A Hydrological Prediction System Based on the SVS Land-Surface Scheme: efficient calibration of GEM-Hydro for streamflow simulation over the Lake Ontario basin.

Étienne Gaborit[1,*], Vincent Fortin[1], Xiaoyong Xu[2], Frank Seglenieks[3], Bryan Tolson[2], Lauren M. Fry[4], Tim Hunter[5], François Anctil[6], and Andrew D. Gronewold[5]

[1]Environment Canada, Environmental Numerical Prediction Research (E-NPR), Dorval, H9P1J3, Canada.
[2]University of Waterloo, Civil and Environmental Engineering Dpt., Waterloo, N2L3G1, Canada.
[3]Environment Canada, Boundary Water Issues, Burlington, L7S1A1, Canada.
[4]U.S. Army Corps of Engineers, Detroit District, Great Lakes Hydraulics and Hydrology Office, Detroit, 48226, U.S.A.
[5]NOAA Great Lakes Environmental Research Laboratory (GLERL), Ann Arbor, 48108, U.S.A.
[6]Civil and Water Engineering department, Université Laval, Québec, G1V0A6, Canada.

Correspondence to: Étienne Gaborit (Etienne.Gaborit@Canada.ca)

Abstract. This work explores the potential of the distributed GEM-Hydro runoff modeling platform, developed at Environment and Climate Change Canada (ECCC) over the last decade. More precisely, the aim is to develop a robust implementation methodology to perform reliable streamflow simulations with a distributed model over large and partly ungauged basins, in an efficient manner. The latest version of GEM-Hydro combines the SVS (Soil, Vegetation and Snow) land-surface scheme and the WATROUTE routing scheme. SVS has never been evaluated from a hydrological point of view, which is done here for all major rivers flowing into Lake Ontario. Two established hydrological models are confronted to GEM-Hydro, namely MESH and WATFLOOD, which share the same routing scheme (WATROUTE) but rely on different land-surface schemes. All models are calibrated using the same meteorological forcings, objective function, calibration algorithm, and basin delineation. GEM-Hydro reveals competitive with MESH and WATFLOOD: NSE √ (Nash-Sutcliffe criterion computed on the square-root of the flows) are for example equal to 0.83 for MESH and GEM-Hydro in validation on the Moira River basin, and to 0.68 for WATFLOOD. A computationally efficient strategy is proposed to calibrate SVS: a simple unit hydrograph is used for routing instead of WATROUTE. Global and local calibration strategies are compared in order to estimate runoff for ungauged portions of the Lake Ontario basin. Overall, streamflow predictions obtained using a global calibration strategy, in which a single parameter set is identified for the whole basin of Lake Ontario, show skills comparable to the predictions based on local calibration: the average NSE √ in validation and over 7 subbasins is of 0.73 and 0.61, respectively for local and global calibrations. Hence, global calibration provides spatially consistent parameter values, robust performance at gauged locations, and reduces the complexity and computation burden of the calibration procedure. This work contributes to the Great Lakes Runoff Inter-comparison Project for Lake Ontario (GRIP-O) which aims at improving Lake Ontario basin runoff simulations by comparing different models using the same input forcings. The main outcome of this study consists in a new generalizable methodology for implementing a distributed

hydrologic model with a high computation cost in an efficient and reliable manner, over a large area with ungauged portions, using global calibration and a Unit Hydrograph to replace the routing component.

Key words. Distributed models, GEM-Hydro, Local and global calibrations, Ungauged basins, Unit hydrograph.

Introduction

Given the continuous increase in precipitation forecast skill of Numerical Weather Prediction (NWP) systems~~, as documented for example over the United States (US) by~~ (Sukovich et al., ~~(~~2014), it became possible to obtain skillful runoff forecasts directly from NWP model outputs, and streamflow forecasts by routing these gridded runoff fields. Indeed, modern NWP models tend to simulate to some extent the snow, vegetation, and soil processes that contribute to the generation of runoff and streamflow. In practice, however, many limitations are still associated with the representation of such processes in NWP systems, which were documented in Clark et al. (2015) and Davison et al. (2016).

Hydrological processes simulated by land-surface schemes (LSS) have been increasingly integrated into ~~–~~NWP models (Balsamo et al., 2009; Masson et al., 2013; Alavi et al., 2016; Wagner et al., 2016), as soil water content and snow water equivalent are recognized as key state variables for streamflow forecasting (Koster et al., 2004; Entekhabi et al., 2010). Environment and Climate Change Canada (ECCC), which provides operational weather and environmental forecasts within its boundary, is currently in the process of implementing a major upgrade to the LSS of the Global Environmental Multi-scale model (GEM), the national model. This new scheme, named SVS for Soil, Vegetation and Snow, has been devised to assimilate space-based soil moisture retrievals as well as surface data, and has proven efficient at simulating soil moisture and brightness temperature (Alavi et al., 2016; Husain et al., 2016). SVS will be used to replace the Canadian version of the ISBA (*Interaction Sol-Biosphère-Atmosphère*) scheme that has been used in GEM since 2001 (Bélair et al., 2003). One of this paper's objectives is to present the first evaluation of the capabilities of the new SVS scheme for streamflow prediction in Canada.

GEM's LSSs can be run either two-way coupled to the atmospheric model or offline, using GEM or other observed atmospheric forcing. The platform for running GEM offline is known as GEM-Surf (Bernier et al., 2011). Runoff obtained from the LSS can then be routed to the outlet of the basin using the WATROUTE routing scheme (Kouwen, 2010). This configuration is known as GEM-Hydro.

Although the SVS scheme typically performed well for soil moisture simulations (e.g. Alavi et al., 2016; Husain et al., 2016), the capabilities of SVS to predict streamflow within the framework of GEM-Hydro, especially for large basins with ungauged portions, have not yet been examined. In this work, we present the calibration and evaluation of GEM-Hydro based upon the SVS scheme for streamflow simulation over the Lake Ontario basin.

The Lake Ontario basin is chosen for the application of GEM-Hydro because ~~: (1)~~ the basin can favor the examination of GEM-Hydro (and SVS) performance for runoff simulation over a wide range of hydrological conditions (mixed vegetation/land cover, natural/regulated regimes, gauged and ungauged portions)~~; (2)~~ , and because there are a large

amount of data available for model set up~~, calibration, and validation~~ for this region.~~; (3) improvements to streamflow and lake level prediction skill can have positive socio-economic impacts since this region is quite populated and industrialized; and (4) this is a Canada-USA transboundary basin co-managed by ECCC and US Army Corps of Engineers (USACE) staff, in accordance with water level management rules set by the International Joint Commission (IJC) for each control structure, including the Moses-Saunders power dam at Cornwall, the outlet of Lake Ontario.~~

Different cascades of interconnected models have been developed over the years to simulate the Great Lakes water levels and thermodynamics, as reported by Wiley et al. (2010), Deacu et al. (2012), and Gronewold et al. (2011), the latter describing the Advanced Hydrologic Prediction System (AHPS), a seasonal water supply and water level forecasting system developed by the National Oceanic and Atmospheric Administration (NOAA) Great Lakes Environmental Research Laboratory (GLERL) in the mid-1990s that has since been employed operationally by the USACE and regional hydropower authorities. Recently, ECCC has developed a short-term (84-h) operational water cycle prediction system (coupled atmospheric, hydrologic, and hydrodynamic modelling) for the Great Lakes and St. Lawrence River (WCPS-GLS, see Durnford et al., in Press). The system uses the version of GEM-Hydro that relies on the simpler ISBA LSS.

To our knowledge, the AHPS and WCPS systems are the only two systems that can provide inflow forecasts for each of the Great Lakes on both sides of the Canada-US border, and neither relies on very sophisticated hydrological models. The need for improving simulations and forecasts of runoff to the Great Lakes has been recognized by both agencies (Gronewold and Fortin, 2012). Multiple additional hydrologic models are indeed available (Coon et al., 2011), but their spatial domains are typically constrained to either the US or Canada. Before embarking on an upgrade of operational systems, GLERL and ECCC agreed to perform a number of intercomparison studies under the umbrella of the Great Lakes Runoff Intercomparison Project (GRIP), in order to better understand the status of existing systems, and to set a benchmark for model performance against which future models could be evaluated. The first study was conducted on the Lake Michigan (GRIP-M) basin by Fry et al. (2014) who compared historical runoff simulations from dissimilar hydrologic models using different calibration frameworks and input data. Amongst the models compared were GLERL's Large Basin Runoff Model (LBRM; Croley and He, 2002) that is part of the AHPS, the NOAA National Weather Service model (NWS; Burnash, 1995), and ECCC's MESH distributed model (*Modélisation Environnementale* – Surface and Hydrology; Pietroniro et al., 2007; Haghnegahdar et al., 2014). A second configuration of MESH was also included, based on Deacu et al. (2012), from which evolved the configuration of GEM-Hydro used by Durnford et al. (in Press) for the operational WCPS-GLS system. The NWS model performed best in terms of NSE, but was positively biased, perhaps because of its typical use as a flood forecasting tool. Overall, it was difficult to attribute any difference in model results to the model structure, given that different forcing data and calibration procedures had been used by each contributor to the project.

The GRIP project was extended next to Lake Ontario (GRIP-O) by Gaborit et al. (2016 a), wh~~ich~~o compared two lumped models, namely LBRM and GR4J (*modèle du Génie Rural à 4 paramètres Journalier*; Perrin et al., 2003), based upon the exact same forcing data and calibration framework. Two precipitation datasets were used as input: the Canadian Precipitation Analysis (CaPA; Lespinas et al., 2015), and a Thiessen polygon interpolation of the Global Historical

Climatology Network - Daily (GHCND; Menne et al., 2012). CaPA is a near real-time quantitative precipitation estimate product from ECCC that is available on a 10-km grid for all of North America: (http://collaboration.cmc.ec.gc.ca/cmc/cmoi/product_guide/submenus/capa_e.html).

The main finding of the first GRIP-O study is that the performance of the models was very satisfactory, resulting in an average NSE √ (Nash-Sutcliffe criterion computed on the square-root of the flows) in validation of 0.86 (over all subbasins and configurations), despite the fact that most tributaries have a regulated flow regime. This satisfactory performance justifies the use of CaPA as a precipitation forcing dataset in later studies, especially for distributed models which require gridded precipitation as input. The performance of lumped models also provides a reference level of performance when evaluating distributed hydrological models.

As an extension of the first GRIP-O study, the present work is focused on the evaluation of distributed hydrologic models for Lake Ontario basin runoff simulations. Distributed models typically have a broader range of applications than lumped ones. For example, GEM-Hydro can be utilized to estimate the Lake Ontario Net Basin Supplies (or NBS, the sum of lake tributary runoff, overlake precipitation, and overlake evaporation: Brinkmann 1983). However, distributed models are more complicated to calibrate and more computationally-intensive, especially for large basins. The present study mainly aims at developing a methodology to improve the calibration efficiency of the distributed GEM-Hydro model for streamflow modelling over the Lake Ontario basin, including its ungauged parts. The proposed methodology is transferable and can be applied to other sophisticated distributed models and large basins with ungauged parts. In order to assess the impact of the SVS land-surface scheme on runoff simulations, the GEM-Hydro model is compared with two other distributed models, which rely on the same routing scheme (WATROUTE) as used in GEM-Hydro but different land-surface schemes. The inter-comparison of the three models could also provide insight into avenues to further improve GEM-Hydro and to capture structural uncertainty in runoff simulations using the multi-model approach.

1 Methodology

1.1 Models

Three different platforms are compared in this study: MESH, WATFLOOD, and GEM-Hydro. MESH and GEM-Hydro have in common a distributed representation of most hydrological processes occurring in a basin and a structure organized around two main components: a LSS for the representation of surface processes (evapotranspiration, infiltration, snow processes, water circulation in the soils), and a river routing scheme for simulating water transport in the streams, which consists of WATROUTE for all models. WATROUTE is a 1-D hydrologic routing model relying mainly on flow directions and elevation data (Kouwen 2010). It routes to the basin outlet the surface runoff and recharge produced by the surface schemes. In WATROUTE, runoff directly feeds the streams while recharge can be provided to an optional Lower Zone Storage (LZS) compartment, representing superficial aquifers, which releases water to the streams. WATFLOOD and

GEM-Hydro make use of the LZS, whereas recharge from MESH feeds directly into the stream. WATFLOOD is not considered to include a LSS because it is not solving the energy balance, only the water balance, but it is distributed.

The version of MESH used in this study relies on version 3.6 of the Canadian LAnd Surface Scheme (CLASS). Each grid cell is subdivided in a number of tiles, and each tile is classified as belonging to one of the five grouped response units (GRUs), based on its land-use/soil type combination. In this paper, we follow the local calibration strategy advocated by Haghnegahdar et al. (2014) for MESH (see section on calibration strategy).

GEM-Hydro is very similar to MESH, but is tied to the LSSs available in GEM: ISBA and SVS. A previous study on the same basin demonstrated the clear superiority of SVS over ISBA, especially in regard to the baseflow component of the streamflow (see Gaborit et al., 2016 b). We thus only use SVS with GEM-Hydro in this paper.

WATFLOOD (Kouwen, 2010) is a distributed model of intermediate complexity that only needs precipitation and temperature as forcing, as opposed to MESH and GEM-Hydro which need additional atmospheric variables (see supplementary material). It relies on the GRUs concept and on many empirical equations. WATFLOOD has been employed by Pietroniro et al. (2007) over the Great Lakes basin.

