[M.L. 2017, Chp. 96, Sec. 2, Subd. 04h] Project Abstract
For the Period Ending June 30, 2020

PROJECT TITLE: Techniques for Water Storage Estimates in Central Minnesota
PROJECT MANAGER: John L. Nieber
AFFILIATION: University of Minnesota
MAILING ADDRESS: 1390 Eckles Ave.
CITY/STATE/ZIP: St. Paul, MN 55108
PHONE: 612-625-6724
E-MAIL: nieber@umn.edu
WEBSITE: http://bbe.umn.edu/people/faculty/johnnieber
FUNDING SOURCE: Environment and Natural Resources Trust Fund
LEGAL CITATION: M.L. 2017, Chp. 96, Sec. 2, Subd. 04h

APPROPRIATION AMOUNT: $250,000
AMOUNT SPENT: $250,000
AMOUNT REMAINING: $0

Sound bite of Project Outcomes and Results
A combination of ground-based monitoring and satellite data were used to quantify freshwater in Central Minnesota. Quantification of stored water is essential to improve water resources management and planning.

Overall Project Outcome and Results
Our freshwater resources reside in surface water bodies (ponds, wetlands, lakes, streams/rivers) and subsurface water reservoirs (soil and groundwater aquifers). Management of these freshwater resources has always been a challenge because we do not have a good idea of how much water is stored in these various entities. The objective of this project was to improve the methods for real-time quantification of the amount of water stored in these entities using existing ground-based measurement networks as well as satellite data. The study region stretched from St. Paul to Moorhead, and encompassed 17 HUC-8 watersheds. The study region has an area of about 53,000 km^3. We collected archived ground-based measurements including streamflows, observation wells, and lake levels for the period 2002-2015. We also acquired satellite data from the GRACE (Gravity Recovery and Climate Experiment), SMOS/SMAP, and Landsat satellites. The GRACE satellite provides data on the total water stored in the earth. The spatial resolution of the data used in this study was 100 km by 100 km. The SMOS/SMAP satellites provide a measure of the surface soil moisture over areas of about 36 km by 36 km. The Landsat satellite provides visual images at a scale of 30 m, and can be used for measuring the surface area of individual lakes; this surface area data can be used to estimate the volume of water stored in a given lake at a given moment in time. The project demonstrated that the variation in total water storage can be monitored by the GRACE satellite, and variations in lake storage can be monitored by the Landsat satellite. For the period 2002-2015 the estimates of time-averaged water storage is 1,500 km^3 for groundwater in the Quaternary (surficial) aquifer, 15 km^3 for lakes, 20 km^3 for soil moisture, and 1.5 km^3 for wetlands. The GRACE satellite became inoperable in late 2017, far exceeding the original planned life for the satellite. However, in May 2018 a new satellite, GRACE-FO (GRACE-Follow On) was launched and it now is providing the same information about total water storage. One of the outcomes of this project is a new research activity to test the utility of water storage information gained from the GRACE-FO satellite to forecast flooding and hydrological droughts in Minnesota.

Project Results Use and Dissemination

The project results have been presented at a number of different forums including the Minnesota Water Resources Conference (October 2019), the Water Resource Sciences Graduate Seminar at the University of Minnesota (September 2019), and the Western Regional Project 4188 Meeting in Las Vegas (January 2020). Two MSc theses were completed based on the work in the project, and the work of two Ph.D. students got started (one to finish in December 2020 and the other to finish in December 2021) based on work in the project.

A methodology for quantifying the volume of water in a lake based on the surface area of a lake was adapted from previous work and was tested during this project for the project study region. This tested methodology was then used in a separate LCCMR funded project in which the volumes of lakes across the State of Minnesota were estimated. This objective of this other project was to use remote sensing to quantify the water quality of lakes and the lake volume estimates were needed to examine lake processes affecting lake water quality.

A methodology was developed for quantifying the volume of water stored in the Quaternary aquifer spanning across the study region. The methodology uses observation well data and lake level data to map the water table across the region. This methodology will be shared with the MNDNR, but also it will also now be used in some immediate future research to assess the water table mapping in quantifying the forecasting of floods, and possibly in forecasting hydrologic droughts.

Environment and Natural Resources Trust Fund (ENRTF)
M.L. 2017 Work Plan, Final Report

Date of Report: September 2, 2020

Final Report

Date of Work Plan Approval: 06/07/2017

Project Completion Date: June 30, 2020

Does this submission include an amendment request? No__

PROJECT TITLE: Techniques for Water Storage Estimates in Central Minnesota

Project Manager: John L. Nieber

Organization: University of Minnesota

Mailing Address: 1390 Eckles Ave.

City/State/Zip Code: St. Paul, MN 55108

Telephone Number: (612)-625-6724

Email Address: nieber@umn.edu

Web Address: http://bbe.umn.edu/people/faculty/johnnieber

Location: 100 mile swath of area lying between the Twin Cities Metro Area and Moorhead, MN.

Total ENRTF Project Budget:		
	ENRTF Appropriation:	**$250,000**
	Amount Spent:	**$250,000**
	Balance:	**$ 0**

Legal Citation: M.L. 2017, Chp. 96, Sec. 2, Subd. 04h

Appropriation Language:
$250,000 the first year is from the trust fund to the Board of Regents of the University of Minnesota to improve water storage estimates in groundwater, soil moisture, streams, lakes, and wetlands through integration of satellite monitoring and ground-based measurements in central Minnesota. This appropriation is available until June 30, 2020, by which time the project must be completed and final products delivered.

I. PROJECT TITLE: Techniques for Water Storage Estimates in Central Minnesota

II. PROJECT STATEMENT:

Minnesota is known as a land of plentiful water – but nobody can tell us how much water there is. **This project will answer the question 'How much water is there?'** It will improve our ability to monitor and quantify the amount of water stored in groundwater aquifers, soils, lakes, wetlands, and streams throughout Minnesota. Knowledge of total water storage is essential to sustainable management and wise use of water resources throughout the state. For purposes of this proposal, water storage is defined as the total water volume at a single point in time in the groundwater aquifers, soil, and surface waters. This differs from water availability, which is a smaller volume, because the total volume stored cannot fully be extracted or used.

Water storage affects the availability of the water for human use (industry, irrigation, power production, domestic), and the availability of the water needed to support ecosystems throughout the state. Currently water storage in aquifers can be estimated using the sparse network of MNDNR observation wells, water storage in soils can be estimated from the very sparse network of soil moisture monitoring sites, and water storage in lakes and wetlands can be estimated from water level measurements at MNDNR/citizen monitoring sites. To our knowledge, to date none of these available data have been used to make estimates of total water storage throughout the state. In fact, to this day, we do not have an estimate of the total water present within Minnesota's borders or even to estimate the storage within a select region of the state.

We will improve the ability to monitor water storage by developing a methodology that joins data from remote sensing and ground-based measurements. Vast amounts of data are available from NASA satellites, but these are underutilized for Minnesota. For our project there are three satellites of particular interest. One is the GRACE satellite which provides data that can be used to quantify the change in storage of all water sources over large, multi-state size areas. A second is the SMAP satellite that provides data on the moisture stored in the soil over intermediate size areas. The third is the World-View3 satellite that provides high resolution images for outlining water levels in lakes and wetlands. Of course, ground-truth data are needed for proper interpretation of satellite-based data, and this is where the network of ground-based monitoring data is essential. The ground-based data sources include observation wells, meteorological stations, lake water levels, stream gages, surface topography, soil maps, and geological maps.

Within the scope of this project the methodology for water storage estimation and mapping will be conducted for a 100-mile-wide swath of area lying along a line between the Twin Cities Metro Area and Moorhead, MN.

The effort in this project should be compared to other ongoing efforts around the world to derive estimates of water storage on the Earth. There are several documented efforts by university researchers and government agency personnel within the U.S., Canada, and countries in Western Europe to derive estimates of storage of water within watersheds in groundwater and surface waters. Examples include the estimation of changes of water storage in the Central Valley of California, Eastern Canada, the Eastern U.S., the Middle East, the Indian Subcontinent, and Mongolia. These documented efforts show that the methods to be used within this project are viable and supported by the success of those other efforts.

III. OVERALL PROJECT STATUS UPDATES:

Project Status as of January 1, 2018:

1. Conducted evaluation of available hydrologic models to be used with the water storage analysis. Here we were interested in being able to consider water balances on a daily time scale, and also at a spatial scale conducive to the footprint of the GRACE total water storage satellite and the SMOS/SMAP soil

moisture satellites. Several models were considered and reviewed. These included the CLM (Community Land Model) model, the NOAA (National Oceanic and Atmospheric Administration) model, and the VIC (Variable Infiltration Capacity) model. Review of the literature showed all of these models to be quite similar in their abilities, but the CLM model appears to have a bit better groundwater recharge component. So, the CLM model was favored from that standpoint. Then in addition, we compared the soil moisture balance over a 10-year period for the Rum River watershed and found that the CLM model provided a better prediction of soil moisture data (1 meter depth) measured within that watershed. Finally, one member of our group (Dr. Griffis) was already heavily invested in using the CLM model so we selected that model for the remainder of our project.

2. The project study area included a 100-mile swath of land between Dakota County and Moorhead, MN. For this area we delineated seventeen (17) HUC-8 watersheds and an image of that area is shown in Figure 1.

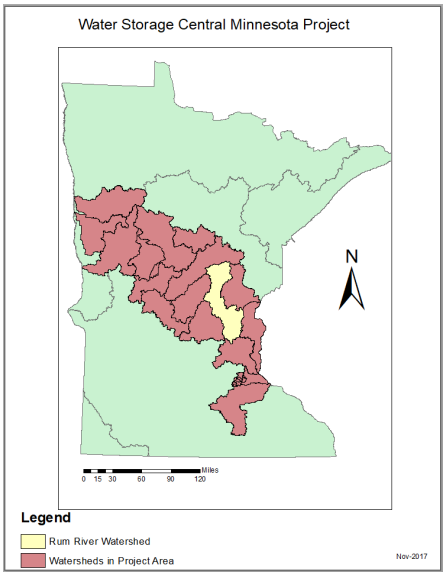


Figure 1. The study area is outlined and is composed of 17 HUC-8 watersheds. The watershed outlined in yellow is for the Rum River watershed. It is highlighted because that watershed was used for trial runs of our water storage estimation reported in the July 2018 update.

3. During this period, we were collecting water data for the study area. The data include streamflow data for 54 gaging stations, average daily precipitation for each of the watersheds, average evapotranspiration for the watersheds, GIS layers for surface topography, soil classifications, and surficial geology, well logs for individual wells located within the study area, monitored water level sin monitoring wells within the study area, and data for the GRACE satellite, SMOS/SMAP satellites, and data for LandSat and WorldView satellites. Except for the groundwater monitoring well data, these data have been processed and transferred to a DVD for distribution to interested parties. We are using these data for our water storage analyses. The groundwater monitoring data are still being processed and once complete those data will be transferred to the DVD for distribution.

Project Status as of July 1, 2018:

1. Soil moisture data from a station located in the Rum River watershed have been used as ground truth for testing the three alternative water balance models, the CLM, NOAA and VIC models. The

comparisons have led us to choose the CLM model for this project. In addition to giving a better prediction of the observed data, one or our team members, Dr. Timothy Griffis has already been using the CLM in his related work.

2. Baseflow recession analysis using the method of Brutsaert and Nieber (1977) has been applied to all of the seventeen watersheds. This analysis yields a residence time parameter that is needed to be able to quantify water storage change using baseflow recession data (Brutsaert, 2008). This storage change analysis has been applied to the Rum River watershed for the period 2001 – 2016. A similar analysis will be conducted for the remaining sixteen watersheds. The results of this analysis will be presented in our January 1, 2019 update. This work is the subject of the M.S. thesis by Mr. Xiang Li, a graduate student in the WRS graduate program.
3. Methods have been developed to quantify the storage of water in lakes and wetlands. For lakes we are using a regression method in which we use hydrography data available for about 900 lakes to establish a relation between lake surface area and water volume. The idea is that we can measure lake surface area from satellite data, and then with the regression relation estimate the volume stored in lakes. A more sophisticated method applies an equation that relates the topography in the buffer around a lake to the stored volume. This second procedure is the subject of a M.S. thesis study by Ms. Chelsea Delaney, graduate student in the WRS graduate program. She will also apply a fractal analysis method, in which surface topography along with fractal analysis procedures will be used to estimate volume. For wetlands we are making use of the National Wetland Inventory (currently using the older inventory since the new one is only being completed now). We are attempting to apply some scale analysis to quantify storage within wetlands.
4. We have applied these methods in a test trial to the Rum River watershed. Our initial estimates of water storage in the watershed are: lakes – 4.86 km^3 , mostly in Lake Millacs (4.3 km^3). 0.1 km^3 in wetlands, 33.4 km^3 in groundwater, and 5.8 km^3 in the unsaturated zone overlying the groundwater. Improved estimates for water storage in the Rum River watershed, as well as the other sixteen watersheds will be completed by September 30, 2018.
5. In a new development we have found that our estimations of lake volume are of significant value to current work being conducted by Dr. Jacques Finlay on a current LCCMR project. In that project they are estimating the total amount (mass) of dissolved organic carbon (DOM) in Minnesota Lakes. They are using satellite data to quantify the concentration of DOM in lakes, but the volume of the lakes are necessary to convert the concentration estimate to mass of DOM. We are thereby collaborating with Dr. Finlay to provide volume estimates for lake across the entire state (not just within our study area), and expect to have this completed by mid-October. Results of their work on DOM mass distribution in lakes across the state has been proposed for presentation at the 2018 Fall meeting of the American Geophysical Union.

Brutsaert, W., 2008. Long-term groundwater storage trends estimated from streamflow records: climatic perspective. Water Resour. Res. 44: W02409. doi:10.1029/2007WR006518.

Brutsaert, W. and J.L. Nieber, 1977. Regionalized drought flow hydrographs from a mature glaciated plateau. Water Resour. Res. 13(3):638-643.

Nieber, J.L., 2018. Minnesota: How much water is there; how is it changing? Open Rivers; Rethinking Water, Place and Community, Issue 10, Spring 2018.

Project Status as of January 1, 2019:

1. We have completed the analysis of lake volumes in lakes within the study regions using two different methods. The methods both provide a relation between lake volume and lake surface area. One method is based on regression analysis using some 900 lakes in the region for which we have quantified lake volume and lake surface area. The other method applies a topographic surface scaling methodology to relate lake surface area to lake volume. This second method does not require lake volume-lake surface area data to develop the predictions, but it is necessary to have such data to verify the output of the method. The two methods are in quite good agreement.
2. Xiang Li continues to conduct analysis of baseflow recession flows as a way of estimating water storage change. He already had much of this completed within the previous period but he continues to refine his analysis as this work is the subject of his M.Sc. thesis.
3. The estimation of groundwater storage within the study region. They have acquired geological data and well data for the region and are using interpolation methods to map water table levels across the study region. At this point they are able to make animated videos of the change in water levels with time during the period of interest (2002 – 2015). Next steps to get water storage will be to incorporate the porosity of geological material. From this mapping we will also be able to quantify the variation in thickness of the unsaturated zone and from that be able to quantify the dynamics of water storage in the unsaturated zone.

Mr. Francisco Lahoud is working with data from the GRACE satellite to quantify changes in water storage at the scale of individual watersheds. We have 17 watersheds in our study region, and the satellite data are being used to track water storage change. At issue is to be able to confirm that the satellite data is really giving water storage change, and not something else. To confirm this, we are working with water balance calculations using precipitation data, evapotranspiration estimates, and streamflow data to calculate water storage changes. Our estimates of precipitation and streamflow are very good, but the evapotranspiration estimates seem to be a problem right now. For evapotranspiration estimation we are using various sources of information (e.g., from NOAA, USGS, University of Minnesota) and attempting to find the method that gives the most reasonable results.

Amendment Request as of June 5, 2019:

I am requesting an extension for this ENRTF funded project. The project was written into statute as running from July 1, 2017 to June 30, 2020, and so the funds for the project need to be spent by June 30, 2020. It has this end date because originally, we planned for a 3-year project, but then when project funds were reduced, we proposed to finish by June 30, 2019. However, due to some circumstances we have now decided we would like to work on the original timeline of completing all project requirements by June 30, 2020. Justifications are given below.

1. The USGS counterpart of the project was not able to get started on the project until October 2017 due to other responsibilities. Secondly, the government shutdown in late fall, 2018-January 2019 stopped all work being conducted by the USGS counterpart. Extending the completion date to June 30, 2020 will allow the USGS counterpart to continue to improve our estimates of water storage based on water well and lake level data.
2. While we have been working on the modeling aspect of the project throughout the duration of the project, there is still significant effort to complete that work. We started the project working with the Community Land Model, but we have later decided to work with the HSPF (Hydrologic System Program

Fortran) model, partly because the MPCA has contracted the development of calibrated HSPF models for all of the watersheds in our project area, thus making it convenient to use that model. We currently have a new graduate student who is working with the HSPF model and the MPCA (Chuck Regan) to get the data for the calibrated models for our study area watersheds. The student is currently on a fellowship and will be conducting the modeling work during this summer (2019), but in the fall of 2019 he will work on funds from the ENRTF funded project. I expect this work will be carried out mostly in this summer and fall but would like to have the flexibility to complete all requirements by June 30, 2020.

Budget: Currently the balance at the University of Minnesota is shown below. I do not know the balance at the USGS subcontract, but the USGS counterpart has indicated that funds are available still to cover his efforts through June 30, 2020.

I am requesting that that $3,000 originally budgeted for project travel be allowed to be used for salaries. We had budgeted the $3,000 in case we might need to visit field sites during the project, but we have found that to be unnecessary, and now find the funds would be of help in our analyses and modeling.

We request a change in scheduling for project activities. The project activities to be delayed include: 1) the enhancement of the methods for quantifying baseline water storage and temporal change in water storage to be completed by January 1, 2020; 2) the validation of the methods used to quantify water storage change based on coupled field measurements, modeling, and satellite measurements, to be completed by March 31, 2020; and 3) completion of the project final report.

Amendment Approved by LCCMR 6/19/2019.

Project Status as of July 1, 2019:

1. Although we had completed the analysis of lake volumes in lakes within the study regions as of January 1, 2019, we continue to test one of the two methods as it is the subject of M.Sc. student Chelsea Delaney. Chelsea will graduate in September 2019 and at that time all improvements in lake volume estimation will be complete. A map of lake volume estimates by HUC-8 watershed is illustrated in Figure A.1. The total lake volume is about 1,600,000 ha-m (4,200 billion gallons). An amazing feature of this is that Lake Mille Lacs composes about 25% of the total lake volume in the study region. This map represents the estimation of lake volume for a given date in the 2002 – 2015 period. We have annual estimates of lake volume.
2. Estimation of volume of water in wetlands in the study region is underway and as of the report date the volume is about 180,000 ha-m of water (475 billion gallons). So the wetland volume is about 10% that of the lakes in the region. This work is being conducted by Chelsea Delaney. The estimation of wetland water volume will be completed by September 30, 2019, the date when Chelsea will complete her M.Sc.
3. Xiang Li has completed his analysis of baseflow recession flows for the seventeen HUC-8 watershed, and has found a good regionalization of baseflow parameters for watersheds in the region. The results for that analysis are being used in another analysis that predicts the change in groundwater storage in the project region. as a way of estimating water storage change. The work is the subject of his M.Sc. thesis. He will defend his thesis before September 2019.
4. The estimation of groundwater storage within the study region is based on the method described in the January 1, 2019 report. The estimates have been completed but there is still some revising to refine the

way of representing the volume of groundwater associated with interactions with lakes, streams and wetlands (we call this active groundwater). A map of the estimation of groundwater storage by HUC-8 watershed in the study region is illustrated in Figure A.2. The estimated volume of active groundwater is about 690,000 ha-m (1,813 billion gallons). This estimation is about 43% of the estimated lake volume in the region. The estimated volume of groundwater within the quaternary aquifer (this includes the active groundwater volume) is 1,620,000,000 ha-m (427,858 billion gallons) is mapped in Figure A.3. Our study has produced annual estimates of groundwater storage. The maps shown in Figures A.2 and A.3 are for the year 2015.

5. Based on the mapping of the groundwater table we will also be able to quantify the variation in thickness of the unsaturated zone and from that be able to quantify the dynamics of water storage in the unsaturated zone. The estimation of water storage in the unsaturated zone is ongoing and will be completed by September 30, 2019.
6. Water storage changes in the study region as estimated by the GRACE satellite signal (specifically, the Mascon data) and the estimated water storage based on the water table mapping across the region are compared in Figure A.4, covering the period from 2004 to 2015. There appears to be good agreement between the two plots of water storage change. We will note here that there are eight GRACE products available to us, and the Mascon product represented in the figure is just one of those products. We will be testing the other products as well during the next several months.
7. We have started to use the HSPF (Hydrologic Simulation Program Fortran) is being used for simulation of the components of the hydrologic cycle with each of the HUC-8 watershed. We selected the HSPF model because the MPCA has developed calibrated HSPF models for most of the HUC-8 watersheds in the state, and we have access to those models through Dr. Chuck Regan at the MPCA. To date we have been working with the HSPF models for the Rum River and for the Wild Rice River. The model provides a consistent way to handle the water balance components in these watersheds and will be comparing the HSPF simulations to the GRACE satellite data. Before selecting the HSPF model we had been conducted the water balances using a simple 'checkbook' method, but that method did not allow us to separate out the lake volume, soil moisture, or groundwater volumes from each other. The HSPF model allows us make this separation in a way such that it is consistent with available meteorological data. The work with the HSPF model will continue into the period from now until March 2020.
8. The project website development began in February 2019 and we now have a project website under construction. The website includes visuals of maps showing the distribution of water storage in lakes and groundwater, and also plots of the variation of water storage for the period 2002 – 2015. The website now includes files of some of the data used for the project and these are available for download to interested users. Much more will be added to the website as it continues to be developed. Additions to the website occur about every other week.

Project Status as of January 1, 2020:

1. Completed analysis of lake volumes in lakes within the study region. The volume of water in lakes (about 43,000 lakes total in the study region), is 15 km^3 , or an equivalent water depth of 280 mm across the study region (53,000 km^2 area). The results are described in detail in the MSc. Thesis of Ms. Chelsea Delaney. She officially graduated in December 2019. A copy of the thesis is given in Appendix E to this report. Note that although this report contains information on the lake volume estimates for an instant in time, we have annual estimates of lake volume as well.

