D. D. Sánchez-Gallegos, J. L. Gonzalez-Compean, S. Alvarado-Barrientos, V. J. Sosa-Sosa, J. Tuxpan-Vargas and J. Carretero, "A containerized service for clustering and categorization of weather records in the cloud," *2018 8th International Conference on Computer Science and Information Technology (CSIT)*, 2018, pp. 26-31.

DOI: 10.1109/CSIT.2018.8486198

A containerized service for clustering and categorization of weather records in the cloud

Dante D. Sánchez-Gallegos
Cinvestav Tamaulipas
Victoria, Mexico
Email: dsanchez@tamps.cinvestav.mx

J. L. Gonzalez-Compean
Cinvestav Tamaulipas
Victoria, Mexico
Email: jgonzalez@tamps.cinvestav.mx

Susana Alvarado-Barrientos
INECOL Ecological Network
Veracruz, Mexico
Email: susana.alvarado@gmail.com

Victor J. Sosa-Sosa
Cinvestav Tamaulipas
Victoria, Mexico
Email: vsosa@tamps.cinvestav.mx

José Tuxpan-Vargas
IPICyT Research Center
SLP, Mexico
Email: jose.tuxpan@ipicyt.edu.mx

Jesus Carretero
Arcos-UC3M
Madrid, Spain
Email: jesus.carretero@uc3m.es

***Abstract*—This paper presents a containerized service for clustering and categorization of weather records in the cloud. This service considers a scheme of microservices and containers for organizations and end-users to manage/process weather records from the acquisition, passing through the prepossessing and processing stages, to the exhibition of results. In this service, a specialized crawler acquires records that are delivered to a microservice of distributed categorization of weather records, which performs clustering of acquired data (the temperature and precipitation) by spatiotemporal parameters. The clusters found are exhibited in a map by a geoportal where statistic microservice also produce results regression graphs on-the-fly. To evaluate the feasibility of this service, a case study based on 33 years of daily records captured by the Mexican weather station network (EMAS-CONAGUA) has been conducted. Lessons learned in this study about the performance of record acquisition, clustering processing, and mapping exhibition are described in this paper. Examples of utilization of this service revealed that end-users can analyze weather parameters in an efficient, flexible and automatic manner.**

***Index Terms*—Clustering, Weather Station records, Workflows, Cloud, microservices, virtual containers**

I. Introduction

The online available weather records and satellite products databases are key for the scientific community because this information represents spatiotemporal snapshots of the earth.

Each provider of databases such as central administrators of weather station networks[1] make the databases available for end-users to get this information and perform studies about climate and environment in an offline manner.

However, end-users that are not familiar and skilled with each platform designed by different agencies spend considerable amounts of initial time to access the data for simple tasks such as visual data exploration and processes this information to obtain useful information from it. Moreover, ironically, the enormous amount of data that is nowadays freely available for downloads such as weather records and satellite data products has not necessarily propelled a proportional widespread use of these data outside of the traditional expert circles.

In this paper is presented a service for clustering and categorization of weather records using a scheme of microservices and containers in the cloud. With this schema, organizations and end-users are able to manage/process weather records from the acquisition to the exhibition of results, including preprocessing and processing stages.

In this service, a specialized crawler acquires records that are delivered to a microservice of distributed categorization of weather records, which performs clustering of acquired data (the temperature and precipitation) by spatiotemporal parameters. The clusters found are exhibited in a map by a geoportal where statistic microservice also produce results regression graphs on-the-fly.

The integration of large weather records within a geographical information system also requires a specific set of skills plus additional time and effort. These limitations create an important roadblock to the creation of hypotheses and research questions about complex datasets such as large time series of weather records. We believe that this roadblock could be largely minimized by the creation of targeted data mining services such the one presented in this work.

A. Applicability of automatic clustering service to real-world scenarios

The main application of the data mining service presented here is the exploration of 33 years (from 1985 to today) worth of daily ground-based observations of temperature and rainfall data collected at 754 different sites instrumented with weather stations across Mexico. As such, the service is a valuable aid to

This work was partially supported by the sectoral fund of research, technological development and innovation in space activities of the Mexican National Council of Science and Technology (CONACYT) and the Mexican Space Agency (AEM), project No.262891.