GEM-Hydro is implemented with a 10 arcmin resolution for the LSS and 0.5 arcmin (≈ 1 km) for the routing. Sensitivity tests (Gaborit et al., 2016 b) revealed that 2 and 10 arcmin resolutions for SVS lead to quite similar performance in terms of streamflow at the outlet, while a substantial amount of computation time is saved when running the coarser resolution (see figure 1). ~~The same was shown for WATROUTE which produces outputs of similar quality be it implemented at a low or high (0.5 arcmin with GEM Hydro) resolution, as long as results are evaluated for large enough basins (i.e., basins which spread over at least a few grid cells). However, the high resolution WATROUTE version is preferred in GEM Hydro for consistency with the WCPS GLS (Durnford et al., submitted) recently developed at ECCC. Moreover, t~~The pre-processing time required by WATROUTE remains almost the same whatever the domain size (Fig. 1), which mitigates the interest of using a coarse resolution to save computation time for WATROUTE. The internal time-step used for GEM-Hydro is 10 minutes~~, which slightly improves streamflow simulations in comparison to a 30 min. time step (see Gaborit et al., 2016 b). Further reducing it does not improve the results~~. See supplementary material for MESH and WATFLOOD implementation details. ~~The internal time step of a model is generally maximized up to the desired output interval, provided that it satisfies numerical stability. In the GEM Hydro version used in this study, a 10 min. time step was required to achieve numerical stability, but a newer version now allows to increase it.~~ Table 1 shows the datasets used for physiographic information.

As the GEM-Hydro suite (including WATROUTE) is quite demanding in terms of computation time, it was decided to test a stand-alone configuration of GEM-Hydro relying on text files only and in which WATROUTE is replaced by a Unit Hydrograph (UH). This version is here forth referred to GEM-Hydro-UH. Figure 1 gives an overview of the relationship between computation time of the different models and the dimension of their domain. Note that GEM-Surf (Land-Surface part of GEM-Hydro) was run on ECCC's supercomputer while GEM-Hydro-UH and WATROUTE were run on a machine

with an AMD Athlon Dual Core Processor 4800+, because GEM-Hydro-UH and WATROUTE are not parallelized yet (their computation time would not change substantially if run on ECCC's supercomputer).

The computation time for the experiment setup described here and when splitting the domain in four on an ECCC supercomputer is about 1.5 min per day for GEM-Surf, provided that the pre-processing of the atmospheric variables was already done (which is the case in calibration: the pre-processing is done only once). WATROUTE (i.e., the routing part of GEM-Hydro) requires 25s per day for the setup described here when running on a local machine. The WATROUTE pre-processing (i.e., preparation of the WATROUTE input files from the SVS outputs, which would need to be done for each new run in calibration) takes about 30s per day and is quite constant whatever the domain size of the inputs fields. One simulation run over the GRIP-O period (4.5 years) therefore currently requires about 2 days with GEM-Hydro and prevents from performing any automatic calibration (which requires at least 400 runs, see below). GEM-Hydro-UH, based on a stand-alone version of SVS, ~~saves a tremendous amount~~requires only about 3% of the GEM-Hydro computation time ~~compared to GEM-Hydro~~ mainly because of the Input/Output processing time: the stand-alone version makes use of text files which are kept open during the simulation and requires only 3s per day on a local machine for this setup (1.2 h for the 4.5 years GRIP-O period or 20 days of calibration with 400 runs if running the whole domain). However, the computation time required by WATROUTE still had to be bypassed to perform automatic calibrations, which was done with the UH concept. The UH (see for example Sherman, 1932) allows the estimation of the streamflow at the basin outlet by partitioning the basin averages of runoff and recharge in time. The same WATROUTE LZS formulation is used in GEM-Hydro-UH in order to estimate stream recharge. The basin averages required for the UH are computed as a weighted average of the SVS grid cells located in the considered basin. The UH only requires a decay parameter corresponding to the lag or response time of the considered basin, which controls the delay between the rainfall event and the resulting streamflow peak. It is estimated with the Epsey method (Almeida et al. 2014), which requires the basin area, perimeter, and the maximum and minimum elevations along the basin main river. The UH lag-time is also used as a free parameter during calibration (Table 2). It is inspired from the UH applied to the routing storage of GR4J (Perrin et al., 2003), but is employed here at an hourly time-step. Finally, this framework allows a considerable reduction of computation time and therefore allows to perform calibration. However, GEM-Hydro-UH is faster than GEM-Hydro as long as the domain size remains of the order of a few thousand points (see Figure 1). Beyond that threshold, not only calibration is not feasible any more with GEM-Hydro-UH, but it is possible that it becomes even slower than GEM-Hydro since the latter can be parallelized. Hydrographs resulting from GEM-Hydro and GEM-Hydro-UH can be very similar (Fig. 3). Finally, the SVS parameters identified by calibrating GEM-Hydro-UH are next transferred to the full version of GEM-Hydro, which then only needs WATROUTE Manning coefficients to be adjusted (if needed) in order to mimic the optimal hydrographs obtained with GEM-Hydro-UH. This last adjustment can be done manually with a few offline WATROUTE runs.

1.2 Study area and data

The GRIP-O spatial framework is defined on Fig. 2. A more detailed description of the area is available in Gaborit et al. (2016 a).

The Lake Ontario basin (Fig. 2) covers 83 000 km^2, of which 19 000 km^2 is the lake surface. All upstream water arriving through the Niagara River is excluded to focus only on the lateral runoff component of Lake Ontario NBS (see Introduction). The US/Canada border follows the Niagara River, the middle of Lake Ontario, and the St-Lawrence River down to the Moses-Saunders dam at Cornwall, Ontario, the Lake outlet. Apart from some major cities (e.g. Toronto), the basin is mostly rural (agriculture, pasture, forest), as shown in Danz et al. (2007).

Streamflow time series were selected based on their duration and proximity to the lake shoreline. Of the 30 selected sites (Fig. 2), 27 have no missing data, 2 are complete at 94%, and one at 80% over the GRIP-O period. Nearly 70% of the total Lake Ontario basin is gauged by the selected sites. Most of the rivers are regulated in some ways, mainly for hydropower and flood mitigation, but regulation generally consists of reservoirs with a simple weir at their outlet (i.e., static control). Therefore, this did not prevent lumped models from reaching good performances in the former GRIP-O study of Gaborit et al. (2016 a). As a consequence, no effort was made to represent in a detailed manner the artificial structures of the region in WATROUTE. Moreover, the small diversions occurring to fill some canals in the region, or even the aquifers which can contribute significantly to baseflow (Singer et al., 2003; Kassenaar and Wexler, 2006), do not prevent lumped models from reaching good performances, which is helpful to this study.

The physiographic data required by the distributed models under study consist of soil texture, land use / land cover, Digital Elevation Model (DEM), and flow direction grids. Table 1 lists the datasets used to provide the physiographic and atmospheric inputs required by the models. 26 land cover classes are defined in GEM-Hydro. Soil textures are from the Global Soil Dataset for Earth system modeling (GSDE; Shangguan et al., 2014), which contains information down to 2.8 m. Soil texture was not calibrated for GEM-Hydro-UH, but some hydraulic parameters, which are derived from soil texture, were calibrated (Table 2). The maximum soil depth is calibrated in GEM-Hydro-UH (Table 2) – see supplementary material for MESH and WATFLOOD configuration details.

Precipitation forcings consist of 24-hourly accumulations from the Canadian Precipitation Analysis (CaPA version 2.4b8). Over the period of interest, CaPA consists of precipitation fields modeled by the Canadian Regional Deterministic Prediction System (RDPS, ≈15 km resolution), corrected by local rain gauge observations (Lespinas et al., 2015). The daily CaPA accumulations were disaggregated on an hourly time-step by following the temporal pattern of hourly precipitation from the RDPS (Carrera et al., 2010). The remaining atmospheric forcings are taken from RDPS outputs, using short-term forecasts having lead time of 6 to 18 h.

1.3 Calibration strategy

The GRIP-O experiment extends from June 1st, 2004 to September 26th, 2011. Calibrating a hydrologic model over a period of four to five years is generally deemed sufficient to achieve reasonable model robustness (e.g. Refsgaard et al., 1996). The calibration period thus ranges from June 1st, 2007 to September 26th, 2011 (4.5 years). Validation is from June 1st, 2005 to June 1st, 2007 (2 years, last one being used as spin-up for calibration), and spin-up from June 1st, 2004 to June 1st, 2005 (1 year). Note that during the automatic calibrations, the spin-up year was simulated only once and for all subsequent runs. The objective function is the Nash-Sutcliffe criterion (Nash and Sutcliffe, 1970) computed on the square-root of the observed and simulated time series (equation 1), in order to avoid over-emphasizing peak-flow events - here forth referred to as "NSE √".

$$NSE\sqrt{} = 1 - \frac{\sum_{i=1,n}(\sqrt{Qobs_i} - \sqrt{Qsim_i})^2}{\sum_{i=1,n}(\sqrt{Qobs_i} - \overline{\sqrt{Qobs}})^2} \quad (1)$$

These decisions are consistent with the lumped modelling decisions made for GRIP-O in Gaborit et al. (2016 a). Other evaluation criteria used in this study consist in the common Nash-Sutcliffe criteria (NSE), the Nash criteria calculated over the log of the flows ("NSE Ln"), and a Percent Bias criteria (PBIAS, equation 2) assessing the simulation's overall water budget fit: a positive value denotes a general tendency to underestimate flows, and vice-versa.

$$PBIAS = \frac{\sum_{i=1,n}(Qobs_i - Qsim_i)}{\sum_{i=1,n}(Qobs_i)} * 100 \quad (2)$$

All metrics are evaluated at the daily time-step. Calibration relies for all models on the Dynamically Dimensioned Search (DDS) algorithm (Tolson and Shoemaker, 2007). Calibration cost did not allow models to be calibrated locally for all GRIP-O subbasins (Fig. 2), but only those shown on Fig. 5. One local calibration takes between 2 and 5 days of computation (400 model runs, see below). Table 2 lists the free parameters of GEM-Hydro. GEM-Hydro-UH was calibrated using multiplicative coefficients that adjust the spatially-varying values of a given parameter, leading to a reasonable number of free parameters (16) while preserving spatial variability – see supplementary material for MESH and WATFLOOD calibration details.

It is important to emphasize that the approach used to calibrate GEM-Hydro may result in unrealistic values for some parameters, as the multiplicative coefficients could bring them beyond the range of physical coherence. More precisely, soil water content thresholds and albedo (Table 2) cannot be higher than 1. Therefore, these values were constrained to realistic ranges after they were adjusted by the calibration algorithm by imposing them a minimum value of 0 and a maximum of 1.

The initial parameter values were set to default ones that generally provide satisfactory results for the model (GEM-Hydro-UH, Table 2). The number of maximum model runs allowed was set to 400 for GEM-Hydro-UH (Sect. 2.2). This maximum appeared sufficient in the sense that the algorithm converged to a good quality final result before reaching 400. This is because the number of GEM-Hydro-UH free parameters is relatively low (16, Table 2). The DDS algorithm is very efficient in the sense that it adjusts the search behavior to the maximum number of objective function evaluations (model

runs) in order to converge to good quality solutions (Tolson and Shoemaker, 2007). The similarity of the performances obtained with GR4J and GEM-Hydro-UH (Fig. 5) supports the choice of the methodology used here, as GR4J was implemented with a maximum of 2000 model runs, three distinct calibration trials, and had an even lower number of free parameters (6, see Gaborit et al., 2016 a).

Even though the three models studied here were not calibrated using the same number of free parameters and the same maximum allowed model runs (see supplementary material), it is assumed that the calibration strategies employed allow each model to come very close to its optimal performance for a given subbasin and the time period considered. Indeed, the strategy used for each of the three models ~~is the result of expert knowledge and~~ always involves parameters affecting the whole range of the main hydrological processes~~, i.e. evaporation, snowmelt, infiltration, soil transfer, and time to peak (channel friction). It is thus logical to use different strategies for each of the models as these do not involve the same parameters, land use classification, or even physical processes.~~ The most important methodological consistencies for achieving a fair comparison between models include, in our view, a common calibration algorithm and objective function, along with common physiographic and forcing data.

Finally, some subbasins in Fig. 2 have more than one major tributary flowing into Lake Ontario. In this case, the most-downstream observed flows on independent tributaries are summed and then extrapolated to the whole subbasin using the Area Ratio Method (ARM; Fry et al., 2014). The resulting "synthetic" flows were considered as observations for GEM-Hydro-UH calibration over the whole subbasin, including its ungauged parts. This methodology was applied to all subbasins with more than one most-downstream gauge (identified with the "N/A" mention for the station attribute in Table 3) for consistency with the calibration experiments performed in the first GRIP-O study (see Gaborit et al., 2016 a), and because lumped models (and GEM-Hydro-UH) can only estimate streamflow at one location. For these subbasins, the true gauged fraction is specified in Table 3.

1.4 Strategy for ungauged areas

The ultimate objective of the GRIP-O project consists in improving simulated Lake Ontario NBS, which calls for estimating runoff from all ungauged areas. To do so, calibration was performed over the GRIP-O gauged area (which includes all GRIP-O gauged subbasins, see Fig. 2), and the resulting parameter set was used in the model implemented over the whole Lake Ontario basin, including its ungauged parts. The "GRIP-O gauged area" is actually gauged at 88.5% due to the strategy used for subbasins with several major tributaries (see end of previous section).

For GR4J, a single (unique) model was used over each of these two areas, requiring a unique calibration and a straightforward parameter transfer. ~~Therefore, the GRIP-O gauged area is represented in GR4J as if it had a unique main river. It was demonstrated in the first GRIP-O paper (see Gaborit et al., 2016 a) that a unique (i.e., single) GR4J model calibrated over a large area could lead to runoff estimates of similar quality than with multiple models implemented over local subbasins, the former strategy being more efficient.~~ Hence for GR4J, local calibration was used but with a unique model for the GRIP-O gauged area.

GEM-Hydro-UH was however implemented locally for each of the gauged GRIP-O subbasins, but a global calibration strategy (see further down) led to a unique calibrated parameter set which was then transferred to a GEM-Hydro model implemented over the whole Lake Ontario basin.

The approach based on calibration for the GRIP-O gauged area and parameter transfer to the whole Lake Ontario basin was preferred to other possible alternatives mainly ~~for two reasons:~~because it allows ~~calibrating the models using close approximations of observed flows (the area used for calibration is gauged at 88.5%, see above) instead of less reliable flow estimations for the whole basin (gauged at 70%), and to~~ takinge into account rainfall over the ungauged areas as well as rainfall over the gauged areas, or, in other words, to use the best approximation of rainfall.

