Proceedings of the 2017 Winter Simulation Conference
W. K. V. Chan, A. D'Ambrogio, G. Zacharewicz, N. Mustafee, G. Wainer, and E. Page, eds.

HYPOTHESIS-DRIVEN EXPERIMENT DESIGN IN COMPUTER SIMULATION STUDIES

Fabian Lorig
Daniel S. Lebherz
Jan Ole Berndt
Ingo J. Timm

Center for Informatics Research and Technology (CIRT)
Trier University
Behringstrasse 21, 54296 Trier, GERMANY

ABSTRACT

In simulation studies, the goal specifies the objective or purpose of the study and thus drives the entire experimentation process. Relevant experiments and respective experiment hypotheses are derived from the study's goal and the model's observed behavior provides evidence whether these hypotheses hold. Current assistance systems do not integrate research hypotheses. Thus, the researcher has to make important design decisions which limits both replicability and reproducibility of the results. In this paper, the process of simulation studies is systematized based on a formally specified hypothesis. By this means, the research hypothesis becomes the key element of the study and guides the entire process. Hypothesis-driven simulation studies allow for the automated design, execution, and evaluation of experiments based on specific research questions. This facilitates documentation and execution as well as replication of simulation studies.

1 INTRODUCTION

Simulation studies aim at answering model-specific research questions by means of experiments (Maria 1997). By executing a model with different parametrizations, the corresponding outputs can be observed, a better understanding of the model's behavior is gained, and respective questions can be answered (Zeigler et al. 2000). Thus, simulation has become a major source of knowledge. It is applied in many scientific disciplines and economic sectors for advancing knowledge and for supporting decisions as it provides efficient, economic, and convenient approaches for analyzing complex systems (Tolk et al. 2013).

Conducting sound simulation studies and identifying relevant experiments is not trivial. The process is extensive and both model- as well as goal-specific adaptations need to be made (Montgomery 2012). The goal of a simulation study specifies the objective or the purpose of the study. Based on this goal, testable hypotheses are specified and simulation experiments are designed to test these hypotheses (Yilmaz et al. 2016). According to Mill (1868), a hypothesis can be scientifically defined as "any supposition which we make [...] in order to endeavour to deduce from it conclusions in accordance with facts which are known to be real". It can be expressed as a logical argument where one or many premises claim to prove a conclusion. In case empirical observations contradict the hypothesis, it is refuted and must be rejected (Whewell 1847). To assist the confirmation of hypotheses in simulation, various procedure models exist that advise the researcher on essential steps for conducting sound and successful simulation studies (Timm and Lorig 2015). Additionally, management frameworks exist that assist reproducible design, execution, and analysis of experiments (Teran-Somohano et al. 2015). Both procedure models and experiment management systems provide valuable and comprehensive support for simulation studies and experiments.

However, using the aforementioned systems and models is no silver bullet. Important design decisions that drive the entire study and that are essential for sound hypothesis testing are still made by the researcher,

e.g., the selection of factors to study. Thus, experimenter bias might occur when these decisions are insufficiently documented or methodologically ungrounded (Uhrmacher et al. 2016). This results in limited replicability and reliability of the results. To overcome this, we believe that the design of simulation experiments for testing hypotheses should not (only) be the researcher's responsibility. Instead, the process of the study must be systematically aligned with the research hypothesis. This includes providing assistance that guides hypothesis testing in simulation studies by extending existing approaches accordingly.

In this paper, we aim at systematizing the process of simulation studies based on a formally specified research hypothesis. We propose a process for *hypothesis-driven simulation studies* which covers two tasks. Relevant experiments are systematically derived from the hypothesis and the simulation outputs are both aggregated and evaluated with respect to proving or disproving the hypothesis. To assist the process, the required formal specification of hypotheses is provided by the *FITS* language which has been proposed by the authors (Lorig et al. 2017). In FITS, hypotheses are expressed as a number of premises that are inferentially linked to a conclusion. Based on a formally specified hypothesis, important factors are identified, relevant experiments as well as the corresponding designs are derived, and the execution of a sufficient number of simulation replications is carried out. By this means, the systematic generation and documentation of credible and reproducible simulation results is facilitated. This approach is complementary to existing ones where the planning of experiments is assisted but important design decisions are met by the experimenter. It is a first step towards facilitating and automating systematic hypothesis testing in simulation studies.

The paper is structured as follows. In Section 2, the background on assistance of computer simulation as well as specification of experiments is introduced. Subsequently, in Section 3, our approach for deriving relevant experiments from formally specified hypotheses is presented. For this purpose, techniques for specifying hypotheses, defining performance measures, identifying important factors, and deriving experimental designs are presented and combined. In Section 4, the approach is evaluated by means of a case study, while Section 5 provides both conclusions as well as an overview of future work.

2 BACKGROUND

The proposed process for assisting and automating the hypothesis-driven design, execution, and evaluation of simulation studies embraces the efforts of automating science. King et al. (2009) underline the importance of automating and recording experiments in sufficient detail to allow for reproducibility as a fundamental pillar of science. Waltz and Buchanan (2009) pick up on this and emphasize the importance of combining experimental design, data collection, and both formation and revision of hypotheses in one process. This demand and potential is also discussed and postulated by the discrete event simulation community (Yilmaz et al. 2014). Hence, the integrated assistance of the entire life-cycle of a simulation study is proposed (Teran-Somohano et al. 2015). Approaches for addressing this demand are twofold: Frameworks and other assistance systems are developed for guiding and facilitating efforts in simulation. Additionally, languages and other formal standards are designed to allow for the specification of experiments. Besides simplifying and standardizing the documentation of simulation, the latter approaches aim at supporting the automation of simulation. Experiment settings become machine-readable and can be processed by assistance systems.