[1]e.g. at national governmental levels like the NCDC for the USA: https://www.ncdc.noaa.gov/cdo-web/datasets, and CONAGUA for Mexico: https://smn.cna.gob.mx/es/emas

researchers, students, and the even interested public, allowing a friendly and direct way to perform initial data exploration of climate change and weather variability, with the potential to spark not examined research questions and hypothesis to be further explored using more in-depth analysis. By allowing the user to set a specific geographical area of interest to perform the exploration, even if only a few 100 meters, unraveling temporal patterns is possible without major effort. These localized temporal patterns might have been occult in studies of much larger spatial scales (e.g. Vose et al. [1], Alexander et al. [2], Senior et al. [3]). Importantly, questions and hypothesis regarding climate change and variability at local and regional scales are still largely unexplored for neotropical environments such as those located in a large proportion of the Mexican Republic [3]–[5]. Therefore, we believe that a service such as presented here is a step forward to put large databases of weather records available to all levels of scientific inquiry.

To evaluate the feasibility of this service, a case study based on 33 years of daily records captured by the Mexican weather station network (EMAS-CONAGUA) has been conducted. Lessons learned in this study about the performance of record acquisition, clustering processing, and mapping exhibition are described in this paper.

Examples of utilization of this service revealed end-users can analyze weather parameters in an efficient, flexible and automatic manner.

II. A CONTAINERIZED SERVICE FOR CLUSTERING AND CATEGORIZATION OF WEATHER RECORDS

In this section, we present the design principles of a containerized service for clustering and categorization of weather records.

The design principles are described in a joint manner with implementation details of this service in a real-world scenario to show the feasibility of this service. Specifically, for this study was considered the application of our service to the processing of online available databases of a weather stations network deployed by Mexican National Weather Commission (CONAGUA) on all the Mexican territory (Figure 1) [6]. This network includes 754 Automatic Weather Stations (EMAS) that, from 1985 to today, daily register the maximum and minimum temperatures, and also the amount of precipitation on a day measured in millimeters.

A. Design principles

The design of this service is based in the construction of patterns to achieve the interconnection applications from the acquisition of data, passing through the preprocessing and processing data to the exhibition of results.

The basic idea is to encapsulate the applications considered in this service into virtual containers and then chaining these containers through input/output interfaces to create processing patterns [7] (e.g. workflows, pipe and filters/pipelines, manager/workers, Master/slave etc). The containers of each stage/phase of these patterns are integrated into black boxes

Fig. 1. Locations of the CONAGUA's weather stations [6]

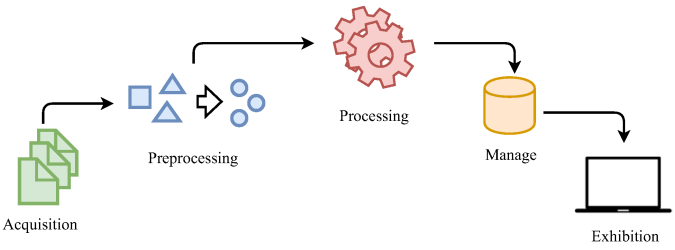

Fig. 2. Workflow created in the containerized service

The design principles are described in a joint manner with implementation details of this service in a real-world scenario to show the feasibility of this service. Specifically, for this study was considered the application of our service to the processing of online available databases of a weather stations network deployed by Mexican National Weather Commission (CONAGUA) on all the Mexican territory (Figure 1) [6]. This network includes 754 Automatic Weather Stations (EMAS) that, from 1985 to today, daily register the maximum and minimum temperatures, and also the amount of precipitation on a day measured in millimeters.

A. Design principles

The design of this service is based in the construction of patterns to achieve the interconnection applications from the acquisition of data, passing through the preprocessing and processing data to the exhibition of results.

