Great Lakes Runoff Inter-comparison Project, Phase 2: Lake Ontario (GRIP-O)

Étienne Gaborit[a,*], Vincent Fortin[a], Bryan Tolson[b], Lauren Fry[c], Tim Hunter[d], and Andrew D. Gronewold[d]

[a] Environment Canada, Environmental Numerical Prediction Research (E-NPR), 2121 transcanadian highway, Dorval H9P1J3, QC, Canada.
[b] University of Waterloo, Civil and Environmental Engineering Dpt., Waterloo N2L3G1, ON., Canada.
[c] U.S. Army Corps of Engineers, Detroit District, Great Lakes Hydraulics and Hydrology Office, 477 Michigan Ave., Detroit 48226, MI,U.S.A.
[d] NOAA Great Lakes Environmental Research Laboratory (GLERL), 4840 S. State Rd. , Ann Arbor 48108, MI.,U.S.A.

* corresponding author.
E-mail address: etienne.gaborit@canada.ca
Postal address: office 453, Environment Canada, 2121 transcanadian highway, Dorval H9P1J3, QC, Canada.
Phone: 1-514-421-5305

Abstract

The Great Lakes Runoff Inter-comparison Project for Lake Ontario (GRIP-O) aims to compare different hydrologic models, using the same settings, in their ability to estimate runoff for the Lake Ontario watershed. The watershed is challenging because many of its tributaries have a regulated flow regime and a significant part remains ungauged. GRIP-O follows the GRIP-M project which focused on Lake Michigan. It involves a comparison between two different sources of precipitation data (CaPA- Canadian Precipitation Analysis and the GHCND- Global Historical Climatology Network - Daily), and focuses here on two lumped models, GR4J (*modèle du Génie Rural à 4 paramètres Journalier*) and LBRM (Large Basin Runoff Model).

Results indicate that both models perform very well, with GR4J performing slightly better than LBRM and the GHCND precipitation dataset resulting in better simulations than CaPA, for this area. Performances are, however, always very satisfactory whatever the combination of model / precipitation data used, even for regulated catchments, and do not show any clear correlation to any of the catchments' properties studied here. Results also tend to confirm that the Area-Ratio Method is appropriate for extrapolating flows from the gauged part of a catchment to the whole catchment including its ungauged parts, as demonstrated in GRIP-M.

Key words: lumped models, regulated watersheds, runoff estimation, Canadian Precipitation Analysis (CaPA), Dynamically Dimensioned Search (DDS) algorithm, local calibration

Introduction

As a freshwater resource, the Laurentian Great Lakes provide many ecosystem services including navigation, potable water supplies, recreational benefits, etc. Associated environmental issues include pollution, erosion, climate change impacts, and flooding. Integrating these considerations to optimize management practices requires the implementation of reliable environmental models. In the Great Lakes, the main physical processes related to the water cycle are stream runoff, over-lake precipitation and evaporation, inter-lake and St. Lawrence channel routing, and groundwater processes. A cascade of distinct models is generally applied to simulate Great Lakes' water levels. Examples are the National Oceanic and Atmospheric Administration (NOAA) Great Lakes Environmental Research Laboratory (GLERL) Advanced Hydrologic Prediction System (AHPS; Gronewold et al., 2011) and Environment and Climate Change Canada's (ECCC) Water Cycle Prediction System (WCPS) for water levels and thermo-dynamics of the Great Lakes (Fortin, ECCC 2016, personal communication). Another good example of interconnected models which simulate many physical processes is provided by Wiley et al. (2010). The study presented here focuses on runoff from terrestrial areas conveyed to the lakes through their tributaries. The goal is to simulate stream flows through hydrologic models.

In the field of hydrologic modeling, simulations are mainly performed by data-driven models (Razavi and Coulibaly, 2012), lumped conceptual models, or (semi-)distributed physical models. Models of the first two categories are relatively easy to implement and generally require few input data and less powerful computational resources, but they are based on conceptual and empirical equations and can represent stream flows only at one location (i.e., one point in a stream). Distributed or semi-distributed physical models, however, aim to represent processes physically (although they often rely on many empirical equations) and provide spatial details for several physical variables. Distributed physical models generally require much more input data and computational time than lumped models.

Many numerical models have been applied to simulate runoff in the Great Lakes watershed (Coon et al., 2011). Applications range from assessing climate change impacts on lake levels (Angel and Kunkel, 2010; Chao, 1999; Croley, 2003; MacKay and Seglenieks, 2013), land-use change effects on water quality (Mao and Cherkauer, 2009; Wiley et al., 2010), short- to mid-term runoff and lake level forecasts (Croley and Lee, 1993; Croley and Hartmann, 1987; Gronewold et al., 2011), or the comparison between different hydrologic platforms and associated implementation strategies (Deacu et al., 2012; Fry et al., 2014; Haghnegahdar et al., 2014). Based on these studies, some rainfall-runoff models provide reliable runoff simulations for Great Lakes watersheds. Lumped models, such as the NOAA GLERL Large Basin Runoff Model (LBRM; Croley and He, 2002) and the NOAA National Weather Service (NWS) model (Burnash, 1995), and distributed physical models such as the MESH (Modélisation Environnementale – Surface and Hydrology; Pietroniro et al., 2007), Watflood (Kouwen, 2010), and the Precipitation-Runoff Modeling System (PRMS; Hay et al., 2011) are examples. In the Lake Ontario watershed, Croley (1983) and Haghnegahdar et al. (2014) calibrated hydrologic models with runoff observations to optimize simulations of this variable. The former demonstrated that the LBRM worked well for simulating weekly flows, while the second evaluated the MESH model on 15 Great Lakes subbasins, including two Lake Ontario subbasins. However, even after calibration, the MESH model did not perform particularly well for validation catchments (Nash-Sutcliffe values below 0.6).

The Great Lakes Runoff Inter-comparison Project for Lake Ontario (GRIP-O) involves a comparison between two lumped models, namely GR4J (modèle du Génie Rural à 4 paramètres Journalier; Perrin et al., 2003) and LBRM. Although GR4J has not been applied yet in the Great Lakes watershed, it was successfully applied in many different climatic conditions (e.g., Pagano et al., 2010; Seiller et al., 2012). LBRM is the reference lumped model for the Great-Lakes, as it is the only model that simulates runoff across the entire Great Lakes basin for operational seasonal water budget forecasting applications. The inter-comparison between distributed models will be addressed in a forthcoming paper (Gaborit et al., in preparation). It includes a comparison with the best performing lumped models from the present study.

In addition to the above comparisons, two different sources for precipitation are compared via their resulting hydrologic performances, namely the Global Historical Climatology Network- Daily (GHCND, a collection of ground observations; Menne et al., 2012) and the Canadian Precipitation Analysis (CaPA) datasets. CaPA has been evaluated using ground measurements of precipitation and generally provided good predictions (Fortin et al., 2015; Lespinas et al., 2015; Mahfouf et al., 2007). CaPA was, however, only compared once to observed precipitation in terms of resulting hydrologic performances (Eum et al., 2014) for a catchment in the Canadian Rocky Mountains and with the Variable Infiltration Capacity model (VIC; Mao and Cherkauer, 2009). Deacu et al. (2012) compared CaPA precipitation and model precipitation from the Canadian Global Environmental Multiscale (GEM) model in terms of the resulting impacts on the Great Lakes' Net Basin Supply simulations (NBS). A Great Lake's NBS is the sum of over-lake precipitation and runoff brought from its watershed, minus over-lake evaporation, and does not take into account streamflow coming from an upstream Great Lake (DeMarchi et al., 2009). Deacu et al. (2012) found that the Great Lakes' NBS was more accurately simulated with model precipitation. However, NBS is the result of many processes with a high level of associated uncertainty, making it difficult to accurately assess the quality of precipitation data in this case.

Given the variety and number of existing models (Coon et al., 2011) and the need to better understand the potential and limitations of hydrologic modeling (Gronewold and Fortin, 2012), a comparison of different runoff models is of interest. Each model has advantages not only of simulation performances or possible applications but also of requirements. The Great Lakes Runoff Inter-comparison Project for Lake Ontario (GRIP-O) further pursues a comparison of runoff modeling tools for the Great Lakes watershed started by Fry et al. (2014). Fry et al. (2014) compared various models in their ability to simulate historical runoff for the Lake Michigan watershed by incorporating different calibration frameworks and input data. The GRIP-O compares different models’ ability to simulate daily runoff from the Lake Ontario watershed (Fig. 1). To achieve a fair comparison, the exact same forcings and calibration framework were used in our simulations. To better manage the freshwater resources of Lake Ontario, the ultimate aim of the GRIP-O is the development of a reliable and efficient daily runoff simulation platform for many environmental applications such as flood alerts and lake-level forecasts.

Site Description

The total watershed area (Fig. 1) is about 83,000 km^2 of which around 19,000 km^2 correspond to the lake surface. Although Lake Ontario receives water from the upstream Great Lakes through the Niagara River, its watershed is defined as the land part which drains directly into the lake; that is, not taking into account the areas which drain in the upstream Great Lakes. This is to focus on the runoff component of Lake Ontario NBS, which does not take into account water provided to the lake by the upstream Great Lakes (DeMarchi et al., 2009). The U.S./Canada border follows the Niagara River, the middle of Lake Ontario, and the St.-Lawrence River down to Cornwall, ON, which consists of the outlet of the Lake Ontario watershed as it is defined here. Major cities inside the catchment include Toronto (subbasin 14), Hamilton (subbasin 15), Rochester (subbasin 3/4bis), Syracuse (subbasin 5), and Kingston (subbasin 9), for a total of about 11 million inhabitants (9 in Canada, 2 on the U.S. side). Apart from the cities, the catchment is mainly rural (agriculture, pasture, forest).