The global calibration of GEM-Hydro-UH consists in finding a unique trade-off parameter set that allows to simultaneously improve performances for all subbasins (Ajami et al., 2004; Haghnegahdar et al., 2014; Gaborit et al., 2015), whereas local calibration consists in finding each subbasin's optimal parameter set. Local calibration logically leads to the optimal performances for a given subbasin, but global calibration may lead to temporal robustness (Gaborit et al., 2015) and spatial consistency of the parameter values, because they are either fixed or adjusted the same way over the whole area under study. Local calibration, on the other hand, because of equifinality and experiment imperfections (model processes, forcing data, observed flows, etc.), may compensate for simulation errors and lead to parameter sets that do not work well when transferred to other (even neighbor) subbasins, as tends to suggest the fact that very different parameter sets were obtained here with the local calibrations of GEM-Hydro-UH (Sect. 2.1 and Table 4). Global calibration is not exempt of equifinality issues either, but to a lower degree than local calibration. Indeed, the use of global parameters constrains parameter values across the basin to be equal and thus provides less freedom to achieve the same overall performance with different parameter sets. Moreover, the attention paid to the parameter ranges used (Table 2) allows to be confident in the physical relevance of the final parameter values.

The objective function associated to global calibration of GEM-Hydro-UH is as follows:

$$OF = \sum_{i=1}^{N}\left(1 - \frac{Nloc_i}{Nglob_i}\right) \quad (2)$$

with $Nloc_i$ the NSE √ value calculated from the local calibration on subbasin i, and $Nglob_i$ the NSE √ calculated from the global calibration on subbasin i. This objective function aims minimizing differences between performances obtained from global and local parameter sets. ~~It does rely on the hypothesis that global performance cannot be higher than local performance, but even if it was the case, this objective function would still be valid.~~ However, as GEM-Hydro-UH was not locally calibrated for all of the 14 GRIP-O subbasins (only those of Fig. 5 because of the computation cost), performances obtained with local GR4J calibrations (Gaborit et al., 2016 a) were used for missing ones, justifying the use of that model in this study. ~~This substitution does make sense considering that firstly, GR4J and GEM-Hydro-UH local performances are similar (Fig. 5), that GR4J local performances were always very satisfactory (see Gaborit et al., 2016 a), and finally that the objective function still makes sense if global performance is higher than the local one.~~.

~~Moreover, a supplemental free parameter was used for GEM-Hydro-UH during global calibration (in addition to those in Table 2), namely the percentage of completely impervious urban areas. This value was fixed to 0.33 during local calibrations, implying that 33% of liquid precipitation or snowmelt over urban covers was automatically considered as runoff with no chance to infiltrate. This value comes from a former study calibrating the SWMM 5 model over urban subbasins in Québec City, Canada (Gaborit et al., 2013). With local calibration, good performances could be reached, using this fixed value, even for "urban" subbasins (such as subbasins 14 and 15 in Sect. 2.1) as the effect of urban surfaces could be accounted for by the other free parameters of Table 2. However, calibrating this parameter helps to distinguish between natural and urban surfaces during global calibration. The calibrated value of the urban cover fraction, which is completely impervious, is equal to 0.69 after global calibration (Table 4). This does make sense as the urban areas around the shore of Lake Ontario generally correspond to high density areas, such as for the city of Toronto. Note also that w~~With global calibration, the response time parameter controlling the UH duration (Table 2) was replaced with a multiplicative factor adjusting the default response times of all local subbasins.

Models were finally implemented over the whole Lake Ontario basin (Fig. 2), and runoff simulations performed with the parameter sets calibrated over the GRIP-O gauged area. GEM-Hydro was selected for this task instead of GEM-Hydro-UH since it was more straightforward and a priori more realistic (see further) to use WATROUTE instead of the simple UH for the ungauged areas of the lake Ontario basin. In GEM-Hydro, standard Manning coefficients were used in WATROUTE, while the lag-time of GEM-Hydro-UH was adjusted during calibration. But it was assessed that simulations with GEM-Hydro (calibrated SVS and LZS parameters and standard Manning values) were very close, both in terms of hydrographs and performances at the gauged sites, to those from the calibrated GEM-Hydro-UH. ~~Performances are generally even slightly better with GEM Hydro (despite the standard Manning values) than with GEM Hydro UH for individual subbasins (not shown), despite the opposite is true when looking at the total GRIP-O gauged area as a whole (see Table 5).~~ Figure 4 summarizes the methodology described here for estimating runoff from the ungauged areas of the Lake Ontario basin with GEM-Hydro.

2 Results

The comparison between GEM-Hydro and GEM-Hydro-UH is first presented to demonstrate the relevance of the UH approach to save the computation time associated with running the routing model of GEM-Hydro. Score improvements obtained by calibrating GEM-Hydro-UH for several subbasins of Lake Ontario basin are then presented, followed by a performance comparison for all models. Finally, the methodology proposed with GEM-Hydro and the lumped GR4J model to simulate streamflows for the ungauged parts of the Lake Ontario basin is evaluated.

Figure 3 presents the hydrographs simulated for the Moira River (subbasin 11 in Fig. 2), with SVS default parameters, standard WATROUTE parameter values in the case of GEM-Hydro, and a UH lag time estimated with the Epsey method in the case of GEM-Hydro-UH. As can be seen from this figure, GEM-Hydro-UH is able to produce streamflow simulations which are very close to those obtained with GEM-Hydro, underlying the relevance of such an

approach to save computation time. Between the uncalibrated GEM-Hydro and GEM-Hydro-UH performances and over the different GRIP-O subbasins, the average absolute difference in NSE √ was 8% with the worst difference being 21% (GEM-Hydro being most of the time better than GEM-Hydro-UH). ~~See also Table 5 for a comparison between the calibrated GEM-Hydro and GEM-Hydro-UH models when looking at performances for the total GRIP-O gauged area (In this case, GEM-Hydro-UH is better because WATROUTE Manning coefficients of GEM-Hydro are still the standard values, see section 2.3).~~ A complete GEM-Hydro run over the GRIP-O calibration period (4.5 years) takes about 48 hours, while the GEM-Hydro-UH version requires only 1.2 hours over the same period.

2.1 GEM-Hydro-UH local calibrations

This section presents GEM-Hydro-UH performances (Fig. 5) either with its default parameter values or after its local calibration on Lake Ontario subbasins, whose characteristics are given in Table 3.

As can be seen from Fig. 5, calibration provides substantial improvements in NSE √ values. Similar results were obtained for NSE and NSE Ln (although these results are not shown), and a lower improvement for PBIAS. Interestingly, all uncalibrated NSE √ are above zero (Fig. 5), and even satisfactory for subbasins 10 and 11. This is encouraging for ungauged subbasin applications. It can also be noticed on Fig. 5 that calibration sometimes inverts the sign of the PBIAS criteria (switching from over- to under-estimation or vice-versa).

Calibration also improves GEM-Hydro-UH Snow Water Equivalent (SWE) simulations but to a lesser degree than for the streamflow. For example, the NSE values for SWE simulations over the 4 consecutive winters of the GRIP-O period improved from -0.12 to 0.42 for the Genessee subbasin, and from 0.49 to 0.68 for the Black River subbasin, respectively before and after calibration (the SWE variable was not used in the computation of the objective function). SWE observations come from the SNow Data Assimilation System (SNODAS, see NOHRSC 2004). Calibration does influence evapotranspiration, but no observations are available to evaluate this model output. For example, for the Moira River, the mean subbasin annual evapotranspiration (over the calibration period) is equal to 527 mm and to 647 mm, before and after calibration respectively. The robustness of the model is also deemed very good, since performances do not substantially deteriorate between calibration and validation (Table 5).

Calibrated parameter values are quite different from one subbasin to the other (even for neighbor subbasins), which may be due to equifinality (different parameter sets can lead to similar simulations) but also to the anthropogenic streamflow regulations. Table 4 presents the ranges of the final parameter values obtained with local calibration. This strongly limits the potential for parameter transferability to ungauged subbasins (Razavi and Coulibaly, 2012; Parajka et al., 2013). As explained in Sect. 1.4, global calibration can help overcoming this by leading to a spatially-coherent parameter set. Results of such an approach are presented in Sect. 2.3.

Calibrated GEM-Hydro-UH performance values are generally very close to those obtained with GR4J and CaPA precipitation (Fig. 5): the mean absolute difference in NSE √ values is 6.1%, with the maximum being 15% (GR4J being generally better). This is very encouraging as the performance benchmark set by GR4J simulations is most of the time quite

high and hard to attain for other models. ~~Therefore, GRIP-O allowed to improve streamflow simulations for the Lake Ontario basin, in comparison to the studies of Croley (1983) and Haghnegahdar et al. (2014), which are the main former studies who proposed the implementation of hydrologic models over this area. Moreover, as new improvements are in progress for SVS (see below), it is probable that GEM-Hydro-UH and GEM-Hydro will even be able to surpass GR4J in terms of performance in the near future.~~

2.2 Inter-comparison of all models

This section aims at comparing MESH, WATFLOOD, and GEM-Hydro-UH, but detailed results specific to MESH and WATFLOOD are only provided in the supplementary materials to this paper. When looking closely at the Moira River hydrographs, for the three calibrated models (Fig. 6), important differences arise. For instance, WATFLOOD has a more flashy behavior and tends to overestimate peak flow events, MESH generally underestimates flows, and GEM-Hydro-UH lays somewhere in between. Peak flow events (even for other subbasins) associated to the spring freshet are generally better represented by MESH, which may be due to a better representation by CLASS of various cold regions hydrological processes, such as snow accumulation and melt, snow interception by vegetation, as well as soil freezing and thawing. NSE $\sqrt{}$ in validation for this basin are respectively equal to 0.83, 0.68, and 0.83 for MESH, WATFLOOD, and GEM-Hydro-UH.

2.3 Runoff estimation for the whole Lake Ontario basin

The parameter values identified from the global calibration are presented in Table 4, along with the ranges resulting from local calibrations. See Sect. 1.4 for more information about methodology related to global calibration. It can be seen from Table 4 that final global parameters generally lay inside the intervals obtained from local calibration, highlighting the trade-off found by global calibration. Moreover, it was noticed (not shown here) that parameter values were very different between local and global calibration procedures, even for basins displaying very similar performances between the two strategies (such as subbasins 3, 5 and 8, see Fig. 7), highlighting the fact that local calibration is more prone to over-calibration (i.e., equifinality).

GEM-Hydro-UH results are given first for each gauged subbasin, in order to compare global calibration, local calibration and default parameters (Fig. 7), followed by GR4J and GEM-Hydro results (with global calibration) for the GRIP-O gauged area and the whole Lake Ontario basin (Table 5 and Figs. 8-9).

GEM-Hydro-UH performances are lower with global calibration than with local calibration, as expected, and sometimes even lower after global calibration than with the default parameters for some subbasins (notably 10 and 11, Fig. 7). However, performances are satisfactory for most of the 14 GRIP-O subbasins with a single parameter set, which confirms that global calibration fulfilled expectations. Given that it takes about 7 days to achieve a local calibration, global calibration, which was completed in 20 days for the 14 subbasins at once, allows to save a substantial amount of computation time. Furthermore, global calibration favors the spatial consistency of parameters and facilitates parameter transfer to ungauged areas, whereas there is no a priori best manner to transfer parameter values obtained from local calibration (Razavi and

Coulibaly, 2012; Parajka et al. 2013). In this study, the strategy related to parameter transfer to the ungauged subbasins is based on spatial proximity, which was already identified as among the best parameter transfer methods for this type of climate in Canada (Razavi and Coulibaly, 2012). Despite a comprehensive assessment of the reliability of the methodology used here for parameter transfer would require the "leave-one-out" framework (see Razavi and Coulibaly, 2012), the satisfying performances and temporal robustness obtained for all GRIP-O subbasins with global calibration, along with the spatial consistency of the unique final parameter set, the homogeneity of the area under study and the spatial proximity of ungauged basins together justify the relevance and a priori reliability of the methodology employed in this study. This statement is moreover supported by the evaluation performed further down for the whole basin.

Performance evaluation for the total GRIP-O gauged area (Table 5) shows that GR4J is better than GEM-Hydro-UH in calibration, but worse in validation. GEM-Hydro-UH leads to a very satisfactory performance, but most importantly to a better streamflow simulation than GR4J in terms of dynamics (see Fig. 8). Note that the smoother GR4J behavior is not due to the single model approach for the whole area, as a similar behavior occurred when aggregating simulations from local GR4J models (Gaborit et al., 2016 a). This smooth behavior seems inherent to the lumped attribute and concepts of GR4J. As depicted in Table 5, performances for the GRIP-O gauged area obtained with GEM-Hydro are close to those obtained with GEM-Hydro-UH, despite being lower for the former, which comes from the standard (uncalibrated) Manning coefficients used with GEM-Hydro, whereas the UH lag time was adjusted during the calibration of GEM-Hydro-UH. WATROUTE coefficients could have been manually tuned in order for GEM-Hydro performance values to reach those of GEM-Hydro-UH in Table 5, but this was not deemed necessary given the already very satisfying performance values obtained with the uncalibrated Manning values.

