# CANOPY, COBWEB, DBSCAN clustering data mining techniques: study and analysis

**Varsha Duhoon***
*University School of Basic and Applied Sciences*
*Non-linear Dynamics Research Lab*
*Guru Gobind Singh Indraprastha University*
*New Delhi*
*India*
*varshaduhoon5@gmail.com*

**Rashmi Bhardwaj**
*University School of Basic and Applied Sciences*
*Non-linear Dynamics Research Lab*
*Guru Gobind Singh Indraprastha University*
*New Delhi*
*India*
*rashmib22@gmail.com*

**Abstract.** Clustering is a process of grouping objects belonging to similar class or kind of objects which are collected and put together in same cluster or else grouped in other cluster based on similarity. The study focuses on the CANOPY, COBWEB and DBSCAN methods of clustering to cluster weather parameters in order to gain insight of the pattern being followed. Clustering technique as an unsupervised form of learning in Data mining helps in providing insight into the distribution of data to visualise and analyze characteristic of each cluster. The data considered in the study is of daily weather parameters for Delhi region from $1^{st}$ January, 2017 to $31^{st}$ October, 2018. The clustering of data is carried out to study the nature using different methods of clustering and the efficiency of these methods are then compared to assess the best suited method based on time taken to form clusters.

**Keywords:** Canopy, Cobweb, DBSCAN, data mining, weather parameters.

## 1. Introduction

Clustering is the process of making group of abstract objects into classes of same objects. The objects are grouped together due to similarity in them, either on the basis of pattern or nature. Cluster of the data values are taken as one group. The algorithm for calculating maximum likelihood from incomplete data at different levels has been done. The study was done on the basis of many examples that are sketched, including missing value situations, applications to grouped, censored or truncated data, finite mixture models, variance component estimation, hyperparameter estimation, iteratively re-weighted least squares and

*. Corresponding author

factor analysis [1]. AI with wavelet decomposition conjuction model was applied for analyzing the river water quality [2]. Hybrid soft computing techniques were also used for studying the behavior of Covid-19 using the data [3]. The machine learning bases assessment of HIV epidemiology was carried out for the Asian region in [4]. Also for nonlinear systems a lot of studies have been carried out using stability analysis in different fields of ecology [6] and physics [5].

For instance different methods such as consensus algorithm, developing a system using Blockchain for prevention of Diabetes, numerical simulation for analyzing the COVID-19 data to understand the situation and analyze, optimizing techniques, evolutionary techniques, model development for forecasting using neural networks, study of effects of magnetic and temperature variation on aluminum oxide was done [7-10, 13-15]. Under time series analysis auto regressive methods have been also used to study the time series of rainfall on daily basis and further the best suited model on the basis of AIC was selected to forecast rainfall [11].

Different parameters of weather were considered to study the pattern of the weather parameters and the classification and clustering techniques were used to study the parameters and the best model among classification and clustering were chosen on the basis of least error in classification techniques and the model taking less time was selected for the clustering methods [12]. The fractal and wavelet to study air and water pollutants behavior were studied by calculating fractal dimension, Hurst exponent and predictability index; further concluding time series showing Brownian behavior [16]. Another study of environmental pollutants was done using nonlinear techniques was done [17] while analysis of spread of COVID-19 for China region was done [18]. Soft computing methods have been used to study water quality in [19]. On the other hand EM clustering algorithm was studied in order to cluster 2 stochastic versions [20]. In [21] Forecasting of weather has been done using K-Means clustering was done while fuzzy randomness was studied in [22]. In article [23] derivation and uses of measures of similarity among 2 hierarchical clusters was studied.

The Density-based clusters are separated from each other by contiguous regions of low density of objects in the article [24]. A general algorithm is studied which includes the application of James-Stein type adjustment further in which e James-Stein shrinkage estimators act as the new centroids in the next clustering iteration until convergence, further the testing of accuracy has been done by using real data example [25]. The clustering methods have been examined in high dimension, EM algorithm significantly outperforms the other methods, and proceed to investigate the effect of various initialization schemes on the final solution produced by the EM algorithm [26]. It is a density-based clustering algorithm because it finds a number of clusters starting from the estimated density distribution of corresponding nodes. DBSCAN is one of the most common clustering algorithms and also most cited in scientific literature [27].