Methods

Models

GR4J is a daily continuous lumped hydrologic model with four free parameters (Perrin et al., 2003). It basically relies on two tanks which represent the soil and routing reservoirs of a catchment and on the Unit Hydrograph theory to produce stream flows. It was coupled with the two free-parameter CémaNeige snow module (Nicolle et al., 2011; Valéry, 2010).

The LBRM is also a daily continuous lumped hydrologic model; it includes a snow module and has a total of nine free parameters by default. Originally developed by NOAA-GLERL, it simulates water transport through a series of cascading tanks (Croley and He, 2002). LBRM has been employed in a variety of research-oriented and operational applications, ranging from hydrodynamic modeling studies (Anderson et al., 2010) to Great Lakes water-level forecasting systems (Gronewold et al., 2011). Both models require daily watershed averages of precipitation and of maximum and minimum temperature, as well as the catchment area. GR4J also requires the mean watershed latitude, used in the potential evapotranspiration formulation, and the mean watershed elevation (up to five elevation classes can be defined in the Cémaneige snow module). It was implemented here using a unique elevation class for each GRIP-O subbasin.

Spatial framework and study area characteristics

The spatial framework for the GRIP-O (Fig. 1) is slightly different from the one made by Croley (1983) for the calibration of the LBRM model. We subdivided some subbasins (4 and 4bis, 10 and 10bis, 13 and 13bis; Fig. 1) in order for the data of the flow gauges to better represent the total subbasin outflow in the case of calibration scheme 1 (i.e., for subbasins containing several flow stations). However, this leaves some areas as ungauged, such as subbasins 2, 4, 13bis, and 9. Subbasin 9 is considered ungauged because of the small area covered by the gauge, despite the fact that it contains a flow station.

The two lumped models studied are subject to a specific calibration for each gauged GRIP-O subbasin identified in Figure 1. This is likely to lead to optimal hydrologic simulations for the different subbasins than if performing global calibration over all the subbasins at once (Gaborit et al., 2015), in which case a unique parameter set would be used for all subbasins.

A detailed and accurate delineation of Great Lakes subbasins, based on 1 arcsecond (arcsec) data, has been published by Wang et al. (2015) as part of the Great Lakes Aquatic Habitat Framework (GLAHF). GRIP-O subbasins rely instead on 30 arcsec flow direction data from HydroSHEDS (Lehner et al., 2006). We used the 30-arcsec data to delineate the watersheds because of the resulting gain in the required computational time for distributed routing models. The same delineation was used for the lumped models. However, GLAHF watershed boundaries were used as a guideline to correct the major discrepancies of the GRIPO delineations based on 30-arcsec data. Cases involving "missing areas" in the GRIP-O subbasins were, however, not addressed, such as for the area located right below the "10 bis" text of Figure 1.

The gauge stations were selected based on their data availability and proximity to the lake shoreline, so that a maximum coverage of the lake watershed could be reached with a minimum of stream gauges with minor missing records, regardless of the type of flow regime (i.e., natural or regulated). Of the 30 stream gauges of Figure 1, 27 have no missing data, two are complete at 94%, and the 20-mile creek gauge (subbasin 1) is complete at 80% over the GRIP-O period.

Regulation within the subbasins (Fig. 1) generally involves artificial reservoirs with dams. Some information is available with regard to the location and size of the structures (Simley and Carswell, 2009). A few consist of major hydraulic structures which are used for hydropower production (such as on the Trent River in subbasin 12) or flooding prevention (such as the Mount Morris dam on the Genesee River). However, no information was found with respect to management policy or operating rules for these structures. Therefore, observed and simulated flow time series downstream of the structures were assessed to gain insights about the effects of the artificial structures which are not accounted for by the models. Canadian structures do not result in frequent and abrupt streamflow fluctuations that the models cannot reproduce, while some U.S. structures do (such as for the Oswego River, which shows strong fluctuations with a period sometimes as short as one day). Overall, most artificial structures on the Lake Ontario watershed seem to mainly involve simple management policies with a smooth effect on downstream stream flows and could be seen as artificial reservoirs with a simple weir at their outlet, hence resulting in artificial lakes. The effect of such structures is relatively well handled by the models during calibration.

Some aquifers do exist in the vicinity of Lake Ontario. The main aquifers are located between Lake Simcoe (west of subbasin 12) and Lake Ontario and consist of two confined aquifers separated by impervious layers and an unconfined aquifer (the Oak Ridges Moraine). These major aquifers are located in subbasins 14, 13, and 12 (Howard et al., 1996). The deepest parts of the confined aquifers can be located down to 150 m below surface. The Oak Ridges Moraine does contribute to stream baseflow for rivers of the aforementioned catchments; Kassenaar and Wexler (2006) estimated that 90% of the Oak Ridges Moraine discharge does contribute to the streams of the northern shore of Lake Ontario. No information was found about the contribution from the deep, major confined aquifers to the streams of the Lake Ontario watershed, but Harvey et al. (2000) demonstrated that water can be brought by aquifers directly to the lake, i.e., without being released first into the streams. However, this flux is generally considered negligible by operational hydrologists in comparison to the other components of the lake's NBS (Lauren Fry, USACE 2015, personal communication). Many other small aquifers do exist in the region (Singer et al., 2003, for an exhaustive review). Therefore, the effect of these aquifers on the streamflow of Lake Ontario tributaries may represent a challenge for hydrologic models, in addition to the regulated flow regime of many rivers in this watershed (Fig. 1).

Finally, some diversions are performed in order to fill the Welland and New-York State Barge (NYSB) canals. The Welland canal (located in subbasin 1) diverts water from Lake Erie into Lake Ontario and involves flows of about 250 m^3/s. The NYSB canal diverts about 30 m^3/s in summer from the Niagara River into the Genesee River (subbasin 3), and a little less than 30 m^3/s is taken from the Genesee River to fill the NYSB canal between the Genesee and Oswego rivers, so this diversion is not expected to significantly affect the Genesee River's overall balance. Finally, the Oswego River gauge station does not take the NYSB canal flow into account. In summary, neither canal nor diversion is supposed to have any significant effect on the Lake Ontario recorded tributary stream flows, but the aquifers may represent an additional challenge for hydrologic models used in the GRIP-O.

Calibration schemes

Two different calibration schemes are used for this project: scheme 1, in which the models are calibrated using an estimation of the whole catchment runoff derived from the gauged observed flow and a simple Area-based Ratio Method (ARM); and scheme 2, which consists of implementing the models over the gauged part of a subbasin. Because the ultimate goal of the GRIP-O is to develop efficient modeling tools to simulate runoff from the whole Lake Ontario watershed, including its ungauged parts, calibration scheme 1 can be viewed as a first step in this direction because it allows simulating runoff for the ungauged area of a gauged subbasin (Fig. 1). The same protocol could also be used to estimate runoff for the whole Lake Ontario watershed, by considering the total Lake Ontario watershed as a single catchment (and applying a unique model over this large area; see section on runoff estimation for the whole GRIP-O area). The ARM has proven reliable for estimating flows for ungauged portions of a watershed, as long as its gauged fraction is higher than a certain threshold of about 40% (Fry et al., 2014). The synthetic flows built this way make it possible to implement the hydrologic model over the whole subbasin at once, thus taking into account rainfall over the ungauged part of the catchment.

Other methods for estimating runoff from a complete subbasin exist but were not as attractive as calibration scheme 1. For example, one could calibrate the model at the gauged sites and then extend the simulated flows to the whole basin using the ARM, but this would imply neglecting the difference in rainfall amounts between the gauged and ungauged areas. One could also implement the model in two steps: a first one to calibrate it at the gauged sites, and a second one to implement it over the whole catchment, using the true rainfall amounts falling over its area and the parameter set derived from the first calibration. Calibration scheme 2 actually corresponds to the first part of this approach, but it requires two implementations of the same model for each subbasin.

To determine which is the best practice among these two different possibilities is beyond the scope of this work, but the preferred two are used in the GRIP-O: both are easy to implement and make use of the "true" forcings. It was first envisioned to only use calibration scheme 1 for all gauged subbasins for consistency with the former GRIP-M (Fry et al., 2014), but calibration scheme 2 corresponds to a practice commonly used in hydrology, i.e., implementing and calibrating the models at the gauged sites (with true observed flows). As a consequence, an arbitrary subdivision was made between the subbasins: when one contains several gauges, calibration scheme 1 is applied (5 cases); when it contains a unique most-downstream gauge (Fig. 1, 8 cases), the second scheme is used. However, these two calibration schemes are not compared here because they were not both applied to each of the subbasins.

Precipitation sources

The first source of precipitation consists of the Canadian Precipitation Analysis (CaPA), a system relying on modeled precipitation fields derived from the Canadian Regional Deterministic Prediction System (RDPS) but corrected with ground-based precipitation observations. More precisely, we chose to use the 24-h CaPA data of the 2.4 b8 version (Lespinas et al., 2015), which consists of gridded fields of 24-hourly accumulated precipitation, corrected by ground stations but not by radar fields, which is the case for CaPA 3.0 (Fortin et al., 2015). No reanalysis was yet available for CaPA 3.0 over the time period of interest to the GRIP-O. Both 6- and 24-h accumulations are available from CaPA 2.4 b8. It was decided to use the 24-h product because more ground stations are generally available in real-time at the 24-h interval than at the 6-h interval, making the 24- h CaPA product more tied to observations than the 6-h one. The resolution of CaPA 2.4 b8 is 0.125 degree (or 450 arcsec, around 15 km near Lake Ontario). CaPA is designed for near real-time application and is thus a fully automated precipitation analysis procedure.

