Data Processing Toolset for the Virtual Watershed

Lisa Palathingal, Rui Wu, Sergiu Dascalu, Frederick C. Harris, Jr.
Computer Science and Engineering
University of Nevada
Reno, NV, USA
{lisa, rui, dascalus, fredh}@cse.unr.edu

Abstract— **In order to understand the hydrologic changes in the watersheds due to climate changes, the EPSCoR jurisdictions of Idaho, Nevada, and New Mexico have collaborated to create the Western Consortium for Watershed Analysis, Visualization and Exploration (WC-WAVE). WC-WAVE will create a Virtual Watershed Platform (VWP) framework for assisting watershed scientists in their research. This software environment will provide a platform to share different data with fellow scientists and run different models in the cloud through web services. The VWP framework capabilities will enable the researchers to acquire and integrate data, visualize the data to obtain results, and identify environmental consequences of hydrologic changes.**

Data Processing Toolset for the Virtual Watershed (DPT-VW) is a user-centric web application that includes a set of new data processing tools for the WC-WAVE project. The application includes tools that convert input data required for running the models as well as output data obtained after running the models from text format into NetCDF format for visualization. The website also includes tools that download datasets of watersheds available from GSTORE and visualize the files in CSV format on 2D graphs. A detailed design and the current implementation of DPT-VW will be discussed in this paper.

Keywords-conversion; download; visualization; virtual watershed; data processing; model

I. INTRODUCTION

A watershed is the area of land where water from precipitation or snow melt flows down into the same place, providing a large proportion of water and ecosystem services to communities in the intermountain west [1]. Climate change is an important factor that greatly affects the watersheds. It may cause impacts on interactions between precipitation, vegetation growth, soil moisture, and other landscape properties. To understand the hydrologic changes in the watersheds, the Experimental Program to Stimulate Competitive Research (EPSCoR) jurisdictions of Idaho, Nevada, and New Mexico have established the Western Consortium for Watershed Analysis, Visualization and Exploration (WC-WAVE). The goal of WC-WAVE is to create a Virtual Watershed (VW) framework for developing and integrating watershed models, thereby allowing the researchers to collect and integrate the data, visualize the data to obtain the results, and understand the environmental consequences of hydrologic changes [2].

The sensors provide information about the surrounding environment of the watershed and allow the researchers to run the models for analyzing the data captured from these sensors. Researchers use this acquired data to develop models that effectively reflect the environment of the watershed. These models can then be used to predict water runoff, precipitation, snow melt, underground water flow, and other environmental factors within a watershed [3].

The data collected by the sensors have to be stored in a suitable format in the VW. The users can then extract the input data, run the model, generate the output data and visualize the results. NetCDF (Network Common Data Form) is considered as a self-describing data format which contains its own metadata. It is a group of software libraries and self-describing, machine-independent data formats which support the creation, access, and sharing of array-oriented scientific data [4].

Developing software tools and frameworks for scientific research can be interesting for several reasons. With the evolution of computing power in recent decades, scientific research is generating more and more data and models independently built by scientific researchers. This is an exciting field where software engineering can facilitate the creation of distributed software systems and frameworks to assist scientists have collaboration on these data and models in a global scale. Being an interdisciplinary field, this field poses various challenges in terms of barriers of communication and team building among different communities involved in the process [5]. Our research aims to help environmental scientists to interact with their model data quickly and effectively.

The rest of the paper is structured as follows. Section II describes the background of WC-WAVE and the Precipitation Runoff Modeling System (PRMS) model to be run in the VW. Section III presents an overview of the proposed DPT-VW web application. Section IV discusses the software system and the results along with the use cases and the visualization graphs. Section V contains a comparison with the related work. Finally, Section VI concludes the paper by summarizing the contributions of the paper along with the future work related to DPT-VW.

