. Furthermore, a pairwise Mean Decrease Accuracy (MDA) permutation metric allowed for the study of variable interactions with Extra-Trees. While the more common random forests algorithm performed well on the benchmark Ishigami-Homma test function, it was less reliable in the more complex cases.

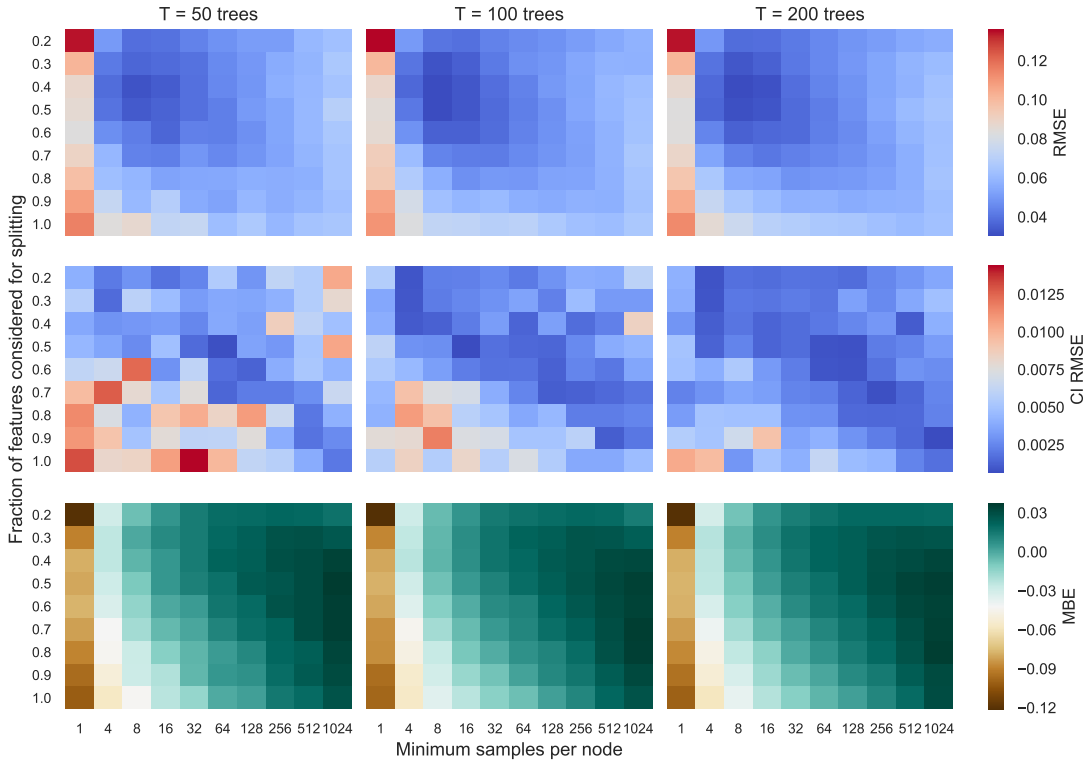


Figure 15: Extra-Trees performance relative to ST across key tuning parameters, for CDICE model (NPV of abatement costs, N=100,000). The figure shows RMSE (top three panels), bootstrapped confidence interval on RMSE (middle), and MBE (bottom panels) across a range of values for the number of trees T, the fraction of features considered for splitting m/p (subplot rows), and the minimum number of samples per node N_{leaf} (subplot columns).

The Extra-Trees estimation of variable importances was systematically assessed across a range of tuning parameters for the algorithm. Based on the case studies and previous literature (Hastie et al., 2009), values of $T = 100$ trees and a number of splitting features $m \approx p/3$ appear to be suitable starting points. The choice of a stopping criterion significantly affects bias, which is especially relevant for a screening application. In order to avoid possible type II errors, a conservative guideline would be to use fully developed trees ($N_{leaf} = 1$) for $N \approx 1000$, then to introduce a stopping criterion $N_{leaf} \propto \sqrt{N}$ for larger samples. Values of 6 and 8 for N_{leaf} thus performed well for $N = 50,000$ and $N = 100,000$ with the H1N1 and CDICE models.