The basic idea is to encapsulate the applications considered in this service into virtual containers and then chaining these containers through input/output interfaces to create processing patterns [7] (e.g. workflows, pipe and filters/pipelines, manager/workers, Master/slave etc). The containers of each stage/phase of these patterns are integrated into black boxes built as microservices that are interconnected through REST APIs.

These patterns enable end-users to create workflows for processing data from the data sources (e.g. weather record databases) passing through a set of preprocessing and processing stages until the user can visualize the results of such a processing procedure.

Basically, the first approach of this service was developed following the stages depicted in Figure 2, where the records are acquired and preprocessed to obtain useful data for a processing module, which performs a distributed clustering [8] with these data. The results of such a processing procedure are exposed as a service by a REST API that is automatically consumed by a geoportal (client) [9].

This design enables the end-users to create workflows and analyze clustering results on-demand and in an automatic manner; as a result, different results can be obtained depending on spatiotemporal parameters [2] chosen by the end-users through a geoportal.

B. Service implementation details

The processing workflow described in Figure 2 is divided into two main stages, the first one performed offline (acquisition and preprocessing) and another one performed online

[2]By Spatiotemporal we refer to coordinates of a coverage land (spatial) in a given period of time (temporal)

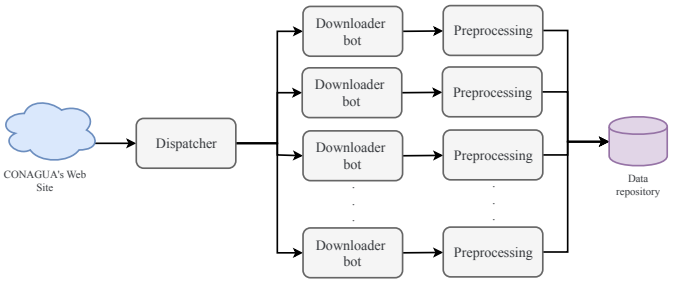

Fig. 3. Acquisition and Preprocessing pattern

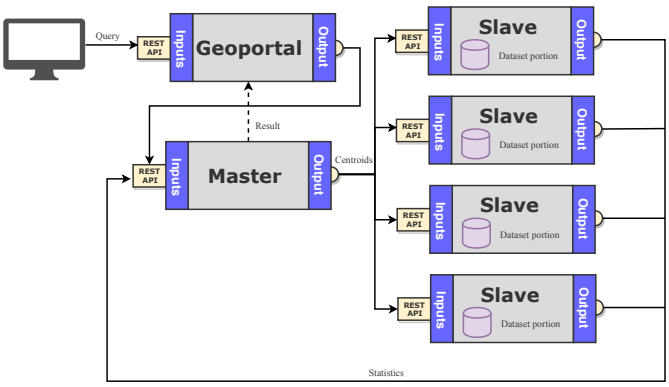

Fig. 4. Distributed clustering service

(cluster processing) that is launched when the user sends a query, by using spatiotemporal parameters, to the geoportal.

1) Acquisition and preprocessing weather records: In the first stage, we constructed a manager/worker (dispatcher/node) pattern, which considers a virtual container called $Dispatcher$ that includes a configurable specialized crawler that automatically obtains all the available data of each station. This container receives URLs as input parameters (that are considered as $tasks$) and distributes t tasks to t containers called $workers$ placed at data preprocessing stage. A worker includes prepossessing applications that perform the downloading of data to later perform transformations, data cleaning, and data integration. These tasks are performed in parallel as workers are launched by the dispatchers at the same time. Figure 3 depicts the pattern that the data follow to be preprocessed in this stage.

For our case study, the crawler, placed in the dispatcher, collects the EMAS metadata from the CONAGUA site and retrieves the URLs associated with each weather station. Each URL delivers a file including the daily records for a period of 33 years captured by a given station, which contains the data captured about temperature and precipitation of each day observed by the station. When the dispatcher has collected the metadata of all the stations, it launches n workers to retrieve the weather station files. The dispatcher distributes the URLs to the workers in a load-balancing way using a two choices technique [10]. Each worker receives a list of sub-set of URLs to download files from the source data. Each time a worker downloads a file, it is immediately sent to a preprocessing container, where the data are transformed into a unified and coherent format and the missed values on the data are filled to match with the unified format by using linear regression. The files are placed in a data repository, that is accessed by the containers in charge of executing the preprocessing to finish each task in the list sent by the dispatcher.