Ören et al. (1984) described the potential for assisting and automating monotonous tasks in simulation, e.g., design, execution, and evaluation of experiments. During the design phase of simulation experiments, assistance can support the selection and preparation of input data and assess the significance that can be expected from a specific set of input data (Lattner et al. 2011). Additionally, the number of replications that are required for given inputs can be estimated (Hoad et al. 2010) and searching optimal values for the input parameters can be assisted (Better et al. 2007). For the execution of simulation experiments, scripts facilitate covering large parameter spaces. Different parametrizations are automatically generated by iterating the variables' values within a given range (Griffin et al. 2002). For evaluating the results of simulation runs, a number of assistance functionalities also exists. These include but are not limited to the automated analysis of simulation output data and the automated statistical evaluation of the results (Robinson 2005). Furthermore, there are integrated frameworks which combine multiple assistance functionalities for guiding

and assisting simulation experiments. This includes domain specific frameworks like SAFE (Perrone et al. 2012) but also cross-domain frameworks like JAMES II (Himmelspach and Uhrmacher 2007). Frameworks that consider hypotheses for addressing specific research questions are limited to natural language (Yilmaz et al. 2016).

For the specification of experiments, domain specific languages (DSL), ontologies, and other formalisms have been proposed. With respect to the reproduction of experiment, the environment or context of the experiment can be described as an *experimental frame* (Zeigler et al. 2000).For the description of the experiment itself, dedicated markup languages, e.g., SED-ML (Köhn and Le Novère 2008), and corresponding guidelines have been proposed (Waltemath et al. 2011). Such approaches are complemented by ontologies for describing scientific experiments (Soldatova and King 2006). Additionally, DSL whose expressive power focuses on individual domains have been proposed for describing experiments, e.g., SESSL (Ewald and Uhrmacher 2014) and the framework proposed by Teran-Somohano et al. (2015). This is of particular interest when models are changed or reused as associated and previously executed simulation experiments can be reused (Peng 2017).

To conclude, a wide range of valuable assistance functionalities, languages, and formalisms exists. Yet, the demand for assistance for the formation and revision of hypotheses in simulation studies is not considered. With regard to this paper's aim, a research gap can be identified as follows. Current assistance systems do not guide the simulation process based on formally specified hypotheses. Furthermore, existing formalisms do not allow for the systematic derivation of relevant experiments for verifying such hypotheses.

3 PROCESS OF HYPOTHESIS-DRIVEN SIMULATION STUDIES

In the process proposed here, existing approaches will be extended, adapted, and combined to systematically take research questions into account when planning and conducting a simulation study. By this means, the hypothesis becomes the key element of the study and guides the entire process. To allow for such *hypothesis-driven simulation studies*, three aspects must be considered: First, the components of simulation studies and their interconnections need to be hierarchically differentiated. Second, the process of simulation studies has to be methodologically aligned with the hypothesis. Finally, the responses received when executing the simulation must be aggregated to apply statistical hypothesis tests to answer the research question.

3.1 Components of Simulation Studies

Simulation studies are conducted to achieve a better understanding of how a system works by performing experiments with a model of the system. In the classical sense, simulation studies consist of two parts, the conception and implementation of the simulation model (*modeling*) and the design, execution as well as evaluation of simulation experiments (*experimentation*) (Law and Kelton 1991). The modeling part requires domain-specific expertise for the creation of a valid model which appropriately represents the real world system. As a result, models are highly individual, it is difficult to compare or assess a model's quality, and the assistance of the modeling part of a simulation study is challenging.

To successfully conduct simulation studies, the modeling part is not always of primary interest as a well-suited model of the studied system may already exist and can be reused. The credibility and replicability of a simulation study instead depends on the design, execution, and evaluation of the experiments. Thus, as a first step towards hypothesis-driven simulation studies, the focus lies on the well-documented, replicable, and assisted experimentation in simulation studies. Experimentation with the model investigates the system's behavior with the objective of providing evidence for or against the simulation study's leading question or assumption (the goal). The model's behavior is determined by its inner structure. However, often the structure is neither accessible nor essential for assessing the model's behavior. Thus, we pursue a black box approach where the experimentation only relies on the observable behavior of the model.

Based on an overall research question, the aim of the approach presented here is to plan and conduct a simulation study with respect to assisting hypothesis-driven experimentation. Simulation studies implement

a hierarchical structure where experiments need to be derived from the study's goal on the one side and consist of complex processes and respective sub-processes on the other side. Thus, a terminological distinction between the components that are relevant in studies and their interdependencies must be made to allow for the systematic assistance of simulation studies (see Fig. 1).

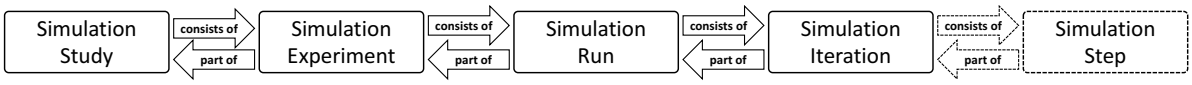

Figure 1: Components of a simulation study.

Winsberg (2010) defines *simulation studies* as inferential processes that investigates complex phenomena by means of computational techniques. To achieve the objective of a simulation study, i.e., address specific research questions, corresponding performance measures are specified and *simulation experiments* are conducted to analyze and compare the influence different parametrizations of the model have on the values of the performance measures. Each experiment describes a series of tests where changes are made to the inputs of the model to observe changes of the outputs (Montgomery 2012). A simulation study consists of at least one simulation experiment per question that has been asked regarding the system. For each simulation experiment one or many *simulation runs* are designed in which the system's response to a specific set of inputs is observed (Maria 1997). In contrast to deterministic simulation models, simulation runs that execute stochastic models with probabilistic inputs demand multiple *simulation iterations* to allow for a statistical assessment of the result's mean value depending on its standard deviation. Each iteration of a stochastic simulation run is initiated with the same parameters but a different stream of random numbers which results in different output values. Finally, depending on how progress is calculated in the simulation model, each simulation iteration may consist of multiple *simulation steps*. Each step (tick) represents progress of the simulation clock and the computation of a new state of the model (Zeigler et al. 2000). For the consideration of warm-up periods or for defining time-based termination of simulation iterations, the concept of simulation steps is reasonable.