Runoff simulations for the whole Lake Ontario basin, including its ungauged areas, are very promising (Table 5). Even if runoff observations actually consist in this case in estimations based on the ARM, computed performances are a priori reliable given that the true gauged fraction of the total area is equal to about 70%, and that the ARM proves reliable starting from a 50% gauged fraction (Fry et al., 2014). ~~GEM-Hydro (and GEM-Hydro-UH) tends to overestimate streamflow total volumes (Table 5, PBIAS), while GR4J achieves a better estimation of the total runoff volumes. The fact that GR4J is better than GEM-Hydro-UH in terms of PBIAS is attributed to the fact that GR4J consists in a single (global) model for the whole area considered, which makes it easier to accommodate the overall water balance (whereas GEM-Hydro-UH is made of several subbasins which are calibrated globally). Indeed, PBIAS values obtained aggregating local GR4J models were poorer (Gaborit et al., 2016 a).~~

~~It is important to emphasize that for the whole basin including its ungauged parts, runoff was estimated with GEM-Hydro instead of GEM-Hydro-UH, which means that streamflow simulations are available at all points inside the domain, whereas GR4J only delivers estimations at the outlet. Moreover, even if the scores are slightly better for GR4J, the streamflow dynamics are generally better represented by GEM-Hydro, as is the case for example for the 2006 summer of Figure 8: GR4J represents a smooth streamflow recession, while GEM-Hydro-UH better follows the small peaks and drops occurring during the recession.~~

It is therefore argued that the methodology proposed here (global calibration of GEM-Hydro-UH and parameter transfer to GEM-Hydro) is relevant, efficient, and reliable, provided that a large enough fraction of the total area is gauged. It could moreover be applied in different climatic contexts, regions, and with different models.

~~Simply extrapolating GEM-Hydro-UH simulated flows from the GRIP-O gauged area to the whole Lake Ontario basin with the ARM leads to the exact same performances as those of the GRIP-O gauged area, because when doing so, we end up with both the simulated and observed flows being extrapolated the same way (i.e., with the ARM), which does not change the scores at all. Based on these scores, it seems that extrapolating the GEM-Hydro-UH flows to the whole Lake Ontario basin leads to better results than transferring the calibrated parameters to GEM-Hydro over the whole Lake Ontario basin (because of the standard WATROUTE coefficients used in GEM-Hydro), but again, GEM-Hydro-UH simulations are only available at the basin outlet, instead of at all GEM-Hydro internal points.~~

Finally, Lake Ontario monthly NBS were estimated with the globally calibrated GEM-Hydro model, and results were compared both to the GLERL residual and component NBS estimates (Fig. 9). Residual NBS rely on the lake observed change in storage and streamflows for the Niagara and St-Lawrence rivers (DeMarchi et al., 2009). Component NBS used here are based on the GLERL Monthly Hydrometeorological Database (GLM-HMD; Hunter et al., 2015), which relies on observed data extrapolated with the ARM for runoff, on observed data interpolated with the Thiessen polygon method for overlake precipitation, and on the Large Lake Thermodynamics lumped Model (LLTM) for overlake evaporation. Component NBS estimates are updated on a regular basis. Data used in this work were updated on August 2nd, 2016. It is still unknown which of these two NBS estimation methods (i.e., residual or component method) is the most accurate (DeMarchi et al., 2009).

It can be seen that the cumulated NBS estimates derived from the calibrated GEM-Hydro model (using global calibration) stand between the component and residual NBS estimates, but are closer to the latter ones. It is however difficult to draw any conclusion regarding the bias of these estimation methods given the uncertainty associated with NBS estimates (DeMarchi et al., 2009). When comparing the GLM-HMD component NBS method to the calibrated GEM-Hydro simulation on a component-by-component basis, the main difference between the two occurs for overlake evaporation, with evaporation from the component method being significantly lower than GEM-Hydro evaporation (not shown). This mainly explains why the NBS estimates from the component method are higher than the other estimates in Figure 9. But again, it is not possible to accurately evaluate overlake evaporation estimates given the lack of observations for this variable. The uncalibrated GEM-Hydro model results in cumulative NBS estimations which are below all other NBS estimations, which tends to suggest that they are underestimated. Therefore, the methodology proposed to calibrate GEM-Hydro seems to improve Lake Ontario NBS simulations.

Discussion and Conclusion

This study explored for the first time the performance of SVS to estimate runoff for a large basin with ungauged portions. Our results indicate that the SVS LSS, as embedded in GEM-Hydro and GEM-Hydro-UH, led to reasonable streamflow simulations for the Lake Ontario basin. According to the inter-comparison experiment conducted for three subbasins (see supplementary material), GEM-Hydro-UH and GEM-Hydro are both competitive to MESH and WATFLOOD. GEM-Hydro has even proven able to produce decent, generally satisfactory runoff simulations with default parameter values, except for areas with a high urban cover fraction. This result is encouraging because SVS is expected to replace ISBA in ECCC operational models in the coming years.

The model inter-comparison study also indicates that there is still room to further improve SVS. For example, adding the soil freeze-thaw processes to the current SVS may improve GEM-Hydro simulations of runoff peaks in spring. Additionally, a new snow module (ISBA-ES) is also being implemented into SVS, which currently relies on a simple force-restore approach. Finally, ~~the~~ work is under way to represent a surface of ponded water in each SVS grid cell, in order to represent subgrid-scale lakes, wetlands, and to better account for the delay associated with surface runoff transfer into the streams.

The calibrated GEM-Hydro-UH performance values are close to GR4J ones (Gaborit et al., 2016 a). The potential benefits of global calibration have been demonstrated here, as for a previous Hydrotel application (Gaborit et al., 2015)~~.~~. It achieves satisfactory performances for a large area with a unique calibration and favors temporal robustness, spatial consistency, and parameter transferability. Therefore, one of this study's main outcomes is the confirmation that global calibration is a very promising and efficient methodology to implement hydrologic models over large areas. It saves computation time and leads to a spatio-temporally robust parameter set that can be transferred to nearby (ungauged) areas. This outcome is important because parameter transfer methods derived from local calibration are still largely prone to failure. More studies still have to be performed with global calibration on other ~~-~~basins and with other hydrological models to confirm the value of this methodology, which worked well for the model and basin studied here. ~~———~~Global calibration of SVS is envisioned in future versions of the WCPS, to assess its benefits in improving weather forecasts, as a calibrated SVS could be coupled to the RDPS atmospheric model, and because a calibrated SVS version should improve surface fluxes representation. Calibrating a LSS based on streamflow and then using it in an atmospheric model to improve weather forecasts has not been reported in the literature so far, to our best knowledge. Another originality of this work which may be of interest to a broad audience is the way the distributed parameters were adjusted during calibration. Instead of regrouping the parameters by GRU as for SA-MESH (see supplementary material) and which led to 60 free parameters during calibration, GEM-Hydro-UH was calibrated with only 16 parameters, which consist mainly of multiplicative factors by which the associated actual parameter values were all multiplied the same way. This allows preserving the spatial variability and coherence of a given parameter, while minimizing the number of free parameters that still affect the whole domain. Of course, additive or exponent factors could be used too, if deemed more relevant. This strategy is moreover suited to using the DSS algorithm, which allows a very fast convergence (in less than 400 iterations) when a limited number of free parameters are used, and therefore contributes to the efficiency of the implementation methodology proposed here. Again,

this could be applied to any distributed hydrologic model. Furthermore, in order to calibrate the GEM-Hydro model, its standard routing part was replaced by a simple UH during calibration of the land-surface scheme, ~~which saved a tremendous amount of computation time~~the simpler setup requiring only 3% of the original computation time. The routing component of GEM-Hydro can be run afterwards, and re-calibrated separately. Once again, the UH can be used with any LSS and on any basin, which allows to calibrate a distributed model when the routing part is time-consuming, as for WATROUTE.

We developed a methodology (global calibration, multiplicative factors used in calibration, and the UH bypass of the routing component) to improve the calibration efficiency and performance of the distributed GEM-Hydro model for streamflow modelling over the Lake Ontario basin, including its ungauged parts. The proposed methodology is transferable and can be useful to the hydrologic community, especially for those who want to use distributed hydrologic models to simulate streamflow for large basins with ungauged parts.

Finally, this work presented the development of an efficient distributed hydrological modeling platform for the Lake Ontario basin, which can be used as a readily testing ground for distributed models. During the preparation and writing of this paper, using the proposed methodology in this study, GEM-Hydro was also applied to the Canadian Nelson, Churchill, and MacKenzie River basins as well as the whole Hudson Bay basin, with satisfactory performance values. This is encouraging given the high degree of regulation involved in some of these basins.

Acknowledgements

This work was supported by the IJC International Watersheds Initiative. The authors wish to thank Dorothy Durnford, Nasim Alavi, and Maria Abrahamowicz, from Environment and Climate Change Canada (ECCC), who provided support with the GEM-Hydro platform implementation, and Djamel Bouhemhem (ECCC), for informatics support. This is NOAA-GLERL contribution No. XXXX.

References

Ajami, N. K., Gupta, H., Wagener, T., and Sorooshian, S.: Calibration of a semi-distributed hydrologic model for streamflow estimation along a river system. J. Hydro., 298(1), 112–135, 2004.

Alavi, N., Bélair, S., Fortin, V., Zhang, S., Husain, S. Z., Carrera, M. L., and Abrahamowicz, M.: Warm Season Evaluation of Soil Moisture Prediction in the Soil, Vegetation and Snow (SVS) Scheme. J. Hydromet., 17(8), 2315–2332, doi: 10.1175/JHM-D-15-0189.1, 2016.

Alfieri, L., Burek, P., Dutra, E., Krzeminski, B., Muraro, D., Thielen, J., and Pappenberger, F.: GloFAS–global ensemble streamflow forecasting and flood early warning. Hydrol. Earth Syst. Sci., 17(3), 1161–1175, doi: 10.5194/hess-17-1161-2013, 2013.

Almeida, I.K., Almeida, A.K., Anache, J.A.A., Steffen, J.L., and Alves Sobrinho, T.: Estimation on time of concentration of overland flow in watersheds: a review. Geociências, 33(4), 661–671, 2014.

Balsamo, G., Beljaars, A., Scipal, K., Viterbo, P., van den Hurk, B., Hirschi, M., and Betts, A. K.: A revised hydrology for the ECMWF model: Verification from field site to terrestrial water storage and impact in the Integrated Forecast System. J. Hydromet., 10(3), 623–643, doi: 10.1175/2008JHM1068.1, 2009.

Bélair, S., Crevier, L. P., Mailhot, J., Bilodeau, B., and Delage, Y.: Operational implementation of the ISBA land surface scheme in the Canadian regional weather forecast model. Part I: Warm season results. J. Hydromet., 4(2), 352–370, doi: 10.1175/1525-7541(2003)4%3C352:OIOTIL%3E2.0.CO;2, 2003.

Bernier, N. B., Bélair, S., Bilodeau, B., and Tong, L.: Near-surface and land surface forecast system of the Vancouver 2010 Winter Olympic and Paralympic Games. J. Hydromet., 12(4), 508–530, doi: 10.1175/2011JHM1250.1, 2011.

Brinkmann, W. A. R.: Association between net basin supplies to Lake Superior and supplies to the lower Great Lakes. Journal of Great Lakes Research, 9(1), 32–39, 1983.

Burnash, R.J.C.: The NWS river forecast system – catchment modelling, in: Computer Models of Watershed Hydrology, Singh, V. (Ed.), Water Resources Publications, Highlands Ranch, CO, 311–366, 1995.

Carrera, M. L., Bélair, S., Fortin, V., Bilodeau, B., Charpentier, D., and Doré, I.: Evaluation of snowpack simulations over the Canadian Rockies with an experimental hydrometeorological modeling system. J. Hydromet., 11(5), 1123–1140, 2010.

Clark, M. P., Fan, Y., Lawrence, D. M., Adam, J. C., Bolster, D., Gochis, D. J., Hooper, R.P., Kumar, M., Leung, L.R., Mackay, D.S, and Maxwell, R. M.: Improving the representation of hydrologic processes in Earth System Models. Water Resources Research, 51(8), 5929–5956, 2015.

Coon, W. F., Murphy, E. A., Soong, D. T., and Sharpe, J. B.: Compilation of watershed models for tributaries to the Great Lakes, United States, as of 2010, and identification of watersheds for future modeling for the Great Lakes Restoration Initiative, US Geological Survey, Troy, NY, Open File Rep. 2011-1202, 2011.

~~Croley, T. E.: Great Lakes basins (U.S.A. Canada) runoff modeling. J. Hydro., 64, 135–158, 1983.~~

Croley, T. E.: Great Lakes climate change hydrologic impact assessment: IJC Lake Ontario-St. Lawrence River regulation study. Great Lakes Environmental Research Laboratory, Ann Arbor, MI, NOAA technical memorandum GLERL-126, 77pp., 2003.

Croley, T. E., and He, C.: Great Lakes large basin runoff modeling, in: Proceedings of the Second Federal Interagency Hydrologic Conference, Las Vegas, NV, July 2002.

Danz, N. P., Niemi, G. J., Regal, R. R., Hollenhorst, T., Johnson, L. B., Hanowski, J. M., Axler, R.P., Ciborowski, J.J., Hrabik, T., Brady, V.J., and Kelly, J. R.: Integrated measures of anthropogenic stress in the US Great Lakes basin. Environmental Management, 39(5), 631–647, 2007.

Davison, B., Pietroniro, A., Fortin, V., Leconte, R., Mamo, M., and Yau, M. K.: What is Missing from the Prescription of Hydrology for Land Surface Schemes? J. Hydromet., doi: 10.1175/JHM-D-15-0172.1, 2016.

Deacu, D., Fortin, V., Klyszejko, E., Spence, C., and Blanken, P. D.: Predicting the net basin supply to the Great Lakes with a hydrometeorological model. J. Hydromet., 13(6), 1739–1759, 2012.

DeMarchi, C., Dai, Q., Mello, M. E., and Hunter, T. S.: Estimation of Overlake Precipitation and Basin Runoff Uncertainty. International Upper Great lakes Study, Case Western Reserve University, Cleveland, OH, 64 pp., 2009.

Durnford, D., Fortin, V., Smith, G., Archambault, B., Deacu, D., Dupont, F., Dyck, S., Martinez, Y., Klyszejko, E., MacKay, M., Liu, L., Pellerin, P., , , Pietroniro, A., Roy, F., Vu, V., Winter, B., Yu, W., Spence, C., Bruxer, J., Dickhout, J. : Towards an operational water cycle prediction system for the Great Lakes and St. Lawrence River. Submitted to the Bulletin of the American Meteorological Society.

