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

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., $(\text{temp} > 0) \wedge (\text{precip} > 0)$) 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^2 to 16 806 km^2.

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) 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`.

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; 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 $\times$ 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 perform 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 ECCC’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 additional 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^2]	Regulation	Obj. 1	Obj. 2	Cal.	Val.
02GA010	NITH RIVER NEAR CANNING	CA	43.1897	-80.4550	1030.0	Natural	x		x	
02GA018	NITH RIVER AT NEW HAMBURG	CA	43.3772	-80.7108	544.0	Natural	x		x	
02GA038	NITH RIVER ABOVE NITHBURG	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 TILLSONBURG	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 SOUTHFIELD 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 FLORENCE	CA	42.6506	-82.0084	1150.0	Natural	x	x	x	
02GG009	BEAR CREEK BELOW BRIGDEN	CA	42.8120	-82.2984	536.0	Natural	x	x	x	
02GG013	BLACK CREEK NEAR BRADSHAW	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 MORAVIAN 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^2]	Regulation	Obj. 1	Obj. 2	Cal.	Val.
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 BEREA 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 Fig.s 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 Fig.s 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).

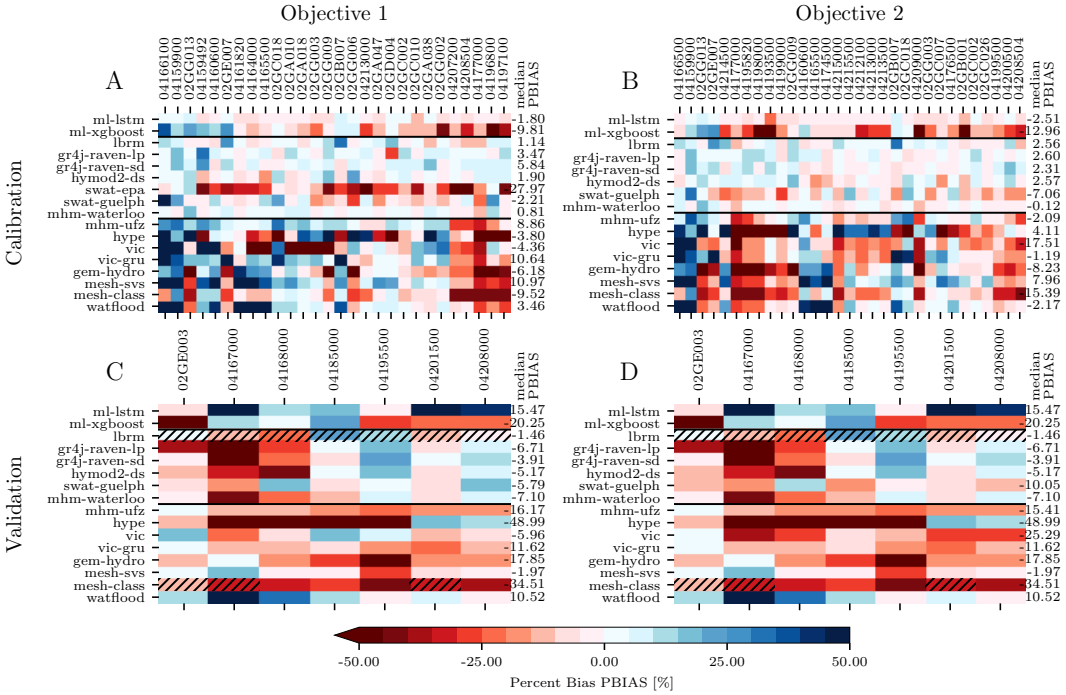

Fig. S1. The performance of the participating models in calibration and validation (A,B and C,D respectively) for each gauging station of objective 1 and 2. The colored tiles indicated the **percent bias (PBIAS)** per gauge and model while the median PBIAS over all gauging stations is displayed to the 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 (*).