II. BACKGROUND

The WC-WAVE is a tri-state consortium composed of researchers and students in Idaho, Nevada, and New Mexico. Its overall aim is to study the impact of climate change on high-mountain watersheds. The WC-WAVE project is composed of three components: Watershed Science,
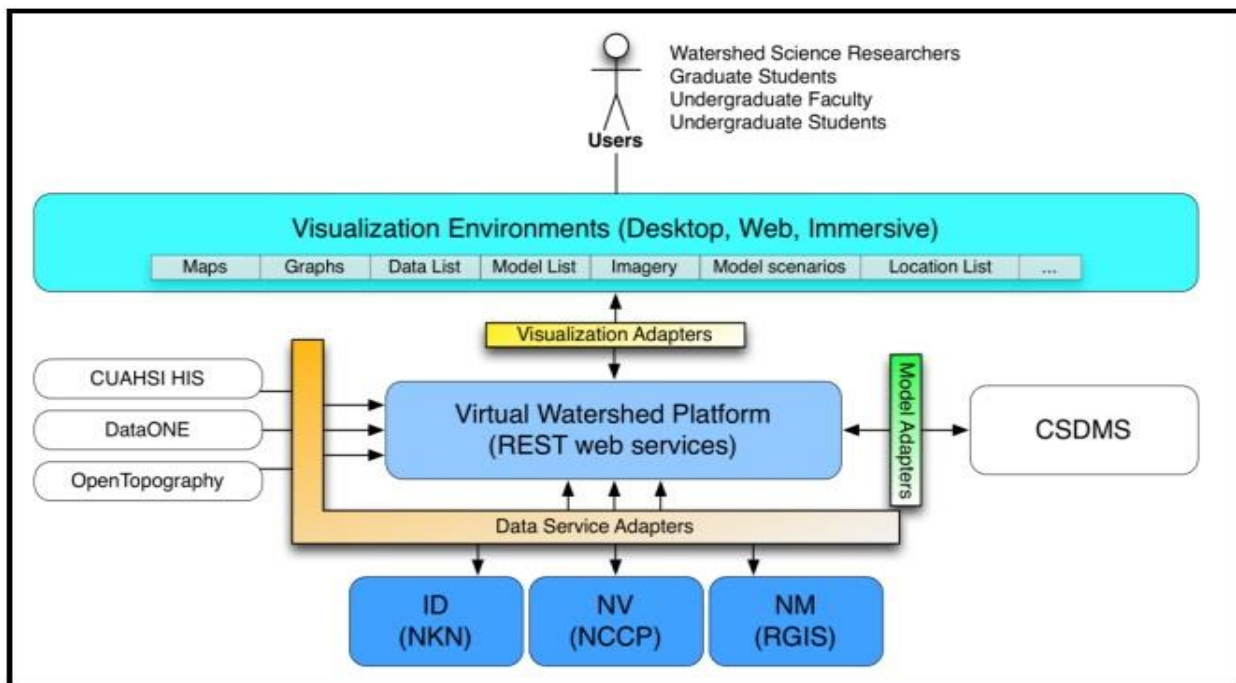
Figure 1: Design of the Virtual Watershed Platform [3]

Visualization and Data Cyberinfrastructure (CI), and Workforce Development. All the participants in these components are collaborating in order to better understand the interactions between precipitation, snow- pack, groundwater flow, and other properties within watersheds. For enhancing the collaborative nature of this project and future work, as well as for advancing data exploration and analysis, the VWP framework will be created which would provide data access and visualization through web-based environments. The VWP will also allow a user to run specific hydrologic models and visualize the results [3]. Fig. 1 illustrates the design of VWP.

To implement and test the VWP, the project participants selected four watersheds throughout the tri-state region to focus on: Dry Creek; Reynolds Creek Experimental Watershed in Idaho; Jemez watershed in New Mexico; and Lehman Creek in Nevada. Members of the Watershed Science component of the WCWAVE project selected four main models to include in the VWP [3]. Among them, one of the models is the PRMS model. PRMS model is developed to simulate and evaluate the watershed response of different combinations of climate and land use [6]. The watershed area is divided into Hydrologic Response Units (HRUs). This division depends upon the hydrologic and physical characteristics such as drainage boundaries, land surface altitude, slope, and aspect; plant type and cover; land use; distribution of precipitation, temperature, and solar radiation. Each HRU is identified by a numerical index. The HRU index starts with 1 [6]. Fig. 2 shows the spatial distribution of HRUs in PRMS output and input files.

The input files required for the PRMS simulation are control file, data file, and parameter file. All the files are in text format. The control file controls the overall simulation of the PRMS model. It specifies the input and output file names, the file path, the simulation starting and ending dates, and the active modules. The data file includes the measured time-series data used in a PRMS simulation. The values of the one-dimensional data vary with time. The time-series data may include the daily precipitation, maximum and minimum air temperatures, solar radiation, pan evaporation, measured streamflow, humidity, wind speed, and snowpack-water equivalent. The parameter file contains the parameters that control each of the hydrologic processes. It includes the values of parameters specified for each module that do not change during a simulation. In the parameter file, the values of the one-dimensional data vary with space [6].

Some parameters are space related. As shown in Figure 1, there are 4704 grid cells where each cell represents one location of the watershed, and there is one value for each cell. There will be 4704 values for the space related parameters in total. These parameters do not have time features. Some parameter are time related. There will be 12 values for the time related parameters in total where each value represents a value for each month. These parameters do not have space features. Some parameters are both space and time related. There will be 12 x 4704 values for the space and time related parameters in total. These parameters have both space and time features.

