Evaluation of Some Validation Measures for Gaussian Process Emulation: a Case Study with an Agent-Based Model

Wim De Mulder
and Geert Molenberghs
and Geert Verbeke

Leuven Biostatistics and
Statistical Bioinformatics Centre
KU Leuven and Hasselt University, Belgium
Email: wim.demulder@cs.kuleuven.be,
geert.molenberghs@uhasselt.be,
geert.verbeke@med.kuleuven.be

Bernhard Rengs
and Thomas Fent

Wittgenstein Centre (IIASA, VID/ÖAW, WU)
VID/ÖAW
Vienna, Austria
Email: bernhard.rengs@oeaw.ac.at,
thomas.fent@oeaw.ac.at

***Abstract*—A common way to evaluate surrogate models is by using validation measures. This amounts to applying a chosen validation measure to a test data set that was not used to train the surrogate model. The selection of a validation measure is typically motivated by diverse guidelines, such as simplicity of the measure, ease of implementation, popularity of the measure, etc., which are often not related to characteristics of the measure itself. However, it should be recognized that the validity of a model is not only dependent on the model, as desired, but also on the behavior of the chosen validation measure. Some, although very limited, research has been devoted to the evaluation of validation measures, by applying them to a given model that is trained on a data set with some known properties, and then evaluating whether the considered measures validate the model in an expected way. In this paper, we perform an evaluation of some statistical and non statistical validation measures from another point of view. We consider a test data set generated by an agent-based model and we successively remove those elements from it for which our previously developed Gaussian process emulator, a surrogate model, produces the worst approximation to the true output value, according to a selected validation measure. All considered validation measures are then applied to the sequence of increasingly smaller test data sets. It is desired that a validation measure shows improvement of a model when test data points on which the model poorly performs are removed, irrespective of the validation measure that is used to detect such data points. Our experiments show that only the considered statistical validation measures have this desired behavior.**

Keywords–Gaussian process emulation; Agent-based models; Validation.

I. Introduction and outline of the paper

In previous work we applied Gaussian process emulation, a surrogate model, to a training data set generated by an agent-based model that we had developed before [1]. Several alternative implementations of the Gaussian process emulation technique were considered and each of these was evaluated according to two different validation measures. Evaluation of the emulators was performed with respect to a test data set of size 500.

In this paper, we consider a research question that is not given proper attention in the literature, namely the evaluation of validation measures themselves. Although some researchers have examined certain characteristics of validation measures, their research is typically limited to the application of several selected validation measures to a given model that is trained on a data set with some known properties, and then evaluating whether these measures are able to validate these properties, see, e.g., [2], [3], [4]. Although such research is, of course, useful, we take here another perspective on the evaluation of validation measures. We consider the influence on validation measures when elements from the test data set are removed in the order proposed by a fixed validation measure. That is, we select a validation measure and we use that measure to find the element in the test data set for which a given surrogate model produces the worst approximation. We will simply refer to the element of a given test data set in which a given surrogate model produces the worst approximation according to a given validation measure as the worst test data point, and we will use the more vague term bad test data point to denote a test data point in which the surrogate model produces a bad approximation according to the given validation measure. It is then clear that the selected validation measure will show improvement when applied with respect to the reduced test data set, i.e., the elements of the test data set that remain after removing the worst test data point. However, an interesting and important research question is how the *other* validation measures will perform on the same reduced test data set. Will they also consider the selected test data point as the most problematic and thus have improved values when they evaluate the surrogate model on the reduced test data set? Or will they have another view on the test data point that is to be considered as the one where the surrogate model performs worst and, therefore, maybe even show *deterioration* of the surrogate model on the reduced data set?

The operation of removing the worst test data point is then repeatedly performed on the remaining test data set such that a graph of the considered validation measures results. This graph shows the evolution of the validation measures on increasingly smaller test data sets, where each test data set in this sequence does not contain the worst test data point of its predecessor. The whole procedure is then repeated by

choosing another validation measure to detect bad test data points and to remove them accordingly. Consequently, another graph of all considered validation measures is produced. These graphs are then analyzed to supply an answer to the following questions. Which validation measures show steady improvement by removing test data points that are designated as bad according to *both* selected validation measures? For which validation measures does the improvement depend on the choice of fixed validation measure that is used to detect bad test data points? Our previously developed Gaussian process emulator that emulates an agent-based model will be used as case study to answer these questions.

The significance of the above research questions is that it is desired to use validation measures whose evaluation of a given model in terms of its performance on a test data set is consistent with respect to other validation measures. That is, if one researcher employs validation measure A and detects a region in input space where the model has low performance, then it is desired that another researcher using validation measure B should see improvement of the model after additional training on points in that region, even though he is using another validation measure. Otherwise, there would be inconsistency between both measures and this would make it impossible to state any justified claim related to the performance of the model. The above described method to evaluate validation measures then simulates the often applied practice of additional training in regions where the given surrogate model performs bad, since such additional training results in improvement in that region. This implies that previously bad test data points will not have that statute anymore and this can be simply simulated by removing them from the test data set.

The outline of the paper is as follows. In Section II, we review Gaussian process emulation and agent-based models to ensure that the paper is self-contained. For the same reason we review our previous work, which is done in Section III. As described above, several validation measures will be considered. Some of them have been developed by statisticians to validate statistical models, such as Gaussian process emulation, while we also consider some validation measures that are popular outside statistical domains and apply to deterministic models. These validation measures are reviewed in Section IV. An in-depth description of and motivation for our experiments is provided in Section V. Results are presented and analyzed in Section VI. Section VII contains a discussion of the experiments, evaluating the implications and meaning of the experimental results.

II. Related work

A short overview of the aspects of our previous work that are relevant for this paper is provided in Section III. In this section, we briefly review Gaussian process emulation and agent-based models.

A. Gaussian process emulation

Gaussian process (GP) emulation provides an approximation to a mapping $\nu : \mathbb{R}^n \to \mathbb{R}$. The approximation to ν, i.e., the emulator, is determined as follows. In the first step, it is assumed that nothing is known about ν. The value $\nu(\mathbf{x})$ for any $\mathbf{x}$ is then modeled as a Gaussian distribution with mean $m(\mathbf{x}) = \sum_{i=1}^{q} \beta_i\, h_i(\mathbf{x})$, where β_i are unknown coefficients and where h_i represent linear regression functions. The covariance between $\nu(\mathbf{x})$ and $\nu(\mathbf{x}')$, with $\mathbf{x}$ and $\mathbf{x}'$ arbitrary input vectors in $\mathbb{R}^n$, is modeled as

$$\mathrm{Cov}\Big(\nu(\mathbf{x}), \nu(\mathbf{x}') \,|\, \sigma^2\Big) \quad = \quad \sigma^2\, c(\mathbf{x}, \mathbf{x}') \tag{1}$$

where σ^2 denotes a constant variance parameter and where $c(\mathbf{x}, \mathbf{x}')$ denotes a function that models the correlation between $\nu(\mathbf{x})$ and $\nu(\mathbf{x}')$. In our previous work, we have used the most common choice for c:

$$c(\mathbf{x}, \mathbf{x}') \quad = \quad \exp\Big[-\sum_i \Big((x_i - x_i')/\delta_i\Big)^2\Big] \tag{2}$$