		Objective	Objective 1: low-human impact			Objective 2: most downstream		
Model	Modeller	*Phase*	Def.	Cal.	Val.	Def.	Cal.	Val.
ML-LSTM	Gauch & Lin	Global calib.	n/a	**0.73**	**0.41**	n/a	**0.54**	**0.41**
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
mHM-Waterloo	McLeod, Kumar & Basu	Local calib.	0.35	**0.76**	**0.68**	0.37	**0.78**	**0.68**
mHM-UFZ	Rakovec, Samaniego & Attinger	Global calib.	0.24	**0.66**	**0.64**	0.27	**0.67**	**0.60**
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

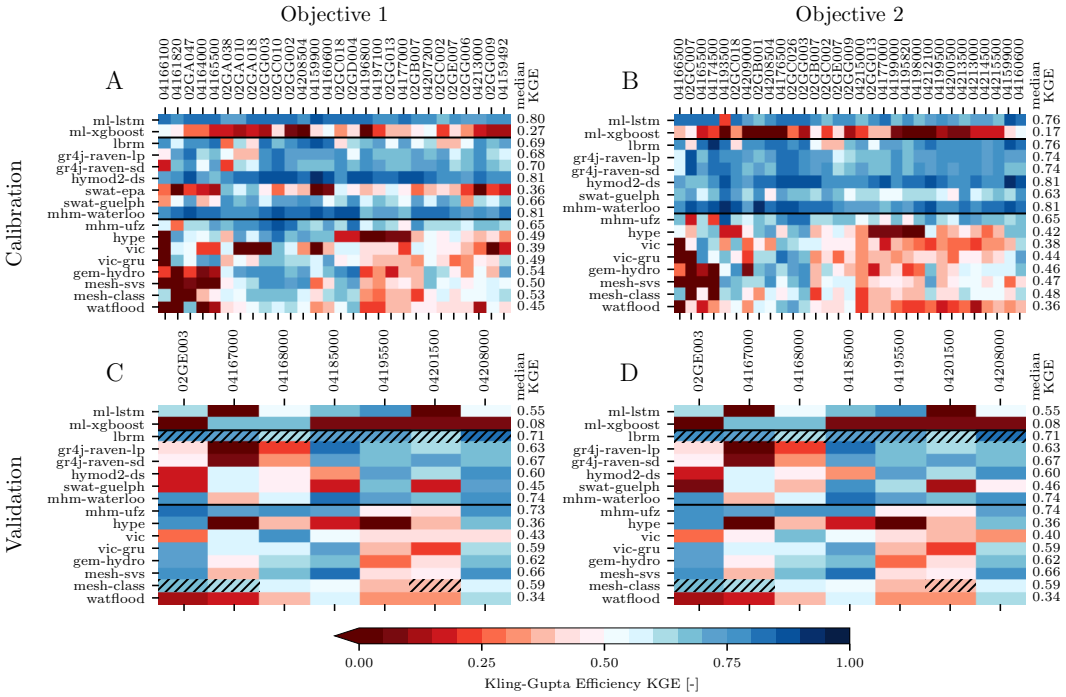

Fig. S2. The performance of the participating models in calibration and validation (A,B and C,D respectively) for each gauging station of objective 1 and 2. The colored tiles indicated the **Kling-Gupta Efficiency (KGE)** per gauge and model while the median KGE over all gauging stations is displayed 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.