2) Clustering and categorization of weather records: The data repository produced by the preprocessing patterns is used as an input parameter by the containers in the processing stage. At this stage, a distributed clustering service was developed in containers organized in the form of a Master/Slave pattern (shown in Figure 4). This pattern is online executed for processing spatiotemporal requests each time the end-user invokes this option available in the geoportal.

The distributed clustering is an implementation of K-Means algorithm, how is described by Cesario and Thalia [8]. In this pattern, the $Master$ node is in charge of centroids initialization and centroids recalculation, while the $slaves$ nodes are in charge of to group each station records with records of other similar stations by using the centroids that were proportioned by the master node. Each slave node, sends the next statistics to the master node: intra-cluster distance, inter-cluster distance, and the number of elements by each cluster. These statistics are used by the master node to recalculate the centroids. This process is ended when the algorithm reaches a maximum number of iterations or the algorithm converges.

This service also includes a R service API for the calculation of the statistic regression of the clusters identified for all the station records of the coverage land selected by end-users. The results produced by this service are consumed by the geoportal and regression graphs of each cluster of stations are also showed and the end-user can view/download these graphs.

For implementation details of our case study, the components of distributed clustering (Master and slave) were encapsulated into virtual containers by using Docker containers and exposed by using a REST API. In this way, the geoportal can consume this service. In the Geoportal, the users can perform online queries drawing a polygon over an area on Mexico (spatial parameter) and obtain the clustering results over the selected area for all period of time considered in this study (33 years is a default temporal parameter but it could be changed by date ranges selected by end-users). Figure 5 shows an example of how a user can draw a polygon and obtain the clustering results for that specific area, which in this case is the Yucatán Peninsula. Please note that the end-users can choose a polygon covering all Mexican territory if required as all the records of the stations of Mexican network were acquired during the acquisition and preprocessing stages.

III. Experimental evaluation methodology

In this section, we describe the methodology used to conduct a proof of concept and an experimental evaluation based on a case study about an automatic acquisition and clustering of the station's weather records of the Mexican weather station network.

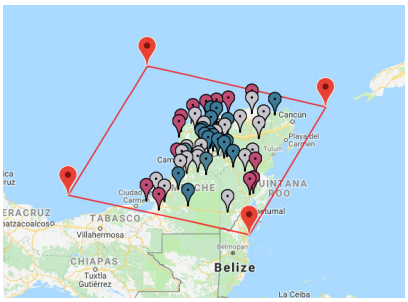

Fig. 5. Example of the clustering service by using a geoportal

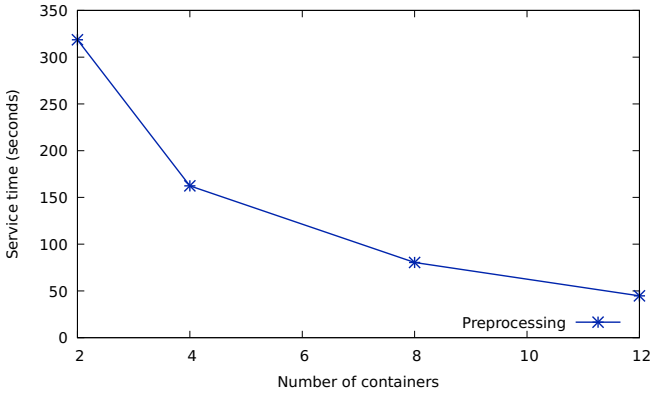

Fig. 6. Service time of the preprocessing schema using different number of container running in parallel

A. Case study metrics

The metrics chosen to evaluate the performance of the prototype were the *service time*, which is the time required by a given service to complete a given task. In order to evaluate this service in a qualitative manner, an example of an application of the service also is described through a simple task of result visualization of the Yucatan peninsula.