with x_i and x_i' the ith component of $\mathbf{x}$ and $\mathbf{x}'$ resp., and where the δ_i represent the so-called correlation lengths. In the second step, training data $(\mathbf{x}_1, \nu(\mathbf{x}_1)), \ldots, (\mathbf{x}_n, \nu(\mathbf{x}_n))$ are used to update the Gaussian distributions to Student's t-distributions via a Bayesian analysis. The mean of the Student's t-distribution in $\mathbf{x}$ is then considered the best approximation to $\nu(\mathbf{x})$. Therefore, we refer to this mean as $\hat{\nu}(\mathbf{x})$. It is given by

$$\hat{\nu}(\mathbf{x}) = m(\mathbf{x}) + U^T(\mathbf{x}) A^{-1} ([\nu(\mathbf{x}_1), \ldots, \nu(\mathbf{x}_n)]^T - H\boldsymbol{\beta}) \tag{3}$$

with

$$\boldsymbol{\beta} \quad = \quad (\beta_1, \ldots, \beta_q)^T \tag{4}$$

(5)

$$H \quad = \quad \begin{bmatrix} h_1(\mathbf{x}_1) & \ldots & h_q(\mathbf{x}_1) \\ \ldots & & \\ h_1(\mathbf{x}_n) & \ldots & h_q(\mathbf{x}_n) \end{bmatrix} \tag{6}$$

and where $U(\mathbf{x})$ contains the correlations, as given by (2), between $\mathbf{x}$ and each of the training data points $\mathbf{x}_i$, and where A is the correlation matrix, containing the correlations between $\mathbf{x}_i$ and $\mathbf{x}_j$ for $i, j = 1, \ldots, n$. The expression (3) shows that the Bayesian analysis adds a correction term to the prior mean $m(\mathbf{x})$ by taking into account the information encapsulated in the training data set. The parameters δ_i can be optimized in terms of maximum likelihood [5], while optimal values for the β_i and for σ^2 can be determined by optimization principles in Hilbert space. For a more detailed account on GP emulation we refer to [6] and [7].

In practical applications, the Student's t-distributions are approximated by Gaussian distributions that are then used for all further operations. The variance of the Gaussian distribution in $\mathbf{x}$, denoted as $v(\mathbf{x})$, gives a measure of the uncertainty in approximating $\nu(\mathbf{x})$ by $\hat{\nu}(\mathbf{x})$. That is, the larger $v(\mathbf{x})$ the more tricky it is to approximate $\nu(\mathbf{x})$ as $\hat{\nu}(\mathbf{x})$. A 95% confidence interval for the true output $\nu(\mathbf{x})$ is given by $[\hat{\nu}(\mathbf{x}) - 2\sqrt{v(\mathbf{x})}, \hat{\nu}(\mathbf{x}) + 2\sqrt{v(\mathbf{x})}]$. An analytical formula for $v(\mathbf{x})$ is given in [7].

The main use of an emulator lies in the critical property that its execution is typically much faster than running the full model ν [8].

An example application of GP emulation is provided in Fig. 1. The model to be approximated is the function $f(x) = x \sin(x)$. The training data points (referred to as observations in the figure) are shown as red dots, while the approximation (called prediction in the figure), given by (3), is denoted by a blue line. A 95% confidence interval can be constructed as outlined above and this is also shown in the figure. It is seen that an emulator is an interpolator, i.e., the approximation is exact in the training data points and the confidence interval

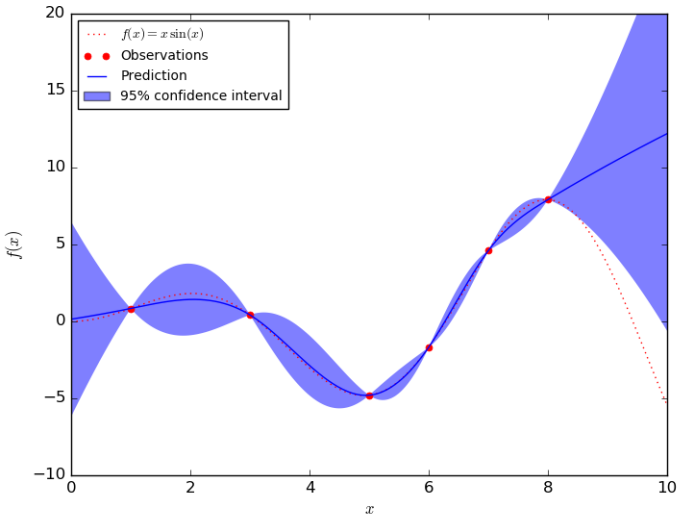

Figure 1. Example application of GP emulation
(*From http://scikit-learn.org/stable/modules/gaussian_process.html*)

in these points have length zero. Another typical property of a GP emulator is clearly noticed from the figure: the length of the confidence intervals increases with increasing distance to the nearest training data point. This property is intuitively clear, since moving away from a training data point means moving away from a point where there is precise information about an output value of the function to be approximated. One final observation is the large discrepancy between $f(x)$ and the emulator over the interval $(0.8, 1]$, which does not contain any training data point. Such a behavior is often observed for approximation techniques and shows that extrapolation should be avoided if possible [9].

B. Agent-based models

An agent-based model (ABM) is a computational model that simulates the behavior of and interactions between autonomous agents. A key feature is that population level phenomena are studied by explicitly modeling the interactions of the individuals in these populations [10], [11]. The systems that emerge from such interactions are often complex and might show regularities that were not expected by researchers in the field who solely relied on their background knowledge about the characteristics of the lower-level entities to make predictions about the higher-level phenomena. In [12], the authors describe situations for which agent-based modeling can offer distinct advantages to conventional simulation approaches. Some include:

- There is a natural representation as agents.
- It is important that agents learn and engage in dynamic strategic behaviors.
- The past is no predictor of the future.
- It is important that agents have a dynamic relationship with other agents, and agent relationships form and dissolve.

Examples of situations where ABMs have been successfully applied are infectious disease transmission [13], the development of risk behaviors during adolescence [14], the simultaneous study of the epidemiological and evolutionary dynamics of Influenza viruses [15], the sector structure of complex financial systems [16] and pedestrian movement [17]. ABMs are especially popular among sociologists who model social life as interactions among adaptive agents who influence one another in response to the influence they receive [18], [19], [20], [21].

Since nonlinear interactions and successive simulation steps are key ingredients of an agent-based model, such models are often computationally expensive. Consequently, if the model has to be executed on a large set of given input points, e.g., to determine parameter values that minimize an error criterion between model output and observed data, this task can often only be accomplished within a reasonable time by relying on emulation. Surprisingly, it is only recently that one has started to realize the use of Gaussian process emulation in analyses with agent-based models [22], [23], [24], [25], [26], [27], [28].

III. Previous work

A. Our agent-based model

In previous work, we developed an ABM to analyze the effectiveness of family policies under different assumptions regarding the social structure of a society [29]. In our model the agents represent the female partner in a household and are heterogeneous with respect to age, household budget, parity, and intended fertility. A network of mutual links connects the agents to a small subset of the population to exchange fertility preferences. The agents are endowed with a certain budget of time and money which they allocate to satisfy their own and their children's needs. We assume that the agent's and their children's consumption levels depend on the household budget but increase less than linearly with household budget. This implies that wealthier households have a higher savings rate. If the household's intended fertility exceeds the actual parity and the disposable budget suffices to cover the consumption needs of another child, the household is subject to the corresponding age-specific fertility. If an additional child is born, other agents may update their intended fertility.

We considered two components of family policies: 1. the policy maker provides a fixed amount of money or monetary equivalent per child to each household and 2. a monetary or nonmonetary benefit proportional to the household income is received by the household. The output on the aggregate level that is simulated by the ABM consists of the cohort fertility, the intended fertility and the fertility gap. Here, as in previous work, we restrict attention to the output component cohort fertility. The input variables include the level of fixed and income dependent family allowances, denoted by b^f and b^v, and parameters that determine the social structure of a society, such as a measure for the agents' level of homophily α, and the strength of positive and negative social influence, denoted by pr_3 and pr_4 resp.

