

University of Nebraska - Lincoln

DigitalCommons@University of Nebraska - Lincoln

---

Civil and Environmental Engineering Faculty  
Publications

Civil and Environmental Engineering

---

2021

## Great Lakes Runoff Intercomparison Project Phase 3: Lake Erie (GRIP-E)

Juliane Mai  
*University of Waterloo*

Bryan A. Tolson  
*University of Waterloo*

Hongren Shen  
*University of Waterloo*

Etienne Gaborit  
*Environment and Climate Change Canada*

Vincent Fortin  
*Environment and Climate Change Canada*

Follow this and additional works at: <https://digitalcommons.unl.edu/civilengfacpub>



Part of <https://digitalcommons.unl.edu/civilengfacpub> Civil Engineering Commons

---

Mai, Juliane; Tolson, Bryan A.; Shen, Hongren; Gaborit, Etienne; Fortin, Vincent; Gasset, Nicolas; Awoye, Herve; Stadnyk, Tricia A.; Fry, Lauren M.; Bradley, Emily A.; Seglenieks, Frank; Temgoua, Andre G.T.; Prncz, Daniel G.; Gharari, Shervan; Haghnegahdar, Amin; Elshamy, Mohamed E.; Razavi, Saman; Gauch, Martin; Lin, Jimmy; Ni, Xiaojing; Yuan, Yongping; McLeod, Meghan; Basu, Nandita B.; Kumar, Rohini; Rakovec, Oldrich; Samaniego, Luis; Attinger, Sabine; Shrestha, Narayan K.; Daggupati, Prasad; Roy, Tirthankar; Wi, Sungwook; Hunter, Tim; Craig, James R.; and Pietroniro, Alain, "Great Lakes Runoff Intercomparison Project Phase 3: Lake Erie (GRIP-E)" (2021). *Civil and Environmental Engineering Faculty Publications*. 235.

<https://digitalcommons.unl.edu/civilengfacpub/235>

This Article is brought to you for free and open access by the Civil and Environmental Engineering at DigitalCommons@University of Nebraska - Lincoln. It has been accepted for inclusion in Civil and Environmental Engineering Faculty Publications by an authorized administrator of DigitalCommons@University of Nebraska - Lincoln.

---

## Authors

Juliane Mai, Bryan A. Tolson, Hongren Shen, Etienne Gaborit, Vincent Fortin, Nicolas Gasset, Herve Awoye, Tricia A. Stadnyk, Lauren M. Fry, Emily A. Bradley, Frank Seglenieks, Andre G.T. Temgoua, Daniel G. Princz, Shervan Gharari, Amin Haghnegahdar, Mohamed E. Elshamy, Saman Razavi, Martin Gauch, Jimmy Lin, Xiaojing Ni, Yongping Yuan, Meghan McLeod, Nandita B. Basu, Rohini Kumar, Oldrich Rakovec, Luis Samaniego, Sabine Attinger, Narayan K. Shrestha, Prasad Daggupati, Tirthankar Roy, Sungwook Wi, Tim Hunter, James R. Craig, and Alain Pietroniro



# Great Lakes Runoff Intercomparison Project Phase 3: Lake Erie (GRIP-E)

Juliane Mai<sup>1</sup>; Bryan A. Tolson<sup>2</sup>; Hongren Shen<sup>3</sup>; Étienne Gaborit<sup>4</sup>; Vincent Fortin<sup>5</sup>; Nicolas Gasset<sup>6</sup>; Hervé Awoye<sup>7</sup>; Tricia A. Stadnyk<sup>8</sup>; Lauren M. Fry<sup>9</sup>; Emily A. Bradley<sup>10</sup>; Frank Seglenieks<sup>11</sup>; André G. T. Temgoua<sup>12</sup>; Daniel G. Princz<sup>13</sup>; Shervan Gharari<sup>14</sup>; Amin Haghnegahdar<sup>15</sup>; Mohamed E. Elshamy<sup>16</sup>; Saman Razavi<sup>17</sup>; Martin Gauch<sup>18</sup>; Jimmy Lin<sup>19</sup>; Xiaojing Ni<sup>20</sup>; Yongping Yuan<sup>21</sup>; Meghan McLeod<sup>22</sup>; Nandita B. Basu<sup>23</sup>; Rohini Kumar<sup>24</sup>; Oldrich Rakovec<sup>25</sup>; Luis Samaniego<sup>26</sup>; Sabine Attinger<sup>27</sup>; Narayan K. Shrestha<sup>28</sup>; Prasad Daggupati<sup>29</sup>; Tirthankar Roy<sup>30</sup>; Sungwook Wi<sup>31</sup>; Tim Hunter<sup>32</sup>; James R. Craig<sup>33</sup>; and Alain Pietroniro<sup>34</sup>

**Abstract:** Hydrologic model intercomparison studies help to evaluate the agility of models to simulate variables such as streamflow, evaporation, and soil moisture. This study is the third in a sequence of the Great Lakes Runoff Intercomparison Projects. The densely populated Lake Erie watershed studied here is an important international lake that has experienced recent flooding and shoreline erosion alongside excessive nutrient loads that have contributed to lake eutrophication. Understanding the sources and pathways of flows is critical to solve the complex issues facing this watershed. Seventeen hydrologic and land-surface models of different complexity are set up over this domain using the same meteorological forcings, and their simulated streamflows at 46 calibration and seven independent validation stations are compared. Results show that: (1) the good performance of Machine Learning models during calibration decreases significantly in validation due to the limited amount of training data; (2) models calibrated at individual stations perform equally well in validation; and (3) most distributed models calibrated over the entire domain have problems in simulating urban areas but outperform the other models in validation. **DOI:** [10.1061/\(ASCE\)HE.1943-5584.0002097](https://doi.org/10.1061/(ASCE)HE.1943-5584.0002097). This work is made available under the terms of the Creative Commons Attribution 4.0 International license, <https://creativecommons.org/licenses/by/4.0/>.

<sup>1</sup>Research Assistant Professor, Dept. of Civil and Environmental Engineering, Univ. of Waterloo, Waterloo, ON, Canada N2L 3G1 (corresponding author). ORCID: <https://orcid.org/0000-0002-1132-2342>. Email: [juliane.mai@uwaterloo.ca](mailto:juliane.mai@uwaterloo.ca)

<sup>2</sup>Professor, Dept. of Civil and Environmental Engineering, Univ. of Waterloo, Waterloo, ON, Canada N2L 3G1.

<sup>3</sup>Ph.D. Student, Dept. of Civil and Environmental Engineering, Univ. of Waterloo, Waterloo, ON, Canada N2L 3G1. ORCID: <https://orcid.org/0000-0002-5979-2159>

<sup>4</sup>Research Scientist, Meteorological Research Div., Environment and Climate Change Canada, Dorval, QC, Canada H9P 1J3. ORCID: <https://orcid.org/0000-0002-9787-9124>

<sup>5</sup>Research Scientist, Meteorological Research Div., Environment and Climate Change Canada, Dorval, QC, Canada H9P 1J3. ORCID: <https://orcid.org/0000-0002-2145-4592>

<sup>6</sup>Physical Scientist, Meteorological Service of Canada, Environment and Climate Change Canada, Dorval, QC, Canada H9P 1J3. ORCID: <https://orcid.org/0000-0003-4542-4408>

<sup>7</sup>Postdoctoral Scholar, Dept. of Geography, Univ. of Calgary, Calgary, AB, Canada T2N 1N4.

<sup>8</sup>Associate Professor, Dept. of Geography, Univ. of Calgary, Calgary, AB, Canada T2N 1N4.

<sup>9</sup>Research Physical Scientist, Integrated Physical and Ecological Modeling and Forecasting Branch, Great Lakes Environmental Research Laboratory, National Oceanic and Atmospheric Administration, Ann Arbor, MI 48108; formerly, US Army Corps of Engineers, Detroit District, Great Lakes Hydraulics and Hydrology Office, Detroit, MI 48226. ORCID: <https://orcid.org/0000-0002-5480-5408>

Note. This manuscript was submitted on October 16, 2020; approved on February 16, 2021; published online on June 30, 2021. Discussion period open until November 30, 2021; separate discussions must be submitted for individual papers. This paper is part of the *Journal of Hydrologic Engineering*, © ASCE, ISSN 1084-0699.

<sup>10</sup>Hydraulic Engineer, US Army Corps of Engineers, Detroit District, Great Lakes Hydraulics and Hydrology Office, Detroit, MI 48226.

<sup>11</sup>Water Resources Engineer, National Hydrological Services, Environment and Climate Change Canada, Burlington, ON, Canada L7S 1A1.

<sup>12</sup>Water Resources Engineer, National Hydrological Services, Environment and Climate Change Canada, Burlington, ON, Canada L7S 1A1.

<sup>13</sup>Coordinator and Officer Environmental Programs, National Hydrological Services, Environment and Climate Change Canada, Saskatoon, SK, Canada S7N 3H5.

<sup>14</sup>Research Associate, Coldwater Laboratory, Univ. of Saskatchewan, Canmore, AB, Canada T1W 3G1.

<sup>15</sup>Associate Member, Global Institute for Water Security, Univ. of Saskatchewan, Saskatoon, SK, Canada S7N 3H5.

<sup>16</sup>Research Scientist, Global Institute for Water Security and Center for Hydrology, Univ. of Saskatchewan, Saskatoon, SK, Canada S7N 3H5. ORCID: <https://orcid.org/0000-0002-3621-0021>

<sup>17</sup>Associate Professor, School of Environment and Sustainability, Dept. of Civil, Geological and Environmental Engineering, and Global Institute for Water Security, Univ. of Saskatchewan, Saskatoon, SK, Canada S7N 3H5.

<sup>18</sup>Ph.D. Student, Institute for Machine Learning, Johannes Kepler Univ., 4040 Linz, Austria; formerly, David R. Cheriton School of Computer Science, Univ. of Waterloo, Waterloo, ON, Canada N2L 3G1.

<sup>19</sup>Professor, David R. Cheriton School of Computer Science, Univ. of Waterloo, Waterloo, ON, Canada N2L 3G1.

<sup>20</sup>Oak Ridge Institute for Science and Education (ORISE) Postdoctoral Fellow, Office of Research and Development, US Environmental Protection Agency, Research Triangle Park, NC 27711. ORCID: <https://orcid.org/0000-0001-6696-4804>

<sup>21</sup>Senior Research Hydrologist, Office of Research and Development, US Environmental Protection Agency, Research Triangle Park, NC 27711.

## Introduction

Hydrologic modeling efforts in the Great Lakes region of North America are a fundamental component for the understanding and management of water resources impacting more than 30 million people in Canada and the United States (USEPA, n.d.). At the larger scale, these models are used for operational forecasting of Great Lakes water levels or for evaluating the impacts of climate change. At the smaller or regional scales, these models can be used for reservoir management, water supply, floodplain mapping, and a multitude of other applications. Multimodel development and careful model intercomparisons can be a critical part of improving simulation or forecasting skills, as well as capturing the uncertainty in any future climate assessments (Wada et al. 2013; Warszawski et al. 2014; McSweeney and Jones 2016; Rosenzweig et al. 2017; Frieler et al. 2017; Huang et al. 2017).

There is a long history of hydrologic model comparison efforts, many of which are outlined below. The following review of past hydrologic model intercomparison studies is focused mainly on studies that (1) compare models in watersheds of at least hundreds of km<sup>2</sup> with a sample size of at least 10 watersheds; and (2) include the participation of multiple independent modeling groups.

The River Forecasting Centers of the National Weather Service (NWS) in the United States launched a Distributed Model Inter-Comparison project (DMIP) in the early 2000s to explore the skill of distributed hydrologic models for river forecasting compared with the operational, lumped models in place at the NWS at that time. In a first phase, 12 distributed models set up by 12 international groups were compared to the lumped Sacramento (Burnash et al. 1973; Burnash and Singh 1995) model (SAC-SMA) set up by the NWS Office of Hydrologic Development (Smith et al. 2004) for flood operations. The effort was initiated to create an organized and controlled experiment to compare streamflow simulations using identical forcing, verification periods, and modeling points. Reed et al. (2004) found that the skill of the model operator has a significant impact on the model performance, and were not able to identify a model that clearly outperforms other models. Similarly, in a very recent paper by Menard et al. (2020) on snow-model intercomparison, the skill of the modeler was found to be an overarching determinant in model performance. In the second phase of the DMIP experiment, DMIP-2 (Smith et al. 2012), data quality and quantity were improved along with the selection of more evaluation points and variables beyond traditional streamflow estimates. Fifteen participating distributed models and two lumped models (SAC-SMA and GR4J) were evaluated in DMIP-2. Results showed that the

distributed models outperform lumped models at modeled interior points of watersheds regarding hourly streamflow. Although not clearly emphasized in the study, one distributed model was superior to other models in several aspects analyzed. That model is the Hydrology Laboratory-Research Distributed Hydrologic Model (HL-RDHM) set up by the NWS Office of Hydrologic Development, referred to as the OHD model in their study.

Another relevant study is the Protocol for the Analysis of Land Surface Models (PALS) for the Land Surface Model Benchmarking Evaluation Project (PLUMBER) study by Best et al. (2015). This study focused on land surface rather than traditional hydrologic models. One novelty of this study was that the complex models were compared against a baseline of purely data-driven regression models. The study surprisingly found that these data-driven models reliably outperformed the complex physically based models regarding simulated sensible and latent heat fluxes, although the reason for this was not explored.

Another relevant multimodel comparison project is the Model Parameter Estimation Experiment (MOPEX), which primarily focused on model calibration techniques for hydrologic models, but also allowed for hydrologic model comparison. One researcher noted that “MOPEX workshops have been convened to bring together interested hydrologists and land surface modelers from all over world to exchange knowledge and experience” (Duan et al. 2006). One important objective of the second and third MOPEX workshops was to compare eight hydrologic models over 12 catchments in the southeastern United States. Core findings of Duan et al. (2006) are that model calibration has a huge potential to improve model performance and that different models represent hydrologic processes differently, hence suggesting that model ensembles may be a way to improve predictions. This set of catchments has since been used as a benchmark set of catchments for many further studies (Kavetski and Clark 2010; Herman 2012; Evin et al. 2014; Cuntz et al. 2016; McInerney et al. 2017; Spieler et al. 2020).

A more recent intercomparison study by de Boer-Euser et al. (2017) focused on the evaluation of multiple models by showing that, although they perform similarly for the entire flow regime, they show clear differences during specific events. In this study, a blind validation was performed to test the models’ transferability in space and time. For these experiments, the modelers were given only forcing data but no discharge observations. The study found that the 11 mostly lumped models performed similarly based on general metrics such as the Nash–Sutcliffe efficiency (NSE). Significant differences in performance could only be diagnosed with a closer look at predictions on an event-by-event basis.

<sup>22</sup>Master Candidate, Earth and Environmental Sciences, Univ. of Waterloo, Waterloo, ON, Canada N2L 3G1.

<sup>23</sup>Associate Professor, Earth and Environmental Sciences and Water Institute, Univ. of Waterloo, Waterloo, ON, Canada N2L 3G1.

<sup>24</sup>Research Associate, Computational Hydrosystems, Helmholtz Centre for Environmental Research–UFZ, Leipzig 04318, Germany.

<sup>25</sup>Research Associate, Faculty of Environmental Sciences, Czech Univ. of Life Sciences, Prague 16500, Czech Republic; Computational Hydrosystems, Helmholtz Centre for Environmental Research–UFZ, Leipzig 04138, Germany. ORCID: <https://orcid.org/0000-0003-2451-3305>

<sup>26</sup>Deputy Head of Department, Computational Hydrosystems, Helmholtz Centre for Environmental Research–UFZ, Leipzig 04138, Germany.

<sup>27</sup>Head of Department, Computational Hydrosystems, Helmholtz Centre for Environmental Research–UFZ, Leipzig 04138, Germany.

<sup>28</sup>Postdoctoral Research Fellow, School of Engineering, Univ. of Guelph, Guelph, ON, Canada N1G 2W1.

<sup>29</sup>Assistant Professor, School of Engineering, Univ. of Guelph, Guelph, ON, Canada N1G 2W1.

<sup>30</sup>Assistant Professor, Dept. of Civil and Environmental Engineering, Univ. of Nebraska–Lincoln, Lincoln, NE 68182.

<sup>31</sup>Research Assistant Professor, Dept. of Civil and Environmental Engineering, Univ. of Massachusetts, Amherst, MA 01003.

<sup>32</sup>Computer Specialist, Integrated Physical and Ecological Modeling and Forecasting Branch, Great Lakes Environmental Research Laboratory, National Oceanic and Atmospheric Administration, Ann Arbor, MI 48108. ORCID: <https://orcid.org/0000-0003-1423-6770>

<sup>33</sup>Associate Professor, Dept. of Civil and Environmental Engineering, Univ. of Waterloo, Waterloo, ON, Canada N2L 3G1. ORCID: <https://orcid.org/0000-0003-2715-7166>

<sup>34</sup>Professor, Dept. Civil Engineering, Univ. of Calgary, Calgary, AB, Canada T2N 1N4; formerly, Executive Director, National Hydrological Services, Environment and Climate Change Canada, Saskatoon, SK, Canada S7N 3H5.

As these previous studies have shown, model intercomparisons can help to understand the differences between models. They can identify locations where some models outperform others, and thus they learn from each other. Model intercomparison projects (MIPs) are usually time-intensive efforts that can be quite challenging because of the wide range of modeling approaches, especially in hydrology. Such differences arise from forcing, basin segmentation, and physical detail differences in the modeling approach. Aggregation or regridding of input forcing data are model specific, with some models requiring many more surface variables. There are also variations in the use of geophysical data—most conceptualized models might not use geophysical data at all (e.g., GR4J, Perrin et al. 2003), while more physically based models classify soil and land cover types (e.g., VIC, Liang et al. 1994; Liang 2003) and others might make direct use of gridded soil and land cover maps at their native resolution (e.g., mHM, Samaniego et al. 2010; Kumar et al. 2013).

After addressing the time-consuming nature and the need for expert knowledge required to set up several models over the same domain, one must then decide which model outputs to compare. In hydrologic models, streamflow is the fundamental and most important and common output variable because many other output variables (e.g., soluble nutrients) depend on the accurate simulation of streamflow. Note that the streamflow at a basin's outlet is the integration of all hydrological processes within the basin. Other hydrologic variables, such as evaporation, soil moisture, and snow depth, are typically measured at the point scale, but often simulated at a variety of spatial and temporal scales within each model, making intercomparison difficult. Moreover, these state variables may or may not be explicitly simulated in some conceptual or Machine Learning models. Last, decisions about model evaluation metrics and calibration strategies need to be made. Several of these decisions are dependent on the region and the research questions in place.

International collaborators conducted a sequence of recent model intercomparison studies in subregions of the Great Lakes, called the Great Lakes Runoff Model Intercomparison Project (GRIP). The sequence started in 2014 with Lake Michigan (GRIP-M), which is entirely located in the United States (Fry et al. 2014). The intercomparison was performed for five models ranging from lumped to distributed physical models. The simulated runoff was compared at 20 stations, while the models used different model inputs and forcing. In 2017, a study was performed for the Lake Ontario watershed (GRIP-O) located on both sides of the Canada-US border (Gaborit et al. 2017a). Two lumped models were compared using two different precipitation forcings, resulting in very satisfactory simulation performances independent of model and forcing used. This was even true when using the simple area-ratio method to extrapolate measured flows to the entire watershed including ungauged locations. Gaborit et al. (2017b) extended the set of models to three distributed models (GEM-Hydro, MESH-CLASS, and WATFLOOD) setup over Lake Ontario, similar to GRIP-O. The models are forced with the same meteorologic variables and geophysical dataset. They share the same calibration algorithm and objective function. The study shows that GEM-Hydro is competitive with MESH and WATFLOOD regarding streamflow.

This sequence of GRIP model intercomparisons is extended in this study to Lake Erie, which is the shallowest of the Great Lakes. Lake Erie is prone to algae blooms with toxic concentration (Michalak et al. 2013; Schmale et al. 2019). The Lake Erie watershed is significantly affected by urban and agricultural runoff such as overfishing, pollution, and eutrophication (Thompson et al. 2019; Ho and Michalak 2017; Dolan and McGunagle 2005) as well as record high water levels resulting in lakeshore flooding, erosion, and record high outflows to Lake Ontario (IJC 2019a, b; USACE 2020).

The Lake Erie watershed is, however, home to about one-third of the population of the entire Great Lakes basin. Approximately 12 million people live in 17 metropolitan areas such as metro Detroit (MI), Cleveland (OH), Toledo (OH), Buffalo (NY), and Orono (ON).

There is significant interest among researchers and operational modelers to have various research and operational models assessed for agility and adequacy across the watershed, as demonstrated by the number of operational and research groups (14) agreeing to participate in the intercomparison. The goals of this study are to assess if models can be identified that outperform others, and if there are locations in the watershed that cannot be modeled reliably; furthermore, a third goal is to evaluate differences between modeling agencies and model classification, such as comparing lumped and distributed models against data-driven models. It is worth noting that previous model intercomparison initiatives in hydrology (except PLUMBER) have largely focused on investigating process-based models (conceptual or physically based) but did not include data-driven models as we did in this study.

In this study, we compared three distinct groups of models, with 17 models in total. These models comprise two data-driven systems serving as baseline benchmarks, seven lumped and locally calibrated, and eight distributed and globally calibrated models. The models are configured and implemented by a group of 33 international collaborators over the Lake Erie watershed using data derived from the same gridded forcing inputs and observed streamflow as the calibration target. These models were evaluated at 46 calibration sites, and seven validation sites were tested in a blind validation similar to de Boer-Euser et al. (2017). The setup used in this study was intended to be a living database, making it easy to add additional models for evaluation if future model revisions lead to more accurate streamflow simulations. The study informs how to design and then formulate research questions for a follow-up model intercomparison over the entire Great Lakes (GRIP-GL) similar to the DMIP and DMIP-2 approach. GRIP-E is also meant to identify a subset of higher quality models for inclusion in the future GRIP-GL study.

## Materials and Methods

This section will introduce the major study organization: the study domain; the comparison of the two objectives regarding the models; and the streamflow gauging stations at which the models are compared. The participating models will then be briefly introduced, while the more detailed model descriptions are provided in the Supplemental Materials. Following this, the meteorologic forcing and geophysical datasets will be introduced. Last, the model comparison procedure will be explained.

### Study Organization and Decision Making

This study represents a significant multiyear effort requiring the collaboration of more than 30 researchers from 20 organizations. Prior to inviting teams to participate in the study, the project leads made three key study design decisions: (1) participating teams would be forced to utilize exactly the same forcing data; (2) the project datasets and model outputs would be archived in a way that this study could serve as a living intercomparison benchmark; and (3) all other study design decisions would be made by consensus by all collaborators in the project. This required participants to agree to monthly meetings, and these meetings continued over a period of 2 years. In addition, as the final study design matured, we continued to advertise open participation widely and brought in several new participating teams, even into the second year of the project. These key decisions and our collaborative approach to study design

created buy-in and trust, and helped motivate participating teams (almost all of which had no explicit funding to participate) to stay involved.

### Modeling Domain

The modeling domain for Lake Erie is specified by the Great Lakes Aquatic Habitat Framework (GLAHF 2016) (Wang et al. 2015) (Fig. 1). The total area of the local Lake Erie watershed is 103,666 km<sup>2</sup>, of which 76,352 km<sup>2</sup> is the land surface draining into the lake. The shapefile of the modeling domain and subwatersheds are available in the folder “shapefiles” on the GitHub associated with this publication (see Data Availability Statement for URL).

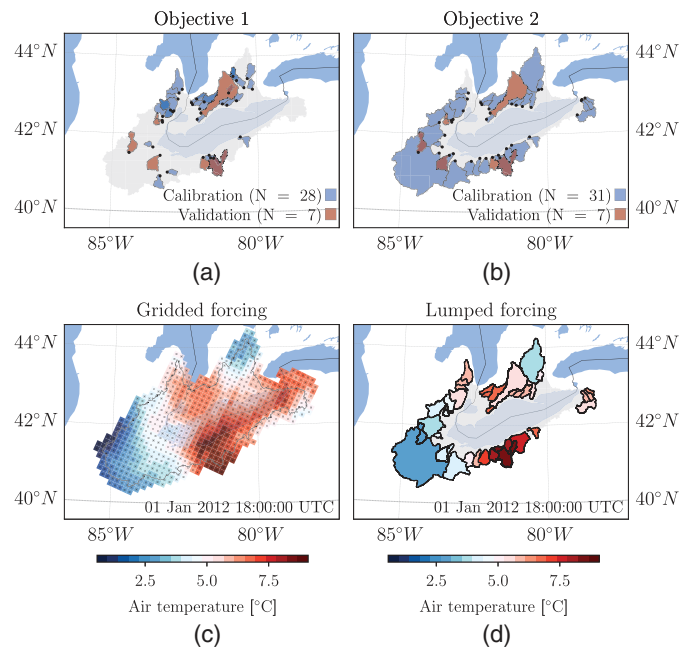
The average elevation along the shoreline is 132 m, while the highest elevation in the local Lake Erie watershed is 697 m south of Buffalo (eastern-most corner of watershed), making this basin relatively low relief. The mean elevation of the study area is 238 m based on the 15 arcsec ( $\approx 500$  m) resolved NASA’s shuttle radar topography mission (SRTM)-conditioned HydroSHEDS digital elevation map (Lehner et al. 2008). Besides water (25.6%), the study area is primarily cropland (51.3%), savannas (e.g., 10%–60% tree cover with canopy >2 m, 12.2%), forest (2.7%), and urban areas (5.9%). The largest urban areas are metro Detroit (MI), Cleveland (OH), Toledo (OH), Buffalo (NY), and London (ON). These estimates were derived using the MODIS/Terra+Aqua Land Cover Type Yearly L3 Global 500m SIN Grid V006 (MCD12Q1\_006) provided by the USGS (Friedl and Sulla-Menashe 2019). The main soil classes are heavy clay (15.7%) mostly north of Lake Erie, some patches of silt loam (4.9%) also north of the lake, and a patch of silty clay (2.8%) northwest of the lake. Everywhere else, the soil is predominately silt (50.0%) or covered with waterbodies (26.6%). These estimates were derived using the Harmonized World Soil Database v1.2.