IV. Results

In a first evaluation, we measure the service time of the preprocessing schema using a different number of preprocessing containers running in parallel. Figure 6 shows, in the vertical axis, the number of containers running on parallel, whereas, in the vertical axis is shown the service time spent by the solution when prepossessing all the files downloaded. Figure 6 shows that duplicating the number of containers produces a speedup of 2x.

Figure 7 shows a comparison of the service time for the distributed clustering using one and 12 containers (vertical axis) for a different number of station groups (horizontal axis). As it can be seen, when using only one container, the service time tends to increase significantly when more stations are considered in a request. This is an expected scenario to be performed by end-users (using a crawler and clustering deployed on a single computer). In turn, with more containers running in parallel, the service time observed by end-users is

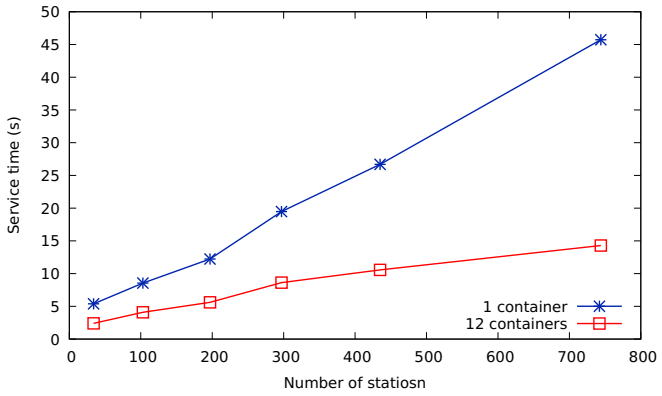

Fig. 7. Service time of the distributed clustering service for different number of stations

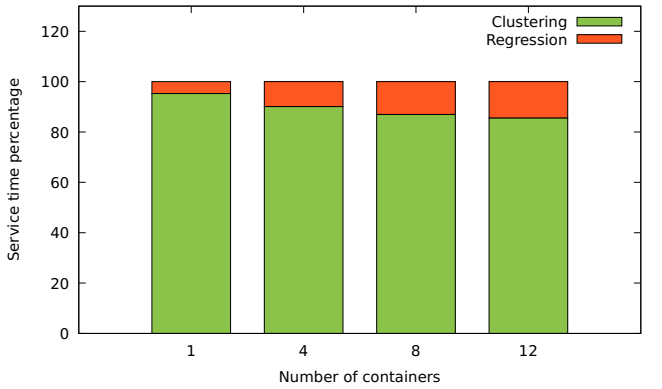

Fig. 8. Percentage of the service time spent by each processing task for different number of containers

reduced significantly making manageable the increment in the amount of load (station records).

Figure 6 and Figure 7 shows both a performance improvement when increasing the number of containers in the patterns without increasing the number of computational resources. This is quite important for real-world scenarios as the idea is to attend requests sent by end-users in on-the-fly and on-demand manner as this service is dynamic not a static one as traditionally performed in the past [11], [12].

Figure 8 shows the percentage of service time that each processing task consumes. The two task contemplates for this graph is the distributed clustering and regressions (linear and polynomial) generated in the master node. While the number of Slaves running in parallel decreases the percentage of the clustering task increments, the percentage of the regression decrements is constant as it is performed by one single container, which reveals the impact of the parallelism patterns on the service performance.