Entekhabi, D., Njoku, E. G., O'Neill, P. E., Kellogg, K. H., Crow, W. T., Edelstein, W. N., Entin, J.K., Goodman, S.D., Jackson, T.J., Johnson, J.,. and Kimball, J.: The soil moisture active passive (SMAP) mission. Proceedings of the IEEE, 98(5), 704–716, doi: 10.1126/science.1100217, 2010.

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.: The Great Lakes Runoff Intercomparison Project Phase 1: Lake Michigan (GRIP-M). J. Hydro., 519(D), 3448–3465, doi: 10.1016/j.jhydrol.2014.07.021, 2014.

Gaborit, É., Fortin, V., Tolson, B., Fry, L., Hunter, T., and Gronewold, D.: Great Lakes Runoff Inter-comparison Project, Phase 2: lake Ontario (GRIP-O). Journal of Great Lakes Research, Available online 5 December 2016, ISSN 0380-1330, 2016 a, doi: 10.1016/j.jglr.2016.10.004..

Gaborit, É., Fortin, V., and Tolson, B.: Great Lakes Runoff Intercomparison Project for Lake Ontario (GRIP-O). Environment and Climate Change Canada, Dorval, QC, internal report, 133 pp., 2016 b. Accessible online at http://collaboration.cmc.ec.gc.ca/science/rpn/publications/pdf/GRIPO_report.pdf

Gaborit, É., Ricard, S., Lachance-Cloutier, S., Anctil, F., and Turcotte, R.: Comparing global and local calibration schemes from a differential split-sample test perspective. Canadian Journal of Earth Sciences, 52(11), 990–999, doi: 10.1139/cjes-2015-0015, 2015.

~~Gaborit, É., Muschalla, D. , Vallet, B. , Vanrolleghem, P.A., and Anctil, F.: Improving the performance of stormwater detention basins by real-time control using rainfall forecasts. Urban Water Journal, 10(4), 230–246, doi:10.1080/1573062X.2012.726229, 2013.~~

Gronewold, A. D., and Fortin, V.: Advancing Great Lakes hydrological science through targeted binational collaborative research. BAMS, 93(12), 1921–1925, 2012.

Gronewold, A.D., Clites, A.H., Hunter, T.S., and Stow, C.A.: An appraisal of the Great Lakes advanced hydrologic prediction system. J. Great Lakes Res., 37, 577–583, 2011.

Haghnegahdar, A.: An improved framework for watershed discretization and model calibration: Application to the Great Lakes basin, Ph.D. thesis, University of Waterloo, 115 pp., 2015.

Haghnegahdar, A., Tolson, B. A., Davison, B., Seglenieks, F. R., Klyszejko, E., Soulis, E. D., Fortin, V., and Matott, L. S.: Calibrating Environment Canada's MESH Modelling System over the Great Lakes Basin. Atmos.-Ocean, 52(4), 281–293, 2014.

Hunter, T.S., Clites, A.H., Campbell, K.B., and Gronewold, A.D.: 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. ISSN 0380-1330, doi: 10.1016/j.jglr.2014.12.006, 2015.

Husain, S. Z., Alavi, N., Bélair, S., Carrera, M., Zhang, S., Fortin, V., Abrahamowicz, M., and Gauthier, N.: The Multi-Budget Soil, Vegetation, and Snow (SVS) Scheme for Land Surface Parameterization: Offline Warm Season Evaluation. J. Hydromet., 17(8), 2293–2313, doi: 10.1175/JHM-D-15-0228.1, 2016.

Kassenaar, J.D.C., and Wexler, E.J.: Groundwater Modelling of the Oak Ridges Moraine Area. York, Peel, Durham, Toronto and The Conservation Authorities Moraine Coalition (YPDT-CAMC), ON, CAMC-YPDT Technical Report #01-06, 2006.

Koster, R. D., Dirmeyer, P. A., Guo, Z., Bonan, G., Chan, E., Cox, P., Gordon, C.T., Kanae, S., Kowalczyk, E., Lawrence, D., and Liu, P.: Regions of strong coupling between soil moisture and precipitation. Science, 305(5687), 1138–1140, doi: 10.1126/science.1100217, 2004.

Kouwen, N.: WATFLOOD/WATROUTE Hydrological model routing & flow forecasting system. Department of Civil Engineering, University of Waterloo, Waterloo, ON, 267 pp., 2010.

Lespinas, F., Fortin, V., Roy, G., Rasmussen, P., and Stadnyk, T.: Performance Evaluation of the Canadian Precipitation Analysis (CaPA). J. Hydromet., 16 (5), 2045–2064, 2015.

Masson, V., Le Moigne, P., Martin, E., Faroux, S., Alias, A., Alkama, R., Belamari, S., Barbu, A., Boone, A., Bouyssel, F., and Brousseau, P.: The SURFEXv7. 2 land and ocean surface platform for coupled or offline simulation of Earth surface variables and fluxes. Geoscientific Model Development, 6, 929–960, doi:10.5194/gmd-6-929-2013, 2013

Menne, M.J., Durre, I., Vose, R.S., Gleason, B.E., and Houston, T.G.: An overview of the Global Historical Climatology Network-Daily Database. J. Atmos. Oceanic Technol., 29, 897–910, 2012.

Nash, J.E., and Sutcliffe, J.V.: River flow forecasting through conceptual models part I — a discussion of principles. J. Hydro., 10 (3): 282–290, 1970.

National Operational Hydrologic Remote Sensing Center (NOHRSC). Snow Data Assimilation System (SNODAS) Data Products at NSIDC, 2009-2011. National Snow and Ice Data Center, Boulder, Colorado, USA, 2004.

Parajka, J., Viglione, A., Rogger, M., Salinas, J.L., Sivapalan, M., and Bloeschl, G.: Hydrograph prediction in ungauged basins – a comparative assessment of studies. Geophysical Research Abstracts, 15, EGU2013-13126, 2013.

Perrin, C., Michel, C., and Andréassian, V.: Improvement of a parsimonious model for streamflow simulation. J. Hydro., 279 (1-4), 275–289, 2003.

Pietroniro, A., Fortin, V., Kouwen, N., Neal, C., Turcotte, R., Davison, B., Verseghy, D., Soulis, E.D., Caldwell, R., Evora, N. , and Pellerin, P.: Development of the MESH modelling system for hydrological ensemble forecasting of the Laurentian Great Lakes at the regional scale. Hydrol. Earth. Syst. Sc. ,11(4), 1279–1294, 2007.

Razavi, T., and Coulibaly, P.: Streamflow prediction in ungauged basins: review of regionalization methods. J. Hydrol. Eng., 18(8), 958–975, 2012.

Refsgaard, J.C., and Knudsen, J.: Operational validation and intercomparison of different types of hydrological models. Water Resources Research, 32(7), 2189–2202, 1996.

Shangguan, W., Dai, Y., Duan, Q., Liu, B., and Yuan, H.: A global soil dataset for earth system modeling. J. Adv. Model. Earth Syst., 6, 249–263, doi:10.1002/2013MS000293, 2014.

Sherman, L. K. Streamow from rainfall by the unit-graph method. Eng. News Record, 108, 501-505, 1932.

Singer, S.N., Cheng, C.K., and Scafe, M.G. (Eds.): The Hydrogeology of southern Ontario, second ed., Environmental monitoring and reporting branch, Ministry of the Environment, Toronto, ON., 240 pp. + appendices, 2003.

Sukovich, E. M., Ralph, F. M., Barthold, F. E., Reynolds, D. W., and Novak, D. R.: Extreme quantitative precipitation forecast performance at the Weather Prediction Center from 2001 to 2011. Weather and Forecasting, 29(4), 894–911, doi: 10.1175/WAF-D-13-00061.1, 2014.

Tolson, B. A., and Shoemaker, C. A.: Dynamically dimensioned search algorithm for computationally efficient watershed model calibration. Water Resources Research, 43(1), W01413 1-16, 2007.

Wagner, S., Fersch, B., Yuan, F., Yu, Z., and Kunstmann, H.: Fully coupled atmospheric-hydrological modelling at regional and long-term scales: Development, application, and analysis of WRF-HMS. Water Resources Research, 52(4), 3187–3211, doi: 10.1002/2015WR018185, 2016.

Wang, L., Riseng, C. M., Mason, L. A., Wehrly, K. E., Rutherford, E. S., McKenna, J. E., Castiglione, C., Johnson, L.B., Infante, D.M., Sowa, S. , and Robertson, M.: 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, 2015.

Wiley, M. J., Hyndman, D. W., Pijanowski, B. C., Kendall, A. D., Riseng, C., Rutherford, E. S., Cheng, S.T., Carlson, M.L., Tyler, J.A., Stevenson, R.J. , and Steen, P. J.: A multi-modeling approach to evaluating climate and land use change impacts in a Great Lakes River Basin. Hydrobiologia, 657(1), 243–262, 2010.

Figures

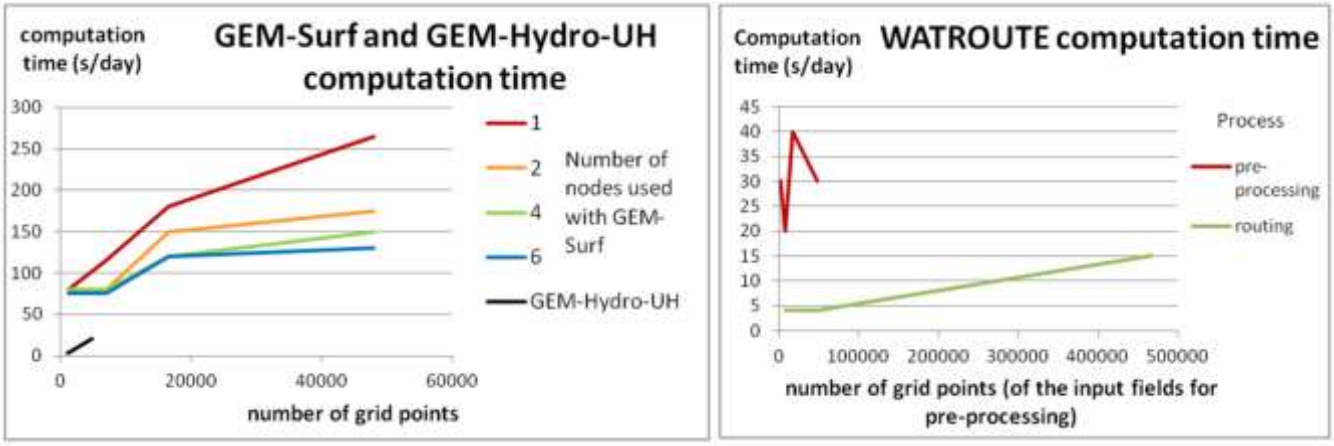


Figure 1: Computation time for GEM-Surf (Land-Surface part of GEM-Hydro), GEM-Hydro-UH, and WATROUTE. See text for details. The number of grid points in this study is 1276 (476000) for GEM-Surf/GEM-Hydro-UH (WATROUTE).

5

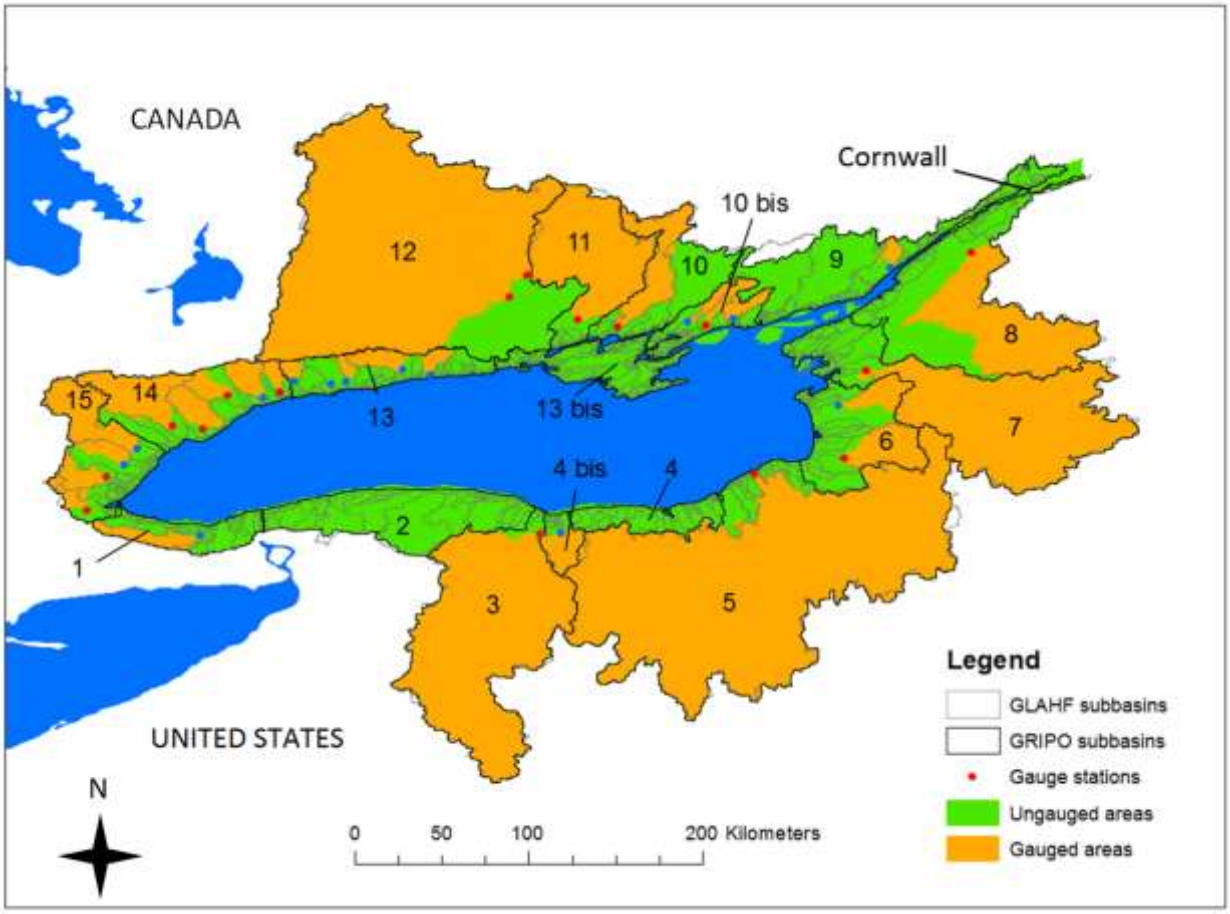


Figure 2: GRIP-O spatial framework: Lake Ontario subbasin delineation (GRIP-O subbasins). GLAHF subbasins are from the Great Lakes Aquatic Habitat Framework (Wang et al., 2015). Dots (blue for natural flow regimes and red for regulated regimes) are the most-downstream flow gauges (i.e., the main tributaries' gauges which are closest to Lake Ontario's shoreline) selected for model calibrations.