The second main source of precipitation data available for the GRIP-O consists of the Global Historical Climatology Network- Daily (GHCND version), developed by the NOAA National Climate Data Center (NCDC), which incorporates many data sources all over the globe, such as weather and climate stations as well as observations from volunteer observers in the U.S. and Canada. The data were interpolated on a 15 arcsec grid (around 450 m near Lake Ontario) using a nearest neighbor (or Thiessen polygons) method.

Table 1 shows the number of ground stations per GRIP-O subbasin (Fig. 1) for each precipitation dataset. The density of the GHCND is higher than that of the stations used in CaPA because CaPA only uses stations for which data is available in real-time at the Canadian Centre for Meteorological and Environmental Prediction (CCMEP). The GHCND network density, however, varies between 2004 and 2011 (the GRIP-O period): it increases in the U.S. basins and decreases in Canada (Table 1). Watershed averages were computed based on the gridded products and the subbasin shapes, either for the whole subbasin in case of calibration scheme 1 or for its gauged area for scheme 2. Both models were provided with daily basin averages of maximum and minimum temperature based on the data contained in the GHCND (same stations as the precipitation ones), even in the case of providing the models with CaPA precipitation.

Calibration details

Because calibrating a hydrologic model over a set of 4 to 5 years is generally enough to achieve reasonable model robustness (e.g., Refsgaard et al., 1996), the calibration period was chosen to range from June 1, 2007 to the end (4.5 years). The validation period extends from June 1, 2005 to June 1, 2007 (2 years), with the first year of data (June 1, 2004 to June 1, 2005) being used for spin-up. The objective function used in calibration consists in the Nash-Sutcliffe Efficiency criterion (NSE; Nash and Sutcliffe, 1970) but computed taking the square-root of the observed and simulated flow time series, in order to avoid overemphasizing peak-flow events, and will be referred to as "NSE √" throughout the paper. Other evaluation criteria used in this study consist in the common Nash-Sutcliffe criterion (NSE), the Nash criterion calculated over the log of the flows (NSE ln), and a bias criterion (in equation 1 below) indicates a simulation's overall water budget fit.

Equation (1) $$\mathrm{PBIAS} = \frac{\sum_{i=1,n}(\mathrm{Qobs}_i - \mathrm{Qsim}_i)}{\sum_{i=1,n}(\mathrm{Qobs}_i)}$$

GR4J and the "Cémaneige" snow module have a total of six free parameters, and LBRM was implemented here using ten free parameters (the upper soil zone capacity, USZ, in addition to its nine conventional ones; Table 2). The calibration algorithm consists in the Dynamically Dimensioned Search (DDS) algorithm (Tolson and Shoemaker, 2007). DDS is designed for calibration problems with many parameters and automatically scales the search within the maximum number of user-specified model runs. A maximum of 2000 model runs was allowed during a calibration trial, and three distinct trials (each starting with a different initial parameter set) were performed for each model, precipitation dataset, and subbasin. For each model and subbasin, the best of the three trials was finally kept to compute the validation performances. Table 2 presents the parameter ranges used in calibration. The initial parameter sets were randomly selected inside the prescribed intervals.

Results and Discussion

Best hydrologic simulations

What are the best possible hydrologic simulation performances for the area under study and how do the models account for regulation impacts on runoff? Among the four different model combinations examined (LBRM or GR4J model, CaPA or GHCND precipitation dataset), the best simulation performances were generally achieved with GR4J and the GHCND dataset (Table 3). Moreover, performances obtained with the latter combination a priori consist of some of the best possible hydrologic simulation performances for the area under study. Indeed, the GHCND dataset consists of a dense network of ground-based stations for precipitation and temperature, and GR4J generally stands among the best hydrologic models in multi-model comparison studies (e.g., Pagano et al., 2010; Seiller et al., 2012).

However, it is always possible to improve local simulation performances, but results presented in this section nevertheless represent the benchmark to which other experiments will be compared.

A few conclusions can be drawn from Table 3. Simulation performances are not related to any of the subbasins' main properties (country affiliation, area, elevation, type of flow regime) or to the calibration scheme or catchment gauged percentage. Very satisfying performances were obtained even in the worst case scenario (influenced flow regime, calibration scheme 1 with synthetic observed flows) and a relatively low gauged fraction (see, for example, subbasin 10bis). At the same time, cases which were a priori expected to lead to some of the best results (natural or less influenced flow regime, calibration scheme 2) often display poorer performances (subbasins 4bis and 6). This was also true for the three other model combinations (LBRM / GR4J and GHCND/ CaPA precipitation).

Simulated hydrographs generally closely follow the observed ones, even for regulated catchments [Fig. 2; see Electronic Supplementary Material (ESM) Fig. S1. Flow regulation in the Lake Ontario watershed mainly consists of artificial reservoirs (see section on spatial framework and study area characteristics) with a weir at their outlet (static control which the models can a priori account for during calibration. However, good performances are obtained even for subbasin 5 (Oswego River, Table 3), despite the fact that it involves a more sophisticated type of control which results in frequent streamflow fluctuations for the observed time-series which the hydrologic models cannot reproduce (see ESM Fig. S1.). For this subbasin, the observed general flow dynamic is, nevertheless, well captured by the smoother simulated flows.

Furthermore, if some aquifers do significantly contribute base flow to some streams of the Lake Ontario watershed (e.g., the northern shore of Lake Ontario), they neither prevent the hydrologic models from successfully reproducing streamflow dynamics, nor do they lead to a strong underestimation of the overall water balance (Table 3).

The GR4J model is well adapted to hydrologic modeling of the Lake Ontario watershed. Model performances obtained here are promising both for the future implementation of distributed models and the estimation of the whole lake's runoff. Indeed, the regulated downstream flow gauges (Fig. 1), together with the gauges on "natural" streams, cover about 74% of the total Lake Ontario watershed land surface. Natural flows represent 20% of the watershed area under study. Therefore, if only natural flows would demonstrate good simulation performances, runoff for the whole watershed could be hardly estimated with the ARM (see Fry et al., 2014, for the reliability of a catchment's runoff estimation as a function of its gauged fraction). But, this is not the case here.

The fact that no calibration scheme systematically leads to better performances than the other, as observed with the three other model combinations, further supports the methodology used to simulate runoff for the ungauged part of a gauged catchment (calibration scheme 1). Using synthetic flows (derived from observed ones and the simple ARM process) to calibrate a model leads to satisfactory simulation performances even in the case where only 40-60% of the subbasin is gauged (see subbasins 6, 10bis, 13, 14 and 15 in Table 3). This also tends to further confirm findings of the GRIP-M (Fry et al., 2014), namely that the ARM can provide reliable runoff estimations for a whole subbasin even if less than half of it is gauged. To further confirm this, calibration scheme 2 was tested for subbasins 6 and 14, so that both schemes could be truly compared for these catchments. The results (see ESM Table S1.) confirm that performances are extremely close for both, sometimes even slightly better with calibration scheme 1.

Validation performances are often lower than in calibration, especially when evaluating the NSE criterion. However, this drop of performances between calibration and validation is smaller when using CaPA precipitation, indicating that the models can be robust and that the drop of performances between calibration and validation is more related to the precipitation data than to the model structure. Although performances are generally lower in validation with GHCND precipitation, they remain satisfactory for most of the subbasins (Table 3).

Finally, the GR4J exhibits satisfactory values for the three different NSE criteria (Table 3) (both in calibration and validation). This result indicates that the different components of streamflow are well reproduced by the model (baseflow, interflow, surface runoff, low-flow periods, floodings). Although no test was performed with using the traditional NSE criterion as the objective function, it is argued that using NSE √ as the objective function is a reasonable choice when using GR4J but not the LBRM (see below). The relevance of using NSE √ as the objective function was already demonstrated, for example by Oudin et al. (2006).

Model and precipitation dataset inter-comparison

Can a relatively new model to Great Lakes hydrological modeling (GR4J) be implemented and demonstrate similar or better success than the LBRM, and which precipitation product results in the better runoff simulations at a subbasin scale? In order to determine if differences in performance observed between the two hydrological models and the two precipitation datasets are statistically significant, a two-sided sign test is used (see for example Walpole and Myers, 1985). The sign test is performed on the sign of the differences in performance, assuming independence across basins. It does not inform us on the magnitude of the mean difference in performance, but rather on whether or not one model or one precipitation dataset is expected to outperform the other on a majority of watersheds. The number of cases for which GR4J outperforms LBRM and the number of cases for which GHCND outperforms CaPA are reported in Table 5, together with the associated p-values.

Results from Table 4 suggest that GR4J performs generally better than LBRM considering NSE √ values (the objective function) in calibration. However, results from Table 5 indicate that for this criterion (NSE √) it cannot be stated with confidence that GR4J is better than LBRM (the difference is not statistically significant). The superiority of GR4J over LBRM is, however, significant for the NSE criterion (Table 5, both in calibration and validation), which puts the emphasis on high streamflow values. Because performing well in peak flow events is one of the main requirements for a hydrologic model, calibrating the LBRM using the NSE √ objective function may not be a good choice, as it generally has difficulty for peak events (Figs. 2 and 3).