2. Estimation of volume of water in wetlands in the study region was completed by Chelsea Delaney and the estimated volume is 1.5 km^3 of water (28 mm equivalent water depth). So the wetland volume is about 10% that of the lakes in the region.
3. Xiang Li has completed his analysis of baseflow recession flows for the seventeen HUC-8 watershed. The result is an estimate of mean groundwater drainage time for each watershed. He found a good relationship between the mean drainage time and geomorphic and hydraulic characteristics of the quaternary aquifer for each the watersheds. This relationship can then be used to determine the mean groundwater drainage time for ungagged watersheds in the region. The derived mean groundwater drainage times were used in an analysis to predict the change in groundwater storage for each of the HUC-8 watersheds, and the results were compared to storage change estimates derived from pointwise measurements. The comparisons were found to be favorable. The details of this work are given in the M.Sc. thesis for Mr. Li. A copy of his completed thesis is given in Appendix F.
4. Estimation of water storage in groundwater within the study region was limited to just the quaternary aquifer. The total groundwater storage was estimated to be 1,500 km^3 , with 33 km^3 being the portion of groundwater we defined as being 'active' groundwater. Active groundwater is defined as being that groundwater storage that directly interacts with surface water (lakes, wetlands, and streams). Details of the methods for estimation of water volume have been given in previous status report sections. Estimates of the uncertainty of the groundwater volumes are still being derived and will be given in the final report.
5. An estimate of water storage in the unsaturated zone is now complete, and the estimated volume is 20 km^3, or an equivalent water depth of 373 mm. The estimate of soil moisture storage was derived using the mapping of water table and the land surface elevation across the study region. It was assumed that the soil moisture content within the unsaturated zone (between the land surface and the water table) was equal to field capacity for the texture of the soil.
6. Comparisons between water storage changes derived by the GRACE satellite data (specifically, the Mascon data) and those from pointwise data as well as from water balance calculations are found to be favorable. The comparisons cover the period from 2004 to 2015.
7. We are continuing to apply the HSPF (Hydrologic Simulation Program Fortran) model for simulation of the components of the hydrologic cycle within each of the HUC-8 watersheds. The objective is to determine how well the HSPF model can match the change in total water storage derived from the GRACE satellite data. The HSPF model provides a consistent way to handle the water balance components in these watersheds, and we will be comparing the HSPF simulations to the GRACE satellite data. Before selecting the HSPF model we had been conducting the water balances using a simple 'checkbook' method for a lumped model representation of the watershed, but that method did not allow us to separate out the lake volume, soil moisture, or groundwater volumes from each other. The HSPF model allows us make this separation in a way such that it is consistent with available meteorological data. The work with the HSPF model will continue into the period from now until June 2020. This work is also being directed to determine whether the combination of the satellite data and the model analysis can be used to improve the forecasting of floods and hydrological droughts.

Project Status as of July 1, 2020:

1. Mr. Xiang Li completed his MSc. Thesis in June. A copy of his thesis is presented in Appendix F. His thesis involved the study of baseflow discharge in each of the 17 HUC-8 watersheds, and regionalization of the baseflow recession parameters for the associated groundwater systems. A good correlation (R^2=0.59)

between measureable watershed characteristics and the baseflow recession response characteristics was found. The derived baseflow recession characteristic (mean travel time) for each watershed was then used in a groundwater storage change analysis for the period 2002 – 2015. As a reference, water table mapping (described below) was used to quantify annual groundwater storage change (in the Quaternary aquifer) for each watershed for that same time period. This tests the question as to whether baseflow discharge recession data alone are sufficient to provide estimates of water storage change in watersheds. The tests showed that for all but three of the seventeen HUC-8 watersheds the agreement between the reference storage change and the change estimated from baseflow recession discharge was quite good. The poorer agreement for the three watersheds was attributed to the poor quality of data during the critical period of the year (winter) when the characteristic baseflow discharge values are selected.

2. The analysis for generating annual groundwater storage change using ground-based observations has been completed. This is for the groundwater storage in the Quaternary aquifer. The analysis for quantifying a reference water storage in the Quaternary aquifer was completed previously, showing that the estimated volume for the entire study area is about 1,500 km^3. The methodology used water levels in observations wells (about 700 in the region), static water levels derived from the County Well Index, and water levels in lakes and streams/rivers (the water table was assumed to be connected to surface waters in the region). The analysis required a very detailed interpolation of the water table across the entire landscape area (53,000 km^2) on a 30 m grid. Several state-of-art methods were applied for the interpolation. The result is a model of the interpolated water table for quantifying the water table map on an annual basis. The annual change in water storage in the Quaternary aquifer is then determined based on this interpolated map. As mentioned above, this change information was used as a reference for comparison of the storage change analysis based on the baseflow recession discharge. A detailed report about this water table analysis is presented in Appendix G.
3. The last task of the project was to apply a physically based watershed model to the watersheds, with the objective to develop a downscaling satellite data that yield storage estimates at too large a spatial scale. While we had previously shown results for water balance calculations for full watersheds using a lumped water balance model, we also wish to know the distribution of water storage within a watershed. The lumped water balance model does not provide this spatial distribution of the storage. For this task we selected to use the HSPF model (Hydrologic System Program Fortran). The model is a semi-distributed physically-based model. Details of reasons for selecting this model were given presented in previous reports. For completion of this project objective it was considered sufficient to show the value of the HSPF model in water storage modeling by modeling only one of the HUC-8 watersheds. For this case the model was applied to the Rum River watershed, as that watershed was analyzed in detail for the lumped water balance calculations outlined in previous reporting. A report on the modeling of the Rum River watershed water storage and watershed discharge with the HSPF model is presented in Appendix H. The output from the HSPF model provides information on the total water storage, the spatial (horizontally and vertically) distribution of storage, and the associated watershed discharge. An example of the spatial distribution of water storage for the Rum River watershed is shown in Figure 2 for January 1995 and January 2005. The modeling has shown the relationship between the total water storage and the watershed discharge. The relationship is not unique (it is hysteretic as seen for 1995 in Figure 3), but it is still useful in terms of being able to forecast flow (normal flow, flood flows, drought flows). The non-unique relation between total water storage and discharge is shown in Figure 4. This demonstrates that with monitoring of water storage, by satellite for instance, it should then be possible to forecast flows from the remotely sensed storage measurements. Carry-forward work that is now ongoing beyond this

project is to relate the watershed discharge to the spatial distribution of water storage (rather than the total storage in the watershed). This effort is ongoing by applying the HSPF model along with machine learning methods (methods from data science) within the scope of National Science Foundation grant. We anticipate this approach will allow the use of large-scale data available from satellites (GRACE-CO, SMAP, and Landsat) to be downscaled for more accurate forecasting, with applications to flood forecasting and drought forecasting.

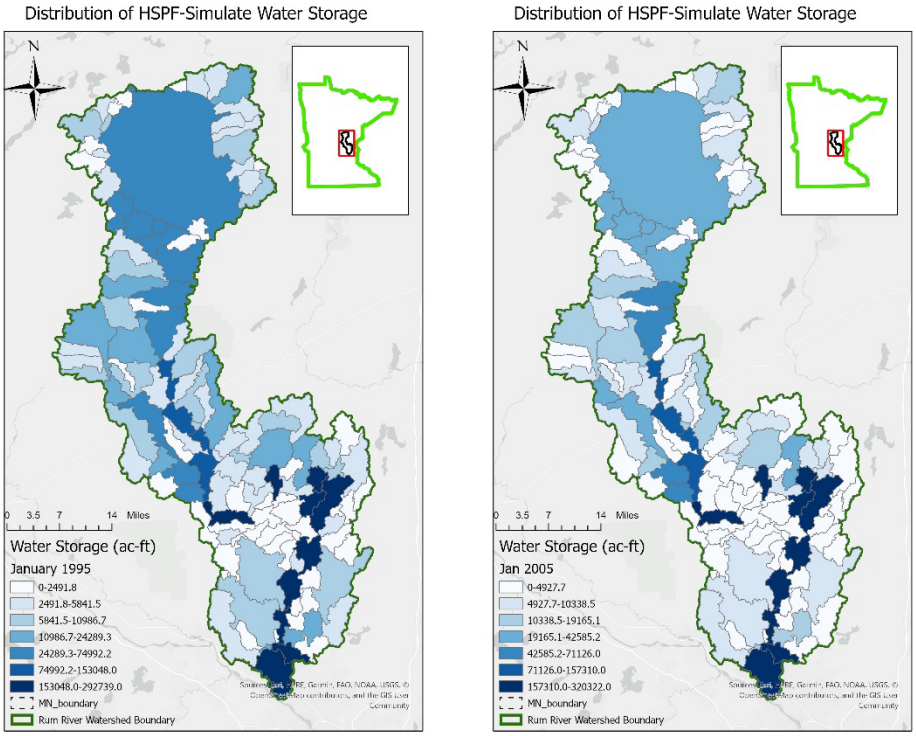


Figure 2. Distribution of water storage in the Rum River on two dates, Jan. 1995 and Jan. 2005.

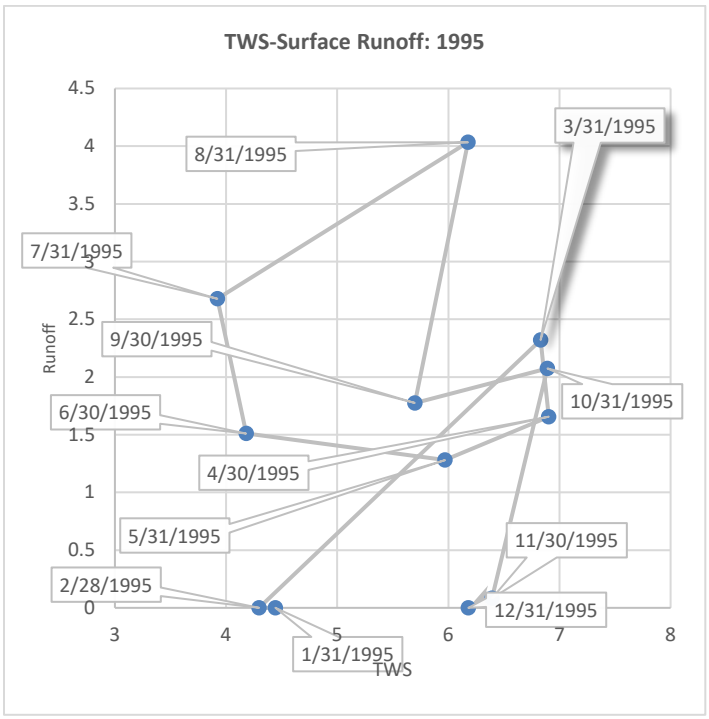


Figure 3. Hysteresis in the storage-discharge relation for the Rum River for 1995.

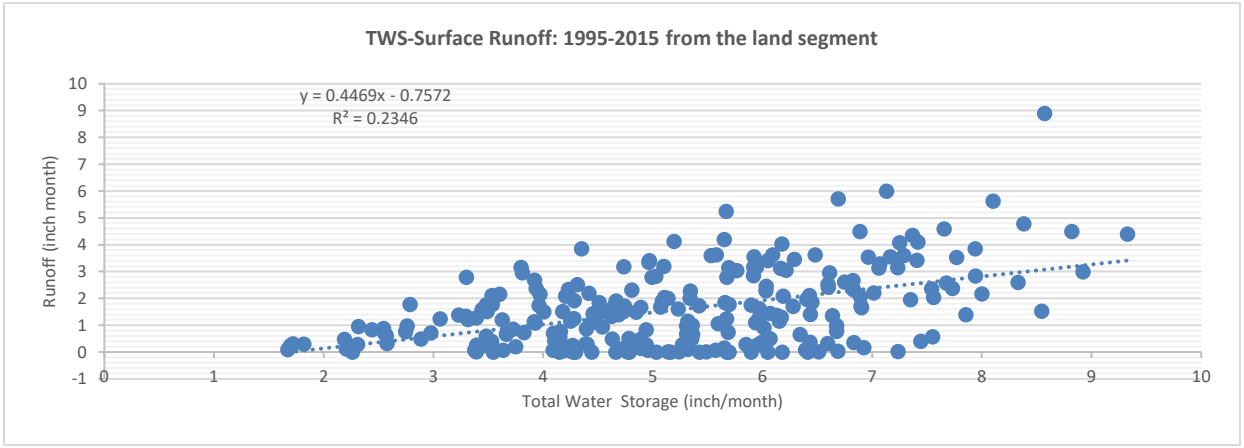


Figure 4. The relation between total water storage and watershed discharge for the Rum River. This is for the period of simulation, 1995 – 2015.

Overall Project Outcomes and Results:

Our freshwater resources reside in surface water bodies (ponds, wetlands, lakes, streams/rivers) and subsurface water reservoirs (soil and groundwater aquifers). Management of these freshwater resources has always been a challenge because we do not have a good idea of how much water is stored in these various entities. The objective of this project was to improve the methods for real-time quantification of the amount of water stored in these entities using existing ground-based measurement networks as well as satellite data. The study region stretched from St. Paul to Moorhead, and encompassed 17 HUC-8 watersheds. The study region has an area of about 53,000 km^3. We collected archived ground-based measurements including streamflows, observation wells, and lake levels for the period 2002-2015. We also acquired satellite data from the GRACE (Gravity Recovery and Climate Experiment), SMOS/SMAP, and Landsat satellites. The GRACE satellite provides data on

the total water stored in the earth. The spatial resolution of the data used in this study was 100 km by 100 km. The SMOS/SMAP satellites provide a measure of the surface soil moisture over areas of about 36 km by 36 km. The Landsat satellite provides visual images at a scale of 30 m and can be used for measuring the surface area of individual lakes; this surface area data can be used to estimate the volume of water stored in a given lake at a given moment in time. The project demonstrated that the variation in total water storage can be monitored by the GRACE satellite, and variations in lake storage can be monitored by the Landsat satellite. For the period 2002-2015 the estimates of time-averaged water storage is 1,500 km^3 for groundwater in the Quaternary (surficial) aquifer, 15 km^3 for lakes, 20 km^3 for soil moisture, and 1.5 km^3 for wetlands. The GRACE satellite became inoperable in late 2017, far exceeding the original planned life for the satellite. However, in May 2018 a new satellite, GRACE-FO (GRACE-Follow On) was launched and it now is providing the same information about total water storage. One of the outcomes of this project is a new research activity to test the utility of water storage information gained from the GRACE-FO satellite to forecast flooding and hydrological droughts in Minnesota.

IV. PROJECT ACTIVITIES AND OUTCOMES:

ACTIVITY 1: Acquire archived data and select hydrologic models.
Description: Hydrologic and meteorological data will be acquired from archived records available from main sources including the Minnesota DNR (MNDNR), Minnesota Pollution Control Agency (MPCA), the Minnesota Department of Health (MDH), the Minnesota Department of Agriculture (MDA), the National Weather Service, the Minnesota Geological Survey (MGS), the State Climatology Office, and the U.S. Geological Survey (USGS). Data will include gauged streamflows, lake water levels, wetland water levels, soil moisture, groundwater levels, and meteorological variables. These data will be used in Activity 2 in establishing the reference levels of water storage for the years 2002 and 2015. We will also conduct an exhaustive search of reports involving analyses of water storage estimates in the area of study.

Within the project we will utilize satellite data in concert with ground-based data and water balance modeling to quantify the changes in water storage over space and time. The satellite data appropriate for this include the GRACE (Gravity Recovery and Climate Experiment) satellite, the SMOS (Soil Moisture and Ocean Salinity) satellite, the SMAP (Soil Moisture Active Passive) satellite, and the World View3 satellite. The satellite data will be acquired for all the area of study for the period 2002-2015 (the GRACE satellite was launched in 2002). These data will be processed, stored, and made ready for use in Activity 3.

A unique feature of this project is that data from different sources will be used to derive the estimates of water storage and estimates of changes in water storage. The data to be used represent different time scales and spatial scales. To combine all the data sources and make the most use of them for estimating water storage an appropriate hydrologic model will be used to fuse the information contained in the data. We will conduct a review of available models and will choose the one most appropriate for this project.

Summary Budget Information for Activity 1:

ENRTF Budget: $ 43,500
Amount Spent: $ 43,500
Balance: $ 0

Outcome	Completion Date
1. Complete set of hydrologic, soil, geologic, groundwater level, meteorological, topographic data, and satellite data prepared for project use. All data will be archived and available on DVD. While used for this project, the summarized data will be useful to state and federal agencies, and to consultancies.	3/31/18
2. Model selected. Documentation of selection procedures and details of tests completed in making the model selection. The selected model will also be of use to consultancies, and for class instruction in hydrology courses.	6/30/18

Activity Status as of January 1, 2018:

1. Delineated seventeen HUC-8 watersheds for the study area. See Figure 1 for a map of the watersheds.
2. This was an active period of data acquisition from different sources, and processing of the data for our needs. The groundwater data acquisition was slow in starting because Jared Trost was delayed in starting his part of the project. However, activity on groundwater picked up in late October and the data acquisition moved along fine.
3. Models identified for use in the project included the CLM, NOAA and the VIC models. These were chosen for evaluation because the models work on a spatial scale conducive to the scale of measurements by the satellites, and also because most agencies and other researchers working with the same satellites use these models. Our criterion for selection is completeness of hydrologic process analysis, and accuracy in predicting soil moisture balances.

Activity Status as of July 1, 2018:

1. The acquired and processed data we transferred to a DVD disk for distribution. The only missing data are the groundwater monitoring well data and some of the aquifer data. These data will be added to the DVD disk once processing is complete.
2. Algorithms have been developed to estimate water storage and water storage change in the watersheds, and these algorithms have been applied to the Rum River watershed, one of the 17 watersheds delineated for the project study area. Estimates for the volumes stored in the lakes, wetlands, soil and groundwater within the watershed are given in the overview report above.
3. Among the three hydrologic models considered for use in the project, the CLM model has been selected as the best of the three. The model is currently being used by our team on the computers at the Minnesota Supercomputer Institute. The report on the application of the model to our watersheds has not been completed yet but will be completed by January 1, 2019. We have also decided to apply the HSPF (Hydrologic System Program Fortran) to the watersheds. This model is currently being used by the MPCA for TMDL studies and the model has been set up and calibrated for most or all of the major watersheds (81 of them) in the state. The data files for model setup are available from the MPCA and we will be acquiring those input data file. Currently we are setting up the HSPF for use in the modeling and acquiring the input files (from the MPCA) for the Rum River watershed.

Activity Status as of January 1, 2019:

1. We have acquired water well data for hundreds of monitored wells located across the study region. Currently we are working with those data to map the dynamics of water levels across the study region. Next, we will be coordinating aquifer property data (in particular porosity) associated with the wells. As we work with these data, we will eventually store those data into the DVD prepared for the July 1, 2018 report. That DVD will then be available upon request.

Activity Status as of July 1, 2019:

1. Although we continue to come across new data available for our study region, we have essentially completed the data acquisition for the project. These data are available on DVD and also the data are being loaded to the project website, so they are available for use by others.

Activity Status as of January 1, 2020:

N/A

Final Report Summary:

1. All project data were made available on DVD, but also available on the project website.
2. The HSPF model was selected for the watershed modeling platform. This model is part of the suite of models associated with the EPA BASINS model toolkit. We have applied it to watersheds within the study region. Model data for all 17 HUC-8 watershed within the study region are available from the MPCA website, and detailed information on the model inputs as applied within this project are available from the project website.

ACTIVITY 2: Develop estimates of baseline water storage for the study region.
Description: The data fusion methodology developed in this project is not intended to provide estimates of the absolute storage of water but instead to quantify the temporal change in water storage relative to a reference baseline. Therefore, a baseline of the water storage will be determined using available pointwise ground-based data. The baseline estimates will be derived for two dates, October 1, 2002 and October 1, 2015. The data to be used for this analysis will be acquired and processed in Activity 1. The storage estimates for these two reference points will then be used in Activity 3 to test the ability of the developed methodology to estimate the change in stored water between those two dates. Brief descriptions of the approach for estimation of storage for the three domains, groundwater, soil, and surface water are given in the following paragraphs.

Groundwater storage. Estimation of groundwater storage will be based on water level measurements from MNDNR monitoring wells and also from monitoring wells operated by others. The MNDNR monitoring well network is the best available data because it is consistently monitored over time. Using the water level measurements at a given location, along with the porosity of the geologic formation at the location the volume stored is simply the product of the porosity times the saturated depth of the formation. The geological maps available from the MGS will be invaluable to deriving the depth and porosity information for the geological features. These estimates will be at the point of the monitoring well and it will be necessary to integrate all the point measurements to get a total volume stored within an area. To achieve this methods of interpolation between points will be adopted, specifically methods that were developed in the mining industry, and applied in the soil science and groundwater hydrology disciplines.

Soil moisture storage. In estimating water storage in the soil profile, it will be assumed that the depth of the soil profile is 5 feet. Where available, soil moisture measurements at point locations within the area of study will be used to estimate the average water stored in the soil profile. As with the groundwater storage estimation, the point estimates (measurements) of soil water storage will be interpolated between points to derive area average values so that total water stored in the landscape can be estimated.

Surface water storage. The storage of water on the land surface includes the water in lakes, wetlands and rivers. To estimate the storage it is necessary to have the water surface elevation within the given water body (lake, wetland, river), and also the bottom topography (or morphometry) of the water body. The morphometry of the lakes and wetlands within the study area will be acquired/derived within Activity 1. Water level data are available from the MNDNR and municipalities for selected lakes and wetlands within the study area. Estimating the volume of water stored in rivers and streams will be made based on the use of stream gauging water surface elevations, and available (directly measured or else derived from LIDAR) channel cross-section data.

Summary Budget Information for Activity 2:

ENRTF Budget: $ 79,000
Amount Spent: $ 79,000
Balance: $ 0

Outcome	Completion Date
1. Estimates of baseline water storage for 2002 and 2015 for specific locations (i.e., point estimates) within the study area. The location specific estimates will be useful to the MNDNR in evaluating the effectiveness of current water management procedures.	9/30/18
2. Maps showing distribution of water storage estimates for surface water, soil moisture, and groundwater across the study area for 2002 and 2015. The mapped area-specific estimates of storage will be useful to the MNDNR in evaluating the effectiveness of current water management procedures.	12/31/18

Activity Status as of January 1, 2018:

N/A.

Activity Status as of July 1, 2018:

1. Estimates for the volumes stored in the lakes, wetlands, soil and groundwater within the Rum River watershed are given in the overview report above. These estimates are for a single period of time. We have also estimated lake volumes and wetland storage volumes for all of the 17 watersheds; however, we are working on verifying those estimates. Estimates for 2002 and 2015 will be completed by September 30, 2019 and included in the January 1, 2019 report.

Activity Status as of January 1, 2019:

1. There are two methods developed for estimating lake volume. The one method developed by Kerry Holmberg is based on regression analysis using observed lake volumes and lake surface area. The model then relates lake volume to lake surface area. Lake surface area can be quantified with aerial remote sensing platforms, so this is the way of keeping track of lake volumes. This method is pretty much completely done at this point and we have estimates of total lake volume for the study region. The second method is being implemented by graduate student Chelsea Delaney, and it is based on land surface topography and ideas about scaling theory. The method is developing well and Chelsea has been able to estimate the total volume of water in lakes within the study region. Her estimate is quite close to the estimate obtained by Kerry Holmberg. Chelsea also assisted an LCCMR project by Dr. Jacques Finley to quantify lake volumes across the State of Minnesota using the methodology she developed. Results of that work were reported at the 2018 Fall meeting of the American Geophysical Union. Chelsea is continuing to refine her analysis as it is the subject of her M.Sc. thesis. We are hoping to be able to extend the work of Chelsea to estimating water storage volumes in seasonal and perennially flooded wetlands.
2. Mr. Xiang Li and Mr. Jared Trost have continued to work on the estimation of groundwater storage within the study region. They have acquired geological data and well data for the region and are using interpolation methods to map water table levels across the study region. At this point they are able to make animated videos of the change in water levels with time during the period of interest (2002 – 2015). Next steps to get water storage will be to incorporate the porosity of geological material. From this mapping we will also be able to quantify the variation in thickness of the unsaturated zone and from that be able to quantify the dynamics of water storage in the unsaturated zone. Estimation of soil

moisture storage will depend on the thickness of the unsaturated zone, and the estimate of soil moisture content within the unsaturated zone. There are a few locations within the study region where soil moisture profiles are available, but we also have near surface soil moisture data available from the SMAT and SMOS satellites.

Activity Status as of July 1, 2019:

1. Estimates of water storage in lakes, wetlands, and groundwater have been derived on an annual basis for the period 2002 to 2015. The estimates for lakes are pretty much finalized, but for wetlands, and groundwater we continue to improve the estimates. The estimate of soil moisture is still being worked on and should be complete by September 30, 2019. We expect that we will continue to tweak the estimates of storages in each of these terrestrial components right up to the end of the project.

Activity Status as of January 1, 2020:

1. Estimates of water storage in lakes and groundwater have been updated and estimates of storage in wetlands and soil moisture have been completed.