Our simulations revealed a positive impact of both fixed and income dependent family allowances on completed cohort fertility and on intended fertility, and a negative impact of fixed and income dependent child supports on the fertility gap. However, several network and social influence parameters are such that they do not only influence fertility itself but also the effectiveness of family policies, often in a detrimental

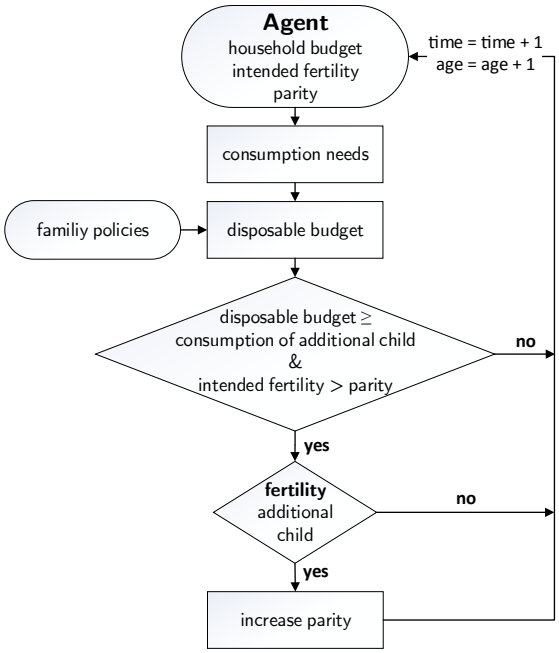

Figure 2. The decision making process in a household

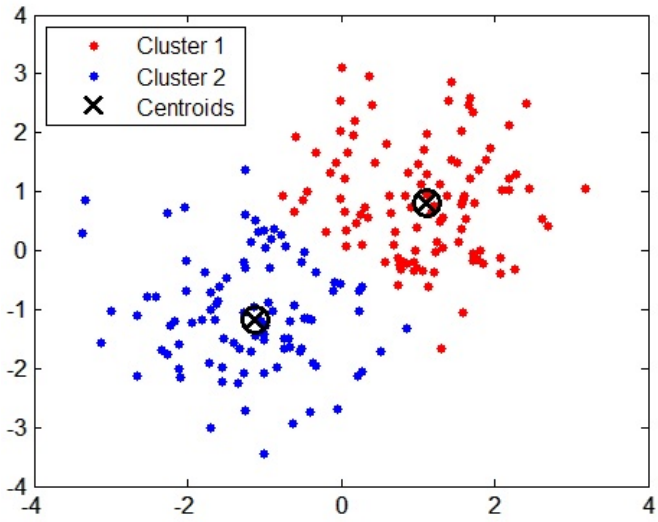

Figure 3. Illustration of k-means for two-dimensional data set with $k = 2$ (*From http://mines.humanoriented.com/classes/2010/fall/csci568/portfolio_exports/mvoget/cluster/cluster.html*)

way. For instance, while a higher degree of homophily among the network partners has a positive effect on fertility, family policies may be less effective in such a society. Therefore, policymakers aiming to transfer a certain policy mix that has proved successful from one country to another one ignoring differences in the social structure may fail. Family policies can only be successful if they explicitly take into account the characteristics of the society they are assigned for.

A flow-chart of the simulations performed by the ABM is provided in Fig. 2. Our model and the sociological hypotheses derived from application of it are extensively described in our previous work [29].

B. Data set generated by agent-based model

The input variables of our ABM are given equidistant values from the input domain and the ABM is applied to generate the corresponding outputs. As input domain we considered the variables b^f, b^v, α, pr_3 and p_4, a selection of the larger amount of variables that were used in the ABM. These five variables were found to have the largest influence on the outcomes. On the output side we restrict attention to one variable, namely cohort fertility. The ABM was applied to 10,732 vectors in the input domain, resulting in a large training data set. A test data set containing 500 input-output pairs was generated, the use of which will be described below.

C. Gaussian process emulation applied to our agent-based model

We applied GP emulation to our ABM. However, the large training data set necessitated us to adapt the originally developed GP emulation technique described in Section II-A. The reason is that the inverse of the correlation matrix is needed in the analytical formulation of the emulator. As this matrix is of quadratic order in the training data set size, it is obvious that the inverse operation cannot be performed (at least not in a numerically stable way). Therefore, we proceeded as follows.

First, we applied k-means [30], a popular cluster analysis algorithm, to subdivide the very large training data set into clusters. Cluster analysis is the unsupervised partitioning of a data set into groups, also called clusters, such that data elements that are member of the same group have a higher similarity than data elements that are member of different groups. Similarity is expressed in terms of a user-defined distance measure, such as the commonly used Euclidean distance which we employed. An illustration of the k-means principle is provided by Fig. 3. The application of k-means to our training data set resulted in 34 clusters with sizes ranging from 15 to 500. Implementation details are described in our previous work [1]. An emulator was then constructed for each of the resulting clusters.

Secondly, values of the parameters of each of the emulators were determined. Determination of the parameters β_i and σ^2 is simple, as analytical expressions exist for their optimal values (see, e.g., [31]). However, such expressions do not exist for the δ_i. These are typically obtained by applying the maximum likelihood principe, as described in [5]. This amounts to optimizing their joint density function which is a nontrivial task here as this function is a $\mathbb{R}^5 \to \mathbb{R}$ mapping (there are five correlation lengths, one for each of the input variables b^f, b^v, α, pr_3 and pr_4), potentially having many local optima. We used genetic algorithms [32] to perform this optimization task. Genetic algorithms are a type of heuristic optimization method that mimics some aspects of the process of natural selection, in that a population of candidate solutions to an optimization problem is evolved toward better solutions. This is done by applying certain operators, called mutation, crossover and reproduction, to the set of candidate solutions. These operators have been inspired by the principles of their biological counterparts and ensure that the population as a whole becomes fitter, i.e., the set of candidate solutions improves gradually according to a chosen error criterion. Fig. 4 illustrates the basic idea of genetic algorithms. Key advantages of genetic algorithms

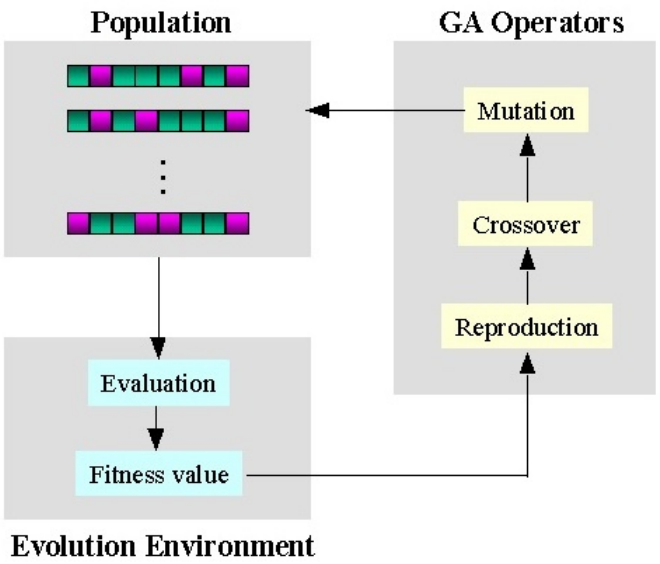

Figure 4. Illustration of genetic algorithms
(*From http://www.ewh.ieee.org/soc/es/May2001/14/Begin.htm*)

are that they only employ function evaluations (and thus not, e.g., information about the derivative, as is required by many other optimization methods, such as, for example, gradient descent) and that they are well suited to avoid getting stuck in local optima [33], [34], [35]. Both characteristics make them particularly useful to optimize the density function of the correlation lengths. For implementation details we again refer to our previous work.

