Hindawi
Complexity
Volume 2018, Article ID 5870987, 15 pages
https://doi.org/10.1155/2018/5870987

*Research Article*

# A Conceptual Approach to Complex Model Management with Generalized Modelling Patterns and Evolutionary Identification

**Sergey V. Kovalchuk**,[1] **Oleg G. Metsker**,[1] **Anastasia A. Funkner**,[1] **Ilia O. Kisliakovskii**,[1] **Nikolay O. Nikitin**,[1] **Anna V. Kalyuzhnaya**,[1] **Danila A. Vaganov**,[1,2] **and Klavdiya O. Bochenina**[1]

[1]*ITMO University, Saint Petersburg, Saint Petersburg, Russia*
[2]*University of Amsterdam, Amsterdam, Netherlands*

Correspondence should be addressed to Sergey V. Kovalchuk; sergey.v.kovalchuk@gmail.com

Received 1 June 2018; Accepted 17 September 2018; Published 1 November 2018

Guest Editor: Rafael Gómez-Bombarelli



Complex systems' modeling and simulation are powerful ways to investigate a multitude of natural phenomena providing extended knowledge on their structure and behavior. However, enhanced modeling and simulation require integration of various data and knowledge sources, models of various kinds (data-driven models, numerical models, simulation models, etc.), and intelligent components in one composite solution. Growing complexity of such composite model leads to the need of specific approaches for management of such model. This need extends where the model itself becomes a complex system. One of the important aspects of complex model management is dealing with the uncertainty of various kinds (context, parametric, structural, and input/output) to control the model. In the situation where a system being modeled, or modeling requirements change over time, specific methods and tools are needed to make modeling and application procedures (metamodeling operations) in an automatic manner. To support automatic building and management of complex models we propose a general evolutionary computation approach which enables managing of complexity and uncertainty of various kinds. The approach is based on an evolutionary investigation of model phase space to identify the best model's structure and parameters. Examples of different areas (healthcare, hydrometeorology, and social network analysis) were elaborated with the proposed approach and solutions.

## 1. Introduction

Today the area of modeling and simulation of complex systems evolves rapidly. A complex system [1] is usually characterized by a large number of elements, complex long-distance interaction between elements, and multiscale variety. One of the results of the area's development is growing complexity of the models used for investigation of complex systems. As a result, contemporary model of a complex system could be easily characterized by the same features as a natural complex system. Usually, a complexity of a model is considered in tight relation to a complexity of a modeling system. Nevertheless, in many cases, the complexity of a model does not mimic the complexity of a system under investigation (at least exactly). It leads to additional issues in managing a complex model during identification, calibration, data assimilation, verification, validation, and application. One of the core reasons for these issues is the uncertainty of various kinds [2, 3] applied on levels of system, data, and model. In addition, complexity is even more extended within multidisciplinary models and models which incorporate additional complex or/and third-party submodels. From the application point of view, complex models are often difficult to support and integrate with a practical solution because of a low level of automation and high modeling skills needed to support and adapt a model to the changing conditions.

On the other hand, recently evolutionary approaches are popular for solving various types of model-centered operations like model identification [4], equation-free methods [5], ensemble management [6], data assimilation [7], and others. Evolutionary computation (EC) provides the ability to implement automatic optimization and dynamic adaptation

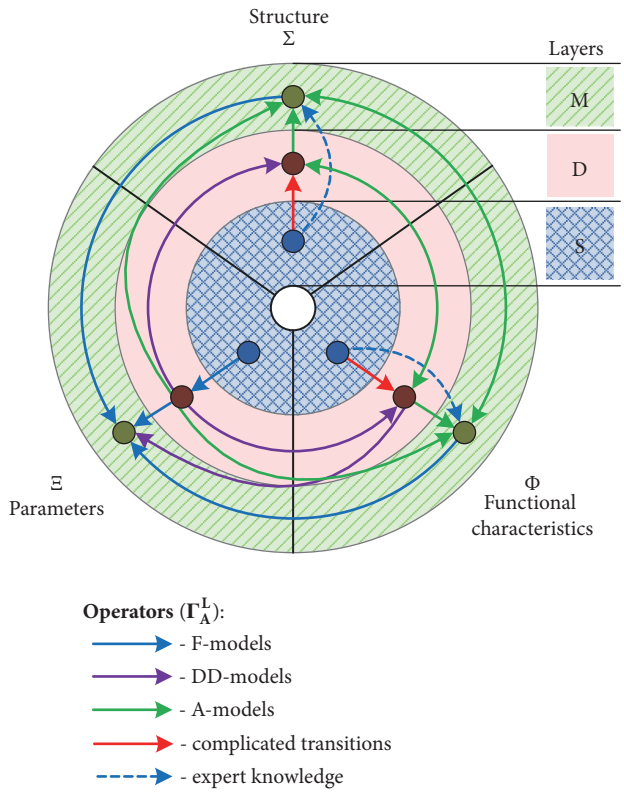



Figure 1: Basic concepts of complex modeling on model ($M$), data ($D$), and system ($S$) layers.

of the system within a complex state space. Still, most of the solutions are still tightly related to the application and modeling system.

Within the current research, we are trying to develop a unified conceptual and technological approach to support core operation with a complex model by distinguishing concepts and operations on model, data, and system levels. We consider a combination of EC and data-driven approaches as a tool for building intelligent solutions for more precise and systematic managing (and lowering) uncertainty and providing the required level of automation, adaptability, and extendibility.

## 2. Conceptual Basis

The proposed approach is based on several key ideas, aimed to extend uncertainty management in complex system modeling and simulation.

(1) Disjoint consideration of model, data, and system in terms of structure, behavior, and quality is aimed toward a system-level review of modeling and simulation process and distinguishes between the uncertainties of various kinds originated from different level [8].

(2) Intelligent technologies like data mining, process mining, machine learning, and knowledge-based approaches are to be hired to fill the gap in automation of modeling and simulation. Key sources for the development of such solution include formalization of various knowledge within composite solutions [9] and data-driven technologies to support the identification of model components.

(3) EC approaches are widespread in modeling and simulation of complex systems [8, 10]. We believe that systematization of this process with separate consideration of spaces for a system (with its subsystems) and a model (with its submodels) could enhance such solutions significantly.

(4) The aim of the approach's development is twofold. First, it is aimed towards automation of modeling operations to extend the functionality of possible model-based applications. Second, working with a combination of EC and intelligent data-driven technologies could be considered as an additional knowledge source for system and model analysis.

Furtherly, this section considers the conceptual basis of the proposed approach with a special focus on the role of EC algorithms and data-driven intelligent technologies for building and exploiting complex models.

*2.1. Core Concepts.* To distinguish between main modeling concepts and operations, we propose a conceptual framework (see Figure 1) for consideration of key processes and operations during modeling of the complex system. The framework may be considered as a generalization and extension of a framework [11, 12] previously defined and used by authors for ensemble-based simulation. Current research extends the concept beyond ensemble-based simulation. It is mainly focused on complex modeling in general with identification of key model management procedures and important

artifacts which can be used for model development and application.

The proposed framework considers three main layers of complex systems' modeling, namely, model ($M$), data ($D$), and system ($S$). Main operations (arrows on the diagram) within the framework are defined within three concepts: quantitative parameters ($\Xi$), functional characteristics ($\Phi$), and structure ($\Sigma$). We denote operations by $\Gamma_L^A$, where $A$ and $L$ stay for concepts and layer (respectively) involved in the operation. Transitions between concepts and between layers are denoted by $A_1 \longrightarrow A_2$ and $L_1 \longrightarrow L_2$ respectively; e.g., operator $\Gamma_{S\longrightarrow D}^{\Xi}$ reflects observation of quantitative parameters and operator $\Gamma_{D\longrightarrow M}^{\Xi}$ stays for basic data assimilation. Also, a set of operators may refer to a single modeling operation; e.g., operators $\Gamma_M^{\Phi\longrightarrow\Xi}$ and $\Gamma_M^{\Sigma\longrightarrow\Xi}$ are often implemented within a single monolithic model. Mainly, operators are related to the specific submodel within a complex model. We consider three key classes of models. F-models are usually classical continuous models developed with knowledge of a system. DD-models are data-driven models based on analysis of available data sets with corresponding techniques (statistics, data mining, process mining, etc.). A-models are mainly intelligent components of a system usually based on machine learning or knowledge-based approaches. Also, we consider EC-based components as belonging to A-models class.