Final Report Summary:

1. The estimates of reference storage of water in lakes, wetlands, groundwater and soil moisture were finalized prior to the January 2020 report and have been reported previously. Details of the estimates of water storage in lakes are presented in Chelsea Delaney's MSc thesis provided in Appendix E.1, and a draft manuscript in Appendix E.2.
2. The methodology for mapping water table distribution across the study region continued to be refined until May 2020, and is now final. The methodology and results are described in detail in a report contained in Appendix G. The model used to do the mapping utilizes observation well data and lake level data. It is available for use by agencies and/or consultants. Water table maps for each year 2002 – 2015 have been developed and are available for viewing on the project website.

ACTIVITY 3: Estimate the changes in water storage over the period 2002 to 2015.
Description: Of main interest is to be able to detect changes in water storage in a watershed or in a region. If there exists a reference value of absolute storage to go along with the change in storage it is then possible to derive the changed absolute storage if so desired. The methodology developed in this project will be used to quantify changes in water storage at spatial scales and time scales of practical interest.

The methodology for quantify changes in water storage will be based on the application of satellite remote sensing data in conjunction with ground-based measurements and water balance modeling. Some details of the steps to be taken in this methodology are presented in the following paragraphs. First the information of the hydrologic water balance model will be presented followed by the presentation of the assimilation of the satellite data.

Hydrologic water balance model. Hydrologic water balance models are typically used to quantify the water balance of an area. Models are used for different scales, ranging from the field-scale (acres) to the river basin scale (hundreds of thousands of acres). In the methodology to be developed in this project a selected hydrologic water balance model will be used with the readily available ground-based data to simulate the water balance of a watershed, and satellite data (GRACE, SMOS/SMAP, World View3) will be used to constrain the results of the water balance to keep the model calculating the water balance accurately. The outcome of this combined modeling and data processing will be accurate measures of water storage change at the spatial scale of interest.

GRACE satellite data. The GRACE satellite was launched into orbit in a collaboration between the German space agency and NASA. The satellite quantifies changes in density of different areas of the Earth. Since the amount of water present in an area is dynamic, it is changes in water storage that will cause a change in the satellite signal as the satellite passes over an area. Other researchers have shown that one can monitor changes in water storage on watersheds as small as 10,000 square miles through a combination of preprocessed GRACE data and hydrologic water balance modeling supplemented with ground-based data.

SMOS and SMAP satellite data. To further improve the ability to constrain the water storage change estimates the project will involve the use of soil moisture data quantified using the SMOS and SMAP satellites. These satellites provide soil moisture estimates on the spatial scale of about 90 square miles.

WorldView3 satellite data. High resolution digital images of the Earth surface are now available through the World View satellite data. With these data it will be possible to quantify the elevation of the surface of open bodies of water, including lakes and wetlands. Changes in storage of these open bodies of water can be determined by using the satellite measure of water surface elevation and the morphometry of the water body.

The developed methodology will be tested by selecting a watershed within the studied area to assess the change in water storage. The estimate of water storage change between 2002 and 2015 available from the analysis of ground-based data in Activity 2 will be used as the reference, or correct value of storage change. The hydrologic model will be run with the input ground-based data and constrained by the satellite data to derive an estimation of the change in storage for the same period of time. If the methodology is sound the reference change and the change calculated by the water balance model should be within reasonable agreement.

Summary Budget Information for Activity 3:

ENRTF Budget: $ 127,500
Amount Spent: $ 127,500
Balance: $ 0

Outcome	Completion Date
1. Completed methodology for estimating the change in water storage within the study area. Documentation on the methodology. The developed methodology will be suitable for publication in the scientific literature and also a part of the graduate student's Ph.D. thesis.	1/1/20
2. Validation of water storage change estimation methodology. The result of the validated methodology will be suitable for publication in the scientific literature and also a part of the graduate student's Ph.D. thesis.	03/31/20
3. Final completion report. The analyses derived from this project will be valuable to state and federal agencies for the tracking of water storage changes in areas of concern within Minnesota.	7/31/20

Activity Status as of January 1, 2018:

N/A.

Activity Status as of July 1, 2018:

1. Work is being conducted on the water storage change analysis for all seventeen HUC-8 watersheds. The storage change analysis that is based on the baseflow recession analysis has been completed for the Rum River watershed, and the analysis for the remaining 16 watersheds will be completed by December 31, 2018.

2. The analysis of water storage change based on modeling (CLM and HSPF models) is underway. The CLM model is being run on the Minnesota Supercomputer Center computing facilities. The HSPF model is currently being tested on PC computers, and the input data files for the HSPF model are being acquired from the MPCA.
3. The water storage variation estimation from the GRACE satellite and the SMOS/SMAP satellites has been completed, but we are continuing to test those results as we complete more of the analysis of the ground-based measurement data, and also as results are completed from the models (CLM and HSPF).

Activity Status as of January 1, 2019:

1. Work with the HSPF model was put on hold during the fall because the graduate student working on it, not directly funded by the project, was busy with taking classes. We expect to be able to pick up on the model again in the last phase of the project and apply the model to the project area. Up until now the water balance analyses conducted has been to use simple water balance calculations. These calculations use our watershed-scale average precipitation, evapotranspiration, and streamflow. Our precipitation and streamflow data are considered to be accurate enough, but the evapotranspiration data used are variable depending on the source of the data. We have sourced data from NOAA, the USGS, and the University of Minnesota. A task during the next period will be to select one of these data sources for our final analysis.
2. Mr. Xiang Li continues to conduct analysis of baseflow recession flows as a way of estimating water storage change. He already had much of this completed within the previous period but he continues to refine his analysis as this work is the subject of his M.Sc. thesis.
3. Mr. Francisco Lahoud is working with data from the GRACE satellite to quantify changes in water storage at the scale of individual watersheds. We have 17 watersheds in our study region, and the satellite data are being used to track water storage change. At issue is to be able to confirm that the satellite data is really giving water storage change, and not something else. To confirm this we are working with water balance calculations using precipitation data, evapotranspiration estimates, and streamflow data to calculate water storage changes. Our estimates of precipitation and streamflow are very good, but the evapotranspiration estimates seem to be a problem right now. As mentioned in item 1 above, for evapotranspiration estimation we are using various sources of information (e.g., from NOAA, USGS, University of Minnesota) and attempting to find the method that gives the most reasonable and consistent results.

Activity Status as of July 1, 2019:

1. We have corrected the source of data used for precipitation within the study region. Previously we were using data acquired from the Oak Ridge National Laboratory (ORNL), but we have found that those data have some bias of being too high. So now we have changed the data source over to the MNDNR source.
2. We have corrected the source of data for large scale evapotranspiration (ET). Previously we were using ET estimated from meteorological stations, but now we have moved to using data acquired from satellites. The source of the new ET estimates is CIGA. We have used ground-based eddy-covariance estimates of ET to compare to the large-scale spatial estimates and found the CIGA estimates to have the highest agreement. Our confirmation (ground-truthing) of these estimates is as good as any tests one will find in the peer-reviewed literature, so we feel confident in our results.

3. The new estimates of spatially distributed precipitation and ET are being applied along with stream gage measurements to quantify changes in water storage (by simple water balance) at the HUC-8 watershed scale, and these calculated changes are being compared to water storage changes estimates derived from the GRACE satellite. We are also comparing the water balance calculations to our analysis of lake, groundwater, wetland and soil moisture storages. Preliminary work with the Rum River watershed and the Wild Rice River watershed indicates that there is good agreement between the satellite derived and the ground-based data derived estimates and the water balance calculations.
4. The work with the HSPF model is continuing and will pick up more in late summer and fall of 2019. While the HSPF model calculation is also for the HUC-8 watershed scale, the advantage of the HSPF water balance calculation is that it will provide a way to separate the water balance estimated storages into the lake, wetland, soil moisture and groundwater components; the current water balance calculations we are conducting lumps all the water together into a total terrestrial water storage within a given HUC-8 watershed.

Activity Status as of January 1, 2020:

1. Water storage calculations using a lumped representation of a watershed, and comparison to estimates of water storage change based on GRACE data and pointwise measurements has been completed.
2. Water storage calculations using the HSFP model is ongoing and will be complete by June 2020.

Final Report Summary:

1. Work on the groundwater storage continued improve the accuracy of storage estimates until May 2020, and is now complete. A detailed report on the groundwater storage is presented in Appendix G.
2. The application of the lumped parameter watershed model to the study region and comparison of the calculated water storage change to the water storage change derived from GRACE data was completed in mid-2019. A brief report on this analysis is presented in Appendix H.1.
3. The HSPF model was applied to the Rum River watershed to analyze the relation between water storage and river discharge. A report on this work is presented in Appendix H.2. The model is shown to be useful to quantify this relation and to quantify the spatial and temporal distribution of water storage within the watershed. In work ongoing beyond the scope of this project we are applying this model to attempt to downscale satellite-derived water storage data (GRACE, SMOS/SMP, Landsat). It is expected that by downscaling the water storage it will be possible to improve the storage-discharge relation, and thereby enhance the ability to forecast floods, hydrologic droughts, and ecosystem flows.

V. DISSEMINATION:

Description:

Project results will be disseminated through seminars conducted within Minnesota and at national and international meetings. Example meetings in Minnesota will include the Minnesota Water Resources Conference held each October in St. Paul, seminars for the Water Resources Sciences Graduate seminar program at the University of Minnesota, other seminars held within the University, a webinar for the UZIG (Unsaturated Zone Interest Group) held quarterly, and an annual presentation at the annual meeting for the regional project W-3188 (“Soil, Water, and Environmental Physics Across Scales”, J.L. Nieber is the University representative) held in Las Vegas. Opportunities to present the results in a seminar format to the MnDNR will also be sought. Seminars at national/international meetings include the American Geophysical Union meeting held in San Francisco each December (travel to national/international meetings will be funded from a University funding source).

A project web site was created on the network server at the University of Minnesota to provide the platform for illustrating the ongoing development of project outcomes. Included there will be visuals of maps, and information about databases created throughout the project. Data will be stored on this server so that interested parties will be able to acquire data compiled by the project activities. Also, reports/manuscripts prepared based on the results will be made available on the server.

Status as of January 1, 2018:

1. Presented a poster of the quantifying soil moisture by satellite; presented at the Minnesota Water Resources Conference, October 2017
2. Presented preliminary finds of the research at the regional project meeting, W-3188 in January 2018.

Status as of July 1, 2018:

1. J.L. Nieber, Minnesota: How much water is there; how is it changing? article in the Open River: Rethinking Water, Place & Community, Issue 10: Spring 2018.
2. Presented a poster on the use of the GRACE satellite to quantify changes in watershed water storage. Poster represented at the 1st Water Resources Assembly meeting held at the University of Minnesota, St Paul campus, January 19, 2018.
3. Presented a poster on the use of the GRACE satellite to quantify changes in watershed water storage. Poster represented at the 11th GEWEX meeting held in Canmore, Canada, May 6 – 11, 2018.

Status as of January 1, 2019:

1. Xiang Li. Baseflow Recession Analysis for 17 Watersheds in Central Minnesota, Seminar presented to the University of Minnesota Water Resources Science seminar, November 30, 2018.
2. Claire G. Griffin, Kerry Holmberg, Chelsea Delaney, Leif G. Olmanson, Patrick L. Brezonik, Jacques C. Finlay, and John L. Nieber. Remote sensing of dissolved organic matter pools in lakes at regional scales, poster presented at the 2018 Fall meeting of the American Geophysical Union, Washington, DC
3. Francisco Lahoud and J.L. Nieber. Quantifying Total Water Storage in the Minnesota River Combining Remote Sensing and Land Use Models Poster presented at the Minnesota Water conference, October 2018.

Status as of July 1, 2019:

A project web site has been created and is currently being populated with project results and accessible project data. One can think of this web site as being 'under construction' and it will continue to be improved incrementally with time (about once every two weeks). Our projection is that the web site will be complete by June 30, 2020. We do expect that the project website will be a useful focal point for this and related research, so we intend to continue to add to and improve the website even beyond the completion of the project in 2020.

As the website is still under construction at this time, we are giving access to a limited number of people, including project team members and any interested LCCMR staff upon request. The link to the website is

https://sites.google.com/s/1VFVvT77dGQvki9gRj-DClaLxTIs-7LV3/p/1Zj1HZo-dOipNlTOfh9ruDUgmb2eJ29mi/edit?userstoinvite=kerry.holmberg%40gmail.com

Status as of January 1, 2020:

1. Presentations at meetings.
    a. Li, X., Drainage Timescale estimates and storage change analysis on a basin scale, Poster, American Geophysical Union Fall meeting, December 2019
    b. Nieber, J.L., Quantifying terrestrial water storage in Central Minnesota, Water Resource Sciences Graduate Student seminar, University of Minnesota, September 2019
    c. Nieber, J.L., Estimating total terrestrial water storage in Central Minnesota, Minnesota Water Resources Conference, St. Paul, October, 2019
    d. Teng, Pai-Feng, The evolution of remote sensing of flood forecasting and potential improvement with Gravity Recovery and Climate Experiment (GRACE): A review, Poster, American Geophysical Union Fall meeting, December 2019

2. Two MSc. Theses.
    a. Ms. Chelsea Delaney, Estimating Lake Water Volume Using Scale Analysis, University of Minnesota, December 2019
    b. Mr. Xiang Li, Drainage Timescale estimates and storage change analysis on a basin scale, June 2020.

3. Manuscript on lake volume estimation. Chelsea Delaney, John Nieber, Kerry Holmberg, Jared Trost, Adam Heathcote, Bruce Wilson, Estimating Lake Water Volume Using Scale Analysis, Draft Manuscript for submission to a scientific journal, September 2020.

4. The results of the lake volume method used in this project have been adopted in a LCCMR funded project managed by Dr. Jacques Finley for a statewide analysis of lake volumes. In that case the lake volume was being used to assess the water quality of the lakes as monitored using satellite data. A draft manuscript for the results of that study has been developed and the manuscript should be submitted for publication in September 2020.

Final Report Summary:

The project has tested the idea of using satellite data, with application to Central Minnesota, to monitor the changes in water storage from space. The GRACE satellite has provided estimates of the change in total water stored in the landscape (surface water, soil moisture and groundwater), while the Landsat satellite has provided data for quantifying the storage of water in lakes (about 43,000 lakes) in the study region. A water table mapping application was developed within the scope of the project to allow mapping of the water table across the study region, and from that mapping the storage of water in the Quaternary aquifer was completed. This mapping provides estimates of water storage change in the groundwater aquifer on an annual basis and estimates of change from this method were compared to the changes derived from the GRACE satellite and the results compared favorably. This water table mapping application is available for use by state agencies and will be of value to land developers, land managers, and water resource planners. This mapping application is being used in a follow-up study to examine the use of the water table mapping to assess the potential of flooding or the potential for hydrologic drought. Three methods for estimating water storage change have been developed/tested within this project; GRACE satellite, water table mapping, and baseflow analysis (work described in the M.Sc. thesis of Xiang Li). All of these methods yield compatible estimates of water storage change, so each can be used independently. The use of all three methods, when data are available for all

methods, is probably the best approach since an ensemble average estimate of water storage change from all three methods will eliminate the shortcomings of any one of the methods.

VI. PROJECT BUDGET SUMMARY:

A. ENRTF Budget Overview:

***This section represents an overview of the preliminary budget at the start of the project. It will be reconciled with actual expenditures at the time of the final report.**

Budget Category	$ Amount	Overview Explanation
Personnel:	$ 198,160	The personnel working on the project from the University of Minnesota include research faculty J. Nieber, J. Baker, B. Wilson, T. Griffis, no funding required; graduate student (F. Lahoud, $79,416 (60% salary, 40% benefits), 0.5 FTE each year for two years); graduate student (TBD), meteorologist, $29,729, (60% salary, 40% benefits), 0.38 FTE for year 1; undergraduate students (TBD), $8,502 (100% salary), 0.5 FTE in summer of both years, and 0.2 FTE during school year of both years; research scientist (B. Hansen, Senior Scientist, $63,590 (73% salary, 27% benefits), 0.6 FTE in year 1 and 0.15 FTE in year 2; R. Kanivetsky, Hydrogeologist; $13,923 (67% salary, 33% benefits), 0.13 FTE for year 1.).
Professional/Technical/Service Contracts: Subcontract with the U.S.G.S.	$ 51,840	Mr. Jared Trost with the Water Resources Center of the U.S.G.S. located in Mounds View, who will work collaboratively with the University of Minnesota research group to accomplish the goals of Activity 1 and Activity 2.
Travel Expenses in MN:	$ 0	Travel within Minnesota is necessary to visit field sites where monitoring of groundwater, soil moisture, streamflow and weather is conducted. The project will not involve collection of field data, but acquisition of data collected by others. However, visiting field sites will be necessary to confirm documented local information about the field sites. Also, travel between St. Paul and Mounds View to conduct project meetings, or to travel within Minnesota to present research results.
Other:	$ 9,909	Summer salary for Tim Griffis received from a NSF/NASA grant that is directly related to the work being done in this project.
TOTAL ENRTF BUDGET:	**$ 250,000**	

Explanation of Use of Classified Staff: N/A

Explanation of Capital Expenditures Greater Than $5,000: N/A

Number of Full-time Equivalents (FTE) Directly Funded with this ENRTF Appropriation: 3.15 FTE

Number of Full-time Equivalents (FTE) Estimated to Be Funded through Contracts with this ENRTF Appropriation: 0.5 FTE

B. Other Funds: $9,909. Funding from a NSF/NASA grant for Tim Griffis which covers one month of his summer salary for both years. This is a match to the ENRTF funding for the project.

VII. PROJECT STRATEGY:
A. Project Partners:
A. Project Team/Partners

No ENRTF funding required:
John L. Nieber, Professor, Department of Bioproducts and Biosystems Engineering, will serve as project principal investigator and will work on all aspects of the project.
Bruce Wilson, Professor, Department of Bioproducts and Biosystems Engineering, will work on the uncertainty analysis of water storage estimates.
Timothy Griffis, Professor, Department of Soil, Water and Climate, will conduct the research related to the water balance (hydrology) model. Dr. Griffis has summer support from a NSF/NASA grant to cover his part of the effort on the project.
John Baker, Professor and Research Leader, USDA-ARS, Department of Soil, Water and Climate, will work alongside Dr. Griffis on the application of the land surface/atmosphere interaction model to estimate soil moisture storage, aquifer recharge, and generation of surface runoff.

ENRTF funding required:
Jared Trost, Hydrologist, USGS, Mounds View, will work on the estimation of water storage in the aquifers of the study area.
Roman Kanivestsky, Adjunct Professor, Department of Bioproducts and Biosystems Engineering, will assist with the interpretation of quaternary and bedrock geologic data for the study area for the estimation of unsaturated zone and aquifer storage.
Kerry Holmberg, Assistant Research Scientist, Department of Bioproducts and Biosystems Engineering, will work on the acquisition of monitoring well data, soil data and lake/wetland data for the study area. She worked on the estimation of water stored within the surface waters of the study area.
Francisco Lahoud, Graduate Student, will work on the satellite data and combining it with ground-based data. He will be involved in acquisition of data, data processing, modeling, and data analysis.
Xiang Li, Graduate Student, will work on the water table mapping, and the analysis of baseflows for quantifying water storage change in aquifers.
Chelsea Delaney, Graduate Student, will work on the estimation of water storage in lakes and wetlands.
Ke Xiao, Graduate Student in meteorology, will work on the modeling of evapotranspiration and will also assist with the water balance (hydrology) model.
Undergraduate Research Assistant, will assist with acquisition of data and data processing, and preparation of visual aids for presentation of results.

B. Project Impact and Long-term Strategy:
Quantification of water storage on the surface and in the subsurface across Minnesota is essential for sustainable management of water use and improvement of the quality of Minnesota water resources. The quantification of water storage will help to reduce the uncertainty about the state of water resources and reduce the potential for conflicts between competing users. It will also assist with accounting for the needs of adequate flows of water in streams and rivers for sustaining aquatic health. In early discussions with the MNDNR it was been established that a method for estimating changes in water storage, regionally and locally will be useful to the MNDNR water allocation planning activities. It is also expected that the methodology will be

helpful to state and federal agencies in the forecasting of potential flooding as well as assessments of impacts of drought on available water supplies, and useful as well to agencies and consultancies in conducting water quality assessments. The scope of the current project is limited to a study of the mid-central region of the state. It is hoped that the methodology will be found valuable enough that additional funding will be available to expand the data-base and methodology to the entire state. In the long-term, but beyond the scope of the current project it is hoped to install the developed methodology, which will be data intensive and modeling intensive, on a server computer at the University of Minnesota, and made available to state agencies for installation on agency computers and use by state agency staff. If that happens the methodology will be taught to water resource managers within the agencies. It is also expected that the methodology will be taught to graduate students at the University of Minnesota. The methodology is currently being used in a follow-up grant from the National Science Foundation to estimate the potential for flooding.

C. Funding History: N/A

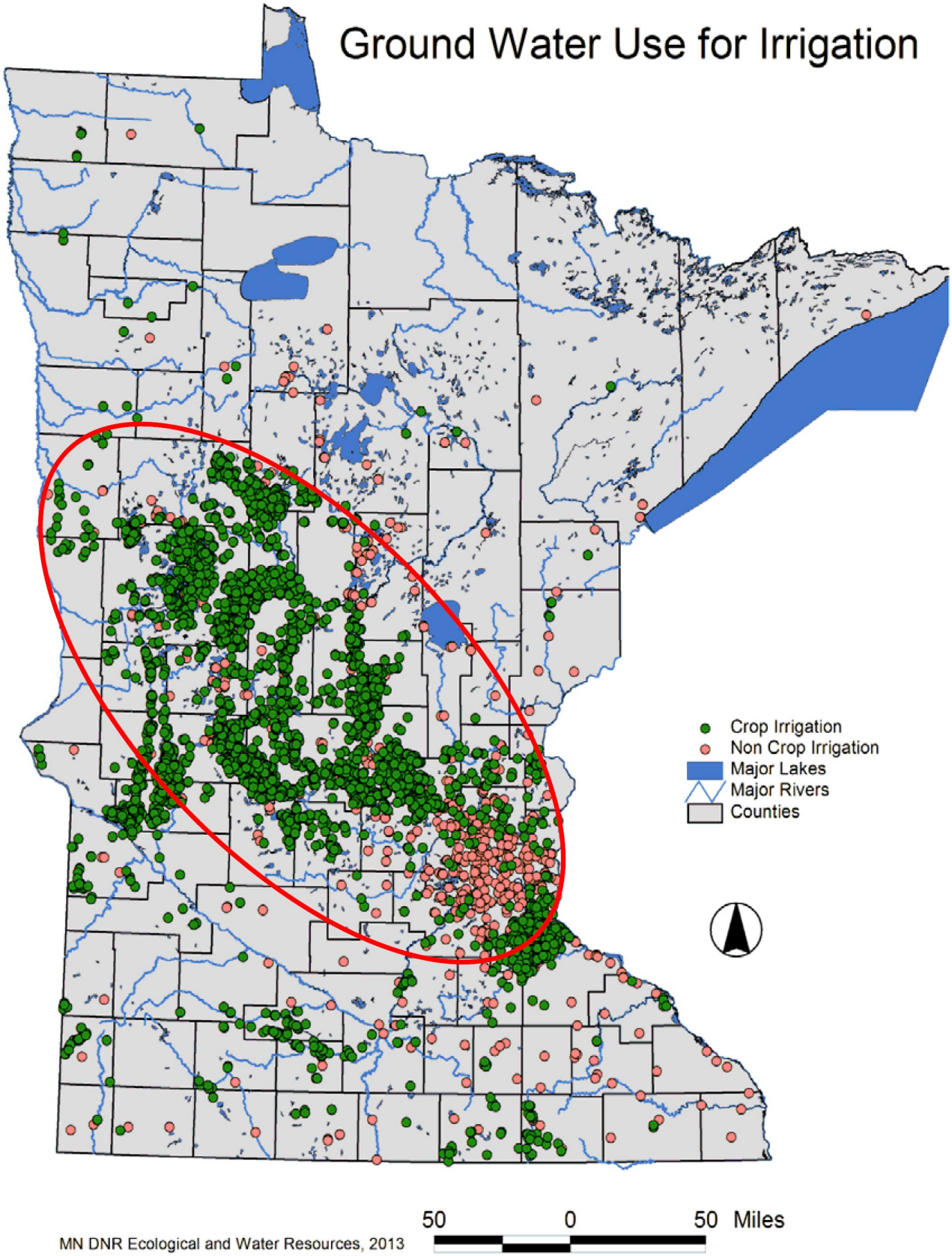
VIII. FEE TITLE ACQUISITION/CONSERVATION EASEMENT/RESTORATION REQUIREMENTS:

A. Parcel List: N/A

B. Acquisition/Restoration Information: N/A

IX. VISUAL COMPONENT or MAP(S):
See map on the next page for the study area.

Area for proposed study for project O18-A.