### Objectives of Model Intercomparison

This study focused on two different objectives to compare the models: (1) model performance in absence of human impacts; and (2) model performance and reliability for simulating lake water budgets. The first objective is regarded to be an easier task for hydrologic models when nonnatural impacts, such as reservoir operations, urban sealed areas, and irrigation do not need to be considered. The second objective was selected to evaluate the models’ performance in generating reliable inflows to the lake to capture the important requirements of understanding the lake water budget. Participating groups were given the option of contributing model results for one or both of these objectives.

### Streamflow Gauging Stations

The streamflow gauging stations were selected based on institutionally important streamflow gauge stations from both NOAA-GLERL (Hunter, personal communication, 2018) and Environment and Climate Change Canada (Seglenieks, personal communication, 2018) and have been used in previous studies (Haghnegahdar 2015). From this proposed set of stations, only stations with drainage areas above 200 km<sup>2</sup> were selected to avoid catchments with flashy behavior. To classify regulation in their databases, the Water Survey Canada (WSC) and USGS labels of nonregulated/regulated and reference/nonreference were used to group stations. Only stations with tag non-regulated and reference were selected for objective 1. For the second objective, the most downstream stations of the set were selected regardless of whether they are regulated or not. The lower limit of a 200-km<sup>2</sup> drainage area was also applied to objective 2. The data were then collected from WSC and USGS and formatted into NetCDF



**Fig. 1.** (Color) The subbasins used in this study for (a) objective 1 (low-human impact watersheds); and (b) objective 2 (most downstream gauges) are highlighted with their colored delineated shape. Blue shapes indicate subbasins used for calibration and red shapes indicate validation basins. The Lake Erie watershed is shown as reference as a light grey shaded area. (c) The 15-km gridded forcing data are shown for one point in time (January 1, 2012, 6:00 p.m. UTC). (d) Some models processed the gridded forcings into, for example, lumped forcings per subbasin.

including unit conversion to ( $\text{m}^3\text{s}^{-1}$ ) for USGS stations that are reported in imperial units ( $\text{ft}^3\text{s}^{-1}$ ). The data are available in the folders, “data/objective\_1” and “data/objective\_2” on the GitHub associated with this publication (see Data Availability Statement for URL). In total, 28 stations were used for objective 1 and 31 stations for objective 2, while 13 appear in both objectives because they are most downstream and of low human impact. A detailed list of these gauges, including their drainage areas and locations, can be found in the Supplemental Materials (Table S1). The delineated subwatersheds draining to the selected stations are shown in Figs. 1(a and b) for objectives 1 and 2, respectively. An additional set of seven streamflow gauge stations was used for spatial validation. Please note that there was no temporal validation performed in this study due to the short period of available meteorologic forcings. All seven stations have a drainage area larger than 200 km<sup>2</sup> and are not separated into objective 1 or objective 2 stations, but rather were used to evaluate both objectives. Based on results obtained during calibration, an improved quality check in selecting these validation stations was applied. All stations were visually inspected and stations in highly urbanized areas were discarded.

### Participating Models

The models participating can be grouped into three groups of approaches: Machine Learning models; models that are calibrated for each individual subwatershed; and models that are calibrated for the entire domain. The individual models in each group are briefly described, respectively, in the following subsections, and a much more comprehensive description of the specific model setups and details about the calibration can be found in the Supplemental

**Table 1.** List of participating models specifying coauthors responsible for model setups, calibration, and runs

Model	Lead	Calibration	Temporal	Spatial
Name	Modeler(s)	Strategy	Resolution	Resolution
<i>Machine Learning models:</i>				
ML-LSTM	Gauch & Lin	Global	Daily	Subwatersheds (46)
ML-XGBoost	Gauch & Lin	Global	Daily	Subwatersheds (46)
<i>Hydrologic models with calibration of each gauge individually (local calibration):</i>				
LBRM	Fry & Bradley	Local	Daily	Subwatersheds (28)
GR4J-lp	Shen & Tolson	Local	Daily	Subwatersheds (46)
GR4J-sd	Shen & Tolson	Local	Daily	Subwatersheds (154 obj. 1; 466 obj. 2)
HYMOD2-DS	Roy & Wi	Local	Daily	Subwatersheds (each containing multiple cells of 15 km)
SWAT-EPA	Ni & Yuan	Local	Daily	HRU
SWAT-Guelph	Shrestha & Daggupati	Local	Daily	HRU
mHM-Waterloo	McLeod, Kumar, & Basu	Local	Hourly	Gridded: 0.125°
<i>Hydrologic and land-surface models with calibration of entire domain (global calibration):</i>				
mHM-UFZ	Rakovec, Samaniego, & Attinger	Global	Hourly	Gridded: 0.125°
HYPE	Awoye & Stadnyk	Global	Daily	Subwatersheds (644)
VIC	Shen & Tolson	Global	Hourly	Gridded: tiles (~1,390)
VIC-GRU	Gharari	Global	Hourly	GRUs (~2,380)
GEM-Hydro	Gaborit	Global	LSS 5 min; Rout. 30–600 s	Gridded: LSS 10 km; Rout. 1 km
MESH-SVS	Gaborit & Princz	Global	LSS 5 min; Rout. 30–600 s	Gridded: LSS 10 km; Rout. 10 km (calib) and 1 km (final run)
MESH-CLASS	Haghnegahdar, Elshamy, & Princz	Global	LSS 30 min; Rout. 30–600 s	Gridded: LSS 10 km; Rout. 1 km
WATFLOOD	Seglenieks & Temgoua	Global	Hourly	Gridded: 10 km (calib); 1 km (final run)

Note: Models were either calibrated for each individual gauging station (local) or had a global setup for all gauges or gauges of one objective. Further, the temporal and spatial resolution for each model was specified. The models were separated into three groups (see italic captions in table): machine learning (ML) models, hydrologic models that are calibrated at each gauge, and models that have a global setup. The datasets used to setup the models are listed in Table 2. A more detailed specification, including version numbers and forcing data preparation, is available in the “Model\_setups\_GRIP.xlsx” on the GitHub associated with this publication (see Data Availability Statement for URL). Please be aware that different resolutions might be used for the land-surface scheme (LSS) and the routing component (Rout.).

Materials. A list of all models including the leading modelers, calibration strategy, and spatiotemporal resolution can be found in Table 1. A much more detailed version of this table, including details on data preparation, can be found online in the “Model\_setups\_GRIP.xlsx” on the GitHub associated with this publication (see Data Availability Statement for URL).

Note that the decision was made that all participating modeling teams would be free to decide how to calibrate their models. Most were calibrated automatically. Some were calibrated stepwise and in a combination of manual and automatic calibration. It was agreed that unifying the calibration methods would not present a fair comparison, given the range in complexity and type of models. Details about the parameters calibrated, calibration budgets, and algorithms and strategies can be found in the Supplemental Materials.

### Machine Learning Models

Two types of data-driven models were used in this study: gradient-boosted regression tree (GBRT) framework XGBoost (Chen and Guestrin 2016) and long short-term memory (LSTM) architecture (Hochreiter and Schmidhuber 1997). The models are called ML-XGBoost and ML-LSTM in this study. The models were trained for the entire domain corresponding to a global calibration similar to the models discussed in the Models Calibrated for Entire Domain: Global Calibration section. The global setup allowed a seamless computation of the spatial validation results. Objectives 1 and 2 streamflow gauging stations were handled together, meaning that the models were trained with all 46 stations together leading to a single final model setup. The Machine Learning models were built and fitted to data just like the traditional hydrologic

models in that streamflow before day  $t$  is not considered an independent variable to predict streamflow on day  $t$ .

### Models Calibrated for Individual Subbasins: Local Calibration

Seven models were set up and calibrated for each of the 46 streamflow gauging station subwatersheds. These models range from conceptual lumped systems to gridded semidistributed approaches; however, they were all calibrated individually for each specific subwatershed.

Participating teams each decided independently that locally calibrated models are either transferred to validation basins using proximate donor basins and an area-ratio method (Fry et al. 2014) or the nearest-neighbor approach. Details on which donor station was used to derive the estimate of each validation station can be found in the README files for each of the models described in this section (e.g., “data/objective\_1/lake-erie/validation/model/HYMOD2-DS/README.md” on the on the GitHub associated with this publication).

The large basin runoff model (LBRM) was described in Croley (1983), with recent modifications described in Gronewold et al. (2017). LBRM is a lumped conceptual model that propagates daily precipitation and temperature into subbasin runoff. LBRM is the only rainfall-runoff model that is used operationally to produce forecasts of runoff for use in water-level forecasts on a seasonal to interannual basis. It is set up and run by the USACE–Detroit District. Hence, the model is discretized into the 29 subwatersheds important for the USACE. Because the primary application of LBRM is simulating total runoff contribution to the Great Lakes, calibration conventionally uses area-ratio estimates of runoff for

each operational subbasin of interest (which are defined by outlets typically downstream of gauging stations) as observations, rather than using streamflow gauge data directly. This approach was used in the current study. In reality, some of the project validation gauges may have been incorporated to the area-ratio estimates used for calibration. As a result, information from those validation gauges may be contained within the calibration dataset. The results for LBRM at the validation stations are, hence, not considered in the results and discussion (although they are shown for completeness).

The GR4J rainfall-runoff model is a parsimonious lumped model with four parameters and is usually operated at a daily scale (Perrin et al. 2003). The original four-parameter GR4J model is coupled with the two-parameter CemaNeige snow module (Valéry et al. 2014). The GR4J models are implemented in the Raven hydrologic modeling framework (Craig et al. 2020). The model is set up in two different modes: a fully lumped version where each of the 46 subwatersheds are handled as one production storage, and a semidistributed version where each subwatershed is discretized into subunits that are then routed to the outlet. In this study, these models are called GR4J-Raven-lp and GR4J-Raven-sd, respectively. The nearest-neighbor method is used to derive the streamflow time series for the validation stations.

HYMOD (Boyle et al. 2000) is a conceptual hydrologic model for catchment-scale simulation of rainfall-runoff processes. In this study, we used a modified version of the original HYMOD model, the new HYMOD2 (Roy et al. 2017) in a distributed setup (DS), named HYMOD2-DS. The 12 parameters of the model were calibrated for each subwatershed independently and validation estimates were derived using the nearest-neighbor method.

The soil and water assessment tool (SWAT) model is a semidistributed process-based hydrologic and water quality model considering the physical characteristics of the watershed including surface elevation, soil type, land use, and factors affecting water routing within the watershed (Arnold et al. 1993; Neitsch et al. 2011). Two setups of SWAT contributed to this study: The first one is set up by USEPA and will be referred to as SWAT-EPA herein. The second one, SWAT-Guelph, is set up by the University of Guelph. Due to personnel adjustments at USEPA during the course of this project, SWAT-EPA only contributed a setup for objective 1 in calibration mode, but was not able to provide objective 2 calibration results and any validation results. The validation results for SWAT-Guelph are derived using the area-ratio method.

The mesoscale hydrologic model (mHM) (Samaniego et al. 2010; Kumar et al. 2013) is a distributed hydrologic model that uses grid cells as a primary hydrologic unit and accounts for a variety of hydrologic processes including canopy interception, root-zone soil moisture, infiltration, evapotranspiration, and runoff generation, as well as river flows along the stream network (Thober et al. 2019). A unique feature of mHM is its novel multiscale parameter regionalization (MPR) scheme to account for the subgrid variability of basin physical properties that allows for the seamless predictions of water fluxes and storages at different spatial resolutions and ungauged locations (Rakovec et al. 2016; Samaniego et al. 2017). The key parts of the MPR scheme are (1) utilization of transfer functions to translate high-resolution spatial data into high-resolution model parameters; and (2) upscaling of high-resolution model parameters to the model spatial scale. For this study, mHM was applied to simulate the water balance only, but the model is capable of simulating land-surface temperature that is relevant for the energy balance calculations (Zink et al. 2018). It should be noted that two versions of mHM were setup independently by two different participating teams in this study. The first version is a locally calibrated setup leading to 46 optimal mHM setups, one for each of the 46 subwatersheds. The nearest-neighbor method was

then used to derive estimates for the validation stations—this version will be called mHM-Waterloo in the following. The second setup, named mHM-UFZ, is the global version of mHM and will be explained briefly in the next section.

### Models Calibrated for Entire Domain: Global Calibration

The third group of modeling approaches are those where models are calibrated for the entire domain, and hence provide one final model setup after calibration, allowing for a seamless derivation of streamflow time series at any point within the study domain. Almost all models were calibrated using the complete set of 46 gauging stations regardless of the objective (low human impact versus most downstream gauges), except for two models (VIC and mHM-UFZ), which were calibrated once with only gauges of objective 1 and a second time with only gauges of objective 2, leading to two final model setups. A brief description of each model is provided below.

The globally calibrated version of the mHM (Samaniego et al. 2010; Kumar et al. 2013) is the mHM-UFZ, set up by UFZ Leipzig. The mHM-UFZ model is the same as the one used for the local calibration (see last paragraph in previous section), but is globally calibrated once for objective 1 stations and once for objective 2 stations. The setup also used different input datasets for a digital elevation map (DEM), soil, and landcover (Table 2) than the locally calibrated version (mHM-Waterloo).

The hydrological predictions for the environment (HYPE) model is an operational hydrologic model developed at the Swedish Meteorological and Hydrological Institute. HYPE includes hydrologic processes such as snow/ice accumulation and melt, evapotranspiration, soil moisture, frozen soil infiltration, groundwater movement and aquifer recharge, surface-water routing through rivers and lakes, and human perturbations through diversion, reservoirs, regulation, irrigation, and water abstractions (Lindström et al. 2010).

The variable infiltration capacity (VIC) model is a macroscale distributed hydrologic model that balances both the water and surface energy budgets (Liang et al. 1994; Liang 2003). VIC simulates land surface-atmospheric fluxes of moisture and energy such as evapotranspiration, surface runoff, baseflow, radiative fluxes, turbulent fluxes of transport, and sensible heat within the grid-cell. The gridded runoff components, comprising surface runoff and baseflow, are then routed to the basin outlet. VIC was represented twice in this study—once in its native setup using grid-cells to discretize watersheds, and in a second, independent setup using the concept of grouped response units (GRUs) (Gharari et al. 2020). The model will hence be called VIC-GRU. VIC was calibrated twice—once with only objective 1 stations and once with only objective 2 stations—while VIC-GRU used all stations and provided only one final, calibrated setup.

The last four models described below (GEM-Hydro, MESH-SVS, MESH-CLASS, WATFLOOD) are closely related in that they all share the same preprocessed model input data and a similar basin segmentation approach.

GEM-Hydro is a physically based, distributed hydrologic model developed at Environment and Climate Change Canada (ECCC). It relies on GEM-Surf (Bernier et al. 2011) to represent five different surface tiles (glaciers, water, ice over water, urban, and land). The land tile is represented with the SVS (soil, vegetation, and snow) hydrologic land surface scheme (HLSS) (Alavi et al. 2016; Husain et al. 2016). GEM-Hydro also relies on WATROUTE (Kouwen 2010), a 1D hydraulic model, to perform 2D channel and reservoir routing. See Gaborit et al. (2017b) for more information on GEM-Hydro.



**Table 2.** Geophysical datasets used to set up models, including their source and native resolution

Name	Resolution	ML- Src	ML- LSTM	ML- XGBoost	LBRM	GR4J- Raven-lp	GR4J- Raven-sd	HYMOD2- DS	SWAT- EPA	SWAT- Guelph	mHM- Waterloo	mHM- UFZ	HYPE	VIC	VIC- GRU	GEM- Hydro	MESH- SVS	MESH- CLASS	MESH- WATFLOOD
<i>Digital elevation model (DEM):</i>																			
HydroSHEDS	3 in. $\approx$ 90 m	[1]						x	x	x									
HydroSHEDS	15 in. $\approx$ 500 m	[1]	x	x		x	x				x		x	x					
HydroSHEDS	30 in. $\approx$ 1 km	[1]														x <sup>a</sup>	x <sup>a</sup>	x <sup>a</sup>	x <sup>a</sup>
Printed topo map	Similar to HUC8	—			x														
Global multiresolution terrain elevation data (GMTED; 2010)	7.5 in. $\approx$ 250 m	[2]										x							
USGS GTOPO30 (1996)	1 km	[3]														x <sup>b</sup>	x <sup>b</sup>	x <sup>b</sup>	
<i>Soil database:</i>																			
FAO HWSD v1.2	30 in. $\approx$ 1 km	[4]									x		x	x					
GSDE	30 in. $\approx$ 1 km	[5]														x	x	x	x
SoilGrids	250 m	[6]										x							
SSURGO (US)	1:12,000	[7]							x										
STATSGO2 (US)	1:250,000	[7]								x									
SLC v3.2 (CA)	1:1 million	[8]							x	x									
<i>Landcover:</i>																			
MODIS MCD12q1 v6	500 m	[9]								x				x					
NALCMS 2005	250 m	[10]													x				
ESA-CCI LC 2015	300 m	[11]														x	x	x	x
ESA-CCI GlobCover 2005-06 v2.2	0.002778° $\approx$ 300 m	[12]										x							
ESA-CCI GlobCover 2009 v2.3	0.002778° $\approx$ 300 m	[12]							x		x			x					
USDA Croplayer database (CDL) 2010 (US)	30 m	[13]							x										
<i>Others:</i>																			
Thickness of soil, regolith, and sediment deposit layers	30 in. $\approx$ 1 km	[14]												x					
Groundwater table depth	30 in. $\approx$ 1 km	[15]												x					
Canadian national hydrology network (NHN)	1:50,000 or better	[16]														x <sup>c</sup>	x <sup>c</sup>	x <sup>c</sup>	
National hydrography dataset of the United States (NHD)	Up to 1:24,000	[17]														x <sup>c</sup>	x <sup>c</sup>	x <sup>c</sup>	
Global depth to bedrock	250 m	[18]																	x

Sources: [1] (HydroSHEDS 2021), [2] (USGS 2010), [3] (USGS 2018), [4] (FAO 2008), [5] (Shangguan et al. 2008), [6] (SoilGrids 2020), [7] (USDA, n.d.), [8] (Agriculture and Agri-Food Canada 2010), [9] (USGS 2019), [10] (CEC 2010), [11] (ESA 2017), [12] (ESA 2005), [13] (USDA 2021), [14] (ORNL and DAAC 2016), [15] (USC 2013), [16] (Natural Resources Canada 2020), [17] (USGS 2021), and [18] (Shangguan et al. 2017).

Note: The datasets might have been postprocessed before being used for individual models. These details can be found online in the “Model\_setups\_GRIP.xlsx” on the GitHub associated with this publication (see Data Availability Statement for URL). All models used the same meteorologic forcings (Table 3).

<sup>a</sup>Dataset used for delineation and routing.

<sup>b</sup>Dataset used for surface/soil slope computation.

<sup>c</sup>Dataset used for computation of drainage density (km/km<sup>2</sup>).

MESH (modélisation environnementale communautaire–surface and hydrology) is a complimentary community hydrologic modeling platform maintained by ECCC (Pietroniro et al. 2007). The MESH framework includes SVS among its HLSSs, as well as the Canadian LAnd Surface Scheme (CLASS) (Verseghy 2000), which is a physically based land surface scheme requiring several forcing data and simulates vertical energy and water fluxes for vegetation, soil, and snow. The two models are referred to as MESH-SVS and MESH-CLASS in this study. MESH uses a grid-based modeling system and accounts for subgrid heterogeneity using the GRU concept from WATFLOOD (Kouwen et al. 1993). GRUs aggregate subgrid areas by common attributes (e.g., soil and vegetation characteristics) and facilitate parameter transferability in space. Like GEM-Hydro, MESH also uses WATROUTE (Kouwen 2010) for channel and reservoir routing. MESH-SVS was used to calibrate GEM-Hydro: MESH-SVS was calibrated first and the SVS parameters were transferred to GEM-Hydro. MESH-CLASS participants inadvertently used two of the stations that are later used for validation during calibration.

The WATFLOOD model is a partially physically based, distributed hydrologic model (Kouwen 1988). The hydrologic processes modeled in WATFLOOD include, but are not limited to, interception, infiltration, evaporation, snow accumulation, interflow, recharge, baseflow, and overland and channel routing. The most important concept of WATFLOOD is the GRU approach, as described above for MESH. The runoff response from each unit with an individual GRU is calculated and routed downstream (Cranmer et al. 2001).

### Meteorologic Forcing Dataset

One key point of this study is that all participating models had to use the same forcing dataset, here the Regional Deterministic Reanalysis System version 1 (RDRS-v1).

The RDRS dataset is a preliminary sample of an atmospheric reforecast and precipitation/ land-surface reanalysis dataset that has recently been developed and released by ECCC (Gasset and Fortin 2017; Gasset et al., forthcoming). The data had previously been provided to this project before its public release. The dataset was chosen because of its high spatial and temporal resolution, and the availability of all variables required to set up hydrologic and

land-surface models. A full list of variables, including units and vertical level, can be found in Table 3. The table also contains the information on which model used which variable. The data are available in CaSPAR (CaSPAR 2017; Mai et al. 2020).

This dataset was obtained from short-term (6-h to 18-h lead time) mesoscale (15-km) integrations of the global environmental multiscale (GEM) atmospheric model coupled to the Canadian Land Data Assimilation System (CaLDAS) and to the Canadian Precipitation Analysis (CaPA), launched every 12 h from initial atmospheric conditions provided by the ERA-Interim reanalysis (Gasset et al., forthcoming). A technical report is available (Gasset et al. 2020).

The RDRS-v1 dataset covers North America with an approximate 15-km by 15-km grid resolution. The hourly dataset is available for January 2010 to December 2014. The entire dataset was used for calibration (2010 used for warm-up). No temporal validation was performed in this study due to the short period of available forcings; instead, a spatial validation was applied (see the section on streamflow gauging stations). The forcing dataset was preprocessed during the GRIP-E project by cropping the full dataset to the domain of the Lake Erie watershed [Fig. 1(c)]. The forcing data were postprocessed for the individual models—for example, some models required lumped forcings [Fig. 1(d)] or needed the data to be aggregated to another spatial or temporal resolution. The spatial and temporal resolutions of each model can be found in Table 1. Further details on forcing data processing, like unit conversions, can be found online in the “Model\_setups\_GRIP.xlsx” on the GitHub associated with this publication (see Data Availability Statement for URL).

### Geophysical Datasets

Besides the meteorologic forcings, most model setups required additional geophysical datasets such as a DEM, soil, and landcover information. Although we tried to unify these datasets across all participating models, it was found infeasible and would have excluded too many models. Hence, in this study, the choice of geophysical datasets was made by the individual modelers. The datasets used are listed in Table 2, specifying which model used which dataset. Please note that some models used additional datasets besides DEM, soil, and landcover for other purposes, for example, to specify the subsurface. These datasets are listed under

**Table 3.** Variables available in the reanalysis dataset, RDRS-v1

Variable	Abbreviation	Long name	Unit	Level	Used by
Precipitation rate	PR0	Quantity of precipitation	(m)	SFC	All models
Air temperature	TT	Air temperature	(°C)	40 m	All models
Inc. shortwave rad.	FB	Downward solar flux	(W/m <sup>2</sup> )	SFC	SWAT-EPA, SWAT-Guelph, VIC, VIC-GRU, GEM-Hydro, MESH-SVS, MESH-CLASS
Inc. longwave rad.	FI	Surf. inc. infrared flux	(W/m <sup>2</sup> )	SFC	VIC, VIC-GRU, GEM-Hydro, MESH-SVS, MESH-CLASS
Atmospheric pressure	P0	Surface pressure	(mbar)	SFC	SWAT-EPA, VIC, VIC-GRU, GEM-Hydro, MESH-SVS, MESH-CLASS
Specific humidity	HU	Specific humidity	(kg/kg)	40 m	SWAT-EPA, SWAT-Guelph, VIC, VIC-GRU, GEM-Hydro, MESH-SVS, MESH-CLASS
Wind components	UU, VV	U/V-comp. of wind (along grid X/Y)	(knots)	40 m	GEM-Hydro
Corr. wind components	UUC, VVC	U/V-comp. of wind (along W-E/S-N direction)	(knots)	40 m	No model
Wind speed	UVC	Wind modulus	(knots)	40 m	SWAT-EPA, SWAT-Guelph, VIC, VIC-GRU, MESH-SVS, MESH-CLASS
Wind direction	WDC	Meteorol. wind direction	(degree)	40 m	No model

Note: The forcings are either available at surface (SFC) or 40-m height. The average spatial resolution is 15 km, while the temporal resolution is 1 h. The data are available for 2010–2014 over North America. The table also specifies which model used which variable.

“Others” in the table. Additional information on postprocessing of these datasets, like aggregation to other spatial resolutions and merging of soil and landcover classes, are available online in the “Model\_setups\_GRIP.xlsx” on the GitHub associated with this publication (see Data Availability Statement for URL).