Table 6 contains some valuable information for comparing the GHCND and CaPA precipitation datasets. Using the GHCND precipitation dataset leads to better calibration performances than when using the CaPA dataset (Tables 4 and 5). This suggests that the daily average catchment precipitation derived from the GHCND data and using the Thiessen Polygon method is closer to the real amounts than the values derived from CaPA are. This also makes sense considering that the GHCND network density is higher than the density of the real-time ground stations used in CaPA (Table 1).

When comparing the U.S. or Canadian side of the Lake Ontario watershed, the model performances are quite different (Table 6). In Canada, model performances resulting from GHCND and CaPA precipitation are close (Table 6), while in the U.S. (median U.S.), the difference resulting from using one source or the other is more pronounced. Because the density of the GHCND is much higher than the density of the ground stations used in CaPA, it can be concluded with some confidence that the daily basin averages of precipitation are well reproduced by CaPA in Canada. This is confirmed when looking at the correlation between CaPA and GHCND daily watershed averages of precipitation (see ESM Table S3.).

However, the differences in performance between GHCND and CaPA are less pronounced in validation than in calibration, and are not statistically significant according to the sign test (Table 5). In fact, for the NSE √ criterion, CaPA outperforms GHCND in validation for a majority of watersheds (9 out of 14). In particular, CaPA leads to better performances than the GHCND in validation for some catchments such as the 4bis, 8, 12 and 15 subbasins. This cannot be attributed to a lower GHCND network density in validation (Table 1). However, the catchments displaying this behavior are in areas where the GHCND density is generally low for the calibration period.

We suggest the lumped models learn to compensate for a spatial misrepresentation of GHCND precipitation during calibration (and for the catchments mentioned earlier) but that this compensation is not suited anymore in validation because the GHCND density changes over time (Table 1), resulting in a better model temporal robustness with CaPA than with GHCND (Table 6). Although not clearly shown here, performances obtained with CaPA precipitation lead to models with a strong temporal robustness (see ESM Table S4.), which therefore supports the relevance of this comparison study.

Parameter values

When looking at parameter values obtained after calibration (Table 2), no clear difference could be identified between using the CaPA or GHCND precipitation dataset for either of the two models. Parameter values were generally close with the two different precipitation sources. The values are not shown here for all of the catchments (see ESM Table S5), but a few statistics are shown for these values as it could be of some interest to initialize future GR4J or LBRM implementations over nearby catchments and/or similar climatic areas. Tables 7 and 8 are devoted to such a purpose.

As GR4J could end up with two main types of parameter sets, depending on if the production store maximum capacity was much lower than the routing store maximum capacity or not, these two cases were distinguished to compute the statistics of Table 8. The production store capacity is supposed to conceptually represent soil depth (Harlan et al., 2010). In some cases (4/14 with GHCND and 5/14 with CaPA), the GR4J production capacity (X1, mm) was very low compared to the routing store capacity (X3, mm) with values generally below 30 mm for the former. This is not considered a realistic value for the Lake Ontario watershed based on observed soil depth data (~1.4 m) but also according to GR4J general range of parameter values (Perrin et al., 2003).

These very low values obtained after calibration with GR4J for the X1 parameter (and accompanied by high values for the X3 routing store capacity) are attributed to flow regulation because they were only observed in the case of partly regulated subbasins, such as subbasins 3, 5, 8, and 12. As regulation in these watersheds is mainly due to artificial reservoirs, the low X1 and high X3 values (Table 8) obtained for these catchments could be seen as an effort made by the model to optimally represent the effect of regulation on simulated stream flows by trying to represent a big routing reservoir (X3). Because this reservoir then captures 90% of runoff generated by the production store (X1), a high X3 value may logically call for a low X1 value to still be able to represent peak flows. Additional tests performed in the case of the very low X1 values revealed that they were not due to local optima found during calibration, as changing the initial parameter values did systematically lead to the same unconventional values. Moreover, the good model temporal robustness obtained with these unconventional X1 and X3 parameter values, along with the very satisfactory simulated hydrographs which they achieve, demonstrates the relevance of the second type of parameter values described in Table 8. Despite the satisfactory performances obtained even for subbasins with regulation structures, it would be interesting to test the methodology proposed by Moulin et al. (2005) to better account for major hydraulic structures in GR4J, but this is dedicated to future work.

Additional model tests

Additional model combinations were performed in order to assess the sensitivity of the evaluation metrics considering the observed and simulated time-series in a different way or in more detail. A test was performed by computing the scores with 7-day averages of flow values to hinder the effect of daily fluctuations induced by regulation on observed flows such as those of the Genesee and Oswego Rivers. Doing so did not lead to a significant increase of performances for the rivers subject to such fluctuations (see ESM Table S6.), and does not mitigate the effect of regulation by slightly changing the time-step of the time-series.

Performances were also assessed on a seasonal basis for the winter (November to May) or summer (June to October) periods, but still using a daily-time-step. The analysis was limited to the calibration period. Winter performances tend to be better than summer ones in terms of all NSE criteria used in this study (see ESM Table S7). This is common in hydrology, as winter processes are generally less chaotic and easier to reproduce than summer ones. For example, winter stream flows in the area considered are frequently governed by temperature via snowmelt processes. In summer, convective storms can be localized over very small areas and can be poorly captured by observation networks. The better winter performances in comparison to summer ones are also due to the NSE criterion itself, which is generally better for higher streamflows (see Martinec and Rango, 1989).

Apart from this, there is not much information which can be derived from this seasonal analysis. It is only highlighted that despite better than summer ones, winter NSE values are almost never higher than overall values for the whole calibration period. In summer, performances remain at least satisfactory, with most of the values comprised between 60 and 80%. The scores related to seasonal or weekly performances can be found in ESM Table S6 and S7.

Runoff estimation for the whole GRIP-O area

What is the best approach for estimating runoff, including its ungauged areas, for the entire Lake Ontario watershed? In order to have a better overview of the performance of Lake Ontario runoff simulations, the sum of runoff from the models implemented locally over each of the 14 GRIP-O gauged catchments (Fig. 1) was evaluated. The total area covered by the models is about 53,460 km^2. Runoff observations used for the 53,460-km^2 area actually correspond to estimations because they are derived from runoff truly observed for the gauged 47,330-km^2 area (brown part of Fig. 1) and the ARM. This estimate should be close to reality as more than 88% of the studied area is gauged.

A test of the model level of detail was made with implementation of a unique GR4J model over the entire 53,460-km^2 area using the GHCND. In this case, the GR4J model was calibrated using precipitation and temperature values corresponding to weighted averages of the 14 subbasins' time-series and the sum of locally observed flows as reference. When using a unique model, the entire area is conceptually interpreted as a unique watershed with one main river, which is far from reality (Fig. 1). The unique model was implemented considering five different elevation classes of equal size and still the NSE √ value as objective function. A unique model is faster to implement than many local ones. This test is inspired from Croley (1983) who showed very promising results by doing so with LBRM.

In general, the simulated whole watershed performances are very good (Table 9) and are, in fact, better than performances obtained for any of the local subbasins- whatever the quality criterion considered (Table 3). This result suggests that the local model biases are compensated when grouping the catchments altogether. Moreover, as the NSE criterion is a skill score comparing simulated values to the mean of the observed flows, the reference is easier to gain the upper hand on when the flow dynamics are very strong, as is the case for the entire area (Fig. 4). When taking the mean of values simulated with the four combination possibilities (two models and two precipitation datasets), performances are as good or better than the best of the four independent combinations, except for the PBIAS criterion which stands somewhere in between the extreme values obtained with the individual combinations.

All four individual combinations of Table 9 display robust performances, which again highlights the compensation of local biases; it can be seen in validation (Table 3). In all cases observed, runoff is generally underestimated by the local models (Tables 3 and 9), which can be more precisely linked to the underestimation of peak flow events with local models (Fig. 4). It is nonetheless evident that both of the two lumped models used here, namely GR4J and LBRM, are able to produce very decent estimations of Lake Ontario runoff, which could be useful to several potential challenges in the field of water management and streamflow / lake level predictions.

Using a unique GR4J model to save computation time and effort for estimating runoff for the whole GRIP-O area described earlier proved to be promising. Although values for the different NSE criteria are slightly smaller for this unique model in comparison to performances obtained by summing all local models together (Table 9, Fig. 4), the opposite is true for the PBIAS criterion. As a consequence and depending on the targeted application, it may be preferable to use a unique model for the whole GRIP-O area than a set of models implemented on local subbasins, such as estimating the general contribution from the entire Lake Ontario area to the lake. In this case, it would a priori be preferable to use a unique model over the entire area and calibrate it with synthetic historic flow records derived from the ARM (i.e., using calibration scheme 1 for the entire Lake Ontario area). This way, the model could benefit from a better approximation of precipitation and temperature over the watershed, by taking into account the values over the ungauged parts of the watershed.

Tests were also performed calibrating the unique model using the conventional NSE criterion, but results were almost exactly the same (not shown but available upon request), which tends to suggest that for GR4J, calibrating on NSE √ or NSE does not change model performances. A calibration test with the NSE instead of NSE √ criterion performed on the Moira watershed (and still with GR4J and the GHCND precipitation) even revealed poorer simulation results with the former conventional criterion than with the objective function used in this work. Using five elevation classes leads to slightly better performances and hydrographs than with a unique elevation class (result available upon request), but this difference may be more significant in the case of watersheds with more pronounced topography.