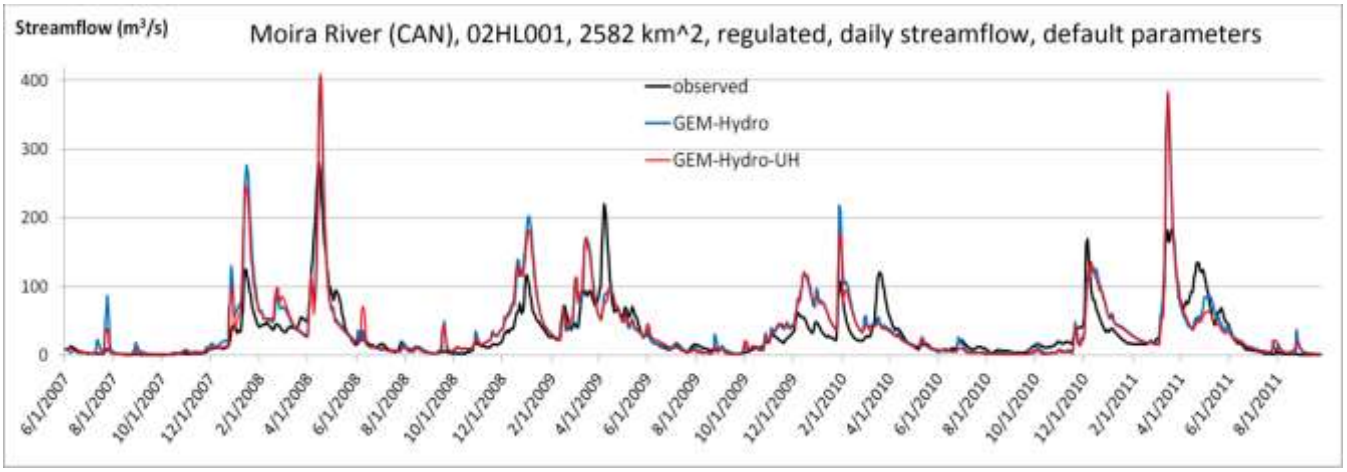


Figure 3: Hydrographs from uncalibrated GEM-Hydro and GEM-Hydro-UH (Moira River - subbasin 11).

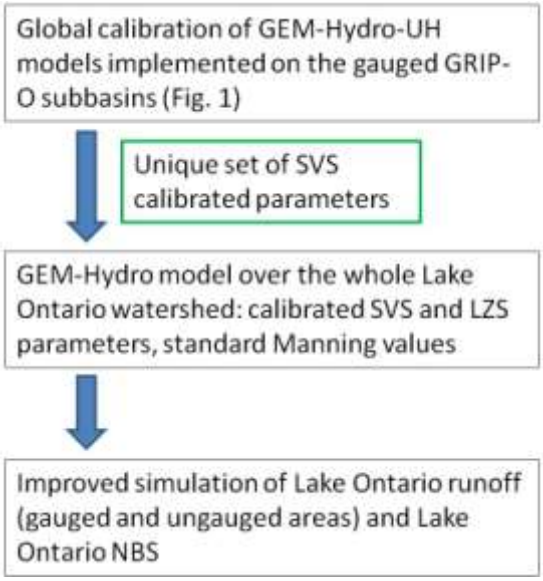


Figure 4: diagram summarizing the methodology employed to simulate Lake Ontario runoff with GEM-Hydro

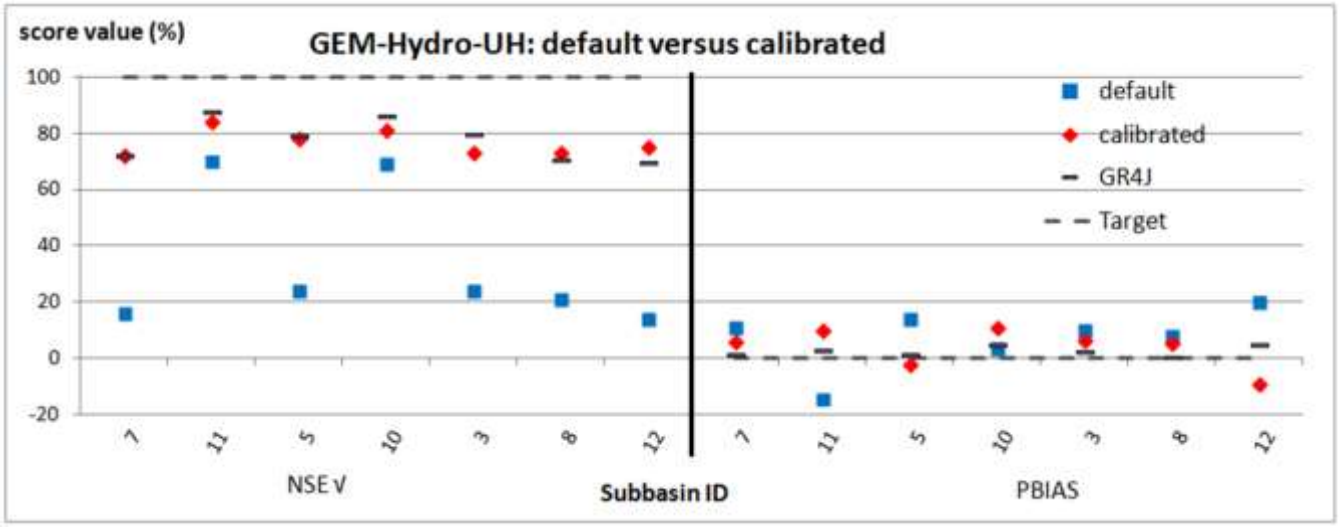


Figure 5: Uncalibrated and calibrated GEM-Hydro-UH performances over the calibration period. Results are presented as NSE √ (left) and PBIAS (right), for many GRIP-O subbasins. The grey dashed line shows perfect scores and GR4J reference is displayed with black markers.

5

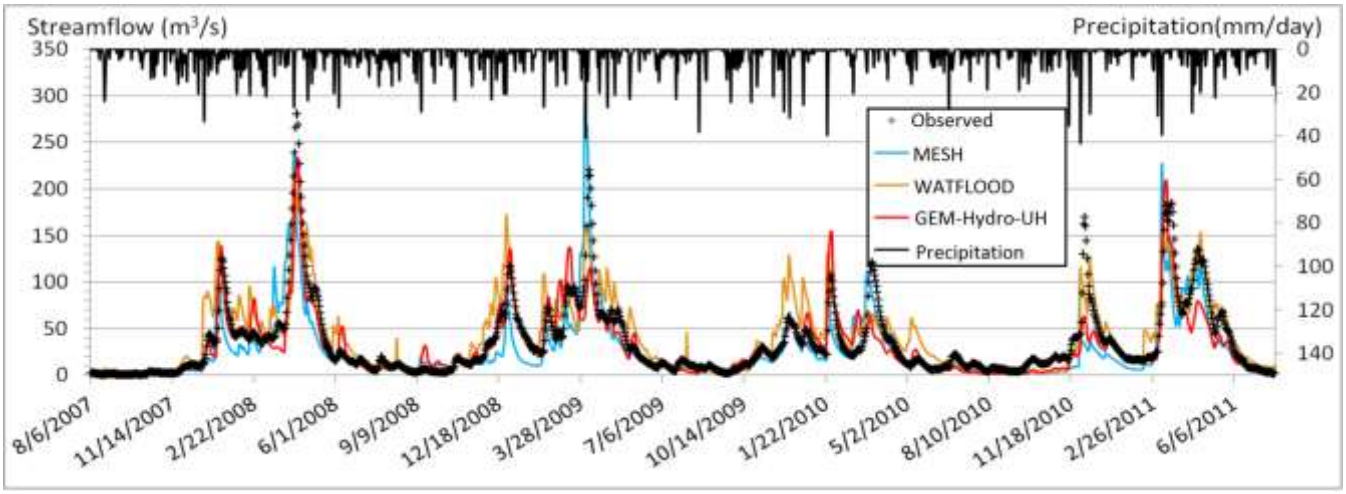


Figure 6: Intercomparison for the Moira River (calibration period, CaPA pecipitation).

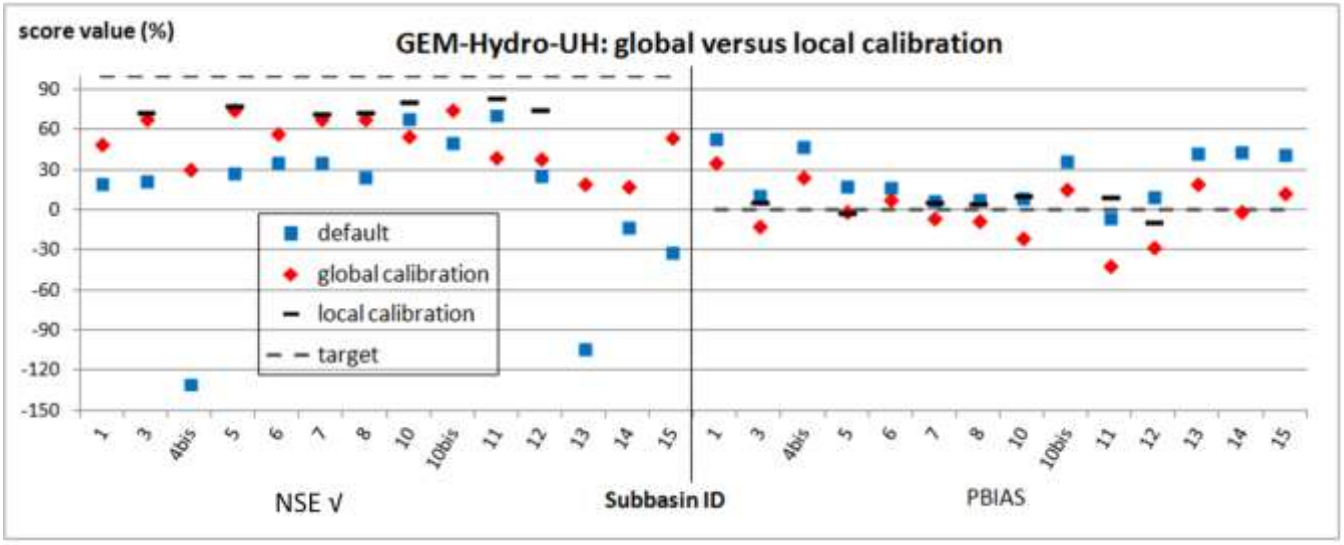


Figure 7: GEM-Hydro-UH performances in validation for the 14 GRIP-O gauged subbasins (Fig. 2) with default, locally, and globally-calibrated parameter values. Perfect scores are shown. Results are presented as NSE √ (left) and PBIAS (right).

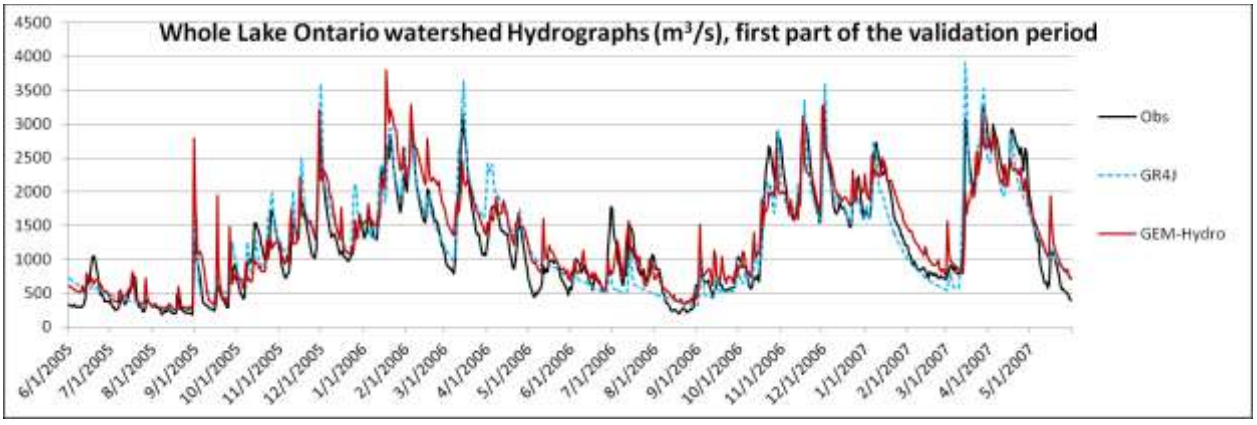


Figure 8: Lake Ontario basin runoff (including its ungauged areas, Fig. 2) for the validation period, comparing GR4J and GEM-Hydro.