### Model Performance Evaluation

The model performances are evaluated for the calibration period from 2011 to 2014, while 2010 was discarded as the model warm-up period. The NSE (Nash and Sutcliffe 1970) is defined by

$$NSE = 1 - \frac{\sum_{t=1}^T (Q_{sim}(t) - Q_{obs}(t))^2}{\sum_{t=1}^T (Q_{obs}(t) - \bar{Q}_{obs})^2} \quad (1)$$

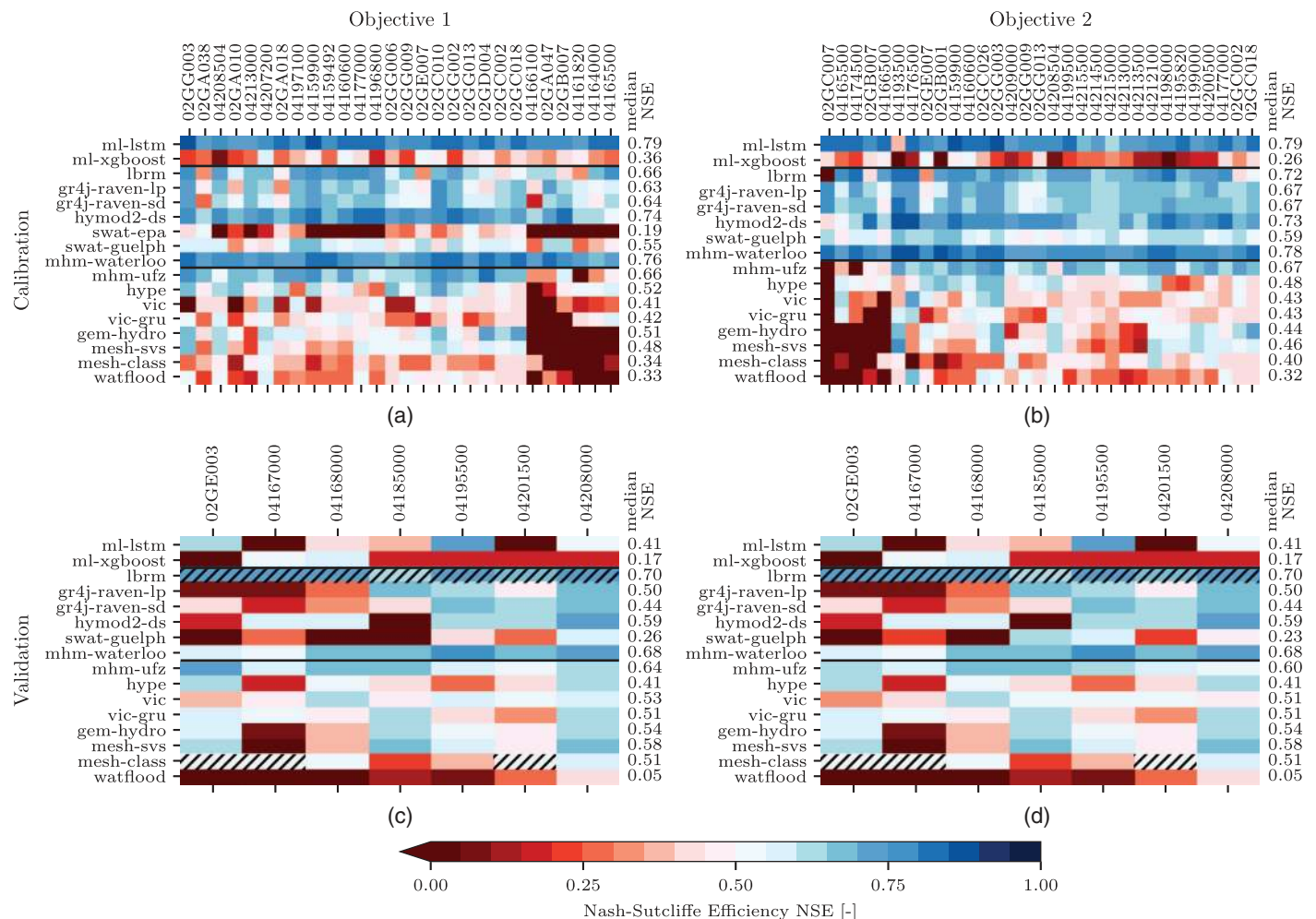
where  $T$  = number of time steps;  $Q_{sim}(t)$  and  $Q_{obs}(t)$  = simulated and observed streamflow at time step  $t$ ; and  $\bar{Q}_{obs}$  = average observed streamflow over the simulation period. The NSE was used to compare the simulated discharge  $Q_{sim}$  with the observed daily streamflow  $Q_{obs}$ . The NSE was chosen based on a survey performed in November 2018 among the eight modeling groups participating at the time. Overall, 75% of the modelers voted for using

NSE, 62.5% for percent bias (PBIAS), 37.5% for NSE of the logarithmic discharge, and 12.5% each for NSE of the square-root of discharge and root-mean-square error regarding discharge. Multiple votes were allowed during that poll. The PBIAS was used as a secondary metric. The analysis of PBIAS, however, did not yield any further insights and results will not be shown. The results of the PBIAS metric and several other metrics can be found in the Supplemental Materials (Figs. S1–S5).

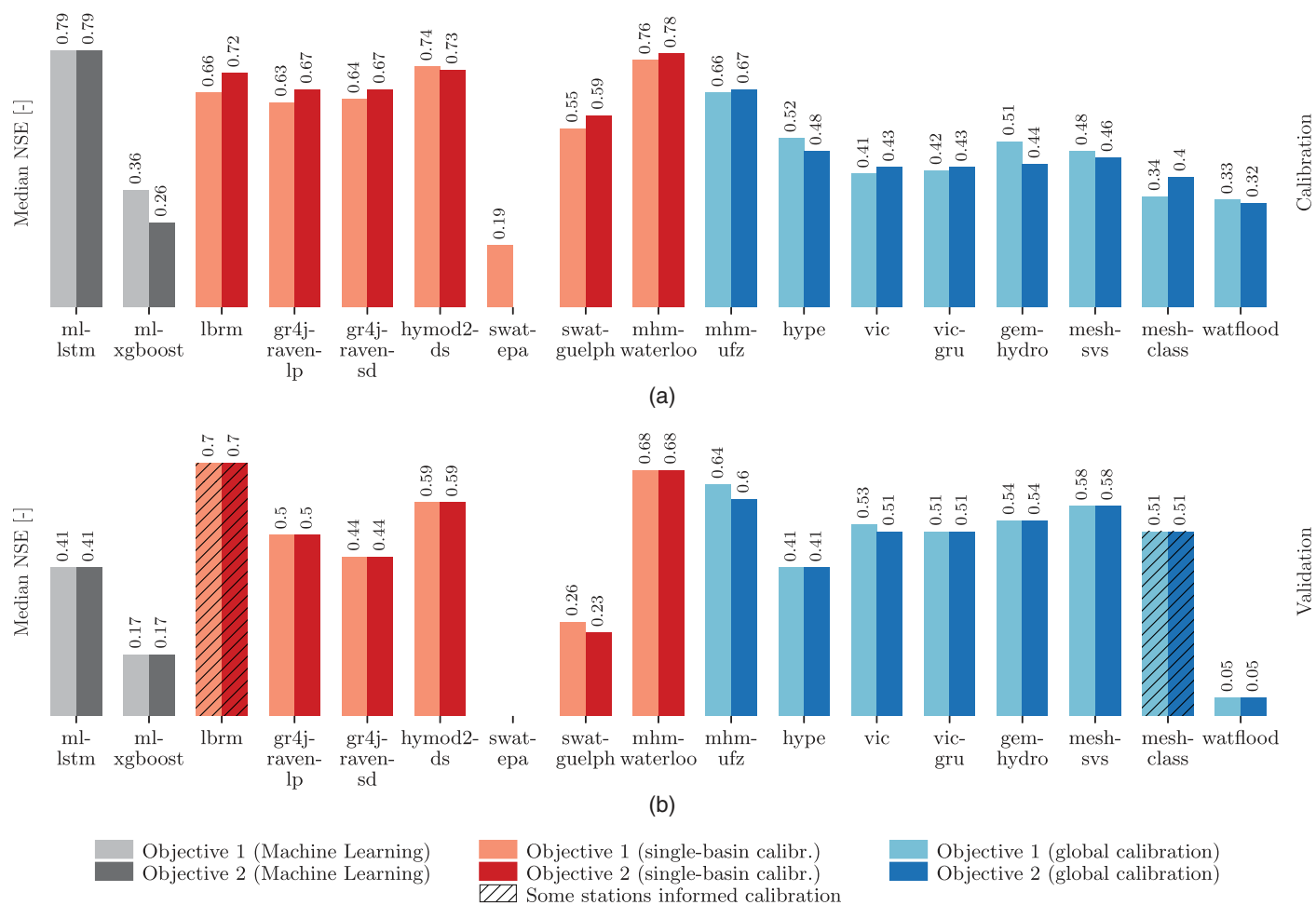
The median of these NSE values was chosen to determine the performance of a model across multiple gauging stations or the performance of a streamflow gauging station across several models. Furthermore, the ensemble of streamflow simulations was analyzed for selected streamflow gauging stations only—the stations with the worst and best median NSE across all models.

This leads to the following four types of analyses:

- Analysis of performance per gauge and model for all stations of objectives 1 and 2 for both calibration and validation (Fig. 2; see Analysis of Model Performance per Streamflow Gauge and Model).
- Analysis of performance per model across all gauging stations for both objectives in calibration and spatial validation (Fig. 3; see Analysis of Model Performance across Gauging Stations).



**Fig. 2.** (Color) Performance of the participating models in calibration and validation (a, b and c, d, respectively) for each gauging station of objectives 1 and 2. The colored tiles indicated the NSE per gauge and model, while the median NSE over all gauging stations is displayed to the right. The black horizontal lines separate (1) Machine Learning models from (2) models that are calibrated at each individual streamflow gauge from (3) models that are calibrated over the entire domain calibrating all streamflow gauges simultaneously. The hatched tiles (validation only) mark gauging stations that have informed the calibration of the corresponding models.



**Fig. 3.** (Color) The median model performance measured using the NSE of streamflow of the Machine Learning models (grey), models using individual basin calibration (red), and models calibrated for the whole domain (blue) for (a) calibration stations; and (b) validation stations. The light-colored bars show results of objective 1, while the dark-colored bars show objective 2 results. The median values are added on top of the bars for reference. Hatched bars indicate that some validation stations might have been used in calibration previously, and hence the validation results are biased. The models that show different values for objective 1 and objective 2 (SWAT-Guelph, mHM-UFZ, and VIC) are the models that provided an optimal model setup for each of the objectives. All other models provided only one final model setup (independent of the objective). The results shown are the median NSE values of the gauge-wise results shown in Fig. 2.

- C. Analysis of performance per gauging station across all models for both objectives in calibration and spatial validation (Fig. 4; see Analysis of Performance per Gauging Station across All Models).
- D. Analysis of simulated streamflow ensembles for the best and worst station of objective 1 for both calibration and validation stations (Fig. 5; see Analysis of Simulated Streamflow Ensembles).

## Results and Discussion

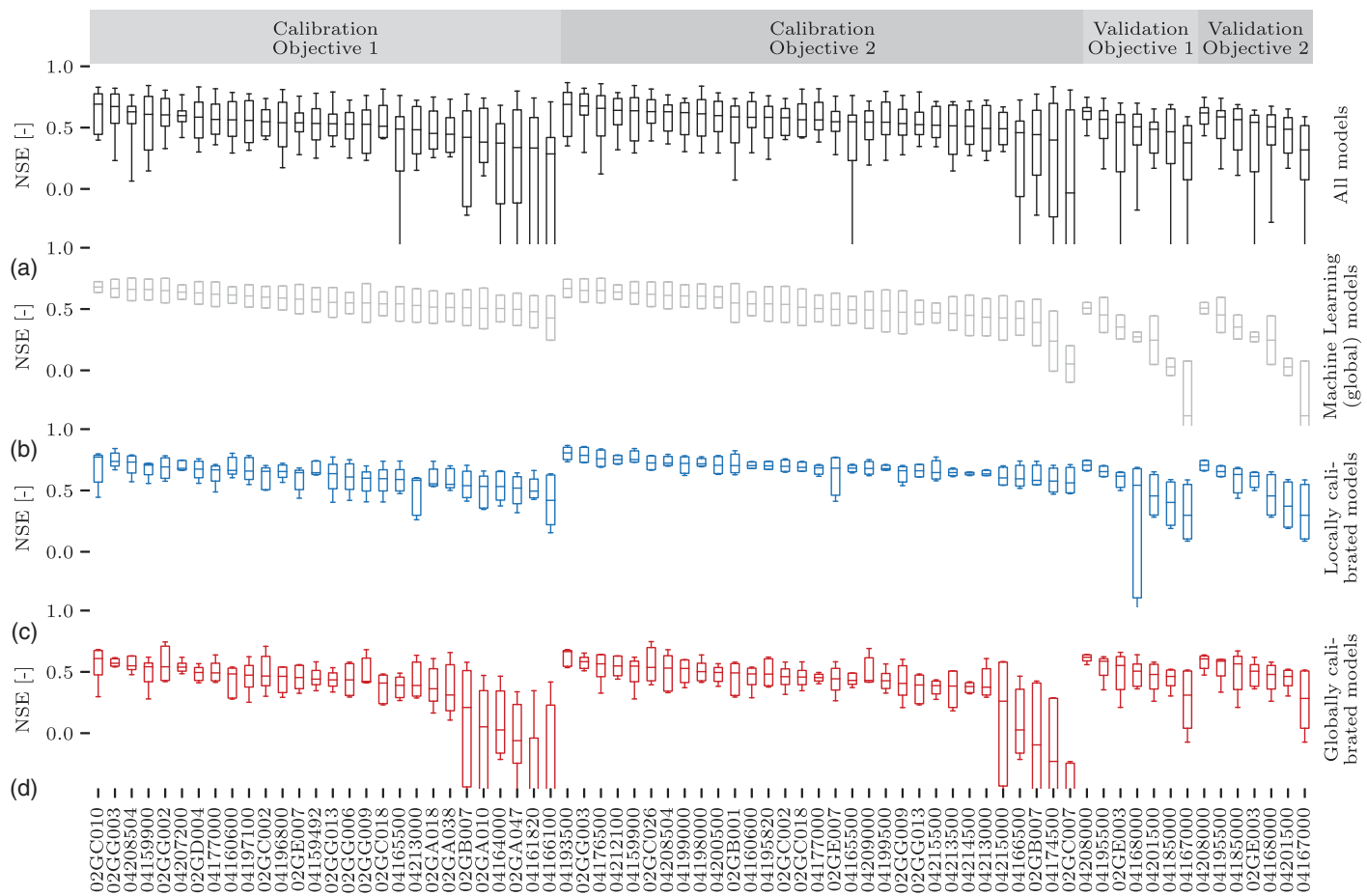
This section will present and discuss the results of the four types of analyses (A–D) mentioned above.

### Analysis of Model Performance per Streamflow Gauge and Model

First, the model performance at each of the 46 streamflow gauging stations and for each of the 17 models regarding NSE of the simulated streamflow was analyzed (Fig. 2). The summary statistics of median NSE over all gauges for the calibration and validation stations for each objective can be found in Table S2. The gauging

stations in Fig. 2 are sorted using a hierarchical clustering such that stations of similar performance patterns across all models appear closer to each other than gauges that show a different pattern. Each of the four panels in the figure is divided by horizontal lines into three blocks separating the two Machine Learning models in the first block from the seven locally calibrated models in the second block and the eight globally calibrated models in the third block.

As expected, the locally calibrated models (second block) in general yield better results than globally calibrated models (third block) in the calibration model [Figs. 2(a and b)]. Most locally calibrated models, however, perform surprisingly well in validation [Figs. 2(c and d)]. It is also surprising that no significant difference between objectives 1 and 2 can be seen. Most models perform equally well in both objectives (see median NSE for each model added as labels right of each panel). The reason for this is that the stations for objective 1 (low human impact watersheds) are chosen purely based on the regulation type as “natural” and “reference” provided by WSC and USGS. The classification, however, is constant in time, and a watershed might have been of no human impact in the past but is no longer (or vice versa). In hindsight,

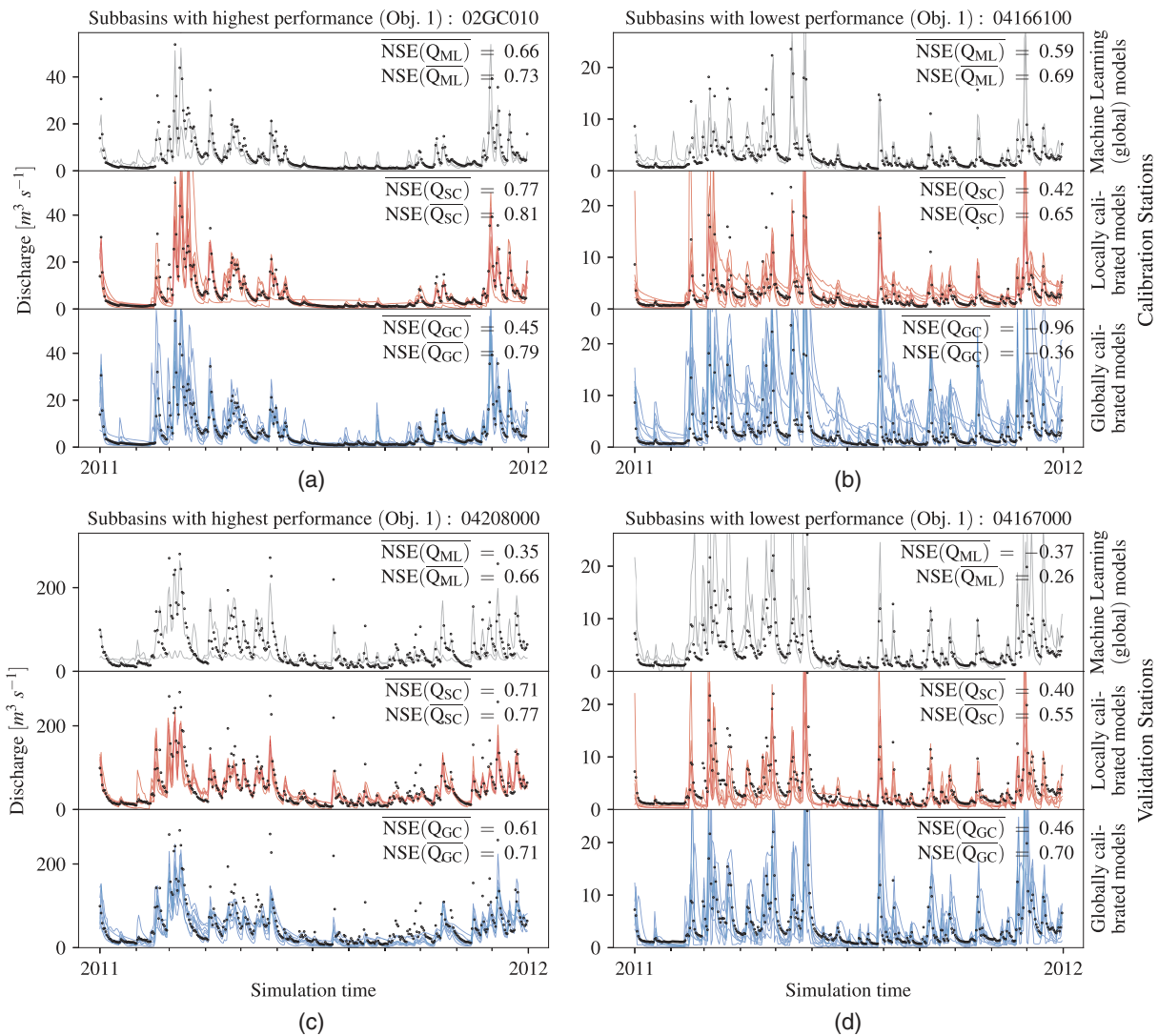


**Fig. 4.** (Color) Performance of the model ensemble per gauge for: (a) all models participating (black boxplots); (b) only the data-driven Machine Learning models (grey boxplots); (c) models that are calibrated at each individual gauge (blue boxplots); and (d) models that are calibrated globally (red boxplots). The horizontal line at the center of each box indicates the median Nash–Sutcliffe performance of the simulated streamflow, while the height of the box is determined by the 25th and 75th percentiles, and the whiskers indicate the 5th and 95th percentiles. No whiskers are shown for the Machine Learning models because there are only two such models. The results are shown for each gauging station (x-axis) for calibration stations of objective 1 (low human impact watersheds) and objective 2 (most downstream gauges) and the stations used for spatial validation of both objectives. The gauges in each of the four categories (calibration/validation and objective 1/objective 2) are sorted according to their decreasing median performance for all models (horizontal lines of each box in panel A).

the stations initially selected as low human impact should have been manually double-checked to ensure that they are indeed not affected by catchment management.

Following the definition of an at least satisfactory performance regarding streamflow of an NSE of 0.5 or higher (Moriassi et al. 2015), the only models with a (median) satisfactory performance in all four setups (panel A–D) are the GR4J-Raven-lp, HYMOD2-DS, mHM-Waterloo, and mHM-UFZ. Satisfactory in both calibration but not in both validation setups (panel A–B) are ML-LSTM, GR4J-Raven-sd, and SWAT-Guelph. LBRM is satisfactory in both calibration setups as well; the validation results are not considered here because some information of the validation stations might have informed calibration already (due to the operational setup of LBRM). The likely reason for the superior performance of these models in validation is that the validation stations were selected after a visual inspection of hydrographs and drainage area to make sure they were not heavily managed and not located in highly urbanized areas. Satisfactory in both validation scenarios but not in both calibration scenarios (panel C–D) are VIC, VIC-GRU, GEM-Hydro, and MESH-SVS. LBRM and MESH-CLASS are not considered for validation because they both used validation stations for calibration.

The machine learning model ML-XGBoost is the only model that performed substantially better for objective 1 (calibration) than for objective 2 ( $\Delta\text{NSE} = 0.10$ ). All other models have very similar median NSEs (in both calibration and validation). However, the second machine learning model, ML-LSTM, performed much better than ML-XGBoost. It has been shown before that ML-XGBoost does not perform well when insufficient training data are available (Gauch et al. 2019). The data-driven ML-LSTM model serves here as a baseline for all hydrologic models, as it can show how much information content is provided in the data and the performance that can be achieved without hydrologic knowledge. Notably, both data-driven models perform equally well (or poorly) in both objectives in calibration (NSE of 0.79 versus 0.79 for ML-LSTM, and NSE of 0.36 versus 0.26 for ML-XGBoost). Even the stations that did not perform well for the globally calibrated models—last six in Fig. 2(a) and first five in Fig. 2(b)—and were deemed likely to be managed did not show decreased performance compared with other stations for the ML models, even though no information on water management policies were used. It may be possible to augment hydrologic models with machine learning estimates to improve the simulations in basins like 04166100 and 02GC007, where almost all globally calibrated models fail to achieve NSEs above 0.



**Fig. 5.** (Color) Streamflow ensemble generated by the three groups of models: Machine Learning models (grey); models that are calibrated for each basin individually (local calibration; red); and models that are calibrated for entire domain (global calibration; blue). The observed streamflow is displayed with black dots in each plot. Results are shown for: (a) best-performing; (b) worst-performing gauging station in calibration; (c) best-performing; and (d) worst-performing station in validation for one example year (2011) out of the entire period (2011–2014). The stations were chosen regarding their overall median NSE (see Fig. 4). All results are regarding calibrations for objective 1 (low human impact watersheds). The median  $NSE(\overline{Q_G})$  for each model group  $G$ , as well as the NSE of the ensemble mean simulated discharge  $NSE(\overline{Q_G})$  of each group  $G$ , are added as a label to each panel. The NSE values are derived for the period from 2011–2014.

mHM-Waterloo is practically as good as the ML-LSTM, showing that most information contained in the forcing and geophysical data can be extracted by the model. HYMOD2-DS and LBRM are ranked second and third in the locally calibrated group. mHM-Waterloo and mHM-UFZ significantly outperform other models in both groups (local and global calibration, respectively). Reasons for mHM's performance might include its unique multiscale parameter regionalization (Samaniego et al. 2010) approach using geophysical data at their native resolution and relating them directly to hydrologic variables without losing the subgrid variability. Another reason is the internal model structure of mHM, which maintains a prescribed nonlinear relationship among internal model variables (Rakovec et al. 2019) being achieved by implementations of the transfer functions of each mHM parameter for most of the processes and elimination of unnecessary parameter interactions. mHM-UFZ significantly outperforms other globally calibrated

models in calibration. It is hard to assign further ranks as three other globally calibrated models—HYPE, GEM-Hydro, and MESH-SVS—perform very similarly in calibration. The results in validation are closer together, but mHM-UFZ is still the best and MESH-SVS is ranked second.

mHM-UFZ, like other globally calibrated models, had problems in performed well for a few stations (i.e., 04166100, 02GA047, 02GB007, 04161820, 0416400, 04165500, 02GC007, 04174500, and 04166500). These stations have been identified to be located in highly urbanized areas such as metropolitan Detroit, which corresponds to fact that models were not a priori informed about any human influence and regulation. Several models, like the land-surface scheme SVS used by both GEM-Hydro and MESH-SVS, as well as WATFLOOD, are known to have problems with flashy responses because tile drainage is not modeled in these models. GEM-Hydro and MESH-SVS perform almost equally in both

objectives in calibration and validation, certainly because they both use exactly the same underlying land-surface scheme, SVS. Furthermore, the calibrated SVS parameters of MESH-SVS were directly used for GEM-Hydro without further calibration. MESH-CLASS, which uses the CLASS land-surface scheme, does not perform as well as GEM-Hydro and MESH-SVS; this might be due to a different calibration strategy or differences in the land-surface schemes.

For the GR4J model, no major improvements can be observed when switching from a lumped setup (GR4J-Raven-lp) to a semi-distributed model setup (GR4J-Raven-sd), which might be due to either the fact that the discretization of the semidistributed version was not detailed enough or the maximum performance of GR4J is already reached when using the lumped version, and an improvement could only be achieved when using additional data such as watershed management rules.