Conclusion

The first phase of the GRIP-O project, involving the implementation of the two lumped GR4J and LBRM hydrological models on the land area of the Lake Ontario watershed, has shown that both models were very efficient in simulating the lake's tributary stream flows. Both models are robust and perform well, whatever the precipitation dataset used as input. However, GR4J performs generally better than LBRM considering the NSE criterion. This suggests that using the NSE √ value as the objective may not be the best choice for implementing LBRM, which would probably achieve better hydrologic simulations using the conventional NSE criterion. For the model combinations performed during this work, GR4J appears to be better than LBRM, but more tests, including additional watersheds, should be performed.

The GHCND dataset leads to better performances than CaPA precipitation in calibration, but not in validation, which is probably linked to the fact that the GHCND network changes over time. Overall, CaPA leads to hydrologic performances very close to the ones obtained with the GHCND for some GRIP-O subbasins, even in the case where the network of ground stations used in CaPA has a significantly lower density than the GHCND stations. This suggests that CaPA is a very useful source of daily precipitation data.

As a consequence, the model combination leading to the best hydrologic performances consists of the GR4J model driven by the GHCND forcings. Numerous possibilities could still be tested, for example, by calibrating the models with the conventional NSE values or by using multi-objective functions that would focus at the same time on the streamflow and snow water equivalent simulations. The results from GR4J and the GHCND forcings will be further used as a benchmark for future hydrological modelling experiments on Lake Ontario, such as an evaluation of distributed models.

The two different calibration schemes used resulted in very satisfactory performances. Using synthetic streamflow time-series for a whole subbasin, derived from observed data and a simple area-ratio method, resulted in a promising and efficient way to estimate the contribution of ungauged parts of a gauged subbasin.

Despite flow regulation affecting most of the GRIP-O subbasins, it did not prevent the models from achieving very satisfactory performances. The models, through calibration, generally mimic the effect of flow regulation, except for some of the U.S. basins (e.g., Oswego River), which involve dynamically sophisticated modifications to the natural flow regime but did not prevent the models from following the general flow trends.

In the light of the performances obtained in this study, promising results are expected in regard of the estimation of runoff for the entire Lake Ontario area (64,000 km^2, Fig. 1). Indeed, performances are even better when looking at the total 53,460-km^2 GRIP-O area as a whole instead of performances for local catchments. Moreover, even a unique GR4J model implemented over the total area resulted in very satisfying performances, especially for the PBIAS criterion, which could save a lot of computation time when interested in runoff for the entire Lake Ontario area. For that, a possibility would be to calibrate a unique GR4J model on the total gauged area of Lake Ontario (Fig. 1) and then to transfer the parameter set to a unique model for the entire Lake Ontario watershed, including its ungauged parts. This would allow taking into account rainfall and temperature data over the ungauged parts of the whole watershed. Runoff simulations for the entire Lake Ontario watershed provide a better understanding of long-term water supply trends (the GRIP-O period can be extended, for example, using CaPA).

Considering the promising results obtained during this work with lumped conceptual models, the second phase of the project will focus on the implementation of a distributed, physically based model developed at Environment and Climate Change Canada and named GEM-Hydro. Physically-based distributed models are interesting because they offer more possible applications than lumped models by representing the physical processes occurring everywhere inside the watershed. The area of Lake Ontario is thus relevant to distributed modeling, as detailed datasets exist for land use/land cover and soil texture. However, the performances of the lumped models will be hard to achieve with distributed models, given that lumped models can, through calibration, more easily accommodate themselves with systematic deficiencies in the forcing data. The good performance of lumped models could serve as a hydrological simulation target for more sophisticated models.

Acknowledgements

This work was supported by the International Joint Commission International Watersheds Initiative. The authors wish to thank Milena Dimitrijevic (E-NPR, Environment Canada, CCMEP Dorval) for providing the scripts that computed the catchment averages of precipitation and temperature from the gridded datasets. This is NOAA-GLERL contribution No. XXXX.

References

Anderson, E.J., Schwab, D.J., Lang, G.A., 2010. Real-time hydraulic and hydrodynamic model of the St. Clair River, Lake St. Clair, Detroit River system. J. Hydraul. Eng. 136, 507–518.

Angel, J. R., Kunkel, K. E., 2010. The response of Great Lakes water levels to future climate scenarios with an emphasis on Lake Michigan-Huron. J. Great Lakes Res. 36(sp2), 51-58.

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

Chao, P., 1999. Great Lakes Water Resources: Climate change impact analysis with transient GCM scenarios. JAWRA 35(6), 1499-1507.

Coon, W. F., Murphy, E. A., Soong, D. T., Sharpe, J. B., 2011. 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 (No. 2011-1202). US Geological Survey.

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

Croley, T., Hartmann, H., 1987. Near Real-Time Forecasting of Large Lake Supplies. J. Water Resour. Plann. Manage. 113(6), 810-823.

Croley, T.E., Lee, D. H., 1993. Evaluation of Great Lakes net basin supply forecasts. JAWRA 29(2), 267-282.

Croley, T.E., He, C., 2002. Great Lakes large basin runoff modeling, in: Second Federal Interagency Hydrologic Modeling Conference, Subcommittee on Hydrology of the Interagency Advisory Committee on Water Data, Las Vegas, NV, July 28-August 1, 12 pp.

Croley, T. E., 2003. Great Lakes climate change hydrologic impact assessment: IJC Lake Ontario-St. Lawrence River regulation study. Great Lakes Environmental Research Laboratory, US Department of Commerce, National Oceanographic and Atmospheric Administration.

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

DeMarchi, C., Dai, Q., Mello, M. E., Hunter, T. S., 2009. Estimation of Overlake Precipitation and Basin Runoff Uncertainty. International Upper Great lakes Study.

Eum, H.-I., Dibike, Y., Prowse, T. Bonsal, B., 2014. Inter-comparison of high-resolution gridded climate data sets and their implication on hydrological model simulation over the Athabasca Watershed, Canada. Hydrol. Process. 28, 4250–4271.

Fortin, V., Roy, G., Donaldson, N., & Mahidjiba, A., 2015. Assimilation of radar quantitative precipitation estimations in the Canadian Precipitation Analysis (CaPA). Journal of Hydrology 531, 296-307.

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., Restrepo, P., 2014. The Great Lakes Runoff Intercomparison Project Phase 1: Lake Michigan (GRIP-M). J. Hydro. 519(D), 3448-3465.

Gaborit, É., Fortin, V., Xiaoyong, X., Seglenieks, F., Tolson, B., Fry, L., Hunter, T., Anctil, F., Gronewold, D. In preparation. A Hydrological Prediction System Based on the SVS Land-Surface Scheme: Implementation and Evaluation of the GEM-Hydro platform on the watershed of Lake Ontario.

Gaborit, É., Ricard, S., Lachance-Cloutier, S., Anctil, F., Turcotte, R., 2015. Comparing global and local calibration schemes from a differential split-sample test perspective. Can. J. Earth Sci. 52(11), 990-999.

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

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

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

Harlan, D., Wangsadipura, M., Munajat, C. M., 2010. Rainfall–runoff modeling of Citarum Hulu River basin by using GR4J, in: Proc. World Congress on Engineering 2010, pp. 1607-1611.

Harvey, F. E., Rudolph, D. L., Frape, S. K., 2000. Estimating ground water flux into large lakes: Application in the Hamilton Harbor, western Lake Ontario. Ground Water 38(4), 550-565.

Hay, L. E., Markstrom, S. L., Ward-Garrison, C., 2011. Watershed-scale response to climate change through the twenty-first century for selected basins across the United States. Earth Interact. 15(17), 1-37.

Howard, K.W.F., Eyles, N., Smart, P.J., Boyce, J.I., Gerber, R.E., Salvatori, S.L., Doughty, M., 1996. The Oak Ridges Moraine of southern Ontario: a groundwater resource at risk. Geoscience Canada 22(3), 101-120.

Kassenaar, J.D.C., Wexler, E.J., 2006. Groundwater Modelling of the Oak Ridges Moraine Area. CAMC-YPDT Technical Report #01-06.

Kouwen, N., 2010. WATFLOOD/WATROUTE Hydrological model routing & flow forecasting system. Department of Civil Engineering, University of Waterloo, Waterloo, Ontario, Canada.

Lehner, B., Verdin, K, Jarvis, A., 2006. HydroSHEDS technical documentation, version 1.0. World Wildlife Fund US, Washington, DC: 1-27.

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

MacKay, M., Seglenieks, F., 2013. On the simulation of Laurentian Great Lakes water levels under projections of global climate change. Climatic Change 117(1-2), 55-67.

Mahfouf, J.-F., Brasnett, B., Gagnon, S., 2007. A Canadian precipitation analysis (CaPA) project: Description and preliminary results. Atmos.-Ocean 45(1), 1-17.

Mao, D., Cherkauer, K. A., 2009. Impacts of land-use change on hydrologic responses in the Great Lakes region. J. Hydro. 374(1), 71-82.

Martinec, J., Rango, A., 1989. Merits of statistical criteria for the performance of hydrological models. Water Resources Bulletin 25(2): 421-432.

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

Moulin, L., Perrin, C., Michel, C., Andréassian, V., 2005. Prise en compte de barrages-réservoirs dans un modèle pluie-débit global: application au cas du bassin de la Seine amont. La Houille Blanche 5, 79-88.