The variable importance metrics provided by the tree-based methods can be assessed in relation to the criteria summarized by Pianosi and Wagener (2015) for an "ideal" sensitivity metric. As such, the MDI and MDA metrics largely meet these criteria, by being suitable for global sampling designs, independent of model structure, relatively easy to implement numerically, and stable across sample sizes and bootstrap resamples. Compared to Sobol indices, a downside of these metrics is the lack of a straightforward mathematical interpretation, as they only provide information about the relative importance of inputs, rather than their direct effect on output variance. However, for practical purposes, the accurate estimation of relative total effects should be sufficient for a factor fixing application. Compared to the $\mu*$ indices for elementary effects (which share this limitation on mathematical interpretability), MDI more accurately estimates the relative values of ST indices, is suitable for non-scalar inputs, and appears more stable at smaller sample sizes. MDA additionally estimates relative pairwise interaction effects, which are not identified by the elementary effects σ indices. MDI and MDA can also be computed from generic Latin Hypercube or Monte Carlo sampling designs. This makes it easier to reuse existing datasets which may have been generated from an uncertainty analysis, or to combine the ensemble methods with other analysis techniques in a multi-method analysis framework.

In parallel, Appendix B compares the total runtime required to compute importance metrics (as well as the total model evaluation runtime), for the more complex H1N1 and CDICE cases. With the software libraries used in this work, the MDI and Morris indices have a similar computation runtime at a given sample size, with the computation runtime largely scaling in proportion to sample size N. The pairwise MDA metric is slightly costlier and scales with the square of the number of input variables p. In the presented cases, the analysis runtime for these metrics was typically small relative to the total evaluation runtime re-