A. Analyzing the application of the service for a real case spatial selection

As an example of an application of the service, we show in Figure 9 the exploration performed by a researcher interested

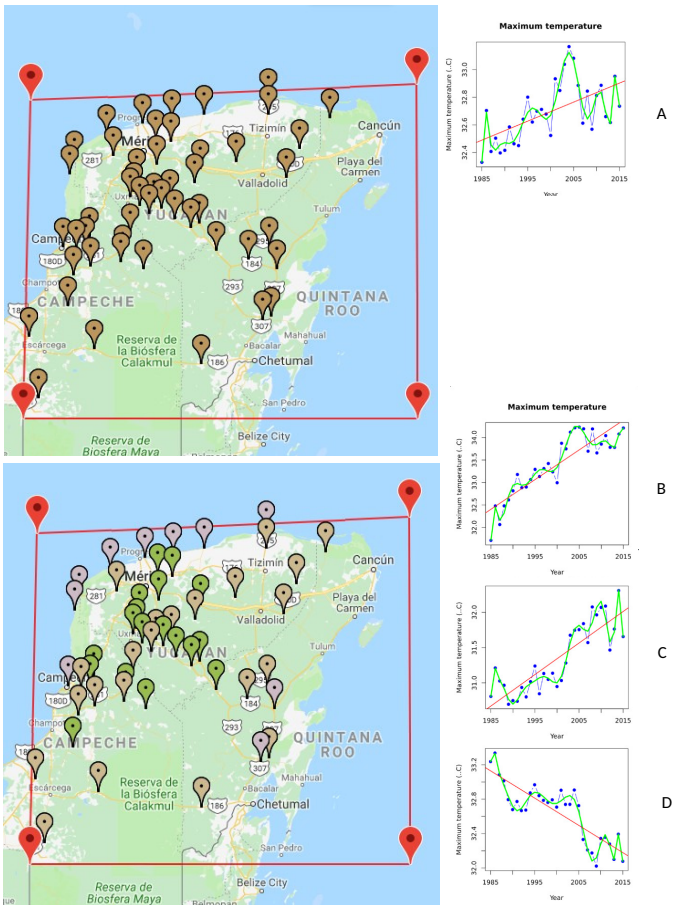

Fig. 9. Temporal patterns of maximum temperature for different weather station groupings: A) one cluster; B-D) three clusters

in the temporal patterns of maximum temperature for the Yucatan Peninsula in southeastern Mexico. When all weather stations are selected (i.e. one cluster), the emergent pattern is a modest increase of less than 1C across the 30-year temperature record (Figure 9a). On the other hand, when three clusters are selected, contrasting patterns emerge: a much stronger increase of about 2C across the 30 years for cluster 0 (Figure 9b), a comparatively slower increase for cluster 1 (Figure 9c) and, surprisingly, a decreasing trend of about 1C for cluster 2 (Figure 9d; note that the y-axis scale is different in all cases). From this initial data exploration enabled by the service, the user now may elaborate working research questions such as: What are the causes of such distinct patterns? Is one factor the proximity to the coast? How important is land cover to explain these patterns?

We consider that a service such as presented here is a step forward to put large databases of weather records available to all levels of scientific inquiry and could represent a useful tool for supporting scientific projects about climate studies.

V. Related work

Data weather analysis to find trends and changes in the temperature through time windows have been proposed in the literature. For example, Ghasemi [12] performs a cluster analysis over the Iran territory to group the weather stations in groups with similar weather conditions, and find the changes in temperature for each cluster through regression analysis. In a similar context, Bravo et. al. [11] performed a clustering analysis over the Mexican territory, whereas Calmanti et. al. proposed a similar analysis over the Italian territory by constructing a model to group the weather stations and by analyzing each cluster generated [13]. However, these studies were created by a specific purpose and for an offline analysis. We consider that this type of work can be improved computationally by the development of an automatic mechanism that performs the clustering algorithm, which receiving spatiotemporal parameters, which is the service proposed in this paper. This could enable experts to process coverage land selections in on-the-fly and on-demand manner and to achieve/interpret results for these specific cover lands. It is important to note, that the service proposed in this paper can be used by two types of scientific end-users to create distributed clustering workflows in a configurable manner and for end-users interested in using visualization tools for the making decisions about a given cover land. The main difference of this service from traditional proposals is the flexibility to create different types of workflows and the dynamic utilization of the data by spatiotemporal parameters, which are not considered in the traditional proposals. Currently, we are working on the building of processing workflows by using different data sources for a weather multidisciplinary project about Mexican territory.