In this study a new clustering technique is introduced which will not just make clusters of data and will also create proper order of data base showing density based clustering structures [28]. The ranking and clustering method is used for aggregating inconsistent information [29]. The study showed that the hierarchical techniques show better results than the K means method in terms of connectivity [30]. The study is about effective methods for spatial data mining [31]. SLINK algorithm carries out single-link (nearest-neighbor) cluster analysis on an arbitrary dissimilarity coefficient and provides a representation of the resultant dendrogram which can readily be converted into the usual tree-diagram [32]. The new hybrid algorithm were compared to existing clustering algorithms on the basis of different measures and it was concluded that hybrid clustering algorithm was more accurate [33].

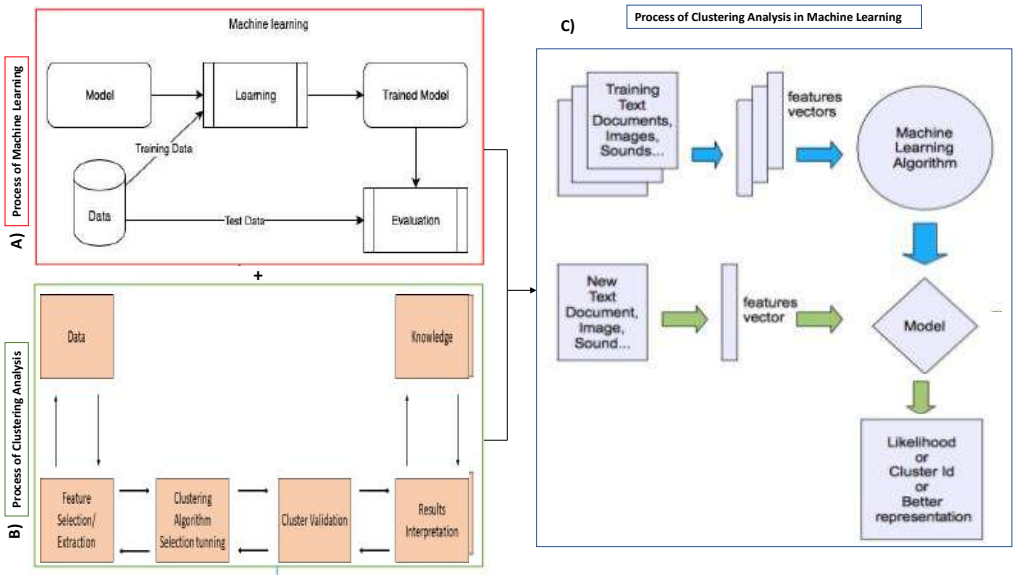
Hierarchical clustering method was studied and its performance was evaluated in [34]. Time series analysis of COVID-19 in Saudi Arabia was carried out [35]. The study introduced new concepts of density threshold and adaptive density using the p-DBSCAN algorithm [36]. In this study the method analyzes source-specific clustering and identifies a consensus set-partition which is as close as possible to all of them [37]. The study includes application of K-Means algorithm and K-prototypes algorithm with application on real world data and both the algorithms showed good results on their application [38].

The study includes application of different methods for modeling the uncertainty of NWP forecasts, various clustering algorithms are applied to group the performance records as the first step, second step was a range of methods are used to fit appropriate probability distributions to errors of each cluster, Results show that incorporating trained uncertainty model outputs into the NWP point predictions provides PI forecasts with higher reliability and skill. This can lead to improvement of decision processes for many applications that rely on these forecasts [39].

In this study different clustering techniques were used to cluster the weather data on the basis of the already provided training data in order to for clusters of different weather parameters on the basis of understanding the nature of different weather parameters from training set, further the best clustering method will be chosen on the basis of least time taking method for both training full set and building model. The clusters were later compared to the actual data of the temperature Maximum and Minimum and rainfall in order to see where the clusters formed were appropriate or not.

Data mining is a procedure of extracting information from existing set of data. Extraction of knowledge about the data in terms of pattern and nature is termed as data mining. Machine learning is divided into 3 main categories such as: Supervised Learning, Unsupervised Learning, and Reinforcement Learning. The basic steps of machine learning are shown in Figure 1 a). Unsupervised learning is a type of learning that helps in learning and understanding which helps us in understanding, discovering patterns in our data related to various features. Two of the main methods used in unsupervised learning are: Principle

Figure 1: Flow Chart Diagram of Analysis Technique in the study




component Analysis and Cluster Analysis. Clustering analysis is a part of data mining which helps in looking into the distribution of data and analyse the characteristics of every cluster formed.