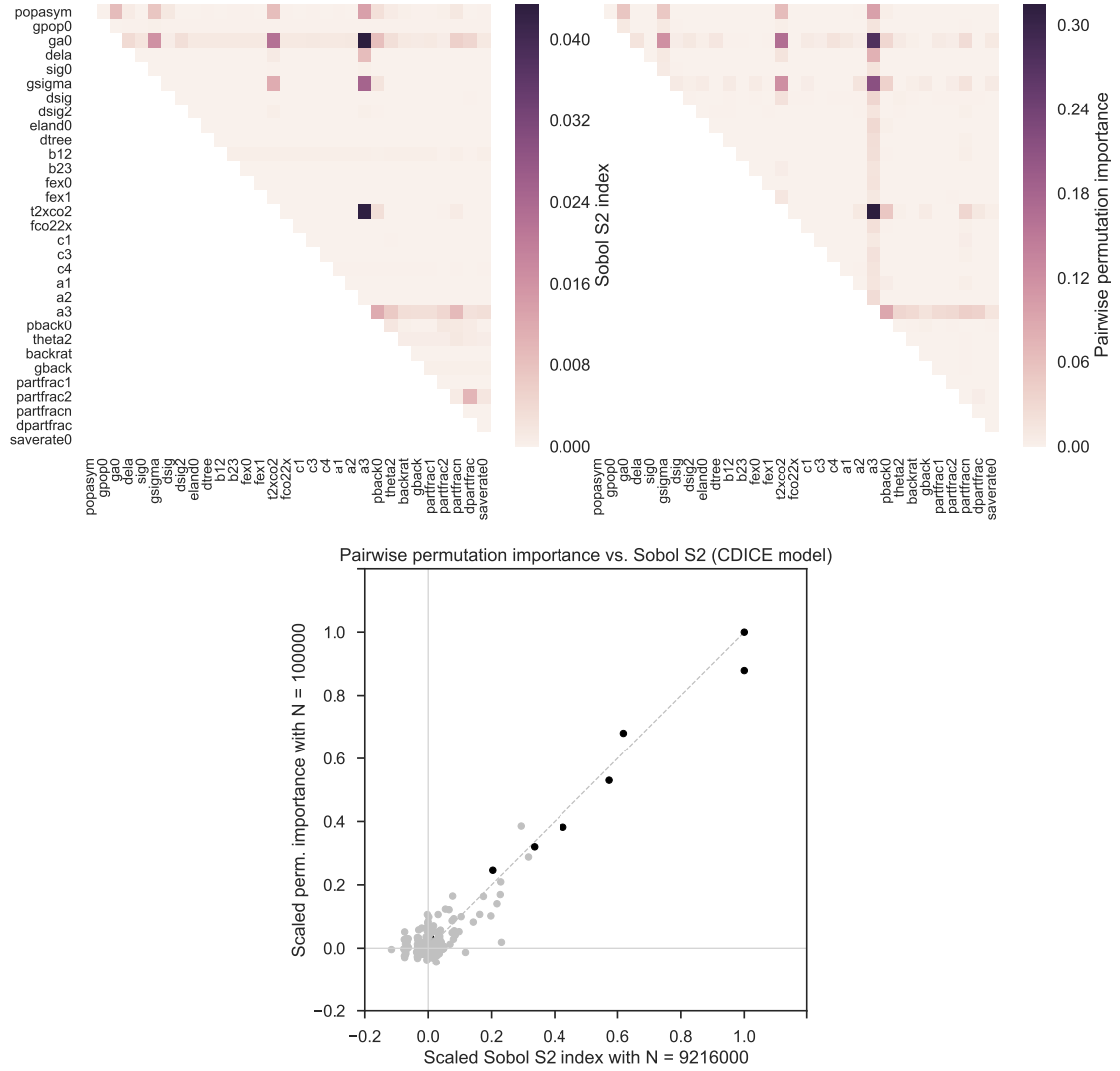


Figure 16: Comparison of pairwise variable interactions in the CDICE model (NPV of abatement costs), using Sobol S2 indices (left) and Extra-Trees pairwise permutation importances (right). The bottom panel plots scaled Sobol S2 and Extra-Trees interaction importances against each other, with light gray markers corresponding to S2 values which are within the confidence bounds estimated by SALib.

quired by the simulation models; it was also significantly smaller than the computation time for Sobol indices. In an analysis setting focused on estimating the relative importance of variables and their interactions, the smaller sample size required by the MDI and MDA metrics can therefore significantly reduce the overall computational cost of the analysis.

In a broader perspective, however, it should be noted that all of the techniques evaluated in this paper followed a variance-based approach to global sensitivity analysis – either by directly calculating variance-based indices with the Sobol method, or by approximating the proportional importance of the latter with elementary effects and tree-based methods. As described by Pianosi and Wagener (2015), variance may not be an appropriate measure of uncertainty for multi-modal or highly skewed output distributions; in these cases, an approach based on the probability density function of the output may be preferable. This property was demonstrated by the authors with a simple non-linear model, for which variance-based GSA did not properly distinguish variable importances. This has clear implications for the cases studied in this paper, as the outputs of the H1N1 and CDICE models showed highly skewed distributions under the uncertainty ranges used to generate input samples. Given the possible limitations of variance-based methods under such conditions, it would be useful to compare the reference Sobol results with a density-based method, and to evaluate the performance of Extra-Trees across a wider range of output distribution shapes.

5. Acknowledgments

This research was supported by the Netherlands Organization for Scientific Research (NWO) under the project Aquifer Thermal Energy Storage Smart Grids (ATES-SG), grant number 408-13-030. We thank three anonymous reviewers for their constructive comments towards improving this paper.

References

Almeida, S., Holcombe, E.A., Pianosi, F., Wagener, T., 2017. Dealing with deep uncertainties in landslide modelling for disaster risk reduction under climate change. Natural Hazards and Earth System Sciences 17, 225.

Altmann, A., Tolosi, L., Sander, O., Lengauer, T., 2010. Permutation importance: a corrected feature importance measure. Bioinformatics 26, 1340–1347. URL: https://academic.oup.com/bioinformatics/article/26/10/1340/193348/Permutation-importance-a-corrected-feature, doi:10.1093/bioinformatics/btq134.

Baroni, G., Tarantola, S., 2014. A General Probabilistic Framework for uncertainty and global sensitivity analysis of deterministic models: A hydrological case study. Environmental Modelling & Software 51, 26–34. URL: http://www.sciencedirect.com/science/article/pii/S1364815213002211, doi:10.1016/j.envsoft.2013.09.022.

Blum, A.L., Langley, P., 1997. Selection of relevant features and examples in machine learning. Artificial Intelligence 97, 245–271. URL: http://www.sciencedirect.com/science/article/pii/S0004370297000635, doi:10.1016/S0004-3702(97)00063-5.

Borgonovo, E., 2007. A new uncertainty importance measure. Reliability Engineering & System Safety 92, 771–784. URL: http://www.sciencedirect.com/science/article/pii/S0951832006000883, doi:10.1016/j.ress.2006.04.015.

Borgonovo, E., Lu, X., Plischke, E., Rakovec, O., Hill, M., 2017. Making the most out of a hydrological model data set: Sensitivity analyses to open the model black box. Water Resources Research 53, 7933–7950. URL: https://agupubs.onlinelibrary.wiley.com/doi/full/10.1002/2017WR020767, doi:10.1002/2017WR020767.

Breiman, L., 2001. Random Forests. Machine Learning 45, 5–32. doi:10.1023/A:1010933404324.

Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J., 1984. Classification and regression trees. URL: http://cds.cern.ch/record/2253780.

Bryant, B.P., Lempert, R.J., 2010. Thinking inside the box: A participatory, computer-assisted approach to scenario discovery. Technological Forecasting and Social Change 77, 34–49. URL: http://www.sciencedirect.com/science/article/pii/S004016250900105X, doi:10.1016/j.techfore.2009.08.002.