VI. Conclusions

This paper presented a containerized service for clustering and categorization of weather records in the cloud. This service considers a scheme of microservices and containers for organizations and users to manage weather records from the acquisition passing through the processing to the exhibition of results in an efficient, flexible and automatic manner. A specialized crawler obtains the records which are delivered to a microservice, which perfoms a distributed clustering and categorization of weather records, grouping the temperature (Max/Min/Average) and precipitation parameters by using spatiotemporal parameters. The results are exhibited in a geoportal where Statistic microservices also produce results graphs on-the-fly. A case study based on 33 years of daily records captured by the Mexican weather station network (EMAS-CONAGUA) was conducted and the lessons learned about the performance of record acquisition, clustering processing, and mapping exhibition were described in this paper. An example of an application of the service described showed the feasibility of this service for end-users and scientific community.

References

[1] Russell S Vose, David R Easterling, and Byron Gleason. Maximum and minimum temperature trends for the globe: An update through 2004. *Geophysical Research Letters*, 32(23), 2005.

[2] LV Alexander, X Zhang, TC Peterson, J Caesar, B Gleason, AMG Klein Tank, M Haylock, D Collins, B Trewin, F Rahimzadeh, et al. Global observed changes in daily climate extremes of temperature and precipitation. *Journal of Geophysical Research: Atmospheres*, 111(D5), 2006.

[3] Rebecca A Senior, Jane K Hill, Pamela González del Pliego, Laurel K Goode, and David P Edwards. A pantropical analysis of the impacts of forest degradation and conversion on local temperature. *Ecology and evolution*, 7(19):7897–7908, 2017.

[4] Edouard L Davin and Nathalie de Noblet-Ducoudré. Climatic impact of global-scale deforestation: Radiative versus nonradiative processes. *Journal of Climate*, 23(1):97–112, 2010.

[5] Roger A Pielke, Andy Pitman, Dev Niyogi, Rezaul Mahmood, Clive McAlpine, Faisal Hossain, Kees Klein Goldewijk, Udaysankar Nair, Richard Betts, Souleymane Fall, et al. Land use/land cover changes and climate: modeling analysis and observational evidence. *Wiley Interdisciplinary Reviews: Climate Change*, 2(6):828–850, 2011.

[6] CONAGUA. Estaciones meteorolgicas automticas (emas), 2018. [Online; accessed May 03, 2018].

[7] J.L. Gonzalez-Compean, Victor Sosa-Sosa, Arturo Diaz-Perez, Jesus Carretero, and Jedidiah Yanez-Sierra. Sacbe: A building block approach for constructing efficient and flexible end-to-end cloud storage. *Journal of Systems and Software*, 135:143 – 156, 2018.

[8] Eugenio Cesario and Domenico Talia. Distributed data mining patterns and services: an architecture and experiments. *Concurrency and Computation: Practice and Experience*, 24(15):1751–1774, 2012.

[9] Dante D. Sánchez-Gallegos, J. L. Gonzalez-Compean, Victor J. Sosa-Sosa, Heidy M. Marin-Castro, and José Tuxpan-Vargas. An interoperable cloud-based geoportal for discovery and management of earth observation products. *Computer Science & Information Technology (CS & IT)*, 2018.

[10] M. Mitzenmacher. The power of two choices in randomized load balancing. *IEEE Transactions on Parallel and Distributed Systems*, 12(10):1094–1104, Oct 2001.

[11] JL Bravo Cabrera, E Azpra Romero, V Zarraluqui Such, C Gay García, and F Estrada Porrúa. Cluster analysis for validated climatology stations using precipitation in mexico. *Atmósfera*, 25(4):339–354, 2012.

[12] Ahmad Reza Ghasemi. Changes and trends in maximum, minimum and mean temperature series in iran. *Atmospheric Science Letters*, 16(3):366–372, 2015.

[13] S Calmanti, A DellAquila, F Maimone, and V Pelino. Evaluation of climate patterns in a regional climate model over italy using long-term records from synop weather stations and cluster analysis. *Climate Research*, 62(3):173–188, 2015.