The output files that are generated after running the PRMS model are the Statistic Variables file, the Animation Variables file, and the Water Budget file. The Statistic Variables file (aka statvar file) includes one dimensional data that vary with time. The Animation file generates the two dimensional temporal and spatial data (that is the data changes with time and space). The Water Budget file generates a summary table of the water budget for a PRMS simulation [6].
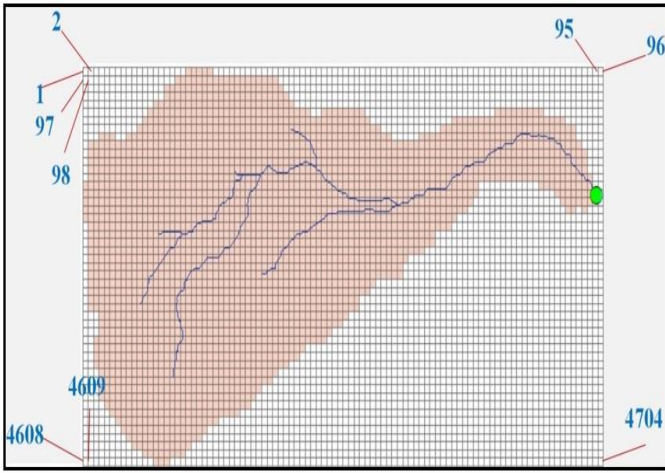
Figure 2. Spatial Distribution of HRUs

III. PROPOSED SYSTEM

Data Processing Toolset for the Virtual Watershed (DPT-VW) is a web application incorporating tools that aids the environmental scientists, students and others in the processing of the data. DPT-VW has four sections: information, tool selection, tool request, and feedback. The system-level diagram is represented in Fig. 3.

The information section presents an overview of DPT-VW with a video describing the features of the website. The tools provided by DPT-VW are useful for: i) converting PRMS input and output files from text into NetCDF format for visualization, ii) downloading the datasets of the watersheds available from the Geographic Storage, Transformation and Retrieval Engine (GSToRE) framework [7] in Georeferenced Tagged Image File Format (GeoTIFF), and iii) visualizing the files in Comma Separated Values (CSV) format on 2D graphs – bar chart and line chart. DPT-VW also includes converters that are already available online such as CSV to JavaScript Object Notation (JSON), CSV to Tab Separated Values (TSV), and TSV to CSV converters. All these tools are available in the tool selection section in DPT-VW. The tool request section permits the users to submit requests for the development of a new tool to the developer. The developer may either accept or reject their requests based on whether he or she would be able to develop the requested tool or not. The users can then view the decision taken by the developer for their submitted requests. DPT-VW also includes a feedback section where users can submit feedback and/or suggestions regarding the tools available in DPT-VW. The developer, after logging into the application, can view all the feedbacks and/or suggestions submitted by the users.

All of the data available through the data portal are stored on the GSToRE platform. GSToRE is a flexible, scalable data management, discovery and delivery platform that supports a combination of Open and Community standards for client access. The platform can support both spatial and non-spatial data. The GSToRE API is the basis for all of the data portals' functionality – dataset search, dataset access, documentation access, and OGC service access. The framework used are Pyramid and SQLAlchemy (Python). The databases used are
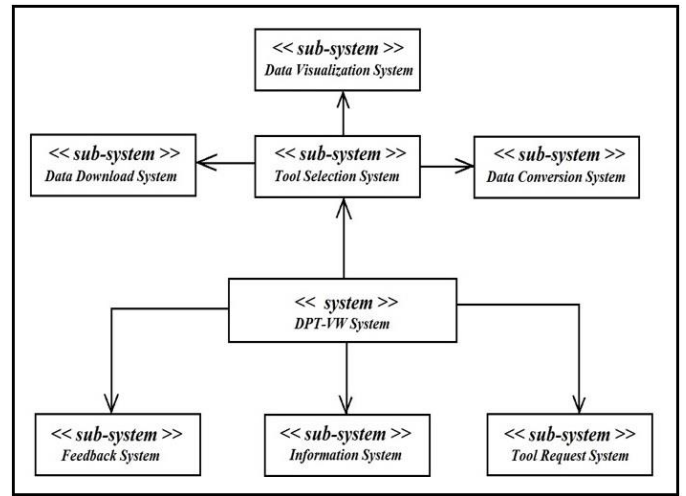

Figure 3. System Level Diagram of DPT-VW

PostgreSQL/PostGIS (canonical data store, spatial data processing), MongoDB (vector and tabular attribute storage), ElasticSearch (search engine) [7].