A flowchart of the process of clustering analysis is shown in Figure 1 b).The process of clustering analysis in Machine Learning is as shown in Figure 3 c). The clustering techniques discussed in the present study are namely CANOPY, COBWEB, DBSCAN under the cluster analysis in order to analyze that whether these tools serve the purpose of clustering data in the correct cluster and how much time does each tool take to cluster the data values so that the study of the patterns can be done. The time taken by the tools is taken is considered in two ways namely Time taken to train full data(T1) and Time taken to make model(T2).

## 2. Methodology

### (i) CANOPY

This algorithm is unsupervised learning algorithm. Objective of the algorithm is to fasten up clustering operations on huge data set, but in terms of handling the large data sets a few may not work or may not provide desired results due to the size. The algorithm of canopy clustering works as follows, where X and Y are two thresholds and X is greater than Y:

- Extract the data set to be used.
- Exclude one value from the set, start new ‘canopy’ having the value.

- For every value remaining, mark it to a new 'canopy' further if distance to first value of canopy is lesser than loose distance X.
- Now, if value is at a distance which is relatively lesser than Y, exclude it from original set.
- Repeat second step till the time no more data values are left to cluster.
- The comparatively poorly clustered canopies can further be sub-clustered using expansive but more precise algorithm.

The canopy algorithm is useful as the number of training data which can be compared at every step is decreasing and also the resulting clusters are improved.

**(ii) COBWEB**

The cobweb clustering algorithm is a simple method of incremental conceptual learning. The algorithm forms a hierarchical group in the form of a classification tree. Every node is a concept and has a probabilistic depiction of the concept. At a given level the sibling nodes do form a partition and for mentioned as follows, the rise in the number of values which can be correctly guessed refers to category utility.

$$\frac{\sum_{k=1}^{n} P(C_k)\left[\sum_i \sum_j P(A_i = v_{ij} \mid C_k)^2 - \sum_i \sum_j P(A_i = v_{ij})^2\right]}{n}. \tag{1}$$

There are four operations that Cobweb uses while making tree. The application of the operations is based on category utility. The operations are as follows:

- Meeting 2 nodes: meeting of two nodes refers to substitute them with a node, sub node of which is the collection of original nodes' set of sub nodes and objects classified under them are summarizes.
- Ripping a node: subnodes substitute the node ripped.
- Adding a new node: When an object is inserted into the tree a node is created correspondingly.
- Shifting values in ladder: objective is to applying COBWEB algo on values and the subtree rooted in the node.

**(iii) DBSCAN**

DBSCAN algo is a clustering algo which is a technique in data mining and machine learning. On the basis of the set of points, DBSCAN algorithm groups values which are near to each other on the basis of measured distance that is EuclideanDistance formula & a least number of points. DBSCAN algorithm needs two parameters:

- Eps: eps determines the distance between the points for them to be considered as part of cluster. This means that the distance in between 2 values less than or equals to this value (eps), these values are taken as neighbours.
- Min points: least number of values to make a dense region. Eg: mark min points parameter to 15, further minimum 15 values are needed to form the dense region.

The estimation of parameter is important in data mining. The basic knowledge about the data is needed to choose the correct parameters. If the value of eps is very small, then a huge portion of the values will be grouped. If the number of values required to form dense region is not satisfied then the values will be removed. Taking too large values will also not be good as maximum values will come in one group and the study will get affected. Choose eps on the basis of distance of data values, but generally eps value is taken small. In general terms, least min points can be calculated from the value of dimensions (d) in the data set, as min points are greater or equals to d+1. More values are considered to be appropriate for data set which has noise and makes better clusters. The least value for the min points should be 3, but the larger the number of points in the data set, larger the min points value that must be chosen. The DBSCAN algorithm is applied to look for associations and structures in data which are hard to allocate manually but it can be easy and quite of use to see if there exists pattern and forecast trends.

## 3. Results and discussion

The daily weather data for different parameters is taken into consideration from 1st January, 2017 to 31st October, 2018 for Delhi region. The parameters taken are Maximum Temperature, Wind Speed, Minimum Temperature, Evaporation, Rainfall and Bright Sunshine whose time series are plotted in Figure 2 a)-f). The study includes clustering of data into clusters to study the nature and also to analyse which method serves as the better method amongst others based the time taken to form clusters. The time taken by the tools is taken is considered in two ways namely Time taken to train full data (T1) and Time taken to make model (T2).