3.2 Decomposition and Execution of Simulation Experiments

After terminologically discriminating between different hierarchical components of simulation studies (cf. Fig. 1), the systematic aggregation and disaggregation of the components must be enabled. To this end, the components must be aligned with the process of the study, linked according to existing procedure models, and the transitions between the components must be technically specified. By this means, we can systematically derive, adapt, and transfer the components with respect to the study's goal. The resulting integrated process (cf. Fig. 2) specifies the links between the components. It closes the gap between the structure of simulation studies and the methodological requirements for systematically answering research questions by means of simulation. To illustrate the extended process and the potential for assisting hypothesis-driven simulation studies, an example from industrial manufacturing process simulation is applied.

According to most simulation procedure models, the necessary first step of a simulation study is the proper definition of the goal (Law and Kelton 1991, Banks 1998). This is essential, as the study's goal specifies the objective as well as the purpose of the study. Thus, it drives the entire experimentation process as relevant experiments need to be derived from the goal in order to achieve it (Conway and McClain 2003). The boundaries for defining reasonable goals are determined by the examined scenario which is given by the model providing the study's context. Based on a manufacturing scenario, a variety of research questions can be thought of, e.g., whether the storage is sufficient in case the order changes (Lattner et al. 2011). A possible question that drives the simulation study could read as follows: *Will the manufacturing cycle efficiency (MCE) increase by more than 10% if the number of machines is increased from 15 to 17?*

For each question which is stated as goal of a simulation study, one or more testable hypotheses need to be constructed that can be verified by means of experiments. Yilmaz et al. (2016) distinguish between three types of experiment hypotheses in simulation-based knowledge generation. In this work, only

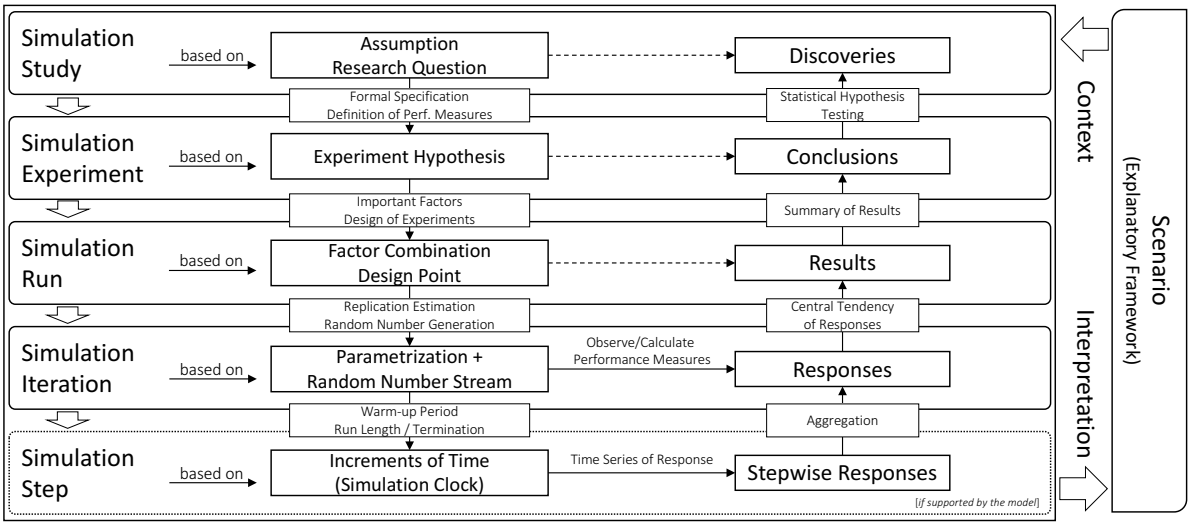

Figure 2: Process for conducting hypothesis-driven simulation studies.

phenomenological hypotheses are relevant as both *mechanistic* and *control hypotheses* make statements about mechanisms of the model which are unavailable in the black box approach pursued here. Phenomenological hypotheses make assertions about the input-output-relationship of a model. They consist of statements regarding the values of outputs in the case of specific input values.

To test phenomenological hypotheses, statistical hypothesis testing approaches can be used. They assess the probability of an observed output in relation to other possible outputs. To make preparations for the application of hypothesis tests, the following requirements must be met. A null hypothesis (H_0) as well as a corresponding alternative hypothesis (H_1) need to be formulated and a data sample needs to be generated. Both H_0 and H_1 need to be derived from the study's goal. Here, H_1 describes the assumption that the model's behavior will change under the defined conditions while H_0 assumes that any changes made to the parametrization of the model or the model itself will not have any effect on the model's response. To define a set of corresponding hypotheses, measures need to be defined first which provide information that can be used for assessing the response of the model.

For some research questions, the outputs of the model can be directly used for measuring the performance of the model but most commonly additional quantitative criteria (*performance measures*) are defined and used to compare the behavior of the model under different parametrizations. In the manufacturing context, key performance indicators (KPI) are suitable performance measures as they enable and facilitate the assessment of manufacturing processes. In the example above, an assumption is made regarding the MCE of a manufacturing process. MCE is an important indicator of process performance and is defined as the ratio between *value-added time* and *manufacturing cycle time* (*throughput time*).