The application is developed using Flask [8], a micro web application framework written in Python along with HTML5, jQuery, JavaScript, CSS5, and Bootstrap. D3 [9] and Dygraph [10] JavaScript libraries are used for the CSV visualization in 2D graphs. The tools for the conversion of PRMS files into NetCDF format are written using netcdf4-python library. SQLite3 [11] is used as the back end.

The flowchart illustrated in Fig. 4 describes the conversion of PRMS parameter input file from text format into NetCDF format. The user wants to upload the file in text format and the location file which includes the latitude and longitude values of each HRU. The location file is required to visualize the values of a parameter on the right location. The user should also input the number of rows, columns and HRUs of PRMS. The number of HRUs is 4704 which is equal to the product of number of rows and number of columns (49 * 96). The tool verifies if the user has uploaded a file of correct format and then converts the file into NetCDF format. After conversion, the NetCDF format file is downloaded and a message is displayed to the user informing that the file has been successfully converted and downloaded. Fig. 5 is an example of visualization of a space and time related parameter, tminf (Minimum air temperature distributed to each HRU).

A NetCDF file is created in python using Dataset constructor. A NetCDF Dataset is a collection of dimensions, variables, and attributes. This collection is used to describe the meaning of data stored in a NetCDF file. The size of all the variables are defined in terms of dimensions. Hence, dimensions used by the variables should be created before the variables are created. A python string is used to set the name of the dimension, and an integer value is used to set the size of the dimension. The createDimension method of Dataset is used to create a NetCDF dimension. A variable can have more than one dimension [12]. The files that contain the temporal (time related) data include time as a dimension in the NetCDF file. The files that contain the spatial (space related) data
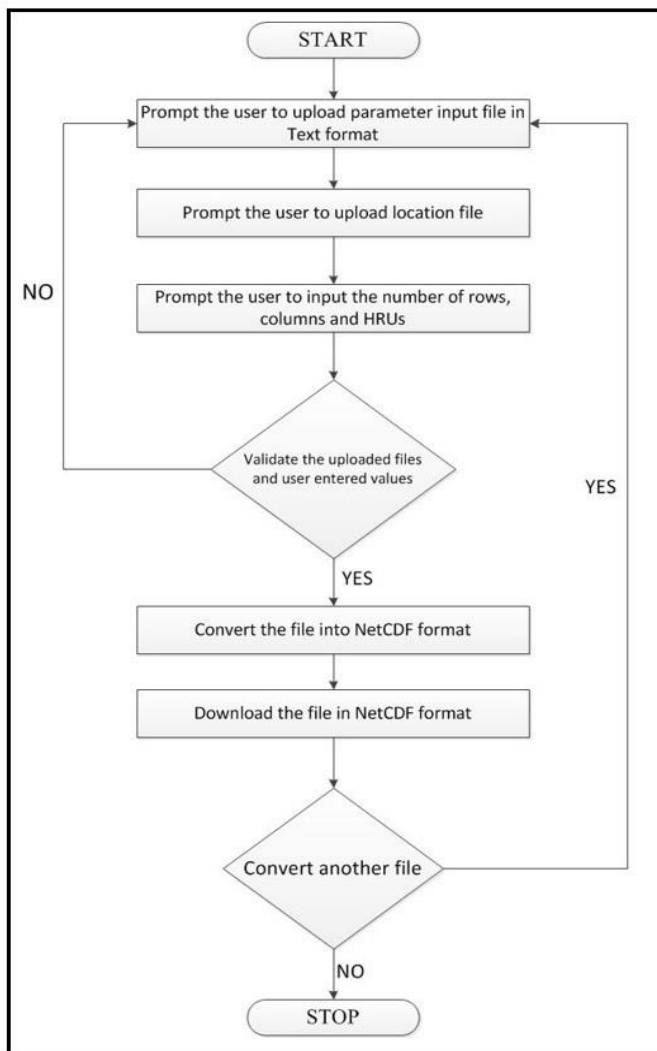
Figure 4. Flow Chart of DPT-VW Conversion Tool

Figure 5. Visualization of Space and Time Related Parameter

include latitude and longitude as the dimensions in the NetCDF file.

The createVariable method of Dataset is used to create a NetCDF variable. The arguments of the createVariable method are the variable name, variable datatype and dimension names of the variable. A python string is used to set the name of the variable. The datatype can be 'f4' (32-bit floating point), 'f8' (64-bit floating point), 'i4' (32-bit signed integer), 'i2' (16-bit signed integer),'i8' (64-bit singed integer), 'i1' (8-bit signed integer), 'u1' (8-bit unsigned integer), 'u2' (16-bit unsigned integer), 'u4' (32-bit unsigned integer), 'u8' (64-bit unsigned integer), or 'S1' (single-character string). The dimensions of the variable are given by a tuple containing the dimension names. The dimensions themselves are usually also defined as variables, called coordinate variables [12].