The Table 1 shows the time taken to train the full data set and to build the model. It is seen very clearly that Canopy took lesser time to make clusters that will help in better and faster study of the clusters. Canopy in total formed twenty one clusters in the lesser time. Figure 2 g)-h) shows that Canopy clustering method takes lesser time in both situations and hence, can be chosen for making clusters. The Table 1 shows the time taken to train the full data set and to build the model. It is seen very clearly that Canopy took lesser time to make clusters that will help in better and faster study of the clusters. Canopy in total formed twenty one clusters in the lesser time.

Figure 2: a)-f) Weather Data Time Series Plot and g)-h) Clustering Technique Performance Comparison Plot

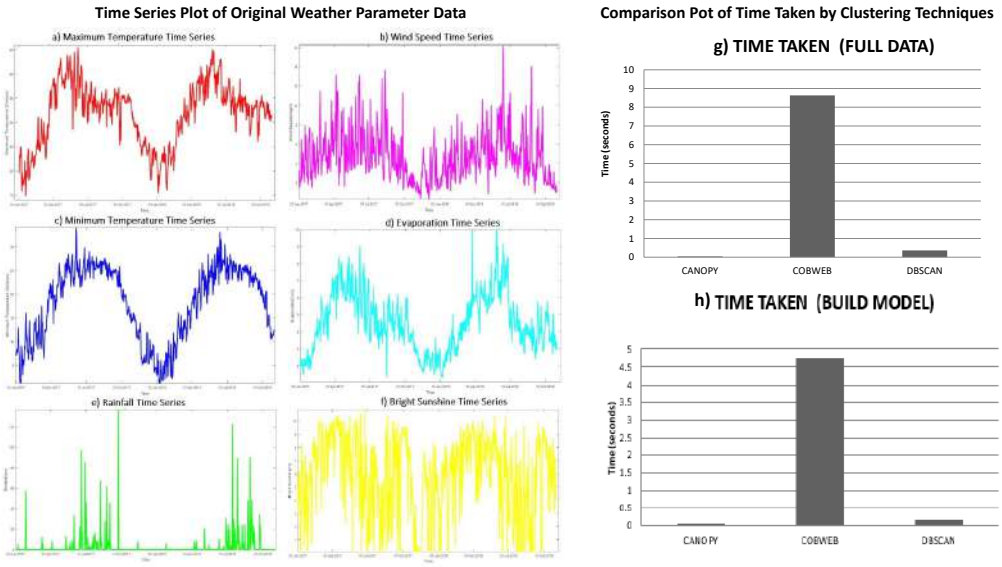



Table 1: Comparison of T1 and T2 for different clustering tools

| Clustering Tools | Time Taken (Full Data: T1) | Time Taken (Build Model: T2) |
|---|---|---|
| CANOPY | 0.04 | 0.01 |
| COBWEB | 8.63 | 4.76 |
| DBSCAN | 0.36 | 0.16 |

## 4. Conclusion

In the field of data analysis Clustering of data is an important problem. Clustering analysis is done in order to study the pattern and nature of the data. The objective of the study was to analyse different clustering methods and to see which among them is best suited for clustering of the weather time series data of different weather parameters. For the study the daily data of different weather parameters for Delhi region from $1^{st}$ January 2017 to $31^{st}$ October 2018 has been considered. Different clustering techniques such as CANOPY, COBWEB and DBSCAN were applied and compared for their efficiency to cluster data based on how much time was taken to build the model and process the full testing data to the assigned clusters. From the comparison of the performance efficiency for all the three methods it was observed that the least processing time was taken by CANOPY method. This result can be further analysed and understood in terms of the faster algorithm of CANOPY method and as the

resulting clusters are improved so better efficiency is observed on using it for clustering of the considered data. Clustering technique extracts knowledge from the data by grouping the data values of same pattern or nature in one group and hence provided insight about the data and hence helps in studying the nature of data. Thus selection of an effective and efficient clustering technique in terms of processing time is essential. It can be concluded from the analysis that CANOPY technique is faster and better in clustering the daily weather data in comparison to other methods compared.

