# On Using Clustering for the Optimization of Hydrological Simulations

Elnaz Azmi
*Steinbuch Centre for Computing*
*Karlsruhe Institute of Technology*
Karlsruhe, Germany
Email: elnaz.azmi@kit.edu

***Abstract*—Accurate water-related predictions and decision-making require a simulation of hydrological systems in high spatio-temporal resolution. However, the simulation of such a large-scale dynamical system is compute-intensive, and hence time consuming. One approach to circumvent these issues is to use landscape properties to reduce model redundancies and computation complexities. This work shows an ongoing project that applies existing clustering methods to identify functionally similar model units and runs the model only on representative model units. The proposed approach consists of several steps, in particular the reduction of dimensionality of the hydrological time series, application of clustering methods, choice of cluster representative, and study of the balance between the uncertainty of the simulation output and the computational effort.**

## I. Introduction

The simulation of hydrological systems and their interactions needs an advanced modeling of water, energy and mass cycles in high spatio-temporal resolution [1] in order to support water-related predictions and decision making. One approach for fast and efficient simulation of distributed and physically based models in high resolution is to use high performance computing (HPC) and parallel processing of the model units [2]. However, parallel running of the whole model is challenging, since the interactions among the model units are not independent. Thus, one can run the processes of independent model units in parallel and the processes of dependent model units sequentially or in parallel using a Message Passing Interface (MPI) for communication and exchange of data between processes. Furthermore, development, test, execution and update of such a model on HPC Clusters involve a potentially large configuration overhead and require advanced programming expertise of domain scientists. Given these points, the main goal of this work is to reduce the computational effort of the model so that it can be run on a desktop computer. The additional objective of the work is to discover underlying patterns of hydrological systems.

## II. Problem Statement

The studied hydrological model in this work is the CAOS (Catchment as Organized Systems) model proposed in [1], which simulates water related dynamics. It provides a high-resolution and distributed process based simulation of water and energy fluxes in the near surface atmosphere, the earth's surface and subsurface. These simulations are generally applicable to the field of hydrological research, agricultural water demand estimation and erosion protection or flood forecasting. In the CAOS model, the landscape is represented as an organized network of model units [1]. The network consists of a hierarchy of different nodes which is abstracted into the network model shown in Fig. 1. The interactions between the nodes are modeled as directed edges. Regarding Fig. 1, the object, vertical container as well as collector nodes interact with the members of the same category. However, the horizontal container nodes function independent of each other and interact with collectors. The blocks of several connected collectors that aggregate an output, generate a large network. The CAOS model simulates water dynamics flowing in the network with a time resolution of five minutes and creates an output time series for any requested node. The water dynamics are influenced by the landscape structure i.e. the properties of nodes, current state of water, and rainfall or radiation (forcing) time series.

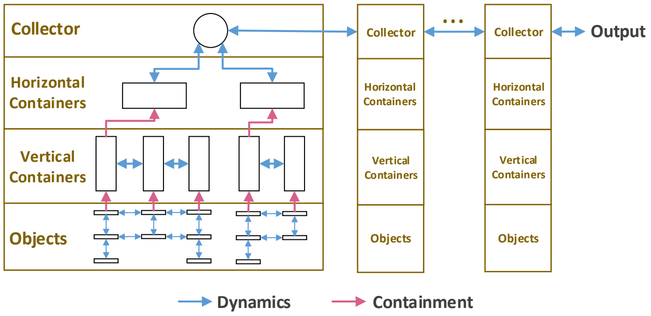

Fig. 1. The hierarchical network model abstracted from the CAOS model

The execution of a high spatio-temporal resolution simulation is very time consuming. For instance, the simulation of the CAOS model in a four square kilometers' area for one year takes about one month on a single CPU. As the available simulation code from the domain scientists can only be run sequentially, first an elementary parallelization of the simulation is done in order to speedup the simulation considering the available number of cores in a test PC. Therefore, I run the simulation concurrently at horizontal container nodes level (the independent nodes) on multi-core processors. Indeed, porting the sequential code to a parallel version that only runs the independent nodes in parallel is more straightforward to implement than parallelization of the dependent nodes and introducing communication between computational parts, e.g. by using MPI. Applying this step, I achieve $9.4$ times speedup using a 16-cores processor.

After this preparation step, I propose a method that can be categorized into Model Order Reduction techniques [3]. Such techniques aim to reduce the computational costs by dimensionality reduction and by computing an approximation of the original model. I exploit the hydrological similarities [4] to reduce the model complexity and computational efforts. The underlying idea of this approach is that nodes with similar properties function similarly if their simulation starts from similar initial states and they are exposed to similar forcing. The main question here is: How to represent hydrological similarity between the nodes in computer science? Since nodes of each level contain a hierarchy of another type of nodes and have their own specific properties, defining unique similarity parameters for all nodes is difficult. One approach to define the similarity is to use the functionality of the container nodes as structure characteristics of them. The functionality data are obtained from a simulation, initiated with full storage of water without application of forcing. The simulation is referred to as drainage test and its output is time series data for each container node. Detecting similar time series leads to recognition of groups of similar nodes. Considering the lack of labeled data (ground truth), a clustering method can be applied to find similar time series. The relevant question at this point is: Which clustering method results in a highly accurate clustering based on the evaluation metrics?