The two types of attributes in a NetCDF file are global and variable. The global attributes provide detailed information about the entire dataset. The variable attributes provide detailed information about one of the variables in a dataset. Gl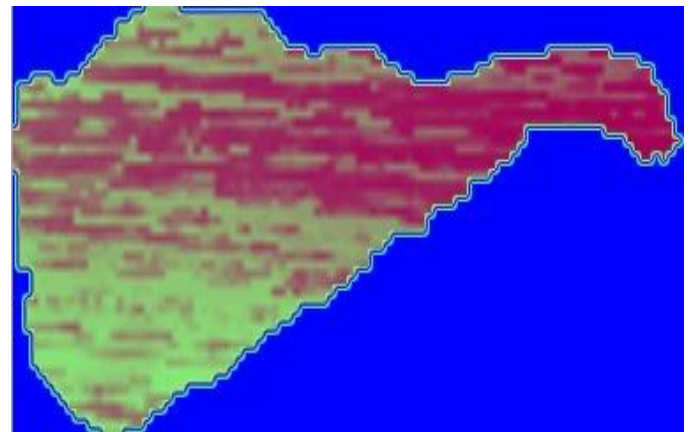obal attributes are set by assigning values to Dataset instance variables. Variable attributes are set by assigning values to Variable instances variables [12].

The SQLite3 database includes four tables: watersheds, datasets, questions and feedbacks. The watersheds table stores the details of the watersheds available from GSTORE. The primary key of the watersheds table is the id. The fields in the table are id, name, and uuid where name is the name of the watershed, uuid (Universally Unique Identifier) is the unique id of the watershed. When the user selects the 'Data Downloaders' option from the 'Tool Selection' section, the names of all the watersheds are displayed. When the user selects a name from the list, the user is redirected to the next web page.

The datasets table stores the details of the datasets of watersheds available from GSTORE. The primary key of the datasets table is the id. The fields in the table are id, wuuid, name, and duuid where wuuid is the unique id of the watershed, name is the name of the watershed, duuid is the unique id of the dataset. When the user selects a name of watershed from the list, the names of the datasets of the selected watershed are extracted from the 'datasets' table and are displayed on the web page.

The questions table stores the requests/suggestions submitted by the users to the developer. The primary key of the questions table is the id. The fields in the table are id, name, email, question, and status. When the user submits the question, the status will be stored as 'N/A'. After the developer logs into the website, he/she can either accept or reject the question. Then, the status will be updated to either 'Accepted' or 'Rejected'.

The feedbacks table store the feedbacks submitted by the users to the developer about the tools available in DPT-VW. The primary key of the feedbacks table is the id. The user will be asked to enter his/her name, feedback, and rating of the website on a scale of 1 to 5, 5 being the highest. The fields in the table are id, name, feedback, scale.

The size of all the PRMS files excluding the animation file is 2 MB. The time required to convert these files is 2.5 seconds. The size of the animation file is 2GB. The time required for the conversion of animation file is 70 seconds.

IV. Software System And Results

We designed our application based on the use cases represented in Fig. 6. The description of each use case is given below.

A. ViewVideo

The home page of DPT-VW gives an overview of the website along with a video to the users. The video gives the users an idea regarding how to use the various data processing tools provided by the website.

B. UploadInputFile

Before using the data processing tools, the user wants to upload an input file of specified format. When the user clicks the 'Upload' button, DPT-VW pops up a window that enables the user to choose a file to upload. If the file is of the supported format, the files is uploaded to the server.

C. ConvertFile

When the user uploads a PRMS file of text format, DPT-VW converts the file into NetCDF format. The PRMS model has three input files (control, data and parameter) and three output files (statistic variables, animation and water budget). All these files are in text format. The converted NetCDF file includes dimensions, variables, and attributes. If the uploaded file is of wrong format, error message is displayed.

D. VisualizeFile

When the user uploads an input file of CSV format, DPT-VW visualizes the data in the file on a line chart and a bar chart. If the file is of unsupported format, DPT-VW displays an error message.

E. DownloadOutputFile

After performing the conversion of PRMS files from text to NetCDF format, DPT-VW allows the user to download the output file of NetCDF format. Also, DPT-VW enables the user to download the datasets of the watersheds available from the GSToRE framework in a GeoTIFF format. Using the existing tools available in DPT-VW, users can download file in JSON (converted from CSV format), TSV (converted from CSV format), and CSV (converted from TSV format) formats.