The traditional VIC model, based on regular grid cells with defined tiles, and VIC-GRU show similar performance. The main differences between the tiles and GRUs is that the latter take into account soil classes in their native resolution while the tiles only use the major soil class per grid cell. This, however, does not make much difference in the Lake Erie watershed because the soil classes do not exhibit high spatial heterogeneity. Hence, the VIC model uses about 1,390 tiles (381 grid cells each with 3–4 vegetation tiles) and VIC-GRU uses about 2,400 grouped response units. Some stations are better for one or the other model, but the overall performance is similar. The fact that VIC/VIC-GRU reaches an upper limit of streamflow performance at a certain discretization level was previously described by Gharari et al. (2020). The differences in performance between the models are likely due to different input datasets used, different routing models, and/or different calibration strategies.

SWAT-EPA provided only results for the calibration of objective 1 stations, but yielded significantly worse performance than any other model, and particularly worse than SWAT-Guelph, which was set up by an independent group. The weak performance of SWAT-EPA is either due to calibration strategy or the datasets used. This shows that model intercomparisons need to take into account modelers' expertise and perseverance in addition to the traditionally considered influence of datasets used, calibration strategies, model discretization, etc.

The performance for all models at the seven validation streamflow gauging stations is displayed in Figs. 2(c and d). In validation, the Machine Learning models show significantly degraded performance relative to calibration, and both hydrologic model groups (locally and globally calibrated) were found to perform better than ML models. This is most likely because there were not enough data (e.g., streamflow gauge data, forcing data, and basin attributes) to train ML models properly, as previously discussed by Gauch et al. (2019). The best model in validation is LBRM, which cannot be considered because it had already used all stations for calibration. Hence, the two versions of mHM (mHM-Waterloo and mHM-UFZ) are ranked first in the local and global model groups, respectively. HYMOD2-DS and MESH-SVS are ranked second in the two groups, respectively.

It should be noted that the locally calibrated models (second group) use either the area-ratio method (Fry et al. 2014) (LBRM and SWAT-Guelph) or the nearest-neighbor method to determine streamflow at the validation stations. The donor basins used to derive those estimates might differ between the models. The list of donor basins used for each validation station can be found in the README files online (e.g., "data/objective\_1/lake-erie/validation/model/HYMOD2-DS/README.md" on the on the GitHub associated with this publication).

## Analysis of Model Performance across Gauging Stations

Fig. 3 shows the median NSE per model across all gauges in objective 1 and objective 2 in calibration [Fig. 3(a)] and validation [Fig. 3(b)], mostly to ease comparison between calibration and validation results (panel A versus panel B) and differences between the two objectives for each model (light versus dark-colored bar per model). A few models report different values for objective 1 and objective 2 in validation (i.e., SWAT-Guelph, mHM-UFZ, and VIC), because they provided an optimal model setup for each of the two objectives. All other models provided only one final model setup (independent of the objective), and hence only provide one modeled streamflow time series for each gauge.

The Machine Learning models show a significant drop in performance between calibration and validation ( $\Delta\text{NSE} \approx -0.38$  for ML-LSTM) caused by the limited training data available. This could be resolved by pretraining the ML models with a larger dataset, for example, CAMELS (Newman et al. 2015; Addor et al. 2017) and then fine-tune with the datasets available for the study domain of Lake Erie. This would partly resolve the imbalance between hydrologic models developed based on expert knowledge over many years or even decades and Machine Learning models only being trained on the current data without benefiting from any form of expert knowledge.

As for the locally calibrated models (red bars), SWAT-Guelph performance also significantly drops ( $\Delta\text{NSE} \approx -0.29$  for objective 1 and  $-0.36$  for objective 2) due to known weak transferability of parameters in SWAT (Heuvelmans et al. 2004). Another reason for the drop in performance might be the usage of the area-ratio method to obtain the estimates at the validation stations, while all other models (except LBRM) used a nearest-neighbor method. The lumped and semidistributed GR4J setups, GR4J-Raven-lp and GR4J-Raven-sd, as well as HYMOD2-DS, decrease in performance but not as much as the ML models ( $\Delta\text{NSE} \approx -0.15$ ,  $\Delta\text{NSE} \approx -0.22$ , and  $\Delta\text{NSE} \approx -0.14$ , respectively); mHM-UFZ maintains almost the same performance in validation as it has in calibration ( $\Delta\text{NSE} \approx -0.11$ ). Almost all globally calibrated models maintain or, surprisingly, even improve performance in validation. This probably a side effect of the much smaller sample size (7 stations versus 28 and 31 in calibration); but, more importantly, it is certainly caused by the large set of urban watersheds in the calibration set for which all global models had problems to obtain good results. Such gauges were avoided in the set of validation stations chosen after we observed that the classification of WSC/USGS needs manual verification. mHM-UFZ maintains the same performance ( $\Delta\text{NSE} \approx -0.05$ ), HYPE decreases slightly ( $\Delta\text{NSE} \approx -0.09$ ); VIC and VIC-GRU show slightly improved performance ( $\Delta\text{NSE} \approx 0.10$  and  $0.08$ ) as do GEM-Hydro, MESH-SVS, and MESH-CLASS ( $\Delta\text{NSE} \approx 0.06$ ,  $0.11$ , and  $0.14$ , respectively). MESH-CLASS, however, is slightly biased, as it already used two of the validation stations for calibration, and thus validation results might appear better than they actually are. WATFLOOD is the only global model that showed drastically decreased performance in validation ( $\Delta\text{NSE} \approx -0.28$ ), which may indicate a flawed validation setup given that WATFLOOD was developed using catchments in the Lake Erie basin (Kouwen et al. 1993), and thus has shown much more robust performance in this region in the past.

In addition, the results in Fig. 3 again highlight that only minor differences exist between the performance of the models for objective 1 and objective 2 stations originally defined as low human impact and most downstream stations, respectively. This is most likely due to the fact that the assignment of low human impact was mainly based on the tags indicated by WSC and USGS in their gauge

information. We see, however, that there are several stations in objective 1 (i.e., 04166100, 02GA047, 02GB007, 04161820, 0416400, and 04165500) where all models had problems in achieving good results, even in calibration, due to their location within highly urbanized areas. These stations, therefore, should not have been classified as objective 1 (although labeled as “natural” by WSC and “reference” by USGS). These stations lowered the performance of objective 1 and led to similar results for both objectives. The validation stations were selected later in the project with more care, making sure they are not located in urban areas. We also visually inspected the observed streamflow to make sure there was no obvious human influence/management in the selected gauges. The void of human-impacted watersheds in validation explains why performance improved in so many models.

### Analysis of Performance per Gauging Station across All Models

Fig. 4 shows the variability of the streamflow simulation performance for each gauging station over sets of models. Panel A considers all models, while the other three panels separate the models into the categories of machine learning (ML) models (panel B), locally calibrated models (panel C), and globally calibrated models (panel D). The gauges ( $x$ -axis) are sorted according to decreasing median NSEs over all models (horizontal line in each box of panel A).

The calibration results show that the gauges with the worst performance have the highest variability regarding the streamflow simulation performance, with only a few exceptions (see panel A). The high variability in performance for calibration stations is due to the high variability of performance of globally calibrated models (panel D), while ML and locally calibrated models (panels B and C, respectively) result in much less variability and, in general, better results for those stations. The stations with greater variability are the stations mentioned above (see the subsections above: Analysis of Model Performance per Streamflow Gauge and Model; and Analysis of Model Performance across Gauging Stations) as being the ones in highly urbanized areas where all the global models, except mHM-UFZ, HYPE, and VIC-GRU, have difficulties in simulating streamflow. All of the above holds for both objectives in calibration mode.

The picture changes when it comes to the performance of the validation stations. Most of the overall variability (panel A) is caused by ML models (panel B) and the locally calibrated models (panel C), which all used either the area-ratio or nearest-neighbor method to transfer parameters of various donor basins to the stations used for validation. The variability of the ML models is a less reliable predictor compared with the results for the locally calibrated models, as it is determined based on only two ML models, while the boxplots for the locally calibrated models represent six models (mostly five, because LBRM should not really be counted as it used all validation stations for calibration). This again shows that the globally calibrated models are superior in validation due to their less subjective approach to model ungauged basins; however, they need overall improvement when it comes to representing urban areas.

The best-performing station across all models in calibration mode of objective 1 is 02GC010 (median NSE = 0.69); the station with lowest performance is 04166100 (median NSE = 0.29). The best-performing station across all models in validation for objective 1 is 04208000 (median NSE = 0.63), while the worst-performing one is 04167000 (median NSE = 0.38). The simulated streamflow time series of these four stations is analyzed in the next section.

### Analysis of Simulated Streamflow Ensembles

In Fig. 5, streamflow ensembles for three groups of models (ML, locally calibrated models, and globally calibrated models) are shown for selected gauging stations in calibration and validation mode only for objective 1 (objective 2 results are similar but not shown). The best and worst objective 1 stations in calibration and validation are used here. The performance is regarding the overall highest and lowest median NSE across all models [Fig. 4(a)].

The ML ensemble consists of only two models (grey hydrograph plots). The two simulated hydrographs in calibration mode [Figs. 5(a and b)] look very realistic—low flows are well represented and some peak flows are missed but most are captured. For the worst station in the calibration set for objective 1 [04166100; Fig. 5(b)], one ML model seems to overestimate the flows consistently, possibly because this is a station with human influence. No data are provided to the ML models to train them regarding human impacts and management, and it is hence expected that these stations are not very well represented with data-driven models (or any other globally calibrated model).

The ensemble of streamflow time series for locally and globally calibrated models (red and blue hydrographs, respectively) look very similar for well-performing stations in both calibration and validation [Figs. 5(a and c)]. This, however, changes for the worst-performing station: The variance in the simulated hydrographs in calibration mode [blue panel in Fig. 5(b)] is much larger for globally calibrated models, while the ensemble of the locally calibrated models looks, expectedly, very consistent. The low flows are well represented, but some peak flows are missed by locally calibrated models [red panel Fig. 5(b)]. This gets surprisingly better for the validation of the overall worst-performing station [blue panel Fig. 5(d)], possibly because these stations are not in urban areas, which are the stations where almost all globally calibrated models had problems already in calibration (namely GEM-Hydro, MESH-SVS, MESH-CLASS, and WATFLOOD, but also VIC and VIC-GRU).

Rather than evaluating the performance of each individual model using the median NSE [added as label  $NSE(Q_G)$  for each group  $G \in (ML, SC, GC)$  to panels in Fig. 5], the model intercomparison now also allows to obtain the performance of the model ensemble. Therefore, the mean ensemble streamflow timeseries  $\bar{Q}_G$  for each model group  $G$  is derived and compared to the observations [added as label  $NSE(\bar{Q}_G)$  to panels in Fig. 5]. For all model groups and both stations, this leads to large improvements in performance, highlighting the strength of using model ensembles rather than individual models, as reported by others (Duan et al. 2007; Muhammad et al. 2018; Darbandsari and Coulibaly 2019).

### Conclusions

Our extensive model intercomparison compares the performance of 17 models, with 14 independent modeling teams building and calibrating these models, for predicting streamflows in the Lake Erie drainage basin. This model intercomparison is archived for future modelers to assess their model performance against the original 17 models included in our study.

The mHM model in its local and global setups (mHM-Waterloo and mHM-UFZ, respectively) were found to produce superior quality hydrographs regarding NSE performance when compared with all other models. For example, when compared with machine learning benchmarks, both mHM setups and the ML-LSTM model produce practically the same quality results for calibration stations, but both mHM setups produce superior quality hydrographs for validation stations. Please note that this finding is highly dependent on the design used in this study. Other studies (e.g., Kratzert et al. 2019)



have shown that the mHM model is outperformed by LSTMs when a large-sample dataset is used for training.

Compared with the 13 traditional hydrologic models (excluding the two ML models and the two mHM setups) mHM, outperforms all of them. This is noteworthy because the models to which mHM is compared are all developed and calibrated by multiple, typically independent teams; thus, in addition to the variation in model type, the model-building strategies also varied. In addition, unlike mHM, five of these traditional models (LBRM, GEM-Hydro, MESH-SVS, MESH-CLASS, and WATFLOOD) were first coded and developed for application in Canada (or the Great Lakes basin in general). This finding could be due to mHM's unique multiscale parameter regionalization (MPR) scheme to account for the subgrid variability of basin physical properties that allows for the seamless predictions of water fluxes and storages at different spatial resolutions and ungauged locations.

Other models with globally calibrated setups performed reasonably well, except possibly HYPE and WATFLOOD, which showed notably lower validation performance levels in comparison with the other globally calibrated models. The VIC, VIC-GRU, GEM-Hydro, MESH-SVS, and MESH-CLASS models all showed similar validation performance levels that were all improved relative to calibration performance levels (note that mHM-UFZ showed a minor decrease in validation performance relative to calibration performance). In contrast, all five locally calibrated models that were validated had degraded validation performance relative to calibration performance. Despite this degradation, two locally calibrated models (HYMOD2-DS and mHM-Waterloo) were still notably better in validation than all globally calibrated models except mHM-UFZ.

Locally calibrated models turn out to be surprisingly powerful in validation especially if parameter sets are transferred wisely—that is, if donor basins are selected that are not necessarily only closest in proximity but also share similar soil and landcover properties. mHM-Waterloo (locally calibrated) turns out to be the best model in the entire study. Both locally calibrated SWAT models perform notably poorly relative to the other five models in the locally calibrated model category.

Many models had problems simulating accurate streamflows in watersheds containing highly urbanized areas or other human impacts such as tile drainage or reservoir operations. The intent was to show this with objective 1 versus objective 2, which turned out to be an imperfect assessment because we initially used only the USGS and WSC flags on natural and managed operations of watersheds. Ultimately, however, we found that stations referred to as “natural/reference” can be highly impacted by human actions, and these impacts are not accounted for in the models. The USGS/WSC flags have the disadvantage of being constant in time for each gauge, making it impossible to account for management or urban development that might have changed over the course of the operational years of a gauge. One solution would be to use the Canadian Dam Inventory, which provides the initiation dates of each regulated location, and use it to classify gauging stations as either regulated or natural. We also recommend manual screening of gauging stations and visual inspection of hydrographs to identify low human impact watersheds. This was done here for the stations used for spatial validation, but unfortunately not for the calibration stations, which led to indistinguishable results in calibration mode for objectives 1 and 2.

## Limitations and Future Work

The work presented here lays the groundwork for a unified model intercomparison for multinational teams. There are, however, several limitations associated with this study:

1. A short period of available forcings (2010–2014) prevented an analysis of model performance of different climatic conditions (e.g., wet versus dry periods) and a temporal model validation. The study presented here performed a spatial model validation only.
2. The comparison of simulated streamflow only does not provide a detailed insight into the performance of (especially) distributed models. It is not clear if models with good streamflow performance sacrifice the quality of other simulated variables.
3. The geophysical datasets utilized were not all consistent; that is, each modeling team was allowed to use the DEM, soil, and landcover information of their choice. This can lead to differences in model performance that cannot be attributed to an actual difference in the models but solely to the quality of the information used in setting up the model.

The main limitation is the short study period of only 5 years for which high-resolution forcing data suitable for driving land-surface hydrologic models are available. One of those years was reserved as warm-up period for each model, leading to only 4 years to train the models and leaving no data for temporal validation. It would be beneficial to use a longer forcing dataset that provides all meteorologic inputs for this region at the same spatiotemporal scale and is available for a longer time period, such that the calibration period could be longer and some data would be still available for temporal validation. A suitable dataset like this was not available during the course of this project. We believe that the conclusions reached in this work are significant regarding (1) weak model performances in urban areas, (2) the major benefit of using model ensembles, and (3) the impact of the modeler's skill on model performance, and that these findings will likely not change even if studied for a longer period. Model performances themselves however are likely to be impacted by a longer and more climatically diverse study period.

It would further be interesting to see if differences in model performance persist if the same geophysical datasets and routing are used across all models, making sure that differences are not due to different input data and their resampling to match the requirements of the respective model. Consistent model setups have been already used for the model family of GEM-Hydro, MESH-SVS, MESH-CLASS, and WATFLOOD, leading to reliable insights that can be exclusively pointed toward model parameterization and conceptualization rather than differences in used geophysical and forcing data. Our study provides motivation for more detailed model diagnosis to investigate why and where some models are outperforming others.

The above-listed limitations have been taken into account for the future GRIP project over the entire Great Lakes (GRIP-GL), which has just ramped up as our GRIP-E work has wrapped up. In GRIP-GL, all these issues are addressed and resolved. GRIP-GL will cover the entire Great Lakes watershed including the Ottawa River. In addition, GRIP-GL will utilize a much longer high-resolution meteorologic input dataset, soon to be released, provided by Environment and Climate Change Canada. The future project will refine the comparison to require models to use consistent geophysical datasets and evaluate model predictive quality against multiple measured response variables beyond streamflow. GRIP-GL will be designed to determine exactly why the mHM model has done so well relative to other models in GRIP-E, and will yield more robust findings with regard to model performance rankings. The larger-scale GRIP-GL study simply could not have been designed or conceived as it has been, nor motivate the modeling teams to participate, without this GRIP-E study preceding it.

## Data Availability Statement

The subwatershed shapes are available at <https://doi.org/10.5281/zenodo.3888690> (Shen et al. 2020). The forcing data (RDRS-v1) are available on CaSPAr ([www.caspar-data.ca](http://www.caspar-data.ca)). Gridded model outputs for GEM-Hydro are available at <https://doi.org/10.5281/zenodo.3890487> (Gaborit et al. 2020), while gridded model outputs for mHM-UFZ are available at <https://doi.org/10.5281/zenodo.3886551> (Rakovec et al. 2020). Further information and documentation are available in the Wiki of the project GitHub <https://github.com/julemai/GRIP-E/wiki>. This GitHub also contains the simulated streamflow of all models for all gauges in calibration and validation model, and all scripts used to prepare the figures presented here are available.

## Acknowledgments

This research was undertaken thanks primarily to funding from the Canada First Research Excellence Fund provided to the Global Water Futures (GWF) Project and the Integrated Modeling Program for Canada (IMPC). The modeling teams for mHM-Waterloo and mHM-UFZ received funding from the Initiative and Networking Fund of the Helmholtz Association through the project Advanced Earth System Modelling Capacity (ESM) ([www.esm-project.net](http://www.esm-project.net)). The work was made possible by the facilities of the Shared Hierarchical Academic Research Computing Network (SHARCNET; [www.sharcnet.ca](http://www.sharcnet.ca)) and Compute/Calcul Canada. The availability of codes, data, and results is summarized in the Data Availability Statement. Although this manuscript has been reviewed and approved for publication by the USEPA, the views expressed in this manuscript are those of the authors and do not necessarily represent the views or policies of USEPA. We thank Drs. Brent Johnson and Heather Golden from USEPA—Office of Research and Development, as well as the three anonymous reviewers for their review and valuable comments. This is NOAA Great Lakes Environmental publication number 1975.

## Supplemental Materials

Tables S1 and S2 and Figs. S1–S5 are available online in the ASCE Library ([www.ascelibrary.org](http://www.ascelibrary.org)).

## References

- Addor, N., A. J. Newman, N. Mizukami, and M. P. Clark. 2017. “The CAMELS data set: Catchment attributes and meteorology for large-sample studies.” *Hydrol. Earth Syst. Sci.* 21 (10): 5293–5313. <https://doi.org/10.5194/hess-21-5293-2017>.
- Agriculture and Agri-Food Canada. 2010. “Soil landscapes of Canada version 3.2.” Accessed March 3, 2021. <http://sis.agr.gc.ca/cansis/nsdb/slc/v3.2/index.html>.
- Alavi, N., S. Bélair, V. Fortin, S. Zhang, S. Z. Husain, M. L. Carrera, and M. Abrahamowicz. 2016. “Warm season evaluation of soil moisture prediction in the soil, vegetation, and snow (SVS) scheme.” *J. Hydrometeorol.* 17 (8): 2315–2332. <https://doi.org/10.1175/JHM-D-15-0189.1>.
- Arnold, J. G., P. M. Allen, and G. Bernhardt. 1993. “A comprehensive surface-groundwater flow model.” *J. Hydrol.* 142 (1–4): 47–69. [https://doi.org/10.1016/0022-1694\(93\)90004-S](https://doi.org/10.1016/0022-1694(93)90004-S).
- Bernier, N. B., S. Bélair, B. Bilodeau, and L. Tong. 2011. “Near-surface and land surface forecast system of the Vancouver 2010 winter Olympic and Paralympic games.” *J. Hydrometeorol.* 12 (4): 508–530. <https://doi.org/10.1175/2011JHM1250.1>.
- Best, M. J., et al. 2015. “The plumbing of land surface models: Benchmarking model performance.” *J. Hydrometeorol.* 16 (3): 1425–1442. <https://doi.org/10.1175/JHM-D-14-0158.1>.
- Boyle, D. P., H. V. Gupta, and S. Sorooshian. 2000. “Toward improved calibration of hydrologic models: Combining the strengths of manual and automatic methods.” *Water Resour. Res.* 36 (12): 3663–3674. <https://doi.org/10.1029/2000WR900207>.
- Burnash, R., R. Ferral, R. McGuire, and R. McGuire. 1973. *A generalized stream flow simulation system. Conceptual modeling for digital computer*. Washington, DC: US Department of Commerce.
- Burnash, R. J. C., and V. Singh. 1995. “The NWS River forecast system—Catchment modeling.” In *Computer models of watershed hydrology*. Englewood, CO: Water Resources Publications.
- CaSPAr. 2017. “The Canadian surface prediction archive.” Accessed March 3, 2021. [www.caspar-data.ca](http://www.caspar-data.ca).
- CEC (Commission for Environmental Cooperation). 2010. “Land cover, 2005 (modis, 250m).” Accessed March 3, 2021. <http://www.cec.org/north-american-environmental-atlas/land-cover-2005-modis-250m/>.
- Chen, T., and C. Guestrin. 2016. “XGBoost: A scalable tree boosting system.” In *Proc., 22nd ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining*, 785–794. New York: Association for Computing Machinery.
- Craig, J. R., et al. 2020. “Flexible watershed simulation with the Raven hydrological modelling framework.” *Environ. Modell. Software* 129 (Jul): 104728. <https://doi.org/10.1016/j.envsoft.2020.104728>.
- Cranmer, A., N. Kouwen, and S.-F. Mousavi. 2001. “Proving watflood: Modelling the nonlinearities of hydrologic response to storm intensities.” *Can. J. Civ. Eng.* 28 (5): 837–855. <https://doi.org/10.1139/01-049>.
- Croley, T. E., II. 1983. “Great lake basins (U.S.A.-Canada) runoff modeling.” *J. Hydrol.* 64 (1): 135–158. [https://doi.org/10.1016/0022-1694\(83\)90065-3](https://doi.org/10.1016/0022-1694(83)90065-3).
- Cuntz, M., J. Mai, L. Samaniego, M. P. Clark, V. Wulfmeyer, O. Branch, S. Attinger, and S. Thober. 2016. “The impact of standard and hard-coded parameters on the hydrologic fluxes in the Noah-MP land surface model.” *J. Geophys. Res. Atmos.* 121 (8): 1–25. <https://doi.org/10.1002/2016JD025097>.
- Darbandsari, P., and P. Coulibaly. 2019. “Inter-comparison of different Bayesian model averaging modifications in streamflow simulation.” *Water* 11 (8): 1707–1727. <https://doi.org/10.3390/w11081707>.
- de Boer-Euser, T., et al. 2017. “Looking beyond general metrics for model comparison—Lessons from an international model intercomparison study.” *Hydrol. Earth Syst. Sci.* 21 (1): 423–440. <https://doi.org/10.5194/hess-21-423-2017>.
- Dolan, D. M., and K. P. McGunagle. 2005. “Lake Erie total phosphorus loading analysis and update: 1996–2002.” Supplement, *J. Great Lakes Res.* 31 (S2): 11–22. [https://doi.org/10.1016/S0380-1330\(05\)70301-4](https://doi.org/10.1016/S0380-1330(05)70301-4).
- Duan, Q., et al. 2006. “Model parameter estimation experiment (MOPEX): An overview of science strategy and major results from the second and third workshops.” *J. Hydrol.* 320 (1–2): 3–17. <https://doi.org/10.1016/j.jhydrol.2005.07.031>.
- Duan, Q., N. K. Ajami, X. Gao, and S. Sorooshian. 2007. “Multi-model ensemble hydrologic prediction using Bayesian model averaging.” *Adv. Water Resour.* 30 (5): 1371–1386. <https://doi.org/10.1016/j.advwatres.2006.11.014>.
- ESA (European Space Agency). 2005. “Globcover.” Accessed March 3, 2021. [http://due.esrin.esa.int/page\\_globcover.php](http://due.esrin.esa.int/page_globcover.php).
- ESA (European Space Agency). 2017. “Land cover CCI climate research data package (CRDP).” Accessed March 3, 2021. <http://maps.elie.ucl.ac.be/CCI/viewer/download.php>.
- Evin, G., M. Thyer, and D. Kavetski. 2014. “Comparison of joint versus postprocessor approaches for hydrological uncertainty estimation accounting for error autocorrelation and heteroscedasticity.” *Water Resour. Res.* 50 (3): 2350–2375. <https://doi.org/10.1002/2013WR014185>.
- FAO (Food and Agriculture Organization). 2008. “Fao soils portal—Harmonized world soil database v 1.2.” Accessed March 3, 2021. <http://www.fao.org/soils-portal/soil-survey/soil-maps-and-databases/harmonized-world-soil-database-v12/en/>.