Bureau, A., Dupuis, J., Falls, K., Lunetta, K.L., Hayward, B., Keith, T.P., Van Eerdewegh, P., 2005. Identifying SNPs predictive of phenotype using random forests. Genetic Epidemiology 28, 171–182. URL: http://onlinelibrary.wiley.com/doi/10.1002/gepi.20041/abstract, doi:10.1002/gepi.20041.

Butler, M.P., Reed, P.M., Fisher-Vanden, K., Keller, K., Wagener, T., 2014. Identifying parametric controls and dependencies in integrated assessment models using global sensitivity analysis. Environmental Modelling & Software 59, 10–29. URL: http://www.sciencedirect.com/science/article/pii/S1364815214001327, doi:10.1016/j.envsoft.2014.05.001.

Campolongo, F., Cariboni, J., Saltelli, A., 2007. An effective screening design for sensitivity analysis of large models. Environmental Modelling & Software 22, 1509–1518. URL: http://www.sciencedirect.com/science/article/pii/S1364815206002805, doi:10.1016/j.envsoft.2006.10.004.

Frey, H.C., Patil, S.R., 2002. Identification and review of sensitivity analysis methods. Risk Analysis: An Official Publication of the Society for Risk Analysis 22, 553–578.

Geurts, P., Ernst, D., Wehenkel, L., 2006. Extremely randomized trees. Machine Learning 63, 3–42. URL: http://link.springer.com/article/10.1007/s10994-006-6226-1, doi:10.1007/s10994-006-6226-1.

Guivarch, C., Rozenberg, J., Schweizer, V., 2016. The diversity of socio-economic pathways and CO2 emissions scenarios: Insights from the investigation of a scenarios database. Environmental Modelling & Software 80, 336–353. URL: http://www.sciencedirect.com/science/article/pii/S1364815216300706, doi:10.1016/j.envsoft.2016.03.006.

Guyon, I., Elisseeff, A., 2003. An Introduction to Variable and Feature Selection. J. Mach. Learn. Res. 3, 1157–1182. URL: http://dl.acm.org/citation.cfm?id=944919.944968.

Hapfelmeier, A., Ulm, K., 2013. A new variable selection approach using Random Forests. Computational Statistics & Data Analysis 60, 50–69. URL: http://www.sciencedirect.com/science/article/pii/S0167947312003490, doi:10.1016/j.csda.2012.09.020.

Harper, E.B., Stella, J.C., Fremier, A.K., 2011. Global sensitivity analysis for complex ecological models: a case study of riparian cottonwood population dynamics. Ecological Applications: A Publication of the Ecological Society of America 21, 1225–1240.

Hastie, T., Tibshirani, R., Friedman, J., 2009. 15 - Random Forests, in: The Elements of Statistical Learning. Springer New York. Springer Series in Statistics, pp. 587–604.

Helton, J.C., Oberkampf, W.L., 2004. Alternative representations of epistemic uncertainty. Reliability Engineering & System Safety 85, 1–10. URL: http://www.sciencedirect.com/science/article/pii/S0951832004000481, doi:10.1016/j.ress.2004.03.001.

Herman, J., Usher, W., 2017. SALib: An open-source Python library for Sensitivity Analysis. The Journal of Open Source Software 2. URL: http://joss.theoj.org/papers/10.21105/joss.00097, doi:10.21105/joss.00097.

Herman, J.D., Reed, P.M., Wagener, T., 2013. Time-varying sensitivity analysis clarifies the effects of watershed model formulation on model behavior. Water Resources Research 49, 1400–1414. URL: http://onlinelibrary.wiley.com/doi/10.1002/wrcr.20124/abstract, doi:10.1002/wrcr.20124.

Homma, T., Saltelli, A., 1996. Importance measures in global sensitivity analysis of nonlinear models. Reliability Engineering & System Safety 52, 1–17. URL: http://www.sciencedirect.com/science/article/pii/0951832096000026, doi:10.1016/0951-8320(96)00002-6.

Hothorn, T., Hornik, K., Strobl, C., Zeileis, A., 2017. party. URL: https://cran.r-project.org/web/packages/party/index.html.

Ishigami, T., Homma, T., 1990. An importance quantification technique in uncertainty analysis for computer models, in: , First International Symposium on Uncertainty Modeling and Analysis, 1990. Proceedings, pp. 398–403. doi:10.1109/ISUMA.1990.151285.

Ishwaran, H., 2007. Variable importance in binary regression trees and forests. Electronic Journal of Statistics 1, 519–537. URL: http://projecteuclid.org/euclid.ejs/1195157166, doi:10.1214/07-EJS039.

Kleijnen, J.P., 2009. Factor Screening in Simulation Experiments: Review of Sequential Bifurcation, in: Advancing the Frontiers of Simulation. Springer, pp. 153–167.