X. RESEARCH ADDENDUM:

The research addendum for this project is presented in Appendix I.

XI. REPORTING REQUIREMENTS:

Periodic work plan status update reports will be submitted no later than January 1, 2018, July 1, 2018, January 1, 2019, and July 1, 2019. A final report and associated products will be submitted between June 30 and August 15, 2019.

XII. RESULT APPENDICES:

Appendices A-D were put here for additional information reported during the duration of the study. Appendices E-H are full reports that are too voluminous to be included here but are associated with this report. Appendix I is the original Research Addendum which is kept here for any future reference.

A. Lake water storage estimation. Statistical approach.

Lake Volumes:
A breakpoint regression model relating lake surface area to lake volume (n=909) is being used to estimate lake volume for all of the lakes in our study area (n=30,000). The Yearly History sub-dataset from the Global Surface Water dataset (https://data.europa.eu/euodp/en/data/dataset/jrc-gswe-global-surface-water-explorer-v1) provides surface water delineations from 1984-2015. Delineations were defined using expert systems, visual analytics and evidential reasoning from big data exploration and information extraction techniques to identify water and land from LandSat images. Rough estimates of total lake volume in the study area from 2001-2015 are given below. Volumes were estimated with surface areas from the GSW dataset using breakpoint regression. River volumes are not included and will be estimated before the next progress report.

Table A.1. Total lake water volume estimates for project area.

w/o Mille lacs

Year	Count	Area m^2	Volume Hect-m
2001	27288	2710323257	1168192
2002	28359	2745875486	1171406
2003	28155	2790395635	1189470
2004	28247	2757821342	1176828
2005	30709	2759010237	1177992
2006	30764	2759010655	1178933
2007	30111	2791131904	1188157
2008	29158	2789532512	1188904
2009	30142	2781118526	1185969
2010	30709	2759846726	1178302
2011	30334	2767956913	1181870
2012	24671	2791264617	1188387
2013	28642	2761926450	1181745
2014	30482	2626386470	1131923
2015	28089	2702117204	1156865

w/Mille lacs

Year	Count	Area m^2	Volume Hect-m
2001	27288	3227649660	1617911
2002	28359	3263326882	1621234
2003	28155	3307836793	1639289
2004	28247	3275229199	1626617
2005	30709	3276665545	1627997
2006	30764	3276623642	1628901
2007	30111	3308726875	1638109
2008	29158	3307030349	1638772
2009	30142	3298503037	1635739
2010	30709	3277215773	1628058
2011	30334	3285346452	1631644
2012	24671	3308726875	1638224
2013	28642	3279437273	1631624
2014	30482	3143516350	1581471
2015	28089	3219112630	1606296

B. Lake water storage estimation. Topographic scaling method.

Methods:

Using the Minnesota Department of Natural Resources (DNR) hydrography data, 40,054 lakes were selected in Minnesota's 17central watersheds to be used to predict their volumes (Figure B.1) (DNR Division of Fish & Wildlife, 2012). 785 lakes with known volumes from the DNR morphology data of the state were also compiled to compare the predicted volumes to the known volumes (Minnesota Department of Natural Resources (DNR), 2015).

Using ArcMap 10.6.1, the area was projected in the NAD_1983_UTM_Zone_15N coordinate system for all layers. Following Heathcote *et al*. (2015) method, each lake had a buffer created around the lake using an equation as seen in the following:

$$D = 2 \cdot \sqrt{(A / \pi)} \quad \text{(Equation 1)}$$

where *D* is the equivalent diameter and *A* is the lake surface area. Once the equivalent diameter was determined, 25% of the equivalent diameter was calculated and used to determine the optimal buffer area. The 25% buffer resulted in the best prediction of lake volume when compared to the set of known volumes.

Topography for each buffer was calculated using a 1/3 arc-second Digital Elevation map (DEM) of Minnesota provided by United States Geological Survey (USGS) and the DNR hydrography layer (U.S. Geological Survey, 2017). Within each lake's buffer, topography was summarized as the minimum, maximum, and mean elevation change which was determined by calculating the difference between the mean elevation within the buffer and the minimum elevation.

Once the topography was determined, we used Heathcote *et al*. (2015) volume equation as a basis for calculating lake volume (V). Using lakes that were in both the hydrography layer and the morphology layer, a regression analysis was conducted using the known volumes from the morphology layer and the mean elevation change and surface area from the hydrography layer. Once run, coefficients for the volume equation was determined as seen in equation 2.

$$\log_{10} V = \log_{10} \text{lake area} \cdot 1.07 + \log_{10} \text{elevation change}_{25} \cdot 0.249 \quad \text{(Equation 2)}$$

After computing lake volumes, a correction for a bias towards back-transformations of values was conducted from Ferguson (1986)

$$\hat{Y}\ corr = 10^{\log_{10}\hat{Y}} \cdot \exp(2.65 \cdot s^2) \quad \text{(Equation 3)}$$

where $\hat{Y}$ corr is the uncorrected predicted value, s^2 is the residual variance from the model, and 2.65 is a constant. This provided a better estimate for the lake volumes calculated.

Comparing the known and predicted lake volumes, the model explained 76% of the variation in lake volume ($R^2 = 0.76$, $F_{[2, 782]} = 1224$, $P < 2.2e\text{-}16$). All statistical analysis was conducted using the statistical software R (RStudio Team, 2016).

Results/ Discussion:

Our results supported our use of the terrestrial slope and surface area to predict the volume of a lake. When comparing the known and predicted lake volumes, the model explained 76% of the variation in lake volume (Figure B.2). The RSE for the model was 0.336 $\log_{10} m^3$. Furthermore, when looking at the total volume of the 785 lakes with a bias correction to a nonbiased correction, we saw that with a bias correction the predicted lake volume was only 0.32% different than the 2.6% difference of the nonbiased total to the observed volume total (Table B.1).

While different models are becoming more abundant in calculating lake volumes, we can see that this model developed by Heathcote *et al*. can in fact significantly predict the volumes of lakes. Even though determining volume could be calculated by field surveying, that is time consuming and expensive. Therefore, there must be a continuation on developing models that can accurately predict lake volumes with only limited data.

Despite the fact there is a significance occurring within this watershed, only 785 lakes were surveyed with a total of over 40,000 lakes within the watersheds. Due to the small sample size of lakes with known volumes, the predicted coefficient is not as accurate as it would be if the sample size was larger. In addition, using a 1/3 arc-second DEM resulted in a loss of smaller lakes when calculating the elevations within each lake buffer. For further research, a better resolution will be used in order to more accurately predict volumes.

While this study is only limited to central Minnesota, a full state calculation has been undergone to determine if an extremely large-scale attempt at using these calculations can accurately predict lake volume across vast distances and ecoregions. Based on preliminary research, the model explained 82% of the variation in lake volume with over 1,000 lakes of 15 acres or larger to compare from. Further research is needed in order to determine if we can better accurately predict lake volumes at different scales.

Figures:

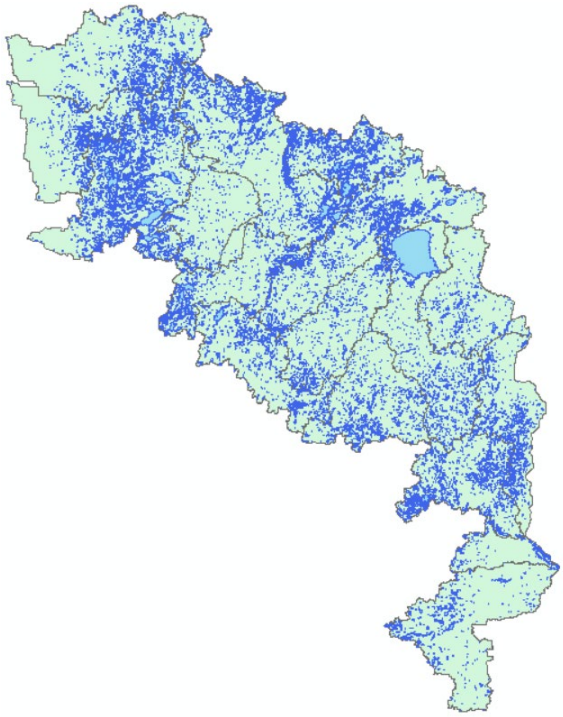


Figure B.1. Map of 40,054 lakes and ponds contained within 17 watersheds of Minnesota showing ones selected for calculating individual and total volume of lakes.

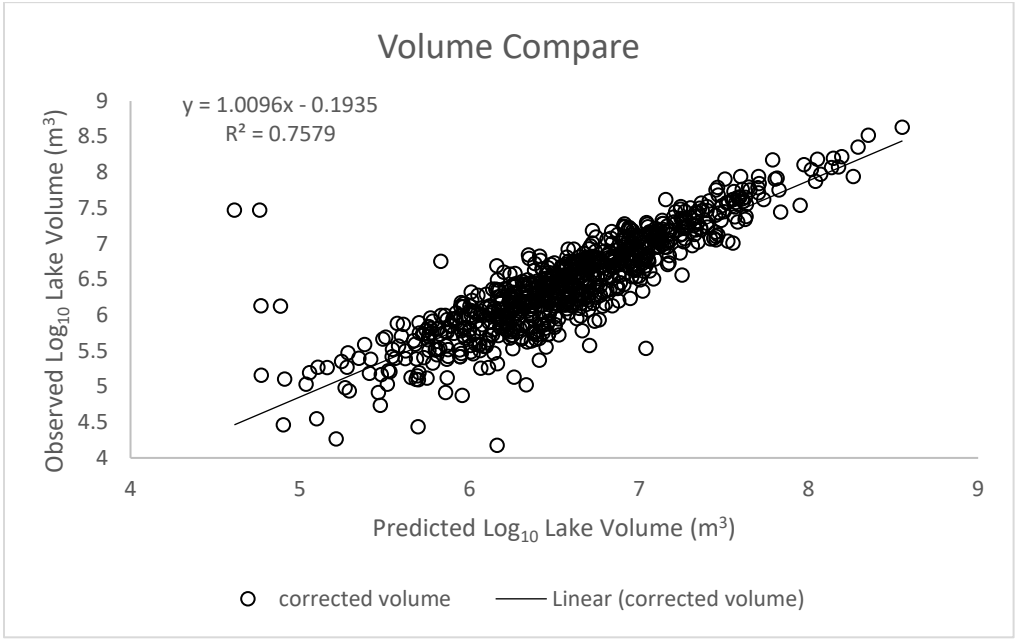


Figure B.2. Observed and predicted lake volume (log_{10} m^3) calculated from a 1/3 arc-second DEM of 17 watersheds in central Minnesota.

Table B.1. Percent differences between the total observed lake volume compared to total predicted lake volume with a statistical correction and without a statistical correction (n=785 lakes).

	Total Volume (hectare)	Percent Difference between Predicted Volume and Observed Volume
Predicted Uncorrected Volume	585875.45	2.6%
Predicted Corrected Volume	789585.65	0.32%
Observed Volume	787038.95	--

References:

DNR Division of Fish & Wildlife - Fisheries Unit. (2012). DNR Hydrography - Lakes and Open Water. Minnesota DNR - Division of Fisheries. ftp://ftp.gisdata.mn.gov/pub/gdrs/data/pub/us_mn_state_dnr/water_dnr_hydrography/metadata/dnr_hydrography_lakes_and_open_water.html

Ferguson, R. I. (1986). River loads underestimated by rating curves. Water resources research, 22(1), 74-76.

Heathcote, A. J., del Giorgio, P. A., & Prairie, Y. T. (2015). Predicting bathymetric features of lakes from the topography of their surrounding landscape. Canadian journal of fisheries and aquatic sciences, 72(5), 643-650.

Minnesota Department of Natural Resources (DNR). 2015. Lake Basin Morphology. Minnesota DNR, Division of Fish and Wildlife. ftp://ftp.gisdata.mn.gov/pub/gdrs/data/pub/us_mn_state_dnr/water_lake_basin_morphology/metadata/metadata.html

RStudio Team (2016). RStudio: Integrated Development for R. RStudio, Inc., Boston, MA URL http://www.rstudio.com/.

U.S. Geological Survey (2017). 1/3rd arc-second Digital Elevation Models (DEMs) - USGS National Map 3DEP Downloadable Data Collection: U.S. Geological Survey.

C. Water table mapping and water storage estimation.

To understand physical changes in water storage, water level data were acquired from the following publicly available sources for the period 2002 to 2015:

1. The Minnesota DNR cooperative groundwater monitoring network (https://www.dnr.state.mn.us/waters/cgm/index.html)
2. The Minnesota DNR lakefinder database (https://www.dnr.state.mn.us/lakefind/index.html)
3. The U.S. Geological Survey (https://waterdata.usgs.gov/mn/nwis/gw/)
4. The Minnesota Department of Agriculture (personal communication)
5. The Minnesota Department of Health Well Index online (http://www.health.state.mn.us/divs/eh/cwi/)

The data were processed to a common horizontal datum (WGS84) and a common vertical datum (NAVD88). Water levels for wells were constrained to the water table aquifer or wells drilled to less than 50 ft in unknown aquifers. Daily mean levels were computed for sites with hourly water level data. The data were reviewed for clear outliers and outliers were removed. For example, a water level of 100 ft was removed from a record for a lake where the remainder of the values were greater than 800 ft. The resulting data set contains 343,580 water

level measurements across the study area for 8,325 measurement locations. Lakes were considered to be surface expressions of the groundwater level.
An annual mean water level for each year from 2002 to 2015 was calculated for each site. Only sites with at least 1 measurement per year were included for the analysis presented below. Only 650 sites met this criteria: 499 lakes and 151 wells. Figure C.1 shows some consistent trends in water levels across the study area during certain periods. For example, most water levels across the study area increased between 2009 and 2011. Water levels were generally above average across the study site in 2002, 2011, and 2014; water levels were generally below average in 2004 and 2007. These large-scale consistent patterns should be apparent in the GRACE satellite data.

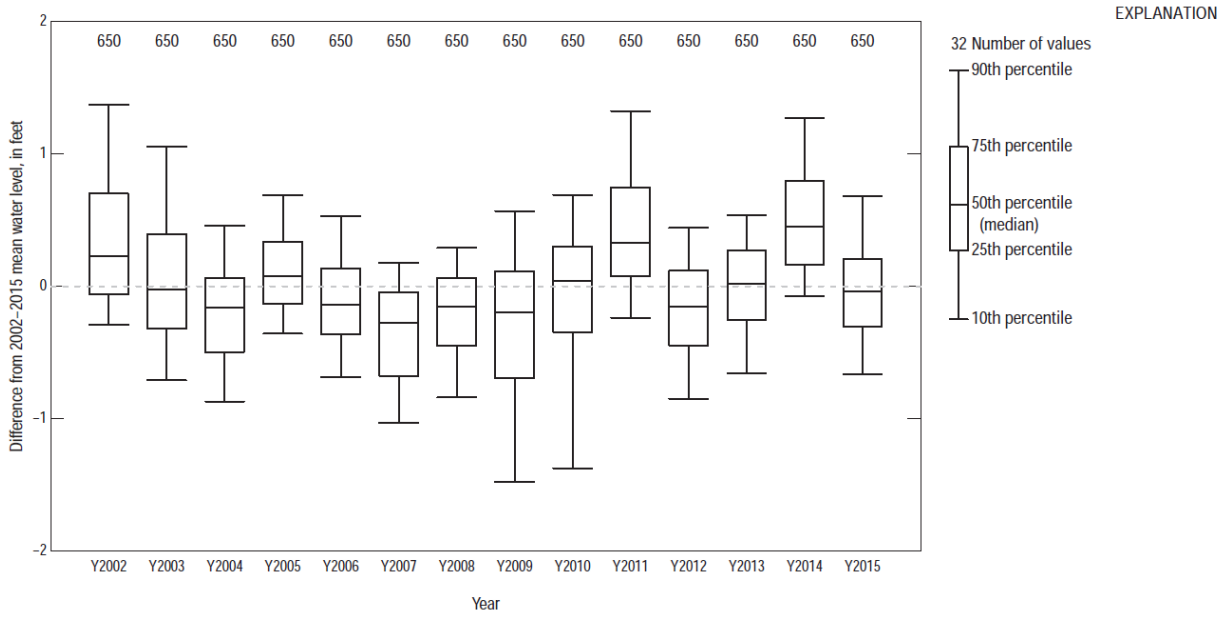


Figure C.1. Boxplot showing water level anomalies by year for all 650 sites with data for each year between 2002 and 2015. Each water level anomaly data point was calculated as the difference between annual mean water level at a site and the 2002-2015 mean annual water level at that site.

Despite the abundance of water level data, substantial portions of the study area have little information. Other data sources are being explored to improve water table mapping through the project area. These include Global Surface Water data from Google Earth Engine, perennial stream centerlines, the National Wetlands Inventory, and a 30-m statewide DEM derived from 1-m LIDAR data. To convert water levels into groundwater storage, the porosity of geologic materials is needed. Porosity is the void space between sediment particles that can be filled with water. Porosity values have been assigned according to USDA texture classes available in the Minnesota Geological Survey's Surficial Geology Map (https://conservancy.umn.edu/handle/11299/191889).

D. Satellite estimation of water storage variations.

This is a brief summary of the progress made in the Water Storage Central Minnesota project in determining terrestrial water balances in each of the watersheds in the study area.

For such a determination the pilot test was started in the Rum River Watershed, with the following results:

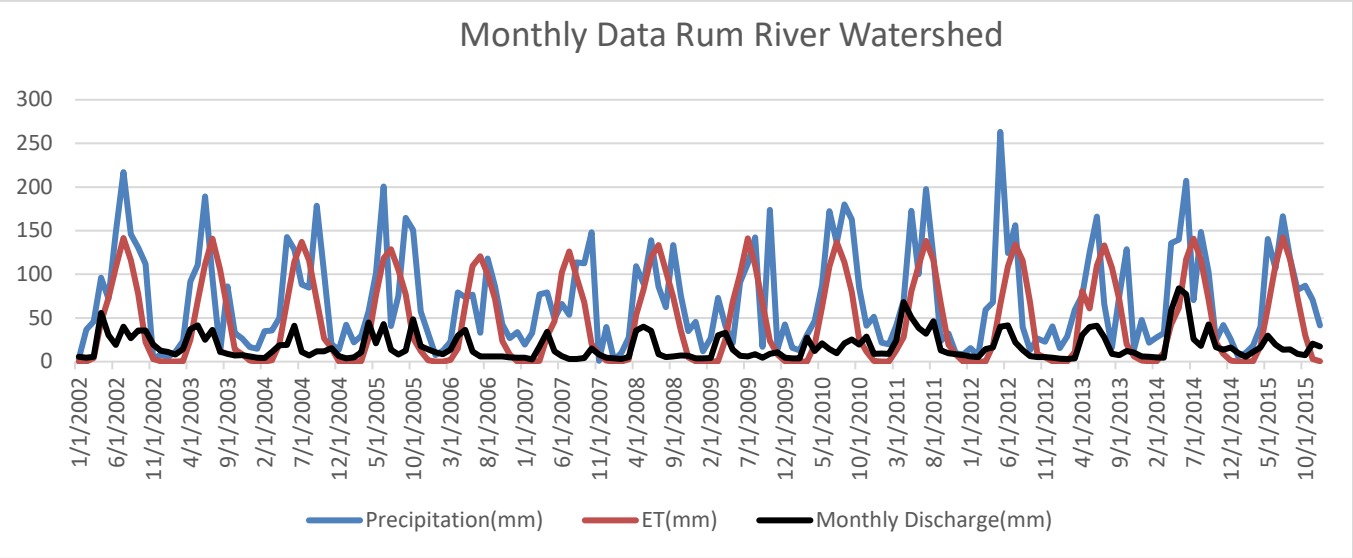
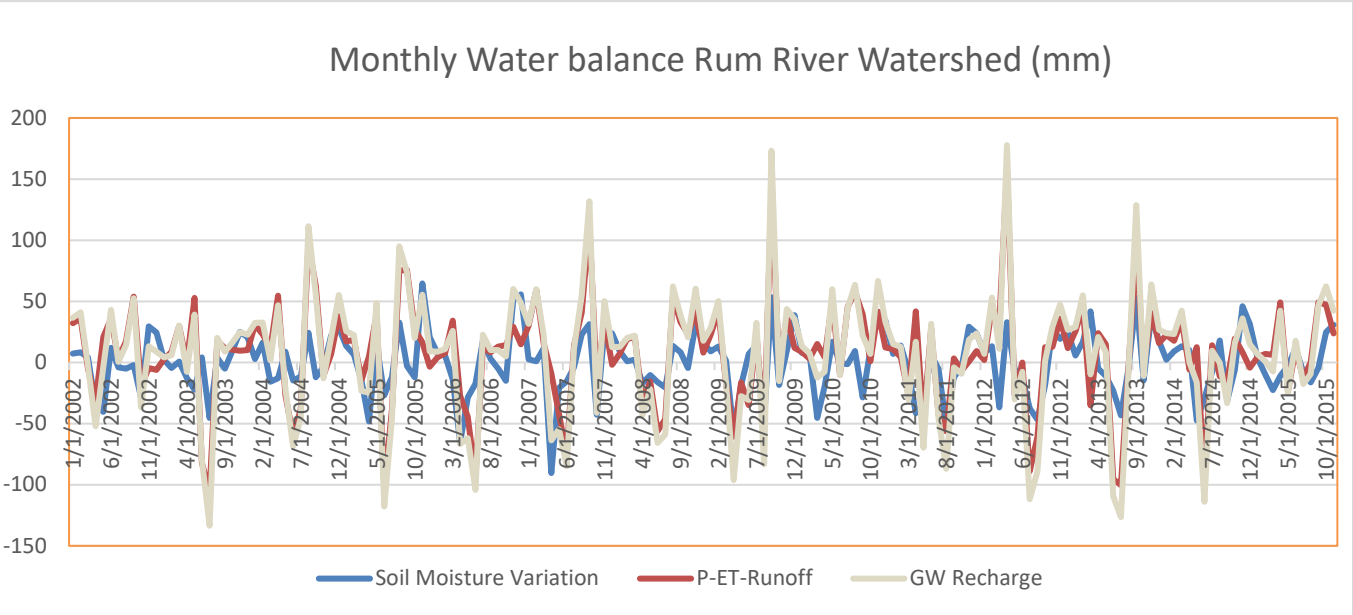
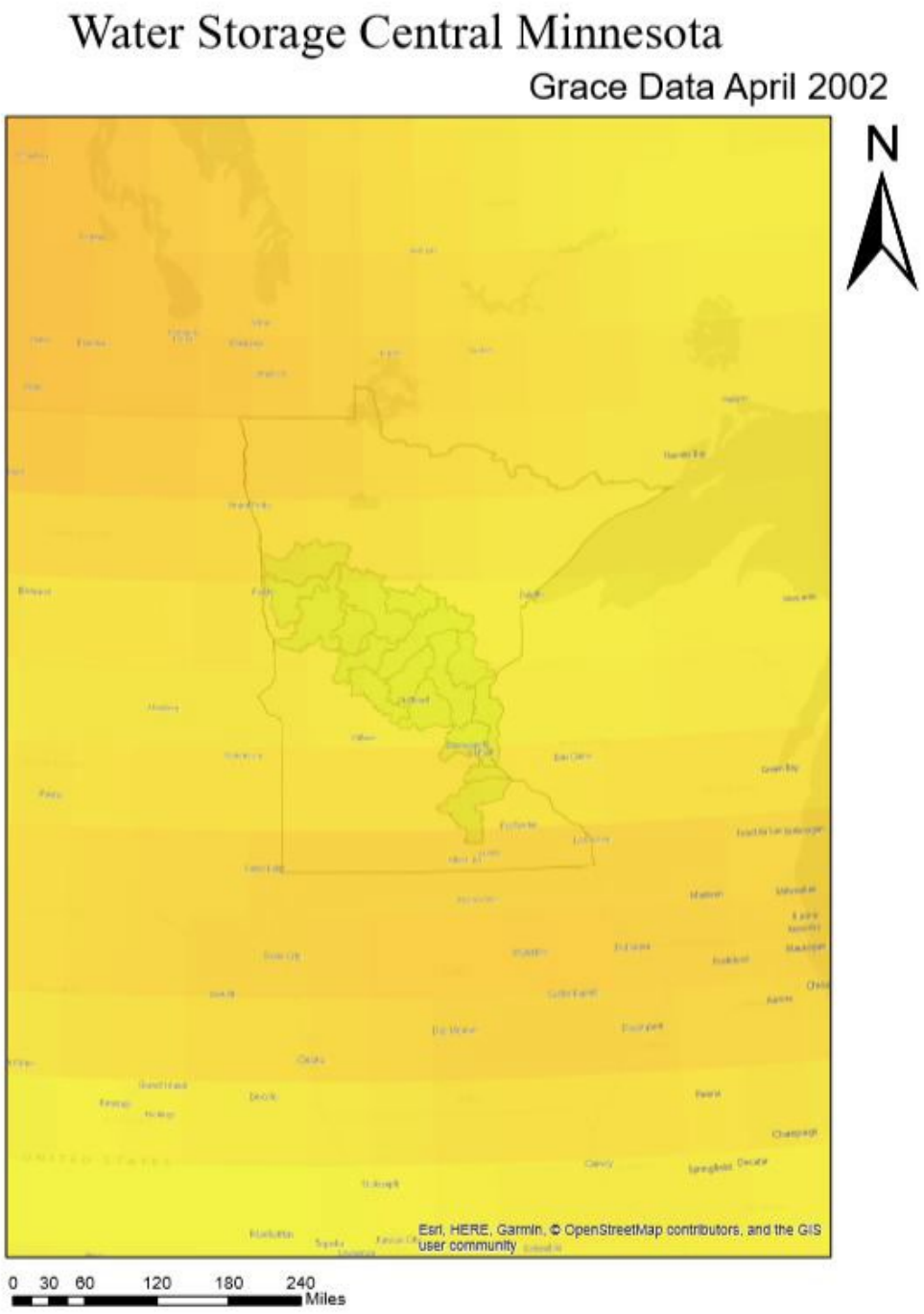
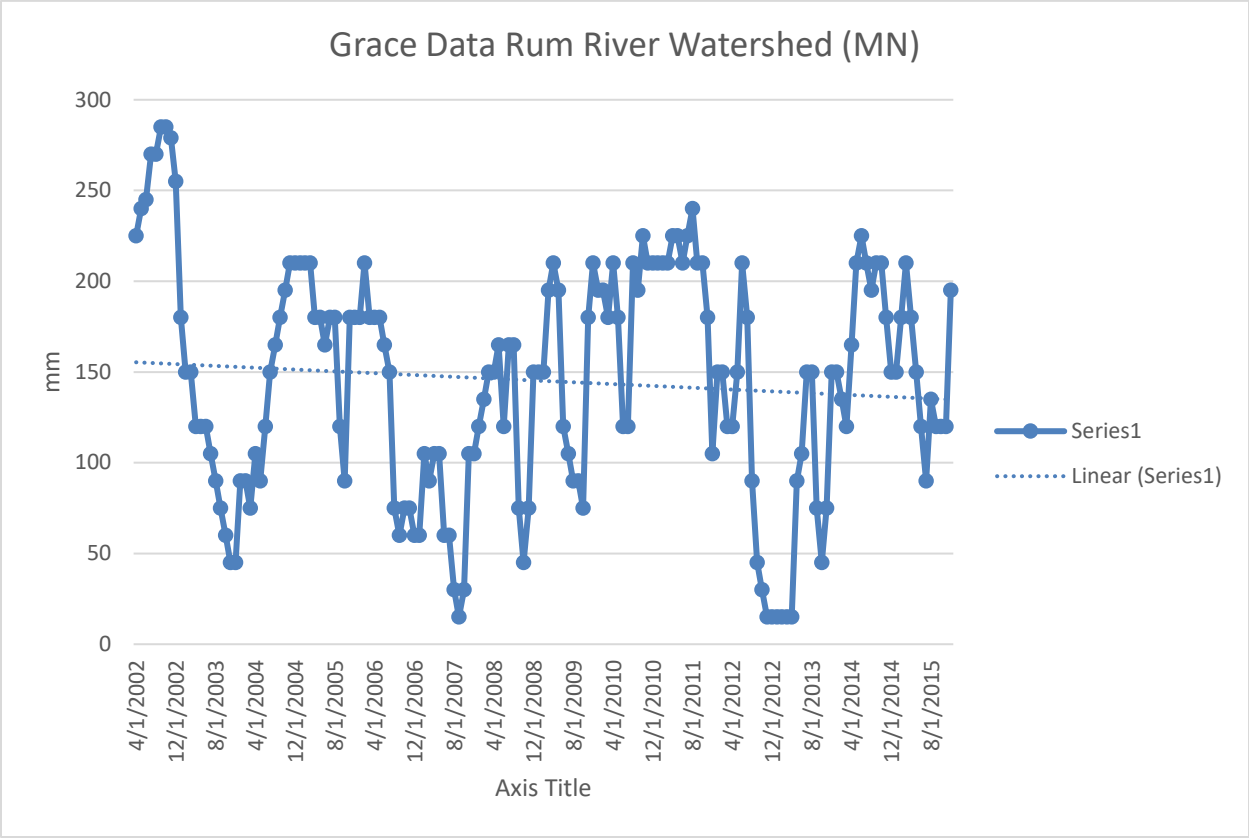


Figure D.1. Time series of precipitation, evapotranspiration and monthly river discharge (all given in mm) for the period 2002 – 2015.

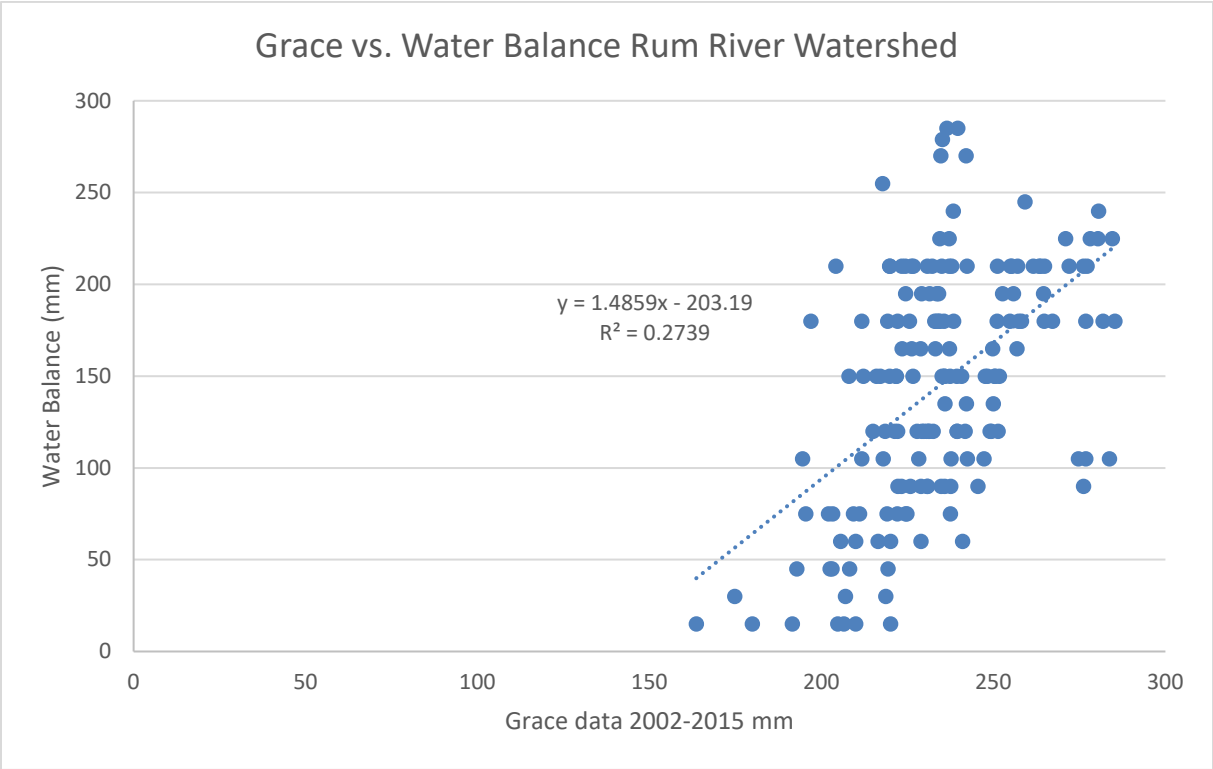


Grace Data for Water Storage Central Minnesota:





There are no negative values for GRACE values because a 150 mm datum was used to avoid them. There is a net decrease of 15 mm in the evaluated period, from 155 mm to 140 mm.



The correlation is acceptable taking into account the Watershed scale and the GRACE satellite measurement scale.

Finally, in the water balance calculation of monthly volumes (using monthly precipitation, evapotranspiration, and streamflow data) there is a net increase of 1170 mm. This result contradicts the result obtained from the GRACE satellite. Thus far we have identified that the probable reason for the discrepancy is that the water balance calculation is very susceptible to the estimation of the watershed-scale estimates of evapotranspiration. For the evapotranspiration data product sources available to us (NOAA, USGS, University of Minnesota) there is significant difference among the estimates. We will be evaluating these data sources in the final part of the project.

July 1, 2019

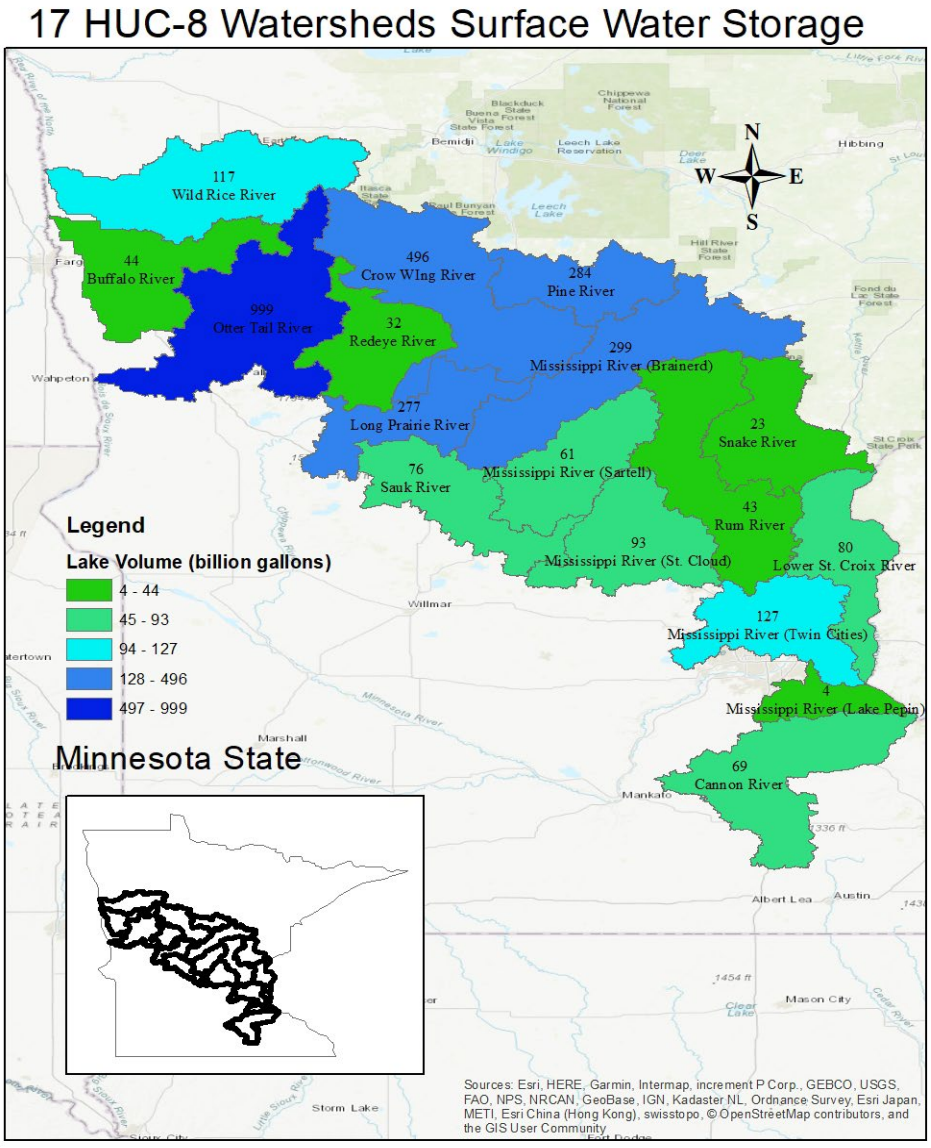


Figure 1. Map of water volume in lakes by HUC-8 watershed within the study region. Volumes are given in billions of gallons.

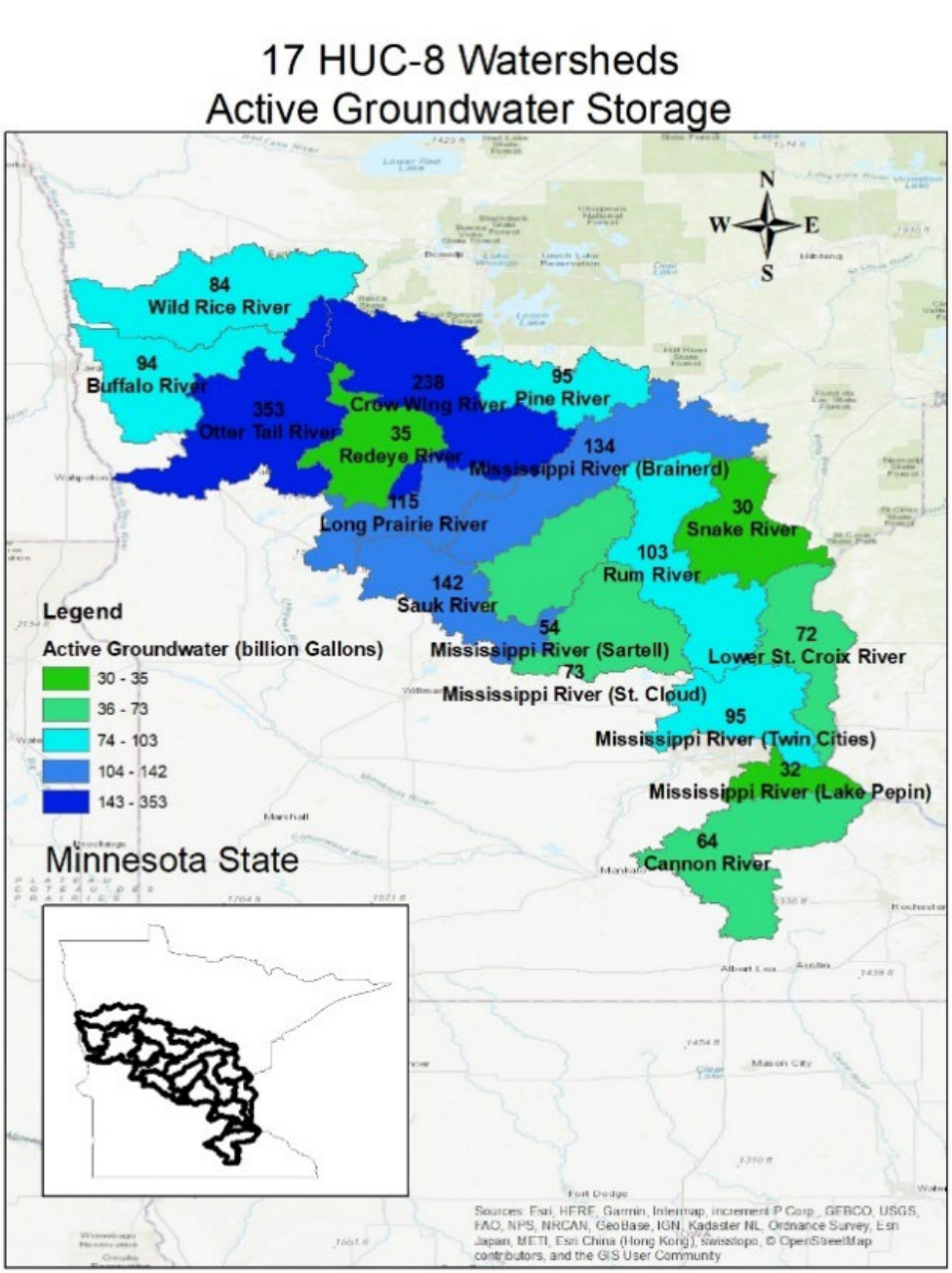


Figure 2. Map of water storage in active groundwater by HUC-8 watershed within the study region for 2015. Volumes are given in billions of gallons.

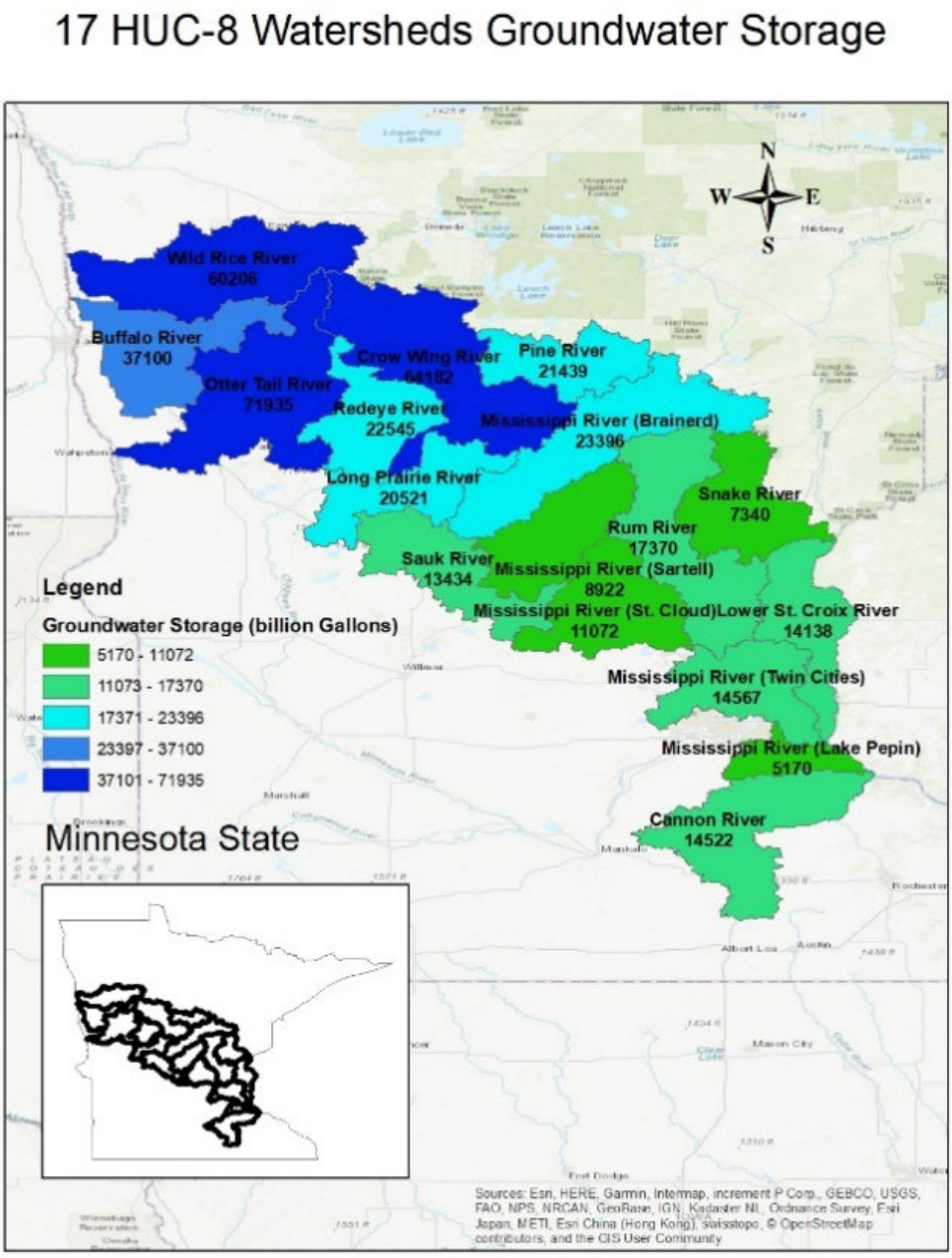


Figure 3. Map of water storage in groundwater by HUC-8 watershed within the study region for 2015. Volumes are given in billions of gallons. The groundwater zone represented is for the quaternary aquifer in the region.