**Acknowledgement**

Authors are thankful to Guru Gobind Singh Indraprastha University (GGSIPU), Delhi(India) for providing research facilities.

**References**

[1] A. P. Dempster, N. M. Laird, D. B. Rubin, *Maximum likelihood from incomplete data via the EM algorithm*, Journal of the Royal Statistical Society Series B (Methodological), 39 (1997), 1-38.

[2] A. Bangia, R. Bhardwaj, K.V. Jayakumar, *River water quality estimation through artificial intelligence conjuncted with wavelet decomposition*, Numerical Optimization in Engineering and Sciences 979, Springer Nature Singapore Pte Ltd., 2020, 107-123.

[3] R. Bhardwaj, A. Bangia, *Data driven estimation of novel COVID-19 transmission risks through hybrid soft-computing techniques*, Chaos, Soliton and Fractals, 140 (2020), 110152.

[4] R. Bhardwaj,A. Bangia, *Machine learned regression assessment of the HIV epidemiology development in Asian region*, Mathematical Modeling and Soft Computing in Epidemiology, Taylor & Francis, 2020, Chapter-4, 51-78.

[5] R. Bhardwaj, M. Chawla, *Convection dynamics of Nanofluids for temperature and magnetic field variations*, International Conference on Innovative Computing and Communication, Advances in Intelligent Systems and Computing, 1165 (2020), 271-289.

[6] R. Bhardwaj, S. Das, *Synchronization of two three-species food chain system with Beddington-DeAngelis functional response using active controllers based on the Lyapunov function*, Italian Journal of Pure and Applied Mathematics, 44 (2020), 57-77.

[7] R. Bhardwaj, D. Datta, *Consensus algorithm*, Decentralised Internet of Things, Studies in Big Data, 71 (2020), 91-107.

[8] R. Bhardwaj, D. Datta, *Development of a recommender system health mudra using blockchain for prevention of diabetes*, Recommender System with Machine Learning and Artificial Intelligence: Particle Tools and Applications in Medical and Agricultural Domains, Wiley & Sons, USA, 2020, Chapter 16, 313-327.

[9] R. Bhardwaj, D. Datta, *Development of epidemiological modeling RD-COVID-19 of Coronavirus infectious disease and its numerical simulation*, Mathematical Modelling and Analysis of Infectious Disease Problems (COVID-19), Springer, 2020.

[10] R. Bhardwaj, D. Datta, *Optimization techniques*, Revista INGLOMAYOR, Ingeniena Global Mayor 18 (A), INGLOMAYOR, 2020, 54-82.

[11] R. Bhardwaj, V. Duhoon, *Auto-regressive integrated moving-averages model for daily rainfall forecasting*, International Journal of Scientific and Technology Research, 9 (2020), 793-797.

[12] R. Bhardwaj, V. Duhoon,*Study and analysis of time series of weather data of classification and clustering techniques*,International Conference on Innovative Computing and Communication, Springer Singapore, 2020, 257-270.

[13] R. Bhardwaj, D. Pruthi, *Evolutionary techniques for optimizing air quality model*, Procedia Computer Science, 167 (2020), 1872-1879.

[14] R. Bhardwaj, D. Pruthi, *Evelopment of model for sustainable nitrogen dioxide prediction using neuronal networks*, International Journal of Environmental Science and Technology, 17 (2020), 2783-2792.

[15] R. Bhardwaj, M. Chawla, A. Bangia, S. Das, J. Goncerzewicz, *Effect of magnetic and temperature variation on Al2O3 nanofluid convection*, Mathematica Applicanda (Matematyka Stosowana), 48-1 (2020), 03-24.

[16] R. Bhardwaj, *Wavelets and fractal methods with environmental applications*, Mathematical Models, Methods and Applications, Springer Singapore, 2016, 173-195.

[17] R. Bhardwaj, *Nonlinear time series analysis of environment pollutants*, Mathematical Modeling on Real World Problems: Interdisciplinary Studies in Applied Mathematics, NOVA New York, 2019, 71-102.

[18] R. Bhardwaj, A. Bangia, J. Mishra, *Complexity analysis of pathogenesis of Coronavirus epidemiology spread in the China region*, Mathematical Modelling and Soft Computing in Epidemiology , Taylor & Francis, 2020, Chapter-13, 240-265.