A key problem within complex system modeling and simulation is related to the absent or at least significantly limited possibility to observe the structure and functional characteristics of the system (operators $\Gamma_{S\longrightarrow D}^{\Phi}$ and $\Gamma_{S\longrightarrow D}^{\Sigma}$) directly. The general solution usually includes implicit substitution of the operators with the expertise of modeler (operators $\Gamma_{S\longrightarrow M}^{\Phi}$ and $\Gamma_{S\longrightarrow M}^{\Sigma}$). Still, the more complex the system under investigation and the model are, the more limited those operations are. To overcome this issue, additional DD-models are involved (operators $\Gamma_D^{\Xi\longrightarrow\Sigma}$ and $\Gamma_D^{\Xi\longrightarrow\Phi}$ for mining in available data, $\Gamma_{D\longrightarrow M}^{\Phi\longrightarrow\Xi}$ for extended discovery of model parameters for various functional characteristics). Also, A-models are hired to extend expert knowledge in discovery of $M$-layer concepts with either formalized knowledge or knowledge discovered in data with machine learning approaches (operators $\Gamma_{D\longrightarrow M}^{\Xi\longrightarrow\Sigma}$, $\Gamma_{D\longrightarrow M}^{\Xi\longrightarrow\Phi}$, $\Gamma_{D\longrightarrow M}^{\Sigma}$, and $\Gamma_{D\longrightarrow M}^{\Phi}$ for direct discovery of structure and functional characteristics directly and operators $\Gamma_D^{\Sigma\longrightarrow\Phi}$, $\Gamma_D^{\Phi\longrightarrow\Sigma}$, $\Gamma_M^{\Sigma\longrightarrow\Phi}$, and $\Gamma_M^{\Phi\longrightarrow\Sigma}$ for interconnection of discovered characteristics in available data and within the used model). In the proposed approach, primary attention is paid to these kinds of solutions where DD- and A-models enable enhancement of complex modeling process with an additional level of automation, adaptation, and knowledge providing.

*2.2. Complex Modeling Patterns.* Considering the defined conceptual framework, we identify several patterns of modeling and simulation of a complex system (see Figure 2). The patterns are defined as combinations in a context of the framework described previously (3 layers, 3 concepts). An essential idea of the proposed patterns is systematization of complex model management approaches with combinations of expertise, intelligent solution (A-models), DD-models, and EC.

The pattern extends the operators described in Section 2.1 for model building with operators for model application (i.e., modelling and simulation) and results analysis (e.g., assessing model quality) required for automated model identification. These additional Operators are denoted with $\Gamma'^{A}_{L}$ and similar notation for indices.

***P1.*** Regular modeling of a system (Figure 2(a)) is a basic pattern usually applied to discover new knowledge on the system under investigation. A model is built using (a) expertise of modeled for identification of structure and functional characteristics of the model ($\Gamma_{S\longrightarrow M}^{\Phi}$ and $\Gamma_{S\longrightarrow M}^{\Sigma}$); (b) available input data usually representing quantitative parameters of a system considered as a static input of the model or source for data assimilation (DA) via operators $\Gamma_{S\longrightarrow D}^{\Xi}$ and $\Gamma_{D\longrightarrow M}^{\Xi}$. Results of model application ($\Gamma'^{\Xi}_{M\longrightarrow D}$, $\Gamma'^{\Sigma}_{M\longrightarrow D}$, and $\Gamma'^{\Phi}_{M\longrightarrow D}$) could be considered from descriptive (mainly structural or quantitative characteristics) or predictive (often forecasting or other functional characteristics). The obtained results are analyzed in comparison to available information about the investigated system ($\Gamma'^{\Xi}_{D\longrightarrow S}$, $\Gamma'^{\Sigma}_{D\longrightarrow S}$, and $\Gamma'^{\Phi}_{D\longrightarrow S}$) forming an optimization loop which can be considered within the scope of all three concepts. Certain limitations within this pattern being applied to complex system modeling and simulation are introduced by two factors. First, working with complex structural and functional characteristics of the model requires a high level of expertise which leads to a limitation of extensibility and automation of model operation. Second, performing optimization in a loop with most algorithms require multiple runs of a model. As a result, computational-intensive models have limitations in optimization-based operations (identification, calibration, etc.) due to performance reasons.

***P2.*** Data-driven modeling (Figure 2(b)) provides an extension to the modeling operation describing the relationship between data attributes. Application of data-driven models may be considered as replacement of actual "full" model, providing (a) information about structure of system and model with data mining (DM) and process mining (PM) techniques ($\Gamma_D^{\Xi\longrightarrow\Sigma}$); (b) generating surrogate models for functional characteristics ($\Gamma_D^{\Xi\longrightarrow\Phi}$); (c) providing estimation of investigated parameters with machine learning (ML) algorithms and models ($\Gamma'^{\Xi\longrightarrow\Xi}_{D}$). In contrast to the previous pattern data-driven models usually operate quickly (although it could require significant time to train the model). Still, such models have lower quality than original "full" models. Nevertheless, combining this pattern with others provides significant enhancement in functionality and performance; e.g., data-driven models can be used in optimization loop (see previous pattern).

***P3.*** Ensemble-based modeling (Figure 2(c)) extends P1 for working with sets of objects (models, data sets, and states) reflecting uncertainty, variability, or alternative solutions (e.g., models). Previously [11] we identified 5 classes of ensembles (see E1-E5 in Figure 2(c)): decomposition

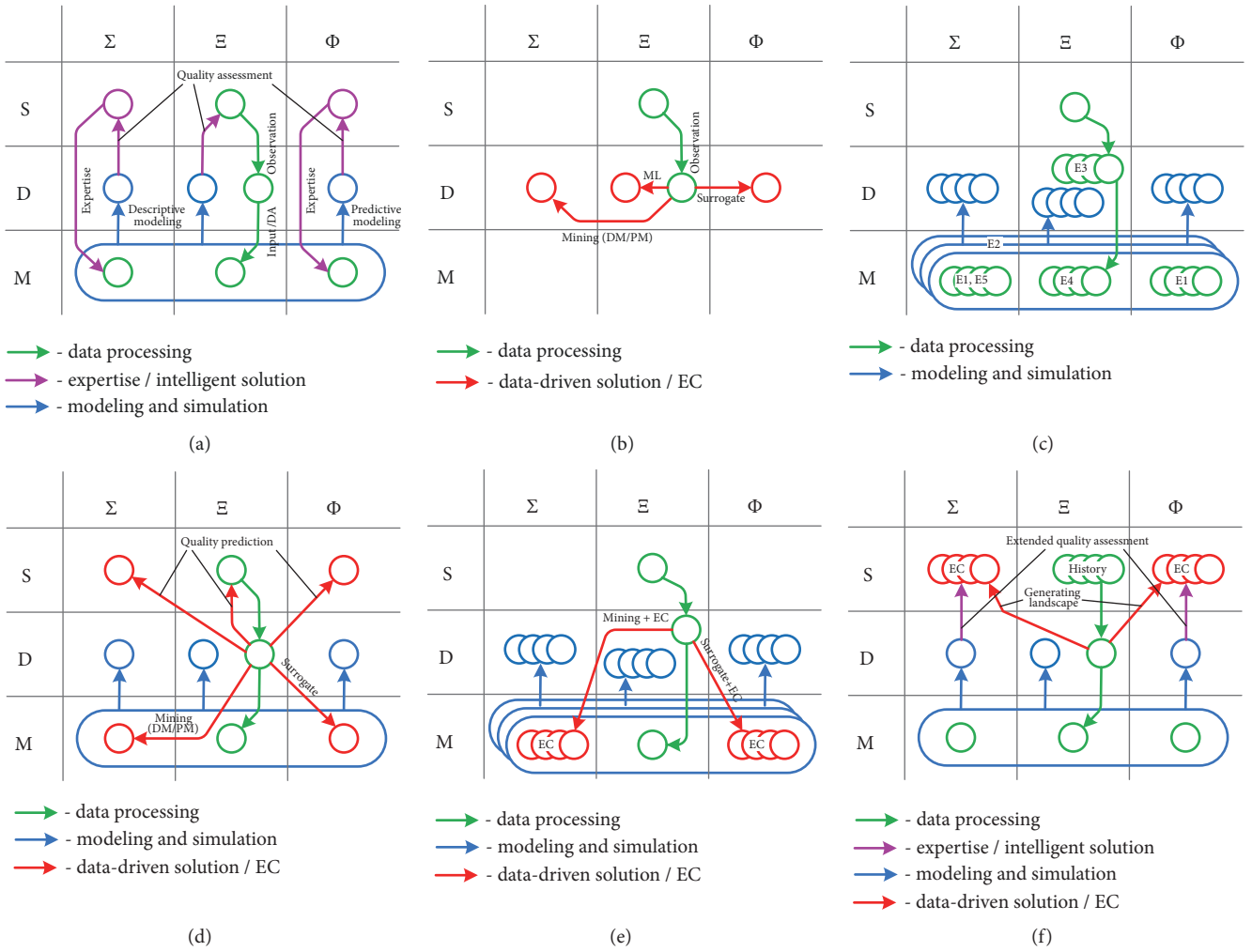



FIGURE 2: Complex modeling patterns: (a) regular modeling; (b) data-driven modeling; (c) ensemble-based modeling; (d) data-driven support of complex modeling; (e) EC in hybrid complex modeling; (f) evolutionary space discovery in hybrid complex modeling.

ensemble, alternative models ensemble, data-driven ensemble, parameter diversity ensemble, and metaensemble. All these patterns can be applied within a context of the proposed framework. Still, an extension of ensemble structure increases structural complexity of the model and thus leads to the need for additional (automatic) control procedures. Moreover, the performance issues of P1 are getting even worthier in ensemble modeling.