Finally, given an input point $\mathbf{x}$ we determine an approximation to the output of the ABM in $\mathbf{x}$ as the output generated by the emulator that corresponds to the cluster closest to $\mathbf{x}$. We define the distance from a point to a cluster as the minimum of all distances from that point to any training data point that is member of the considered cluster. Obviously, there are other ways to combine the 34 emulators into one approximator. However, experimental results in our previous work demonstrated that the described approximator performs better than some alternative methods to combine the emulators.

In summary, when we speak of the output of the emulator in $\mathbf{x}$ we refer to the output of the emulator that was trained with the part of the full training data set that constitutes the cluster to which $\mathbf{x}$ is closest in terms of the described minimum distance. The notation $\nu(\mathbf{x})$ is used to denote the output of the ABM in $\mathbf{x}$, while $\hat{\nu}(\mathbf{x})$ refers to the output of the emulator in that input point.

IV. Validation measures

We consider several validation measures that can evaluate the performance of a given emulator. Two of them are related to popular measures in statistics, namely the average interval score and the average absolute individual standardized error. They take the uncertainty in the approximation generated by the emulator into account. Five other measures (Nash-Sutcliffe efficiency, coefficient of determination, index of agreement, relative Nash-Sutcliffe efficiency and relative index of agreement) are non statistical measures and have been used in a variety of fields. A final, extremely simple measure, is just the average of the absolute relative differences between approximations and true outputs. The values of the measures are determined with respect to a given test data set T.

A. Average interval score

The quality of a confidence interval $[l(\mathbf{x}), u(\mathbf{x})]$ around $\hat{\nu}(\mathbf{x})$ can be evaluated using the interval score described in [36]. Given an $(1-\alpha)\%$ confidence interval $[l(\mathbf{x}), u(\mathbf{x})]$, with $\alpha = 0.05$ chosen in this paper, the interval score is defined as

$$\begin{aligned} IS(\mathbf{x}) &= \Big(u(\mathbf{x}) - l(\mathbf{x})\Big) + \frac{2}{\alpha}\Big(l(\mathbf{x}) - \nu(\mathbf{x})\Big)\mathbf{1}_{\{\nu(\mathbf{x})<l(\mathbf{x})\}} \\ &\quad + \frac{2}{\alpha}\Big(\nu(\mathbf{x}) - u(\mathbf{x})\Big)\mathbf{1}_{\{\nu(\mathbf{x})>u(\mathbf{x})\}} \end{aligned} \tag{7}$$

where $\mathbf{1}_{\{expr\}}$ refers to the indicator function, being 1 if expression $expr$ holds and 0 otherwise. This scoring rule rewards narrow intervals, while penalizing lack of coverage. The lower its value, the higher the quality of the confidence interval. In terms of the average interval score, the given emulator is perfect when the value of the average interval score equals zero. This can only happen when $l(\mathbf{x}) = u(\mathbf{x}) = \nu(\mathbf{x})$. The first equality implies that the confidence interval is reduced to a single point, and if this is combined with the other equality we find that the value of this single point equals the value of the emulator. Thus, the perfect case occurs when the estimate equals the true value and when, at the same time, there is no uncertainty about how well the predicted value approximates the true one. Or in other words: the estimate equals the true value and we *know* that this is the case. The average interval score is simply the average of $IS(\mathbf{x})$ over all considered test points $\mathbf{x}$. An important advantage of the average interval score is that, unlike many other validation measures, this measure simultaneously evaluates the uncertainty in the approximation as given by the confidence interval, and the quality of the approximation. The first term in (7) evaluates the amount of uncertainty in the approximation: the larger the uncertainty related to the approximation, the larger the first term. The second and third term evaluate the quality of the approximation. If the true value is outside the confidence interval, and thus far from the approximation in a certain sense, one of both terms will be large. For some other work where this measure is used, we refer to [37] and [38].

B. Average absolute individual standardized error

Given $\mathbf{x}$, the corresponding individual standardized error [39] is given by

$$SE(\mathbf{x}) = \frac{\nu(\mathbf{x}) - \hat{\nu}(\mathbf{x})}{\sqrt{v(\mathbf{x})}} \tag{8}$$

This measure takes both the approximation and the constructed confidence interval into account, just as the average interval score discussed in Section IV-A. The measure SE, given by equation (8), is very useful since it allows to evaluate the magnitude of SE in a rather straightforward way. As outlined in Section II-A, the distributions of the approximations are approximately Gaussian. This implies that if the emulator properly represents ν, the distribution of SE is approximately standard normal. Thus, we expect that about 95% of SE values are smaller than 2 in absolute value. That is, if there are a considerable number of test points $\mathbf{x}$ for which the absolute

value of $SE(\mathbf{x})$ is larger than 2, then this is a clear warning that the emulator might not perform well. This convenient evaluation of a given emulator is an important advantage over the average interval score, where we do not have such reference values. The average interval score is only useful when at least two different emulators are to be compared to each other, while SE can be used to evaluate a single emulator. On the other hand, the average interval score has the benefit of not making any assumption about the distribution of the approximations. Taking absolute values and averaging over all considered test data points, we obtain our average absolute individual standardized error.

C. Nash-Sutcliffe efficiency

The Nash-Sutcliffe efficiency (NSE), proposed in [40], is determined as

$$NSE = 1 - \frac{\sum_{\mathbf{x}\in T}\Big(\nu(\mathbf{x}) - \hat{\nu}(\mathbf{x})\Big)^2}{\sum_{\mathbf{x}\in T}\Big(\nu(\mathbf{x}) - \overline{\nu}\Big)^2} \quad (9)$$

with $\overline{\nu}$ the average of $\nu(\mathbf{x})$ over all elements of T. The range of NSE lies between 1.0 (perfect fit) and -∞. An NSE of lower than zero indicates that $\overline{\nu}$ would have been a better predictor than the calculated approximations $\hat{\nu}(\mathbf{x})$. The fact that the Nash-Sutcliffe efficiency squares differences between true and estimated values implies that large values have large influence while small values are almost neglected, which might or might not be desired for the application at hand [4]. Furthermore, while the NSE is a convenient and normalized measure of model performance, it does not provide a reliable basis for comparing the results of different case studies [41]. Nevertheless, NSE is a popular measure for the evaluation of models, especially of hydrological models [42].

D. Coefficient of determination

The coefficient of determination r^2 is the square of the Pearson correlation coefficient:

$$r^2 = \left(\frac{\sum_{\mathbf{x}\in T}(\nu(\mathbf{x}) - \overline{\nu})(\hat{\nu}(\mathbf{x}) - \overline{\hat{\nu}})}{\sqrt{\sum_{\mathbf{x}\in T}(\nu(\mathbf{x}) - \overline{\nu})^2}\sqrt{\sum_{\mathbf{x}\in T}(\hat{\nu}(\mathbf{x}) - \overline{\hat{\nu}})^2}} \right)^2 \quad (10)$$

where $\overline{\hat{\nu}}$ refers to the averages of $\hat{\nu}(\mathbf{x})$ over the test data points. The measure is widely applied by statisticians [43].