[19] S. Bhardwaj, A. Khanna, D. Gupta, *Water quality evaluation using soft computing method*, Advances in Intelligent Systems and Computing, volume 1166, Springer, 2020, 1043-1052.

[20] G. Celeux, G. Govaert, *A classification EM algorithm for clustering and two stochastic versions*, Computational Statistics and Data Analysis, 14 (1992), 315–332.

[21] S. Chakraborty, N. Nagwani, L. Dey, *Weather forecasting using incremental K-means clustering*, CiiT International Journal of Data Mining and Knowledge Engineering, 4 (2012), 214-219.

[22] D. Datta, R. Bhardwaj, *Fuzziness-randomness modeling of plasma disruption in first wall of fusion reactor using type I fuzzy random set*, An Introduction to Fuzzy Sets, Nova New York, 2020, Chapter 5, 91-113.

[23] E.B. Fawlkes, C.L. Mallows, *A method for comparing two hierarchical clusterings*, Journal of the American Statistical Association, 78 (1983), 553–584.

[24] H.P. Kriegel, P. Kroger, J. Sander, A. Zimek, *Density-based clustering*, WIREs Data Mining and Knowledge Discovery, 1 (2011), 231–240.

[25] J. Gao, D. B. Hitchcock, *James-Stein shrinkage to improve K-means cluster analysis*, Computational Statistics & Data Analysis, 54 (2010), 2113-2127.

[26] M. Meila, D. Heckerman, *An experimental comparison of several clustering and intialization methods*, UAI'98: Proceedings of the Fourteenth conference on Uncertainty in Artificial Intelligence, 1998, 386–395.

[27] A. Ram, S. Jalal, A.S. Jalal, M. Kumar, *A density based algorithm for discovering density varied clusters in large*, International Journal of Computer Applications, 3-6 (2010), 1:4.

[28] M. Ankerst, M. M. Breunig, H.P. Kriegel, J. Sander, *OPTICS: ordering points to identify the clustering structure*, ACM SIGMOD Record, 28 (1999), 49–60.

[29] N. Ailon, M. Charikar, A. Newman, *Aggregating inconsistent information: ranking and clustering*, Journal of ACM, 55 (2005), 23:1-23:27.

[30] N. Shobha, T. Asha, *Monitoring weather based meteorological data: clustering approach for analysis*, 2017 International Conference on Innovative Mechanisms for Industry Applications (ICIMIA), 2017, 75-81.

[31] R. T. Ng, J. Han, *Efficient and effective clustering method for spatial data mining*, VLDB '94: Proceedings of the $20^{th}$ International Conference on Very Large Data Bases, 1994, 144-155.

[32] R. Sibson, *SLINK: an optimally efficient algorithm for the single-link cluster method*, The Computer Journal, 16 (1973), 30–34.

[33] S. Kumar, M. Singh, *A novel clustering technique for efficient clustering of big data in Hadoop ecosystem*, Big Data Mining and Analytics, 2 (2019), 240-247.

[34] S. Dasgupta, *Performance guarantees for hierarchical clustering*, COLT'02: Proceedings of the 15th Annual Conference on Computational Learning Theory, 2002, 351-363.

[35] S.K. Sharma, S. Bhardwaj, R. Bhardwaj, M. Alowaidi, *Nonlinear time series analysis of pathogenesis of COVID-19 pandemic spread in Saudi Arabia*, Computers, Materials and Continua, 66 (2021), 805-825

[36] S. Kisilevich, F. Mansmann, D. Keim, *P-DBSCAN: a density based clustering algorithm for exploration and analysis of attractive areas using collections of geo-tagged photos*, COM.Geo '10: Proceedings of the 1st International Conference and Exhibition on Computing for Geospatial Research & Application, 38 (2010), 1-4.

[37] V. Filkov, S. Kiena, *Integrating microarray data by consensus clustering*, International Journal on Artificial Intelligence Tools, 13 (2004), 863–880.

[38] Z. Huang, *Extensions to the k-means algorithm for clustering large data sets with categorical values*, Data Mining and Knowledge Discovery, 2 (1998), 283–304.

[39] A. Zarnani, P. Musilek, J. Heckenbergerova, *Clustering numerical weather forecasts to obtain statistical prediction intervals*, Climate Resilience and Sustainibility, Royal Meterological Society, 21 (2013), 605-618.

Accepted: November 23, 2020