***P4.*** One of the key ideas of the proposed approach is an implementation of data-driven analysis of model states, structure, and behavior. To implement it within a conceptual framework we propose pattern for data-driven complex modeling (Figure 2(d)). It includes identification and prediction of a model structure through DM and PM techniques ($\Gamma_{D\longrightarrow M}^{\Xi\longrightarrow\Sigma}$) and generation of surrogate models for injection into the complex model ($\Gamma_{D\longrightarrow M}^{\Xi\longrightarrow\Phi}$). In addition, it is possible to use data-driven techniques to predict the quality of the considered model and use it for model optimization ($\Gamma'^{\Xi}_{D\longrightarrow S}$, $\Gamma'^{\Xi\longrightarrow\Sigma}_{D\longrightarrow S}$, and $\Gamma'^{\Xi\longrightarrow\Phi}_{D\longrightarrow S}$).

***P5.*** A key pattern for EC implementation is presented in Figure 2(e). Here EC is used to identify a model structure ($\Gamma_{D\longrightarrow M}^{\Xi\longrightarrow\Sigma}$) and surrogate submodels ($\Gamma_{D\longrightarrow M}^{\Xi\longrightarrow\Phi}$) with a consideration of population of models. As a result, modeling result is also (as well as in P3) presented in multiple instances which may be analyzed, **filtered end evolved** within consequent iterations over changing time (and processing of coming observations of the system) or within a single timestamp (and fixed observation data).

***P6.*** Finally, last presented pattern (Figure 2(f)) is aimed at investigation of system phase space using DD-models and/or EC to reflect unobservable landscape for estimation of model positioning, assessing its quality in inferring of (sub-)optimal structural ($\Gamma'^{\Xi\longrightarrow\Sigma}_{D\longrightarrow S}$ and $\Gamma'^{\Sigma}_{M\longrightarrow D}$) and functional ($\Gamma'^{\Xi\longrightarrow\Phi}_{D\longrightarrow S}$ and $\Gamma'^{\Phi}_{M\longrightarrow D}$) characteristics of the actual system.

These patterns could be easily combined to obtain better results within a specific application. Especial interest from the point of view of EC is attracted to the patterns where a set of models (or sub-model) instances is considered (P5,

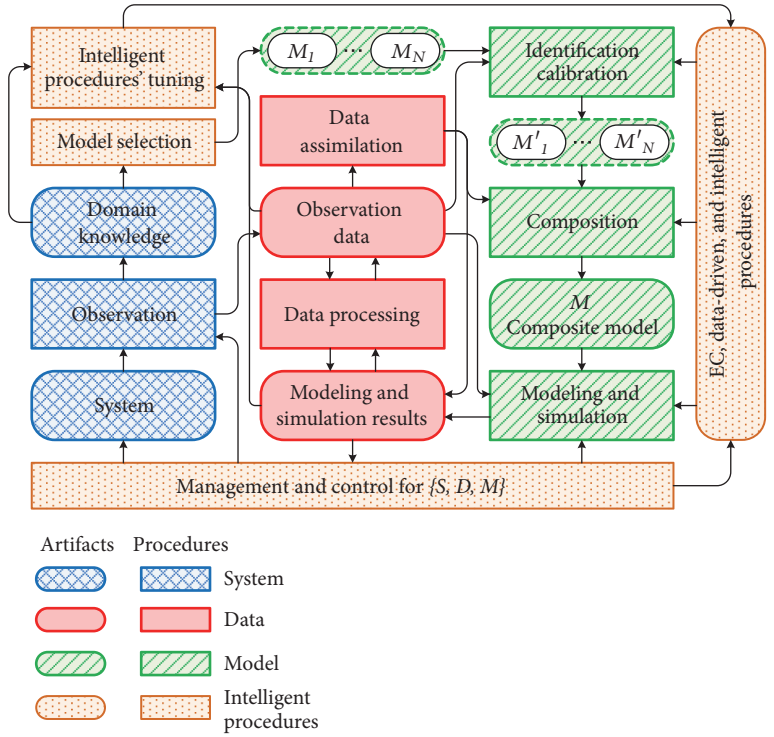



Figure 3: Artifacts and procedures within a typical composite solution.

P6). It is possible to consider ensemble-based techniques (P3) in a fashion of EC, but within our approach, we prefer consideration of ensemble as a composite model with several submodels. In that case, ensemble management refers to the concept of complex model structure.

Several important goals may be reached within the presented patterns:

(i) automation of complex model management with intelligent solutions, DD-models, and EC;

(ii) optimization of model structure and application under defined limitations in precision and performance;

(iii) enhanced ways of domain knowledge discovery for applications and general investigation of a system.

*2.3. Composite Solution Development.* The proposed structure of core concepts and patterns may be applied in various ways to form a solution which combine operators with original implementation within the solutions or implemented as external model calls. Figure 3 shows the essential elements (artifacts and procedures) in a typical composite solution within the proposed conceptual layers ($S$, $D$, and $M$). $S$-layer includes actual system's state which can be assessed through the observation procedure and described by explicit domain knowledge. $D$-layer includes datasets divided into observation data and simulation/modeling data with procedures for data processing and data assimilation. Finally, $M$-layer includes a set of available basic models $M_1 \dots M_N$ which may be identified, calibrated with available data having tuned models $M'_1 \dots M'_N$ as a result. Here, essential elements are model composition (which may be performed either automatically, or by the modeler) and application of the model.

The key benefit of the approach is an application of a combination of EC, data-driven and intelligent procedures to manage the whole composite solution including data processing, modeling, and simulation to lower uncertainty in $\Sigma \times \Xi \times \Phi$. Within the shown structure these procedures may be applied:

(i) to rank and select alternative models;

(ii) to support model identification, calibration, composition, and application;

(iii) to manage artifacts on various conceptual layers in a systematic way;

(iv) to infer implicit knowledge from available data and explicitly presented domain knowledge.

The shown example draws a brief view on the composite solution development while the particular details may differ depending on a particular application. Key important procedures within the proposed composite solution are the implementation of intelligent procedures to support model identification and systematic management of composite model are considered in Sections 2.4 and 2.5.

*2.4. Evolutionary Model Identification.* Implementing evolution of models within a complex modeling task structure, functional and quantitative parameters are usually considered as genotype whereas model output (data layer) are

considered as phenotype. Within the proposed approach we can adapt basic EC operations definition within genotype-phenotype mapping [13]:

(i) epigenesis as model application: $f_1 : S \times M \longrightarrow D$;

(ii) selection: $f_2 : S \times D \longrightarrow D$;

(iii) genotype survival: $f_3 : S \times D \longrightarrow M$;

(iv) mutation: $f_4 : M \longrightarrow M$.

In addition, we consider quality assessment usually treated as fitness for selection and survival (or in more complex algorithms for controlling of other operations like mutation):

(i) data quality: $q_d : S \times D \longrightarrow Q_d$;

(ii) model quality: $q_m : S \times M \longrightarrow Q_m$.

Here $Q_d$ and $Q_m$ are often considered as $\mathbb{R}^N$ with some quantitative quality metrics. Model quality usually are considered through data quality, i.e., $q_m \sim q_d(s, f_1(s, m))$, but within our approach this separation is considered as important because in addition we introduce supporting operations with data-driven procedures as in complex modeling many of these functions (first of all $f_1$, $q_m$, and $q_d$) have significant difficulties to be applied directly (some of these issues are considered in relationship with patterns). Data driven operations (first of all, $f_1$ and $q_m$) can be introduced as substitution of previously introduced basic operations (see also patterns P2, P4, and P6):

(i) epigenesis as DD-model application: $f_1^d : S \times M \longrightarrow D$;

(ii) model generation: $g^d : S \times D \longrightarrow M$;

(iii) model quality prediction: $q_m^d : S \times M \longrightarrow Q_m$;

(iv) space discovery: $w^d : S \times D \longrightarrow S$.

Operation $w^d$ could be used within an intelligent extension within selection or survival operations ($f_2$ and $f_3$). It becomes especially important in case of lack of knowledge in system's structure or functional characteristics. Operation $g^d$ at the same time could be used as a part of mutation operation $f_4$ (or initial population generation). Having this extension, we can implement enhanced versions of EC algorithms (e.g., genetic algorithms, evolution strategies, and evolutionary programming) with data-driven operations to overcome or at least to lower complex modeling issues.