- Friedl, M., and D. Sulla-Menashe. 2019. "MCD12Q1 MODIS/Terra+Aqua land cover type yearly L3 global 500m SIN Grid V006." Accessed August 1, 2020. <https://lpdaac.usgs.gov/products/mcd12q1v006/>.
- Frieler, K., et al. 2017. "Assessing the impacts of 1.5°C global warming—Simulation protocol of the Inter-Sectoral Impact Model Intercomparison Project (ISIMIP2b)." *Geosci. Model Dev.* 10 (12): 4321–4345. <https://doi.org/10.5194/gmd-10-4321-2017>.
- Fry, L. M., et al. 2014. "The Great Lakes runoff intercomparison project phase 1: Lake Michigan (GRIP-M)." *J. Hydrol.* 519: 3448–3465. <https://doi.org/10.1016/j.jhydrol.2014.07.021>.
- Gaborit, E., D. G. Princz, V. Fortin, D. Durnford, and J. Mai. 2020. "GEM-hydro gridded simulations for the Great Lakes runoff inter-comparison project for Lake Erie (GRIP-E)." Accessed March 19, 2021. <https://doi.org/10.5281/zenodo.3890487>.
- Gaborit, É., V. Fortin, B. Tolson, L. Fry, T. Hunter, and A. D. Gronewold. 2017a. "Great Lakes runoff inter-comparison project, phase 2: Lake Ontario (GRIP-O)." *J. Great Lakes Res.* 43 (2): 217–227. <https://doi.org/10.1016/j.jglr.2016.10.004>.
- Gaborit, É., V. Fortin, X. Xu, F. Seglenieks, B. Tolson, L. M. Fry, T. Hunter, F. Ancil, and A. D. Gronewold. 2017b. "A hydrological prediction system based on the SVS land-surface scheme: Efficient calibration of GEM-Hydro for streamflow simulation over the Lake Ontario basin." *Hydrol. Earth Syst. Sci.* 21 (9): 4825–4839. <https://doi.org/10.5194/hess-21-4825-2017>.
- Gasset, N., et al. Forthcoming "A 10 km North American precipitation and land surface reanalysis based on the GEM atmospheric model." *Hydrol. Earth Syst. Sci. Discuss.* <https://doi.org/10.5194/hess-2021-41>.
- Gasset, N., L. Benyahya, M. Dimitrijevic, G. Roy, and V. Fortin. 2020. Evaluation of a capa reanalysis over North America and Canada on summer & winter 2010–2014." Accessed March 3, 2021. [https://collaboration.cmc.ec.gc.ca/science/outgoing/capa.grib/hindcast/capa\\_hindcast\\_rdrs\\_v1/Evaluation\\_CaPA\\_Reanalysis\\_North\\_America\\_2010-2014.pdf](https://collaboration.cmc.ec.gc.ca/science/outgoing/capa.grib/hindcast/capa_hindcast_rdrs_v1/Evaluation_CaPA_Reanalysis_North_America_2010-2014.pdf).
- Gasset, N., and V. Fortin. 2017. "Toward a 35-years North American precipitation and surface reanalysis." In *Proc., American Geoscientist Union Falls Meeting*. New Orleans: American Geophysical Union.
- Gauch, M., J. Mai, and J. Lin. 2019. "The proper care and feeding of CAMELS: How limited training data affects streamflow prediction." *Environ. Modell. Software* 135: 104926. <https://doi.org/10.1016/j.envsoft.2020.104926>.
- Gharari, S., M. P. Clark, N. Mizukami, W. J. M. Knoben, J. S. Wong, and A. Pietroniro. 2020. "Flexible vector-based spatial configurations in land models." *Hydrol. Earth Syst. Sci. Discuss.* 24 (2): 5953–5971. <https://doi.org/10.5194/hess-24-5953-2020>.
- GLAHF (Great Lakes Aquatic Habitat Framework). 2016. "Great Lakes aquatic habitat framework." Accessed March 3, 2021. <https://www.glahf.org>.
- Gronewold, A. D., T. Hunter, J. Allison, L. M. Fry, K. A. Kompoltowicz, R. A. Bolinger, and L. Pei. 2017. "Project documentation Report for Great Lakes seasonal and inter-annual water supply forecasting improvements project phase I: Research and development." Accessed March 19, 2021. <https://www.glerl.noaa.gov/pubs/fulltext/2018/20180020.pdf>.
- Haghnegahdar, A. 2015. "An improved framework for watershed discretization and model calibration: Application to the Great Lakes Basin." Ph.D. thesis, Dept. of Civil and Environmental Engineering, Univ. of Waterloo.
- Herman, J. D. 2012. "Time-varying sensitivity analysis reveals impacts of watershed model choice on the inference of dominant processes." Ph.D. thesis, Pennsylvania State Univ., Graduate School.
- Heuvelmans, G., B. Muys, and J. Feyen. 2004. "Analysis of the spatial variation in the parameters of the SWAT model with application in Flanders, Northern Belgium." *Hydrol. Earth Syst. Sci.* 8 (5): 931–939. <https://doi.org/10.5194/hess-8-931-2004>.
- Ho, J. C., and A. M. Michalak. 2017. "Phytoplankton blooms in Lake Erie impacted by both long-term and springtime phosphorus loading." *J. Great Lakes Res.* 43 (3): 221–228. <https://doi.org/10.1016/j.jglr.2017.04.001>.
- Hochreiter, S., and J. Schmidhuber. 1997. "Long short-term memory." *Neural Comput.* 9 (8): 1735–1780. <https://doi.org/10.1162/neco.1997.9.8.1735>.
- Huang, S., et al. 2017. "Evaluation of an ensemble of regional hydrological models in 12 large-scale river basins worldwide." *Clim. Change* 141 (3): 381–397. <https://doi.org/10.1007/s10584-016-1841-8>.
- Husain, S. Z., N. Alavi, S. Bélair, M. Carrera, S. Zhang, V. Fortin, M. Abrahamowicz, and N. Gauthier. 2016. "The multibudget soil, vegetation, and snow (SVS) scheme for land surface parameterization: Offline warm season evaluation." *J. Hydrometeorol.* 17 (8): 2293–2313. <https://doi.org/10.1175/JHM-D-15-0228.1>.
- HydroSHEDS. 2021. "Hydrosheds website." Accessed March 3, 2021. <https://www.hydrosheds.org/>.
- IJC (International Joint Commission). 2019a. "Causes of the 2019 high water event." Accessed March 3, 2021. <https://ijc.org/en/loslrb/watershed/causes-2019-high-water-event>.
- IJC (International Joint Commission). 2019b. "Lake Ontario outflow sets records in 2019, further increases expected in new year." Accessed March 3, 2021. <https://www.ijc.org/en/loslrb/lake-ontario-outflow-sets-records-2019-further-increases-expected-new-year>.
- Kavetski, D., and M. P. Clark. 2010. "Ancient numerical daemons of conceptual hydrological modeling: 2. Impact of time stepping schemes on model analysis and prediction." *Water Resour. Res.* 46: W10511. <https://doi.org/10.1029/2009WR008896>.
- Kouwen, N. 1988. "Watflood: A micro-computer based flood forecasting system based on real-time weather radar." *Can. Water Resour. J.* 13 (1): 62–77. <https://doi.org/10.4296/cwrj1301062>.
- Kouwen, N. 2010. *WATFLOOD/ WATROUTE hydrological model routing & flow forecasting system*. Waterloo, ON: Univ. of Waterloo.
- Kouwen, N., E. D. Soulis, A. Pietroniro, J. Donald, and R. A. Harrington. 1993. "Grouped response units for distributed hydrologic modeling." *J. Water Resour. Plann. Manage.* 119 (3): 289–305. [https://doi.org/10.1061/\(ASCE\)0733-9496\(1993\)119:3\(289\)](https://doi.org/10.1061/(ASCE)0733-9496(1993)119:3(289)).
- Kratzert, F., D. Klotz, G. Shalev, G. Klambauer, S. Hochreiter, and G. Nearing. 2019. "Towards learning universal, regional, and local hydrological behaviors via machine learning applied to large-sample datasets." *Hydrol. Earth Syst. Sci.* 23 (12): 5089–5110. <https://doi.org/10.5194/hess-23-5089-2019>.
- Kumar, R., L. Samaniego, and S. Attinger. 2013. "Implications of distributed hydrological model parameterization on water fluxes at multiple scales and locations." *Water Resour. Res.* 49 (1): 360–379. <https://doi.org/10.1029/2012WR012195>.
- Lehner, B., K. Verdin, and A. Jarvis. 2008. "New global hydrography derived from spaceborne elevation data." *EOS Trans. Am. Geophys. Union* 89 (10): 93–94. <https://doi.org/10.1029/2008EO100001>.
- Liang, X. 2003. "A new parameterization for surface and groundwater interactions and its impact on water budgets with the variable infiltration capacity (VIC) land surface model." *J. Geophys. Res.* 108 (16): 1989. <https://doi.org/10.1029/2002JD003090>.
- Liang, X., D. P. Lettenmaier, E. F. Wood, and S. J. Burges. 1994. "A simple hydrologically based model of land surface water and energy fluxes for general circulation models." *J. Geophys. Res. Atmos.* 99 (7): 14415–14428. <https://doi.org/10.1029/94JD00483>.
- Lindström, G., C. Pers, J. Rosberg, J. Strömquist, and B. Arheimer. 2010. "Development and testing of the HYPE (hydrological predictions for the environment) water quality model for different spatial scales." *Hydrol. Res.* 41 (3–4): 295–319. <https://doi.org/10.2166/nh.2010.007>.
- Mai, J., K. C. Kornelsen, B. A. Tolson, V. Fortin, N. Gasset, D. Bouhemem, D. Schäfer, M. Leahy, F. Ancil, and P. Coulibaly. 2020. "The Canadian surface prediction archive (CaSPAr): A platform to enhance environmental modeling in Canada and globally." *Bull. Am. Meteorol. Soc.* 101 (3): E341–E356. <https://doi.org/10.1175/BAMS-D-19-0143.1>.
- McInerney, D., M. Thyer, D. Kavetski, J. Lerat, and G. Kuczera. 2017. "Improving probabilistic prediction of daily streamflow by identifying Pareto optimal approaches for modeling heteroscedastic residual errors." *Water Resour. Res.* 53 (3): 2199–2239. <https://doi.org/10.1002/2016WR019168>.

- McSweeney, C. F., and R. G. Jones. 2016. "How representative is the spread of climate projections from the 5 CMIP5 GCMs used in ISI-MIP?" *Clim. Serv.* 1: 24–29. <https://doi.org/10.1016/j.ciser.2016.02.001>.
- Menard, C. B., et al. 2020. "Scientific and human errors in a snow model intercomparison." *Bull. Am. Meteorol. Soc.* 102 (1): 1–46. <https://doi.org/10.1175/BAMS-D-19-0329.1>.
- Michalak, A. M., et al. 2013. "Record-setting algal bloom in Lake Erie caused by agricultural and meteorological trends consistent with expected future conditions." *Proc. Natl. Acad. Sci. U.S.A.* 110 (16): 6448–6452. <https://doi.org/10.1073/pnas.1216006110>.
- Moriasi, D. N., M. W. Gitau, N. Pai, and P. Daggupati. 2015. "Hydrologic and water quality models: Performance measures and evaluation criteria." *Trans. ASABE* 58 (6): 1763–1785. <https://doi.org/10.13031/trans.58.10715>.
- Muhammad, A., T. Stadnyk, F. Unduche, and P. Coulibaly. 2018. "Multi-model approaches for improving seasonal ensemble streamflow prediction scheme with various statistical post-processing techniques in the Canadian Prairie region." *Water* 10 (11): 1604–1618. <https://doi.org/10.3390/w10111604>.
- Nash, J. E., and J. V. Sutcliffe. 1970. "River flow forecasting through conceptual models: Part I—A discussion of principles." *J. Hydrol.* 10 (3): 282–290. [https://doi.org/10.1016/0022-1694\(70\)90255-6](https://doi.org/10.1016/0022-1694(70)90255-6).
- Natural Resources Canada. 2020. "National hydro network—NHN—Geobase series." Accessed March 03, 2021. <https://open.canada.ca/data/en/dataset/a4b190fe-e090-4e6d-881e-b87956c07977>.
- Neitsch, S. L., J. G. Arnold, J. R. Kiniry, and J. R. Williams. 2011. *Soil and water assessment tool theoretical documentation version 2009*. College Station, TX: Texas Water Resources Institute.
- Newman, A. J., et al. 2015. "Development of a large-sample watershed-scale hydrometeorological data set for the contiguous USA: Data set characteristics and assessment of regional variability in hydrologic model performance." *Hydrol. Earth Syst. Sci.* 19 (1): 209–223. <https://doi.org/10.5194/hess-19-209-2015>.
- ORNL and DAAC. 2016. "Global 1-km gridded thickness of soil, regolith, and sedimentary deposit layers." Accessed March 3, 2021. [https://daac.ornl.gov/SOILS/guides/Global\\_Soil\\_Regolith\\_Sediment.html](https://daac.ornl.gov/SOILS/guides/Global_Soil_Regolith_Sediment.html).
- Perrin, C., C. Michel, and V. Andréassian. 2003. "Improvement of a parsimonious model for streamflow simulation." *J. Hydrol.* 279 (1–4): 275–289. [https://doi.org/10.1016/S0022-1694\(03\)00225-7](https://doi.org/10.1016/S0022-1694(03)00225-7).
- Pietroniro, A., et al. 2007. "Development of the MESH modelling system for hydrological ensemble forecasting of the Laurentian Great Lakes at the regional scale." *Hydrol. Earth Syst. Sci.* 11 (4): 1279–1294. <https://doi.org/10.5194/hess-11-1279-2007>.
- Rakovec, O., R. Kumar, J. Mai, M. Cuntz, S. Thober, M. Zink, S. Attinger, D. Schäfer, M. Schrön, and L. Samaniego. 2016. "Multiscale and multivariate evaluation of water fluxes and states over European river basins." *J. Hydrometeorol.* 17 (1): 287–307. <https://doi.org/10.1175/JHM-D-15-0054.1>.
- Rakovec, O., R. Kumar, M. McLeod, J. Mai, and L. Samaniego. 2020. "mHM-UZF gridded simulations for the Great Lakes Runoff Intercomparison Project for Lake Erie." Accessed March 19, 2021. <https://doi.org/10.5281/zenodo.3886551>.
- Rakovec, O., N. Mizukami, R. Kumar, A. J. Newman, S. Thober, A. W. Wood, M. P. Clark, and L. Samaniego. 2019. "Diagnostic evaluation of large-domain hydrologic models calibrated across the contiguous United States." *J. Geophys. Res. Atmos.* 124 (24): 13991–14007. <https://doi.org/10.1029/2019JD030767>.
- Reed, S., V. Koren, M. Smith, Z. Zhang, F. Moreda, and D. J. Seo. 2004. "Overall distributed model intercomparison project results." *J. Hydrol.* 298 (1–4): 27–60. <https://doi.org/10.1016/j.jhydrol.2004.03.031>.
- Rosenzweig, C., et al. 2017. "Assessing inter-sectoral climate change risks: The role of ISIMIP." *Environ. Res. Lett.* 12 (1): 010301. <https://doi.org/10.1088/1748-9326/12/1/010301>.
- Roy, T., H. V. Gupta, A. Serrat-Capdevila, and J. B. Valdes. 2017. "Using satellite-based evapotranspiration estimates to improve the structure of a simple conceptual rainfall–runoff model." *Hydrol. Earth Syst. Sci.* 21 (2): 879–896. <https://doi.org/10.5194/hess-21-879-2017>.
- Samaniego, L., et al. 2017. "Toward seamless hydrologic predictions across spatial scales." *Hydrol. Earth Syst. Sci.* 21 (9): 4323–4346. <https://doi.org/10.5194/hess-21-4323-2017>.
- Samaniego, L., R. Kumar, and S. Attinger. 2010. "Multiscale parameter regionalization of a grid-based hydrologic model at the mesoscale." *Water Resour. Res.* 46 (5): W05523. <https://doi.org/10.1029/2008WR007327>.
- Schmale, D. G. III, A. P. Ault, W. Saad, D. T. Scott, and J. A. Westrick. 2019. "Perspectives on harmful algal blooms (HABs) and the cyberbiosecurity of freshwater systems." *Front. Bioeng. Biotechnol.* 7 (128): 1–7. <https://doi.org/10.3389/fbioe.2019.00128>.
- Shangguan, W., Y. Dai, Q. Duan, B. Liu, and H. Yuan. 2008. "The global soil dataset for earth system modeling." Accessed March 3, 2021. <http://globalchange.bnu.edu.cn/research/soilw>.
- Shangguan, W., T. Hengl, J. de Jesus, H. Yuan, and Y. Dai. 2017. "A global depth to bedrock dataset for earth system modeling." Accessed March 3, 2021. <http://globalchange.bnu.edu.cn/research/dtb.jsp>.
- Shen, H., J. Mai, B. A. Tolson, and M. Han. 2020. "Watershed shapes for the Great Lakes Runoff Intercomparison Project for Lake Erie (GRIP-E)." Accessed March 3, 2021. <https://doi.org/10.5281/zenodo.3888690>.
- Smith, M., V. Koren, S. Reed, Z. Zhang, Y. Zhang, F. Moreda, Z. Cui, N. Mizukami, E. A. Anderson, and B. A. Cosgrove. 2012. "The distributed model intercomparison project—Phase 2: Motivation and design of the Oklahoma experiments." *J. Hydrol.* 418–419: 3–16. <https://doi.org/10.1016/j.jhydrol.2011.08.055>.
- Smith, M. B., D.-J. Seo, V. I. Koren, S. M. Reed, Z. Zhang, Q. Duan, F. Moreda, and S. Cong. 2004. "The distributed model intercomparison project (DMIP): Motivation and experiment design." *J. Hydrol.* 298 (1–4): 4–26. <https://doi.org/10.1016/j.jhydrol.2004.03.040>.
- SoilGrids. 2020. "Soilgrids250m 2.0." Accessed March 3, 2021. <https://soilgrids.org/>.
- Spieler, D., J. Mai, J. R. Craig, B. A. Tolson, and N. Schütze. 2020. "Automatic model structure identification for conceptual hydrologic models." *Water Resour. Res.* 56 (9): e2019WR027009. <https://doi.org/10.1029/2019WR027009>.
- Thober, S., M. Cuntz, M. Kelbling, R. Kumar, J. Mai, and L. Samaniego. 2019. "The multiscale routing model mrm v1.0: Simple river routing at resolutions from 1 to 50 km." *Geosci. Model Dev.* 12 (6): 2501–2521. <https://doi.org/10.5194/gmd-12-2501-2019>.
- Thompson, S. A., R. L. Stephenson, G. A. Rose, and S. D. Paul. 2019. "Collaborative fisheries research: The Canadian Fisheries Research Network experience." *Can. J. Fish. Aquat. Sci.* 76 (5): 671–681. <https://doi.org/10.1139/cjfas-2018-0450>.
- USACE. 2020. "Great Lakes water levels still setting records." Accessed March 3, 2021. <https://www.lre.usace.army.mil/Media/News-Releases/Article/2217758/great-lakes-water-levels-still-setting-records/>.
- USC (Universidad de Santiago de Compostela). 2013. "Global patterns of groundwater table depth." Accessed March 3, 2021. <http://thredds-gfml.usc.es/thredds/catalog/GLOBALWTDFTP/catalog.html>.
- USDA. n.d. "Natural resources conservation service soils—Soil geography." Accessed March 3, 2021. <https://www.nrcs.usda.gov/wps/portal/nrcs/detail/soils/survey/geol/>.
- USDA. 2021. "Cropscape and cropland data layer." Accessed March 3, 2021. [https://www.nass.usda.gov/Research\\_and\\_Science/Cropland/SARS1a.php](https://www.nass.usda.gov/Research_and_Science/Cropland/SARS1a.php).
- USEPA. n.d. "Facts and figures about the Great Lakes." Accessed March 3, 2021. <https://www.epa.gov/greatlakes/facts-and-figures-about-great-lakes>.
- USGS. 2010. "USGS eros archive—digital elevation-global multi-resolution terrain elevation data." Accessed March 3, 2021. <https://www.usgs.gov/centers/eros/science/usgs-eros-archive-digital-elevation-global-multi-resolution-terrain-elevation>.
- USGS. 2018. "USGS eros archive—digital elevation global 30 arc-second elevation (gtopo30)." Accessed March 3, 2021. <https://www.usgs.gov/centers/eros/science/usgs-eros-archive-digital-elevation-global-30-arc-second-elevation-gtopo30>.
- USGS. 2019. "Mcd12q1 v006 modis/terra+aqua land cover type yearly 13 global 500 m sin grid." Accessed March 3, 2021. <https://lpdaac.usgs.gov/products/mcd12q1v006/>.
- USGS. 2021. "National hydrography dataset." Accessed March 3, 2021. <https://www.usgs.gov/core-science-systems/ngp/national-hydrography/national-hydrography-dataset>.

- Valéry, A., V. Andréassian, and C. Perrin. 2014. “‘As simple as possible but not simpler’: What is useful in a temperature-based snow-accounting routine? Part 1—Comparison of six snow accounting routines on 380 catchments.” *J. Hydrol.* 517 (Sep): 1166–1175. <https://doi.org/10.1016/j.jhydrol.2014.04.059>.
- Verseghy, D. L. 2000. “The Canadian land surface scheme (CLASS): Its history and future.” *Atmos. Ocean* 38 (1): 1–13. <https://doi.org/10.1080/07055900.2000.9649637>.
- Wada, Y., et al. 2013. “Multimodel projections and uncertainties of irrigation water demand under climate change.” *Geophys. Res. Lett.* 40 (17): 4626–4632. <https://doi.org/10.1002/grl.50686>.
- Wang, L., et al. 2015. “A spatial classification and database for management, research, and policy making: The Great Lakes aquatic habitat framework.” *J. Great Lakes Res.* 41 (2): 584–596. <https://doi.org/10.1016/j.jglr.2015.03.017>.
- Warszawski, L., K. Frieler, V. Huber, F. Piontek, O. Serdeczny, and J. Schewe. 2014. “The inter-sectoral impact model intercomparison project (ISI-MIP): Project framework.” *Proc. Natl. Acad. Sci. U.S.A.* 111 (9): 3228–3232. <https://doi.org/10.1073/pnas.1312330110>.
- Zink, M., J. Mai, M. Cuntz, and L. Samaniego. 2018. “Conditioning a hydrologic model using patterns of remotely sensed land surface temperature.” *Water Resour. Res.* 54 (4): 2976–2998. <https://doi.org/10.1002/2017WR021346>.

# SUPPLEMENTAL MATERIALS

*ASCE Journal of Hydrologic Engineering*

## Great Lakes Runoff Intercomparison Project Phase 3: Lake Erie (GRIP-E)

Juliane Mai, Bryan A. Tolson, Hongren Shen, Étienne Gaborit, Vincent Fortin, Nicolas Gasset, Hervé Awoye, Tricia A. Stadnyk, Lauren M. Fry, Emily A. Bradley, Frank Seglenieks, André G. Temgoua, Daniel G. Princz, Shervan Gharari, Amin Haghnegahdar, Mohamed E. Elshamy, Saman Razavi, Martin Gauch, Jimmy Lin, Xiaojing Ni, Yongping Yuan, Meghan McLeod, Nandita Basu, Rohini Kumar, Oldrich Rakovec, Luis Samaniego, Sabine Attinger, Narayan K. Shrestha, Prasad Daggupati, Tirthankar Roy, Sungwook Wi, Tim Hunter, James R. Craig, and Alain Pietroniro

**DOI:** 10.1061/(ASCE)HE.1943-5584.0002097

© ASCE 2021

U.S. government works are not subject to copyright.

## S.1 Detailed Description of Model Setups

### S.1.1 Machine Learning Models

The following gives a brief and somewhat simplified introduction to our machine learning models; we refer to [Gauch et al. \[2019\]](#) for more details on our training procedures. We use two types of models: the gradient-boosted regression tree (GBRT) framework XGBoost [[Chen and Guestrin, 2016](#)] and a Long Short-Term Memory (LSTM) architecture [[Hochreiter and Schmidhuber, 1997](#)]. Both models take as input the average temperature and total precipitation of the previous thirty days, concatenated with static basin attributes (area, mean river slope and length, mean basin slope, mean river bank width and depth, mean elevation, Manning’s river flood plain coefficient, mean streamflow, Manning’s river channel coefficient, and a binary variable indicating whether the basin is regulated). All inputs are lumped for each basin, and we train one model on all basins combined.

GBRT are based on regression trees. Regression trees construct a directed, tree-like graph during calibration. Each root-to-leaf path maps a distinct conjunction of input properties (e.g.,  $(temp > 0) \wedge (precip > 0)$ ) to a predicted streamflow value. To generate a prediction for a given day, we search the path that evaluates to True for that day’s input and predict the corresponding streamflow value. GBRT iteratively train multiple regression trees to predict the previous iteration’s error and additionally apply regularization techniques.

The LSTM is a type of neural network that processes lumped time series. As we ingest the time series, the network updates internal memory states that it considers when generating predictions. Unlike process-based hydrologic models, these states do not have any semantic interpretation but are learned by the model. To tune the network parameters (weights), we generate predictions with the current weights, calculate their error, and update the parameters to improve the predictions.

### S.1.2 LBRM

The Large Basin Runoff Model (LBRM, described in [Croley II \[1983\]](#), with recent modifications described in [Gronewold et al. \[2017\]](#)) is a lumped conceptual model that propagates daily precipitation and temperature into subbasin runoff. LBRM was developed by NOAA Great Lakes Environmental Research Laboratory (GLERL) specifically for use in simulating total runoff contribution to the Great Lakes. It is one of a suite of models run by the U.S. Army Corps of Engineers – Detroit District for simulating historical runoff into the Great Lakes as well as informing seasonal net-basin supply forecasts as part of the U.S. contribution to the internationally coordinated 6-month forecast of Great Lakes water levels. LBRM is the only rainfall-runoff model that is used operationally to produce forecasts of runoff for use in water level forecasts on a seasonal to interannual basis. As it is configured at USACE-Detroit, the model operates on 121 subbasins throughout the Great Lakes, 21 of which are located in the Lake Erie basin. The Lake Erie subbasins range in size from 119 km<sup>2</sup> to 16 806 km<sup>2</sup>.