The values of r^2 are between 0 and 1. The measure describes how much of the observed dispersion is explained by the estimation. A value of zero means no correlation at all, whereas a value of 1 means that the dispersion of the estimations is equal to that of the true values. Although many authors consider the coefficient of determination a useful measure of success of predicting the dependent variable from the independent variables [44], the fact that only the dispersion is quantified is a major drawback of r^2. A surrogate model that systematically over- or underestimates all the time can still result in good r^2 values close to 1.0 even if all estimations are critically wrong [45], [46].

E. Index of agreement

The index of agreement d was proposed in [47] to overcome the insensitivity of NSE and r^2 to differences in the true and estimated means and variances. It is defined as:

$$d = 1 - \frac{\sum_{\mathbf{x}\in T}\Big(\nu(\mathbf{x}) - \hat{\nu}(\mathbf{x})\Big)^2}{\sum_{\mathbf{x}\in T}\Big(|\hat{\nu}(\mathbf{x}) - \overline{\nu}| + |\nu(\mathbf{x}) - \overline{\nu}|\Big)^2} \quad (11)$$

Due to the mean square error in the numerator, d is also very sensitive to large values and rather insensitive to small values, as is the case for NSE. The range of d is $[0, 1]$ with 1 denoting perfect fit.

Practical applications of d show that it has some disadvantages [45]. First, relatively high values, say more than 0.65, may be obtained even for poor surrogate model fits. Secondly, systematic over- or underestimation can, as with the coefficient of determination, be masked by high values of d. There exist several variations on the above definition of the index of agreement, for example, by considering absolute differences instead of squared differences [48] or by removing the approximations $\hat{\nu}(\mathbf{x})$ from the denominator [49].

F. Relative Nash-Sutcliffe efficiency

The NSE described above quantifies the difference between the original model and the surrogate model in terms of absolute values. As a result, an over- or underestimation of higher values has, in general, a greater influence than those of lower values. Therefore, one has introduced the following *relative* NSE [45]:

$$NSE_{rel} = 1 - \frac{\sum_{\mathbf{x}\in T}\left(\frac{\nu(\mathbf{x}) - \hat{\nu}(\mathbf{x})}{\nu(\mathbf{x})}\right)^2}{\sum_{\mathbf{x}\in T}\left(\frac{\nu(\mathbf{x}) - \overline{\nu}}{\overline{\nu}}\right)^2} \quad (12)$$

Some recent research where the relative NSE is used include [50] and [51].

G. Relative index of agreement

The same idea can be applied to the index of agreement, resulting in the *relative* index of agreement [45]:

$$d_{rel} = 1 - \frac{\sum_{\mathbf{x}\in T}\left(\frac{\nu(\mathbf{x}) - \hat{\nu}(\mathbf{x})}{\nu(\mathbf{x})}\right)^2}{\sum_{\mathbf{x}\in T}\left(\frac{|\hat{\nu}(\mathbf{x}) - \overline{\nu}| + |\nu(\mathbf{x}) - \overline{\nu}|}{\overline{\nu}}\right)^2} \quad (13)$$

H. Average absolute relative difference

Given a test data point $\mathbf{x}$, we can evaluate the quality of the approximation as the absolute relative difference between $\nu(\mathbf{x})$ and $\hat{\nu}(\mathbf{x})$ as follows:

$$RD(\mathbf{x}) = \left|\frac{\hat{\nu}(\mathbf{x}) - \nu(\mathbf{x})}{1/2(\hat{\nu}(\mathbf{x}) + \nu(\mathbf{x}))}\right| \quad (14)$$

The average absolute relative difference, denoted ARD, is then the average of $RD(\mathbf{x})$ over all considered test data points.

This measure has the disadvantage of being unbounded, which makes it difficult to evaluate whether the obtained value is, e.g., large or very large. However, the fact that this measure is very simple makes it easy to interpret.

V. Description of the experiments

We experimentally evaluate how the described validation measures evolve when we successively remove elements from the test data set. Three methods are considered to remove elements. First, removal in terms of the absolute individual standardized error. That is, we calculate all validation measures for the full test data set T consisting of 500 test points. Then we remove the element with the largest absolute individual standardized error and calculate the validation measures again with respect to this reduced test data set. This procedure is repeated until only two elements remain (we do not calculate the measures for a test data set consisting of one element since this makes some measures, such as r^2, undefined due to division by zero). Secondly, removal in terms of the absolute relative difference, where the element with the largest absolute relative difference is removed first, then the element with the second largest absolute relative difference, etc. The third removal method discards elements in a purely random way.

Our experiments are related to the well established practice of evaluating a model with respect to some test data set and enlarging the training data set if the evaluation indicates poor performance. Preferably, the training data set is extended with bad points, i.e., points for which a chosen validation measure indicates large discrepancy between the true output value and the generated approximation, since it is intuitive to consider such points as lying in regions of input space where training was not performed properly. The points with which the training data set is extended should then be removed from the test data set. However, our purpose here is not to consider the influence of the extension of the training data set on the performance of the model, since it is clear that overall performance will, in general, be improved by extending learning to regions that were not given proper attention in a previous learning step. Rather, our goal is to assess the influence of removing bad data points from the test data set on our validation measures. Of course, it is obvious that removing the element with the largest absolute individual standardized error will result in an improvement of the average absolute individual standardized error. What is less obvious, however, is how this will affect the other validation measures. Thus, a first research question is to what extent the values of the described validation measures are sensitive to the choice of criterion that is used to describe a test data point as bad. From another perspective, this research question asks if the validation measures are compatible. That is, if a test point is regarded as bad by a certain measure, do all the other measures agree with this, in the sense that removing such an element improves their value? This research question is of the utmost importance, as it is desired that our evaluation of the goodness-of-fit of a model is only, or at least mainly, dependent on the model and not on the choice of validation measure. Furthermore, even when it would hold that all validation measures improve by removing test points that are bad according to a certain measure, they might not improve to the same extent. Some measures might improve very significantly when one bad point is removed, while other measures might encounter only a marginal benefit. It is also important to detect such differences, if they exist, between validation measures, since an overly optimism in the improvement of a model after having extended training might not be justified if the improvement according to other measures would only show incremental improvement. Indeed, such a case would point to an artifact of the chosen validation measure rather than to inherent characteristics of the improved model.

The random removal of elements serves as a benchmark case: validation measures should improve much more in response to the removal of bad points according to a well chosen validation measure than according to a removal that is completely random.

VI. Results

The results are shown in Figs. 5-12. Each figure displays the evolution of one of the eight considered validation measures, described in Section IV, as elements are progressively discarded from the test data set, and this for each of the three removal methods (i.e., according to the absolute individual standardized error, according to the absolute relative difference and via random removal).

It is seen that, at first sight, the average interval score, the average absolute individual standardized error, the Nash-Sutcliffe efficiency, the coefficient of determination, and the index of agreement behave as desired: they all gradually improve as the worst element of the current test data set is removed. However, a closer look at Figs. 7-9 reveals that the Nash-Sutcliffe efficiency, the coefficient of determination and the index of agreement evaluate the emulator as becoming worse for the removal of, approximately, the first 20 elements when removal is done according to the absolute relative difference.

Although the relative versions of the Nash-Sutcliffe efficiency and of the index of agreement have been developed to compensate certain deficiencies of these measures, we observe that these extensions do not result in unequivocally better behavior in our experiments. Their behavior with respect to the removal of elements according to the absolute individual standardized error is quite erratic, almost indiscernible from their behavior when removal is random. On the other hand, these relative measures show more consistent behavior in terms of removal according to the absolute relative difference. Whereas the non relative Nash-Sutcliffe efficiency and the non relative index of agreement become worse by removing the approx. first 20 elements and only steadily increase after having reduced the test data set by these 20 elements, the relative counterparts increase steadily from the removal of the first element on.