F. UseExistingTools

DPT-VW provides existing data processing tools to the end user. All the existing data processing tools are available in the tool selection section. Some examples of existing tools available in the tool selection section are CSV to JSON, CSV to TSV, and TSV to CSV converters.

G. SubmitRequest

If the user wants additional tools on the DPT-VW website, he/she can submit a request for the new tools to the developer. When the user selects the 'Tool Request' option, he/she will be redirected to a web page which includes three questions – name, email and question. The user wants to fill out the answers for these questions and select the 'Submit' button. All the answers will be stored in the database.
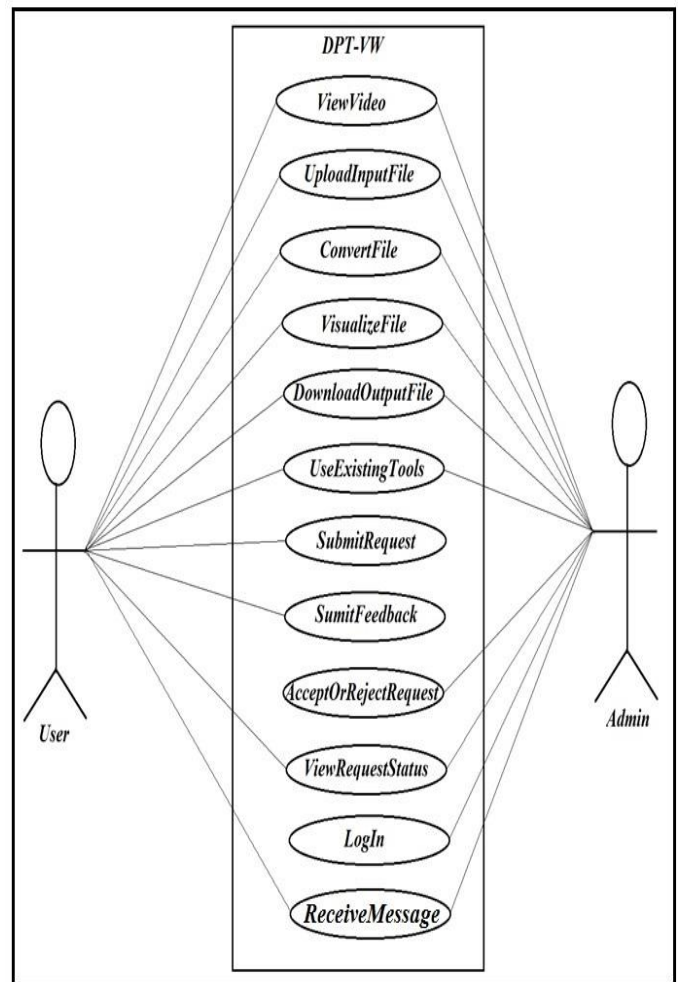

Figure 6. Use case diagram of DPT-VW

H. SubmitFeedback

If the user wants to let the developer know his/her opinions and/or suggestions about the DPT-VW website, he/she can submit the feedback to the developer. When the user selects the 'Feedback' option, he/she will be redirected to a web page which includes two questions – name and feedback. The user will also be asked to rate the website on a scale of 1 to 5. The user wants to fill out the answers for these questions and select the 'Submit' button. All the answers will be stored in the database.

I. AcceptOrRejectRequest

The developer, after logging into the DPT-VW website, can view all the tool requests submitted by the users. The developer can either accept or reject the tool requests. The status (Accepted/Rejected) of the tool requested will be added to the database. The user can then view the status of the tool requests by selecting the 'List of New Tool Requests' option.

J. ViewRequestStatus

When the user selects the 'List of New Tool Requests' option, he/she can view the list of tool requests submitted by all the users. It includes the total number of questions/requests,

number of requests accepted and rejected by the developer, and the number of requests not viewed by the developer yet. It will also display the list of tool requests.

K. LogIn

The developer has the option to log in into the DPT-VW website and view the tool requests and feedbacks submitted by the users. The developer has to submit the correct username and password to successfully log in and view the tool requests and/or feedbacks.

L. ReceiveMessage

DPT-VW website displays messages to the user in various situations. Successful messages will be displayed after (i) converting the input file of text format into a NetCDF format (ii) downloading the datasets of the watersheds available from GSTORE. Error messages will be displayed if (i) the developer enters a wrong username and/or password, (ii) the user uploads a file of format not supported by DPT-VW.