Performance indicators are goal-specific and often not part of the outputs of a simulation model. Instead, output variables need to be mathematically combined to new (*target*) variables which can then be used to assess the model's performance. In the example used here, all output variables that represent non-value-added times, i.a., *wait time* and *queue time*, need to be summed up to a new variable *manufacturing cycle time*. Such (*intermediate*) variables are created artificially with respect to assessing the performance of the model but are not directly used as performance measures (Ören et al. 1984). As a next step, the output value *process time* can be divided by the intermediate variable *manufacturing cycle time* to receive the target variable *manufacturing cycle efficiency*.

In the example given here, the formulated question consists of a relative statement regarding the model's behavior; a 10% increase of the MCE when two machines are added. When applying statistical hypothesis tests, the definition of a test statistic that summarizes the dataset is required. As the study's goal consists

of the assessment of the MCE, the mean of the MCE's distribution is a suitable test statistic. Consequently, a possible pair of experiment hypotheses can read as follows.

H_0: If the no. of machines is increased from 15 to 17, the mean MCE will not increase by more than 10%.

H_1: If the no. of machines is increased from 15 to 17, the mean MCE will increase by more than 10%.

To provide systematic assistance to this process, the goal of the study, the respective experiment hypotheses, and the process for deriving the hypotheses from the study's goal need to be formally specified. For this purpose, the authors have proposed *FITS*, a language which enables the formal specification of statistical hypotheses in simulation (Lorig et al. 2017). This allows for both automated parametrization and evaluation of simulation studies. Necessary experiments are systematically derived from the hypothesis, resulting simulation runs are executed, and outputs are analyzed with a suitable hypothesis test.

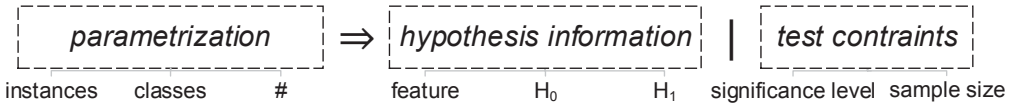

Figure 3: Structure of an experiment hypothesis in *FITS*.

In Fig. 3, the structure of an experiment hypothesis in *FITS* language is shown. A *FITS* expression consists of three parts: the parametrization of the model, information on the statistical hypothesis, and additional test constraints. In the parametrization part, specific values or ranges of values are assigned to the independent variables of the model. As the closed-world assumption applies in *FITS*, the #-operator (*ceteris paribus*) is used to assign standard values to all remaining input variables of the model which have not been explicitly declared. Based on the parametrization of the model, detailed information on both null (H_0) and alternative hypothesis (H_1) are provided in the *hypothesis information* part of the expression. First, a number of output or target variables are defined to serve as performance measures and respective statistical measures are required for determining the central tendency of these measures. Subsequently, both H_0 and H_1 are formulated based on values or value ranges of the performance measures. Finally, *test constraints* as significance level and sample size are stated. As the proper definition of test constraints is challenging, we aim to provide assistance for this step of the simulation study, too, which is presented later on. Excluding test constraints at first, the following expression illustrates how H_0, H_1, and the two competing parametrizations from the example above can be formally specified using *FITS*:

$$ParSet1(machines(17)) \wedge ParSet2(machines(15)) \wedge \#$$
$$\Rightarrow \mu_1(MCE) \wedge \mu_2(MCE) \wedge (H_0(\mu_1 - \mu_2 \leq 10\%) \vee H_1((\mu_1 - \mu_2 > 10\%))$$

As a next step, experiments are conducted to generate outputs and to test whether the experiment hypothesis holds. According to the definition of experiments (cf. Sec. 3.1), conducting experiments includes changing the model's inputs for observing outputs. The *Design of Experiments* (DOE) is challenging, as decisions have to be made regarding which input variables (factors) are altered and which values (levels) are relevant and feasible. This results in a trade-off between computational complexity and coverage of the parameter space. Furthermore, the selection of a suitable hypothesis test is necessary at this stage. The experimentation process needs to be aligned with the hypothesis test's requirements to be able to observe and record the values of the test statistic during experimentation. For all of the resulting relevant parametrizations, individual simulation runs are executed to obtain comprehensive and sound results.

In the presented example, the simulation study consists of a single experiment with a corresponding pair of hypotheses. Based on these hypotheses, a two-sample test is required to test whether or not the null hypothesis holds. In this case, a *two-sample t-test* for normal populations and independent observations is suitable for testing the hypotheses. Hence, individual samples have to be drawn for each parameter set (i.e., 15 and 17 machines) by means of simulation. To execute simulation runs, specific values need to be assigned to all input variables. This also includes those whose factor levels are not explicitly part of the hypothesis. In general, the model's responses for all possible factor level combinations are of interest

to fully assess the response surface of the performance measure. However, full factorial designs are not feasible for large models with many factors due to the combinatorial explosion of simulation runs.

In this case, the identification of a minor set of factors that have a major impact on the performance measure is advisable according to the *parsimony principle* (*Occam's razor*) (Kleijnen 2008). For this purpose, different *factor screening* approaches have been proposed like the Morris method (Morris 1991) and sequential bifurcation (SB) (Bettonvil and Kleijnen 1997), while SB requires a smaller number of simulation runs compared to the Morris method. Each approach requires an individual set of conditions to be met by the model, e.g., whether correlations between factors are allowed. Thus, and because of the required mathematical understanding, it is reasonable to assist the selection and application of factor screening in simulation studies. Technically, SB pursues a divide and conquer approach for identifying factor effects by systematically altering the input factor levels.

After identifying a set of factors that is important with respect to the study's performance measures, the systematic variation of the factor levels by means of individual simulation runs needs to be planned. To decrease the number of required simulation runs as well as the computational efforts, fractional factorial designs have been proposed in the DOE field. Examples are 2^k *factorial designs* or *latin hypercube designs* which define a subset of levels for each factor to be tested during the simulation (Sanchez 2005). After applying SB to identify important factors, the use of a 2^k *factorial design* seems suitable as both approaches use two levels (low and high) for each factor. Each possible combination of the identified factor levels results in a specific parametrization of the model and consequently defines an individual simulation run.