Our simplest validation measure, the average absolute relative difference, decreases steadily if removal is with respect to the absolute relative difference. But this is of course a trivial observation, as it is obvious that a measure improves if elements are discarded that are bad according to that same measure. Much more relevant is that the average absolute relative difference shows undesired behavior when elements are removed according to their absolute individual standardized error. Although its global trend is decreasing until about 350 elements are deleted, it suddenly starts to increase after that turning point.

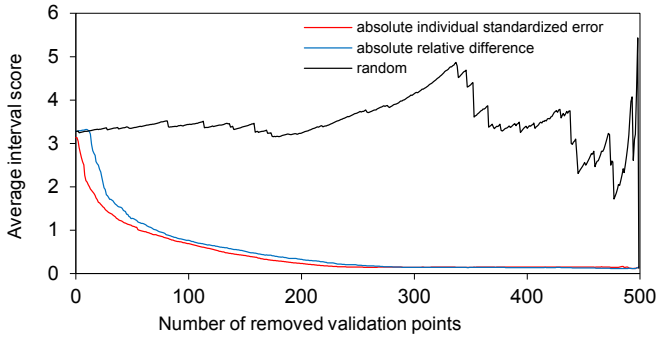

Figure 5. Average interval score

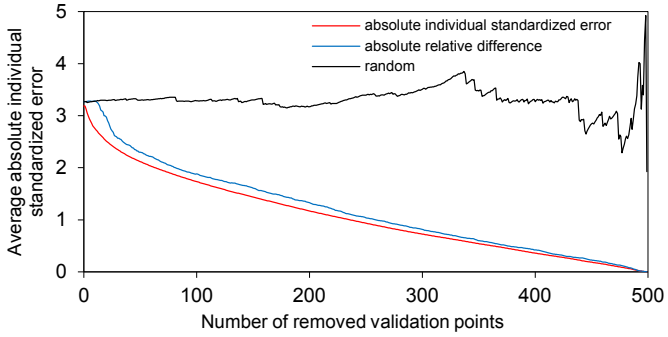

Figure 6. Average absolute individual standardized error

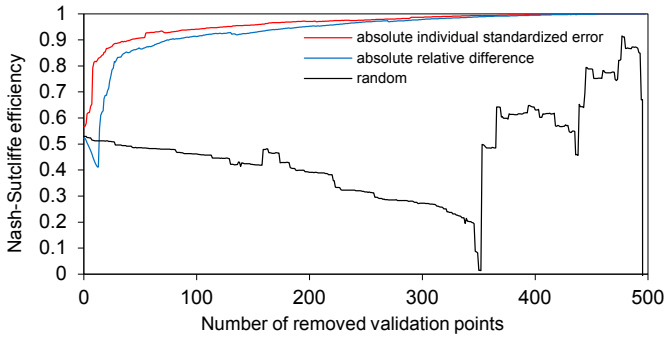

Figure 7. Nash-Sutcliffe efficiency

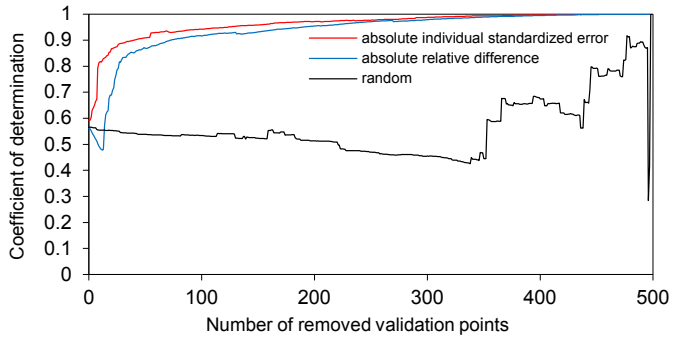

Figure 8. Coefficient of determination

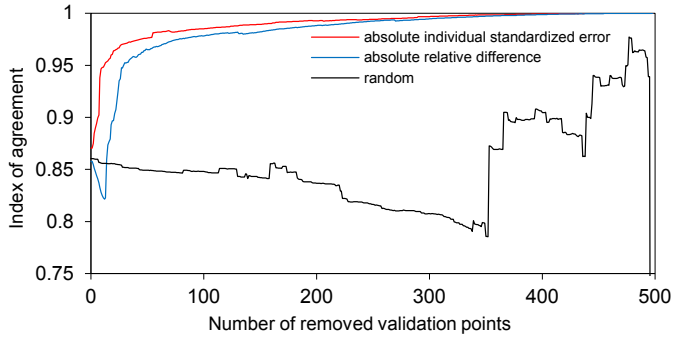

Figure 9. Index of agreement

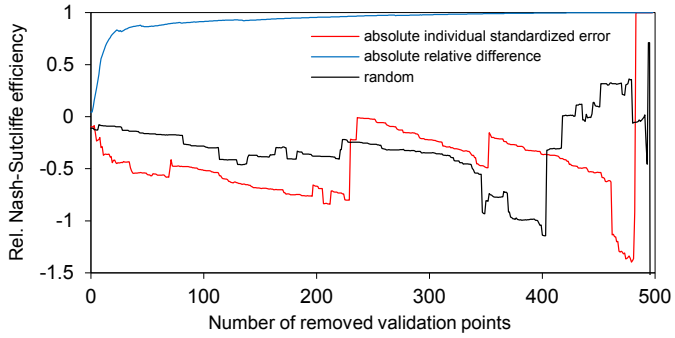

Figure 10. Relative Nash-Sutcliffe efficiency

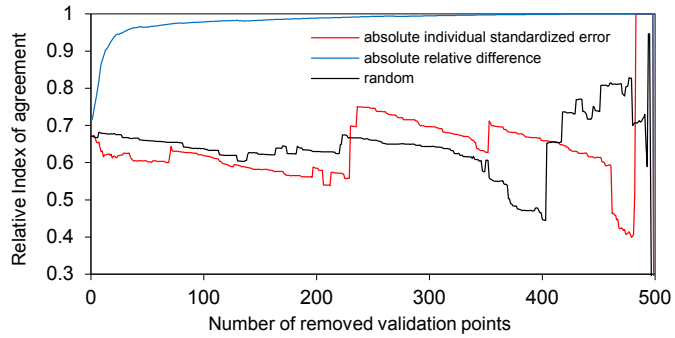

Figure 11. Relative index of agreement

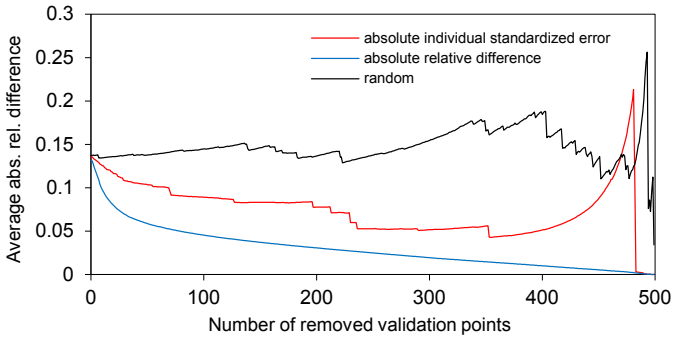

Figure 12. Average absolute relative difference

VII. Discussion

The experiments indicate that the average interval score and the average absolute individual standardized error have the most desired behavior. Whether a point is labeled 'bad' according to its absolute individual standardized error or according to its absolute relative difference, removing the worst element from the test data set results in better values of both measures. We remind that only these two measures take the uncertainty in the approximation into account (see Section IV). Thus, our experiments suggest that statistical surrogate models, such as Gaussian process emulation, have certain benefits over deterministic surrogate models, such as polynomial approximation, in particular that the uncertainty in the approximation is also modeled. This uncertainty measure should then be taken into account in validating the model.