The application of clustering methods requires some preprocessing of the input data or setting of initial parameters. For time series clustering, four major components namely dimensionality reduction, distance measurement, clustering algorithm and evaluation are applied [5]. To make time series compatible with the conventional clustering algorithms, I convert them into equal length feature vectors. I test my approach with some standard clustering methods like $K$-means and DBSCAN. Applying the clustering to the feature vectors, the accuracy and validation of the extracted clusters are evaluated. This is a challenging issue without having a ground truth. Therefore, heuristic arguments are used to judge the quality of the results [6]. Using the output of variously configured clustering algorithms, I run the simulation only on the representative of each cluster and map the output of the representative simulation to the other members of every cluster. As a result, the uncertainty of the approximation of the original simulation outputs can be controlled by clustering parameters e.g number of clusters and the simulation computation time. With this approach, I save the computation of similar and redundant nodes and calculate an estimation of the original simulation outputs.

## III. Related Work

Classification and clustering are the mostly used methods in environmental science in order to detect patterns in data sets, make decisions and extract the required information by using similarity measurements [7]. To analyze the uncertainty of weather situations, $K$-means, Clara, HClust and Fuzzy clustering algorithms were studied in [8]. They proposed a method to decrease the RMSE of point forecasts by up to $10\%$. To predict the minimum and maximum weather-based meteorological data, the application of $K$-means was compared with Hierarchical clustering using internal validation measures in [9]. The spectral clustering was used to determine the coherent precipitation regime regions in [10]. They obtained spatial patterns of the precipitation regions that provide a new hydroclimatological insight to understand the hydrological systems.

## IV. Summary and Outlook

This work introduces an approach to make use of landscape structure to reduce redundancies of the hydrological model and its computation complexities. It uses time series clustering to initially cluster the nodes based on their structure. The approach shows promising results using $K$-means clustering. As a forward step, the clustering approach will be extended to consider also current state of the model units as well as forcing time series in the simulation model.

## Acknowledgment

I would like to thank the CAOS Research Unit (http://www.caos-project.de/) for making the data set available.

## References

[1] E. Zehe, U. Ehret, L. Pfister, T. Blume, B. Schroeder, M. Westhoff, C. Jackisch, S. J. Schymanski, M. Weiler, K. Schulz *et al.*, "Hess opinions: From response units to functional units: a thermodynamic reinterpretation of the hru concept to link spatial organization and functioning of intermediate scale catchments," *Hydrology and Earth System Sciences*, vol. 18, no. 11, pp. 4635–4655, 2014.

[2] R. Maxwell, L. Condon, and S. Kollet, "A high-resolution simulation of groundwater and surface water over most of the continental us with the integrated hydrologic model parflow v3," *Geoscientific model development*, vol. 8, no. 3, p. 923, 2015. [Online]. Available: https://www.geosci-model-dev.net/8/923/2015/

[3] P. Benner and H. Faßbender, "Model order reduction: Techniques and tools," *Encyclopedia of Systems and Control*, pp. 1–10, 2013.

[4] U. Ehret, E. Zehe, U. Scherer, and M. Westhoff, "Dynamical grouping and representative computation: a new approach to reduce computational efforts in distributed, physically based modeling on the lower mesoscale," *presented at the AGU Chapman conference, 23–26 September, 2014*, no. Abstract 2093, 2014.

[5] S. Aghabozorgi, A. S. Shirkhorshidi, and T. Y. Wah, "Time-series clustering–a decade review," *Information Systems*, vol. 53, pp. 16–38, 2015.

[6] T. Hastie, R. Tibshirani, and J. Friedman, "Unsupervised learning," in *The elements of statistical learning*. Springer, 2009, pp. 485–585.

[7] Á. Arroyo, V. Tricio, E. Corchado, and Á. Herrero, "A comparison of clustering techniques for meteorological analysis," in *10th International Conference on Soft Computing Models in Industrial and Environmental Applications*. Springer, 2015, pp. 117–130.

[8] A. Zarnani, P. Musilek, and J. Heckenbergerova, "Clustering numerical weather forecasts to obtain statistical prediction intervals," *Meteorological Applications*, vol. 21, no. 3, pp. 605–618, 2014.

[9] N. Shobha and T. Asha, "Monitoring weather based meteorological data: Clustering approach for analysis," in *Innovative Mechanisms for Industry Applications (ICIMIA), 2017 International Conference on*. IEEE, 2017, pp. 75–81.

[10] M. Türkeş and H. Tatlı, "Use of the spectral clustering to determine coherent precipitation regions in turkey for the period 1929–2007," *International Journal of Climatology*, vol. 31, no. 14, pp. 2055–2067, 2011.