Finally, when executing the designed simulation runs, the impact stochastic inputs have on the variation of simulation outputs has to be considered. By replicating simulation runs, a larger sample is drawn so that statistical measures of dispersion can be applied, the variance can be quantified, and statements regarding the parent population can be made. To define the number of replications needed for a sufficient estimate of the performance measure's mean, different approaches exist: rule of thumb, graphical methods, or confidence intervals with specified precision (Hoad et al. 2010). For the assistance and automation of the simulation process, a confidence interval-based approach is suitable as it is the least dependent on the expertise of the analyst, makes use of statistical inference, and thus can be algorithmically described based on a given significance level (Lattner et al. 2011). Each simulation iteration that results from the same simulation run shares the same parametrization but the outputs differs due to the generated random numbers.

For some types of models, simulation iterations can be divided into individual simulation steps. An example is *discrete-event simulation* where the states of the model change at discrete points in time and a simulation clock keeps track of the model's current time. In contrast to real world time progress, simulation time skips periods where no events occur and instantly jumps to the point of time the next event takes place (Fujimoto 2015). Each simulated point in time is an individual step of the simulation. By this means, time constraints such as warm-up periods or temporal termination conditions can be easily implemented. Additionally, the simulation analyst can keep track of the response variables' progress over time.

3.3 Aggregation and Interpretation of Simulation Results

After conducting all simulation runs and all respective iterations, the response of each iteration must be aggregated to draw conclusions and to confirm or disconfirm the assumptions of the study. If the model implements a stepwise progress of time, the output most likely will have the shape of a time series and thus a number of decisions needs to be made. The response of the iteration can either be defined as the value of the model's outputs at a specific point in time, e.g., the last step of the simulation, or the central tendency of the outputs over a period of time. When averaging output values, both the consideration of the values of all steps or the resetting of statistical measures after a defined warm-up period are feasible. Alternatively, the outputs of each simulation step can be seen as an individual sample and analyzed accordingly.

To overcome the gap between the model's outputs (V_o) and the performance measures that are part of the study's goal, target variables (V_t) were specified based on the model's output variables in a previous step. First, each iteration's response (V_o^{It}) is extracted from the time series of output data in step-based

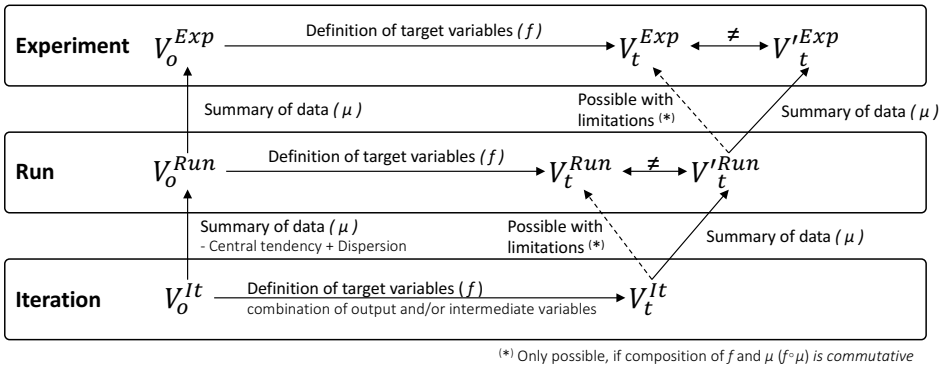

Figure 4: Aggregation of performance measures in simulation studies.

simulation models or directly by observing the output values in unspecified models. Then, to obtain each run's results, the values of the specified target variables (V_t^{Run}) need to be calculated based on the model's responses (f) and aggregated for each simulation run (μ). Here, the order in which the two steps are applied is of particular relevance as they are not interchangeable without the risk of producing biased or incorrect results. The functions f and μ do not per se commute with each other. Thus, $f \circ \mu \neq \mu \circ f$ must be assumed and a differentiation between V_t^{Run} and $V_t'^{Run}$ needs to be made (cf. Fig. 4). This is important to avoid misinterpretations of the results due to an (inadvertently) incorrect aggregation of the outputs.

V_t^{Run} expresses average values of output variables that are independently calculated over a number of iterations and related afterwards. $V_t'^{Run}$, in contrast, is applied when interdependencies exist between the outputs and a reliable estimation of the performance measure is only provided when calculated for each scenario. The same differentiation applies for the aggregation of results from different simulation runs for drawing conclusions from an experiment. To avoid this pitfall and to achieve sound results, thorough planning and intelligent assistance of the process of aggregating performance measures is necessary.

Finally, after aggregating the responses of both the executed iterations and runs, statistical hypothesis tests need to be applied to the data of each experiment for proving or disproving the previously defined pair of hypotheses. The selection of an appropriate test has taken place at an earlier stage of the process. This is to ensure all samples that are required to perform the hypothesis test are correctly drawn and to design the experiments accordingly. After rejecting the null or alternative hypothesis based on the results of the hypothesis test, the initial assumption or research question of the study can be assessed or answered and the simulation study is completed. However, the discoveries of the study may only be interpreted with respect to the study's scenario. It serves as an explanatory framework and is given by the model.

4 CASE STUDY: APPLICATION AND EVALUATION OF THE PROCESS

In Sec. 3, we presented an integrated process for the systematic conduction of hypothesis-driven simulation studies. For evaluation purposes, we apply this process to a simulation study. After introducing the *NetLogo* model we use here, the goal of the study is formulated, experiments with their respective hypotheses, corresponding hypothesis tests, and important factors are derived, simulation runs are designed, the number of iterations is estimated, and the simulation is executed. Subsequently, the mean values of the performance measures are calculated based on the iterations' outputs, the responses are statistically aggregated for each run, and hypothesis tests are performed to prove or disprove the experiment's hypotheses. By this means, we systematically provide an answer to the initial assumption stated as a goal of the simulation study.