For this study, LBRM’s nine parameters are calibrated for each Lake Erie subbasin using a Dynamically Dimensioned Search algorithm (DDS) [[Tolson and Shoemaker, 2007](#)] encoded within the Ostrich optimization software package [[Matott, 2017](#)]. The DDS algorithm is run for 300 iterations for each of the 21 USACE subbasins simulating runoff for the period from January 1, 2010 to December 31, 2014 while discarding the first year as warm-up. The RDRS dataset was lumped to the subbasins and used as forcings. The objective during calibration is to maximize the Nash-Sutcliffe Efficiency score between simulated subbasin runoff and area-ratio-derived estimates of subbasin runoff provided by GLERL [[Hunter et al., 2015](#)]. This calibration is performed for each of the 21 USACE sub-basins. The Area-Ratio method [[Fry et al., 2014](#)] is subsequently used to retrieve the hydrographs for the 46 objective 1 and objective 2 streamflow gauges based on the 21 optimal USACE setups.

### S.1.3 Lumped and semi-distributed GR4J

The GR4J rainfall-runoff model is a parsimonious lumped model with four parameters and is usually operated at a daily scale [Perrin et al., 2003]. This model has been widely used in hydrologic modelling studies for both operational and investigative purposes, and it has shown good performance in streamflow simulation [Seiller et al., 2017; Arsenault et al., 2018; Wright et al., 2018]. The original GR4J model is comprised of a runoff production store, a routing store and two unit hydrographs. Since the simulation of the snow processes is necessary in this study, the original 4-parameter GR4J model was coupled with the Cema-Neige snow module [Valéry et al., 2014]. This snow module simulates the snow cover and snowmelt processes with two parameters. Further, the 6-parameter GR4J model is emulated by the Raven model [Craig et al., 2020]. The GR4J model requires daily precipitation, air temperature and potential evapotranspiration series as inputs. The potential evapotranspiration is estimated using the Hargreaves equation. This empirical approach is based on the air temperature and incoming solar radiation, where the incoming solar radiation is calculated using the equation based on Dingman [2015]. In this study, there are two versions of the GR4J model used in this study both emulated in Raven: (1) Lumped GR4J (GR4J-lp) model, which is established at the lumped watershed scale, thus simulating hydrological processes for the entire watershed. The GR4J-lp model ignores river channel routing process. (2) Semi-distributed GR4J (GR4J-sd) model, which requires a discretization of watershed. It simulates the runoff production processes within each sub-watershed independently, and then routes water from sub-watersheds to the outlet. Since a watershed is discretized into several sub-watersheds based on the topographical characteristics, river channel routing processes can be better simulated in this model version for large watersheds. Meteorological forcing, i.e., daily precipitation and air temperature, are aggregated from the grid-cell to sub-watershed scale based on the RDRS data set (Table 2). The runoff production and potential evapotranspiration calculation in both GR4J model versions are the same. The GR4J-lp model uses six parameters (four for the original GR4J model and two for the snow module) for calibration. The Manning's coefficient of the river channel is additionally used in the GR4J-sd model for river routing calibration. In this study, both the GR4J-lp and GR4J-sd are calibrated in the 46 catchments and validated in seven catchments (Section S.2). A single-gauge calibration strategy is applied for the model parameter calibration that the parameters are independently tuned in each catchment, and thus yielding 46 different parameter sets after optimization. The Dynamically Dimensioned Search (DDS) algorithm is employed to auto-calibrate model parameters [Tolson and Shoemaker, 2007]. The DDS is used here by employing the optimization software toolkit Ostrich [Matott, 2017]. The Nash-Sutcliffe efficiency (NSE) is utilized as the calibration objective function. The auto-calibration at each catchment terminates when the maximum budget of 1000 model evaluations is reached. The calibration at each catchment is repeated for ten independent trails to eliminate the influence of randomness. The best result out of these ten trails is reported.

### S.1.4 HYMOD2-DS

HYMOD [Boyle et al., 2000] is a conceptual hydrological model for catchment-scale simulation of rainfall-runoff processes. The model uses precipitation and potential evapotranspiration (PET) as inputs to generate streamflow and actual evapotranspiration (AET) as outputs. The model is based on the probability-distributed storage capacity concept of Moore [1985], which represents the vertical soil moisture accounting process. The original HYMOD is lumped in nature, where the horizontal routing is carried out by a Nash Cascade (leaky linear reservoirs connected in series to represent surface and subsurface flows across the watershed) and a leaky linear reservoir (to represent baseflow).

In this study, we used a modified version of the original HYMOD model. The soil moisture accounting process is based on the new HYMOD2 [Roy et al., 2017], which has an improved parameterization for the evaporation process. HYMOD2 is coupled with a river routing model to be suitable for modeling a distributed watershed system as described in Wi et al. [2015]. Additionally, we also coupled the Degree Day Snow model [Martinec, 1975] with



HYMOD2 since snow is an important factor for several of the catchments under consideration. PET is derived based on the Hamon method [Hamon, 1961], in which, daily PET is computed as a function of daily mean temperature and hours of daylight. We call this modified model, the HYMOD2-DS, where "DS" is used to denote the distributed version of HYMOD2. In this version of the model, we calibrated a total of 12 parameters, one from the PET module, one from the snow module, six from HYMOD2, and four from the routing module.

The model was calibrated using the Shuffled Complex Evolution (SCE) algorithm [Duan et al., 1992] with two complexes. In total 25 loops equaling about 2000 model evaluations are used as the calibration budget to minimize the mean squared error of each of the 46 sub-watersheds.

### S.1.5 SWAT-EPA

The Soil and Water Assessment Tool (SWAT) model is a semi-distributed process based hydrologic model considering the physical characteristics of the watershed including surface elevation, soil type, land use, and factors affecting water routing within the watershed [Arnold et al., 1993; Neitsch et al., 2011]. Moreover, it contains modules that simulate agricultural activities such as irrigation and fertilization. Given that 74% of Lake Erie basin are covered by agricultural cropland [of Canada, 2015; Survey, 2014] providing large quantity of nutrient load into Lake Erie [Dolan and Chapra, 2012], we constructed a SWAT model to investigate agriculture activity affects on stream water quality in Lake Erie basin. Streamflow as a part of the output of the model was provided to GRIP-E project comparing with other hydrological model. The model was setup by the co-authors at the U.S. Environmental Protection Agency (EPA) and will be called SWAT-EPA hereafter.

In this project, ArcSWAT 2012.10.21 with SWAT Rev. 670 (<https://swat.tamu.edu/software/arcswat/>) was used to construct a SWAT model with inputs shown in Table 3. The Lake Erie basin was delineated based on a DEM with area of 9057 ha. The study area was divided into 176 sub-watersheds and contained 3398 Hydrological Response Units (HRUs) based on soil type, land use, and slope length with thresholds of 5%, 5%, and 10%, respectively. The model simulated daily stream flow from 2010 to 2014 with two years warming period. Calibration process was achieved by using the SWAT-CUP SUFI2 algorithm [Abbaspour et al., 2004]. The calibration process was performed on three gauge stations (04159492, 02GG006, and 04213000) to maximize Nash-Sutcliffe Efficiency (NSE) [Nash and Sutcliffe, 1970] and the fitted values were further applied to the rest of stations in objective 1. SWAT-EPA did not contribute to objective 2.

### S.1.6 SWAT-Guelph

Soil and Water Assessment Tool (SWAT) [Arnold et al., 1998] is a physically based, semi-distributed continuous long-term simulation model developed by United States Department of Agriculture (USDA). The model was setup by the co-authors at the University of Guelph and will be called SWAT-Guelph hereafter. The model divides a watershed into sub-basins which are further divided into hydrological response units (HRUs) – the computational units of the model. The conditioned HydroSHEDS digital elevation model (DEM) with spatial resolution of 90m (<https://hydrosheds.cr.usgs.gov/index.php>) was used to derive stream network and 699 sub-basins making sure that considered streamflow gauging location corresponds to a sub-basin. The Soil Landscapes of Canada (SLC) version 3.2 (<http://sis.agr.gc.ca/cansis/nsdb/slc/v3.2/base.html>) at 1:1million scale was merged with the digital general soil map of the US (STATSGO), available at ArcSWAT database (<https://swat.tamu.edu/software/arcswat/>) at 250 m spatial resolution to create a combined soil map. The soil map was overlaid with the Terra and Aqua combined MODIS Land Cover Type (MCD12Q1) version 6 ([https://lpdaac.usgs.gov/dataset\\_discovery/modis/modis\\_products\\_table/mcd12q1\\_v006](https://lpdaac.usgs.gov/dataset_discovery/modis/modis_products_table/mcd12q1_v006)) with spatial resolution of 500 m, and the DEM derived slope map to create 7777 HRUs. In the next step, the

Regional Deterministic Reforecast System (RDRS) based daily precipitation, maximum and minimum temperature, solar radiation, relative humidity and wind speed data were used to create a functional SWAT model. The Curve number (CN) method [USDA-SCS, 1986] was used to calculate surface runoff and infiltration. Similarly, Penman-Monteith equation was used to calculate evapotranspiration and variable storage routing [Williams, 1969] was used to channel routing. The SWAT-CUP and its SUFI2 algorithm [Abbaspour et al., 2004] was used to conduct multi-site calibration of 17 model parameters known to influence streamflow in similar watershed [Zhang et al., 2018] in a daily time step, considering Nash-Sutcliffe Efficiency (NSE) [Nash and Sutcliffe, 1970] as the objective function. In the process, a single iteration of 2000 model runs was conducted.

It need to be noted that some model parameters in SWAT are global (e.g., SMTMP.bsn) while others are local (e.g., SURLAG.hru). This is leading to the fact that validation results for the same station depend on the objective meaning that two validation time series for the same station can be derived. This can be seen, for example, in Fig. 2C and 2D of the main manuscript where the performance of the stations is different for the two objectives.

### S.1.7 mesoscale Hydrologic Model (mHM)

The mesoscale Hydrologic Model (mHM) [Samaniego et al., 2010; Kumar et al., 2013] is a distributed hydrologic model that use grid cell as a primary hydrologic unit and accounts for variety of hydrologic processes including canopy interception, root-zone soil moisture, infiltration, evapotranspiration, runoff generation as well as river flows along the stream network [Thober et al., 2019]. mHM reached the technology readiness level 9 with the Copernicus Climate Change Service proof-of-concept (pre-operation) project EDgE [Samaniego et al., 2020]. This model is also used operationally in the German Drought Monitor [Zink et al., 2016]. The model is forced with the daily gridded fields of at least precipitation, and minimum/ maximum temperature. mHM provides several evapotranspiration (PET) parameterization that depend on the data availability (e.g., Hargreaves-Samani, Penman-Monteith, Priesley-Taylor). It uses a novel multiscale parameter regionalization (MPR) scheme to account for the sub-grid variability of basin physical properties that allows for the seamless predictions of water fluxes and storages at different spatial resolutions and ungauged locations [Rakovec et al., 2016; Samaniego et al., 2017]. The model has been extensively evaluated in several studies, the code is open source, available on online repository [git.ufz.de/mhm](https://git.ufz.de/mhm).

mHM was run in two versions with efforts combining two different organizations – hereafter named as “mHM-Waterloo” and “mHM-UFZ”. While the both model versions uses the same source code ([www.ufz.de/mhm](http://www.ufz.de/mhm); release version 5.10), they differ in their usages of underlying basins physiographical data-sets (see Table 3 for more details) and in general model set-ups. The main difference between the two models is the parameters estimation approach. In mHM-Waterloo the transfer parameters are estimated for each basin independently whereas in the mHM-UFZ, they are found as a compromise solution for all basins. Another reason of different model performances is the selection of the datasets used to setup the models (see Table 2 in the main manuscript). The “mHM-Waterloo” version was established for each basin separately with parameters being specifically calibrated to each study basin; whereas the “mHM-UFZ” was established as a regional model with a single set of model parameters being applicable to each individual group of study basins (i.e., objective 1 and objective 2). In this way, we designated the “mHM-Waterloo” version more to the lumped category (i.e., basin specific) and “mHM-UFZ” to the distributed one. Both model versions were calibrated using Nash-Sutcliffe Efficiency (NSE) as the objective functions and the Dynamically Dimensioned Search (DDS) [Tolson and Shoemaker, 2007] as the optimization algorithm with 1000 model iterations. The model in general has up to 50 free calibration parameters; however a controlled calibration experiment carried out by [Rakovec et al., 2019] showed similar performance of mHM over 492 US basins using the full range calibration parameters and subset of randomly selected 14 parameters.

### S.1.8 HYPE

The Hydrological Predictions for the Environment (HYPE) model is an operational hydrologic model developed at the Swedish Meteorological and Hydrological Institute. HYPE includes hydrological processes above ground, land routines and deep processes such as snow/ice accumulation and melt, evapotranspiration, soil moisture of up to three soil layers and flow paths, frozen soil infiltration, groundwater movement and aquifer recharge, surface-water routing through rivers and lakes, and human perturbations through diversion, reservoirs, regulation, irrigation and water abstractions [Lindström et al., 2010]. Though HYPE is more conceptual in nature, it operates at a sub-basin scale and integrates physiographic characteristics related to elevation, land cover/land use, and soil types which control the spatial variation of the processes represented. HYPE is used for operational hydrological forecasts in Europe [Pechlivanidis et al., 2014] and was also adapted for large-scale applications across climate regions [Arheimer et al., 2020; Bajracharya et al., 2020; Pechlivanidis and Arheimer, 2015; Strömqvist et al., 2009]. The initial set of parameters used for the development of the HYPE model for the Lake Erie Basin is taken from the Arctic-HYPE configuration described in Stadnyk et al. [2020]. We used HYPE version 5.7.0 available at <https://sourceforge.net/projects/hype/files/>. An objective function is optimized for multiple stations using daily streamflow and a composite criterion combining the average of the Nash-Sutcliffe efficiencies of selected sub-basins and the average of their relative bias. Thus, a stepwise automatic calibration approach based on the Differential Evolution Markov Chain method [Braak, 2006] is employed to derive a set of optimal values for the model parameters. These parameters are general or linked to land use/land cover and soil types. The HYPE model code is open source and supported by wiki documentation (<http://www.smhi.net/hype/wiki/doku.php?id=start>) and users' discussion forum.

### S.1.9 VIC

The Variable Infiltration Capacity (VIC) model is a macro-scale distributed hydrological model that balances both the water and surface energy budgets [Liang et al., 1994; Liang, 2003]. This model has been extensively applied in hydrology such as streamflow simulation [Reed et al., 2004; Gao et al., 2010; Livneh et al., 2013]. VIC simulates land surface-atmospheric fluxes of moisture and energy such as evapotranspiration, surface runoff, baseflow, radiative fluxes, turbulent fluxes of transport, and sensible heat within the grid-cell. The gridded runoff components, comprising surface runoff and baseflow, are then routed to the basin outlet. In this study, the image version 5.1.0 of VIC model is used, which can be retrieved at <https://vic.readthedocs.io/en/master/Development/ModelDevelopment>. The VIC model is built for the RDRS forcing grid-cells with a resolution of 15 km × 15 km in the Lake Erie basin. The DEM, soil, and land cover data specified in Table 3 are utilized for VIC parameterization. These data are all aggregated to the 15 km grid-cell scale. In addition, VIC requires sub-daily meteorological drivers from the RDRS forcing data set, i.e. precipitation, air temperature, atmospheric pressure, incoming shortwave radiation, incoming longwave radiation, vapor pressure, and wind speed. Since the VIC version 5.1.0 does not internally contain a routing module, the Raven model is employed as an independent routing module for the VIC. The VIC-generated runoff and baseflow fluxes at the grid-cell scale are first aggregated to the sub-watershed scale, and then routed to the catchment outlet in terms of hillslope routing and river channel routing processes. There are multiple user-calibrated parameters in VIC model furnished by Gao et al. [2010]. In this study, seven parameters are selected for model calibration based on Xie et al. [2007] and Wen et al. [2011]. The seven parameters selected are the exponent of the variable infiltration capacity curve  $b$ , the maximum velocity of baseflow  $Ds_{max}$ , fraction of  $Ds_{max}$  where non-linear baseflow begins  $Ds$ , fraction of maximum soil moisture where non-linear baseflow occurs  $Ws$ , thickness of the top thin layer  $d_1$ , the middle layer  $d_2$ , and the bottom layer  $d_3$ . More details of the parameter definitions and value ranges can be found in Gao et al. [2010]. In this study, the VIC model is calibrated in 46 catchments and validated in seven catchments (Section S.2). A global gauge calibration strategy is applied for the model parameter calibration that the parameters are concurrently

tuned in each catchment, thus yielding only one parameter set after optimization. The VIC model is auto-calibrated using Ostrich [Matott, 2017] using the Dynamically Dimensioned Search algorithm (DDS) [Tolson and Shoemaker, 2007]. The median Nash-Sutcliffe efficiency (NSE) of streamflow at the 46 calibration stations is utilized as the calibration objective function. The global auto-calibration terminates when the maximum allowable limit of 1000 model evaluations is reached. To eliminate influence of randomness in calibration and take into account the computational burden, five trials are applied in the global calibration. The best result out of the five trials is reported.

#### S.1.10 VIC-GRU

VIC-GRU is a vector-based implementation of the VIC model using the concept of Grouped Response Units (GRUs). Traditionally, the VIC model is set up at grid scale while sub-grid variability and spatial extents are smeared. VIC-GRU instead is set up for the exact spatial extent of units that are identified to have similar soil type or vegetation cover.

For the Lake Erie basin and given the soil and land cover classes, eight GRUs are identified that can be classified with four soil types on two general land cover of forested and non-forested areas. When the GRUs are forced at the resolution of the RDRS data set it results in 2380 computational units. The model then simulates the states and fluxes for each computational unit. A full description of VIC-GRU implementation is provided by Gharari et al. [2020]. VIC-GRU includes representation of subsurface preferential flow to further enhance the model capabilities to reproduce flashier hydrographs [Gharari et al., 2019]. For the routing model, the mizuRoute stand-alone routing model is used [Mizukami et al., 2016]. The computational units output have been passed to the sub-basins and are then routed.

The model is calibrated to maximize the sum of NSE values for all the available gauges for calibration (Sec. S.2). The ten parameters representing the variable infiltration capacity, saturated hydraulic conductivity, slope of water retention curve, depth of the soil layer (first and second layers), fraction of subsurface macropore water movement, baseflow coefficient, LAI and stomatal resistance scaling factor, depth scaling factor for the vegetated areas which are none-forested in comparison to the forested area are calibrated. For the routing model, velocity and diffusivity are also calibrated as they are often sensitive parameters in large scale modeling [Haghnegahdar and Razavi, 2017]. The calibration is performed with total budget of 1000 simulations using the genetic algorithm by Yoon and Shoemaker [2001] provided in the Ostrich framework [Matott, 2017].

#### S.1.11 GEM-Hydro, MESH-SVS, and MESH-CLASS

GEM-Hydro is a physically-based, distributed hydrologic model developed at Environment and Climate Change Canada (ECCC). It relies on GEM-Surf [Bernier et al., 2011] to represent five different surface tiles (glaciers, water, ice over water, urban, land). The land tile is represented with the SVS (Soil, Vegetation and Snow) Hydrologic Land Surface Scheme (HLSS). See Alavi et al. [2016] and Husain et al. [2016] for more information on SVS. GEM-Hydro also relies on Watroute [Kouwen, 2010], a 1-D Hydraulic model, to perform 2-D channel and reservoir routing. See Gaborit et al. [2017] for more information on GEM-Hydro.

MESH (Modélisation Environnementale communautaire - Surface and Hydrology) is a complimentary community hydrologic modelling platform maintained by ECCC [Pietroniro et al., 2007]. The MESH framework includes SVS among its HLSSs, as well as the Canadian Land Surface Scheme (CLASS) [Verseghy, 2000]. CLASS is another model developed at ECCC and used in the Canadian Global Climate Model (GCM). Within MESH and GEM-Hydro, the HLSS is responsible for coupled energy and water balance in the vertical dimension, while Watroute is used for routing the runoff, lateral flow and drainage generated by the HLSS from one grid cell to the next through a 2-D horizontal grid. MESH can operate in distributed and semi-distributed modes.

One significant difference between CLASS and SVS is that CLASS discretizes the soil in vertical layers for both soil moisture and temperature whereas the version of SVS used here relies on a Force-Restore method for predicting soil and snow temperature, relying on a vertical discretization of the soil only for representing the soil moisture profile and fluxes. Because of this, the soil layer discretization used for the energy balance is the same as used for the water balance in CLASS, while it is simplified to two layers, representing a thin near-surface soil and thick deep soil, for the energy balance in SVS. The version of SVS used in this study does not represent soil freezing and thawing processes, while CLASS does. The implementation of CLASS in MESH includes sloped hydrology, which is not included in the CLASS version of the Canadian GCM. Both CLASS and SVS use the same multi-layer soil discretization and algorithms for sloped interflow calculations but additionally sloped CLASS includes overland runoff routing to an assumed river within each grid cell [Soulis et al., 2011]. When computing the energy balance, both CLASS and SVS partition a grid cell into four subareas, based on the presence or absence of snow and of tall vegetation. However, CLASS includes more detailed physics for the energy and water budget in the presence of vegetation, as it computes a separate energy budget for the canopy and accounts for interception of rain and snow by the vegetation. While SVS has more vegetation classes to choose from with parameters stored in look-up tables, CLASS has only four vegetation sub-types (in addition to urban and barren land) and different parameters have to be assigned if sub-classes of these are to be distinguished (e.g. SVS has different classes for crops such as rice, cotton, etc. while CLASS has one cropland canopy type and to differentiate by crop type, one needs to create separate GRUs with different parameters for each).

In this work, two MESH configurations are used, one emulating GEM-Hydro with SVS as the HLSS (called "MESH-SVS"), and one using CLASS as the HLSS (called "MESH-CLASS"). MESH is distinguished from GEM-Hydro as it can be run outside of ECCO's computational infrastructure (i.e. in stand-alone mode), which however means that MESH cannot be run in a fully-coupled mode with the GEM 3-D atmospheric model. Therefore, MESH-SVS requires less computation time than GEM-Hydro to run hydrologic experiments over small to medium-sized basins, because GEM-Hydro is a component of a much larger modelling infrastructure and system. Moreover, when using SVS, MESH can only represent one (the land) out of the five surface tiles represented in GEM-Surf. Because of these differences, GEM-Hydro and MESH-SVS are treated here as two different hydrological models.

MESH-SVS was used to calibrate SVS and Watroute parameters, some of which were then transferred into GEM-Hydro (see further). Despite the above differences between MESH-SVS and GEM-Hydro, simulation differences in terms of total Lake Erie daily inflow (see further down) were judged to be within an acceptable margin between the two platforms, thus justifying the methodology employed here.

The geophysical databases used for the HLSS in GEM-Hydro and MESH-SVS consist of the Gridded Soil Dataset for Earth System modelling (GSDE, 1 km resolution, eight layers reaching a total depth of 2.3m, see [Shangguan et al., 2014] for soil texture, the ESA CCI LC 2015 Global Map (European Space Agency Climate Change Initiative Land Cover) for land cover (300 m resolution), and the USGS 1 km GTOPO30 Digital Elevation Model to derive surface and soil slopes. The 30 arcsec (1 km) HydroSHEDS dataset [Lehner et al., 2008] was used to derive 1 km flow directions, drainage areas, and channel properties for Watroute in both models. The same databases were used for MESH-CLASS except that the eight GSDE soil layers were aggregated to four to avoid too thin layers (< 10 cm thick) that can cause numerical instabilities and then another fifth layer was added to reach a depth of 5.5 m by repeating the properties of the last GSDE layer. Additionally, the depth to bedrock (or depth of permeable layer - *SDEP*) was derived for MESH-CLASS from the spatially distributed dataset by Shangguan et al. [2017]. For most grid cells, the *SDEP* was below the modelled soil column. The three models employed here (MESH-SVS, MESH-CLASS, and GEM-Hydro) use a 5 arcmin resolution ( $\approx 10$  km) for the surface part, and a 30 arcsec resolution ( $\approx 1$  km) for the routing. For MESH-SVS, the routing was run using two different resolutions: a 10 km

resolution was used during calibration, while a 1 km resolution was used to perform the final run, in order to obtain more realistic hydrology at streamflow gauge locations. Running the routing at a 10 km resolution resulted in some significant inaccuracies in terms of river network delineation but allowed to save a significant amount of computation time and led to similar streamflow simulations in terms of total daily inflows to Lake Erie. In total, 14 SVS parameters were calibrated using MESH-SVS following the methodology employed in [Gaborit et al. \[2017\]](#) using a maximum of 300 simulations. Four Watroute parameters were also calibrated: two baseflow parameters, and two types of Manning coefficients. This methodology consists of a global calibration strategy (see [Gaborit et al. \[2015\]](#)), since the same calibrated parameter set is then used for all subbasins of the Lake Erie watershed. These calibrated parameters were then transferred to GEM-Hydro, except the Manning coefficients. Indeed, MESH can only use fixed values for Manning coefficients and a given river class, while GEM-Hydro can use temporally and spatially-varying Manning coefficients which vary for each grid cell according to slope, vegetation, and month of the year to account for vegetation growth and potential ice effects. Since these default variable Manning coefficients led to better performances than the fixed, calibrated values, the former were preserved in the final GEM-Hydro setup. The total daily Lake Erie inflows were used as the objective function to calibrate MESH-SVS, which consists of the Nash-Sutcliffe efficiency criterion. The time-series of observed daily inflows to Lake Erie were estimated based on the total observed streamflow entering the lake, which were then extrapolated using the Area-Ratio-Method (ARM, see [Fry et al. \[2014\]](#)) to account for ungauged areas of the Lake Erie watershed. Inflows from Saint-Claire River were disregarded to ignore the influence of the other upstream Great Lakes.