Comparing Fig. 5 and Fig. 6, the main difference between the average interval score and the average absolute individual standardized error is that the first one reacts much more pronounced to the removal of elements, at least concerning the removal of about the first half of all elements. The decrease of the average interval score appears to be of exponential order, while the average absolute individual standardized error seems to improve only linearly except for the first dozen or so elements. This indicates that one should be careful to report an improvement in a model as very significant when the average interval score is used as validation measure, since part of the improvement might be solely due to characteristics inherent in that validation measure. It is advised to validate the model in terms of both the average interval score and the average absolute individual standardized error.

The other measures do not show steady improvement with respect to either the average interval score or the average absolute individual standardized error. Remarkably, each of these other measures *do* improve steadily in terms of *one* of these measures. The Nash-Sutcliffe efficiency, the coefficient of determination and the index of agreement improve consistently when removal of elements is performed according to the absolute individual standardized error, as is seen from Figs. 7, 8, and 9. On the other hand, the relative Nash-Sutcliffe efficiency, the relative index of agreement and, of course, the average absolute relative difference show steady improvement in terms of the absolute relative difference. This implies that these measures are sensitive to the criterion that is used to measure the quality of the approximation in a certain point. A point that is designated as bad, i.e., low quality of approximation in that point, according to the absolute individual standardized error might not be recognized as such by the aforementioned six non statistical validation measures. The same applies to measuring the quality of approximation in a point by the absolute relative difference.

VIII. Conclusion and future work

In this paper, we have evaluated eight validation measures for surrogate models: the average interval score, the average absolute individual standardized error, the Nash-Sutcliffe efficiency, the coefficient of determination, the index of agreement, the relative Nash-Sutcliffe efficiency, the relative index of agreement and the average absolute relative difference. The first two measures are statistical in nature, taking into account the uncertainty of the approximation generated by the surrogate model. The other measures are solely based on the generated approximation values. The evaluation was performed using a Gaussian process emulator that was applied to an agent-based model. We developed both Gaussian process emulator and agent-based model in previous work.

Our method of evaluating validation measures has, as far as we are aware of, not been applied yet. We consider a test data set and successively remove those elements from it for which our emulator produces the worst approximation to the true output value, in terms of the absolute individual standardized error. The considered validation measures are then applied to the sequence of increasingly smaller test data sets. The same procedure is applied with removal of test data points in terms of the absolute relative difference. It is desired that a validation measure shows improvement of a model when test data points on which the model poorly performs are removed, irrespective of the measure that is used to detect such data points. Our experiments indicate that only the average interval score and the average absolute individual standardized error have this desired behavior.

Our work has some practical implications:

- Statistical surrogate models, which not only produce an approximation to or estimation of the output in a given input point but also a measure for the uncertainty in the approximation, are preferred over deterministic models. Evaluation of such a model should then be done by a statistical validation measure that takes this uncertainty measure into account, such as the average interval score and the average absolute individual standardized error.
- It is bad practice to evaluate a given model in terms of a single validation measure, as the value of this measure might not only reflect the performance of the model but also certain inherent artifacts of the measure itself. Evaluating a model using several measures ensures different perspectives on the performance of the model, and thus avoids an overly optimistic or pessimistic view on its performance that might not be justified.

As future research, it would be interesting to evaluate other validation measures according to our evaluation procedure. Especially recently developed validation measures that are meant to extend or improve previously developed measures should be evaluated. Examples include:

- A relatively recent alternative to the index of agreement that is dimensionless, bounded by -1.0 and 1.0 and for which the authors claim that it is more rationally related to model accuracy than are other existing indices [52].
- Another alternative to the index of agreement that is also dimensionless and bounded [53]. The authors demonstrate the use and value of their index on synthetic and real data sets, but an evaluation in line with our procedure would increase justification of their claims.
- A bounded version of the Nash-Sutcliffe efficiency [54].

Our experiments show that such an additional evaluation is not superfluous, as modifications to existing measures that in terms of analytical formulation seemingly compensate some

clear drawbacks of the existing measure might not show as consistent behavior in practice as one is inclined to anticipate.

ACKNOWLEDGMENT

The authors acknowledge funding from the KU Leuven funded Geconcerteerde Onderzoeksacties (GOA) project New approaches to the social dynamics of long-term fertility change [grant 20142018;GOA/14/001].

REFERENCES

[1] W. De Mulder, B. Rengs, G. Molenberghs, T. Fent, and G. Verbeke, "Statistical emulation applied to a very large data set generated by an agent-based model," in Proceedings of the 7th International Conference on Advances in System Simulation, SIMUL, 2015, pp. 43–48.

[2] Y. Liu, Z. Li, H. Xiong, X. Gao, and J. Wu, "Understanding of internal clustering validation measures," in Proceedings of the IEEE International Conference on Data Mining. IEEE, 2010.

[3] M. Brun, C. Sima, J. Hua, J. Lowey, B. Carroll, E. Suh, and E. Dougherty, "Model-based evaluation of clustering validation measures," Pattern Recognition, vol. 40, 2007, pp. 807–824.

[4] D. R. Legates and G. J. McCabe Jr., "Evaluating the use of goodness-of-fit measures in hydrologic and hydroclimatic model validation," Water Resources Research, vol. 35, 1999, pp. 233–241.

[5] I. Andrianakis and P. G. Challenor, "The effect of the nugget on Gaussian process emulators of computer models," Computational Statistics and Data Analysis, vol. 56, 2012, pp. 4215–4228.

[6] A. O'Hagan, "Bayesian analysis of computer code outputs: A tutorial," Reliability Engineering & System Safety, vol. 91, 2006, pp. 1290–1300.

[7] J. Oakley and A. O'Hagan, "Bayesian inference for the uncertainty distribution of computer model outputs," Biometrika, vol. 89, 2002, pp. 769–784.

[8] J. Gómez-Dans, P. Lewis, and M. Disney, "Efficient emulation of radiative transfer codes using Gaussian processes and application to land surface parameter inferences," Remote Sensing, vol. 8, 2016, doi:10.3390/rs8020119.

[9] D. Larose and C. Larose, Eds., Data mining and predictive analytics. Wiley, 2015.

[10] N. Gilbert, Ed., Agent-based models: quantitative applications in the social sciences. SAGE Publications, Inc, 2007.

[11] F. C. Billari, T. Fent, A. Prskawetz, and J. Scheffran, Eds., Agent–Based Compuational Modelling: Applications in Demography, Social, Economic, and Environmental Sciences, ser. Contributions to Economics. Springer, 2006.

[12] C. Macal and M. North, "Agent-based modeling and simulation: Abms examples," in Proceedings of the 2008 Winter Simulation Conference, 2008, pp. 101–112.

[13] L. Willem, Agent-based models for infectious disease transmission: exploration, estimation & computational efficiency. PhD thesis, 2015.

[14] N. Schuhmacher, L. Ballato, and P. van Geert, "Using an agent-based model to simulate the development of risk behaviors during adolescence," Journal of Artificial Societies and Social Simulation, vol. 17, no. 3, 2014.