To ensure the replicability of the results presented in this paper, both a model and a simulation framework which are publicly available are used. The supply chain model by Gil (2012) can be downloaded via the *Modeling Commons* repository and was developed in *NetLogo*. The model implements an artificial market with four different types of participants, i.e., customers, retailers, distributors, and factories. Following this order, each participant has an individual stock, a resulting demand for a product, and purchases the

product from the subsequent participant. The customer purchases from the retailer et cetera. Thus, the model implements a single product supply chain and can be used to analyze how stock levels and demand calculation change under different forecast and inventory management strategies.

The chosen model aims to illustrate the *bullwhip effect*, where minor variations of the customers' demand escalate and result in major variations in the factories production volume due to forecast uncertainties and safety stocks. Thus, we state the following assumption as a goal of the simulation study: "*If the average demand of the customers increases by 10 units, the retailers' average EOQ (economic order quantity) per customer will increase by more than 10 units.*" Applying the *FITS* approach (Lorig et al. 2017), the following combined experiment hypothesis can be derived from the aforementioned goal. As the increase of the customers average demand is not stated as an absolute value, a reference must be provided based on which the increase is quantified as an absolute value. Accordingly, *Welch's t-test* for two independent samples is well suited for verifying the resulting pair of hypotheses:

$$ParSet1(Demand\_W(20)) \ \wedge \ ParSet2(Demand\_W(10)) \ \wedge \ \#$$
$$\Rightarrow \mu_1(EOQ \div Clients\_N) \ \wedge \ \mu_2(EOQ \div Clients\_N) \ \wedge \ (H_0(\mu_1 - \mu_2 \leq 10) \ \vee \ H_1((\mu_1 - \mu_2 > 10))$$

In this example, the *EOQ* (a standard method for inventory management) is part of the model and provided as an output variable. Yet, the target variable ($EOQ \div Clients\_N$) which was chosen as a performance measure is not part of the model and needs to be calculated. Additionally, the experiment's stated hypothesis defines two parameter sets (*ParSet*) in each of which a specific value is assigned to the input variable *Demand_W*. No assignments are made for the remaining variables which indicated by the #-operator (*FITS* syntax for *ceteris paribus*). The model consists of 14 factors and each factor is defined by a range of admissible values. In this example, we limit the scenario to 1 factory, 3 distributors, and 7 retailers. Furthermore, we only consider the *(s,Q)* inventory policy and customers purchasing daily. Still, the amount of possible factor value combinations is too high for a full coverage of the parameter space. Hence, important factors are identified and experimental designs applied for reducing the computational efforts. When screening for important factors, correlations between the inputs can occur. For an unbiased estimation of the effect groups of factors have on a performance measure, the use of a screening technique which can handle two-factor interactions is advisable. We assume that a first-order polynomial with two-factor interactions is a suitable metamodel for approximating the inputs' effects on the performance measure. Therefore, we chose *sequential bifurcation (SB) with fold over design* as applied in (Kleijnen et al. 2003).

Compared to the estimated total effect ($\beta_{4-12} = 91.75$), the factors *Clients_N* ($\beta_4 = 21.77$), *Demand_W* ($\beta_8 = 15.89$), and *Product_cost* ($\beta_{12} = 32.50$) constitute more than 75% of the main effect. In contrast, the effect of the remaining factors is negligible. One of the important factors, the demand of the customers, is part of the experiment hypothesis and thus the factors levels of interest are given. Both the number of clients and the product costs are not part of the hypothesis. Yet, as they were identified as being important for the selected performance measure, they need to be altered during the simulation, too. The importance of these factors was identified using SB with high and low values for each factor. Consequently, experimental designs which discriminate between low and high values are preferable. Applying a 2^k *factorial design* results in four different parametrizations which need to be simulated for each of the two parameter sets defined in the hypothesis. Consequently, eight individual simulation runs emerge (cf. Table 1).

For each simulation run, the number of necessary replications for achieving satisfactory results has to be individually estimated. We applied the confidence interval method with specified precision as described by Hoad et al. (2010) with a 1% accepted deviation of the confidence interval about the mean. As the resulting values of mean and standard deviation originate from different sample sizes, *Welch's t-test* for two samples and both unequal variances and sample sizes is applied for a pairwise verification of the experiment hypothesis.

The results show that H_0, i.e., the average EOQ per customer will not increase by more than 10 units through an increasing demand, can only be rejected if at least one factor level is high. To interpret these results, the negative effect both factors have on the performance measure needs to be taken into account.

Table 1: Pairwise factor level combinations, simulation replications and results as well as statistical hypothesis tests and implications of all executed simulation runs.

	Factor levels			**Simulation results**			**Hypothesis test**	
	Demand	*Prod._Cost*	*Clients*	# replications	cumulative mean	standard deviation	Welch's t-test	rejection
Run 1	10	low	low	28	17.857	0.488	t = -14.281	t $\leq$ 1.678
Run 5	20	low	low	30	25.600	0.675	df = 47.142	$\Rightarrow$ **accept** H_0
Run 2	10	high	low	26	56.731	1.402	t = 29.815	t $>$ 1.679
Run 6	20	high	low	26	80.962	1.990	df = 44.911	$\Rightarrow$ **reject** H_0
Run 3	10	low	high	30	33.567	0.898	t = 16.042	t $>$ 1.674
Run 7	20	low	high	30	48.167	1.289	df = 51.777	$\Rightarrow$ **reject** H_0
Run 4	10	high	high	32	106.094	2.878	t = 40.428	t $>$ 1.674
Run 8	20	high	high	30	152.200	4.021	df = 52.266	$\Rightarrow$ **reject** H_0