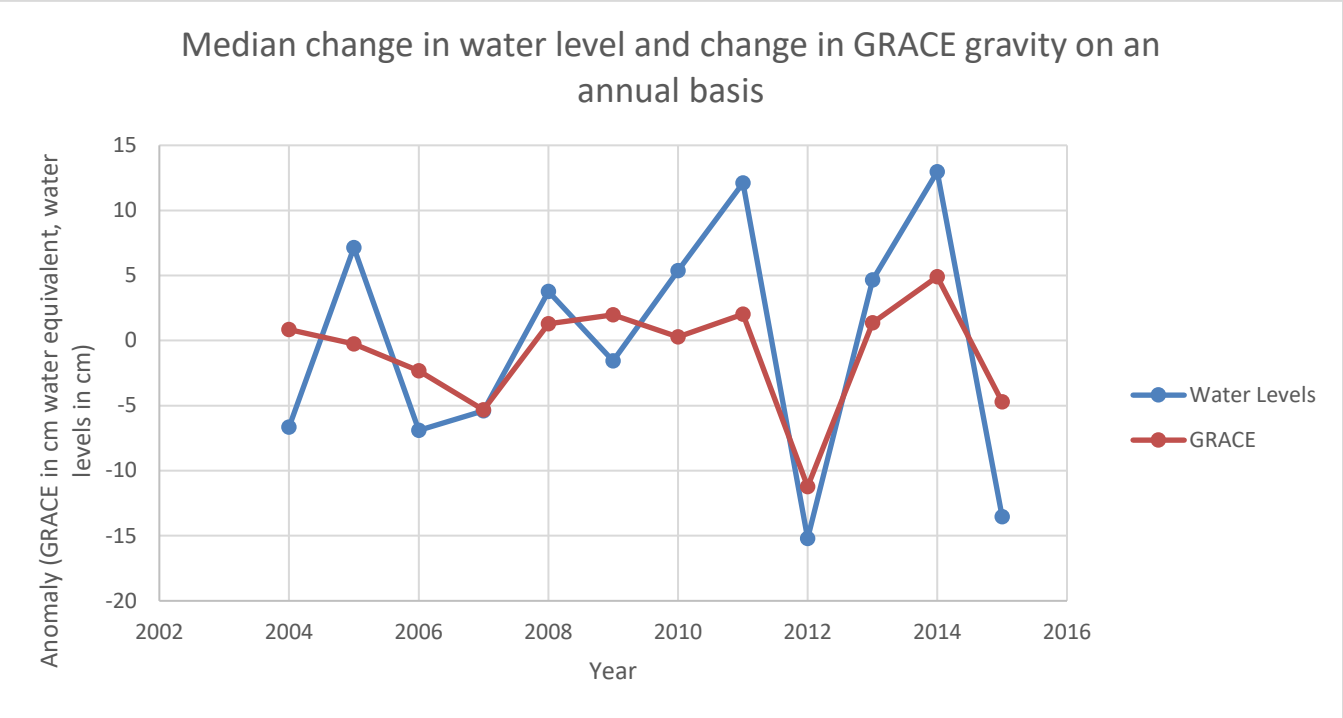


Figure 4. Comparison of two estimates of water storage change in the study region for the period 2004 – 2015. One of the estimates is from the water table mapping approach that uses observation well and lake level data to quantify the water stored in groundwater. The other estimate is from one of the available GRACE satellite products (Mascon).

Final Report Appendices
August 2020

The final report contains five appendices, E - I. These accompany this report in the form of MSWORD files.

Environment and Natural Resources Trust Fund
M.L. 2017 Project Budget

Project Title: *Techniques for Water Storage Estimates in Central Minnesota*
Legal Citation: $250,000 the first year is from the trust fund to
Project Manager: John L. Nieber
Organization: University of Minnesota
M.L. 2017 ENRTF Appropriation: $*250,000*
Project Length and Completion Date: 3 *Years, June 30, 2020*
Date of Report: July 1, 2020

ENVIRONMENT AND NATURAL RESOURCES TRUST FUND BUDGET		Activity 1 Budget; 06/19/19	Amount Spent	Activity 1 Balance	Activity 2 Budget; 6/19/19	Amount Spent	Activity 2 Balance	Activity 3 Budget; 6/19/19	Amount Spent	Activity 3 Balance	TOTAL BUDGET	TOTAL SPENT	TOTAL BAL.
BUDGET ITEM													
Personnel (Wages and Benefits)		$27,948	$27,948	$0	$42,712	$42,712	$0	$127,500	$127,500	$0	$198,160	$198,160	$0
Roman Kanivetsky, Hydrogeologist: (67% salary, 33% benefits); 0.065 FTE for first year.	*$5,109*		$5,109			$0			$0				
Research Scientist, Kerry Holmberg: (73% salary, 27% benefits); 0.25 FTE in both years	*$48,619*		$5,400			$5,800			$37,419				
Francisco Lahoud, Graduate Student: (60% salary, 40% benefits); 0.5 FTE for year 1 and 0.25 FTE for year 2.	*$58,106*		$9,689			$21,292			$27,125				
Ke Xiao, Graduate Student, Meteorologist: (60% salary, 40% benefits); 0.25 FTE for year 1 of the project.	*$11,166*		$7,500			$3,666			$0				
Xiang Li, Graduate Student: (60% salary, 40% benefits), 0.5 FTE each year for year 1; (84% salary, 16% benefits), 0.5 FTE for year 2.	*$50,063*		$0			$7,107			$42,956				
Chelsea Delaney, Graduate Student: (60% salary, 40% benefits), 0.25 FTE for 2018; (84% salary, 16% benefits), 0.25 FTE for 2019.	*$21,055*		$0			$4,000			$17,055				
Pai-Feng Teng, Graduate student: 0.25 FTE for 2019; (60% salary, 40% benefits)	*$2,945*		$0			$0			$2,945				
TBD, Undergraduate assistant: $8,502 (100% salary); 0.5 FTE in summer of both years, and 0.2 FTE during school year of both years.	*$1,097*		$250			$847			$0				
Professional/Technical/Service Contracts		$15,552	$15,552	$0	$36,288	$36,288	$0	$0		$0	$51,840	$51,840	$0
Jared Trost, Hydrogeologist, Water Resources Center, U.S.G.S., Mounds View: $51,840 (75% salary, 25% benefits); 0.0.35 FTE for first year and 0.15 FTE for second year. Jared will assist with acquisition of data on monitoring wells and interpretation of geologic formations, and he will work on the estimates of storage in the subsurface.													
Travel expenses in Minnesota		$0	$0.00	$0	$0	$0	$0	$0	$0	$0	$0	$0	$0
Travel will be conducted to visit field sites for locations of monitoring wells, soil moisture measurements, and streamflow measurements. Also travel to project meetings between St. Paul and Mounds View.													

frontiers | Frontiers in Water

ORIGINAL RESEARCH
published: 16 June 2022
doi: 10.3389/frwa.2022.886964

Estimating Lake Water Volume With Regression and Machine Learning Methods

Chelsea Delaney[1], *Xiang Li*[1], *Kerry Holmberg*[1], *Bruce Wilson*[1], *Adam Heathcote*[2] *and John Nieber*[1]*

[1] *Department of Bioproducts and Biosystems Engineering, University of Minnesota, St. Paul, MN, United States,* [2] *St. Croix Watershed Research Station, Science Museum of Minnesota, Marine on St. Croix, MN, United States*

The volume of a lake is a crucial component in understanding environmental and hydrologic processes. The State of Minnesota (USA) has tens of thousands of lakes, but only a small fraction has readily available bathymetric information. In this paper we develop and test methods for predicting water volume in the lake-rich region of Central Minnesota. We used three different published regression models for predicting lake volume using available data. The first model utilized lake surface area as the sole independent variable. The second model utilized lake surface area but also included an additional independent variable, the average change in land surface area in a designated buffer area surrounding a lake. The third model also utilized lake surface area but assumed the land surface to be a self-affine surface, thus allowing the surface area-lake volume relationship to be governed by a scale defined by the Hurst coefficient. These models all utilized bathymetric data available for 816 lakes across the region of study. The models explained over 80% of the variation in lake volumes. The sum difference between the total predicted lake volume and known volumes were <2%. We applied these models to predicting lake volumes using available independent variables for over 40,000 lakes within the study region. The total lake volumes for the methods ranged from 1,180,000- and 1,200,000-hectare meters. We also investigated machine learning models for estimating the individual lake volumes and found they achieved comparable and slightly better predictive performance than from the three regression analysis methods. A 15-year time series of satellite data for the study region was used to develop a time series of lake surface areas and those were used, with the first regression model, to calculate individual lake volumes and temporal variation in the total lake volume of the study region. The time series of lake volumes quantified the effect on water volume of a dry period that occurred from 2011 to 2012. These models are important both for estimating lake volume, but also provide critical information for scaling up different ecosystem processes that are sensitive to lake bathymetry.

Keywords: bathymetry, lake volume, scale analysis, machine learning, Minnesota

OPEN ACCESS

Edited by:
Tongren Xu,
Beijing Normal University, China

Reviewed by:
Georgia A. Papacharalampous,
Czech University of Life Sciences, Czechia
Alban Kuriqi,
Universidade de Lisboa, Portugal

***Correspondence:**
John Nieber
nieber@umn.edu

Specialty section:
This article was submitted to Water and Hydrocomplexity, a section of the journal Frontiers in Water

Received: *01 March 2022*
Accepted: *06 April 2022*
Published: *16 June 2022*

Citation:
Delaney C, Li X, Holmberg K, Wilson B, Heathcote A and Nieber J (2022) Estimating Lake Water Volume With Regression and Machine Learning Methods. Front. Water 4:886964. doi: 10.3389/frwa.2022.886964

INTRODUCTION

Fresh water is a crucial resource to humans. With an ever-changing environment, we need to be better prepared to protect it. One of the most important freshwater bodies are lakes. While the surface area of all lakes covers <4% of the global landmass and the total volume of water is a small fraction of total terrestrial freshwater, they are home to a wide range of biodiverse ecosystems (McDonald et al., 2012). The ecosystem functioning of lakes provides tangible ecologic and economic value, yet key information such as lake datasets that contain basic morphological and hydrologic characteristics needed to determine these functions are missing (Hollister et al., 2011; Crétaux et al., 2016). Lake volume and maximum lake depth are vital components in many lake functions related to the physical, biological, and chemical processes within a lake. For example, the volume of a lake can affect the water residence time which in turn can affect the nutrient dynamics and primary productivity (Sobek et al., 2011) as well as the zooplankton dynamics of a lake (Obertegger et al., 2007). With missing or inaccurate data, the prediction of these functions is not as precise as they could be, making it more difficult to quantify the changes that may occur within these environments (Sobek et al., 2011; Crétaux et al., 2016; Messager et al., 2016).

As two important parameters determining the nature of circulation processes and biogeochemical processes in lakes, data on lake volume and lake depth are scarce. Even the available data in many parts of the world are merely present for only a very small fraction of the total number of lakes. For instance, in Minnesota, 'the land of lakes,' the number of lakes with detailed bathymetric data is <2% of the total number of lakes in the state. Given that current technology makes it impractical to directly measure bathymetric information at large scales, it becomes necessary to develop predictive models for these parameters using the information that is available. At present, a widely used approach is to estimate lake volume with lake surface area data. Models for lake volume using lake surface area were among the first models developed and include the work by Håkanson and Karlsson (1984). Improvements in lake volume models were made by including a second prediction variable that involved some measure of the land surface topography in the area surrounding a lake. The idea of this second variable is that the topography of the surface surrounding a lake would reflect the topography of the lake bottom. Studies that involved a prediction variable representing the surrounding topography include Håkanson and Peters (1995), Hollister et al. (2011), and Sobek et al. (2011).

A modification of the lake buffer topography variable was proposed by Heathcote et al. (2015). In this study, they used the change in surface elevation in a buffer area surrounding the lake, with the buffer area scaled according to lake surface area. Heathcote et al. applied this model to the data for 433 lakes located in different geographic regions in the southern part of the Province of Quebec (Canada). In doing so, the model explained 95% of the variation in lake volume.

While the Heathcote et al. (2015) method predicted lake volumes using self-similar scaling, the Cael et al. (2017) method developed a model assuming that the land surface is self-affine. The scaling of such surfaces has been shown theoretically to be related to the Hurst coefficient. Since lake water fills in the depressions of the land surface, a description of the surface as being self-affine should provide a theoretical background for predicting the volume of water in the depressions. According to the theory of such self-affine surfaces, the volume of the depression will be proportional to the depressional surface area raised to some exponent. This exponent can be shown to be calculated from the Hurst coefficient, which itself can be related to the fractal dimension of the surface. For the earth surface, the Hurst coefficient has been determined to be about 0.4 ± 0.1 for the spatial scale relevant to lakes (see for example Renard et al., 2013).

In their study, Cael et al. (2017) predicted lake volumes on a global scale with vastly different regions and topographic features. The model is meant to be used to predict the total volume and mean depth of a collection of lakes. However, the model can be used to estimate the volume for individual lakes, but these are determined on a statistical basis. Their estimate of the total volume of lakes globally was 199,000 km^3, which is lower than previous estimates of 210,000 km^3.

Both Heathcote et al. (2015) and Cael et al. (2017) methods estimate lake volume in a statistical regression model. Statistical models are elegant in their solid theoretical foundation, interpretability, and easy implementation. Nevertheless, their ability to handle non-linearity and complex prediction problems are also constrained by their simple model architecture. Recently, machine learning (ML) methods have become a popular approach to model complex non-linearities from scientific data and their contributions to tackle water-related problems have been previously acknowledged (Shen et al., 2018). Despite the wide applicability of machine learning, their use in lake volume prediction, to our knowledge, has not been explored. Thus, we have additionally developed and applied machine learning models to predict lake volume using limited lake bathymetric data and compared this technique to the performance of the regression models.

The ability of ML to solve predictive problems (Sejnowski, 2020) has already made its scientific applications span a diversity of fields. Among them, ML applications in hydrology have also experienced unprecedented progress (Shen et al., 2018). Kratzert et al. (2018, 2019) built machine learning models to predict catchment scale streamflow using weather forcing data and achieves state-of-the-art performance, which also scientifically advances the development in hydrologic regionalization. Jia et al. (2020) coupled physical knowledge into machine learning and builds knowledge-guided machine learning models to model lake temperature. Shukla et al. (2022) applied machine learning methods and Gaussian process modeling techniques to predict discharge with hydrologic knowledge in complex stage-discharge relationships. Additionally, machine learning has also been applied to map lake spreading areas (Deoli et al., 2021) and flooding regions (Avand et al., 2022).

For this paper, we tested the ability of several methods to predict lake volumes in the central region of Minnesota (USA), a region that has over 40,000 lakes (Delaney, 2019). The objective

TABLE 1 | Averages of morphology traits from 816 surveyed lakes provided by Minnesota Department of Natural Resources (DNR).

Lake Volume (m^3)	Number of lakes	Average size (m^2)	Average max depth (m)	Average depth (m)	Average volume (m^3)	Average surface area (m^2)
$\leq 10^4$	11	124,239	2.3	0.8	55,032	124,156
10^4-10^5	192	3,338,656	6.7	2.2	530,127	333,926
10^5-10^6	436	1,054,611	11.0	4.0	3,653,086	1,054,480
10^6-10^7	166	4,477,037	18.9	6.3	25,600,190	4,477,055
$>10^7$	11	25,148,785	31.2	7.9	190,669,424	25,148,967

of this study was to predict volumes of lakes to better understand lake processes using readily available, remotely sensed data. The methods included a model using just lake surface area, the model of Heathcote et al. (2015) using lake surface area and near lake topography, the model of Cael et al. (2017) using lake surface area and assuming self-affine surfaces, and methods based on conventional machine learning tools with lake surface area and near lake topography as independent variables. In addition to testing the ability of these models to predict lake volume for one point in time, we also applied the lake surface area regression model to determine the temporal variability in total lake water volume for the entire region for the period 2002–2015.

METHODS

The database used for developing the regression models and the machine learning models was derived from archived lake data available from the Minnesota Department of Natural Resources (Minnesota Department of Natural Resources, 2017). The available data was for 816 lakes, with known volumes and a shapefile of each lake with corresponding bathymetric data. These 816 lakes ranged from volumes of 10^4 to greater than 10^7 m^3 with known maximum depths, average depths, and surface areas for each lake. A summary of each lake size category is given in **Table 1**. The developed regression models were then applied to the other lakes in the study region. The hydrography data for these other lakes, absent depth or volumes, were also available from the MnDNR (Department of Natural Resources Division of Fish Wildlife, 2014). The distribution of these lakes, a total of 40,054, is illustrated in **Figure 1**. The boundaries of the Hydrologic Unit Code (HUC)-8 watersheds, 17 of them, in the region are shown in the illustration. Lake Mille Lacs is noticeable in the northeastern part of the study region by its large size. It contains nearly 25% of the total lake water volume in the 17 HUC-8 watersheds. To better illustrate the range and variability in predictions of lake volume, and because the bathymetric data of large lakes such as Lake Mille Lacs are usually well-defined due to their large economic and recreational value, this lake was excluded from the estimates of total regional lake water volume that follow.

Data for the temporal variation of lake surface area was acquired from satellite data provided by the Global Surface Water (GSW) observations program (Pekel et al., 2016). These data were acquired for the period 2002–2015.

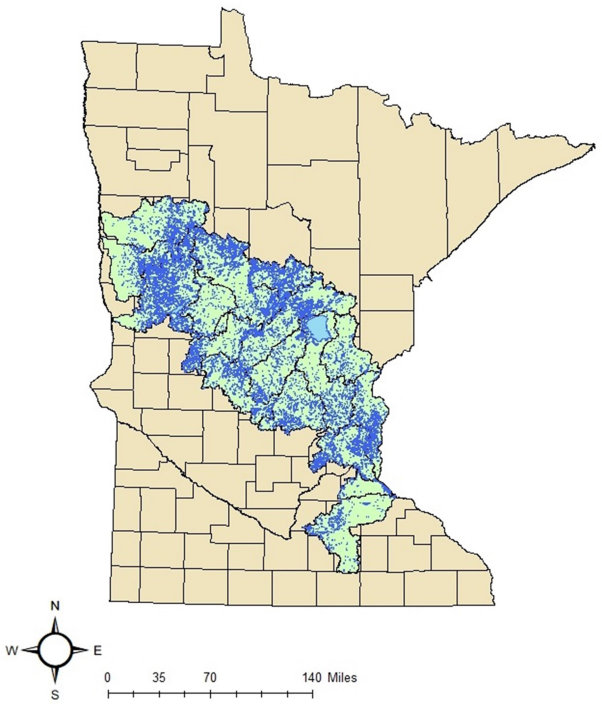


FIGURE 1 | Map of 40,054 lakes contained within 17 HUC-8 watersheds of central Minnesota used in lake volume prediction.

Method Using Surface Area Alone

The first model developed was one using lake surface area only, with the equation being

$$V = aA^b \tag{1}$$

where a and b are empirical constants. A regression model was developed by regressing log-transformed lake volume on log-transformed lake surface area for the 816 lake dataset.

Heathcote Method

The original concept of using lake surface area and land surface slope came from Håkanson and Peters (1995) who suggested using an empirical model that calculated lake volume from lake surface area and maximum slope of the catchment from 95 lakes in Sweden. While the lake volume model was able to explain a

high percentage of the variability in volume, the model requires catchment area data which may not always be available in some locations. Sobek et al. (2011) improved this concept by using surface area and the maximum slope of the land surface within a static buffer of 50 m around each lake, for a total of 6,130 lakes, to calculate lake volume within Sweden resulting in the lake volume model that explained 92% of the variability in volume. Heathcote et al. (2015) further developed this method to predict lake volume as well as maximum depth, by using the lake surface area and the average change in land surface elevation within a near-lake buffer with the buffer length dependent on the lake surface area. This allowed the lake's buffer area to be proportional to the size of the lake rather than a static buffer distance as done by Sobek et al. (2011).

Heathcote et al. (2015) found the average change in elevation between the surrounding terrestrial landscape and the lake surface to be the best predictor of bathymetric properties, lake volume, and maximum depth (Heathcote et al., 2015). The terrestrial buffer surrounding the lake was used because of the assumption that the elevation change surrounding the lake was formed by the same geomorphic process forming the elevation change within the lake and that the slope of the surrounding topography is near to that of the slope of the lake bottom (Hollister et al., 2011). Due to not being able to calculate the slope occurring under the water because that information is not available, the method uses elevation change surrounding the lake as an independent variable to predict the lake volume. The concept is that by studying the relationship between the morphology of a lake and the surrounding area, lake volumes can be predicted without detailed bathymetric data. Based on their empirical testing, Heathcote et al. (2015) found that the length of buffer should be 25% of the equivalent diameter (D) of the lake surface, where $D = 2\sqrt{\frac{A}{\pi}}$ and A is the lake surface area. In our application of the Heathcote et al. (2015) method, the topography for each buffer of Minnesota lakes was calculated using a 1/3 arc-second Digital Elevation map (DEM) (∼10 m) (U.S. Geological Survey, 2017).

The prediction equation for lake volume based on the Heathcote et al. (2015) approach is given by

$$V = A^c D E_{25}^d \tag{2}$$

where DE_{25} corresponds to the average elevation change within the buffer of length equal to 25% of the equivalent lake surface diameter, and c and d are empirical parameters. This regression equation, in log transformed form, $log_{10}\, V = c\, log_{10} A + d\, log_{10} D\, E_{25}$, was fit to the data for the 816 lakes. According to Heathcote et al., this log transformation helps to prevent heteroscedasticty. Due to there being a bias introduced when estimates are being back transformed from regressions, corrections were conducted based on Ferguson (1986) to prevent variables from being underestimated (Ferguson, 1986). The Pearson's partial correlation coefficient and the Akaike information criteria (AIC) (Akaike, 1974) test were run to determine the strength in relationship and to assess the predictive power of the regression model between variables (surface area and elevation change). All statistical analysis was conducted using the statistical software R (RStudio Team, 2016) and the "ppcor" package was used to calculate the partial correlation coefficient (Kim, 2015).

Due to the size range and the variability of lake formation within the region further testing was conducted to determine whether or not pooling the lakes within the region into groups of similarity might improve the accuracy of lake volume prediction (Delaney, 2019). Two group selections were tested: grouping by lake size and grouping by the HUC-8 watershed within which a set of lakes are located.

Lakes were categorized by surface area size into the following size ranges: $<10^4$, 10^4-10^5, 10^5-10^6, 10^6-10^7, and $>10^7$ m^2. Due to the lack of known volumes of lakes with a surface area $<10^4$ m^2, those lakes were assigned a depth of 0.5 m in order to calculate volume by multiplying the depth and surface area. This depth was chosen because known lake morphology within the region for lakes within a surface area between 10^4 and 10^5 had an average depth of 0.8 m (**Table 1**) and we assumed that the average depth of lakes with a surface area $<10^4$ m^2 would be smaller than that of lakes with a surface area between 10^4 and 10^5. Each of the size groups had their own regression analysis conducted following the Heathcote et al. (2015) method.

Lakes were also segregated by HUC-8 watersheds to examine whether geographic location played a role in the lake volume relation. Each watershed with its own lakes had a regression analysis conducted following the protocol above.

With all individual lake volumes calculated, the volume of the 40,054 lakes with known surface area and elevation change was calculated to find a sum total of water storage for each of the different lake groupings.

Cael Method

Cael et al. (2017) proposed a volume-surface area scaling method to estimate the cumulative volume of a collection of lakes. They provided theoretical background on the relationship by proposing that when scaling self-affine surfaces, the volume and area of a lake existing on that surface has a relationship through the use of the Hurst coefficient. Through this theoretical approach, the lake volume is given by

$$V \alpha A^{1+\frac{H}{2}} \tag{3}$$

where H is the Hurst coefficient. For the surface of the earth, the Hurst coefficient has been determined to be 0.4 ± 0.1. Rather than accounting for the near-lake surface topography as done in the Heathcote et al. (2015) approach, the Cael et al. method already has the surface topography accounted for in the use of the Hurst coefficient. This approach facilitated the prediction of lake volumes across diverse regions and topography with limited bathymetric data. Of course, the equation above is a theoretical result and it requires empirical data to test whether the theory applies. To test this, we fitted the empirical equation (Equation 4) to lake surface area and corresponding volume data for the 816 lakes in the data set for Central Minnesota, where ζ is the volume-area scaling exponent, κ is a proportionality coefficient, and ε is an error term.

$$V = 10^{\kappa+\varepsilon} A^{\zeta} \tag{4}$$

The regression analysis was conducted using log transformed surface area and volume to compare known volumes to predicted volumes derived from the empirical formula. The ζ and κ were determined by the regression analysis from the slope and intercept. To consider the variability in lake volumes within the study area, confidence intervals from bootstrapping resampling procedures were calculated (Leschinski, 2019). These two procedures were used to account for the different sources of uncertainty within ζ and κ. The error (ε) within the equation was determined from the root-mean-square error (RMSE) of the residuals of the scaling relationship.

The equation was then applied to all the lakes within the study area (40,054) and then summed to determine the total lake volume. All statistical analyses were conducted using the statistical software R (RStudio Team, 2016) and the "pracma" package was used to calculate the Hurst coefficient (Borchers, 2019).

