Assessing Water Budget Sensitivity to Precipitation Forcing Errors in Potomac River Basin Using the VIC Hydrologic Model

CyberTraining: Big Data + High-Performance Computing + Atmospheric Sciences

Team 1: Reetam Majumder[1], Redwan Walid[2], Jianyu Zheng[3],
Research assistants: Carlos Barajas[1], Pei Guo[2], Chamara Rajapakshe[3],
Faculty mentors: Aryya Gangopadhyay[2], Matthias K. Gobbert[1],
Jianwu Wang[2], Zhibo Zhang[3]
Clients: Kel Markert[4], Amita Mehta[5], Nagaraj K. Neerchal[1]

[1]Department of Mathematics and Statistics, UMBC
[2]Department of Information Systems, UMBC
[3]Department of Physics, UMBC
[4]University of Alabama Huntsville / NASA-SERVIR
[5]Joint Center for Earth Systems Technology, UMBC

Technical Report HPCF–2019–11, hpcf.umbc.edu > Publications

Abstract

The Potomac River Basin is a watershed located on the East Coast of the USA across West Virginia, Virginia, Pennsylvania, Maryland, and the District of Columbia. Inter-annual variations in precipitation makes it challenging to plan for water allocation within the basin. Therefore, understanding seasonal to inter-annual variations in water availability within the basin is important for planning water resources management. We set up on a distributed-memory cluster and used the hydrologic model Variable Infiltration Capacity (VIC) to estimate the water budget components for the Potomac river basin from April to September 2017. We also assessed the effect of precipitation forcing errors and its variability on the water balance for the same time period. We were able to identify April and May as the months where the water balance was most sensitive to variability. Sub-basins with the highest sensitivity over the course of the six months of interest were also identified, and variability in water balance increased as we increased the variability in precipitation.

1 Introduction

The Potomac River Basin, also known as the Potomac watershed, is located on the East Coast of the USA across West Virginia, Virginia, Pennsylvania, Maryland, and the District of Columbia. Figure 1.1 from NASA's "Blue Marble" image[1] provides a sketch of the Potomac river basin's extent. The Interstate Commission on the Potomac River Basin (ICPRB)[2] manages the water basin, which has a drainage area of 14,670 square miles. According to the ICPRB, about 600 million gallons per day (mgd) is used for water supply, of which 500 mgd is for the Washington area. About 1.6 billion gallons, most of which is returned to

[1]Sourced using the `Basemap` package in `Python` from https://visibleearth.nasa.gov/
[2]Interstate Commission on the Potomac River Basin. https://www.potomacriver.org

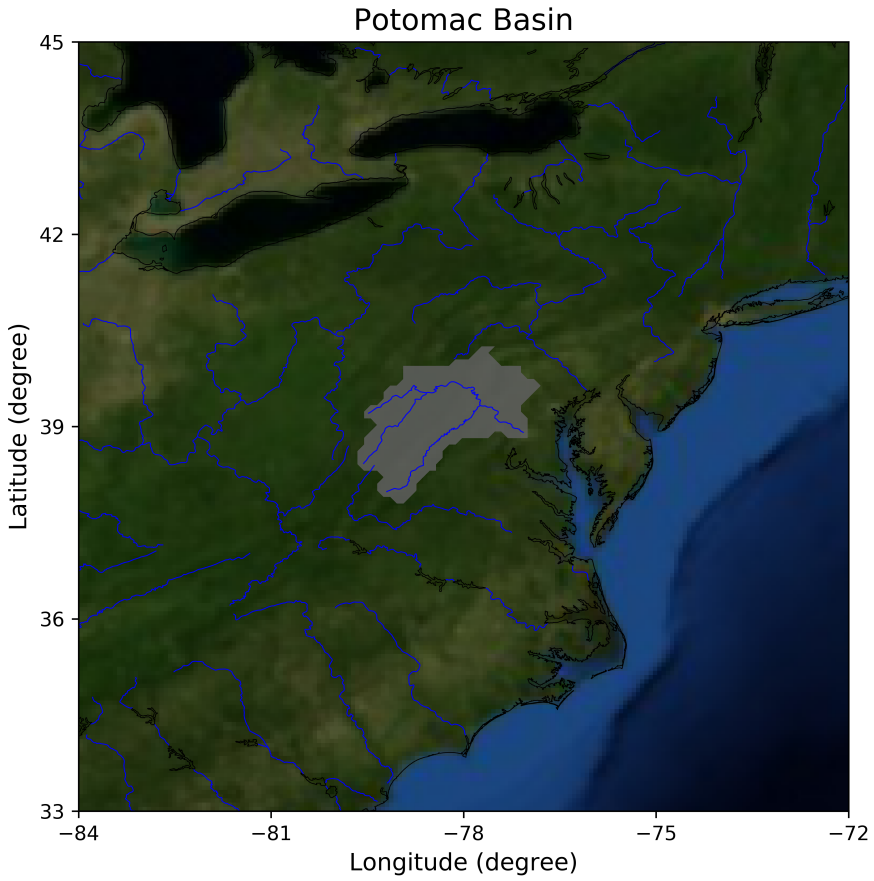


Figure 1.1: Extent of the Potomac river basin on the East Coast of the USA indicated by the gray shape; rivers are represented by blue lines. The Potomac originates at two separate sources in West Virginia and Virginia, and drains into the Chesapeake Bay which connects to the Atlantic ocean on the Eastern side of the map.

the streams, is used daily for power plant cooling and other industrial uses.[3] Agricultural and commercial fishing practices are also common. Therefore, sufficient quantities of clean water are essential for human needs and also for ecosystem health.[4] To meet the demand for drinking water in the District of Columbia and the growing suburbs in Maryland and Virginia, the Cooperative Water Supply Operations on the Potomac (CO-OP)[5] is in place since 1979. Accordingly, agreements among water utilities of D.C., Maryland, and Virginia

[3]Potomac Basin facts. https://www.potomacriver.org/potomac-basin-facts

[4]Drinking Water and Water Resources. https://www.potomacriver.org/focus-areas/water-resources-and-drinking-water

[5]CO-OP. https://www.potomacriver.org/focus-areas/water-resources-and-drinking-water/cooperative-water-supply-operations-on-the-potomac

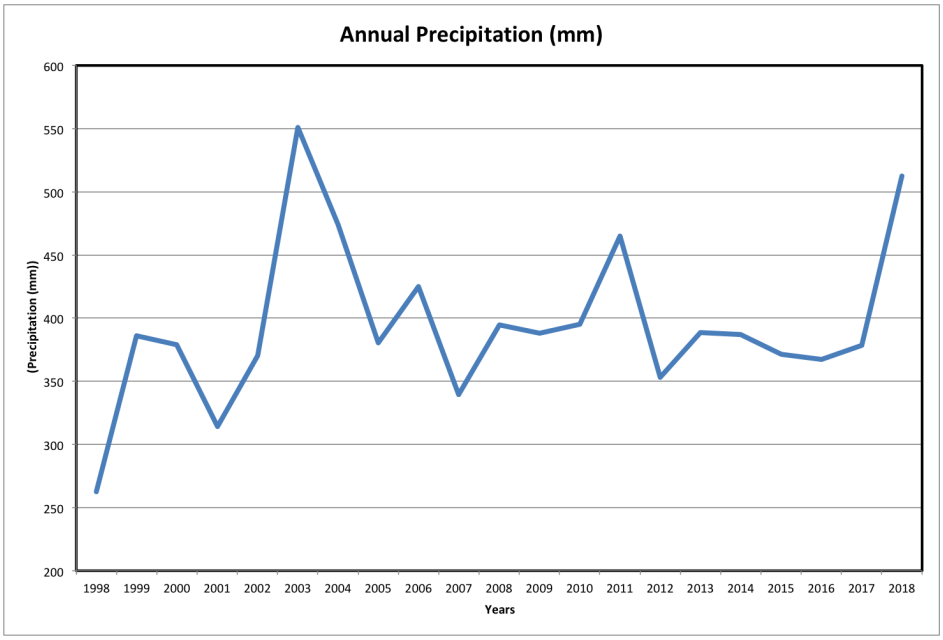


Figure 1.2: Basin-averaged annual precipitation time series from NASA's Tropical Rainfall Measuring Mission (TRMM) Multi-satellite Precipitation Analysis (TMPA) from 1998 to 2018.

ensure water allocation based on water availability and water usage. In addition to increasing demand on water, climate variability plays an important role in the region's water supply. In particular precipitation, the main source of water in the Potomac Basin, varies inter-annually and makes it challenging to plan for water allocation within the basin. Therefore, understanding seasonal to inter-annual variations in water availability within the basin due to climate variability is important for planning water resources management.