Data downloaders allow the users to download datasets of watersheds available from GSTORE. First, DPT-VW displays a list of all the watersheds available from GSTORE. When the user selects a name of watershed from the list and clicks the 'Select' button, he/she will be redirected to a page which displays a list of all the datasets available for the selected watershed. When the user selects a name of dataset from the list and clicks 'Download' button, the dataset gets downloaded in GeoTIFF format. The user will receive a 'Successfully downloaded' message once the dataset download has been successfully completed.

Figures 7-9 represent the visualization of CSV file uploaded by the users on a bar chart and a line chart. When the user uploads a CSV file and selects 'Upload' button, DPT-VW displays the names/labels of each column in the file. The user wants to select two checkboxes and select 'Visualize'. The data will be displayed on a bar chart as well as on a line chart. The color of each bar changes as the user moves the pointer as shown in Fig. 7. Similarly, as illustrated in Fig. 8, DPT-VW visualizes the CSV file on a sortable bar chart where the bars get sorted when the user selects the 'Sort values' option. Fig. 9 shows the visualization of CSV file on a line chart.
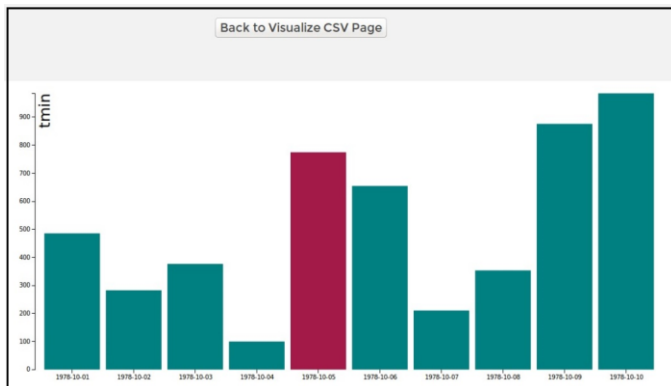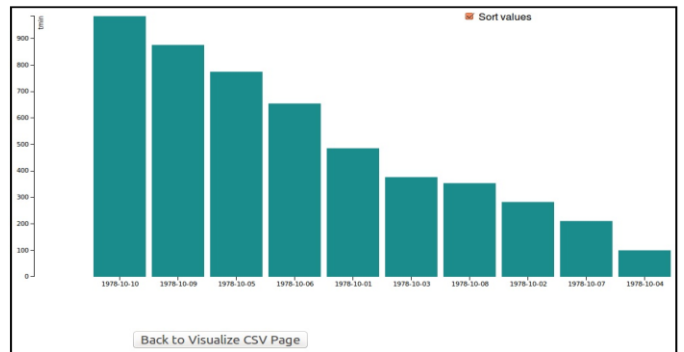

Figure 7. Visulization of CSV file on a bar chart

Figure 8. Visulization of CSV file on a sortable bar chart

V. COMPARISON WITH RELATED WORK

Mr. Data Converter [13] is a web application that converts a file in CSV format into various web-friendly formats such as HTML, JSON and XML. The user wants to copy and paste the input data on the space provided and select the output format required from the dropdown list. The application is available online and is free to use. However, the application does not provide an option to upload and/or download the file(s). If the user wants to convert a file, he/she has to copy and paste the data from the file on the space provided in the website. After the data conversion, the user again has to copy and paste the converted data into a file. The tool also does not show the progress of file conversion which would affect the user satisfaction. DPT-VW allows the user to upload the file for conversion and visualization. After conversion, the user can download the file. DPT-VW also shows the progress of file conversion.

MATLAB (MATrix LABoratory) is a desktop application used by millions of users worldwide for data visualization and processing. MATLAB can visualize data in 2D and 3D because it provides built-in 2D and 3D plotting functions [14]. The users have to write MATLAB scripts to visualize and process the data which would be difficult for the researchers and the students who have never learned programming. A well written documentation about MATLAB is available online which includes scripts and built-in functions required for visualization and processing. The MATLAB license is around $100. Because it is a desktop application, MATLAB cannot be accessed by the users online. DPT-VW does not include 3D visualization tools. We are planning to extend our project by adding 3D visualization tools in the future.