Machine Learning Method

Although there is a large pool of ML model options to utilize, we selected the artificial neural network (ANN) (Dreyfus, 1990) as one important baseline approach to investigate. ANN became one of the popular sub-families of ML in recent years because of its internal function and architectural advantages in capturing non-linearities in data. In particular, a few ANN variants have already made tremendous improvements to address some difficult computer science challenges, such as computer vision (LeCun et al., 2015) and natural language processing (Hirschberg and Manning, 2015). Details of ANN are explained in the section below.

In addition to ANN, we also explored other traditional ML alternatives which were once popular ML baseline options before neural networks arose. Those other ML models we investigate include support vector regression (SVR) (Cortes and Vapnik, 1995), random forests (RF) (Breiman, 2001), and Gradient boosted regression tree (GT) (Chen and Guestrin, 2016). Because we emphasize the application of ANN to represent ML in this paper, we will not provide as much detail on these alternative models, however, we wanted to highlight our consideration of other ML candidate models.

By nature, ML models are data hungry (Adadi, 2021) and a generalizable ML model requires training process involving abundant data, which in reality often leads to a challenge for data collection. In other words, the ML model using limited data will learn behaviors in a way that is hard to generalize to out of sample scenarios. For this reason, ML models should not be trained and evaluated using the same dataset because they can easily overfit the data during the training process but achieve unsatisfactory performance for unseen testing data. Thus, evaluating ML models based on seen training data without testing on unseen data gives a biased model assessment. To evaluate the ML model on unseen testing data, we performed 5-fold cross-validation to evaluate the ML models. The whole dataset was split into five equal-sized chunks, each time one portion of it was dropped as the testing data while the remaining four chunks were used for training the ML model. For each ML model, we will only assess its testing performance and report those statistics across five different trainings as the model evaluation metrics.

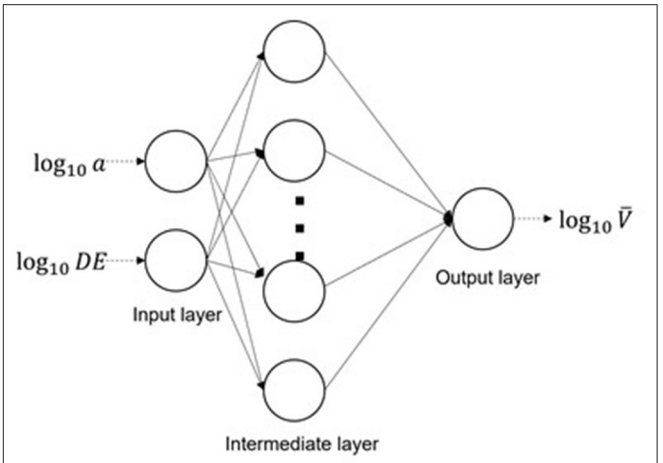


FIGURE 2 | Illustration of an artificial neural network. Each circle is a neuron, basic computation unit, in the network. Each arrow represents a computation connection from the neurons in last layer to the one in current layer.

In this paper the machine learning models were compared only to the Heathcote et al. (2015) regression model. To provide a fair comparison between the machine learning results and regression results, the Heathcote method was also subjected to the 5-fold cross validation.

Artificial Neural Networks

ANN maps input data (x_i) into the output target variable (x_o). It is a computation architecture stacking multiple layers of neurons. Neurons are basic computation units in the ANN and store numbers to proceed to the next step of computation. Layers are a collection of neurons whose computation occurs at the same stage. We use a simple three-layer artificial neural network for illustration purposes (**Figure 2**), which consists of input layers, intermediate layers, and output layers. The input data enters the ANN *via* the input layer and is then transformed into intermediate layer output (x_m), the dimension of which has been predefined. This transformation (Equation 5) firstly linearly transforms x_i and then often adapts a non-linear operator (σ) that takes a non-linear function to introduce non-linearity into the system. x_m is then transformed to yield the final prediction (x_o) as the output in the output layer (Equation 6).

$$x_m = \sigma(W_i^m x_i + b_i) \tag{5}$$

$$x_o = \sigma\left(W_m^O x_m + b_m\right) \tag{6}$$

$$L\left(W_i^m,\ W_m^O, b_i,\ b_m\right) = \frac{1}{N}\sum\nolimits_N \left(x_o - y\right)^2 \tag{7}$$

$$x_i = [log_{10}a,\ log_{10}DE] \tag{8}$$

The predicted output is compared against the given observed data (y) and a loss value is calculated through a loss function L (Equation 7) that often takes a form of root mean squared error for numeric prediction problems, consistent with most regression

problems. N is the number of input data records. Note that the loss L is a function of trainable parameters in an ANN. For this illustration network, there are four trainable parameters—W_i^m, W_m^O, b_i, and b_m. W_i^m and b_i denotes the linear transformation matrix and intercept term that maps x_i to x_m, respectively. W_m^O and b_i functionalize to map x_m to x_o, respectively. Training an ANN will update those trainable parameters until the L reaches a minimum, a process called optimization that adopts specific algorithms to search for optimal trainable parameters.

For the lake volume prediction problem, x_i is a 2-dimensional vector of surface area and lake elevation change in a log scale (Equation 8). x_o is $\log_{10}\tilde{V}$, the predicted volume (log scale), and y is the observed lake volume in log scale ($\log_{10} V$). Although the illustrated ANN architecture adopts a three-layer ANN, in practice, the depth of ANN and the number of neurons of intermediate layers can also vary and is determined empirically. For details of the ANN architecture we used, and other implementation details, please refer to **Appendix A1**.

Note that compared to statistical models, the parameters of ANN models are difficult to interpret mechanistically. Regression coefficients quantify the relationship between independent variables and target variables. In contrast, the learned parameters in ANN functionalize collectively without an explicit interpretation to understand the relationship between input features to outputs. Although some research has attempted to unveil its black-box mechanism (Montavon et al., 2018), its internal functions are still not as transparent and understandable as regression models and thus merits further research efforts to advance its progress.

Temporal Variation of Total Lake Volume

Using the lake volume estimation model based on lake surface area alone, the temporal variation in the total volume of water in the region's lakes was determined using data from the Global Surface Water (GSW) observations program, which is based on LANDSAT imagery at a 30-meter resolution. The first regression formulation, Equation (1) was applied with surface areas derived from the digitized lake maps taken from the GSW data set for the period 2002 through 2015. An example of a digitized map image for two lakes for two dates (one in 2012 and one in 2015) is illustrated in **Figure 3**. The digital cells show the locations where the satellite sensed the presence of water. The blue colored cells show the presence of water in both 2012 and 2015, while the magenta colored cells show the presence of water in 2015, but not in 2012. The surface areas for each lake in the region was determined for each year (for the month of June), the areas were substituted into the regression model (Equation 1) to estimate the volume for each lake, and the total of water volume in the region was calculated by the sum of volumes for all lakes.

RESULTS

Lake Surface Area Model

The bathymetric data for the 816 lakes were used to perform a regression by fitting to the measured surface area and the lake volume calculated from the bathymetric information. The model fit yielded

$$V = 0.256A^{1.13} \quad A \geq 1.25\ km^2 \tag{9}$$

$$V = 0.0328A^{1.236} \quad A < 1.25\ km^2 \tag{10}$$

This model explained 83% of the variability of the lake volume. A plot of the predicted and observed lake volume using this regression is presented in **Figure 4**.

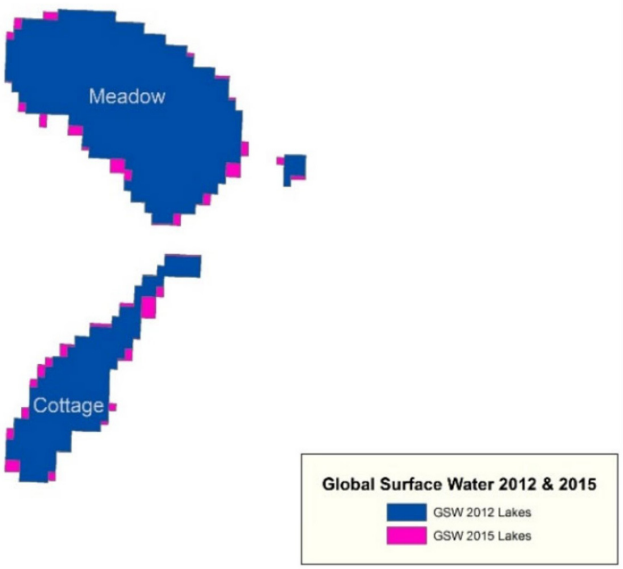


FIGURE 3 | The digital map for two lakes showing the presence of water in the cells; the magenta cells are locations where water was present in 2015 but not in 2012.

Heathcote Method

All Lakes Pooled

Equation (2) represents the Heathcote et al. (2015) model for lake volume. The independent variables in this equation were determined as the best predictors based on the Pearson partial correlation coefficient and AIC test (**Tables 2**, **3**). When comparing the known and predicted lake volumes, the model explained 82% of the variation in lake volume [$R^2 = 0.82$, $F_{(1,812)} = 3{,}811$, $p < 2.2e{-}16$] (**Figure 5**). The surface area and elevation change accounted for 82% and 2% of the variation within the model, respectively. The RSE for the model was 0.282 $\log_{10}$ m^3. The coefficients for the all-lakes pooled data model were $c = 1.17$ and $d = 0.07$. The total lake volume predicted by the model for the pooled lakes was 7.5% different from the known total volume for the 816 lakes (**Figure 5**).

Lakes Grouped by Size

Splitting the lake regression analysis by surface area resulted in an 83% explanation of the variation in lake volume [$R^2 = 0.83$, $F_{(1,812)} = 3{,}835$, $p < 2.2e{-}16$] (**Table 4**). The RSE for the model was 0.281 $\log_{10}$ m^3. The coefficients c and d were different for

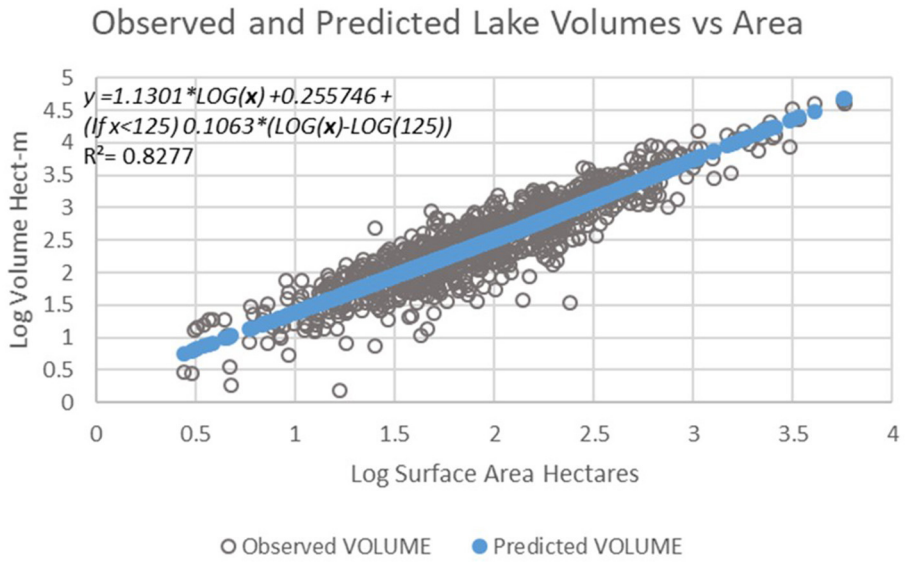


FIGURE 4 | Plot of predicted lake volume vs. observed lake volume for the 816 lakes using the simple regression on lake surface area model.

TABLE 2 | Pearson partial correlation coefficient tested to determine correlation strength of independent variables to lake volume.

Coefficient variables	Lake volume
Surface area	0.90
Elevation change	0.15

TABLE 3 | Akaike information criteria (AIC) and ΔAIC for the different predictive models tested for determining lake volume.

Model variables	AIC	ΔAIC
Surface area + elevation change	317.79	0.0
Surface area	334.89	17.1

each category of lake area, with c ranging from 0.78 to 1.26, and d ranging from −0.04 to 0.75. The total lake volume predicted by the size-segregated regression equations was 1.9% different from the known total lake volume (**Table 4**).

Lakes Grouped by Watershed

Grouping the lakes by watershed resulted in the model explaining 84% of the variation in lake volume [$R^2 = 0.84$, $F_{(1,814)} = 4{,}342$, $p < 2.2e{-}16$] (**Table 5**). The RSE for the model was 0.269 $\log_{10}$ m^3. The coefficients c and d were different for each category of watershed, with c ranging from 0.91 to 1.67, and d ranging from −0.30 to 0.78. The total lake volume predicted by the watershed-segregated regression equations was 2.6% different from the known total lake volume (**Table 5**).

Using the different groupings of lakes, the total volumes were calculated for the 40,054 lakes within the region (**Table 6**). When comparing the three lake groupings, surface size grouping resulted in the highest lake volume with 1,236,436 hectare-meters while the model with all the lakes pooled yielded the lowest lake volume with 1,179,284 hectare-meters, a 4.7% difference (99% confidence interval 1,152,266–1,247,112 hectare-meters).

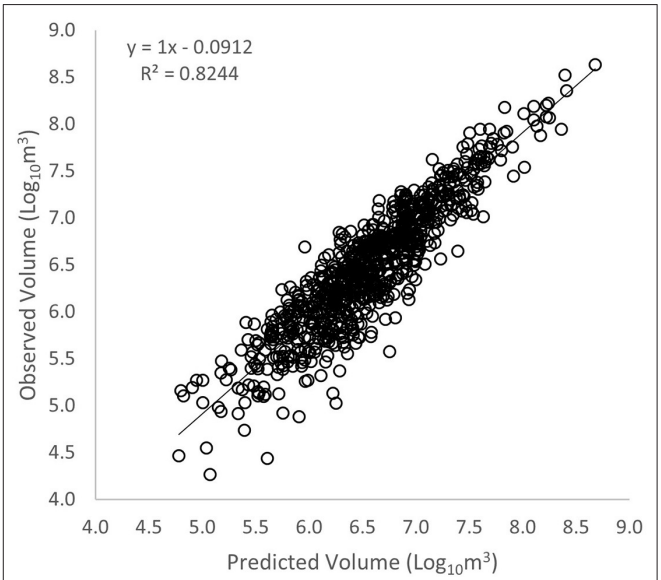


FIGURE 5 | Heathcote method of observed vs. predicted lake volumes for a linear regression model [$R^2 = 0.82$, $F_{(1,812)} = 3{,}811$, $p < 2.2e{-}16$].

Cael Method

The Cael et al. (2017) method uses surface area which is the most significant variable to determine lake volume as seen in the Pearson partial correlation coefficient (**Table 2**). The analysis

TABLE 4 | Total predicted volume by each lake size in the study area based on Heathcote et al. (2015) model.

Size	Known volume (m^3)	Predicted volume (m^3)	Percent difference	Number of lakes (*n*)	c Coefficient	d Coefficient
10^4-10^5	4,761,038	4,272,646	10.3%	25	0.78	0.08
10^5-10^6	806,186,183	726,584,526	9.9%	438	1.12	0.05
10^6-10^7	4,538,884,627	4,478,136,636	1.3%	333	1.26	0.08
$>10^7$	2,692,298,510	2,677,799,591	0.5%	20	0.94	0.75
Total	8,042,130,358	7,886,793,399	1.9%	816	1.17	0.07

Regression analysis using surface area and elevation change in terrestrial buffer was conducted for each size [$R^2 = 0.83$, $n = 816$, $F_{(1,812)} = 3{,}835$, $p < 2.2e{-}16$).

TABLE 5 | Total predicted volume by each watershed in study area based on Heathcote et al. (2015) model.

Watersheds	Known volume (m^3)	Predicted volume (m^3)	Percent differences	Number of lakes (n)	c Coefficient	d Coefficient
Buffalo river	36,350,941	37,246,650	2.4%	19	1.13	0.19
Cannon river	220,456,231	242,749,474	9.6%	32	0.95	−0.04
Crow Wing river	1,359,708,747	1,278,342,131	6.2%	112	1.18	0.75
Long Prairie river	952,569,927	837,593,468	12.8%	51	1.21	0.10
Lower St. Croix river	203,413,119	165,149,343	20.7%	46	1.15	0.02
Mississippi River- Brainerd	624,888,496	676,090,233	7.9%	60	0.91	0.41
Mississippi river—Lake Pepin	13,431,687	14,608,058	8.4%	5	1.05	−0.04
Mississippi river—Sartell	121,815,236	127,774,030	4.8%	35	1.22	−0.17
Mississippi river—St. Cloud	284,449,132	266,029,558	6.7%	85	1.07	−0.08
Mississippi river—Twin Cities	332,693,906	317,921,490	4.5%	110	1.04	0.58
Ottertail river	2,509,976,730	2,597,468,051	3.4%	82	1,12	0.25
Pine river	902,738,899	822,873,334	9.3%	79	1.04	0.54
Redeye river	51,419,386	50,341,558	2.1%	8	1.67	0.18
Rum river	104,355,324	82,893,697	22.9%	25	1.35	0.78
Sauk river	147,591,348	157,164,850	6.3%	49	0.92	0.30
Snake river	61,919,196	55,757,269	10.5%	7	1.57	−0.10
Wild Rice river	114,352,053	104,835,498	8.7%	11	1.40	−0.30
Total	8,042,130,358	7,834,838,692	2.6%	816	1.17	0.07

Regression analysis using surface area and elevation change in terrestrial buffer was conducted for each watershed [$n = 816$, $R^2 = 0.84$, $F_{(1,814)} = 4{,}342$, $p < 2.2e{-}16$].

TABLE 6 | Comparison of total volume of the 40,054 lakes based on three approaches of Heathcote et al. (2015) method (Mille Lacs Lake not included).

Distribution of lakes	Total volume (m^3)
Project area	11,792,840,000
Size	12,364,360,000
Watershed	11,833,470,000

of Cael et al. was for lakes sampled from the US, Canada, and Sweden, and their analysis yielded a Hurst coefficient of 0.41. In our study of the 816 lakes the Cael et al. model yielded

$$V = 10^{-0.498+\varepsilon} A^{1.17} \tag{11}$$

For this model result, the Hurst coefficient is 0.34 which is within the theoretical range (0.4 ± 0.1) for the earth's surface.

When comparing the known and predicted lake volumes based on Equation 11, the model explained 82% of the variation in volume for individual lake volumes [$R^2 = 0.82$, $F_{(1,812)} = 3{,}697$, $p < 2.2e{-}16$] (**Figure 6**). For this same regression equation, the total observed volume to the predicted volumes of the 816 lakes were compared. Our predictions were 1.4% different than that of the observed volume total (**Table 7**). The RSE for the model was 0.296 $\log_{10}$ m^3. After calculating the total volume with the 40,054 lakes by both methods, the difference between the Heathcote et al. (2015) and the Cael et al. (2017) methods for all the lakes pooled was 3% (**Table 8**).

Machine Learning Method

All ML models were trained using the lake surface area and land surface elevation change, both of which are used in the Heathcote method while the Cael method uses only surface area. Therefore, we benchmarked ML methods against the Heathcote method. Without further grouping lake data based on the watershed location or lake surface area size, we used the full dataset for the purpose of investigating ML modeling ability in contrast to statistical regression models. To allow a fair comparison between machine learning methods and regression methods,

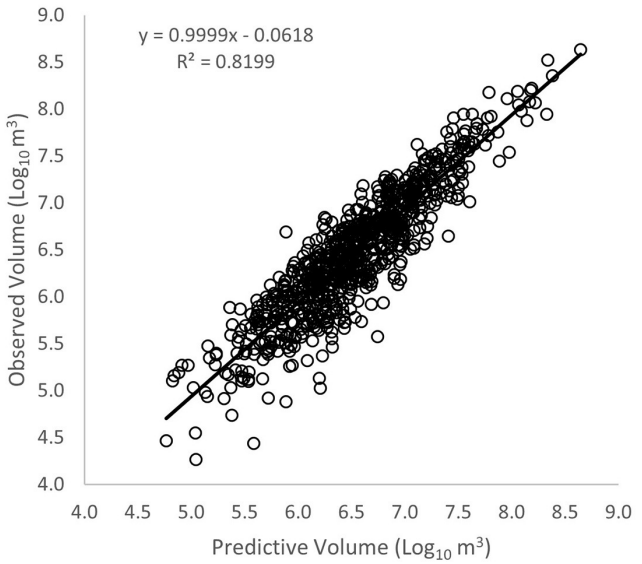


FIGURE 6 | Cael method of observed vs. predicted lake volumes for a linear regression model [$R^2 = 0.82$, $F_{(1,812)} = 3{,}697$, $p < 2.2e{-}16$].

TABLE 7 | Using Cael et al. (2017) method, percent difference between the 816 observed lake volumes and the predicted volumes.

Comparison	Total volume (m3)	Percent difference
Observed volume	8,042,130,358	-
Predicted volume	7,928,754,557	1.4%

TABLE 8 | Heathcote et al. (2015) vs. Cael et al. (2017) total volume comparison for all lakes pooled.

Method	Total volume (m^3)
Heathcote, all lakes pooled	11,792,840,000
Cael, all lakes pooled	12,113,930,000

the Heathcote method is evaluated using the cross-validation approach as well. Note that 5-fold cross-validation will evaluate the models using five different portions of the testing data and thus yield five different testing metrics. The less variant those testing metrics are, the more stable the corresponding models behave. As shown in **Table 9**, averages of the R^2 and RMSE values across 5-fold validations are reported. Meanwhile, the standard deviation across 5-fold validations is also reported to show the stability of the model performance. Among all the ML models, although the ANN testing performance is less stable during cross-validation than the Heathcote method, ANN exhibits the best predictive performance with a RMSE of 0.286 and a R^2 of 0.819 in contrast to the Heathcote method (0.296 RMSE and 0.811 R^2). Besides, SVR (0.291 RMSE score and 0.819 R^2) also achieves slightly better predictive performance than the Heathcote method. Both RF and GT yield a predictive performance slightly worse than, if not comparable to, the

TABLE 9 | ML methods comparison against the Heathcote method results in a 5-fold cross validation test.

Models	R^2	RMSE
Heathcote method	0.811 (0.11)	0.296 (0.009)
ANN	0.819 (0.041)	0.286 (0.035)
SVR	0.819 (0.026)	0.291 (0.017)
RF	0.789 (0.004)	0.311 (0.020)
GT	0.809 (0.024)	0.296 (0.017)

Both R^2 and RMSE shows the average of testing performance. The number in the parentheses is the standard deviation.

Heathcote method. The result is that the ANN model yielded the best predictive performance.

Temporal Variation of Total Lake Volume in Central Minnesota Region

The data acquired from the GSW observation program was used to determine the surface areas of lakes on an annual basis for the study region. Those surface areas for the over 40,000 lakes were substituted into the regression model (Equation 1) and the volumes summed for all lakes. The resulting temporal variation of the total lake water stored (in equivalent mm) in the region is illustrated in **Figure 7**. There is a clear drop in water stored in the lakes in 2011–2012. Those years corresponded to a period of rainfall deficit.

DISCUSSION

Regression Methods

All three regression models, the simple regression given by Equation (1), the regression given by the Heathcote et al. (2015) model (Equation 2), and the Cael et al. (2017) model (Equation 3), provided fairly accurate predictions of the lake volumes for the 816 surveyed lakes. Among these, the Heathcote et al. model provided the best representation of the known individual lake volumes, while the Cael et al. model provided the best representation of the total volume of lake water in the region.