Figure 1.2 shows a basin-averaged annual precipitation time series from NASA's Tropical Rainfall Measuring Mission (TRMM) Multi-satellite Precipitation Analysis (TMPA) from 1998 to 2018 [7]. Clearly, substantial year-to-year variations of precipitation can be seen, with the mean around 394 mm. Moreover, as seen in Figure 1.3, the mean precipitation map from TMPA shows spatial pattern of precipitation over the basin. The central basin receives relatively lower precipitation compared to the eastern and western portions of the basin. In this study, we focus on assessing the spatial and temporal variability of water availability in the Potomac Basin using a hydrologic model. The water budget equation for a river basin, considering that there is no surface, sub-surface, or groundwater net inflow/outflow in the watershed, and that surface runoff and baseflow contribute to discharge can be expressed as

$$\Delta S = Pr - ER - RO - \text{Baseflow}$$

where

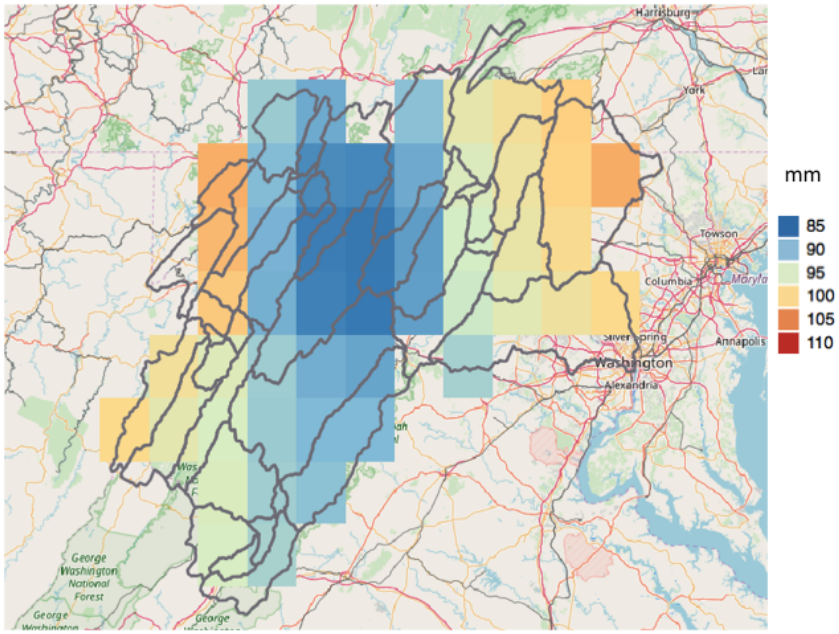


Figure 1.3: Map of the Potomac river basin showing mean precipitation from 1998 to 2018 based on TMPA.

- Δ**S** is the change in water storage in the basin surface (snow amount, soil moisture) and sub-surface (root zone moisture, groundwater) water storage components,
- **Pr** is Precipitation,
- **ET** is Evapotranspiration, a combination of evaporation from bare soil and wet canopy, and transpiration from each vegetation type,
- **RO** is the Surface Runoff, a function of soil wetness and soil infiltration capacity, and
- **Baseflow**, or the Sub-surface Runoff, is the portion of the precipitation that is sustained between precipitation events, and is a function of surface and sub-surface soil moisture.

The available water over the basin, referred to as the water budget or the water balance, can be estimated by subtracting ET, surface and sub-surface runoff from the precipitation.

We use Variable Infiltration Capacity (VIC)[6], a hydrologic model, to estimate the water budget components. As described in Section 2, VIC is forced by precipitation and provides daily ET, RO, and Baseflow as outputs. VIC uses daily precipitation, surface temperature and wind data as input. In this study we use precipitation from the Global Precipitation Measurement (GPM) mission; specifically, we use Integrated Multi-satellitE Retrievals for

[6]VIC Macroscale Hydrologic Model. https://vic.readthedocs.io/en/master

GPM (IMERG) data derived by merging passive microwave measurements from a constellation of satellites, calibrated with those from the GPM Core Observatory, and adjusted with measurements from global rain gauge networks.[7]

The primary goal of this project is to estimate water budget components for the Potomac river basin using VIC forced by the daily IMERG precipitation. We focus on the rainy season (April to September) of year 2017. It is important to note that merging of data from different satellites with varying spatial resolution and temporal sampling in IMERG leads to errors in the precipitation estimates. Generally, the quality of the estimates can be broken into errors resulting from bias (departure of the estimates from the true values) and random fluctuations around the true values due to measurement-algorithm effects and sampling errors [5]. In case of IMERG data, the random error is dominant, and the bias error is smaller. Further, while the bias is compensated for, the random error cannot be corrected [6]. We aim to assess how intra-seasonal variability of precipitation within the basin affects water budget estimates and how the errors in the IMERG precipitation influence water budget components, both spatially and temporally. The latter is done using a sensitivity analysis.

Sensitivity analysis (SA) can be defined as "the study of how uncertainty in the output of a model (numerical or otherwise) can be apportioned to different sources of uncertainty in the model input" [16]. A few common purposes, or settings, discussed in [15] are:

- **Factor Prioritization** aims to rank input factors by their relative contribution to the output variability
- **Factor Fixing** identifies if any factor has a negligible effect on output variability
- **Mapping** aims to determine the region of input variability that significantly alters output, producing extreme values as an example.

A detailed review of SA methods in the context of environmental models can be found in [13], which also states that "the simplest type of SA varies (perturbs) the input factors of the simulation model from their nominal values one-at-a-time (OAT) and assesses the impacts on the simulation results via visual inspection." Since precipitation is the only immediate factor of interest, our setup is already OAT. We conduct a Local SA, where Local entails that we only consider the variation of daily precipitation around a specific value instead of variations over the entire possible space of variability. We fit a distribution to the daily precipitation data based on its spatio-temporal features; the parameters of the distribution provide us the nominal value around which input variability is considered. Then, by changing the standard deviation of the distribution without affecting its mean, we inspect the output standard error and interquartile range of the water balance to establish the months and sub-basins that are the most sensitive to this change.

To achieve these goals we first execute (run) the VIC model with daily IMERG precipitation for April to September 2017 and monthly water budget components are analyzed over the basin. For the Sensitivity Analysis we conduct multiple runs of VIC forced by IMERG adjusted by a range of errors; this provides us the distribution of the water balance based

[7]GPM-IMERG. https://pmm.nasa.gov/gpm/imerg-global-image

on the resampling distribution of the precipitation. We examine how variability in precipitation errors impact the water budget components. More broadly, the goal is also to set up a framework for running VIC on the distributed-memory cluster taki in the UMBC High Performance Computing Facility (hpcf.umbc.edu) that can scale to conduct analyses over a significantly longer duration of time, and over larger river basins. This results in a Big Data problem, since raw data for each year is over 100 GB and requires significant pre-processing before model runs. Thus, this project is an example of a problem that encompasses all areas of our CyberTraining initiative:

(i) HPC expertise is required to set VIC up to run on a sophisticated distributed-memory cluster like taki, to use the mandatory job scheduler effectively, and for using the pre-processing tools and scripts.