For MESH-CLASS, the routing was run at the 1 km resolution. Some manual editing of drainage directions was necessary to bring the drainage areas of most sub-basins inline with reported areas (by WSC and USGS). The model was calibrated to the time-series of observed daily streamflow at 35 stations designated to be the most downstream ones. The basin was masked for the watersheds corresponding to those gauges to reduce the calibration run time. Calibrations of HLSS and routing parameters were done independently in two stages. First, selected HLSS parameters for dominant GRUs (broadleaf forest, short grass, long grass, crops and urban) were calibrated (70 in total) to minimize the sum of absolute percent bias ( $\sum |PBIAS| \rightarrow Min$ ) across all stations to ensure the water balance is closed to within acceptable limits. The calibration started with 49 stations (all most downstream) but the number of stations was later reduced because the aggregation led to sub-optimal performance for most stations as the optimizer kept trying to improve the least performing ones similar to what is reported above for VIC-GRU. The calibration budget was initially assigned 5000 iterations but was later increased to 12 000 because some of the parameter combinations crashed and convergence was slow due to the large number of parameters. For this second calibration attempt with 35 stations, the best performing parameter sets from the 49 one were used to speed up the process further.

Then, routing parameters were calibrated using the Nash-Sutcliffe and Kling-Gupta efficiency (KGE) criteria (as alternatives) with a budget of 5000 model evaluation for each. The KGE calibration resulted in slightly better NSE values for most gauges. A total of 25 routing parameters were calibrated. These are roughness coefficients for overland flow for the five dominant GRUs in addition to roughness and baseflow parameters for five river classes. River classes were introduced in the MESH-CLASS setup to improve the representation of spatial variability of channel roughness as small rivers are generally different from large ones, especially with the lack of temporal variability. The classes were assigned based on visual analysis of the histogram of the “bankfull” capacity of river channels (log-transformed to reduce the range) which is direct function of the cumulative drainage area (DA) at each grid cell. The larger the value of DA, the larger the value of bankfull and thus the larger the river. Arbitrary thresholds were applied to the histogram to differentiate the river classes but this can be done more objectively using quantiles as done for the Yukon river basin by [Elshamy et al. \[2020\]](#). Routing parameters could interact with HLSS parameters but the impact was assessed

by comparing the PBIAS values before and after calibrating the routing and differences are negligible.

To perform the calibrations of both MESH configurations, the Ostrich toolkit was used [Matott, 2017], using the Dynamically Dimensioned Search algorithm (DDS) [Tolson and Shoemaker, 2007]. More information on application of MESH-CLASS for the entire Great Lakes can be found in Haghnegahdar et al. [2014] and Xu et al. [2015].

### S.1.12 WATFLOOD

The WATFLOOD model was first created in 1973 and is a partially physically-based, distributed hydrological model which has been used for not only long flood forecasting, but also for long-term hydrological simulation of watersheds for such applications as climate change [Kouwen, 1988]. The model was designed to run using the easily available input variables of temperature and precipitation. These input variables can be derived from station data, weather radar, numerical weather models, or climate change scenarios.

The hydrological processes modelled in WATFLOOD include, but are not limited to, interception, infiltration, evaporation, snow accumulation and ablation, interflow, recharge, baseflow, and overland and channel routing. The most important concept of WATFLOOD is the grouped response unit (GRU) approach which is a conceptual grouping of land surface areas with similar land use that are expected to have similar hydrological response. The runoff response from each unit with an individual GRU is calculated and routed downstream [Cranmer et al., 2001]. River channels are classified allowing for different flow characteristics depending on the nature of the river channel. WATFLOOD computes infiltration using the Philip formula, which represents physical aspects of the infiltration process.

WATFLOOD has been employed by over the Great Lakes basin [Pietroniro et al., 2007] and other basins across Canada [Bomhof et al., 2019; Unduche et al., 2018]. The format of the land surface characteristics database for WATFLOOD was designed to be the same as for the GEM-Hydro model described in the previous section. Thus for this study the same geophysical databases were used for soil texture, land cover, slope, flow direction, drainage area, and channel properties.

The version of the WATFLOOD model used for this study was 9.8 and it was calibrated at a 10 km resolution using all the stations in the calibration list (Section S.2) at the same time. This resolution was chosen to reduce the computation needed for the calibration. The parameters that were calibrated were coefficients related to flow between the different storages (i.e., surface to upper zone to lower zone), snowmelt, evaporation, and river channel roughness. The Nash-Sutcliffe efficiency criterion was used as the objective function for the calibration. The Ostrich toolkit was used for all calibrations [Matott, 2017], using the Dynamically Dimensioned Search algorithm (DDS) [Tolson and Shoemaker, 2007] with a maximum of 500 iterations. Final streamflow was then calculated using the 1 km version of the geophysical database.

## S.2 Information on Gauge Stations for Calibration and Validation

Table S1 lists all gauge stations in the Lake Erie watershed used for this study. The spatial distribution can be found in Fig. 1A and 1B in the main manuscript.

**Table S1.** Gauges used for calibration and spatial validation in this study. The location, drainage area, country, regulation, gauge name, and ID are given as well as if the station is used for objective 1 (low-human impact) and/or objective 2 (most-downstream gauges). In total 46 stations are used for calibration (28 for objective 1 and 31 for objective 2; 13 gauges in both objectives) and 7 stations are used for validation (all stations used for both objectives).

Gauge ID	Gauge Name	Ctry	Lat [deg]	Lon [deg]	Area [km <sup>2</sup> ]	Regu- lation	Obj.		Cal.	Val.
							1	2		
02GA010	NITH RIVER NEAR CANNING	CA	43.1897	-80.4550	1030.0	Natural	x		x	
02GA018	NITH RIVER AT NEW HAM- BURG	CA	43.3772	-80.7108	544.0	Natural	x		x	
02GA038	NITH RIVER ABOVE NITH- BURG	CA	43.4839	-80.8350	326.0	Natural	x		x	
02GA047	SPEED RIVER AT CAMBRIDGE	CA	43.4219	-80.3327	762.0	Natural	x		x	
02GC010	BIG OTTER CREEK AT TILL- SONBURG	CA	42.8573	-80.7236	354.0	Natural	x		x	
02GD004	MIDDLE THAMES RIVER AT THAMESFORD	CA	43.0591	-80.9949	306.0	Natural	x		x	
02GG002	SYDENHAM RIVER NEAR ALVINSTON	CA	42.8308	-81.8517	701.0	Natural	x		x	
02GG006	BEAR CREEK NEAR PETROLIA	CA	42.9058	-82.1191	249.0	Natural	x		x	
04159492	BLACK RIVER NEAR JEDDO MI	US	43.1525	-82.6241	1197.8	Natural	x		x	
04161820	CLINTON RIVER AT STERLING HEIGHTS MI	US	42.6145	-83.0266	802.8	Natural	x		x	
04164000	CLINTON RIVER NEAR FRASER MI	US	42.5773	-82.9513	1143.0	Natural	x		x	
04166100	RIVER ROUGE AT SOUTH- FIELD MI	US	42.4478	-83.2977	224.3	Natural	x		x	
04196800	TYMOCHTEE CREEK AT CRAWFORD OH	US	40.9228	-83.3488	608.1	Natural	x		x	
04197100	HONEY CREEK AT MELMORE OH	US	41.0223	-83.1096	388.2	Natural	x		x	
04207200	TINKERS CREEK AT BEDFORD OH	US	41.3845	-81.5273	222.1	Natural	x		x	
02GB007	FAIRCHILD CREEK NEAR BRANTFORD	CA	43.1474	-80.1546	389.0	Natural	x	x	x	
02GC002	KETTLE CREEK AT ST. THOMAS	CA	42.7777	-81.2140	331.0	Natural	x	x	x	
02GC018	CATFISH CREEK NEAR SPARTA	CA	42.7461	-81.0569	295.0	Natural	x	x	x	
02GE007	MCGREGOR CREEK NEAR CHATHAM	CA	42.3835	-82.0951	204.0	Natural	x	x	x	
02GG003	SYDENHAM RIVER AT FLO- RENCE	CA	42.6506	-82.0084	1150.0	Natural	x	x	x	
02GG009	BEAR CREEK BELOW BRIG- DEN	CA	42.8120	-82.2984	536.0	Natural	x	x	x	
02GG013	BLACK CREEK NEAR BRAD- SHAW	CA	42.7624	-82.2592	213.0	Natural	x	x	x	
04159900	MILL CREEK NEAR AVOCA MI	US	43.0545	-82.7346	438.7	Natural	x	x	x	
04160600	BELLE RIVER AT MEMPHIS MI	US	42.9009	-82.7691	390.9	Natural	x	x	x	
04165500	CLINTON RIVER AT MORA- VIAN DRIVE AT MT. CLEMENS MI	US	42.5959	-82.9088	1892.6	Natural	x	x	x	

*Continued on next page*



Table S1 – Continued from previous page

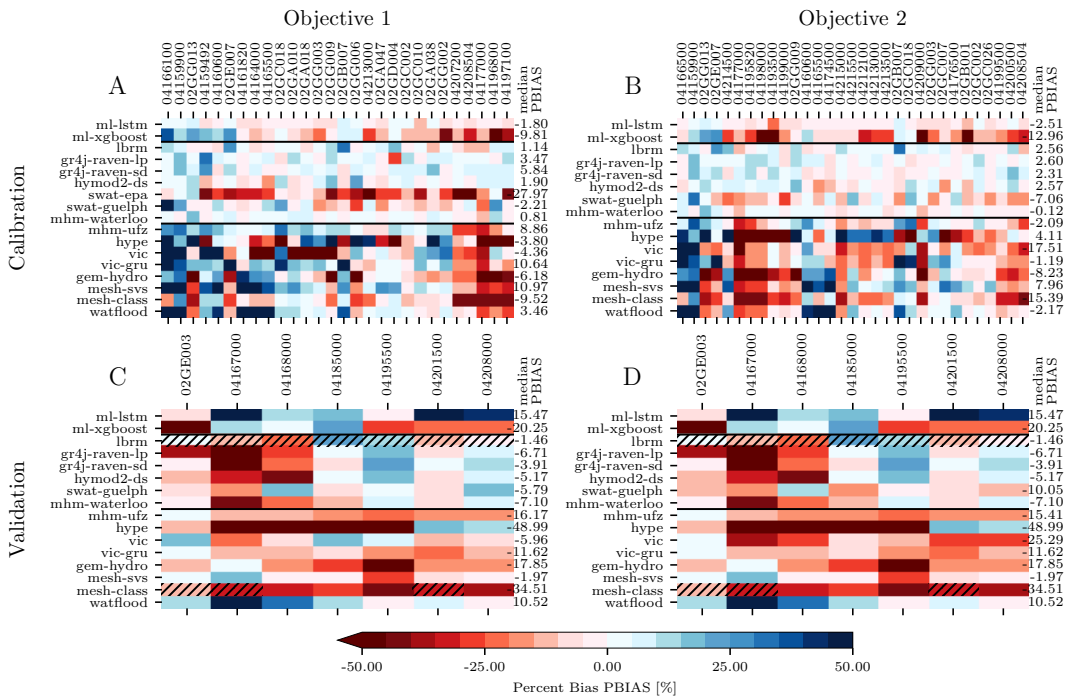
Gauge ID	Gauge Name	Ctry	Lat [deg]	Lon [deg]	Area [km <sup>2</sup> ]	Regu- lation	Obj.		Cal.	Val.
							1	2		
04177000	OTTAWA RIVER AT UNIVERSITY OF TOLEDO TOLEDO OH	US	41.6597	-83.6125	343.8	Natural	x	x	x	
04208504	CUYAHOGA RIVER NEAR NEWBURGH HEIGHTS OH	US	41.4626	-81.6810	2043.6	Natural	x	x	x	
04213000	CONNEAUT CREEK AT CONNEAUT OH	US	41.9270	-80.6040	455.3	Natural	x	x	x	
02GB001	GRAND RIVER AT BRANTFORD	CA	43.1327	-80.2673	5200.0	Regulated		x	x	
02GC007	BIG CREEK NEAR WALSINGHAM	CA	42.6856	-80.5385	567.0	Regulated		x	x	
02GC026	BIG OTTER CREEK NEAR CALTON	CA	42.7107	-80.8408	665.0	Regulated		x	x	
04166500	RIVER ROUGE AT DETROIT MI	US	42.3723	-83.2555	476.0	Regulated		x	x	
04174500	HURON RIVER AT ANN ARBOR MI	US	42.2861	-83.7333	1928.2	Regulated		x	x	
04176500	RIVER RAISIN NEAR MONROE MI	US	41.9606	-83.5310	2686.0	Regulated		x	x	
04193500	MAUMEE RIVER AT WATERVILLE OH	US	41.5001	-83.7127	16409.4	Regulated		x	x	
04195820	PORTAGE RIVER NEAR ELMORE OH	US	41.4912	-83.2246	1266.2	Regulated		x	x	
04198000	SANDUSKY RIVER NEAR FREMONT OH	US	41.3078	-83.1588	3243.8	Regulated		x	x	
04199000	HURON RIVER AT MILAN OH	US	41.3017	-82.6068	947.4	Regulated		x	x	
04199500	VERMILION R NR VERMILION OH	US	41.3820	-82.3168	672.4	Regulated		x	x	
04200500	BLACK RIVER AT ELYRIA OH	US	41.3803	-82.1046	1026.9	Regulated		x	x	
04209000	CHAGRIN RIVER AT WILLOUGHBY OH	US	41.6309	-81.4034	637.4	Regulated		x	x	
04212100	GRAND RIVER NEAR PAINESVILLE OH	US	41.7189	-81.2279	1784.9	Regulated		x	x	
04213500	CATTARAUGUS CR AT GOWANDA NY	US	42.4640	-78.9350	1128.8	Regulated		x	x	
04214500	BUFFALO CREEK AT GARDENVILLE NY	US	42.8548	-78.7550	368.4	Regulated		x	x	
04215000	CAYUGA CREEK NR LANCASTER NY	US	42.8901	-78.6450	248.0	Regulated		x	x	
04215500	CAZENOVIA CREEK AT EBENEZER NY	US	42.8298	-78.7750	350.7	Regulated		x	x	
02GE003	THAMES RIVER AT THAMESVILLE	CA	42.5449	-81.9673	4370.0	Regulated	x	x		x
04167000	MIDDLE RIVER ROUGE NEAR GARDEN CITY MI	US	42.3481	-83.3116	229.5	Regulated	x	x		x
04168000	LOWER RIVER ROUGE AT INKSTER MI	US	42.3006	-83.3002	219.2	Regulated	x	x		x
04185000	Tiffin River at Stryker OH	US	41.5045	-84.4297	1061.9	Natural	x	x		x
04195500	PORTAGE R AT WOODVILLE OH	US	41.4495	-83.3613	1108.5	Regulated	x	x		x
04201500	ROCKY R NR BERE A OH	US	41.4075	-81.8826	691.5	Regulated	x	x		x
04208000	CUYAHOGA R AT INDEPENDENCE OH	US	41.3953	-81.6298	1831.1	Regulated	x	x		x

### S.3 Model Performance per Model and Objective

Table S2 summarizes the model performances as median Nash-Sutcliffe Efficiency over all gauges for each model. The results are shown for both objectives and for a default model setup and the final model setup after calibration. The calibration results are shown as a barchart in Fig. 3A in the main manuscript. The results per gauge station are shown in Figs 2A and 2B of the main manuscript. The validation results are shown as a barchart in Fig. 3B in the main manuscript. The results per gauge station are shown in Figs 2C and 2D of the main manuscript.

### S.4 Model Performance regarding other metrics

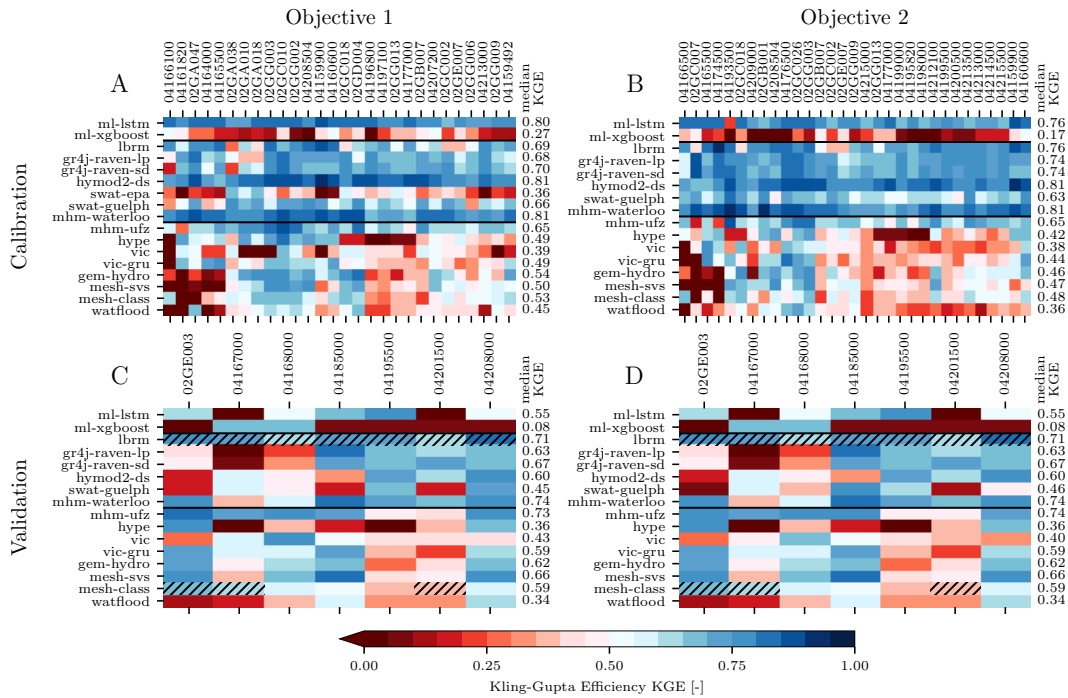
All models are primarily analysed regarding the Nash-Sutcliffe efficiency. However, other metrics were tested but are not presented in the main manuscript as they did not yield any further insight. The auxiliary metrics tested are percent bias (PBIAS; Figure S1), Kling-Gupta Efficiency (KGE; Figure S2), and the three components of the Kling-Gupta Efficiency, i.e. relative variability (KGE\_a; Figure S3), bias (KGE\_b; Figure S4), and Pearson Correlation Coefficient (KGE\_r; Figure S5).



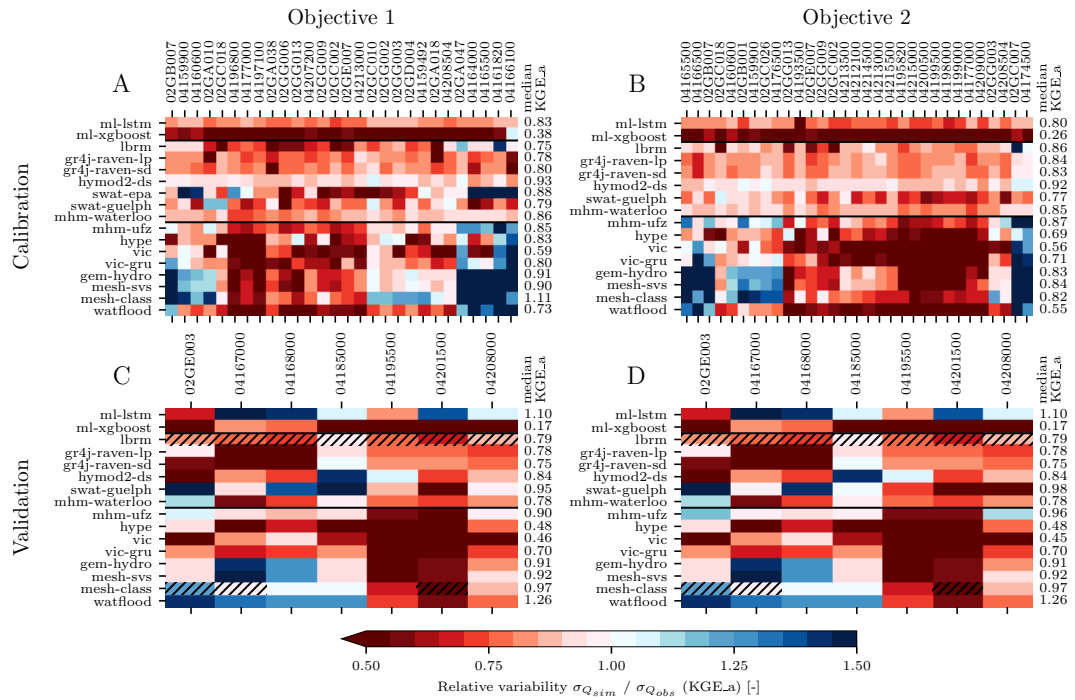
**Fig. S1.** The performance of the participating models in calibration and validation (A,B and C,D respectively) for each gauging station of objective 1 and 2. The colored tiles indicated the **percent bias (PBIAS)** per gauge and model while the median PBIAS over all gauging stations is displayed to the right. The black horizontal lines separate i) Machine Learning models from ii) models that are calibrated at each individual streamflow gauge from iii) models that are calibrated over the entire domain calibrating all streamflow gauges simultaneously. The hatched tiles (validation only) mark gauging stations that have informed the calibration of the corresponding models.

**Table S2.** Model performance as median Nash-Sutcliffe efficiency (NSE) over all gauges for each model in either pre-calibration (Def.) mode or after calibration (Cal.) or spatial validation (Val.) for both objectives considered in this study. Major improvements of all models were achieved by the automatic calibration. All models calibrated at individual gauges (local calibration; e.g., GR4J, LBRM, mHM-Waterloo) outperform the models that are calibrated at all N gauging stations together (global calibration; e.g., mHM-UFZ, GEM-Hydro). The latter leads to one final model setup (one parameter set) while the individual calibrations lead to N model setups (N parameter sets) and are hence more difficult to be transferred to other domains. Best performing models in each of the three groups are highlighted with bold font. LBRM and MESH-CLASS used some of the validation stations already in calibration. These results can therefore not be considered proper validation results and are indicated with an asterisk (\*).

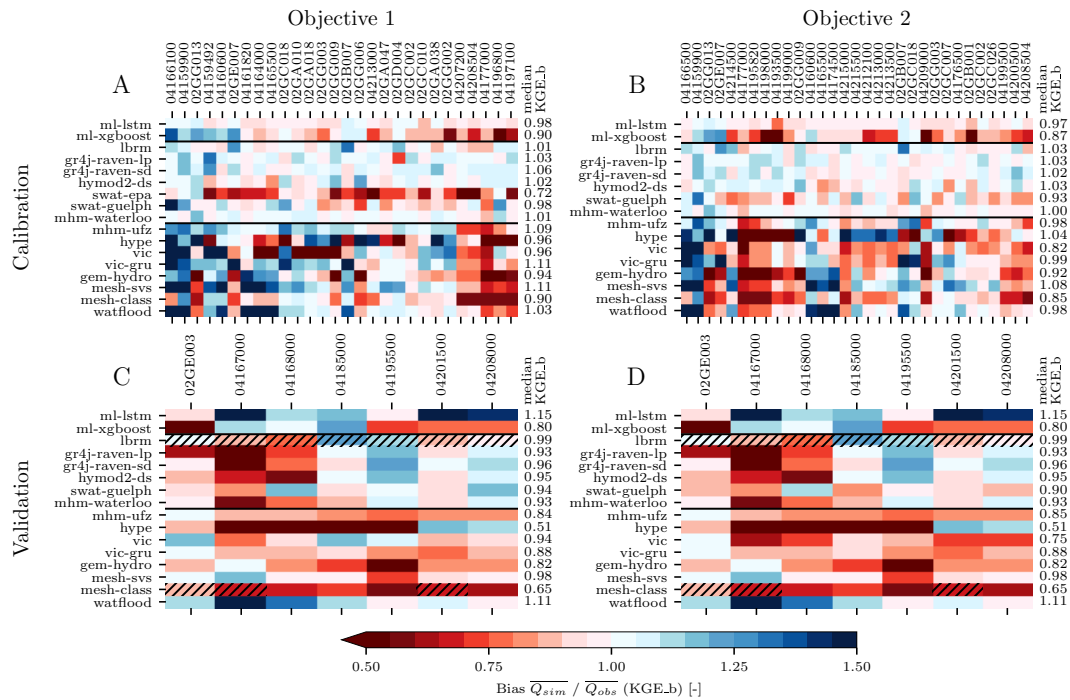
Model	Modeller	<i>Objective</i> <i>Phase</i>	Objective 1: low-human impact			Objective 2: most downstream		
			Def.	Cal.	Val.	Def.	Cal.	Val.
<b>ML-LSTM</b>	Gauch & Lin	Global calib.	n/a	<b>0.73</b>	<b>0.41</b>	n/a	<b>0.54</b>	<b>0.41</b>
ML-XGBoost	Gauch & Lin	Global calib.	n/a	0.37	0.17	n/a	0.22	0.17
LBRM	Fry & Bradley	Local calib.	0.41	0.66	0.70*	0.57	0.72	0.70*
GR4J-Raven-lp	Shen & Tolson	Local calib.	0.08	0.63	0.50	0.07	0.67	0.50
GR4J-Raven-sd	Shen & Tolson	Local calib.	0.07	0.64	0.44	0.06	0.67	0.44
HYMOD2-DS	Roy & Wi	Local calib.	-1.52	0.74	0.59	-0.19	0.73	0.59
SWAT-EPA	Ni & Yuan	Local calib.	-0.17	0.19	n/a	-0.22	n/a	n/a
SWAT-Guelph	Shrestha & Daggupati	Local calib.	-0.19	0.55	0.26	-0.39	0.59	0.23
<b>mHM-Waterloo</b>	McLeod, Kumar & Basu	Local calib.	0.35	<b>0.76</b>	<b>0.68</b>	0.37	<b>0.78</b>	<b>0.68</b>
<b>mHM-UFZ</b>	Rakovec, Samaniego & Attinger	Global calib.	0.24	<b>0.66</b>	<b>0.64</b>	0.27	<b>0.67</b>	<b>0.60</b>
HYPE	Awoye & Stadnyk	Global calib.	0.08	0.52	0.41	0.08	0.48	0.41
VIC	Shen & Tolson	Global calib.	0.22	0.41	0.53	0.37	0.43	0.51
VIC-GRU	Gharari	Global calib.	-0.11	0.42	0.51	0.10	0.43	0.51
GEM-Hydro	Gaborit	Global calib.	0.38	0.51	0.54	0.36	0.44	0.54
MESH-SVS	Gaborit & Princz	Global calib.	0.33	0.44	0.58	0.33	0.45	0.58
MESH-CLASS	Haghn., Elshamy & Princz	Global calib.	-0.05	0.34	0.51*	0.16	0.40	0.51*
WATFLOOD	Seglenieks & Temgoua	Global calib.	-1.09	0.33	0.05	-0.50	0.32	0.05



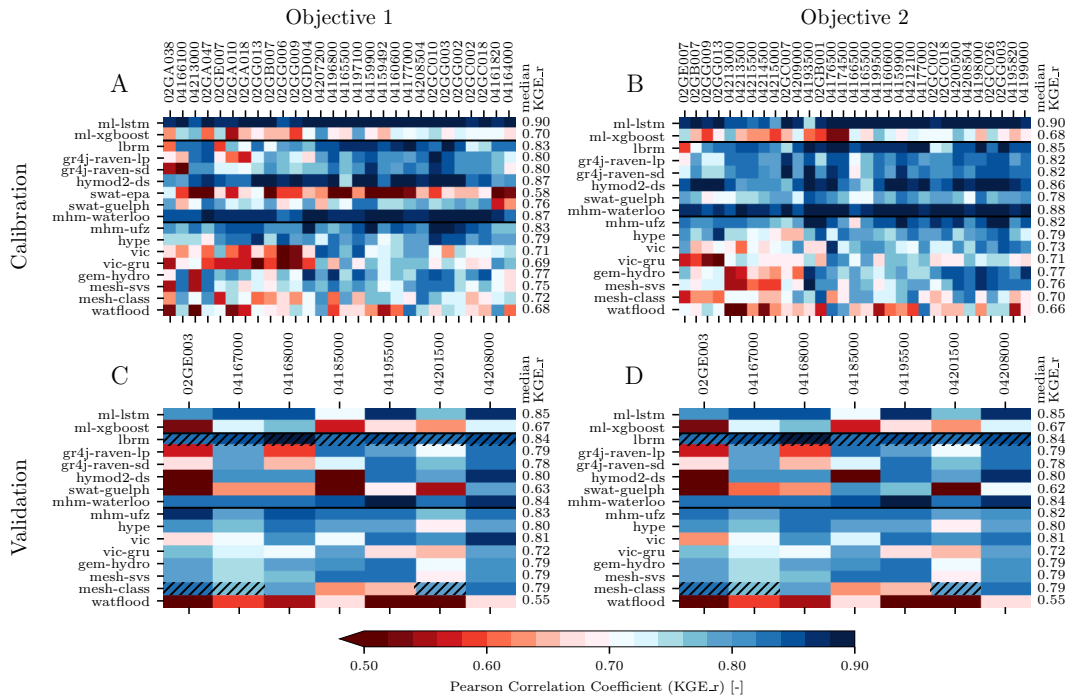
**Fig. S2.** The performance of the participating models in calibration and validation (A,B and C,D respectively) for each gauging station of objective 1 and 2. The colored tiles indicated the **Kling-Gupta Efficiency (KGE)** per gauge and model while the median KGE over all gauging stations is displayed to the right. The black horizontal lines separate i) Machine Learning models from ii) models that are calibrated at each individual streamflow gauge from iii) models that are calibrated over the entire domain calibrating all streamflow gauges simultaneously. The hatched tiles (validation only) mark gauging stations that have informed the calibration of the corresponding models.