*2.5. Model Management Approach and Algorithm.* By model management we assume operations with models within problem domain solution development and application. This includes identification, calibration, DA, optimization, prediction, and forecasting. To systematize the model management in the presented patterns we propose an approach for explicit consideration of spaces $S$, $D$, and $M$ within hybrid modeling with EC and DD-modeling. To summarize complex modeling procedures within the approach, we developed a high-level algorithm which includes series of steps to be implemented within a context of complex model management.

*Step 1* (space discovery). This step identifies the description of phase space (in most cases, $S$) in case of lack of knowledge or for automation purposes. For example, the step could be applied in the discovery of system state space or model structure. Space description may include (a) distance metrics; (b) proximity structure (e.g., graph, clustering hierarchy, and density); (c) positioning function. One of the possible ways to perform this step is an application of DM and EC algorithm to available data (see pattern P6).

*Step 2* (identification of supplementary functions). Data-driven functions ($\Phi$) are applied to work in model evolution with consideration of space (landscape) representation as available information.

*Step 3* (evolutionary processing of a set of models). This step is described by a combination of basic EC operations (population initialization, epigenesis, selection, mutation, and survival) with supplementary functions. A form of combination depends on (a) selected EC algorithm; (b) application requirements and restrictions; (c) model-based issues (e.g., performance, quality of surrogate models, etc.).

*Step 4* (assimilation of updated data and knowledge). This step is applied for automatic adaptation purposes and implement DA algorithm. DA can be applied to (a) set of models, (b) EC operations (e.g., affecting selection function); (c) supplementary functions (as they are mainly data-driven); (d) phase space description (if descriptive structure is identified from changed data or/and knowledge).

The steps can be repeated in various combination depending on an application and implemented pattern. Also, the steps are general and could be implemented in various ways. Several examples are provided in the Section 3.

*2.6. Available Building Blocks of a Composite Solutions.* EC proposes a flexible and robust solution to identify complex model structures within a complex landscape with possible adaptation towards changing condition and system's state (including new states without prior observation. A significant additional benefit is an ability to manage alternative solutions simultaneously with possible switching and various combination of them depending on the current needs. Still, within the task of model identification and management, the EC (and also many metaheuristics) have certain drawbacks which require additional steps to implement the approach within particular conditions:

(i) high computational cost due to the multiple runs of a model;

(ii) low reproducibility and interpretability of obtained results due to randomized nature of the searching procedure;

(iii) complicated tuning of hyperparameters for better EC convergence;

(iv) indistinct definition of genotype boundaries;

(v) complicated mapping of genotype to phenotype space.

To overcome these issues, the proposed approach involves two options. First, the intelligent procedures may be used to tune EC hyperparameters (P5), predict features of genotype-phenotype mapping, boundaries, etc. (P4), and discover interpretable states and filters (for system, data, and model) to control convergence and adaptation of population (P2, P4, and P5) with interpretable and reproducible (through the defined control procedure). Second, the composite model may use various approaches, methods, and elements to obtain better quality and performance of the solution:

(i) surrogate models (P2, P4, P5) which may increase performance (for example, within preliminary and intermediate optimization steps);

(ii) ensemble models (P3) which may be considered as interpretable and controllable population;

(iii) interpretation and formal inference using explicit domain-specific knowledge and results of data mining to feed procedures of EC and infer parameters in both models and EC.

(iv) controllable space decomposition (P6) with predictive models for possible areas and directions of population migration in EC to explicitly lower uncertainty and obtain additional interpretability;

Finally, an essential feature of the proposed approach is a holistic analysis of a composite solution with possible coevolution models (submodes within a composite model) and data processing procedures.

## 3. Application Examples

This section presents several practical examples where the proposed approach, patterns, or some of their elements were applied. The examples were intentionally selected from diverse problem domains to consider generality of the approach. The considered problems are developed in separated projects which are in various stages. Problem #1 (ensemble metocean simulation) was investigated in a series of projects (see, e.g., [11, 14, 15]). Within this research we are trying to extend model calibration and DA with EC techniques to develop more flexible and accurate multimodel ensembles. Problem #2 (clinical pathways (CPs) modelling) is important in several ongoing project aimed to model-based decision support in healthcare (see, e.g., [16–18]). The proposed approach plays important role by enabling deeper analysis of clinical pathways in various scenarios (interactive analysis of available CPs with identification of clusters of similar patients, DA in predictive modelling of ongoing cases, etc.). Finally, Problem #3 shows very early results in recently started project in online social network analysis.

*3.1. Problem #1: Evolution in Models for Metocean Simulation.* The environmental simulation systems usually contain several numerical models serving for different purposes (complementary simulation processes, improving the reliability of a system by performing alternative results, etc.). Each model typically can be described by a large number of quantitative parameters and functional characteristics that should be adjusted by an expert or using intelligent automatized methods (e.g., EC). Alternative models inside the environmental simulation system can be joined in ensemble according to complex modeling pattern based on evolutionary computing (a combination of P3 and P5 patterns). In the current case study, we introduce an example illustrated an ensemble concept in forms of the alternative models ensemble, parameter diversity ensemble, and metaensemble. For identification of parameters of proposed ensembles (in a case of model linearity) least square method or (in a case of nonlinearity) optimization methods can be used. As we need to take into account not only functional space $\Phi$ and space of parameters $\Xi$ for a single model but also perform optimal coexistence of models in the system (i.e., $\Sigma$), evolutionary and coevolutionary approaches seem to be an applicable technique for this task. It is worth mentioning that coevolutionary approach can be applied to independent model realizations through an ensemble as a connection element. In this case parameters (weights) in the ensemble can be estimated separately from the coevolution procedure in a constant form or dynamically. As a case study of complex environmental modeling we design ensemble model that consists of the SWAN (http://swanmodel.sourceforge.net/) model for ocean wave simulation based on two different surface forcings by NCEP (https://www.esrl.noaa.gov/psd/data/gridded/data.ncep.reanalysis.html) and ERA Interim (https://www.ecmwf.int/en/forecasts/datasets/archive-datasets/reanalysis-datasets/era-interim). Thus, different implementations of SWAN model were connected in the form of an alternative models ensemble with least-squares-calculated coefficients defining structure of the complex model. Two parameters—wind drag and whitecapping rate (WCR)—were calibrated using evolutionary and coevolutionary algorithms implementing $\Gamma_{D\longrightarrow M}^{\Xi\longrightarrow\Phi}$ in P5 (for detailed sensitive analysis of SWAN see [19]). Case of coevolutionary approach can be represented in a form of parameter diversity ensemble, where each population is constructed an ensemble of alternative model results with different parameters. Also, we can add ensemble weights to model parameters diversity and get metaensemble that can be identified in a frame of coevolutionary approach.

In a process of model identification and verification, measurements from several wave stations in Kara sea were used. Fitness function represents the mean error (RMSE) for all wave stations. For results verification MAE (mean absolute error) and DTW (dynamic time wrapping) metrics were used.

Figure 4(a) represents surface (landscape) of RMSE in the space of announced parameters (drag and WCR) for implementations SWAN+ERA and SWAN+NCEP. It can be seen that the evolutionary-obtained results converge to the minimum of possible error landscape. The landscape was obtained by starting the model with all parameters variants

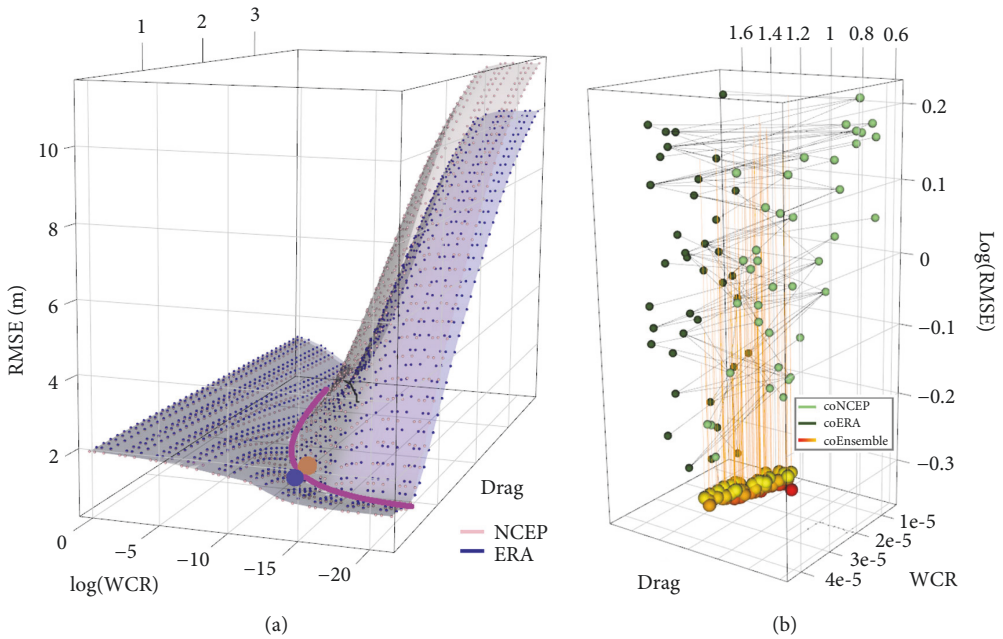



Figure 4: Metocean simulation: (a) error landscape for wave height simulation results using ERA and NCEP reanalysis as input data and (b) Pareto frontier of coevolution results for all generations.