(ii) Big Data skills are brought in to obtain and pre-process the different input datasets. Most of the raw data are in netCDF format and have to be converted to binary/ASCII, aggregated in some cases, and need to be spatially and temporally aligned; an understanding of atmospheric physics is leveraged for these steps.

(iii) Domain knowledge is vital to analyze the spatial variations in monthly water budget components over sub-basins within the Potomac basin, to assess the relative contributions of the various sub-basins to overall water availability.

(iv) Statistical techniques are used to measure the effect of precipitation errors and its variability on the water budget components.

Using primarily the HPC and Big Data tools, we were able to set up and run VIC on the distributed-memory cluster taki at UMBC. This required the installation of multiple software, getting data from different sources and aligning them. The entire process was streamlined and a scalable framework established. Next, we estimated the water budget components for the Potomac river basin using VIC for the months of April to September 2017; water balance was found to be positive for May and July, and negative for April, June, and September. The distribution of water balance shows a pattern similar to the spatial distribution of rainfall, especially for April and May. We believe that an increase in vegetation increases evapotranspiration in the basin, resulting in higher loss of water as we go from June to September. Using parametric resampling, we ran a preliminary sensitivity analysis for the precipitation data. The goal was to see if VIC's outputs are reflective of the dispersion in the input precipitation data. We were able to identify April and May as the months where the water balance was most sensitive to variability. Sub-basins with the most sensitivity over the course of the six months of interest were also identified, and variability in water balance scaled as we increased the variability in precipitation.

The remainder of this report is organized as follows. In Section 2, we discuss model details and the datasets utilized in our study. Then we focus on the tools used to conduct data pre-processing before we can run the VIC model; the results of the model run are outlined in Section 3. In Section 4, we discuss the procedure and results of the sensitivity analysis. Finally, Section 5 provides the conclusions and outlines ideas for future work.

2 The VIC Model and Related Tools

2.1 Model Description and Features

The VIC model, shown in Figure 2.1, is a semi-distributed, macro-scale hydrologic model (MHM) which can be used to understand hydrological processes in almost real-time [9]. According to Hamman et al. [3], it is a research model developed around 1994, "and has since been used extensively for basin- to global-scale applications that include hydrologic dataset construction, trend analysis of hydrologic fluxes and states, data evaluation and assimilation, forecasting, coupled climate modeling, and climate change impact assessment." VIC can be run in either a water balance mode or a water-and-energy balance mode; both modes evaluate time before space while modeling. The model also requires a minimum of one year of spin-up time. VIC simulates the water and the energy fluxes that occur near the land surface and provides necessary information regarding the water budget of any region at a particular time. The water and the surface energy budgets are computed independently for each grid cell; variations at the sub-grid level are modeled statistically. It is assumed that there is no channel flow, sub-surface flow, or recharge to soil from the river; water is assumed to enter a grid cell via the atmosphere (precipitation) only. Also, VIC does not

[1] https://arset.gsfc.nasa.gov/water/webinars/VIC18. Image from the Open Access VIC Documentation

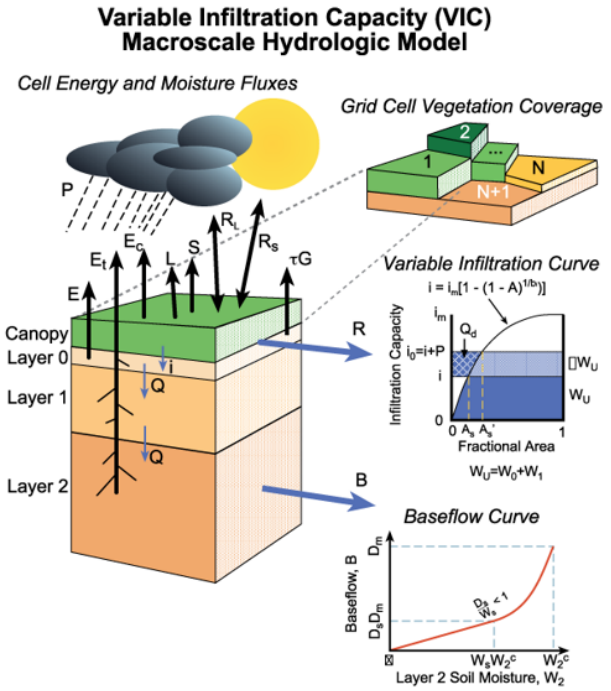


Figure 2.1: The structure of the Variable Infiltration Capacity (VIC) Macroscale Hydrologic Model and some of the physical processes considered while modeling.[1]

Table 2.1: Input variables required by VIC, along with their source and temporal resolution.

Variable	Source	Temporal Resolution
Precipitation [4]	GPM[1]- IMERG[2]	Daily
Surface air temperature [2]	MERRA-2[3]	Daily
Surface wind speed	MERRA-2	Daily
Land cover type [1]	MODIS[4]	Yearly
Leaf Area Index [12]	MODIS	Monthly
Shortwave Albedo [17]	MODIS	Monthly
Initial soil moisture conditions	Precipitation climatology	Static
Soil Characteristics [18]	HWSD[5]	Static
Annual Precipitation [4]	IMERG	Static
Elevation [8]	SRTM[6]	Static

include groundwater within the model. It is assumed that groundwater flow is relatively small compared to surface and sub-surface flow. These assumptions are generally valid for grid areas larger than 3 km and smaller than 2°. For simulating stream flow, VIC results are post-processed with a separate routing model, typically the one in Lohmann et al. [11]. We run VIC in water balance mode.

2.2 Input Data for VIC

VIC requires several sets of input data which need to be specified using a set of parameter files. These are the global parameter file, meteorological forcing files, soil parameter file, vegetation library file, and the vegetation parameter file. Additionally, there are a few optional parameters pertaining to the initial state, elevation bands, snow parameters, and lake/wetland parameters. The main input for the model is the global parameter file which has links to the other parameter files and initial states; the location and format of each input dataset is specified in its respective parameter file.

All datasets for running the model were downloaded for $76° - 80°W$, $37° - 41°N$, between January 2016 and September (in some cases, December) 2017. The model uses daily precipitation, surface air temperature, and wind speeds as atmospheric forcing. For land boundary conditions climatological means of Leaf Area Index (LAI) and surface albedo are specified. In addition, annual climatology of IMERG is used to derive initial soil moisture condition. The model uses Shuttle Radar Topography Mission (SRTM) elevation and slope provide terrain and Harmonized World Soil Database (HWSD) is used to specify soil types

[1]Global Precipitation Measurement
[2]Integrated Multi-satellitE Retrievals for GPM
[3]Modern-Era Retrospective analysis for Research and Applications, Version 2
[4]MOderate Resolution Imaging Specroradiometer
[5]Harmonized Soil World Database
[6]Shuttle Radar Topography Mission

and characteristics within the basin. The daily forcing data from January 2016 to March 2017 is used for spin-up, and daily water budget components were converted to monthly data and analyzed for April through September 2017. There are a total of 387 grid points/cells for our dataset, and we have 639 days of data for each of them.

2.3 Data Preparation and Pre-processing

The daily datasets were collated and rearranged such that at each grid point, we have a time series for each variable. The set of parameter files (global, soil, precipitation etc.) provides VIC instructions on how to read and utilize the data. Pre and post-processing scripts written in Python were available at `github.com/KMarkert/servir-vic-training`. Some packages required were `Python 2` specific, and we installed our own copy for the purposes of the project.