**Fig. S3.** The performance of the participating models in calibration and validation (A,B and C,D respectively) for each gauging station of objective 1 and 2. The colored tiles indicated the **relative variability (KGE\_a)** per gauge and model while the median KGE\_a over all gauging stations is displayed to the right. The black horizontal lines separate i) Machine Learning models from ii) models that are calibrated at each individual streamflow gauge from iii) models that are calibrated over the entire domain calibrating all streamflow gauges simultaneously. The hatched tiles (validation only) mark gauging stations that have informed the calibration of the corresponding models.



**Fig. S4.** The performance of the participating models in calibration and validation (A,B and C,D respectively) for each gauging station of objective 1 and 2. The colored tiles indicated the **bias** ( $KGE_b$ ) per gauge and model while the median  $KGE_b$  over all gauging stations is displayed to the right. The black horizontal lines separate i) Machine Learning models from ii) models that are calibrated at each individual streamflow gauge from iii) models that are calibrated over the entire domain calibrating all streamflow gauges simultaneously. The hatched tiles (validation only) mark gauging stations that have informed the calibration of the corresponding models.



**Fig. S5.** The performance of the participating models in calibration and validation (A,B and C,D respectively) for each gauging station of objective 1 and 2. The colored tiles indicated the **Pearson Correlation Coefficient (KGE<sub>r</sub>)** per gauge and model while the median KGE<sub>r</sub> over all gauging stations is displayed to the right. The black horizontal lines separate i) Machine Learning models from ii) models that are calibrated at each individual streamflow gauge from iii) models that are calibrated over the entire domain calibrating all streamflow gauges simultaneously. The hatched tiles (validation only) mark gauging stations that have informed the calibration of the corresponding models.

## References

- Abbaspour, K. C., Johnson, C. A., and van Genuchten, M. T. (2004). "Estimating Uncertain Flow and Transport Parameters Using a Sequential Uncertainty Fitting Procedure." *Vadose Zone Journal*, 3(4), 1340–1352.
- Alavi, N., Bélair, S., Fortin, V., Zhang, S., Husain, S. Z., Carrera, M. L., and Abrahamowicz, M. (2016). "Warm Season Evaluation of Soil Moisture Prediction in the Soil, Vegetation, and Snow (SVS) Scheme." *Journal of Hydrometeorology*, 17(8), 2315–2332.
- Arheimer, B., Pimentel, R., Isberg, K., Crochemore, L., Andersson, J. C. M., Hasan, A., and Pineda, L. (2020). "Global catchment modelling using World-Wide HYPE (WWH), open data, and stepwise parameter estimation." *Hydrology and Earth System Sciences*, 24(2), 535–559.
- Arnold, J. G., Allen, P. M., and Bernhardt, G. (1993). "A comprehensive surface-groundwater flow model." *Journal of hydrology*, 142(1-4), 47–69.
- Arnold, J. G., Srinivasan, R., Mutiah, R. S., and Williams, J. R. (1998). "Large area hydrologic modeling and assessment part i: Model development1." *JAWRA Journal of the American Water Resources Association*, 34(1), 73–89.
- Arsenault, R., Brissette, F., and Martel, J.-L. (2018). "The hazards of split-sample validation in hydrological model calibration." *Journal of Hydrology*, 566, 346–362.
- Bajracharya, A., Awoye, H., Stadnyk, T., and Asadzadeh, M. (2020). "Time Variant Sensitivity Analysis of Hydrological Model Parameters in a Cold Region Using Flow Signatures." *Water*, 12(4), 961–24.
- Bernier, N. B., Bélair, S., Bilodeau, B., and Tong, L. (2011). "Near-Surface and Land Surface Forecast System of the Vancouver 2010 Winter Olympic and Paralympic Games." *Journal of Hydrometeorology*, 12(4), 508–530.
- Bomhof, J., Tolson, B. A., and Kouwen, N. (2019). "Comparing single and multi-objective hydrologic model calibration considering reservoir inflow and streamflow observations." *Canadian Water Resources Journal / Revue canadienne des ressources hydriques*, 44(4), 319–336.
- Boyle, D. P., Gupta, H. V., and Sorooshian, S. (2000). "Toward improved calibration of hydrologic models: Combining the strengths of manual and automatic methods." *Water Resources Research*, 36(12), 3663–3674.
- Braak, C. J. F. T. (2006). "A Markov Chain Monte Carlo version of the genetic algorithm Differential Evolution: easy Bayesian computing for real parameter spaces." *Statistics and Computing*, 16(3), 239–249.
- Chen, T. and Guestrin, C. (2016). "XGBoost: A scalable tree boosting system." *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '16*, ACM, 785–794.
- Craig, J. R., Brown, G., Chlumsky, R., Jenkinson, R. W., Jost, G., Lee, K., Mai, J., Serrer, M., Sgro, N., Shafii, M., Snowdon, A. P., and Tolson, B. A. (2020). "Flexible watershed simulation with the Raven hydrological modelling framework." *Environmental Modelling & Software*, 129, 104728.
- Cranmer, A., Kouwen, N., and Mousavi, S.-F. (2001). "Proving watflood: Modelling the nonlinearities of hydrologic response to storm intensities." *Canadian Journal of Civil Engineering*, 28, 837–855.
- Croley II, T. E. (1983). "Great lake basins (u.s.a.-canada) runoff modeling." *Journal of Hydrology*, 64(1), 135–158.
- Dingman, S. L. (2015). *Physical hydrology*. Waveland Press, Long Grove, Illinois, third edition. edition.
- Dolan, D. M. and Chapra, S. C. (2012). "Great lakes total phosphorus revisited: 1. loading analysis and update (1994–2008)." *Journal of Great Lakes Research*, 38(4), 730–740.
- Duan, Q. Y., Sorooshian, S., and Gupta, H. V. (1992). "Effective and efficient global optimization for conceptual rainfall-runoff models." *Water Resources Research*, 28(4), 1015–1031.



- Elshamy, M., Loukili, Y., Princz, D., Richard, D., Tesemma, Z., and Pomeroy, J. W. (2020). “Yukon River Basin Streamflow Forecasting System.” *Report submitted to Government of Yukon*, n/a.
- Fry, L. M., Gronewold, A. D., Fortin, V., Buan, S., Clites, A. H., Luukkonen, C., Holtschlag, D., Diamond, L., Hunter, T., Seglenieks, F., Durnford, D., Dimitrijevic, M., Subich, C., Klyszejko, E., Kea, K., and Restrepo, P. (2014). “The Great Lakes Runoff Intercomparison Project Phase 1: Lake Michigan (GRIP-M).” *Journal of Hydrology*, 519, 3448–3465.
- Gaborit, É., Fortin, V., Xu, X., Seglenieks, F., Tolson, B., Fry, L. M., Hunter, T., Anctil, F., and Gronewold, A. D. (2017). “A hydrological prediction system based on the SVS land-surface scheme: efficient calibration of GEM-Hydro for streamflow simulation over the Lake Ontario basin.” *Hydrology and Earth System Sciences*, 21(9), 4825–4839.
- Gaborit, É., Ricard, S., Lachance-Cloutier, S., Anctil, F., and Turcotte, R. (2015). “Comparing global and local calibration schemes from a differential split-sample test perspective.” *Canadian Journal of Earth Sciences*, 52(11), 990–999.
- Gao, H., Tang, Q., Shi, X., Zhu, C., Bohn, T., Su, F., Pan, M., Sheffield, J., Lettenmaier, D., and Wood, E. (2010). *Water Budget Record from Variable Infiltration Capacity (VIC) Model*. UNSPECIFIED, 120–173.
- Gauch, M., Mai, J., and Lin, J. (2019). “The proper care and feeding of CAMELS: How limited training data affects streamflow prediction.
- Gharari, S., Clark, M. P., Mizukami, N., Knoben, W. J. M., Wong, J. S., and Pietroniro, A. (2020). “Flexible vector-based spatial configurations in land models.” *Hydrology and Earth System Sciences Discussions*, 1–40.
- Gharari, S., Clark, M. P., Mizukami, N., Wong, J. S., Pietroniro, A., and Wheeler, H. S. (2019). “Improving the representation of subsurface water movement in land models.” *Journal of Hydrometeorology*, 20(12), 2401–2418.
- Gronewold, A. D., Hunter, T., Allison, J., Fry, L. M., Kompoltowicz, K. A., Bolinger, R. A., and Pei, L. (2017). “Project Documentation Report for Great Lakes seasonal and inter-annual water supply forecasting improvements project Phase I: Research and Development.” *Report no.*, NOAA-GLERL, Ann Arbor, MI, <<https://www.glerl.noaa.gov/pubs/fulltext/2018/20180020.pdf>> (October).
- Haghnegahdar, A. and Razavi, S. (2017). “Insights into sensitivity analysis of Earth and environmental systems models: On the impact of parameter perturbation scale.” *Environmental Modelling & Software*, 95, 115–131.
- Haghnegahdar, A., Tolson, B. A., Davison, B., Seglenieks, F. R., Klyszejko, E., Soulis, E. D., Fortin, V., and Matott, L. S. (2014). “Calibrating Environment Canada’s MESH Modelling System over the Great Lakes Basin.” *ATMOSPHERE-OCEAN*, 52(4), 281–293.
- Hamon, W. R. (1961). “Estimating Potential Evapotranspiration.” *Journal of the Hydraulics Division*, 87, 107–120.
- Hochreiter, S. and Schmidhuber, J. (1997). “Long short-term memory.” *Neural computation*, 9(8), 1735–1780.
- Hunter, T. S., Clites, A. H., Campbell, K. B., and Gronewold, A. D. (2015). “Development and application of a North American Great Lakes hydrometeorological database — Part I: Precipitation, evaporation, runoff, and air temperature.” *Journal of Great Lakes Research*, 41(1), 65–77.
- Husain, S. Z., Alavi, N., Bélair, S., Carrera, M., Zhang, S., Fortin, V., Abrahamowicz, M., and Gauthier, N. (2016). “The Multibudget Soil, Vegetation, and Snow (SVS) Scheme for Land Surface Parameterization: Offline Warm Season Evaluation.” *Journal of Hydrometeorology*, 17(8), 2293–2313.
- Kouwen, N. (1988). “Watflood: a micro-computer based flood forecasting system based on real-time weather radar.” *Canadian Water Resources Journal / Revue canadienne des ressources hydriques*, 13(1), 62–77.
- Kouwen, N. (2010). *WATFLOOD/ WATROUTE Hydrological Model Routing & Flow Forecasting System*. Department of Civil Engineering, University of Waterloo, Waterloo, ON, Department of Civil Engineering, University of Waterloo, Waterloo, ON.

- Kumar, R., Samaniego, L., and Attinger, S. (2013). "Implications of distributed hydrologic model parameterization on water fluxes at multiple scales and locations." *Water Resources Research*, 49(1), 360–379.
- Lehner, B., Verdin, K., and Jarvis, A. (2008). "New Global Hydrography Derived From Spaceborne Elevation Data." *Eos, Transactions American Geophysical Union*, 89, 93–94  
Data is available at [www.hydrosheds.org](http://www.hydrosheds.org).
- Liang, X. (2003). "A new parameterization for surface and groundwater interactions and its impact on water budgets with the variable infiltration capacity (VIC) land surface model." *Journal of Geophysical Research*, 108(D16), 1989–17.
- Liang, X., Lettenmaier, D. P., Wood, E. F., and Burges, S. J. (1994). "A simple hydrologically based model of land surface water and energy fluxes for general circulation models." *Journal of Geophysical Research: Atmospheres*, 99(D7), 14415–14428.
- Lindström, G., Pers, C., Rosberg, J., Strömqvist, J., and Arheimer, B. (2010). "Development and testing of the HYPE (Hydrological Predictions for the Environment) water quality model for different spatial scales." *Hydrology Research*, 41(3-4), 295–319.
- Livneh, B., Rosenberg, E. A., Lin, C., Nijssen, B., Mishra, V., Andreadis, K. M., Maurer, E. P., and Lettenmaier, D. P. (2013). "A long-term hydrologically based dataset of land surface fluxes and states for the conterminous United States: Update and extensions." *Journal of Climate*, 26(23), 9384–9392.
- Martinez, J. (1975). "Snowmelt-runoff model for stream flow forecasts ." *Nordic Hydrology*, 6(3), 145–154.
- Matott, L. S. (2017). *OSTRICH – An Optimization Software Toolkit for Research Involving Computational Heuristics Documentation and User's Guide*. State University of New York at Buffalo Center for Computational Research, 17.12.19 edition.
- Mizukami, N., Clark, M. P., Sampson, K., Nijssen, B., Mao, Y., McMillan, H., Viger, R. J., Markstrom, S. L., Hay, L. E., Woods, R., Arnold, J. R., and Brekke, L. D. (2016). "mizuroute version 1: a river network routing tool for a continental domain water resources applications." *Geoscientific Model Development*, 9(6), 2223–2238.
- Moore, R. J. (1985). "The probability-distributed principle and runoff production at point and basin scales." *Hydrological Sciences Journal*– . . . , 30(2), 273–297.
- Nash, J. E. and Sutcliffe, J. V. (1970). "River flow forecasting through conceptual models: Part I - A discussion of principles." *Journal of Hydrology*, 10, 282–290.
- Neitsch, S. L., Arnold, J. G., Kiniry, J. R., and Williams, J. R. (2011). "Soil and water assessment tool theoretical documentation version 2009." *Report no.*, Texas Water Resources Institute.
- of Canada, G. (2015). "Government of Canada." *Government of Canada*, <<https://open.canada.ca/data/en/dataset/9e1efe92-e5a3-4f70-b313-68fb1283eadf>> (Mar).
- Pechlivanidis, I. G. and Arheimer, B. (2015). "Large-scale hydrological modelling by using modified PUB recommendations: the India-HYPE case." *Hydrology and Earth System Sciences*, 19(11), 4559–4579.
- Pechlivanidis, I. G., Bosshard, T., Spångmyr, H., Lindström, G., Gustafsson, D., and Arheimer, B. (2014). "Uncertainty in the Swedish operational hydrological forecasting systems." *Vulnerability, Uncertainty, and Risk: Quantification, Mitigation, and Management*, American Society of Civil Engineers, 253–262.
- Perrin, C., Michel, C., and Andréassian, V. (2003). "Improvement of a parsimonious model for streamflow simulation." *Journal of Hydrology*, 279, 275–289.
- Pietroniro, A., Fortin, V., Kouwen, N., Neal, C., Turcotte, R., Davison, B., Verseghy, D., Soulis, E. D., Caldwell, R., Evora, N., and Pellerin, P. (2007). "Development of the MESH modelling system for hydrological ensemble forecasting of the Laurentian Great Lakes at the regional scale." *Hydrology and Earth System Sciences*, 11(4), 1279–1294.
- Rakovec, O., Kumar, R., Mai, J., Cuntz, M., Thober, S., Zink, M., Attinger, S., Schäfer, D., Schrön, M., and Samaniego, L. (2016). "Multiscale and multivariate evaluation of water fluxes and states over European river basins." *J. Hydrometeorol.*, 17, 287–307.

- Rakovec, O., Mizukami, N., Kumar, R., Newman, A. J., Thober, S., Wood, A. W., Clark, M. P., and Samaniego, L. (2019). “Diagnostic Evaluation of Large-Domain Hydrologic Models Calibrated Across the Contiguous United States.” *Journal of Geophysical Research: Atmospheres*, 124(24), 13991–14007.
- Reed, S., Koren, V., Smith, M., Zhang, Z., Moreda, F., and Seo, D. J. (2004). “Overall distributed model intercomparison project results.” *Journal of Hydrology*, 298(1-4), 27–60.
- Roy, T., Gupta, H. V., Serrat-Capdevila, A., and Valdes, J. B. (2017). “Using satellite-based evapotranspiration estimates to improve the structure of a simple conceptual rainfall&ndash;runoff model.” *Hydrology and Earth System Sciences*, 21(2), 879–896.
- Samaniego, L., Kumar, R., and Attinger, S. (2010). “Multiscale parameter regionalization of a grid-based hydrologic model at the mesoscale.” *Water Resources Research*, 46(5), W05523.
- Samaniego, L., Kumar, R., Thober, S., Rakovec, O., Zink, M., Wanders, N., Eisner, S., Müller Schmied, H., Sutanudjaja, E. H., Warrach-Sagi, K., and Attinger, S. (2017). “Toward seamless hydrologic predictions across spatial scales.” *Hydrol. Earth Syst. Sci.*, 21(9), 4323–4346.
- Samaniego, L., Thober, S., Wanders, N., Pan, M., Rakovec, O., Sheffield, J., Wood, E. F., Prudhomme, C., Rees, G., Houghton-Carr, H., Fry, M., Smith, K., Watts, G., Hisdal, H., Estrela, T., Buontempo, C., Marx, A., and Kumar, R. (2020). “Hydrological Forecasts and Projections for Improved Decision-Making in the Water Sector in Europe.” *Bulletin of the American Meteorological Society*, 100(12), 2451–2472.
- Seiller, G., Roy, R., and Anctil, F. (2017). “Influence of three common calibration metrics on the diagnosis of climate change impacts on water resources.” *Journal of Hydrology*, 547, 280–295.
- Shangguan, W., Dai, Y., Duan, Q., Liu, B., and Yuan, H. (2014). “A global soil data set for earth system modeling.” *Journal of Advances in Modeling Earth Systems*, 6(1), 249–263.
- Shangguan, W., Hengl, T., Mendes de Jesus, J., Yuan, H., and Dai, Y. (2017). “Mapping the global depth to bedrock for land surface modeling.” *Journal of Advances in Modeling Earth Systems*, 9(1), 65–88.
- Soulis, E. D., Craig, J. R., Fortin, V., and Liu, G. (2011). “A simple expression for the bulk field capacity of a sloping soil horizon.” *Hydrological Processes*, 25(1), 112–116.
- Stadnyk, T., MacDonald, M., Tefs, A., Awoye, O., Déry, S., Gustafsson, D., Isberg, K., and Arheimer, B. (2020). “Hydrological modelling of the freshwater discharge into Hudson Bay using HYPE..” *Elementa*, under review.
- Strömqvist, J., DAHNE, J., DONNELLY, C., Lindström, G., Rosberg, J., Pers, C., YANG, W., and Arheimer, B. (2009). “Using recently developed global data sets for hydrological predictions.” *Symposium HS. at the Joint IAHS IAH Convention, Hyderabad, India, September*, IAHS Publ, 121–127 (September).
- Survey, U. G. (2014). “Multi-resolution land characteristics (mrlc) consortium.” *Multi-Resolution Land Characteristics (MRLC) Consortium*, <<https://www.mrlc.gov/data/nlcd-2011-land-cover-conus-0>> (Mar).
- Thober, S., Cuntz, M., Kelbling, M., Kumar, R., Mai, J., and Samaniego, L. (2019). “The multiscale routing model mrm v1.0: simple river routing at resolutions from 1 to 50 km.” *Geoscientific Model Development*, 12(6), 2501–2521.
- Tolson, B. A. and Shoemaker, C. A. (2007). “Dynamically dimensioned search algorithm for computationally efficient watershed model calibration.” *Water Resources Research*, 43(1), W01413.
- Unduche, F., Tolossa, H., Senbeta, D., and Zhu, E. (2018). “Evaluation of four hydrological models for operational flood forecasting in a canadian prairie watershed.” *Hydrological Sciences Journal*, 63(8), 1133–1149.
- USDA-SCS (1986). “Urban hydrology for small watersheds.” *Report No. TR-55*, US Department of Agriculture-Soil Conservation Service (USDA-SCS), <[https://www.nrcs.usda.gov/Internet/FSE\\_DOCUMENTS/stelprdb1044171.pdf](https://www.nrcs.usda.gov/Internet/FSE_DOCUMENTS/stelprdb1044171.pdf)>.
- Valéry, A., Andréassian, V., and Perrin, C. (2014). ““As simple as possible but not simpler”: What is useful in a temperature-based snow-accounting routine? Part 1 - Comparison of six

- snow accounting routines on 380 catchments.” *Journal of Hydrology*, 517(C), 1166–1175.
- Verseghy, D. L. (2000). “The Canadian land surface scheme (CLASS): Its history and future.” *Atmosphere-Ocean*, 38(1), 1–13.
- Wen, L., Lin, C. A., Wu, Z., Lu, G., Pomeroy, J., and Zhu, Y. (2011). “Reconstructing sixty year (1950-2009) daily soil moisture over the Canadian Prairies using the Variable Infiltration Capacity model.” *Canadian Water Resources Journal*, 36(1), 83–102.
- Wi, S., Yang, Y. C. E., Steinschneider, S., Khalil, A., and Brown, C. M. (2015). “Calibration approaches for distributed hydrologic models in poorly gaged basins: implication for streamflow projections under climate change.” *Hydrology and Earth System Sciences*, 19(2), 857–876.
- Williams, J. R. (1969). “Flood routing with variable travel time or variable storage coefficients.” *Transactions of the ASAE*, 12(1), 0100–0103.
- Wright, D. P., Thyer, M., Westra, S., and McInerney, D. (2018). “A hybrid framework for quantifying the influence of data in hydrological model calibration.” *Journal of Hydrology*, 561, 211–222.
- Xie, Z., Yuan, F., Duan, Q., Zheng, J., Liang, M., and Chen, F. (2007). “Regional Parameter Estimation of the VIC Land Surface Model: Methodology and Application to River Basins in China.” *Journal of Hydrometeorology*, 8(3), 447–468.
- Xu, X., Tolson, B. A., Li, J., Staebler, R. M., Seglenieks, F., Haghnegahdar, A., and Davison, B. (2015). “Assimilation of SMOS soil moisture over the Great Lakes basin.” *Remote Sensing of Environment*, 169, 163–175.
- Yoon, J.-H. and Shoemaker, C. A. (2001). “Improved real-coded ga for groundwater bioremediation.” *Journal of Computing in Civil Engineering*, 15(3), 224–231.
- Zhang, B., Shrestha, N. K., Daggupati, P., Rudra, R., Shukla, R., Kaur, B., and Hou, J. (2018). “Quantifying the impacts of climate change on streamflow dynamics of two major rivers of the northern lake erie basin in canada.” *Sustainability*, 10(8).
- Zink, M., Samaniego, L., Kumar, R., Thober, S., Mai, J., Schäfer, D., and Marx, A. (2016). “The German drought monitor.” *Environmental Research Letters*, 11(7), 074002.