Kohavi, R., John, G.H., 1997. Wrappers for feature subset selection. Artificial Intelligence 97, 273–324. URL: http://www.sciencedirect.com/science/article/pii/S000437029700043X, doi:10.1016/S0004-3702(97)00043-X.

Kwakkel, J.H., 2017. The Exploratory Modeling Workbench: An open source toolkit for exploratory modeling, scenario discovery, and (multi-objective) robust decision making. Environmental Modelling & Software 96, 239–250.

Kwakkel, J.H., Jaxa-Rozen, M., 2016. Improving scenario discovery for handling heterogeneous uncertainties and multinomial classified outcomes. Environmental Modelling & Software 79, 311–321. URL: http://www.sciencedirect.com/science/article/pii/S1364815215301092, doi:10.1016/j.envsoft.2015.11.020.

Lazar, C., Taminau, J., Meganck, S., Steenhoff, D., Coletta, A., Molter, C., Schaetzen, V.d., Duque, R., Bersini, H., Nowe, A.,

2012. A Survey on Filter Techniques for Feature Selection in Gene Expression Microarray Analysis. IEEE/ACM Transactions on Computational Biology and Bioinformatics 9, 1106–1119. doi:10.1109/TCBB.2012.33.

Liu, Q., Homma, T., 2009. A new computational method of a moment-independent uncertainty importance measure. Reliability Engineering & System Safety 94, 1205–1211. URL: http://www.sciencedirect.com/science/article/pii/S0951832008002573, doi:10.1016/j.ress.2008.10.005.

Louppe, G., 2014. Understanding Random Forests: From Theory to Practice. PhD thesis. Universite de Liege. URL: http://orbi.ulg.ac.be/handle/2268/170309.

Morris, M.D., 1991. Factorial Sampling Plans for Preliminary Computational Experiments. Technometrics 33, 161–174. URL: http://www.tandfonline.com/doi/abs/10.1080/00401706.1991.10484804, doi:10.1080/00401706.1991.10484804.

Nordhaus, W., 2007. Accompanying Notes and Documentation on Development of DICE-2007 Model: Notes on DICE-2007. delta. v8 as of September 21, 2007. Yale University, New Haven, NE, USA .

Nossent, J., Elsen, P., Bauwens, W., 2011. Sobol sensitivity analysis of a complex environmental model. Environmental Modelling & Software 26, 1515–1525. URL: http://www.sciencedirect.com/science/article/pii/S1364815211001939, doi:10.1016/j.envsoft.2011.08.010.

Pappenberger, F., Beven, K.J., Ratto, M., Matgen, P., 2008. Multi-method global sensitivity analysis of flood inundation models. Advances in Water Resources 31, 1–14. URL: http://www.sciencedirect.com/science/article/pii/S0309170807000747, doi:10.1016/j.advwatres.2007.04.009.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E., 2011. Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research 12, 2825–2830.

Pianosi, F., Iwema, J., Rosolem, R., Wagener, T., 2017. Chapter 7 - A Multimethod Global Sensitivity Analysis Approach to Support the Calibration and Evaluation of Land Surface Models, in: Petropoulos, G.P., Srivastava, P.K. (Eds.), Sensitivity Analysis in Earth Observation Modelling. Elsevier, pp. 125–144. URL: https://www.sciencedirect.com/science/article/pii/B9780128030110000070, doi:10.1016/B978-0-12-803011-0.00007-0.

Pianosi, F., Wagener, T., 2015. A simple and efficient method for global sensitivity analysis based on cumulative distribution functions. Environmental Modelling & Software 67, 1–11. URL: http://www.sciencedirect.com/science/article/pii/S1364815215000237, doi:10.1016/j.envsoft.2015.01.004.

Plischke, E., Borgonovo, E., Smith, C.L., 2013. Global sensitivity measures from given data. European Journal of Operational Research 226, 536–550. URL: http://www.sciencedirect.com/science/article/pii/S0377221712008995, doi:10.1016/j.ejor.2012.11.047.

Pruyt, E., Hamarat, C., 2010. The influenza H1n1v pandemic: an exploratory system dynamics approach, in: Proceedings of the 28th International Conference of the System Dynamics Society, Seoul, Korea, 25-29 July 2010.

Qi, Y., Bar-Joseph, Z., Klein-Seetharaman, J., 2006. Evaluation of different biological data and computational classification methods for use in protein interaction prediction. Proteins 63, 490–500. doi:10.1002/prot.20865.

Saltelli, A., 2002a. Making best use of model evaluations to compute sensitivity indices. Computer Physics Communications 145, 280–297.