2.3.1 Projection and aggregation

MODIS data is available in sinusoidal grids ($10^\circ \times 10^\circ$ tile) in HDF file format, which needs to be converted into geographical (WGS84) projection. An HDF-EOS To GeoTIFF Conversion tool (HEG)[7] was required to process MODIS data which re-projects MODIS data into WGS84 format. HEG has a Graphic User Interface (GUI) but there are limitations on how many files can be edited in batch using the GUI. So instead we first generated a parameter file for each of Land Cover, LAI and Shortwave (SW) Albedo from the GUI; the parameter file generated for Land Cover was then used for all Land Cover files, and so on. HEG was run from the command line using the parameter files we generated to loop over all the files using a bash script.

LAI and SW Albedo were converted to monthly format - one value for each month for each grid point. Since they do not vary a lot, monthly averaged data was used instead of daily data. LAI has a native temporal resolution of 8 days; we used weighted averages to get monthly values. SW Albedo is available daily and it was averaged to get monthly data. Northward (v) and Eastward (u) wind component variables (from MERRA-2) were combined as wind speed by $\sqrt{u^2 + v^2}$ for each spatio-temporal component of the data using the pre-processing scripts.

2.3.2 Pre-processing and final steps

The datasets were pre-processed to only cover the Potomac Basin. Grid template of $0.1^\circ \times 0.1^\circ$ resolution was created for the Potomac basin shape file. The pixels were aligned spatially. The raster datasets (SRTM, Elevation and Slope, IGBP[8] land cover, soil and annual precipitation) were aligned to the grid template. The input files were formatted to generate the various parameter files to be read by VIC. Finally, the global parameter file was updated to run the VIC model on taki.

[7] https://newsroom.gsfc.nasa.gov/sdptoolkit/HEG/HEGHome.html

[8] International Geosphere-Biosphere Programme

2.4 Execution Times

VIC took under 2 minutes to execute (run) for our use case. However, we expect it to scale with the duration and the number of grid points in our data. The major time bottleneck was the pre-processing part of the data; The MODIS pre-processing took around 1 hour and the meteorological forcings took around 2 hours to be generated. The remaining data pre-processing (usually involving the parameter files and raster creation) each took the order of seconds.

3 Water Budget Outputs from VIC

Our main focus is on assessing the variability of the water budget components between April and September 2017. Evapotranspiration (ET), runoff and baseflow are part of VIC's outputs for each grid point at the daily level. We looked at these components both for each month, and for each grid point.

As shown in Figure 3.1, precipitation in May, July, and August is much higher than that for April, June, and September during 2017, while runoff and baseflow have the highest values in May but lower in July and August. ET is lower than precipitation in April, May and July. ET in the basin would depend on vegetation cover and surface energy fluxes while runoff depends on both vegetation cover and soil moisture conditions. Compared to ET and runoff, precipitation has a higher intra-seasonal variation, which will lead to variation in water balance.

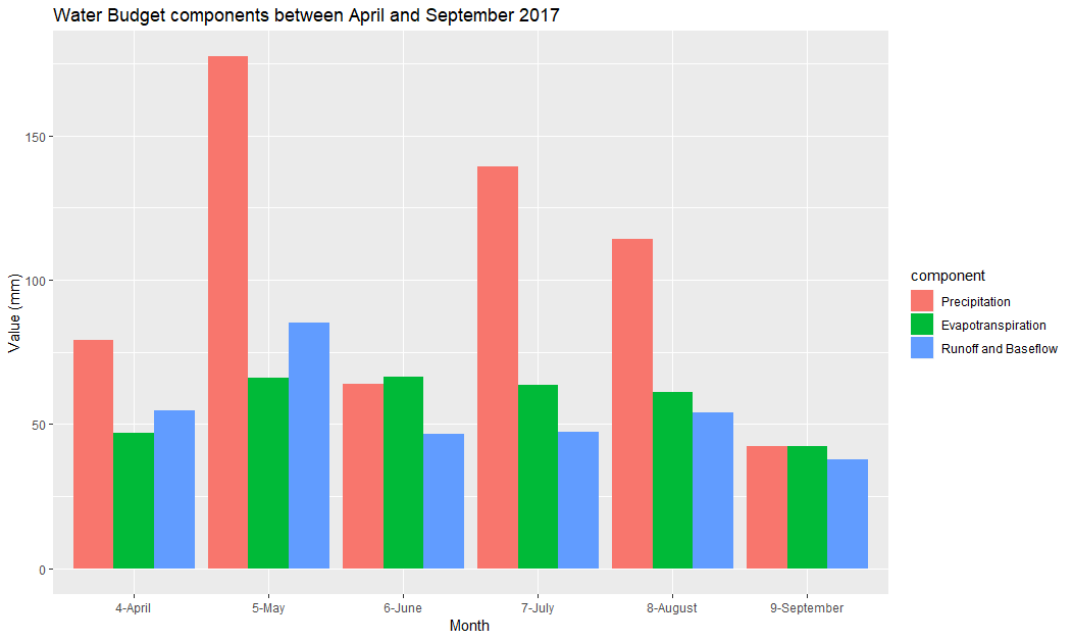


Figure 3.1: Basin-averaged monthly mean precipitation (red), evapotranspiration (green) and runoff and base flow (blue) between April and September 2017.

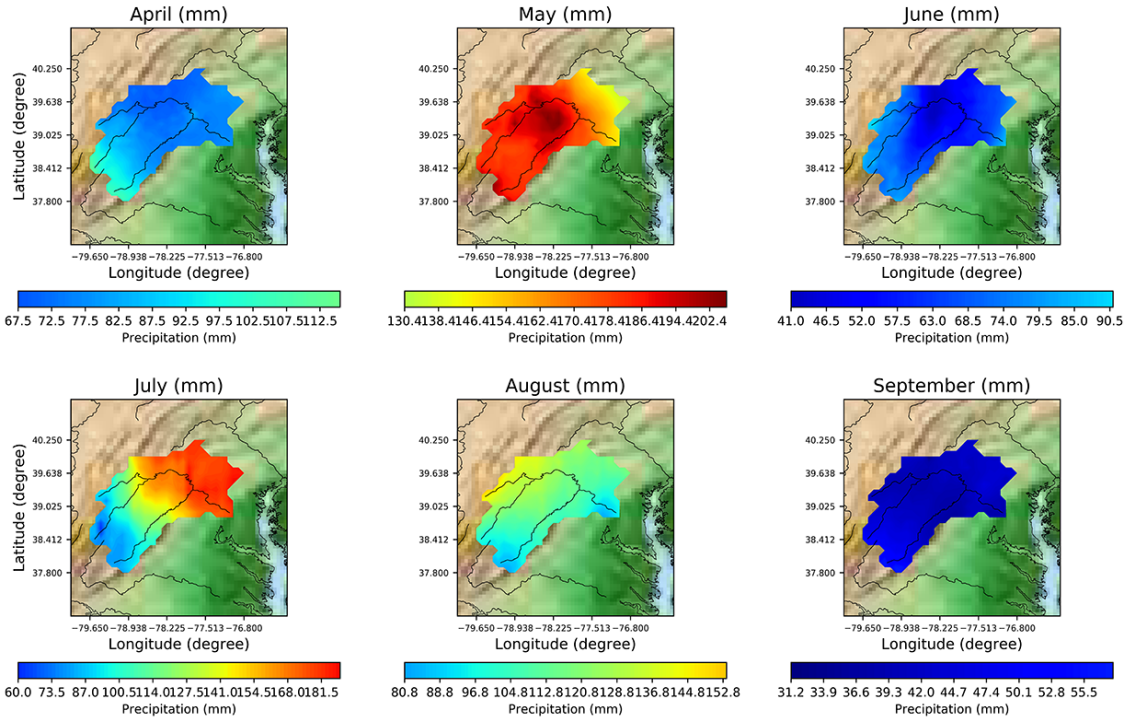


Figure 3.2: Monthly precipitation (mm) in the Potomac basin from April to September 2017.