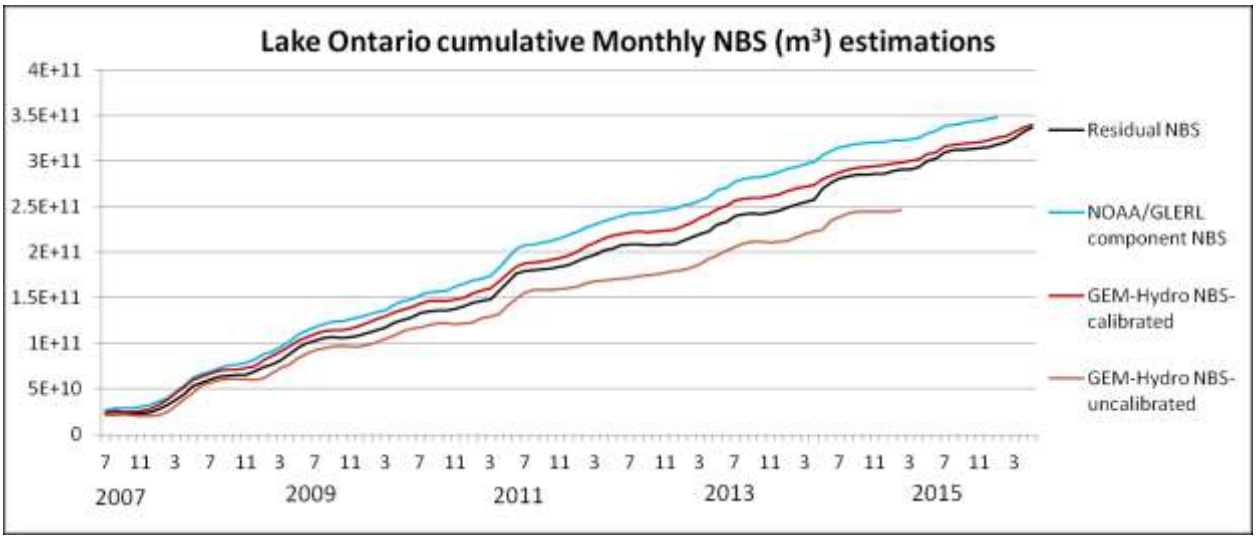


Figure 9: cumulative Lake Ontario NBS (Net Basin Supplies) estimates. Months are shown on the x-axis. See text for further details.

Tables

Table 1: Data sources; NA: North America

Dataset/origin	Type of data	Coverage	Resolution/scale	Source
GSDE	soil texture	Global	~ 1km (30")	Shangguan *et al.* 2014
GLOBCOVER 2009	land cover	Global	300m (10")	ESA 2009
HydroSheds	Flow directions	Global	~ 1km (30")	USGS and WWF 2006
SRTM	DEM	Global	90m (3")	NGA and NASA 2000

HyDAT	Gauge stations	CAN	N/A	ECCC
NWIS	Gauge stations	US	N/A	USGS
CaPA v2.4b8	Precipitation	NA	~ 15 km	ECCC
RDPS	Atmospheric forcings	NA	15/10 km	ECCC

Table 2: Information on GEM-Hydro-UH 16 free parameters; LZS: Lower Zone Storage; coeff. : coefficient; mult. : multiplicative; precip. : precipitation; param.: parameter; min.: minimum; max.: maximum.

Param. \ range	description	initial	Min.	Max.	Param. \ range	description	initial	Min.	Max.
HU_decay	response time (h)	60.0	20.0	400.0	LAI	Leaf-Area Index mult. coeff.	1.0	0.2	5.0
FLZCOEFF	LZS mult. coeff.	1.0E-05	1.0E-07	1.0E-04	Z0M	roughness length mult. coeff.	1.0	0.2	5.0
PWR	LZS exponent coeff.	2.8	1.0	5.0	TBOU	boundary between liquid and solid precip. (˚C.)	0.0	-1.0	1.5
MLT	coeff. To divide snowmelt amount	1.0	0.5	2.0	EVMO	evaporation resistance mult. coeff.	1.0	0.1	10.0
GRKM	Horizontal conductivity mult. coeff.	1.0	0.1	30.0	KVMO	vertical conductivity mult. coeff.	1.0	0.1	30.0
SOLD	soil depth (m)	1.4	0.9	6.0	PSMO	soil water suction mult. coeff.	1.0	0.1	10.0
ALB	albedo mult. coeff.	1.0	0.2	5.0	BMOD	slope of retention curve mult. coeff.	1.0	0.1	10.0
RTD	root depth mult. Coeff.	1.0	0.2	5.0	WMOD	threshold soil moisture contents mult. coeff.	1.0	0.1	10.0

5

Table 3: GRIP-O subbasins characteristics.

country	Subbasin #	Station	%_gauged	Area(km^2)	Flow regime	mean elev. (m)
CAN	1	20_mile	N/A	307	natural	198
USA	3	Genessee	N/A	6317	regulated	418
USA	4bis	Irondequoit	N/A	326	natural	172
USA	5	Oswego	N/A	13287	regulated	259
USA	6	N/A	40	2406	mixed	264
USA	7	Black River	N/A	4847	regulated	471
USA	8	Oswegatchie	N/A	2543	regulated	250
CAN	10	Salmon_CA	N/A	912	regulated	196
CAN	10bis	N/A	44.2	944	mixed	115
CAN	11	Moira	N/A	2582	regulated	228
CAN	12	N/A	88	12515.5	regulated	282
CAN	13	N/A	40.3	1537.5	natural	178
CAN	14	N/A	61.3	2689.4	mixed	209
CAN	15	N/A	63	2245.8	mixed	263

Table 4: Final parameter values or ranges after calibration; for global calibration, HU_decay consists of a multiplicative coefficient. See Table 2 for parameter definition.

		HU_decay	FLZCOEFF	PWR	MLT	GRKM	SOLD	ALB	RTD
global calibration		0.5 (mult)	7.1E-07	2.3	0.7	6.7	0.9	1.0	3.7
local calibration range	min	46.0	1.4E-07	1.1	0.4	1.5	0.9	0.4	1.1
	max	142.7	8.5E-05	4.2	1.5	13.1	4.6	2.0	3.9
		LAI	Z0M	TBOU	EVMO	KVMO	PSMO	BMOD	WMOD
global calibration		1.9	3.9	0.4	1.8	2.9	1.5	0.7	1.4
local calibration range	min	0.6	0.2	-0.9	0.6	1.0	1.2	0.6	0.6
	max	4.6	3.8	0.5	3.5	9.4	9.4	1.5	2.8

Table 5: performances for the GRIP-O gauged area and the whole Lake Ontario basin (Fig. 2) with GR4J and globally-calibrated GEM-Hydro-UH and GEM-Hydro models. Cal., val.: calibration and validation periods, respectively.

	GRIPO gauged area: 53459.2 km^2						Lake Ontario basin: 68214.8 km^2			
	GR4J		GEM-Hydro-UH		GEM-Hydro		GR4J		GEM-Hydro	
Scores (%)	cal	val	cal	val	cal	val	cal	val	cal	val
NSE	82.4	84.6	80.1	83.4	79.8	80.5	82.9	85.5	81.8	82.0
NSE √	84.7	85.5	83.0	86.6	78.5	82.4	84.4	85.0	80.5	83.7
NSE Ln	83.3	84.0	82.1	87.2	74.4	82.3	82.4	82.8	76.8	83.7
PBIAS	-0.3	1.5	-9.0	-8.1	-13.1	-10.9	-2.2	-1.2	-10.3	-8.2

5

Supplementary material: intercomparison of MESH, WATFLOOD, and GEM-Hydro-UH

1.~~1~~ Models

Three different platforms are compared in this study: MESH, WATFLOOD, and GEM-Hydro. They have in common a distributed representation of most hydrological processes occurring in a basin and a structure organized around two main components: a LSS for the representation of surface processes (evapotranspiration, infiltration, snow processes, water circulation in the soils), and a river routing scheme for simulating water transport in the streams, which consists of WATROUTE for all models. WATROUTE is a 1-D hydraulic model relying mainly on flow directions and elevation data (Kouwen 2010). It routes to the ~~catchment~~ basin outlet the surface runoff and recharge produced by the surface schemes. In WATROUTE, runoff directly feeds the streams while recharge can be provided to an optional Lower Zone Storage (LZS) compartment, representing superficial aquifers, which releases water to the streams. WATFLOOD and GEM-Hydro make use of the LZS, whereas recharge from MESH feeds directly into the stream.

The version of MESH used in this study relies on version 3.6 of the Canadian LAnd Surface Scheme (CLASS). Each grid cell is subdivided in a number of tiles, and each tile is classified as belonging to one of the five grouped response units (GRUs), based on its land-use/soil type combination. In this paper, we follow the local calibration strategy advocated by Haghnegahdar et al. (2014) for MESH (see section on calibration strategy).

GEM-Hydro is very similar to MESH, but is tied to the LSSs available in GEM: ISBA and SVS. A previous study on the same ~~watershed~~ basin demonstrated the clear superiority of SVS over ISBA, especially in regard to the baseflow component of the streamflow (see Gaborit et al., 2016 b). We thus only use SVS with GEM-Hydro in this paper.

WATFLOOD (Kouwen, 2010) is a distributed model of intermediate complexity that only needs precipitation and temperature as forcing, as opposed to MESH and GEM-Hydro which need additional atmospheric variables (Table 1). It relies on the GRUs concept and on many empirical equations. WATFLOOD has been employed by Pietroniro et al. (2007) over the Great Lakes ~~watershed~~basin.

In this project, WATFLOOD and MESH are implemented with a 10 arcmin (≈ 20 km) spatial resolution (both for their LSS and routing schemes), while GEM-Hydro is implemented with a 10 arcmin resolution for the LSS and 0.5 arcmin (≈ 1 km) for the routing. Sensitivity tests (Gaborit et al., 2016 b) revealed that 2 and 10 arcmin resolutions for SVS lead to quite similar performance in terms of streamflow at the outlet, while a substantial amount of computation~~al~~ time is saved when running the coarser resolution (almost proportionally if using the same number of nodes). The same was shown for WATROUTE which produces outputs of similar quality be it implemented at a low (10 arcmin for MESH and WATFLOOD) or high (0.5 arcmin with GEM-Hydro) resolution, as long as results are evaluated for large enough catchments (i.e., catchments which spread over at least a few grid cells). However, the high-resolution WATROUTE version is preferred in GEM-Hydro for consistency with the WCPS-GLS (Durnford et al., submitted) recently developed at ECCC. Hence, the higher resolution GEM-Hydro's routing scheme is not expected to give GEM-Hydro any advantage in comparison to MESH and WATFLOOD.

The internal time-step used for GEM-Hydro is 10 minutes, which slightly improves streamflow simulations in comparison to a 30 min. time-step (see Gaborit et al., 2016 b). Further reducing it does not improve the results. The internal time-steps used for MESH and WATFLOOD are respectively equal to 30 and 60 minutes. The internal time-step of a model is generally maximized up to the desired output interval, provided that it satisfies numerical stability. In the GEM-Hydro version used in this study, a 10-min. time-step was required to achieve numerical stability, but a newer version now allows to increase it. Table 1 summarizes the main specificities of the models and the required forcing data. Table 2 shows the datasets used for physiographic information.

The physiographic data required by the distributed models under study consist of soil texture, land use / land cover, Digital Elevation Model (DEM), and flow direction grids. Table 2 lists the datasets used to provide the physiographic and atmospheric inputs required by the models. 26 land cover classes are defined in GEM-Hydro, while WATFLOOD and MESH rely only on 7 of them, which are aggregations of GEM-Hydro classes. Soil textures are from the Global Soil Dataset for Earth system modeling (GSDE; Shangguan et al., 2014), which contains information down to 2.8 m. However, soil texture is calibrated for MESH (Table 5). Soil texture was not calibrated for GEM-Hydro-UH, but some hydraulic parameters, which are derived from soil texture, were calibrated (Table 3). WATFLOOD does not need soil texture information ~~(Table 2)~~. By default, the maximum soil depth was set to 1.4 m in GEM-Hydro (for the area under study), 4.1 m in MESH, and is not defined in WATFLOOD. The maximum soil depth is calibrated in GEM-Hydro and MESH (Table 3 to Table 5). The parameter ranges of Tables 3-5 were generally chosen as wide as possible while remaining physically realistic, in order to let more freedom to the optimization algorithm, which may a priori increase the chances of finding optimal parameter sets during calibration.

2~~1.3~~ Calibration strategy

Different paradigms were used to calibrate them. GEM-Hydro-UH was calibrated using multiplicative coefficients that adjust the spatially-varying values of a given parameter, leading to a reasonable number of free parameters (16) while preserving spatial variability. MESH was implemented calibrating the 12 free parameters of its 5 different GRUs in an independent manner, thus resulting in 60 free parameters. WATFLOOD had the lowest number of free parameters during calibration, and involved calibrating parameter values which are valid for the entire subbasin (no spatial variability) or for one of the three main land cover types considered inside the model, i.e. bare ground, snow covered ground, or other grounds (Table 4).

It is important to emphasize that the approach used to calibrate GEM-Hydro may result in unrealistic values for some parameters, as the multiplicative coefficients could bring them beyond the range of physical coherence. More precisely, soil water content thresholds and albedo (Table 3) cannot be higher than 1. Therefore, these values were constrained to realistic ranges after they were adjusted by the calibration algorithm by imposing them a minimum value of 0 and a maximum of 1.

The initial parameter values were either set to default ones that generally provide satisfactory results for the model (GEM-Hydro-UH, Table 3) or to random values (WATFLOOD, MESH). The number of maximum model runs allowed depends on the model being used. For example, 400 runs revealed sufficient for GEM-Hydro-UH (Sect. 2.2) in the sense that no significant performance improvement was achieved beyond. This is because the number of GEM-Hydro-UH free parameters is relatively low (16, Table 3). The DDS algorithm is very efficient in the sense that it adjusts the search behavior to the maximum number of objective function evaluations (model runs) in order to converge to good quality solutions (Tolson and Shoemaker, 2007). The similarity of the performances obtained with GR4J and GEM-Hydro-UH (Fig. 3 in main document~~4~~) supports the choice of the methodology used here, as GR4J was implemented with a maximum of 2000 model runs, three distinct calibration trials, and had an even lower number of free parameters (6, see Gaborit et al., 2016 a).