VISTED (VISualization Toolset for Environmental Data) is a client/server web application, the goal of which is to aid the researchers in the visualization of the climate data [15]. The features of VISTED are data selection, extraction, download, conversion, and visualization. The tool is easy to use and the visualization results are interactive. The users will be able to download the selected data in ASCII, binary and NetCDF formats. However, users can visualize and download datasets generated by the NCAR/WRF climate model [16]. DPT-VW includes the features such as data upload, conversion, download, and visualization.
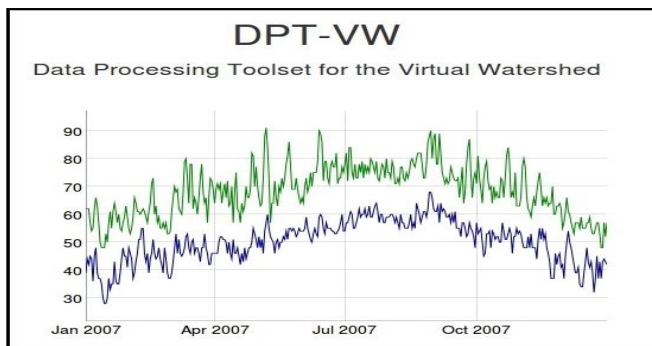
Figure 9. Visulization of CSV file on a line chart

VI. CONCLUSION AND FUTURE WORK

The paper describes about DPT-VW, a new web application that provides data processing tools effective for the WC-WAVE project. The main features of DPT-VW are data conversion, data download, and data visualization. The tools can convert PRMS files from text to NetCDF format, CSV to TSV and JSON formats and CSV to TSV format. The application enables the users to download datasets from GSToRE. It also supports visualization of CSV files on a bar chart and a line chart. It also allows the users to submit feedback and requests for the development of a new tool to the developer.

In this paper, the basic and core functionalities of DPT-VW are presented. In the future, DPT-VW will include more data conversion, download, 2D and 3D visualization techniques. The tool request section will also incorporate additional features where the users receive status of their tool request (accepted/rejected with a valid reason) via their email. The developer will also inform the users the time period required to develop the requested tool.

ACKNOWLEDGMENT

REFERENCES

[1] What is a watershed? [Online]. Available: http://water.epa.gov/type/watersheds/whatis.cfm/. [Accessed: 14-February-2016].

[2] The western consortium for water analysis, visualization and exploration [Online]. Available: http://www.nmepscor.org/science/western-consortium-water-analysis-visualization-and-exploration. [Accessed: 14-February -2016].

[3] C. Carthen, T. Rushton, C. Johnson, A. Hesson, D. Nielson, B. Worrell, J. Anderson, R. Lew, N. Wood, M. Ziegler, D. Delparte, W. Johansen, S. Dascalu, and F. Harris, Jr. "Design of a virtual watershed client for the WC-WAVE project", Proceedings of The 2015 International Conference on Collaboration Technologies and Systems (CTS-2015), Atlanta, June 2015, pp 90-96.

[4] NetCDF [Online]. Available: http://www.unidata.ucar.edu/software/netcdf/docs/. [Accessed: 14-February -2016].

[5] M. Hossain, S. Dascalu, and F. Harris, "A Software Environment for Watershed Modeling", Proceedings of the 24th International Conference on Software Engineering and Data Engineering (SEDE 2015), San Diego, CA, October 2015, pp X-XX.

[6] PRMS-IV, the precipitation runoff modeling system [Online]. Available: http://pubs.usgs.gov/tm/6b7/pdf/tm6-b7.pdf/. [Accessed: 14- February -2016].

[7] Virtual Watershed API Documentation [Online]. Available: http://vwp-dev.unm.edu/docs/. [Accessed: 14- February -2016].

[8] M. Grinberg, Flask web development: developing web applications with python, 1st ed, O'Reilly Media, CA, USA, 2014.

[9] A. Jain, "Data visualization with the d3.js javascript library," Journal of Computing Sciences in Colleges, vol. 30, no. 2, pp. 139-141, 2014.

[10] Dygraphs [Online]. Available: http://dygraphs.com/. [Accessed: 14-February -2016].

[11] G. Allen and M. Owens, The definitive guide to sqlite, 2nd ed, Apress Berkely, CA, USA, 2010.

[12] netCDF4 module [Online]. Available: http://unidata.github.io/netcdf4-python/#netCDF4.Dataset. [Accessed: 14-February-2016].

[13] Mr. data converter [Online]. Available: https://shancarter.github.io/mr-data-converter/. [Accessed: 14- February -2016].

[14] B. R. Hunt, R. L. Lipsman and J. M. Rosenberg, A Guide to MATLAB for Beginners and Experienced Users, Cambridge University Press, 2014.

[15] L. Ravi, S. Dascalu, F. Harris, Jr., J. Mejia and N. Belkhatir, "VISTED: a visualization toolset for environmental data," Proceedings of The 2015 International Conference on Computers and Their Applications (CATA 2015), Honolulu, HI, March 2015, pp. 335-342.

[16] Nevada climate change portal [Online]. Available: http://sensor.nevada.edu/. [Accessed: 14- February -2016].