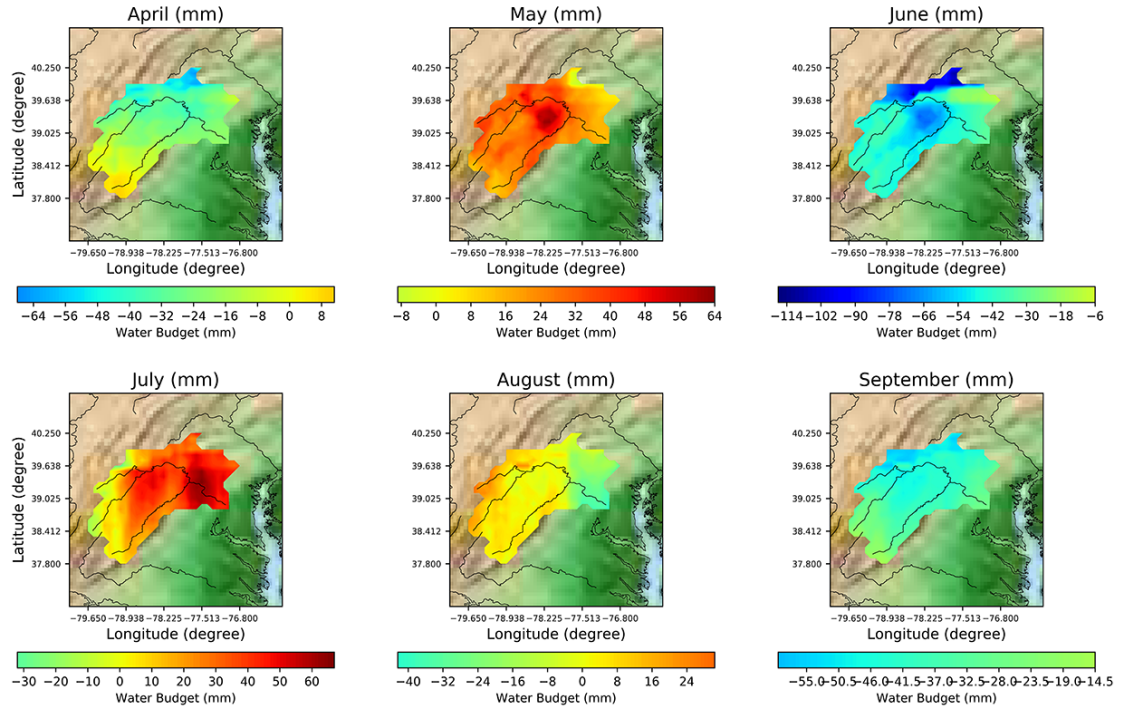
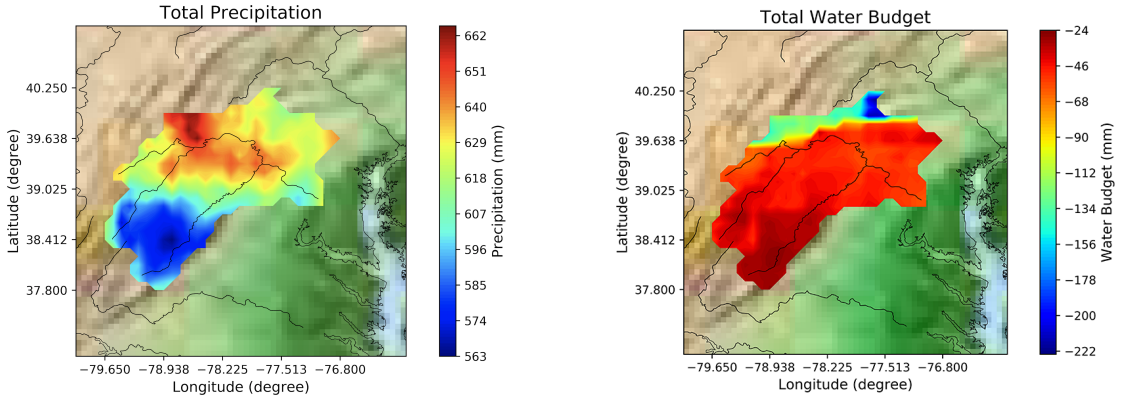


Figure 3.3: Monthly water balance (mm) in the Potomac basin from April to September 2017.

Figure 3.2 shows spatial variability of precipitation. During this period, the central and northern parts of the basin appears to have received larger amounts of precipitation in May, July, and August.

Figure 3.3 shows the distribution of water balance from April to September. Positive water balance is observed in May and July, which also have a high rate of monthly precipitation as shown in Figure 3.2. The negative water balance in April, June, and September is consistent with the lower rate of precipitation in those months while ET and runoff values remain comparable or even higher compared to other months. Also, the distribution of water balance has a similar pattern with the spatial distribution of rainfall, especially the one in April and May. During June to September, we believe that increase in vegetation would increase ET within the basin, resulting in higher loss of water to the basin.



(a) (b)

Figure 3.4: (a) Total precipitation (mm) and (b) total water balance (mm) in the Potomac basin from April to September 2017.

Total precipitation and water balance from April to September are shown in Figure 3.4 respectively. Figure 3.4 (a) shows a higher rate of precipitation in the western-central region of Potomac basin and a lower value in the southwest corner. Figure 3.4 (b) on the other hand shows water deficit across the entire basin, including a visible band at the northern part of the Potomac basin, where much larger negative values can be seen. We will be examining ET, Runoff, terrain, and snow conditions to explain the presence of this band. Further, a sensitivity analysis of precipitation can give us an idea of how the distribution of the estimated water balance from VIC can change across the river basin.

4 Water Budget Sensitivity to Precipitation Forcing Errors

4.1 Spatio-temporal features of the data

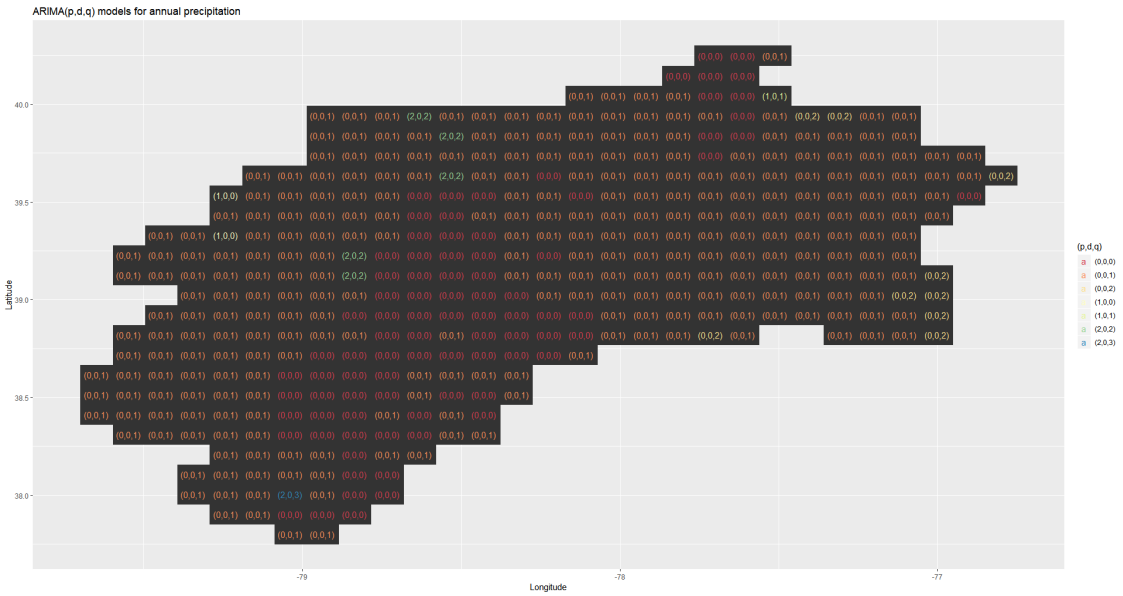


Figure 4.1: Best ARIMA(p,d,q) models for precipitation between Jan 2016 and Sep 2017 based on R's `auto.arima()` function, fitted independently at each grid point.