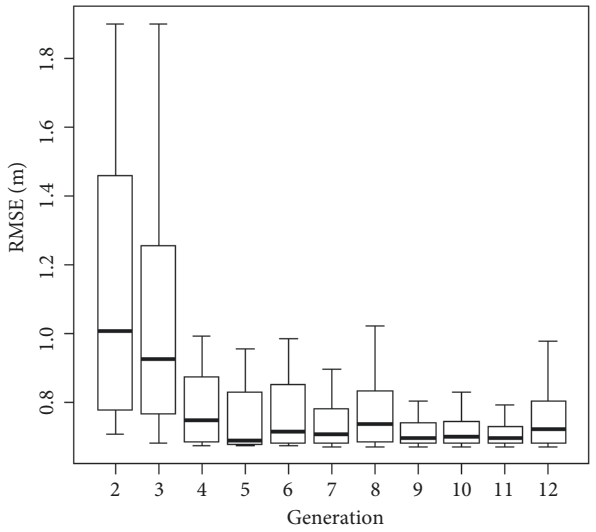



Figure 5: Coevolution convergence of diversity parameters ensemble for metocean models.

from full 30x30 grid (i.e., 900 runs), while evolutionary algorithm was converged in 5 generations with 10 individuals (parameters set) in population (50 runs) that allows performing identification two orders faster. The convergence of co-evolution for SWAN+ERA+NCEP case is presented in Figure 5.

Although error landscapes for a pair of implementations SWAN+ERA and SWAN+NCEP are close to each other, separated evolution does not consider optimization of ensemble result. For this purpose, we apply coevolutionary approach that produces the set of Pareto-optimal solutions for each generation. Figure 4(b) shows that the error of each model in the ensemble is significant (coNCEP and coERA for models along), but the error of the whole ensemble (coEnsemble) converges to minimum very fast.

Obtained result can be analyzed from the uncertainty reduction point of view. Model parameters optimization helps to reduce parameters uncertainty that can be estimated through error function. But when we apply an ensemble approach to evolutionary optimized results, it is suitable to talk about reduction of the uncertainty connected with input data sources (NCEP and ERA) as well. Moreover, metaensemble approach allowed reduction of uncertainty, connected with ensemble parameters.

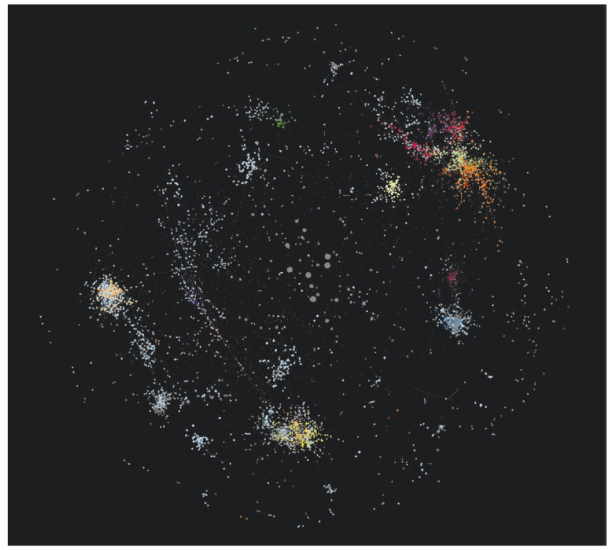


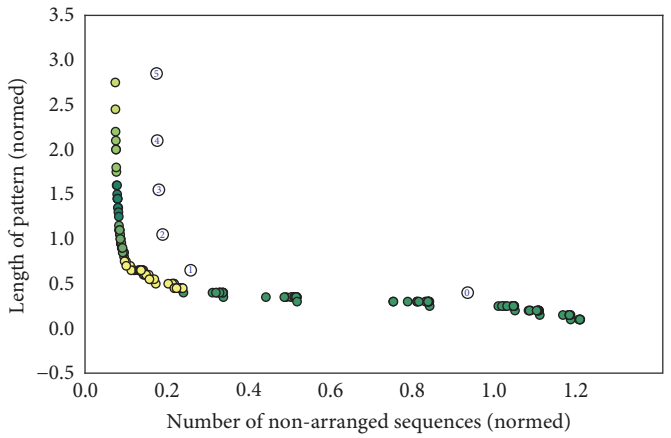
Figure 6: Graph-based representation of processes space in healthcare (interactive view) (Demonstration available at https://www.youtube.com/watch?v=EH74f1w6EeY).

Summarizing results of the metocean case study we can denote that EC approach shows significant efficiency up to 120 times compared with grid search without accuracy losses. According to this experimental study, quality of ensemble with evolutionary optimized models is similar to results of the grid search and MAE metric is equal to 0.24 m and DTW metric – 51. Also, we can mention that coevolutionary approach provides 10 % accuracy gain compared with results of single evolution of model implementations, but this is still similar to ensemble result with evolutionary optimized models. Nevertheless, coevolutionary approach allowed to achieve 200 times acceleration. Within the context of the proposed approach space $\Phi$ were investigated using defined structure of the model in space $\Sigma$ for the purpose of model calibration.

*3.2. Problem #2: Modeling Health Care Process.* Modeling healthcare processes are usually related to the enormous uncertainty and variability even when modeling single disease. One of the ways to identify a model of such process is PM [20]. Still, direct implementation of PM methods does not remove a major part of the uncertainty. Within current research, we applied the proposed approach for identification purposes both in the analysis of historical cases and prediction of single process development. Here we consider processes of providing health care in acute coronary syndrome (ACS) cases which is usually considered as one of the major death causes in the world. We used a set of 3434 ACS cases collected during 2010-2015 in Almazov National Medical Research Centre one of the leading cardiological centers in Russia. The data set contains electronic health records of these patients with all registered events and characteristics of a patient.

To simplify consideration of multidimensional space of possible processes ($\Gamma'^{\Xi\longrightarrow\Sigma}_{D\longrightarrow S}\Gamma^{\Xi}_{S\longrightarrow D}$ for analysis of $\Sigma$ on layer $S$) we introduced graph-based representation of this space with vertices representing cases and edges representing proximity of cases. Analysis of such structure enables easy discovering of common cases (e.g., as communities in graph). Such discovering enables explicit interpretable structuring of the space and representation of further landscape for EC in terms of P6 pattern. Moreover, direct interactive investigation of visual representation of such structure (see Figure 6) provides significant insights for medical researchers.

We have developed evolutionary-based algorithm for patterns identification and clustering in such representation with two criteria to be optimized (see Figure 7). Here processes were represented by a sequence of labels (symbols) denoting key events in PM model. Typical patterns were then selected for Pareto frontier. The convergence process is demonstrated in Figure 8 (10 best individuals from Pareto frontier according to the integral criterion were selected). As a result, this solution may refer to P5 pattern and operator $\Gamma^{\Xi\longrightarrow\Sigma}_{D\longrightarrow M}$ while discovering model structure. Figure 9 shows an example of typical process model (i.e., structural characteristic of the model) for one of the identified clusters. Detailed description of the approach, algorithms, and results on CPs discovering, clustering, and analysis including comparison of three version of CP discovery algorithms with performance comparison can be found in [10]. An important outcome of the approach being applied in this application is interpretability of the clusters and identified patterns. For example, 10 clusters and corresponding CPs obtained interpretation by cardiologists from Almazov National Medical Research Centre. The obtained interpretation and further discovering and application with CP structure are presented in [17]. Another important benefit given by such space structure discovering is lowering uncertainty of patient's treatment trajectory by a hierarchical positioning of an evolved process (selection of a cluster and selection of position within the cluster). For example, discrete-event simulation model




0 AFIFNE
1 AFNIFEIFDNE
2 AFENIFEENIFEDNFIEFE
3 AFNIFEDINIFENDDFFEIDNFEEDFIFE
4 AEFNINFDEDFNEIDFDFNEFDIEDNFNEIFDDNIEFFDE
5 AFANIFADIFEFNIEFDEDEINFDEFDEDFFDNIFENDFIEIDAFEEDEIFEDNA

Figure 7: Pareto frontier for CP patterns discovery.