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

Nicolle, P., Ramos, M.H., Andréassian, V., Valery, A., 2011. Mieux prévoir les crues nivales : évaluation de prévisions probabilistes de débit sur des bassins versants de montagne Français, in: actes du colloque SHF: "L'eau en montagne, mieux observer pour mieux prévoir", 163-170. 16th of March 2011, Lyon : France.

Oudin, L., Andréassian, V., Mathevet, T., Perrin, C., Michel, C., 2006. Dynamic averaging of rainfall-runoff model simulations from complementary model parameterizations. Water Resources Research 42(7), W07410. DOI: 07410.01029/02005WR004636.

Pagano, T., Hapuarachchi, P., Wang, Q., 2010. Continuous rainfall-runoff model comparison and short-term daily streamflow forecast skill evaluation. CSIRO Tech. Rep. EP103545, 70 pp., CSIRO, Australia.

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

Pietroniro, A., Fortin, V., Kouwen, N., Neal, C., Turcotte, R., Davison, B., ... , Pellerin, P., 2007. 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.

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

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

Seiller, G., Anctil, F., Perrin, C., 2012. Multimodel evaluation of twenty lumped hydrological models under contrasted climate conditions. Hydrol. Earth Syst. Sci. 16, 1171–1189.

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

Simley, J. D., Carswell Jr, W. J., 2009. The National Map—Hydrography: US Geological Survey Fact Sheet 2009–3054. US Geological Survey National Center, Reston, VA.

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

Valéry, A., 2010. Modélisation précipitations-débit sous influence nivale, élaboration d'un module neige et évaluation sur 380 bassins versants. Thesis (PhD). ENGREF, Cemagref, Paris, 405 pp.

Walpole, R. E., Myers, R. H., 1985. Probability and Statistics for Scientists and Engineering. MacMillan Publishing Co., NewYork, NY, 17.

Wang, L., Riseng, C. M., Mason, L. A., Wehrly, K. E., Rutherford, E. S., McKenna, J. E., ... , Robertson, M., 2015. A spatial classification and database for management, research, and policy making: The Great Lakes aquatic habitat framework. J. Great Lakes Res. 41(2), 584-596.

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

List of Tables:

Table 1: Number of precipitation ground stations per GRIPO subbasin (see Figure 1) for the GHCND and CaPA datasets.

Table 2: Parameter ranges used in calibration; one of the three GR4J trials was performed with GR4J default initial parameter values indicated here; other trials were made with a random initial parameter set; coeff: coefficient; UH: Unit Hydrograph; USZ: Upper Soil Zone; LSZ: Lower Soil Zone.

Table 3: Evaluation metrics obtained with GR4J and the GHCND precipitation dataset. All values are in percent; optimal value is 100 except for PBIAS (0). NSE √, NSE Ln: NSE values computed with either the square-root or log of the flows, respectively. See section on *methods* for more details. NSE=Nash-Sutcliffe Efficiency index, PBIAS=Percent Bias, CA=Canada.

Table 4: NSE (Nash-Sutcliffe Efficiency index) values obtained with GR4J, minus NSE values from LBRM. CALIBR=calibration,VALID=validation, GHCND=Global Historical Climatology Network-Daily, CaPA=Canadian Precipitation Analysis. Values are in percent, so that a value of 5 in the Table corresponds for example to a NSE of 75% for GR4J and 70% for LBRM.

Table 5. The number of positive cases associated to the inter-comparison study, i.e., the number of cases with the GR4J model or GHCND precipitation showing better performances than LBRM or CaPA, respectively. For the comparison of GR4J vs LBRM or GHCND vs CaPA, the number of cases displayed in Table 6 was derived from the mean of performances along the two precipitation datasets or models, respectively. The p-value of the two-sided sign test is displayed in parenthesis. A number equal or greater than 12 out of 14 cases is considered significant (i.e., a p-value of less than 5%). CALIBR=calibration, VALID=validation, GHCND=Global Historical Climatology Network- Daily, CaPA=Canadian Precipitation Analysis.

Table 6: NSE √ and NSE (Nash-Sutcliffe Efficiency index) values obtained with the GHCND precipitation dataset, minus NSE values obtained with CaPA precipitation, in calibration (CALIBR) or validation (VALID), and with the LBRM or GR4J models. Values are in percent, so that a value of 5 corresponds, for example, to a NSE of 70 %with CaPA and of 75 % with GHCND precipitation.

Table 7: Statistics for the LBRM (Large-Basin Runoff Model) parameters after calibration taking all cases into account (28: 14 catchments and 2 precipitation datasets). Lin= linear, res=

reservoir, coeff= coefficient, evap= evaporation, perco= percolation, USZ= Upper Soil Zone, std= standard deviation, LSZ= Lower Soil Zone, Tbase = base temperature, val. = value.

Table 8: Statistics for the GR4J (*modèle du Génie Rural à 4 paramètres Journalier*) parameters after calibration separating the 28 final parameter sets in two cases depending on the relative values of X1 and X3, either of similar magnitude (or X1 > X3), or with X1 way lower than X3 (X1 << X3).

Table 9: Daily runoff simulation performances when assessed for the 57,460-km^2 Lake Ontario area (see text). CAL= calibration, VAL= validation, NSE= Nash-Sutcliffe Efficiency index. The mean of runoff simulated with the four different combination possibilities (mean GR4J-LBRM-CAPA-GHCND) was assessed. The last column shows performances obtained with a unique GR4J model applied to this large area using GHCND precipitation.