Our precipitation data consists of 387 grid points, each having data from Jan 2016 to September 2017. We fitted an ARIMA model for each grid point for the whole duration of the available data; Figure 4.1 shows the best possible ARIMA model fit for the time series at each location. We found that for over 90% of the grid points, there either existed no temporal structure or the data could be modeled as an MA(1) process, often with a unit root; i.e. each day's data is correlated with the innovation/error component of the previous day. Based on this, we assumed the data to be temporally independent and identically distributed. Of course, this assumption can be made only because the duration of our data is small. Should we stretch the model back to a few more years' worth of data, we might be looking at a noticeable temporal structure.

For spatial dependence, we took a simple approach. We plot a correlogram for each day across the 387 data points. The second point on the plot, with a distance of 0.12° corresponded approximately to the 0.1° resolution of our data, and that is the value we were most interested in.

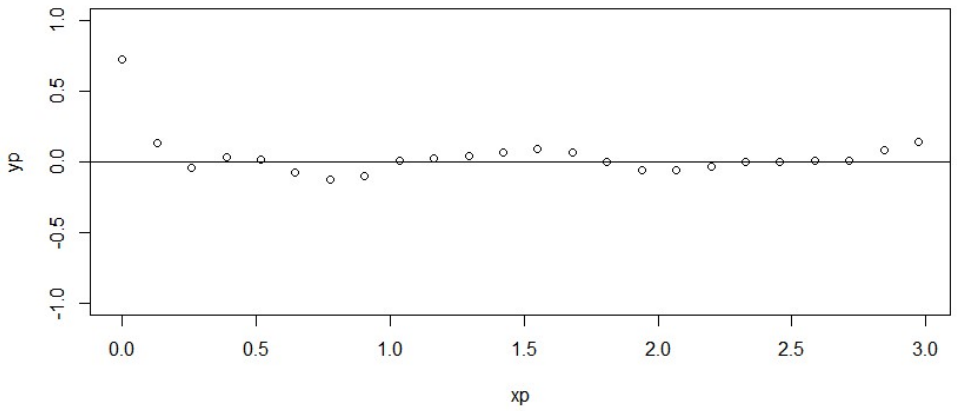


Figure 4.2: Correlogram for January 29, 2016 across 387 grid points of the Potomac basin.

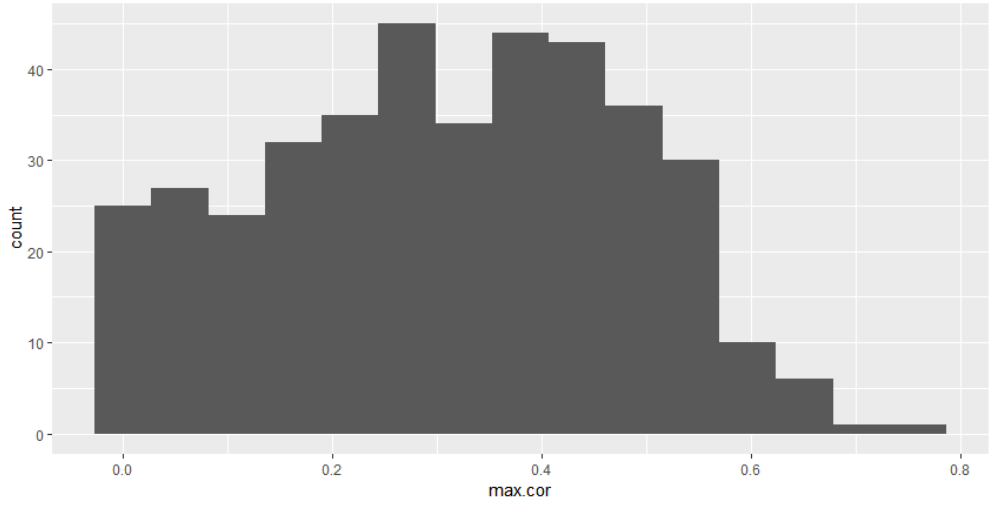


Figure 4.3: Histogram of approximate spatial correlation between adjacent grid points.

In Figure 4.2, the x-axis shows distances, and the y-axis shows the correlation of precipitation between any 2 points with that distance. We see that points very close to each other have correlated precipitation, and it falls off with increasing distance. As mentioned before, we collect the 2^{nd} point from each of the 387 such correlograms and plot them as a histogram.

From Figure 4.3 we see that the correlation is very low in most cases, though there are some high values. The truncation on the left corresponds to the days without precipitation. Based on this, we make a further assumption of the lack of spatial correlation within our data. However, we have not yet been able to explore more rigorous spatial analysis like a variogram or an anisotropic analysis of the correlation structure.

4.2 Resampling procedure and Sensitivity Analysis

Assuming from the previous section that precipitation is independently and identically distributed, its positive component can be modeled using either a Gamma or lognormal distribution [14] [10]. If we go with a Gamma distribution, we can formulate the model in a way that allows changing the variance as a proportion of the estimated mean without actually changing the mean [10]. Operationally, this is achieved by increasing the coefficient of variation without changing the mean. We propose the following resampling and analysis procedure:

- For each grid point in our data,
  - Estimate the proportion of zeros for the cell as the sample proportion, $\hat{p_0}$, corresponding to the proportion of days with no rainfall
  - Fit a Gamma distribution to the positive part of the data, parameterized by its mean $\hat{\mu_0}$ and coefficient of variation $\hat{CV_0}$
  - Generate $k = 100$ i.i.d. samples each covering the 639 days, based on the above parameters
  - Run the model with each sample, and estimate the water budget based on the output
  - Calculate the Standard Error (SE) and Interquartile Range (IQR) at each grid point
- Plot summary statistics and visualize the resampling distribution of the water balance between April and September 2017.
- Inflate the Coefficient of Variation by 50% and 100%. Plot its effect on the variability in the water balance measured in terms of Standard Error and IQR.

The resampling procedure provides us with an ensemble of precipitation forcing conditions for three different values of dispersion, allowing us to run a VIC ensemble to measure the distributional properties of the water balance. Sensitivity of the model is gauged by comparing the variability of the water budget (output, in terms of SE and IQR) as a function of the variability of precipitation (input, in terms of CV), which we inflate.

4.3 Results from the Sensitivity Analysis

From Figures 4.4 and 4.5 we see that spatially, the northern part of the basin as well as a diagonal channel in the southwest area are most sensitive to forcing errors. Since precipitation has a right-tailed distribution containing extreme values, the standard error of the estimates is less representative as we inflate the variability in the model. So while both SE and IQR show us which areas are most sensitive to precipitation forcing errors for the original variability of the data, the IQR is more meaningful to interpret as we inflate variability.