In this case, SB demands the inversion of the factors such that switching the factor level from low to high has a positive overall effect. A low factor level represents a large number of customers with respectively high product costs and vice versa. Consequently, when increasing the customers' demand from 10 to 20, a significant increase of the study's performance measure of more than 10 units can be confirmed when the model is executed with small values for the number of clients, the product cost, or both factors but not for large values of both factors. Without further knowledge of the bullwhip effect it can be assumed that an increasing demand from customers results in a proportional increase of the retailers' order quantity. In contrast to this, the simulation shows that retailers will adjust their order quantities disproportionately. In terms of the study's scenario (the supply chain model) this discovery can be interpreted as indicating that retailers forecast less conservatively when ordering cheaper products. Furthermore, a smaller amount of customers increases the forecast uncertainty, too, resulting in higher order quantities.

5 CONCLUSIONS

In this paper, we have presented an integrated process for the systematic conduction of hypothesis-driven simulation studies. The aim of this process is to enable the assisted and automated design, execution, and evaluation of simulation experiments based on the study's goal. To achieve this, the proposed process assists the formal specification of one or several experiment hypotheses based on the goal of the study. By formally specifying hypotheses and corresponding tests, necessary experiments, resulting parametrizations, and the number of required iterations per simulation, runs can be systematically derived. This facilitates the conduction of simulation studys as the selection and application of suitable techniques is assisted. Furthermore, the replicability and reproducibility of the study's results is improved as the process allows for a detailed and automated documentation of the procedure as well as an effortless repetition of the entire process, i.e., when modifying the research question or the model.

This work is a first step towards the development of a research assistance system capable of assisting hypothesis-driven simulation studies by systematically deriving relevant parametrizations. The presented evaluation illustrates that the automation of this process is feasible. The selection criteria for choosing adequate and applicable methods and for connecting them in a suitable way do exist. Yet, algorithms for automating the selection process based on ontologies or decision trees need to be implemented. Furthermore, the proposed process is designed in a modular way so existing assistance services for individual steps of the process can easily be integrated and exchanged, e.g., approaches presented in Sec. 2.

Future work will focus on the integration as well as systematic and hypothesis-specific selection and application of further methods and techniques for each step of the process. This extension includes but is not limited to a more detailed differentiation of statistical hypothesis tests, factor screening approaches, experimental designs, and techniques for the estimation of replications. The approach's limitation to phenomenological aspects of experiments can be compensated by enriching the hypothesis specification.

REFERENCES

Banks, J. 1998. *Handbook of Simulation*. John Wiley & Sons.

Better, M., F. Glover, and M. Laguna. 2007. "Advances in Analytics: Integrating Dynamic Data Mining with Simulation Optimization". *IBM Journal of Research and Development* 51 (3.4): 477–487.

Bettonvil, B., and J. P. Kleijnen. 1997. "Searching for Important Factors in Simulation Models with many Factors: Sequential Bifurcation". *European Journal of Operational Research* 96 (1): 180–194.

Conway, R. W., and J. O. McClain. 2003. "The Conduct of an Effective Simulation Study". *INFORMS Transactions on Education* 3 (3): 13–22.

Ewald, R., and A. M. Uhrmacher. 2014. "SESSL: A Domain-Specific Language for Simulation Experiments". *ACM Transactions on Modeling and Computer Simulation (TOMACS)* 24 (2): 11.

Fujimoto, R. 2015. "Parallel and Distributed Simulation". In *Proceedings of the 2015 Winter Simulation Conference*, edited by C. M. Macal, M. D. Rossetti, L. Yilmaz, I.-C. Moon, W. K. Chan, and T. Roeder, 45–59. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Gil, A. 2012. *Artificial Supply Chain*. www.modelingcommons.org/browse/one_model/3378, accessed 03/17.

Griffin, T. G., S. Petrovic, A. Poplawski, and B. Premore. 2002. *SOS*. www.ssfnet.org/sos, accessed 03/17.

Himmelspach, J., and A. M. Uhrmacher. 2007. "Plug'n Simulate". In *ANSS*, 137–143. IEEE.

Hoad, K., S. Robinson, and R. Davies. 2010. "Automated Selection of the Number of Replications for a Discrete-Event Simulation". *Journal of the Operational Research Society* 61 (11): 1632–1644.

King, R. D., J. Rowland, S. G. Oliver, M. Young, W. Aubrey, E. Byrne, M. Liakata, M. Markham, P. Pir, L. N. Soldatova et al. 2009. "The Automation of Science". *Science* 324 (5923): 85–89.

Kleijnen, J. P. 2008. *Design and Analysis of Simulation Experiments*, Volume 20. Springer.

Kleijnen, J. P., B. Bettonvil, and F. Persson. 2003. "Finding the Important Factors in Large Discrete-Event Simulation: Sequential Bifurcation and its Applications". *CentER Discussion Paper; Vol.* 2003-104.

Köhn, D., and N. Le Novère. 2008. "SED-ML–an XML Format for the Implementation of the MIASE Guidelines". In *Computational Methods in Systems Biology*, 176–190. Springer.

Lattner, A. D., T. Bogon, and I. J. Timm. 2011. "An Approach to Significance Estimation for Simulation Studies". In *Agents and Artificial Intelligence*, 177–186. ICAART.

Lattner, A. D., H. Pitsch, I. J. Timm, S. Spieckermann, and S. Wenzel. 2011. "AssistSim-Towards Automation of Simulation Studies in Logistics". *Simulation Notes Europe* 21 (3–4): 119.

Law, A. M., and W. D. Kelton. 1991. *Simulation Modeling and Analysis*, Volume 2. McGraw-Hill.