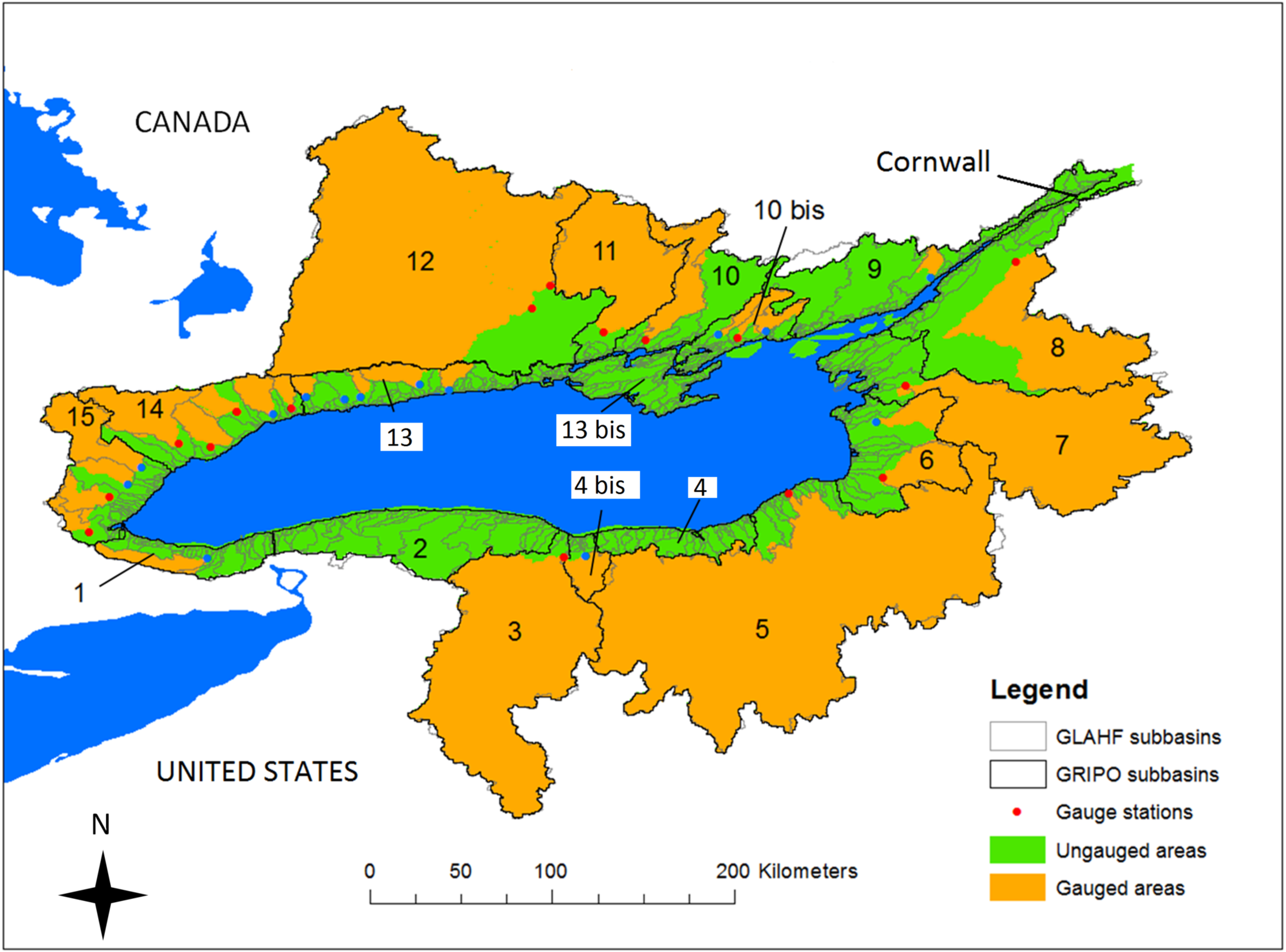
Figure captions

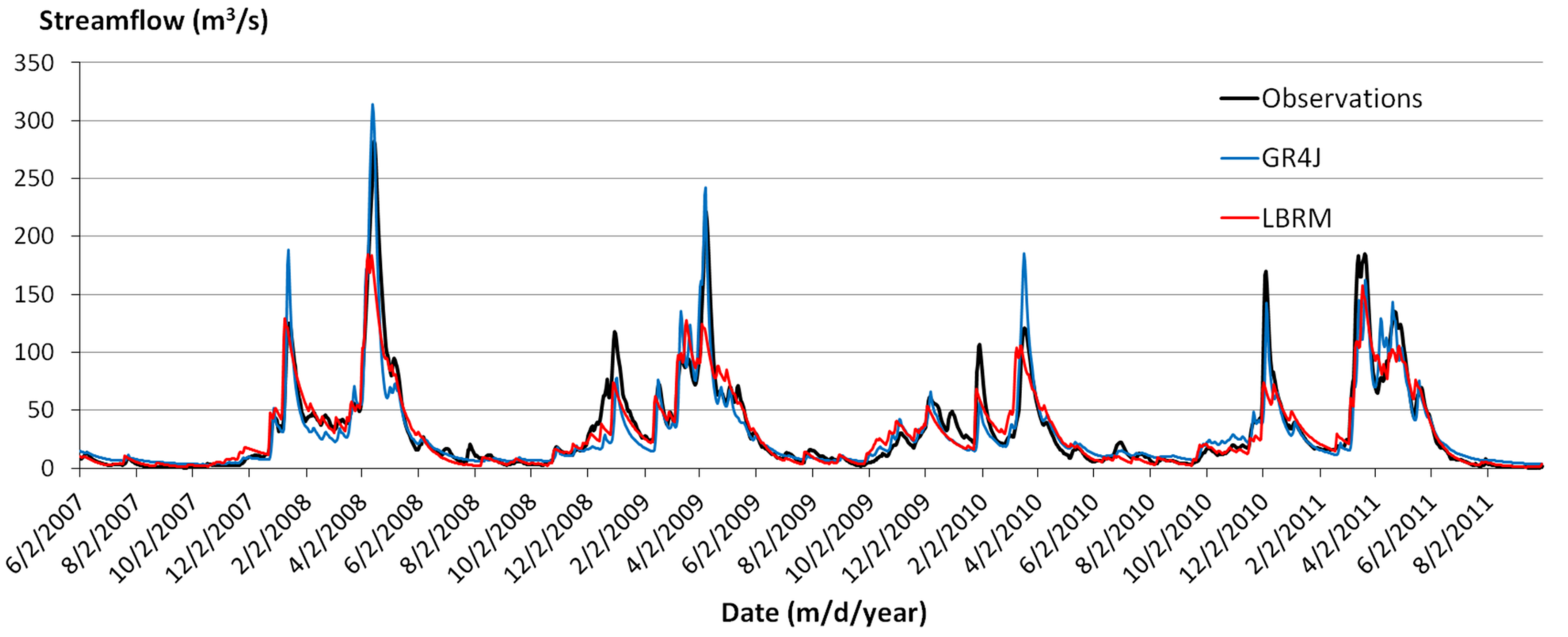
Figure 1: Lake Ontario subbasin delineation (GRIP-O subbasins). GLAHF subbasins depict the delineation performed by the Great Lakes Aquatic Habitat Framework. The main watershed outlet is located at Cornwall, ON. All dots correspond to most-downstream flow gauges selected for model calibrations; blue ones correspond to rivers with natural flow regimes while red ones are located on regulated rivers. The light (orange in online version) area represents the gauged area (about 74% of the total - dark (green in on-line version) - Lake Ontario watershed).

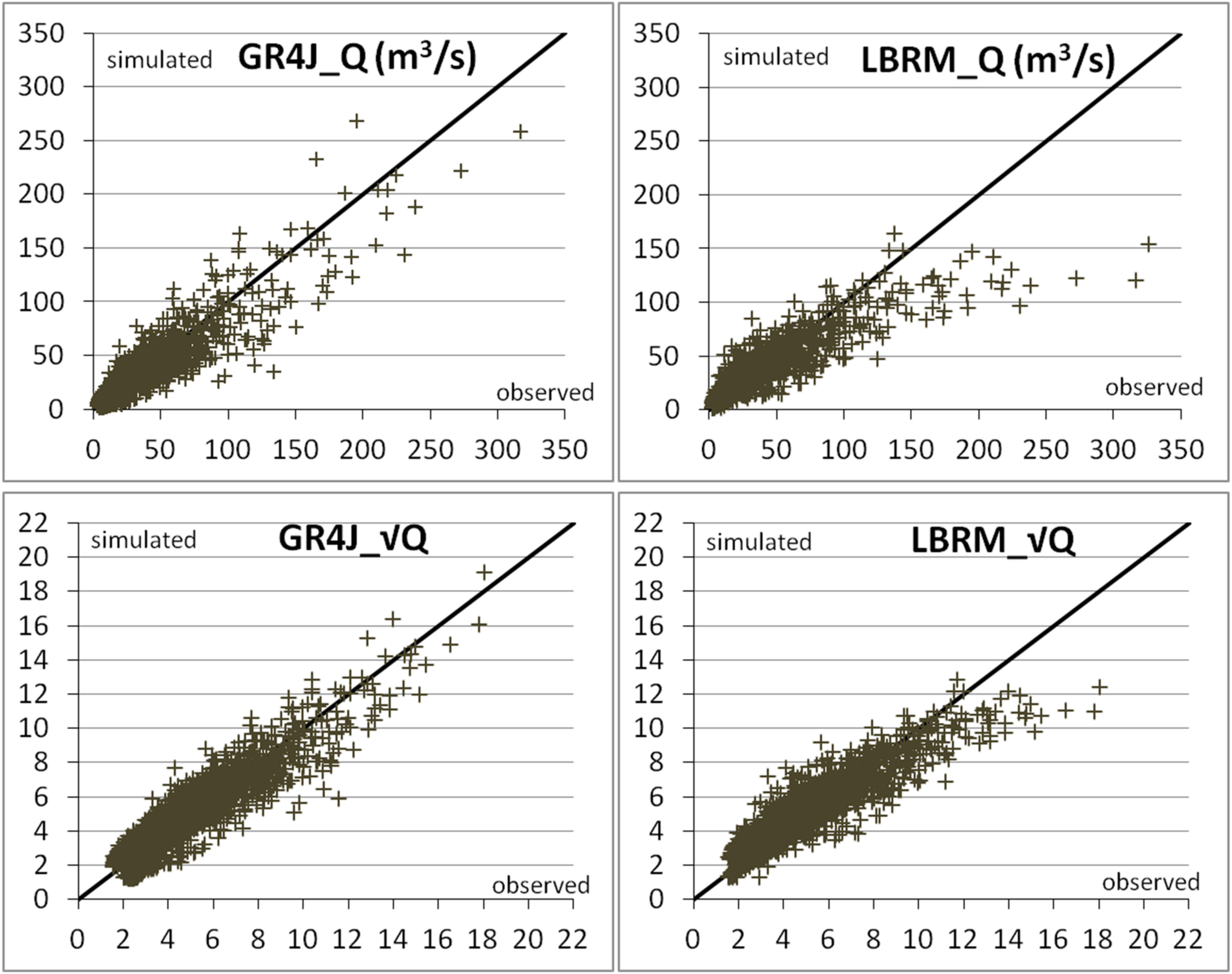
Figure 2: Hydrographs in calibration for the Moira River (subbasin 11), derived from empirical measurements, the GR4J model (*modèle du Génie Rural à 4 paramètres Journalier*), and the LBRM model (Large-Basin Runoff Model). GHCND (Global Historical Climatology Network- Daily) precipitation data used.

Figure 3: Subbasin 15 plots of simulated versus observed flows (Q) or square-root of the flows (√Q) for the whole GRIP-O period and with the GHCND precipitation data. GR4J = *modèle du Génie Rural à 4 paramètres Journalier*, LBRM = Large-Basin Runoff Model.

Figure 4: comparison between hydrographs for the 57,460 km^2 Lake Ontario area obtained either with summing runoff from the 14 local GR4J models (GR4J_locals) or with the unique GR4J model (GR4J_unique) calibrated over the whole area. In both cases, runoff was produced with using the GHCND precipitation.