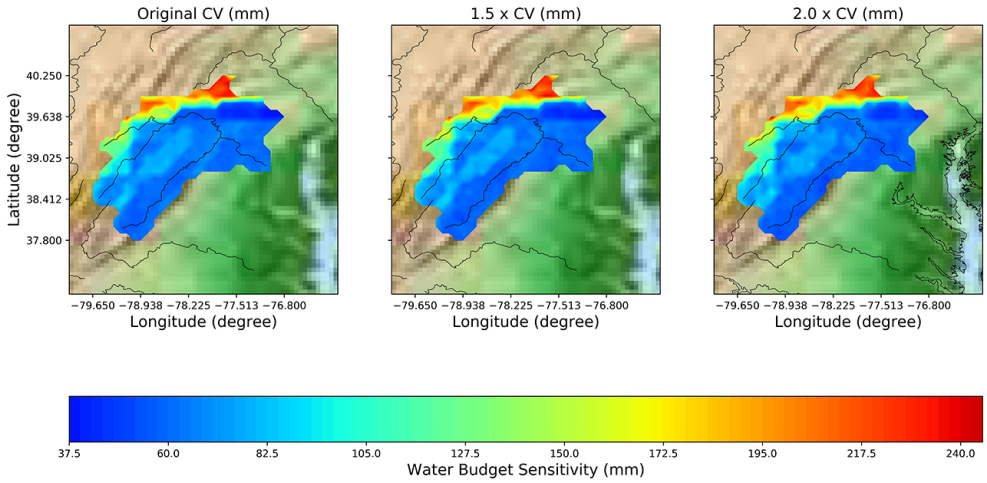


Figure 4.4: Standard Errors (SE) of the resampling distribution of the total water balance.

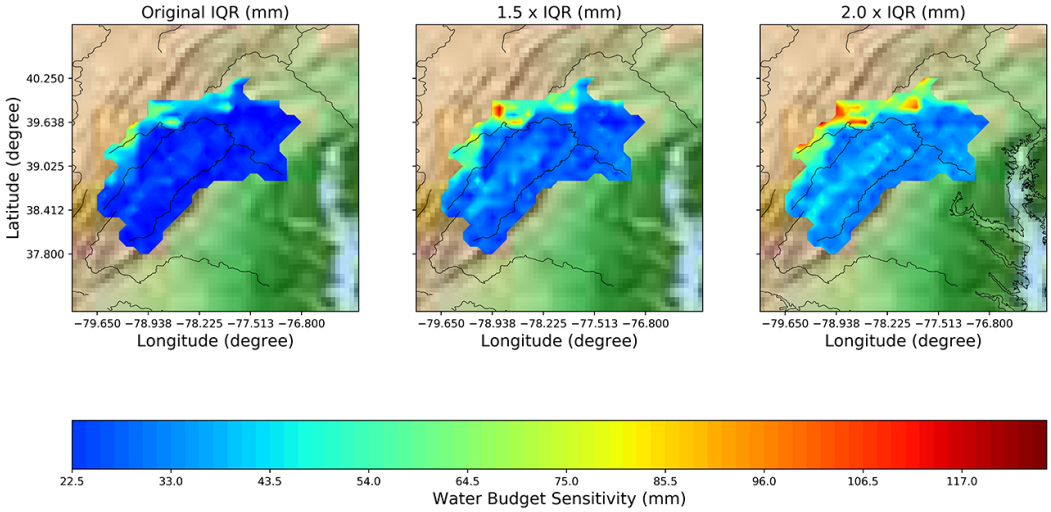


Figure 4.5: Interquartile Range (IQR) of the resampling distrubution of total water balance.

In Figure 4.6 and Table 4.1 we notice some bias in our estimates suggesting that our model assumptions have not been accurate. They do however follow the same general patterns from the observed data. We also see some interesting patterns when it comes to the variability of the estimates. For example, the April and May estimates seem to be most sensitive to the variability in the data, given by their significantly longer confidence intervals compared to the other months.

We also notice that higher the variability in our original data (higher values of CV in this case), the higher the variability of our estimates. While it might seem obvious, it must be noted that precipitation data from different sources are often assimilated to come up with a single estimate (our IMERG dataset is one such case), a process which can lead to an underestimation or overestimation of its dispersion. VIC will carry over these features in its

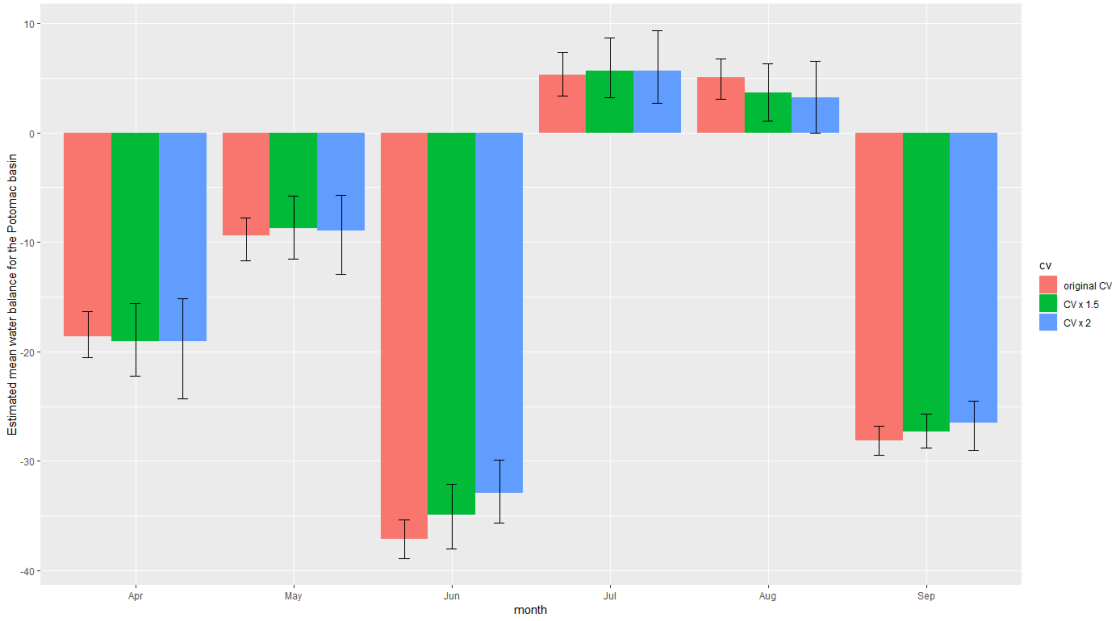


Figure 4.6: Mean and 95% empirical confidence intervals for monthly water budget.

Table 4.1: Lengths of the empirical 95% Confidence intervals.

CV	Apr	May	Jun	Jul	Aug	Sep
Original	4.25	3.93	3.54	3.98	3.68	2.65
CV x 1.5	6.69	5.71	5.92	5.44	5.27	3.10
CV x 2	9.17	7.28	5.77	6.64	6.64	4.48

output ensembles, which is motivation to choose a precipitation input that best reflects the systemic variability of the process.

5 Conclusions and Future Work

In the first half of this CyberTraining project, we were able to set up and run VIC, a full scale hydrologic model, on the distributed-memory cluster taki at UMBC. This required the installation of multiple tools, getting data from different sources and aligning them. We streamlined the entire process and established a framework which allows scalability. Next, we estimated the water budget components for the Potomac river basin using VIC for the months of April to September 2017. The results have been visualized at a monthly level as well as through a heatmap, i.e., both temporally and spatially. For a calibrated model these outputs can be used for post-hoc analyses.

In the second half of our project, we ran a preliminary sensitivity analysis for the precipitation data. The goal was to see if VIC's outputs are reflective of the dispersion in the input precipitation data. We were able to identify April and May as the months where the

water balance was most sensitive to variability, as well as sub-basins with the most sensitivity over the course of the 6 months of interest. However, some bias was present in our outputs, indicating that the precipitation data would be better modeled by a spatio-temporal distribution.

The next step for the first part of this project would be to calibrate the model using data for at least five years in duration. Of course, this presents data challenges, so we are looking into alternative methods to procure data so that a lot of the pre-processing can happen directly at source, saving us overhead on taki. A well calibrated model will improve the quality of the model output and will allow us to conduct volumetric analysis on the water budget components. For the calibration process we will first include a runoff routing model [11] in the VIC post-processing step to get river discharge (streamflow) as output. We will calibrate the model by comparing VIC streamflow with the streamflow measured by the United States Geological Survey (USGS) in the Potomac river channel and its tributaries using stream gauges. We also plan to time each step so we have a better idea of bottlenecks and how the runtimes might scale for different use cases.