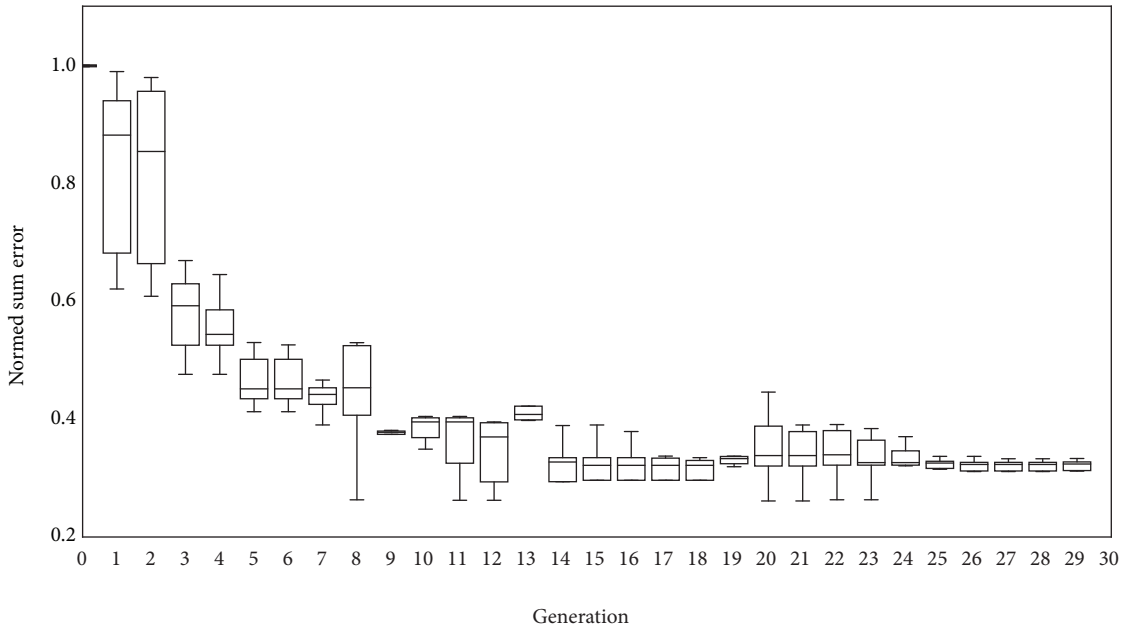



Figure 8: Evolutionary convergence during CP pattern discovery.

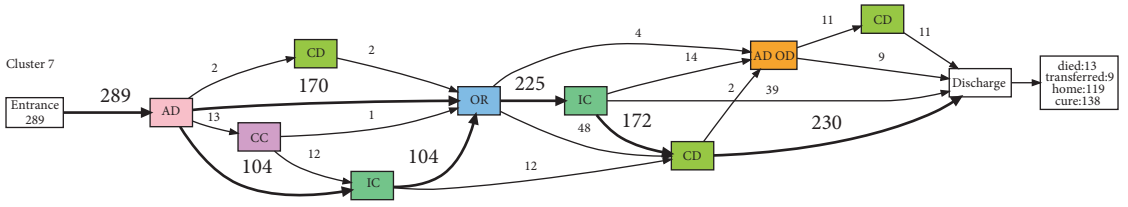



Figure 9: Example of process model showing transfers between hospital's departments.

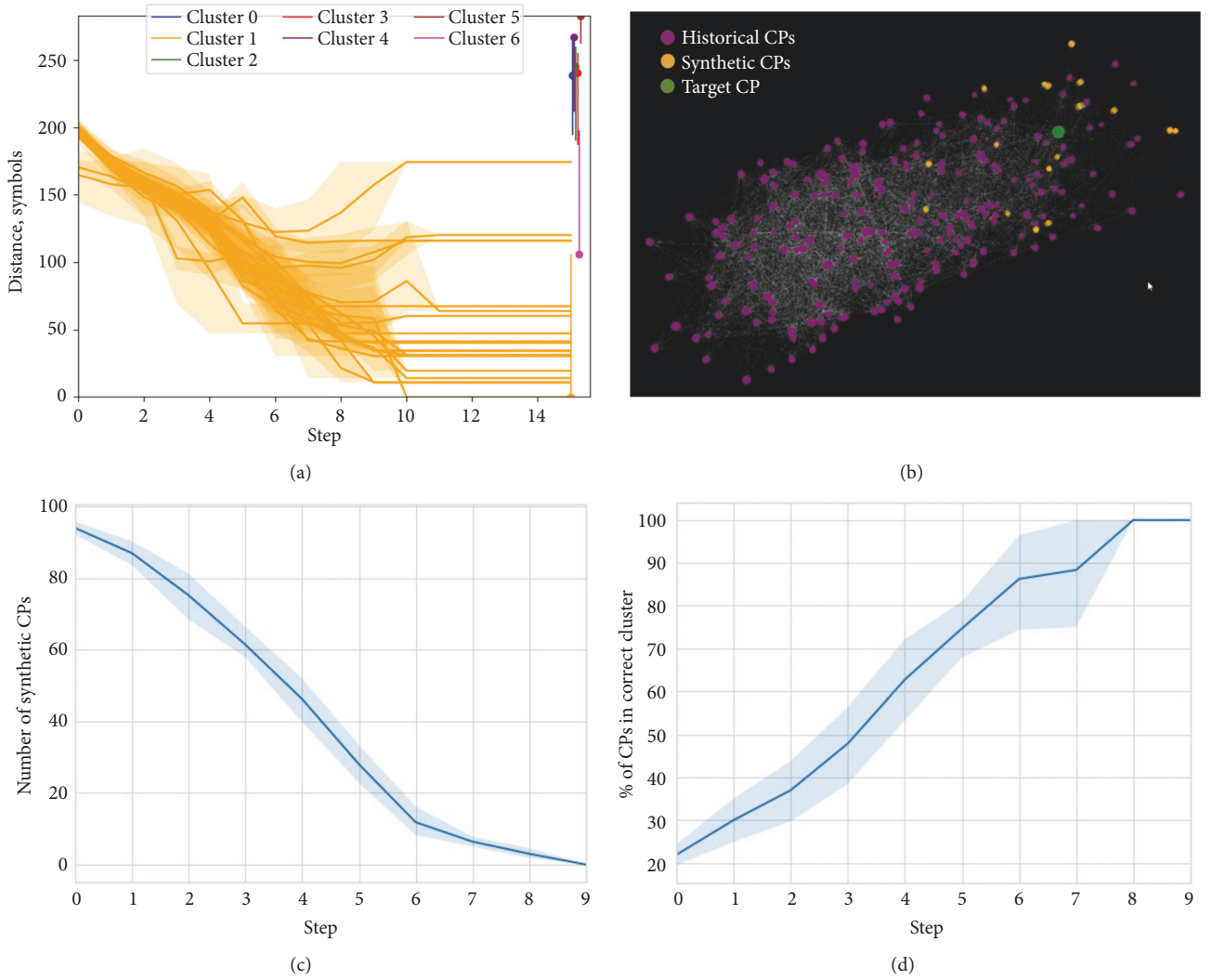



Figure 10: Evolution of synthetic CPs: (a) CP population convergence; (b) evolution of possible CP (demonstration available at https://www.youtube.com/watch?v=twvfX9zKsY8); (c) number of synthetic CPs; (d) % of CPs in correct cluster.

described in [17] provides a more appropriate length of stay distribution within simulation with discovered classes of CPs (Kolmogorov-Smirnov statistics decreased by 51% (from 0.255 to 0.124).

Furtherly we propose an algorithm to dynamically generate possible development of the process in healthcare using identified graph-based space representation with evolutionary strategies, assimilating incoming data (events) within a case (${\Gamma'}^{\Sigma}_{M\longrightarrow D}$ in P5 and $\Gamma_D^{\Xi\longrightarrow\Sigma}$ in P2). We consider convergence (Figure 10(a)) of the introduced synthetic continuation of the processes to the right class (identified clusters of typical cases were used) with mapping to the graph-based space representation with proximity measures (Figure 10(b)). As a result, the appearance of the CP's events decreases the number of synthetic CPs and increases percentage of CPs positioned in the correct cluster (see an example in Figure 10(c) and Figure 10(d) correspondingly). This enables interpretable positioning and uncertainty lowering in predicting further CP's development for a particular patient.

Here a combination of patterns P2, P5, and P6 in the implementation of the proposed algorithm (see Section 2.5) enables interactive investigation of processes space and data assimilation into a population of possible continuations of a single process during its evolving. This solution can be applied in exploratory modeling and simulation of patient flow processing as well as decision support in specialized medical centers.

*3.3. Problem #3: Mining Social Media.* Nowadays social media analysis (that began with static network models emphasizing a topology of connections between users) strives to explore dynamic behavioral patterns of individuals which can be recovered from their digital traces on the web. The prediction of social media activities requires combining analytical and data-driven models as well as identifying the optimal structure and parameters of these models according to the available data. Here we show an example of the problem in this field involving evolutionary identification of a model.