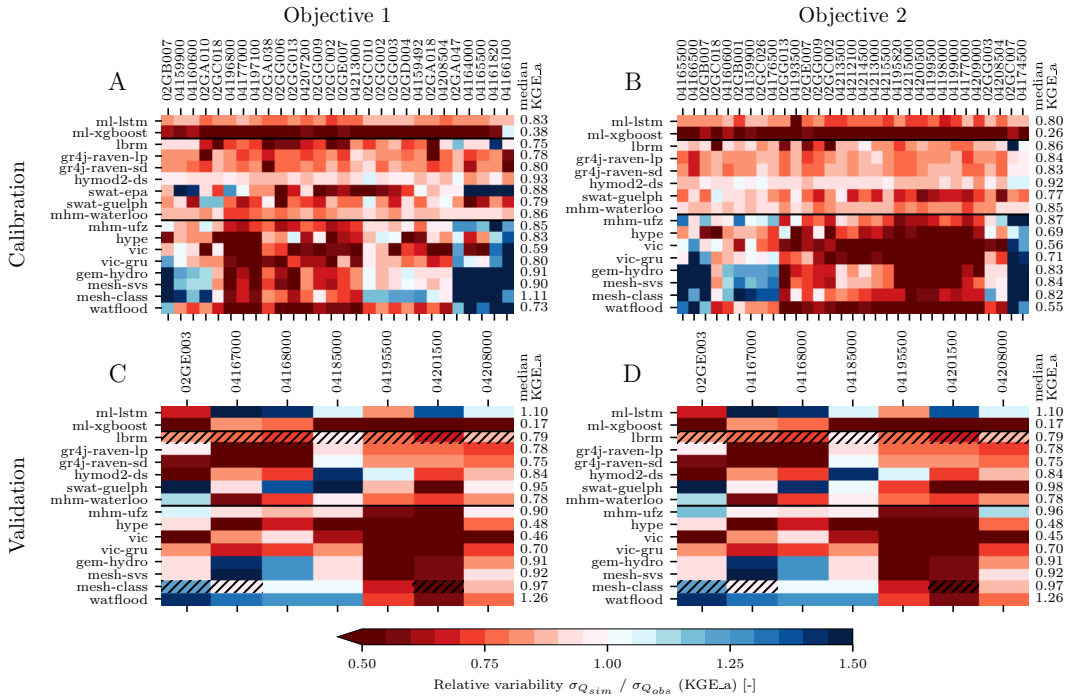

Fig. S3. The performance of the participating models in calibration and validation (A,B and C,D respectively) for each gauging station of objective 1 and 2. The colored tiles indicated the **relative variability (KGE_a)** per gauge and model while the median KGE_a over all gauging stations is displayed 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.