Saltelli, A., 2002b. Sensitivity Analysis for Importance Assessment. Risk Analysis 22, 579–590. URL: http://onlinelibrary.wiley.com.tudelft.idm.oclc.org/doi/10.1111/0272-4332.00040/abstract, doi:10.1111/0272-4332.00040.

Saltelli, A., Annoni, P., 2010. How to avoid a perfunctory sensitivity analysis. Environmental Modelling & Software 25, 1508–1517. URL: http://www.sciencedirect.com/science/article/pii/S1364815210001180, doi:10.1016/j.envsoft.2010.04.012.

Saltelli, A., Ratto, M., Andres, T., Campolongo, F., Cariboni, J., Gatelli, D., Saisana, M., Tarantola, S., 2008. Global sensitivity analysis: the primer. John Wiley & Sons.

Saltelli, A., Tarantola, S., 2002. On the Relative Importance of Input Factors in Mathematical Models. Journal of the American Statistical Association 97, 702–709. URL: http://dx.doi.org/10.1198/016214502388618447, doi:10.1198/016214502388618447.

Simm, J., de Abril, I., 2015. ExtraTrees. URL: http://github.com/jaak-s/extraTrees.

Singh, R., Wagener, T., Crane, R., Mann, M.E., Ning, L., 2014. A vulnerability driven approach to identify adverse climate and land use change combinations for critical hydrologic indicator thresholds: Application to a watershed in Pennsylvania, USA. Water Resources Research 50, 3409–3427. URL: https://agupubs.onlinelibrary.wiley.com/doi/full/10.1002/2013WR014988, doi:10.1002/2013WR014988.

Sobol, I.M., 2001. Global sensitivity indices for nonlinear mathematical models and their Monte Carlo estimates. Mathematics and Computers in Simulation 55, 271–280. URL: http://www.sciencedirect.com/science/article/pii/S0378475400002706, doi:10.1016/S0378-4754(00)00270-6.

Strobl, C., Boulesteix, A.L., Kneib, T., Augustin, T., Zeileis, A., 2008. Conditional variable importance for random forests. BMC Bioinformatics 9, 307. URL: http://dx.doi.org/10.1186/1471-2105-9-307, doi:10.1186/1471-2105-9-307.

Strobl, C., Boulesteix, A.L., Zeileis, A., Hothorn, T., 2007. Bias in random forest variable importance measures: Illustrations, sources and a solution. BMC Bioinformatics 8, 25. URL: http://dx.doi.org/10.1186/1471-2105-8-25, doi:10.1186/1471-2105-8-25.

Strobl, C., Hothorn, T., Zeileis, A., 2009. Party on! A new, conditional variable importance measure available in the party package. The R Journal , 14–17URL: https://epub.ub.uni-muenchen.de/31165/.

Tang, T., Reed, P., Wagener, T., Van Werkhoven, K., 2006. Comparing sensitivity analysis methods to advance lumped watershed model identification and evaluation. Hydrology and Earth System Sciences Discussions 3, 3333–3395. URL: https://hal.archives-ouvertes.fr/hal-00298785.

Touzani, S., Busby, D., 2014. Screening Method Using the Derivative-based Global Sensitivity Indices with Application to Reservoir Simulator. Oil & Gas Science and Technology - Revue d'IFP Energies nouvelles 69, 619–632. doi:10.2516/ogst/2013195.