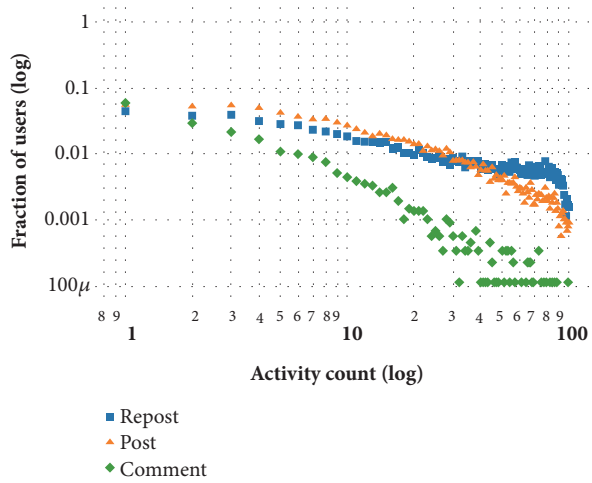



Figure 11: Distribution of posts, reposts, and comments on personal walls of subscribers of bank community.

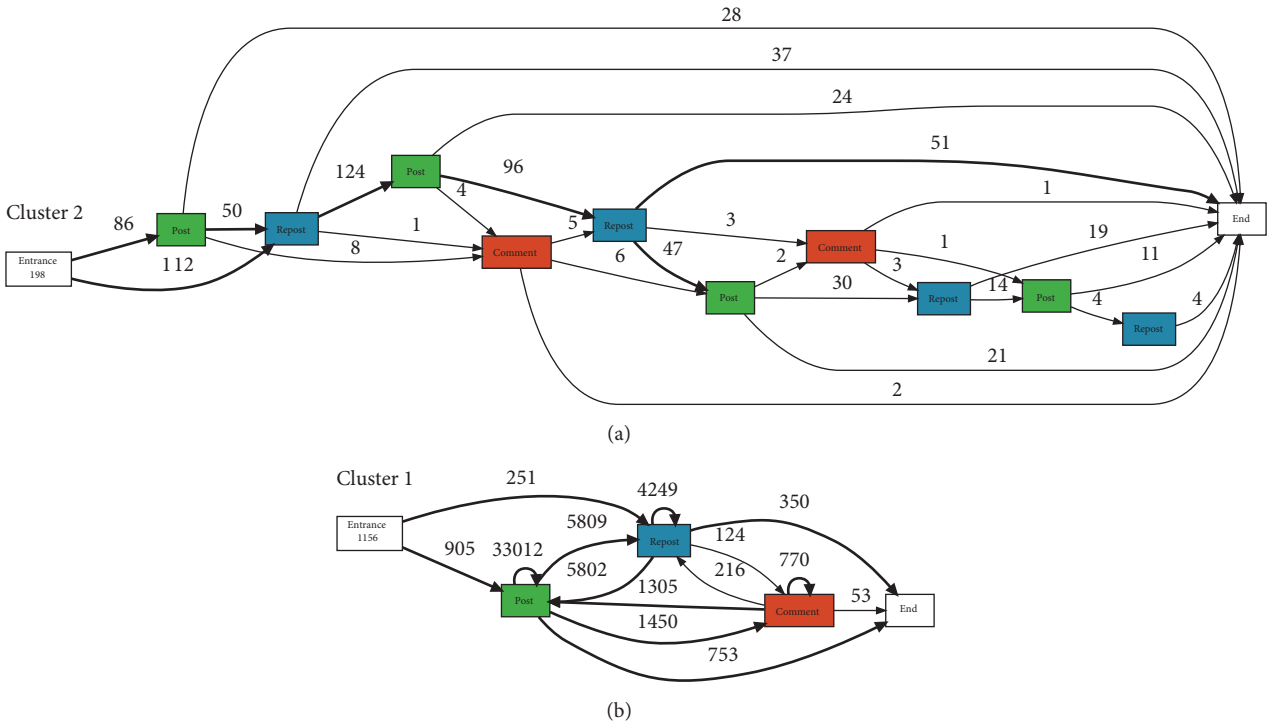



Figure 12: Example of process model (a) with expanded cycles and (b) with collapsed cycles.

A digital trace of a user in an online social network (OSN) is a sequence (chain) of observed activities separated with time gaps. Each OSN supports different types of "hidden" and observable activities. For example, in a largest Russian social network vk.com (further is denoted as VK) a user has a personal page (wall) with three types of activities: post (P)—when a user makes a record by himself; repost (R)—when the user copies the record of another user or community to his or her wall and comment (C)—when the user comments the record on his or her wall. Figure 11 illustrates the distribution of these activities for subscribers of large Russian bank community in VK. The collected dataset consists of 100 (or less if unavailable) last entries (posts or reposts), and comments for the entries for 8K user walls in a period January 2017–December 2017. Comments are much less common than posts and reposts. The distributions of the posts and reposts are similar, but there is a group of "spreaders" with a significant number of reposts.

We applied the technique described in Section 3.2 to analyze the processes. Still, the considered process has significantly different structure. By default, it is continuous with random repetition of events, while healthcare process in ACS cases has finite and more "strong" structure. Figure 12 shows a typical process structure identified with EC-based approach

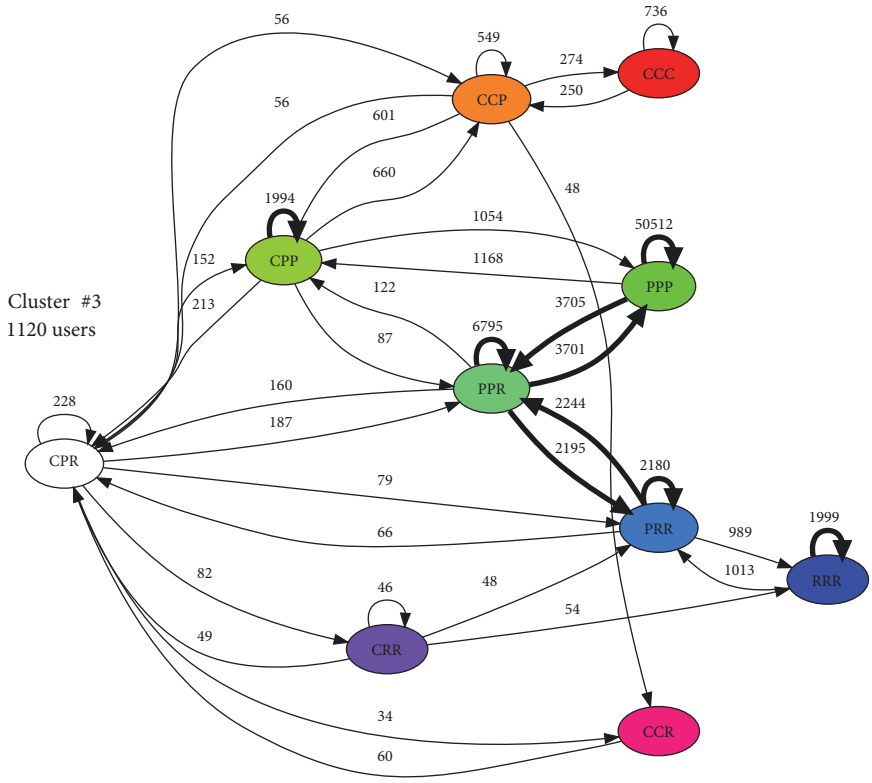



Figure 13: Example of process model for cluster #3 using n-gram analysis.

Table 1: Mean of activities' combinations for users' clusters.

| Cluster | Size | CCC | CCP | CCR | CPP | CPR | CRR | PPP | PPR | PRR | RRR |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 5238 | 0.52 | 0.74 | 0.24 | 1 | 0.72 | 0.42 | 6.68 | 6.4 | 7.56 | 8.53 |
| 2 | 2110 | 0.11 | 0.13 | 0.16 | 0.16 | 0.36 | 0.59 | 2.09 | 3.95 | 12.32 | 60.3 |
| 3 | 1120 | 0.91 | 1.38 | 0.14 | 3.62 | 0.75 | 0.18 | 50.02 | 11.78 | 5 | 2.77 |

and visualized with expanded cycles (a) and with collapsed cycles (b). The second one could be considered as more relevant than the first one which is significantly affected by a length of selected history. It is natural to consider it as a random process or state-transition model. In that case, three identified clusters (characterized by various frequencies of transitions) could be interpreted as typical behavior models.