A major shortcoming of our VIC simulations arose from an over-simplification of snow parameterization. We modeled snow to be a function of surface air temperature alone and not elevation. The Potomac basin includes the central Appalachian mountain range and Potomac highland (a part of Allegheny plateau) with varying elevation. We plan to run VIC with improved snow parameterization for more accurate water cycle representation in the basin.

Most of the knowledge base for VIC available to us was for VIC-4 and not the current version viz. VIC-5;[1] consequently, we used VIC-4 for the project. Parallelization in VIC-4 is achieved by manually running the model for different grid points across multiple processes, since each grid point is modeled independently of the others. The current version VIC-5 has a classic C driver which uses ASCII or binary I/O and is the most similar to what we used, and also a modern Image driver which uses netCDF for I/O and MPI for parallel processing. There are additional experimental Python and CESM drivers; all implementations are open source. Running the MPI-optimized VIC-5 would also be a way to test improvements in model runtimes.

For the sensitivity analysis, more comprehensive statistical modeling on the precipitation data is required. Using data for a longer duration would help in this case as well. We are also considering using the error variable that is provided by GPM-IMERG, which are errors from its internal data merging process. Better estimation tools in form of model calibration for VIC, spatio-temporal modeling in the sensitivity analysis, as well as parallelizing the data pre-processing and model runs can all be extended into further research topics.

Acknowledgments

This work is supported by the grant CyberTraining: DSE: Cross-Training of Researchers in Computing, Applied Mathematics and Atmospheric Sciences using Advanced Cyberinfras-

[1] VIC Release Notes. https://vic.readthedocs.io/en/master/Development/ReleaseNotes

tructure Resources from the National Science Foundation (grant no. OAC–1730250). The hardware in the UMBC High Performance Computing Facility (HPCF) is supported by the U.S. National Science Foundation through the MRI program (grant nos. CNS–0821258, CNS–1228778, and OAC–1726023) and the SCREMS program (grant no. DMS–0821311), with additional substantial support from the University of Maryland, Baltimore County (UMBC). See hpcf.umbc.edu for more information on HPCF and the projects using its resources. Co-authors Reetam Majumder and Carlos Barajas were supported as HPCF RAs. The VIC hydrologic model used in this study was developed at the University of Washington Department of Civil and Environmental Engineering Computational Hydrology group, and the model code was obtained by from https://github.com/UW-Hydro/VIC. The VIC model overview and input data information were obtained from https://arset.gsfc.nasa.gov/water/webinars/VIC18. NASA-SERVIR VIC training documentation and scripts developed by Kel Markert from https://github.com/KMarkert/servir-vic-training were used as background and for data pre-processing.

References

[1] M. Friedl and D. Sulla-Menashe. MCD12Q1 MODIS/Terra+Aqua Land Cover Type Yearly L3 Global 500m SIN Grid V006[Data set]. NASA EOSDIS Land Processes DAAC, 2019. doi: 10.5067/MODIS/MCD12Q1.006.

[2] Ronald Gelaro et al. The Modern-Era Retrospective Analysis for Research and Applications, Version 2 (MERRA-2). *Journal of Climate*, 30(14):5419–5454, 2017. doi: 10.1175/JCLI-D-16-0758.1.

[3] Joseph J Hamman, Bart Nijssen, Theodore J Bohn, Diana R Gergel, and Yixin Mao. The Variable Infiltration Capacity model version 5 (VIC-5): infrastructure improvements for new applications and reproducibility. *Geoscientific Model Development*, 11(8):3481–3496, 2018.

[4] G. Huffman. GPM IMERG Final Precipitation L3 Half Hourly 0.1 degree x 0.1 degree V05. Greenbelt, D, Goddard Earth Sciences Data and Information Services Center (GES DISC), 2017. Accessed: 04.25.2019, 10.5067/MODIS/MCD12Q1.006.

[5] George J. Huffman. Estimates of root-mean-square random error contained in finite sets of estimated precipitation. *Journal of Applied Meteorology*, 36:1191–1201, 1997.

[6] George J. Huffman, David T. Bolvin, Eric J. Nelkin, and Jackson Tan. Integrated MultisatellitE Retrievals for GPM (IMERG) Technical Documentation. Technical report, NASA GSFC, 2019.

[7] George J. Huffman, David T. Bolvin, Eric J. Nelkin, and David B. Wolff. The TRMM Multisatellite Precipitation Analysis (TMPA): Quasi-Global, Multiyear, Combined-Sensor Precipitation Estimates at Fine Scales. *Journal of Hydrometeorology*, 8:38–55, 2 2017.

[8] NASA JPL. NASA Shuttle Radar Topography Mission Global 1 arc second [Data set]. NASA EOSDIS Land Processes DAAC, 2013. doi: 10.5067/MEaSUREs/SRTM/SRTMGL1.003.

[9] Xu Liang, Dennis P Lettenmaier, Eric F Wood, and Stephen J Burges. A simple hydrologically based model of land surface water and energy fluxes for general circulation models. *Journal of Geophysical Research: Atmospheres*, 99(D7):14415–14428, 1994.

[10] Hongli Liu, Antoine Thiboult, Bryan Tolson, Franois Anctil, and Juliane Mai. Efficient treatment of climate data uncertainty in ensemble Kalman filter (EnKF) based on an existing historical climate ensemble dataset. *Journal of Hydrology*, 568, 11 2018.

[11] Dag Lohmann, Ralph Nolte-Holube, and Ehrhard Raschke. A large-scale horizontal routing model to be coupled to land surface parametrization schemes. *Tellus A*, 48(5):708–721, 1996.

[12] R. Myneni, Y. Knyazikhin, and T. Park. MCD15A2H MODIS/Terra+Aqua Leaf Area Index/FPAR 8-day L4 Global 500m SIN Grid V006 [Data set]. NASA EOSDIS Land Processes DAAC, 2015. doi: 10.5067/MODIS/MCD15A2H.006.

[13] Francesca Pianosi, Keith Beven, Jim Freer, Jim W. Hall, Jonathan Rougier, David B. Stephenson, and Thorsten Wagener. Sensitivity analysis of environmental models: A systematic review with practical workflow. *Environmental Modelling & Software*, 79:214 – 232, 2016.

[14] Sai. K. Popuri. *Prediction methods for semi-continuous data with applications in climate science*. PhD thesis, University of Maryland, Baltimore County, 2017.

[15] Andrea Saltelli, Marco Ratto, Terry Andres, Francesca Campolongo, Jessica Cariboni, Debora Gatelli, Michaela Saisana, and Stefano Tarantola. *Global sensitivity analysis: the primer*. John Wiley & Sons, 2008.

[16] Andrea Saltelli, Stefano Tarantola, Francesca Campolongo, and M Ratto. *Sensitivity Analysis in Practice: A Guide to Assessing Scientific Models*. John Wiley & Sons, 06 2004.

[17] C. Schaaf and Z. Wang. MCD43A1 MODIS/Terra+Aqua BRDF/Albedo Model Parameters Daily L3 Global - 500m V006[Data set]. NASA EOSDIS Land Processes DAAC, 2015. doi: 10.5067/MODIS/MCD43A1.006.

[18] W. R. Wieder, J. Boehnert, G. B. Bonan, and M. Langseth. Regridded Harmonized World Soil Database v1.2. Data set. Available on-line [http://daac.ornl.gov] from Oak Ridge National Laboratory Distributed Active Archive Center, Oak Ridge, Tennessee, USA, 2014. http://dx.doi.org/10.3334/ORNLDAAC/1247.