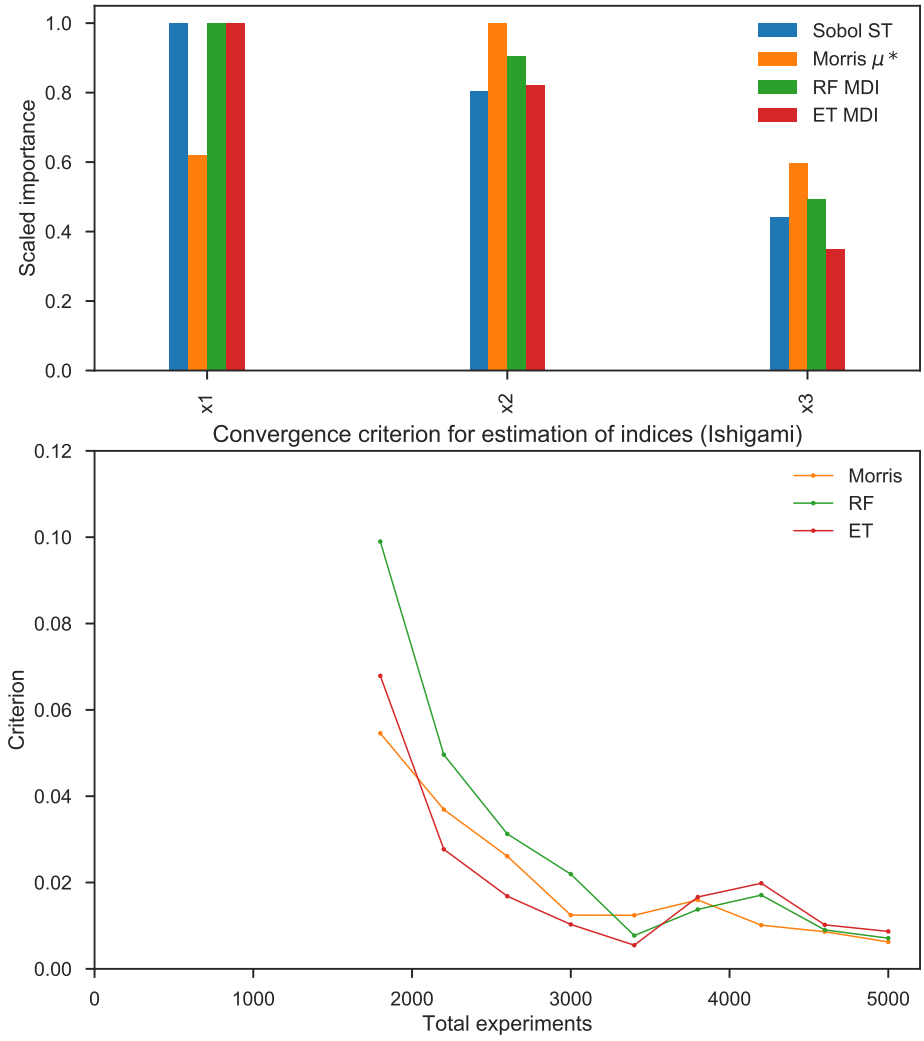
Wright, M.N., Ziegler, A., Konig, I.R., 2016. Do little interactions get lost in dark random forests? BMC Bioinformatics 17. URL: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4815164/, doi:10.1186/s12859-016-0995-8.

Appendix A. Variable importance metrics

Figures A.17, A.18 and A.19 present detailed results for the estimation of scaled variable importances in each case study.

Appendix B. Comparison of analysis runtimes

Table B.3 presents representative runtimes for each of the key analyses presented in the paper, using the EMA Workbench 1.1 library to sample and simulate experiments, and SALib 1.1.3 (for Sobol and Morris) and scikit-learn 0.18.1 (for Extra-Trees) to compute sensitivity indices. The analyses were performed on an Intel Xeon E5-2620 CPU with the Anaconda 3.6 Python 64-bit distribution.