[15] B. Roche, J. M. Drake, and P. Rohani, "An agent-based model to study the epidemiological and evolutionary dynamics of influenza viruses," BMC Bioinformatics, vol. 12, no. 87, 2011.

[16] J.-J. Chen, L. Tan, and B. Zheng, "Agent-based model with multi-level herding for complex financial systems," Scientific Reports, vol. 5, 2015.

[17] A. Crooks, A. C. X. Lu, S. Wise, J. Irvine, and A. Stefanidis, "Walk this way: improving pedestrian agent-based models through scene activity analysis," ISPRS International Journal of Geo-Information, vol. 4, no. 3, 2015, pp. 1627–1656.

[18] M. Macy and R. Willer, "From factors to factors: computational sociology and agent-based modeling," Annual Review of Sociology, vol. 28, 2002, pp. 143–166.

[19] F. Bianchi and F. Squazzoni, "Agent-based models in sociology," Wiley Interdisciplinary Reviews: Computational Statistics, vol. 7, 2015, pp. 284–306.

[20] F. Squazzoni, Ed., Agent-based computational sociology. Wiley, 2012.

[21] A. El-Sayed, P. Scarborough, L. Seemann, and S. Galea, "Social network analysis and agent-based modeling in social epidemiology," Epidemiologic Perspectives & Innovations, vol. 9, 2012, doi: 10.1186/1742-5573-9-1.

[22] G. Dancik, D. Jones, and K. Dorman, "Parameter estimation and sensitivity analysis in an agent-based model of Leishmania major infection," Journal of Theoretical Biology, vol. 262, 2010, pp. 398–412.

[23] J.-S. Lee, T. Filatova, A. Ligmann-Zielinska, B. Hassani-Mahmooei, F. Stonedahl, and I. e. a. Lorscheid, "The complexities of agent-based modeling output analysis," Journal of Artificial Societies and Social Simulation, vol. 18, 2015, doi: 10.18564/jasss.2897.

[24] J. Sexton and Y. Everingham, "Global sensitivity analysis of key parameters in a process-based sugarcane growth model: a Bayesian approach ," in Proceedings of the 7th International Congress on Environmental Modelling and Software, 2014.

[25] A. Heppenstall, A. Crooks, L. See, and M. Batty, Eds., Agent-based models of geographical systems. Springer, 2011.

[26] D. Heard, G. Dent, T. Schifeling, and D. Banks, "Agent-Based models and microsimulation," Annual Review of Statistics and Its Application, vol. 2, 2015, pp. 259–272.

[27] J. Bijak, J. Hilton, and E. Silverman, "From agent-based models to statistical emulators," in Joint Eurostat/UNECE Work Session on Demographic Projections, 2013.

[28] J. Castilla-Rho, G. Mariethoz, M. Rojas, R. Andersen, and B. Kelly, "An agent-based platform for simulating complex human-aquifer interactions in managed groundwater systems," Environmental Modelling & Software, vol. 73, 2015, pp. 305–323.

[29] T. Fent, B. Aparicio Diaz, and A. Prskawetz, "Family policies in the context of low fertility and social structure," Demographic Research, vol. 29, 2013, pp. 963–998.

[30] A. Jain and R. Dubes, Eds., Algorithms for clustering data. Prentice Hall College Div, 1988.

[31] J. Oakley and A. O'Hagan, "Bayesian inference for the uncertainty distribution of computer model outputs," Biometrika, vol. 89, 2002, pp. 769–784.

[32] M. Mitchell, Ed., An introduction to genetic algorithms. MIT Press, 1998.

[33] P. Fleming and A. Zalzala, Eds., Genetic algorithms in engineering systems. The Institution of Engineering and Technology, 1997.

[34] E. Sanchez, Ed., Genetic algorithms and fuzzy logic systems: soft computing perspectives. Wspc, 1997.

[35] L. Chambers, Ed., The practical handbook of genetic algorithms: new frontiers. CRC Press, 1995.

[36] T. Gneiting and A. Raftery, "Strictly proper scoring rules, prediction, and estimation," Journal of the American Statistical Association, vol. 102, 2007, pp. 359–378.

[37] N. Cahill, A. Kemp, B. Horton, and A. Parnell, "Modeling sea-level change using errors-in-variables integrated Gaussian processes," The Annals of Applied Statistics, vol. 9, 2015, pp. 547–571.

[38] C. Lian, C. Chen, Z. Zeng, W. Yao, and H. Tang, "Prediction intervals for landslide displacement based on switched neural networks," IEEE Transactions on Reliability, vol. 65, 2016, pp. 1483–1495.

[39] L. Bastos and A. O'Hagan, "Diagnostics for Gaussian process emulators," Technometrics, vol. 51, 2009, pp. 425–438.

[40] J. Nash and J. Sutcliffe, "River flow forecasting through conceptual models, Part I - A discussion of principles," Journal of Hydrology, vol. 10, 1970, pp. 282–290.

[41] B. Schaefli and H. Gupta, "Do Nash values have value?" Hydrological Processes, vol. 21, 2007, pp. 2075–2080.

[42] H. Gupta, H. Kling, K. Yilmaz, and G. Martinez, "Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling," Journal of Hydrology, vol. 377, 2009, pp. 80–91.

[43] N. Sören Blomquist, "A note on the use of the coefficient of determination," The Scandinavian Journal of Economics, vol. 82, 980, pp. 409–412.

[44] N. Nagelkerke, "A note on a general definition of the coefficient of determination," Biometrika, vol. 78, 1991, pp. 691–692.

[45] P. Krause, D. Boyle, and F. Bäse, "Comparison of different efficiency criteria for hydrological model assessment," Advances in Geosciences, vol. 5, 2005, pp. 89–97.

[46] J. Liao and D. McGee, "Adjusted coefficients of determination for logistic regression," The American Statistician, vol. 57, 2003, pp. 161–165.

[47] C. Willmot, "On the validation of models," Physical Geography, vol. 2, 1981, pp. 184–194.

[48] C. Willmott, S. Ackleson, R. Davis, Feddema, K. Klink, D. Legates, J. O'Donnell, and C. Rowe, "Statistics for the evaluation and comparison of models," Journal of Geophysical Research, vol. 90, 1985, pp. 8995–9005.

[49] C. Willmott, S. Robeson, and K. Matsuura, "A refined index of model performance," International Journal of Climatology, vol. 32, 2012, pp. 2088–2094.

[50] M. Barbouchi, R. Abdelfattah, K. Chokmani, N. B. Aissa, R. Lhissou, and A. E. Harti, "Soil salinity characterization using polarimetric InSAR coherence: Case studies in Tunisia and Morocco," IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, vol. 8, 2015, pp. 3823–3832.

[51] D. M. M. Ahouansou, S. K. Agodzo, S. Kralisch, L. O. Sintondji, and C. Fürst, "Analysis of the hydrological budget using the J2000 model in the Pendjari River Basin, West Africa," Journal of Environment and Earth Science, vol. 5, 2015, pp. 24–37.

[52] C. Willmott, S. Robeson, and K. Matsuura, "A refined index of model performance," International Journal of Climatology, vol. 32, 2011, pp. 2088–2094.

[53] G. Duveiller, D. Fasbender, and M. Meroni, "Revisiting the concept of a symmetric index of agreement for continuous datasets," Scientific Reports, vol. 6, 2016, doi:10.1038/srep19401.

[54] T. Mathevet, C. Michel, V. Andréassian, and C. Perrin, "A bounded version of the Nash-Sutcliffe criterion for better model assessment on large sets of basins," in IAHS Red Books Series no. 307. IAHS, 2006.