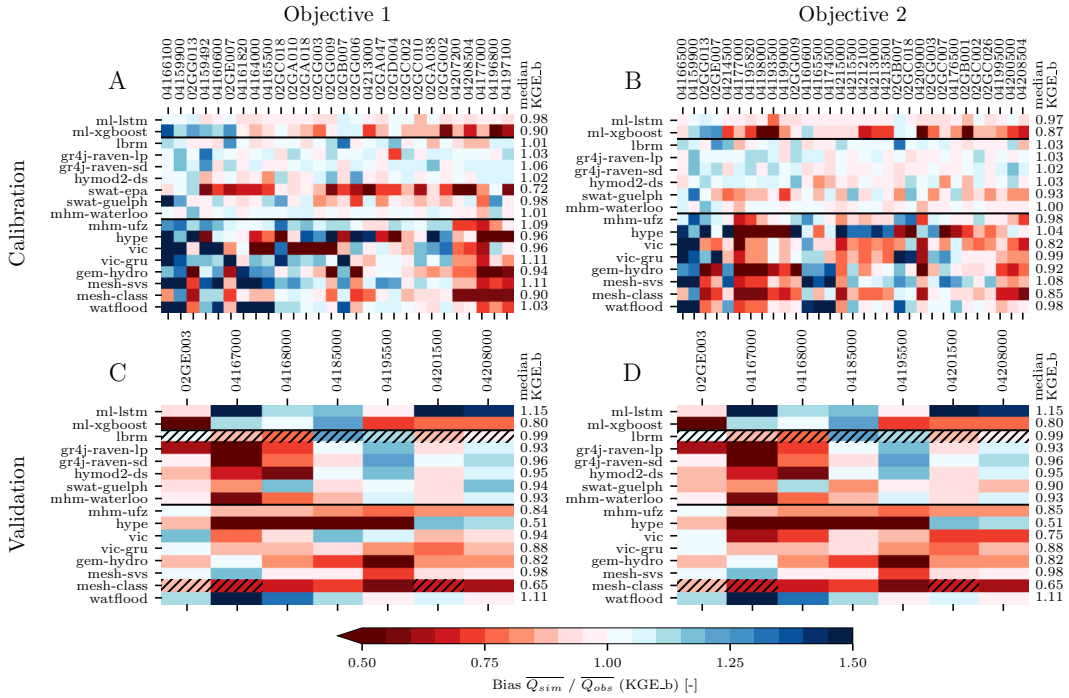

Fig. S4. The performance of the participating models in calibration and validation (A,B and C,D respectively) for each gauging station of objective 1 and 2. The colored tiles indicated the **bias (KGE_b)** per gauge and model while the median KGE_b over all gauging stations is displayed to the right. 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.

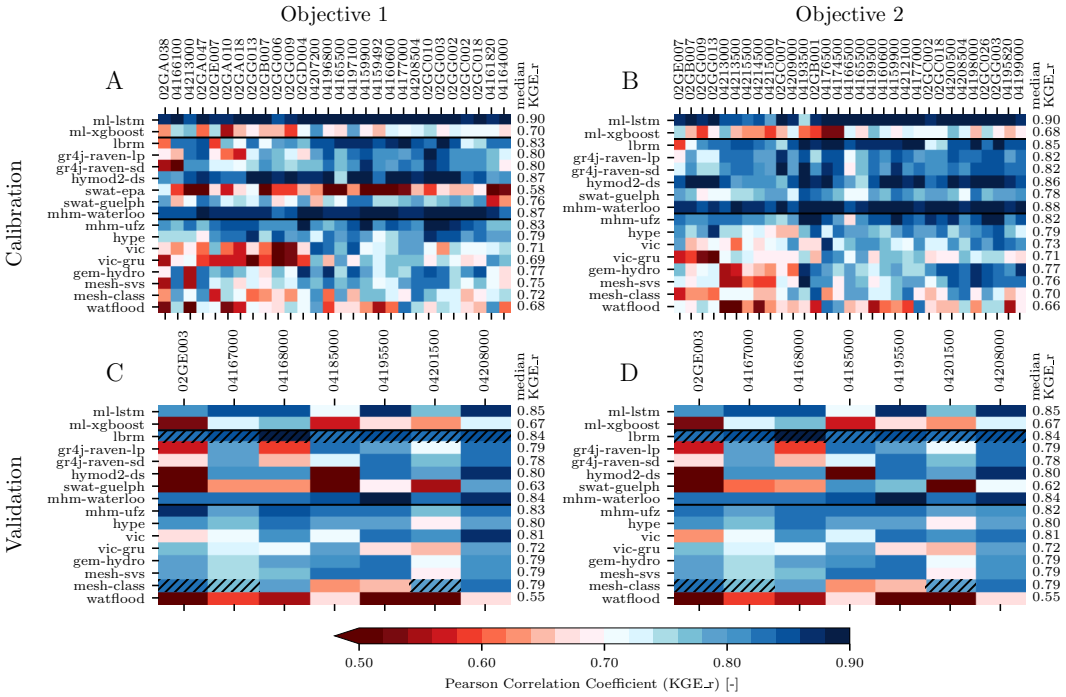

Fig. S5. The performance of the participating models in calibration and validation (A,B and C,D respectively) for each gauging station of objective 1 and 2. The colored tiles indicated the **Pearson Correlation Coefficient (KGE_r)** per gauge and model while the median KGE_r over all gauging stations 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., Muttiah, 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 Wheater, 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.

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.

Martinec, 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–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>.

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.