A maximum of 1000 model runs was used to calibrate MESH and of 1500 for WATFLOOD. Finally, the calibration strategy used for MESH consists of an improved and reliable strategy based on the work of Haghnegahdar et al. (2014). Despite the random initial values used for MESH and WATFLOOD, only one calibration trial was performed for each of the models on a given subbasin. Even though the three models studied here were not calibrated using the same number of free parameters and the same maximum allowed model runs, it is assumed that the calibration strategies employed allow each model to come very close to its optimal performance for a given subbasin and the time period considered. Indeed, the strategy used for each of the three models is the result of expert knowledge and always involves parameters affecting the whole range of the main hydrological processes, i.e. evaporation, snowmelt, infiltration, soil transfer, and time to peak (channel friction). It is thus logical to use different strategies for each of the models as these do not involve the same parameters, land use classification, or even physical processes. The most important methodological consistencies for achieving a fair comparison between models include, in our view, a common calibration algorithm and objective function, along with common physiographic and forcing data.

~~2.2~~3 Results~~Inter-comparison of all models~~

This section aims at comparing MESH, WATFLOOD, and GEM-Hydro-UH performance values. The calibration strategy used for each of them is described in Sect. 1.3. Note that MESH was only calibrated on the Moira and Black Rivers, and WATFLOOD on the Moira, Black, and Salmon Rivers. Calibration and validation performances are presented in Fig. 1~~5~~ and calibrated hydrographs, in Fig. 2~~6~~.

It was deemed uninformative to present the calibrated parameter values since they are highly location dependant and subject to the equifinality issue (see previous section). Table 4~~7~~ of the main document however highlights the final parameter ranges for GEM-Hydro-UH. Overall, GEM-Hydro-UH outperforms MESH and WATFLOOD, both in calibration and validation (Fig. 1~~5~~). The robustness of the models is generally quite good, but less so for MESH on the Black River (subbasin 7 in Fig. 1~~5~~).

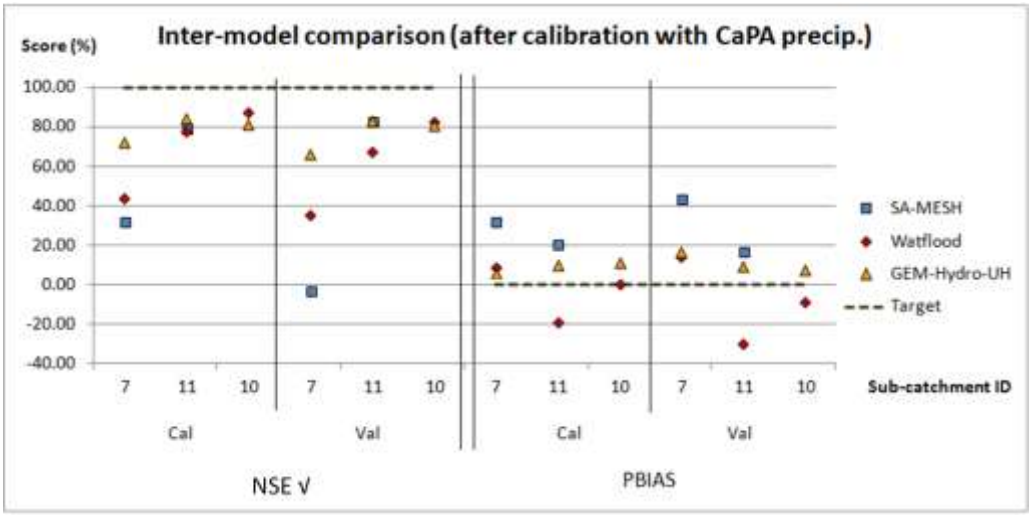
When looking closely at the Moira River hydrographs (Fig. 2~~6~~), important differences arise between the models. For instance, WATFLOOD has a more flashy behavior and tends to overestimate peak flow events, MESH generally underestimates flows, and GEM-Hydro-UH lays somewhere in between. Peak flow events (even for other subbasins) associated to the spring freshet are generally better represented by MESH, which may be due to a better representation by CLASS of various cold regions hydrological processes, such as snow accumulation and melt, snow interception by vegetation, as well as soil freezing and thawing.

It is possible that the differences in model performance may be explained by the different calibration strategies used for each model, and that better performances could be obtained with MESH and WATFLOOD for these watersheds, although the calibration details were in each case determined by an expert user of each model. The optimal calibration strategy, as well as the number of free parameters, could be revisited for each model in order to see if this explains the above differences, but this is quite beyond the scope of the paper.

Even if the intercomparison is obviously limited in the number of available test cases, it allows highlighting the mandatory need of calibrating hydrologic models, that models have unique behaviors that translate in substantial differences in hydrographs, and that each of the models could benefit from some strengths of its competitors. For example, SVS would likely benefit from the implementation of the soil freezing and melting processes that are present in CLASS.

Results however strongly indicate that SVS can compete with more established Canadian models for simulating streamflow. In the coming years, after SVS becomes operationally implemented within ECCC's GEM-based NWP systems, it will be possible to obtain useful streamflow predictions by simply post-processing the runoff output from GEM using a unit hydrograph, or by routing these time series using a more sophisticated routing scheme.

Many distributed models do exist worldwide, each one possessing its own advantages and drawbacks, but also its own optimal implementation and calibration methodology, which makes a perfectly fair inter-comparison quite challenging, if not unrealistic.



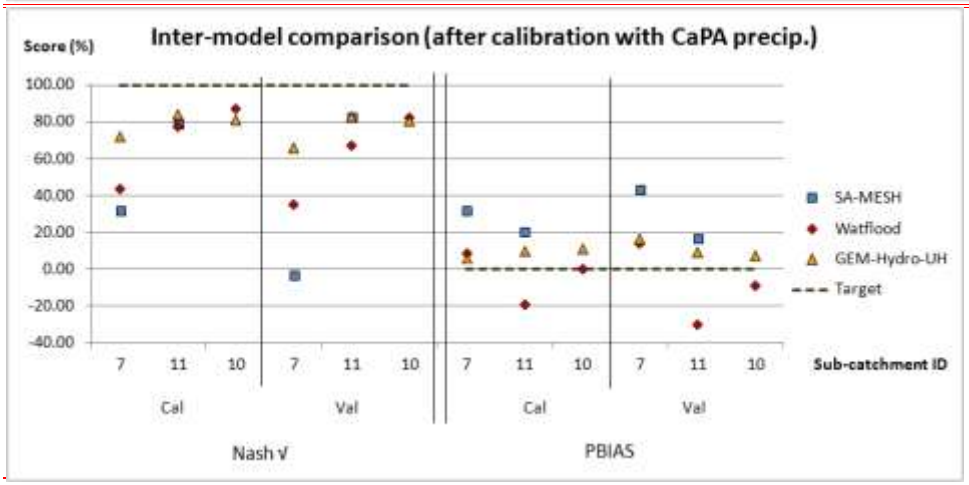


Figure 15: Intercomparison for three GRIP-O subbasins (Table 36 in main document). MESH was not implemented on subbasin 10. Cal, Val: calibration and validation periods, respectively. Scores that would be achieved if models provided a perfect fit to observations are indicated by the dashed line and labelled "Target".

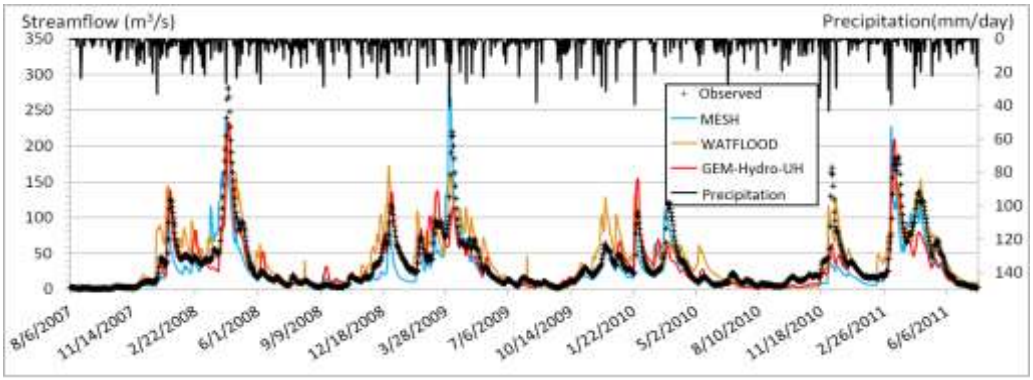


Figure 26: Intercomparison for the Moira River (calibration period, CaPA pecipitation).

Table 1: Data requirements and model specificities. P: precipitation, T: temperature, H: humidity, R:, radiative forcings (short- and long-wave incoming radiations), W: wind, Ps: pressure; LULC: Land Use / Land Cover, Topo: elevation data, Flow Dir: flow directions. Brackets indicate time-step used in this study.

Model name	Underlying theory	Spatial distribution	Time-step [min]	Forcing data	Physiographic data
WATFLOOD	Physical/Conceptual	Semi-distributed	Flexible [60]	P, T	LULC, Topo, Flow Dir
GEM-Hydro	Physical	Semi-distributed	Flexible [10]	P, T, H, R, W, Ps	LULC , Soil, Topo, Flow Dir
MESH	Physical	Semi-distributed	Flexible [30]	P, T, H, R, W, Ps	LULC, Soil, Topo, Flow Dir

Table 2: Data sources; NA: North America

Dataset/origin	Type of data	Coverage	Resolution/scale	Source
GSDE	soil texture	Global	~ 1km (30")	Shangguan *et al.* 2014
GLOBCOVER 2009	land cover	Global	300m (10")	ESA 2009

HydroSheds	Flow directions	Global	~ 1km (30")	USGS and WWF 2006
SRTM	DEM	Global	90m (3")	NGA and NASA 2000
HyDAT	Gauge stations	CAN	N/A	ECCC
NWIS	Gauge stations	US	N/A	USGS
CaPA v2.4b8	Precipitation	NA	~ 15 km	ECCC
RDPS	Atmospheric forcings	NA	15/10 km	ECCC

Table 3: Information on GEM-Hydro-UH 16 free parameters: LZS: Lower Zone Storage; coeff. : coefficient; mult. : multiplicative; precip. : precipitation; param.: parameter; min.: minimum; max.: maximum.

Param. \ range	description	initial	Min.	Max.	Param. \ range	description	initial	Min.	Max.
HU_decay	response time (h)	60.0	20.0	400.0	LAI	Leaf-Area Index mult. coeff.	1.0	0.2	5.0
FLZCOEFF	LZS mult. coeff.	1.0E-05	1.0E-07	1.0E-04	Z0M	roughness length mult. coeff.	1.0	0.2	5.0
PWR	LZS exponent coeff.	2.8	1.0	5.0	TBOU	boundary between liquid and solid precip. (˚C.)	0.0	-1.0	1.5
MLT	coeff. To divide snowmelt amount	1.0	0.5	2.0	EVMO	evaporation resistance mult. coeff.	1.0	0.1	10.0
GRKM	Horizontal conductivity mult. coeff.	1.0	0.1	30.0	KVMO	vertical conductivity mult. coeff.	1.0	0.1	30.0
SOLD	soil depth (m)	1.4	0.9	6.0	PSMO	soil water suction mult. coeff.	1.0	0.1	10.0

ALB	albedo mult. coeff.	1.0	0.2	5.0	BMOD	slope of retention curve mult. coeff.	1.0	0.1	10.0
RTD	root depth mult. Coeff.	1.0	0.2	5.0	WMOD	threshold soil moisture contents mult. coeff.	1.0	0.1	10.0

Table 44: Information on WATFLOOD 14 free parameters; LZS: Lower Zone Storage; coeff. : coefficient; mult. : multiplicative.

parameter	minimum	maximum	parameter	minimum	maximum
channel Manning's N	0.01	1.0	upper zone retention (mm)	1.0	300.0
LZS mult. coeff.	1.0E-09	1.0E-05	infiltration coefficient bare ground	0.8	0.99
LZS exponent coeff.	2.0	3.0	infiltration coefficient snow covered ground	0.8	0.99
melt factor (mm/dC/hour)	0.1	3.0	overland flow roughness coefficient bare ground	1.0	75.0
interflow coefficient	1.0	100.0	overland flow roughness coefficient snow covered ground	1.0	75.0
interflow coefficient bare ground	1.0	200.0	Interception evaporation factor	0.1	75.0
interflow coefficient snow covered ground	1.0	200.0	base temperature (dC)	-3.0	3.0

Table ~~5~~5: Information on MESH 60 free parameters: independent values are sought for each of the 5 model Grouped Response Units (GRUs; source: Haghnegahdar, 2015).

parameter	description	vegetation or river class (5)	minimum	maximum
ROOT	Annual maximum rooting depth of vegetation category [m]	crop and grass	0.2	1.0
		Forest	1.0	3.5
RSMN	Minimum stomatal resistance of vegetation category [$s.m^{-1}$]	Crop	60.0	110.0
		Grass	75.0	125.0
		Forest	100.0	150.0
VPDA	Vapour pressure deficit coefficient	All	0.5	1.0
SDEP	Soil permeable (Bedrock) depth [m]	All	0.35	4.1
DDEN	Drainage density [km/km^2]	All	2.0	100.0
SAND	Percent sand content [%]	All	0.0	100.0
CLAY	Percent clay content [%]	All	0.0	100.0
RATIO	The ratio of horizontal to vertical saturated hydraulic conductivity	All	2.0	100.0
ZSNL	Limiting snow depth below which coverage is less than 100% [m]	All	0.05	1.0
ZPLS	maximum water ponding depth for snow-covered areas [m]	All	0.02	0.15
ZPLG	maximum water ponding depth for snow-free areas [m]	All	0.02	0.15
WFR2	Channel roughness factor	All	0.02	2.0

Mis en forme : Largeur : 11", Haut