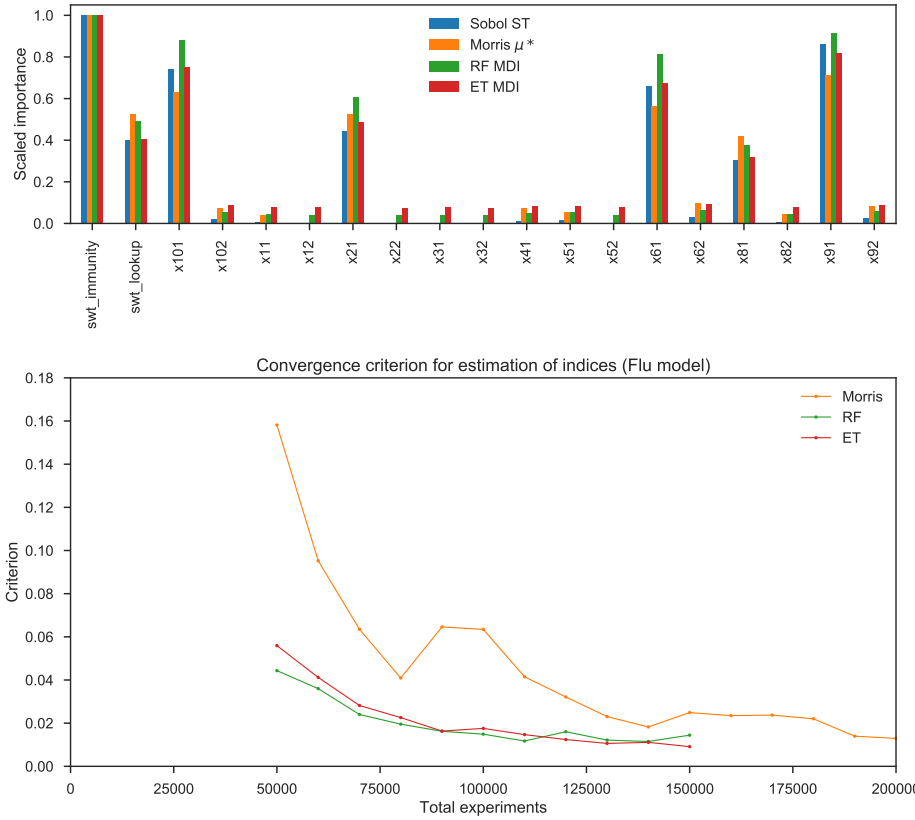


	Sobol ST (N=17000)	Morris $\mu*$ (N=5000)	RF MDI (N=5000)	ET MDI (N=5000)
RMSE	-	0.250	0.069	0.051
MBE	-	-0.016	-0.054	0.021

Figure A.17: Scaled variable importances and error measures for Ishigami function.

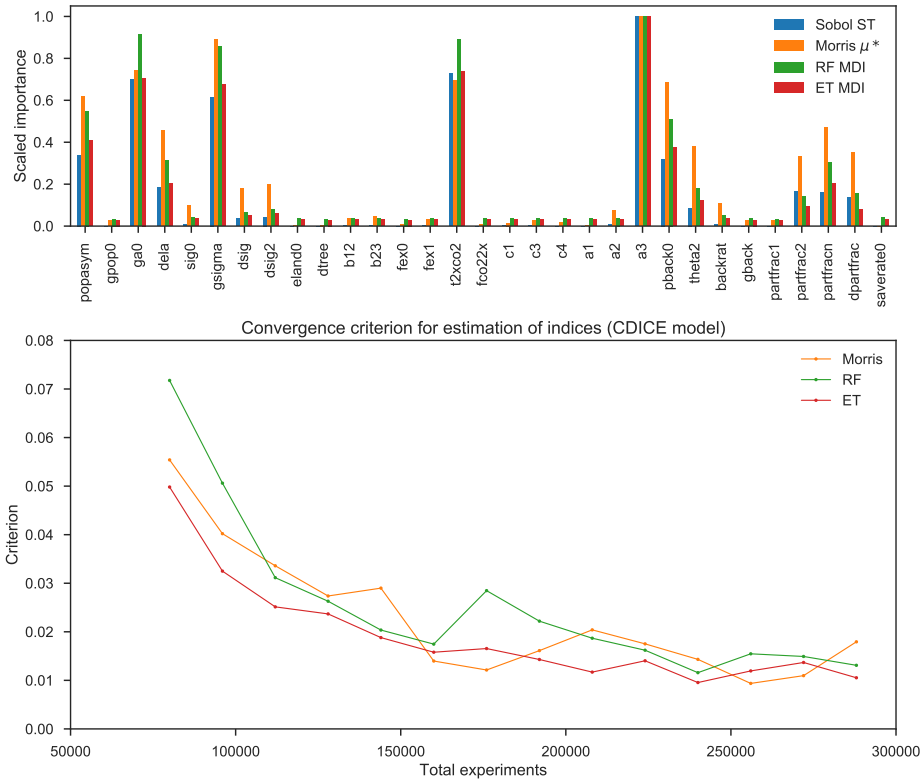
	H1N1 flu model		CDICE model	
	Model evaluation (s)	Analysis (s)	Model evaluation (s)	Analysis (s)
Sobol (S1, S2, ST)	8778 (N=8e5)	105	4661 (N=9.22e6)	735
Morris ($\mu*, \sigma$)	2131 (N=2e5)	5.8	126 (N=2.88e5)	7.8
ET (MDI importances)	1614 (N=1.5e5)	6.4	128 (N=2.88e5)	9.6
ET (pairwise MDA)	1059 (N=5e4)	16.4	42.4 (N=1e5)	51.8

Table B.3: Representative runtimes for the evaluation of the H1N1 and CDICE test cases, and for the computation of sensitivity indices. The sample size used in each model/analysis combination is indicated in parentheses.



	Sobol ST (N=8e5)	Morris $\mu*$ (N=2e5)	RF MDI (N=1.5e5)	ET MDI (N=1.5e5)
RMSE	-	0.081	0.084	0.056
MBE	-	0.017	-0.056	-0.041

Figure A.18: Scaled variable importances and error measures for H1N1 flu model.



	Sobol ST (N=9.22e6)	Morris $\mu*$ (N=2.88e5)	RF MDI (N=2.88e5)	ET MDI (N=2.88e5)
RMSE	-	0.149	0.094	0.035
MBE	-	-0.099	-0.064	-0.023

Figure A.19: Scaled variable importances and error measures for CDICE model (NPV of abatement costs).