When comparing our research to the Heathcote et al. (2015) research, the lake surface area of lakes in Central Minnesota has a larger correlation to lake volume than that of the buffer elevation difference. This may be because of there being a smaller range of elevation within the study area, being a relatively flat region, resulting in the elevation difference in the buffer having a weaker relationship. The Heathcote et al. (2015) study compared 433 lakes selected from five different regions, two of which were situated in a mountainous region. When comparing the five regions, the mountainous region models produced the most accurate lake volumes as well as the highest R^2 ($R^2 > 0.90$). The regions with less elevation change such as the Eastmain region resulted in R^2 similar to the results reported herein for Central Minnesota's R^2 ($R^2 \approx 0.80$). This affirms the hypothesis that when the elevation has a larger range, the estimate of lake volume will have a stronger correlation to surface elevation change (Heathcote et al., 2015).

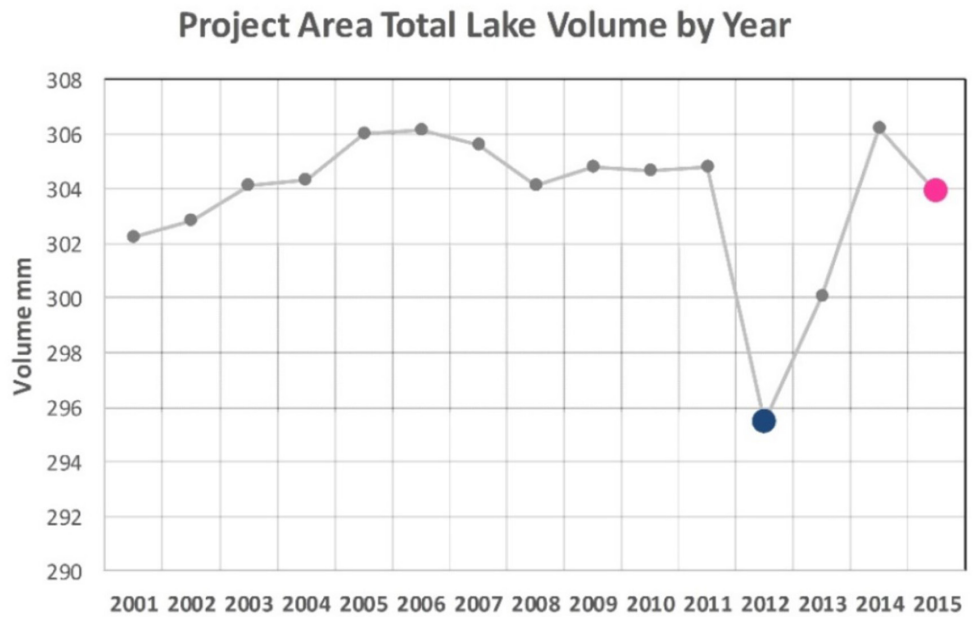


FIGURE 7 | The temporal variation in total water stored in the lakes of the region. Note the sharp drop in volume in the year 2011–2012, and the gradual recovery in following 2 years.

Among the three groupings of lakes using the Heathcote et al. (2015) procedure, determining lake volume by watershed resulted in the best prediction. A reason why the watershed grouping was the best prediction when compared to the known volumes is most likely that the lakes within a given subregion (or watershed) are almost all formed by the same geomorphic process, resulting in the lakes' formation being similar. Like Heathcote et al. (2015), we assumed that similar processes formed the lake and their landscape.

One issue with the analysis for all of the methods, regression and machine learning is that no bathymetric data exist for lakes smaller than 10^4 m^2 surface area. To fill in this data, it was assumed that a lake smaller than 10^4 m^2 surface area had an average depth of 0.5 m. This, of course, imposes an error in the data for a very large number of lakes that exist in the region. The predominance of larger lakes in the bathymetric data set is clear from **Table 1**, and it is clear from the estimates of total lake volume for the region that most of the total volume, about 66%, is contained in the 816 recorded lakes. The remaining 39,000+ lakes for which estimates were made contained the smaller fraction of the total volume. One improvement that could be made for the development of the prediction models would be to increase the amount of bathymetric data for the lakes in the small size range.

Another source of error in the analysis for the Heathcote et al. (2015) model was the use of a 1/3 arc-second DEM (U.S. Geological Survey, 2017). This approach essentially eliminated elevation data for lakes smaller than 10^4 m^2 due to the lakes being too small for the DEM to pick up the elevation difference. For further research, DEM data with better resolution should be used in order to predict volumes more accurately by obtaining the buffer elevations from the smaller lakes.

It is not clear why the regression coefficient for the elevation change was negative for some of the data sets involving watershed groupings and lake area groupings. Theoretically, the coefficients should be positive. Perhaps the resulting negative coefficients occurred from less accurate elevation measurements resulting from the coarse DEM resolution. Further analysis is needed to determine the cause of the negative coefficients.

While this study is only limited to central Minnesota, an independent study covering the full state was completed to determine if the Heathcote et al. (2015) approach can accurately predict the lake volume for lakes across the entire state of Minnesota. For example, using the surface area and elevation change for lakes >4,047 m^2 across the entire state of Minnesota, Griffin et al. (2018) and Finlay (2019) used the Heathcote et al. (2015) method to estimate lake volumes for the purpose of quantifying the regional variability of DOM pools in the water column of the region's lakes. Based on preliminary research, the model explained 82% of the variation in the lake volume with over 1,000 lakes of 4,047 m^2 or larger. This research reaffirmed that using the lake's surface area and surrounding landscape can be used to accurately predict a lake's volume and can be used in diverse geographic areas with little morphologic and bathymetric data available.

The results for the Cael et al. (2017) model yielded a Hurst coefficient of 0.34 for the lakes in the Central Minnesota region. Cael et al. applied the method to four regions some of which had topographic features more like the Central Minnesota landscapes (Sweden, Wisconsin, some parts of Quebec), while

others were more mountainous, for example the Adirondack region of New York. The resulting Hurst coefficients derived for these different regions reflected the topography of the individual regions. The Hurst coefficients derived by Cael et al. were 0.24 for the Wisconsin region, 0.32 for Sweden, 0.33 for Quebec (data included mountainous as well as more flat regions), and 0.48 for the Adirondack region. With all regions combined the derived Hurst coefficient was 0.40. This demonstrated that the Hurst coefficient picks up the topographic features through the relationship formed between lake surface area and lake volume.

In order to see whether, like the Heathcote et al. (2015) method, the Cael et al. (2017) method can have its lakes grouped by size and watershed, the lakes were grouped by the same categories. The significance of predicting total lake volumes when comparing the total volume of known lakes within the region was decreased when splitting into groups. Meaning that grouping the lakes by surface area size and watershed did not produce any significant results. Therefore, having a larger set of lakes when comprising the Cael et al. (2017) model improves the predictability of total lake volume. Even though the Cael et al. (2017) method was unable to significantly predict total lake volume when grouped by surface area and watershed, both the Heathcote et al. (2015) and the Cael et al. (2017) method were both able to significantly predict volumes when pooling all lakes together.

While both methods predict lake volume, the Cael et al. (2017) method, by design, is better suited to predict a group of lakes rather than individual lakes. The Heathcote et al. (2015) method is better at predicting volume and depth for individual lakes and therefore can be used when calculating individual lake processes. Consequently, one method may be more advantageous than the other depending on what future research questions are being asked.

Machine Learning Method

Although the popularity of ML seemingly makes it a strong candidate approach for our lake volume predictions, a drastic improvement of the lake volume prediction accuracy is not observed in our case. Even though, among them, the ANN yields the best performance and suggests that its modeling ability to capture complex data patterns is more pronounced than the other three alternative models.

RF yields the relatively worst performance, which is likely caused by the low dimensions of input features (2-D data of lake surface area and elevation change) and its data hungry characteristics. Prediction tasks often benefit from the RF modeling because RF automatically finds uniform input feature subspace. However, given a 2-D input feature, the advantage of subspace searching is not leveraged. Further, a collection of 816 lakes is not a rich dataset for RF and would easily make RF overfit the training data and produce worse testing performance.

GT and SVR yielded comparable performance to the regression method. ANN exhibited the best performance among the selected ML methods and is slightly better than regression approaches. The reason for such a negligible performance improvement is possibly because the Heathcote method has achieved a performance satisfactory enough

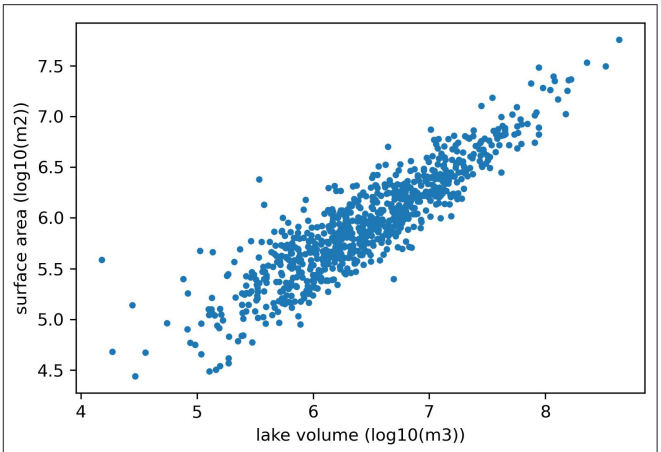


FIGURE 8 | The relationship between lake volume (log scale) and lake surface area (log scale).

that the performance improvement room for ANN is too small. As shown in **Figure 8**, the correlation between $\log_{10}(A)$ and $\log_{10}(V)$ is as high as 0.90, which suggests a limited non-linear complexity between input data and lake volume. Such a limited non-linear data pattern constrained the ANN predictive performance improvement in contrast to the Heathcote method.

All ML models show relatively more variant testing performance in contrast to the Heathcote method, which suggests the randomness in machine learning models and the uncertainty in its trainable parameters. On the contrary, the regression style Heathcote method preserves consistent testing performance (lower standard deviation of the testing performance in the cross-validation evaluation), which implies that linear regression models' generalization performance is more stable than ML for this problem.

Although ANN shows relatively better prediction accuracy, it does not have well-understood mechanisms underlying its explanatory power. For the Heathcote method, regression coefficients can offer sufficient interpretation to understand models. The positive regression coefficient of lake surface area and its statistical significance indicates the significant contribution of the surface area variable to lake volume estimation. However, this insight is missing for the ANN model.

Additionally, ANN only takes a 2-dimension input, which collectively groups all lakes together without any distinguishment among individual lakes. The model lacks distinct lake awareness information that might help more accurately predict volumes. It is likely that lake surface area and elevation change does not contain sufficient additional information for the volume prediction that is not already captured in the linear regression models. Therefore, it would be necessary to provide more physical information of lakes, such as, more lake geometry information, and surrounding land surface features, to further improve lake volume prediction accuracy.

Although the benefit of applying a machine learning model is not obvious for lake volume in our results, other bathymetric characterization of lakes, such as, lake depth may gain more from this approach. Heathcote et al. (2015) reported that a statistical model for predicting maximum lake depth only explains half of the system variance, which suggests that the majority of lake depth variance is difficult for statistical models to explain. Converse to the linear relationship between lake area and lake volume, relationships among other lake morphology features may be more complex. We hypothesize that this complexity is also accompanied with hidden non-linearities, which provides another research opportunity for implementing machine learning models and exploring their predictive capability in the future.

CONCLUSION

We predicted lake volume through Central Minnesota using readily available morphologic data and a variety of previously published and novel methods. Three regression-based analysis methods and four machine learning methods were applied to develop predictions of lake volumes for over 40,000 lakes located in the central section of Minnesota. The methods were developed using detailed lake bathymetric data for 816 lakes located in the same region. The resulting prediction methods estimated the total volume of lake water in the region to be in the range of about 12 ± 0.2 km^3.

The regression models included a regression on lake surface area, a model based on the Heathcote et al. (2015) model that included lake surface area and mean elevation change in a designated buffer area outside the lake area, and a model based on the Cael et al. (2017) model that utilized the theory of self-affine surfaces. Among the machine learning models, the ANN performed the best, and it was found that the ANN performance was slightly better than any of the regression models. The small incremental benefit in performance of the ANN method over the regression models is explained by the fact that the relation between log-transformed lake surface area and log-transformed lake volume is nearly linear. If the relation were more non-linear, the ML methods might have been able to provide a larger increase in performance. This is the power of ML approaches, in that they facilitate the development of data-driven models when the relations between variables are complex and non-linear. One immediate future need is to evaluate the ability of ML methods for prediction of lake maximum depth.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: Lake data: https://hdl.handle.net/11299/211726, Global surface water: https://global-surface-water.appspot.com/download.

AUTHOR CONTRIBUTIONS

JN conceived the project idea and acquired the funding. CD conducted statistical method prediction and prepared the first draft of this manuscript with the guidance kindly offered from AH. XL conducted the ML experiments and wrote their counterparts in the manuscript. KH developed the volume-surface area regression model and analyzed the annual trend analysis. BW supervised the statistical analysis design. CD, XL, AH, and JN all proofread and revised the manuscript. All authors contributed to the article and approved the submitted version.

FUNDING

This study was financially supported by the Legislative-Citizen Commission on Minnesota Resources (M.L. 2017, Chp. 96, Sec. 2, Subd. 04h).

ACKNOWLEDGMENTS

JN effort on this project was partially supported by the USDA National Institute of Food and Agriculture, Hatch/Multistate Project MN 12-109. BW effort on this project was partially supported by the USDA National Institute of Food and Agriculture, Hatch/Multistate Project MN 12-069.

REFERENCES

Adadi, A. (2021). A survey on data-efficient algorithms in big data era. *J. Big Data.* 8, 1–54. doi: 10.1186/s40537-021-00419-9

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Trans. Automat. Contr.* 19, 716–723. doi: 10.1109/TAC.1974.1100705

Avand, M., Kuriqi, A., Khazaei, M., and Ghorbanzadeh, O. (2022). DEM resolution effects on machine learning performance for flood probability mapping. *J. Hydro-Environ. Res.* 40, 1–16. doi: 10.1016/j.jher.2021.10.002

Borchers, H. W. (2019). *Package "pracma"* (2.2.5). R Foundation for Statistical Computing, Vienna, Austria. Available online at: http://CRAN.Rproject.org/package=pracma

Breiman, L. (2001). Random forests. *Mach. Learn.* 45, 5–32. doi: 10.1023/A:1010933404324

Cael, B. B., Heathcote, A. J., and Seekell, D. A. (2017). The volume and mean depth of Earth's lakes. *Geophys. Res. Lett.* 44, 209–218. doi: 10.1002/2016GL071378

Chen, T., and Guestrin, C. (2016). XGBoost: a scalable tree boosting system. *Knowled. Discov. Data Mining* 16, 651–662. doi: 10.1145/2939672.2939785

Cortes, C., and Vapnik, V. (1995). Support-vector networks. *Mach. Learn.* 20, 273–297. doi: 10.1007/BF00994018

Crétaux, J. F., Abarca-del-Río, R., Bergé-Nguyen, M., Arsen, A., Drolon, V., Clos, G., et al. (2016). Lake volume monitoring from space. *Surv. Geophys.* 37, 269–305. doi: 10.1007/s10712-016-9362-6

Delaney, C. O. (2019). *Estimating lake water volume using scale analysis* (M.S.Thesis). University of Minnesota, St Paul, MN, United States.

Deoli, V., Kumar, D., Kumar, M., Kuriqi, A., and Elbeltagi, A. (2021). Water spread mapping of multiple lakes using remote sensing and satellite data. *Arab. J. Geosci.* 14, 1–15. doi: 10.1007/s12517-021-08597-9

Department of Natural Resources Division of Fish and Wildlife (2014). *DNR Hydrography - Lakes and Open Water*. Available online at: http://www.mngeo.state.mn.us/committee/standards/mgmg/metadata.htm%0A (accessed November 29, 2018).

Dreyfus, S. E. (1990). Artificial neural networks, back propagation, and the Kelley-Bryson gradient procedure. *J. Guid. Control Dyn.* 13, 926–928. doi: 10.2514/3.25422

Ferguson, R. I. (1986). River loads underestimated by rating curves. *Water Resour. Res.* 22, 74–76. doi: 10.1029/WR022i001p00074

Finlay, J. (2019). *Assessment of surface water quality with satellite sensors.* Final Report, project funded by the Legislative-Citizens Committee on Minnesota Resources, Legal Citation: M.L. 2016, Chp. 186, Sec. 2, Subd. 04i.

Griffin, C. G., Holmberg, K., Delaney, C., Olmanson, L. G., Brezonik, P. L., Nieber, J., et al. (2018). "Remote sensing of dissolved organic matter pools in lakes at regional scales," in *IGU Fall Meeting Conference, Vol. 2018* (Washington, DC: American Geophysical Union).

Håkanson, L., and Karlsson, B. (1984). On the relationship between regional geomorphology and lake morphometry-A Swedish example. Geografiska Annaler: Ser. *A, Phys. Geograph*. 66, 103–119. doi: 10.1080/04353676.1984.11880102

Håkanson, L., and Peters, R. H. (1995). *Predictive Limnology*. Amsterdam: SPB Academic.

Heathcote, A. J., del Giorgio, P. A., and Prairie, Y. T. (2015). Predicting bathymetric features of lakes from the topography of their surrounding landscape. *Can. J. Fish. Aquat. Sci.* 72, 643–650. doi: 10.1139/cjfas-2014-0392

Hirschberg, J., and Manning, C. D. (2015). language processing. *Science.* 349, 261–266. doi: 10.1126/science.aaa8685

Hollister, J. W., Milstead, W. B., and Urrutia, M. A. (2011). Predicting maximum lake depth from surrounding topography. *PLoS ONE* 6, e25764. doi: 10.1371/journal.pone.0025764

Jia, X., Willard, J., Karpatne, A., Read, J. S., Zwart, J. A., Steinbach, M., et al. (2020). Physics-guided machine learning for scientific discovery: an application in simulating lake temperature profiles. *ACM IMS Trans. Data Sci.* 2, 1–25. doi: 10.1145/3447814

Kim, S. (2015). *Package "ppcor"* (1.1). R Foundation for Statistical Computing, Vienna, Austria. Available online at: https://cran.r-project.org/web/packages/ppcor/ppcor.pdf

Kratzert, F., Klotz, D., Brenner, C., Schulz, K., and Herrnegger, M. (2018). Rainfall – runoff modelling using Long Short-Term Memory (LSTM) networks. *Hydrol. Earth Syst. Sci.* 22, 6005–6022. doi: 10.5194/hess-22-6005-2018

Kratzert, F., Klotz, D., Shalev, G., Klambauer, G., Hochreiter, S., and Nearing, G. (2019). Towards learning universal, regional, and local hydrological behaviors via machine learning applied to large-sample datasets. *Hydrol. Earth Syst. Sci.* 23, 5089–5110. doi: 10.5194/hess-23-5089-2019

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature* 521, 436–444. doi: 10.1038/nature14539

Leschinski, C. H. (2019). *Package 'MonteCarlo'* (1.0.6). R Foundation for Statistical Computing, Vienna, Austria. Available online at: https://CRAN.R-project.org/package=MonteCarlo

McDonald, C. P., Rover, J. A., Stets, E. G., and Striegl, R. G. (2012). The regional abundance and size distribution of lakes and reservoirs in the United States and implications for estimates of global lake extent. *Limnol. Oceanogr.* 57, 597–606. doi: 10.4319/lo.2012.57.2.0597

Messager, M. L., Lehner, B., Grill, G., Nedeva, I., and Schmitt, O. (2016). Estimating the volume and age of water stored in global lakes using a geo-statistical approach. *Nat. Commun.* 7, 1–11. doi: 10.1038/ncomms13603

Minnesota Department of Natural Resources (2017). *Lake Basin Morphology*. St Paul, MN: Minnesota DNR, Division of Fish and Wildlife.

Montavon, G., Samek, W., and Müller, K. R. (2018). Methods for interpreting and understanding deep neural networks. *Digit. Signal Process. A Rev. J.* 73, 1–15. doi: 10.1016/j.dsp.2017.10.011

Obertegger, U., Flaim, G., Braioni, M. G., Sommaruga, R., Corradini, F., and Borsato, A. (2007). Water residence time as a driving force of zooplankton structure and succession. *Aquat. Sci.* 69, 575–583. doi: 10.1007/s00027-007-0924-z

Pekel, J.-F., Cottam, A., Gorelick, N., and Belward, A. S. (2016). High-resolution mapping of global surface water and its long-term changes. *Nature* 540, 418–422. doi: 10.1038/nature20584

Renard, F., Candela, T., and Bouchaud, E. (2013). Constant dimensionality of fault roughness from the scale of micro-fractures to the scale of continents. *Geophys. Res. Lett.* 40, 83–87. doi: 10.1029/2012GL054143

RStudio Team (2016). *RStudio: Integrated Development for R*. Boston, MA: RStudio, Inc. (1.1.456).

Sejnowski, T. J. (2020). The unreasonable effectiveness of deep learning in artificial intelligence. *Proc. Natl. Acad. Sci. U. S. A.* 117, 30033–30038. doi: 10.1073/pnas.1907373117

Shen, C., Laloy, E., Elshorbagy, A., Albert, A., Bales, J., Chang, F. J., et al. (2018). HESS opinions: incubating deep-learning-powered hydrologic science advances as a community. *Hydrol. Earth Syst. Sci.* 22, 5639–5656. doi: 10.5194/hess-22-5639-2018

Shukla, R., Kumar, P., Vishwakarma, D. K., Ali, R., Kumar, R., and Kuriqi, A. (2022). Modeling of stage-discharge using back propagation ANN-, ANFIS-, and WANN-based computing techniques. *Theor. Appl. Climatol.* 147, 867–889. doi: 10.1007/s00704-021-03863-y

Sobek, S., Nisell, J., and Folster, J. (2011). Predicting the volume and depth of lakes from map-derived parameters. *Inland Waters* 1, 177–184. doi: 10.5268/IW-1.3.426

U.S. Geological Survey. (2017). *1/3rd Arc-Second Digital Elevation Models (DEMs) – USGS National Map 3DEP Downloadable Data Collection*. Reston, VA: U.S. Geological Survey.

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

APPENDIX A1

Hyper-parameters of the applied ML models are listed below. Those values were determined after hyper-parameter tuning.

Artificial Neural Network

Activation function for each layer: ReLu.
Model architecture: input (2d) -> 4d -> 16d -> 32d -> (output) 1d
Optimization algorithm: Adam optimizer (learning rate: 0.001).

Random Forest

Number of trees: 100
Maximum tree depth: 8.

Support Vector Machine

Radial basis function kernel.

Gradient Boosted Regression Tree

Number of trees: 80.

APPENDIX A2

Abbreviation Glossary

MnDNR, Minnesota Department of Natural Resources.
HUC-8, Hydrologic unit codes.
GSW, Global surface water.
DEM, Digital Elevation map.
AIC, Akaike information criteria.
RMSE, Root-mean-square error.
RSE, Relative standard error.
ML, Machine Learning.
ANN, Artificial neural network.
SVR, Support vector regression.
GT, Gradient boosted regression tree.
RF, Random forests.
CDOM, colored dissolved organic matter.
DOC, dissolved organic carbon.
LANDSAT, Satellite that studies and photographs the surface by using remote-sensing techniques.

Variable Glossary

V, Volume.
a, b, c, d, empirical constants.
A, Lake surface area.
D, Buffer distance from the shoreline outward.
DE25, 25% of the average elevation changes within the buffer.
H, Hurst Coefficient.
ζ, volume-area scaling exponent.
κ, proportionality coefficient.
ε, error term.
x_i, input data.
x_{o}, output target variable.
x_{m}, intermediate layer output.
σ, non-linear operator.
$\mathbf{W_i^m}$, $\mathbf{W_m^o}$, trainable parameters (weight matrix in neural network layers).
$\mathbf{b_m}$, $\mathbf{b_o}$, trainable parameters (bias terms in neural network layers).
L, Loss Function.
N, Number of input data records.
y, observed data.