Lorig, F., C. A. Becker, and I. J. Timm. 2017. "Formal Specification of Hypotheses for Assisting Computer Simulation Studies". In *TMS/DEVS at SpringSim Multiconferenece*, 1180–1191. SCS.

Maria, A. 1997. "Introduction to Modeling and Simulation". In *Proceedings of the 1997 Winter Simulation Conference*, edited by D. H. Withers, B. L. Nelson, S. Andradóttir, and K. J. Healy, 7–13. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Mill, J. S. 1868. *System of Logic: Vol. II*. Longmans, Green, Reader, and Dyer.

Montgomery, D. C. 2012. *Design and Analysis of Experiments*, Volume 8. John Wiley & Sons.

Morris, M. 1991. "Factorial Sampling Plans for Preliminary Experiments". *Technometrics* 33 (2): 161–174.

Ören, T. I., B. P. Zeigler, and M. S. Elzas. 1984. *Simulation and Model-Based Methodologies: An Integrative View*, Volume 10. NATO ASI Series F: Computer and System Sciences.

Peng, D. 2017. *Reusing Simulation Experiments for Model Composition and Extension*. Rostock University.

Perrone, L. F., C. S. Main, and B. C. Ward. 2012. "SAFE: Simulation Automation Framework for Experiments". In *Proceedings of the 2012 Winter Simulation Conference*, edited by O. Rose, A. Uhrmacher, M. Rabe, C. Laroque, R. Pasupathy, and J. Himmelspach, 2825–2836. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Robinson, S. 2005. "Automated Analysis of Simulation Output Data". In *Proceedings of the 2005 Winter Simulation Conference*, edited by F. Armstrong, J. Joines, N. Steiger, and M. Kuhl, 763–770. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Sanchez, S. M. 2005. "Work Smarter, not Harder". In *Proceedings of the 2005 Winter Simulation Conference*, edited by F. Armstrong, J. A. Joines, N. Steiger, and M. E. Kuhl, 69–82. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Soldatova, L. N., and R. D. King. 2006. "An Ontology of Scientific Experiments". *Journal of the Royal Society Interface* 3 (11): 795–803.

Teran-Somohano, A., A. E. Smith, J. Ledet, L. Yilmaz, and H. Oğuztüzün. 2015. "A Model-Driven Engineering Approach to Simulation Experiment Design and Execution". In *Proceedings of the 2015 Winter Simulation Conference*, edited by C. M. Macal, M. D. Rossetti, L. Yilmaz, I.-C. Moon, W. K. Chan, and T. Roeder, 2632–2643. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Timm, I. J., and F. Lorig. 2015. "A Survey on Methodological Aspects of Computer Simulation as Research Technique". In *Proceedings of the 2015 Winter Simulation Conference*, edited by C. M. Macal, M. D. Rossetti, L. Yilmaz, I.-C. Moon, W. K. Chan, and T. Roeder, 2704–2715. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Tolk, A., B. L. Heath, M. Ihrig, J. J. Padilla, E. H. Page, E. D. Suarez, C. Szabo, P. Weirich, and L. Yilmaz. 2013. "Epistemology of Modeling and Simulation". In *Proceedings of the 2013 Winter Simulation Conference*, edited by R. Hill, M. Kuhn, R. Pasupathy, S.-H. Kim, and A. Tolk, 1152–1166. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Uhrmacher, A., S. Brailsford, J. Liu, M. Rabe, and A. Tolk. 2016. "Reproducible Research in Discrete Event Simulation". In *Proceedings of the 2016 Winter Simulation Conference*, edited by T. Huschka, S. Chick, J. Jiminez, P. Frazier, T. Roeder, R. Szechtman, and E. Zhou, 1301–1315. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Waltemath, D., R. Adams, D. A. Beard, F. T. Bergmann, U. S. Bhalla, R. Britten et al. 2011. "Minimum Information about a Simulation Experiment (MIASE)". *PLoS Comput Biol* 7 (4): e1001122.

Waltz, D., and B. G. Buchanan. 2009. "Automating Science". *Science* 324 (5923): 43–44.

Whewell, W. 1847. *The Philosophy of the Inductive Sciences*. J. W. Parker.

Winsberg, E. 2010. *Science in the Age of Computer Simulation*. University of Chicago Press.

Yilmaz, L., S. Chakladar, and K. Doud. 2016. "The Goal-Hypothesis-Experiment Framework". In *Proceedings of the 2016 Winter Simulation Conference*, 1001–1012. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Yilmaz, L., S. J. Taylor, R. Fujimoto, and F. Darema. 2014. "Panel: The Future of Research in Modeling & Simulation". In *Proceedings of the 2014 Winter Simulation Conference*, edited by S. J. Buckley, J. A. Miller, A. Tolk, L. Yilmaz, S. Y. Diallo, and I. O. Ryzhov, 2797–2811. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Zeigler, B. P., H. Praehofer, and T. G. Kim. 2000. *Theory of Modeling and Simulation*. Academic press.

AUTHOR BIOGRAPHIES

FABIAN LORIG received Master's degree in 2014 and works as research assistant focusing on intelligent assistance for the design and execution of simulation experiments. His email address is lorigf@uni-trier.de.

DANIEL S. LEBHERZ studied Mathematics and Politics, focusing on applied mathematics. As research assistant, he works on mathematical analysis of experiments. His email address is lebherz@uni-trier.de.

JAN OLE BERNDT received Diploma degree (2009) and PhD (2015) in computer science from University of Bremen. He is senior researcher, focusing on ABM and ABSS. His email address is berndt@uni-trier.de.

INGO J. TIMM received Diploma degree (1997), PhD (2004), and *venia legendi* (2006) in computer science from University of Bremen. He is a full professor and both founded and heads the CIRT and its Research Lab on Simulation. His email address is itimm@uni-trier.de.