N-grams analysis is often used to detect patterns in people's behaviors [21, 22]. N-grams analysis is based on counting frequencies of combinations or sequences of activities. We collected all sorted 3-grams (so called 3-sets) for each user's sequence to analyze the frequency of event combinations. As a result, three clusters of vectors with 3-sets chains were identified with k-means clustering method. Figure 13 shows all combinations and transitions between them for cluster #3 as an example. Using Figure 12 and Table 1, it is possible to see that cluster #3 includes users who often make new records (P) and sometimes comment records (C). So, cluster #3 mostly consists of "bloggers." Cluster #2 includes "spreaders" who copy other records (R) frequently. And the biggest cluster #1 consists of people who make new records and copy other ones equally but less intensively comparing to other clusters. That may be considered as a typical behavior for user of OSN. N-grams analysis allows detecting typical behavioral patterns and obtaining process models for social media activities using chains of different lengths as input data. Thus, this type of data-driven modeling is more appropriate to research continuous processes. Figure 14 shows a graph-based representation of process space with of all users' patterns.

This subsection provides very early results. Next step within application of the proposed approach in this application includes an extension of process model structure (a) with temporal labeling (gaps between events); (b) considering process within a sliding time window to get more structured processes; (c) linking the model with causal inference; (d) introduction of DM techniques for EC positioning of ongoing processes in model space. We believe that these extensions could enhance discovery of model structure (P4) and provide deeper insight on social media activity investigation.

## 4. Conclusion and Future Work

The development of the proposed approach is still an ongoing project. We aimed for further systematization and detailing

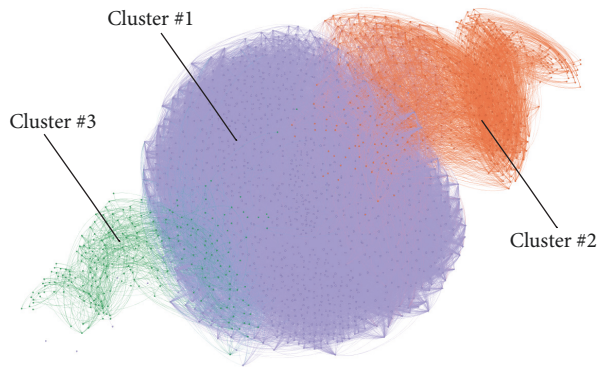



Figure 14: Graph-based representation of processes' space for three clusters in social media activity.

of the proposed concepts, methods, and algorithms, as well as more comprehensive and deeper implementation of EC-based applications. Further work of the development includes the following directions:

(i) dualization on the role of data-driven and intelligent operations in proposed approach and described patterns;

(ii) extended analysis of various EC techniques applicable within the approach;

(iii) investigation on EC-based discovery for models of complex systems with lack or inconsistent observations;

(iv) detailed formalization of expertise and knowledge-based methods within the approach;

(v) extending the approach with interactive user-centered modelling and phase space analysis;

(vi) development of multilayered approach for decision support and control of system and process $S$, available data $D$, and complex model $M$.

## Data Availability

The data used in the examples presented in Section 3 were initially obtained within complimentary projects performed by the authors of the paper. The data is available from the corresponding author upon request after explicit claiming of the purpose and plan of requested data usage to check for possible violation of the corresponding projects' rules.

## Conflicts of Interest

The authors declare that they have no conflicts of interest.

## Acknowledgments

This paper presents an extension and further development of the work [23]. This research is financially supported by The Russian Scientific Foundation, Agreement #14-11-00823 (15.07.2014).

## References

[1] N. Boccara, *Modeling complex systems*, Graduate Texts in Physics, Springer, New York, Second edition, 2010.

[2] H. McManus and D. Hastings, "A framework for understanding uncertainty and its mitigation and exploitation in complex systems," *IEEE Engineering Management Review*, vol. 34, no. 3, pp. 81–94, 2006.

[3] W. Walker, P. Harremoës, J. Rotmans et al., "Defining uncertainty: a conceptual basis for uncertainty management in model-based decision support," *Integrated Assessment*, vol. 4, no. 1, pp. 5–17, 2003.

[4] J. Yan and J. R. Deller, "NARMAX model identification using a set-theoretic evolutionary approach," *Signal Processing*, vol. 123, pp. 30–41, 2016.

[5] I. G. Kevrekidis, C. W. Gear, and G. Hummer, "Equation-free: The computer-aided analysis of complex multiscale systems," *AIChE Journal*, vol. 50, no. 7, pp. 1346–1355, 2004.

[6] H. Ihshaish, A. Cortés, and M. A. Senar, "Parallel Multi-level Genetic Ensemble for Numerical Weather Prediction Enhancement," *Procedia Computer Science*, vol. 9, pp. 276–285, 2012.

[7] G. Dumedah, "Formulation of the Evolutionary-Based Data Assimilation, and its Implementation in Hydrological Forecasting," *Water Resources Management*, vol. 26, no. 13, pp. 3853–3870, 2012.

[8] V. V. Kashirin, A. A. Lantseva, S. V. Ivanov, S. V. Kovalchuk, and A. . Boukhanovsky, "Evolutionary simulation of complex networks' structures with specific functional properties," *Journal of Applied Logic*, vol. 24, no. part A, pp. 39–49, 2017.

[9] S. V. Kovalchuk, P. A. Smirnov, K. V. Knyazkov, A. S. Zagarskikh, and A. V. Boukhanovsky, "Knowledge-Based Expressive Technologies Within Cloud Computing Environments," in *Practical Applications of Intelligent Systems*, vol. 279 of *Advances in Intelligent Systems and Computing*, pp. 1–11, Springer, Berlin, Germany, 2014.

[10] A. A. Funkner, A. N. Yakovlev, and S. V. Kovalchuk, "Towards evolutionary discovery of typical clinical pathways in electronic health records," *Procedia Computer Science*, vol. 119, pp. 234–244, 2017.

[11] S. V. Kovalchuk and A. V. Boukhanovsky, "Towards Ensemble Simulation of Complex Systems," *Procedia Computer Science*, vol. 51, pp. 532–541, 2015.

[12] S. V. Kovalchuk, A. V. Krikunov, K. V. Knyazkov, and A. V. Boukhanovsky, "Classification issues within ensemble-based simulation: application to surge floods forecasting," *Stochastic Environmental Research and Risk Assessment*, vol. 31, no. 5, pp. 1183–1197, 2017.

[13] D. Fogel, "Phenotypes, genotypes, and operators in evolutionary computation," in *Proceedings of the 1995 IEEE International Conference on Evolutionary Computation*, p. 193, Perth, WA, Australia.

[14] S. V. Ivanov, S. V. Kovalchuk, and A. V. Boukhanovsky, "Workflow-based Collaborative Decision Support for Flood Management Systems," *Procedia Computer Science*, vol. 18, pp. 2213–2222, 2013.

[15] A. Gusarov, A. Kalyuzhnaya, and A. Boukhanovsky, "Spatially adaptive ensemble optimal interpolation of in-situ observations into numerical vector field models," in *Proceedings of the 6th International Young Scientist Conference on Computational Science, YSC 2017*, pp. 325–333, Finland, November 2017.

[16] A. V. Krikunov, E. V. Bolgova, E. Krotov, T. M. Abuhay, A. N. Yakovlev, and S. V. Kovalchuk, "Complex data-driven predictive modeling in personalized clinical decision support for Acute Coronary Syndrome episodes," in *Proceedings of the International Conference on Computational Science, ICCS 2016*, pp. 518–529, USA, June 2016.

[17] S. V. Kovalchuk, A. A. Funkner, O. G. Metsker, and A. N. Yakovlev, "Simulation of patient flow in multiple healthcare units using process and data mining techniques for model identification," *Journal of Biomedical Informatics*, vol. 82, pp. 128–142, 2018.

[18] A. Yakovlev, O. Metsker, S. Kovalchuk, and E. Bologova, "Prediction of in-hospital mortality and length of stay in acute coronary syndrome patients using machine-learning methods," *Journal of the American College of Cardiology*, vol. 71, no. 11, p. A242, 2018.

[19] A. Nikishova, A. Kalyuzhnaya, A. Boukhanovsky, and A. Hoekstra, "Uncertainty quantification and sensitivity analysis applied to the wind wave model SWAN," *Environmental Modelling & Software*, vol. 95, pp. 344–357, 2017.

[20] E. Rojas, J. Munoz-Gama, M. Sepúlveda, and D. Capurro, "Process mining in healthcare: A literature review," *Journal of Biomedical Informatics*, vol. 61, pp. 224–236, 2016.

[21] T. Sinha, P. Jermann, N. Li, and P. Dillenbourg, "Your click decides your fate: Inferring Information Processing and Attrition Behavior from MOOC Video Clickstream Interactions," in *Proceedings of the EMNLP 2014 Workshop on Analysis of Large Scale Social Interaction in MOOCs*, pp. 3–14, Doha, Qatar, October 2014.

[22] C. Marceau, "Characterizing the Behavior of a Program Using Multiple-Length N-Grams," Defense Technical Information Center, 2005.

[23] S. V. Kovalchuk, O. G. Metsker, A. A. Funkner et al., "Towards management of complex modeling through a hybrid evolutionary identification," in *Proceedings of the the Genetic and Evolutionary Computation Conference Companion*, pp. 255-256, Kyoto, Japan, July 2018.