CANADA
Cornwall
10 bis
12
11
10
9
8
7
6
15
14
13
13 bis
4 bis
4
2
1
3
5
UNITED STATES
N
0
50
100
200 Kilometers
Legend
GLAHF subbasins
GRIPO subbasins
Gauge stations
Ungauged areas
Gauged areas


Streamflow (m³/s)
350
300
250
200
150
100
50
0
Observations
GR4J
LBRM
6/2/2007
8/2/2007
10/2/2007
12/2/2007
2/2/2008
4/2/2008
6/2/2008
8/2/2008
10/2/2008
12/2/2008
2/2/2009
4/2/2009
6/2/2009
8/2/2009
10/2/2009
12/2/2009
2/2/2010
4/2/2010
6/2/2010
8/2/2010
10/2/2010
12/2/2010
2/2/2011
4/2/2011
6/2/2011
8/2/2011
Date (m/d/year)


simulated
GR4J_Q (m^3/s)
observed
0
50
100
150
200
250
300
350
simulated
LBRM_Q (m^3/s)
observed
0
50
100
150
200
250
300
350
simulated
GR4J_√Q
observed
0
2
4
6
8
10
12
14
16
18
20
22
simulated
LBRM_√Q
observed
0
2
4
6
8
10
12
14
16
18
20
22

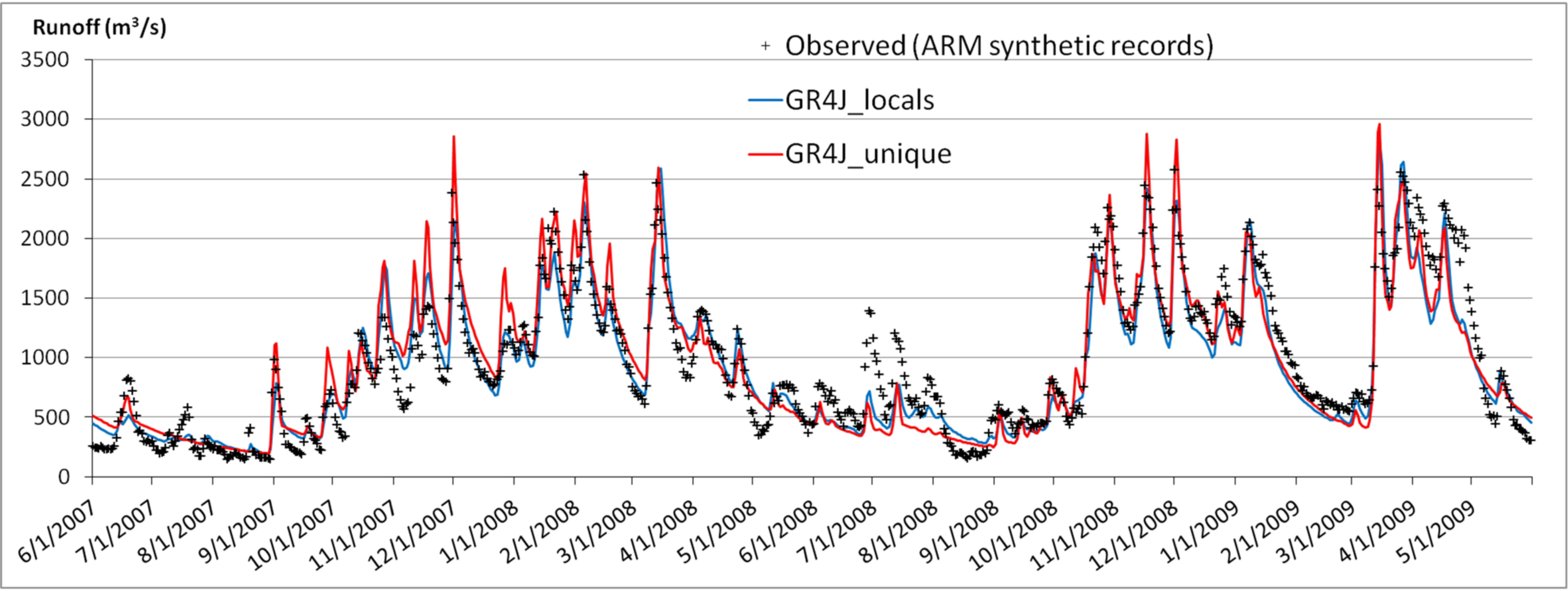

Runoff (m³/s)
+ Observed (ARM synthetic records)
GR4J_locals
GR4J_unique
3500
3000
2500
2000
1500
1000
500
0
6/1/2007
7/1/2007
8/1/2007
9/1/2007
10/1/2007
11/1/2007
12/1/2007
1/1/2008
2/1/2008
3/1/2008
4/1/2008
5/1/2008
6/1/2008
7/1/2008
8/1/2008
9/1/2008
10/1/2008
11/1/2008
12/1/2008
1/1/2009
2/1/2009
3/1/2009
4/1/2009
5/1/2009

Table 1: Number of precipitation ground stations per GRIPO subbasin (see Figure 1) for the GHCND and CaPA datasets.

			number of rain gauges			density (gauges/1000 km^2)		
		Area						
Subbasin	Country	(km^2)	CaPA	GHCND (2004)	GHCND (2011)	CaPA	GHCND (2004)	GHCND (2011)
1	CA	1087	5	7	3	5	6	3
3	USA	6455	15	17	22	2	3	3
4bis	USA	450	0	2	5	0	4	11
5	USA	13928	20	27	74	1	2	5
6	USA	2406	0	2	3	0	1	1
7	USA	5917	4	9	15	1	2	3
8	USA	4977	3	4	6	1	1	1
10	CA	2688	0	4	2	0	1	1
10bis	CA	2062	1	2	1	0	1	0
11	CA	2853	1	5	3	0	2	1
12	CA	12516	2	15	9	0	1	1
13	CA	1538	2	4	2	1	3	1
14	CA	2689	5	11	7	2	4	3
15	CA	2246	2	7	5	1	3	2

Table 2: Parameter ranges used in calibration; one of the three GR4J trials was performed with GR4J default initial parameter values indicated here; other trials were made with a random initial parameter set; coeff: coefficient; UH: Unit Hydrograph; USZ: Upper Soil Zone; LSZ: Lower Soil Zone.

LBRM free parameters (10)	unit	low bound	upper bound	GR4J free parameters (6)	unit	low bound	upper bound
Tbase	˚C	1.00E-04	10	x1: Capacity of production store	mm	10	2500
Snowmelt factor	cm/˚C/day	1.00E-04	5	x2: Water exchange coeff	mm	-15	10
linear reservoir coeff: percolation	day^{-1}	1.00E-04	100	x3: Capacity of routing store	mm	10	700
partial linear reservoir coeff: USZ evap	m^{-3}	1.00E-09	1.00E-06	x4: UH time base	day	0	7
linear reservoir coeff: interflow	day^{-1}	1.00E-04	10	x5: degree-day factor	-	1	30
linear reservoir coeff: deep percolation	day^{-1}	1.00E-05	10	x6: snowpack inertia factor	-	0	1
partial linear reservoir coeff: LSZ evap.	m^{-3}	1.00E-12	5.00E-07				
linear reservoir coeff: groundwater	day^{-1}	1.00E-06	1				
linear reservoir coeff: surface flow	day^{-1}	1.00E-03	50				
USZ Capacity	cm	1.00E+00	100				

Table 3: Evaluation metrics obtained with GR4J and the GHCND precipitation dataset. All values are in percent; optimal value is 100 except for PBIAS (0). NSE √, NSE Ln: NSE values computed with either the square-root or log of the flows, respectively. See section on *methods* for more details. NSE=Nash-Sutcliffe Efficiency index, PBIAS=Percent Bias, CA=Canada.

								CAL				VAL			
Subbasin		Cal.				Flow	mean		NSE	NSE			NSE	NSE	
#	country	scheme	Station	% gauged	Area(km^2)	regime	elev. (m)	NSE	√	Ln	PBIAS	NSE	√	Ln	PBIAS
1	CA	2	20 mile	N/A	307	natural	198	77.0	79.8	76.5	16.3	79.1	82.7	76.3	10.6
3	USA	2	Genessee	N/A	6317	regulated	418	81.0	83.1	79.0	2.4	81.6	83.9	82.9	-1.2
4bis	USA	2	Irondequoit	N/A	326	natural	172	66.8	71.4	65.7	2.4	55.3	61.5	62.6	-0.5
5	USA	2	Oswego	N/A	13287	regulated	259	83.7	83.2	79.4	2.2	69.3	71.1	69.1	3.9
6	USA	1	N/A	40	2406	mixed	264	71.4	76.2	75.8	2.3	63.5	70.6	74.2	12.8
7	USA	2	Black river	N/A	4847	regulated	471	79.2	81.9	81.7	2.3	72.5	74.6	74.7	3.4
8	USA	2	Oswegatchie	N/A	2543	regulated	250	78.5	82.1	83.7	1.6	68.2	76.9	79.9	-1.6
10	CA	2	Salmon CA	N/A	912	regulated	196	86.2	88.3	81.1	7.0	82.8	83.4	72.2	1.7
10bis	CA	1	N/A	44.2	944	mixed	115	82.2	88.7	84.8	7.9	76.6	82.4	81.2	1.8
11	CA	2	Moira	N/A	2582	regulated	228	88.2	89.6	84.3	4.3	83.6	82.7	75.5	3.0
12	CA	1	N/A	88	12515.5	regulated	282	73.1	72.8	66.1	4.3	62.0	67.2	62.7	-5.0
13	CA	1	N/A	40.3	1537.5	natural	178	62.2	66.4	63.1	2.8	57.7	62.4	61.3	8.3
14	CA	1	N/A	61.3	2689.4	mixed	209	78.8	78.3	70.9	4.6	71.5	76.0	76.4	6.9
15	CA	1	N/A	63	2245.8	mixed	263	80.4	81.7	78.2	-0.5	66.5	68.3	67.0	-0.4

Table 4: NSE (Nash-Sutcliffe Efficiency index) values obtained with GR4J, minus NSE values from LBRM. CALIBR=calibration,VALID=validation, GHCND=Global Historical Climatology Network- Daily, CaPA=Canadian Precipitation Analysis. Values are in percent, so that a value of 5 in the Table corresponds for example to a NSE of 75% for GR4J and 70% for LBRM.

		GHCND precipitation				CAPA precipitation			
		NSE √		NSE		NSE √		NSE	
Subbasin	Country	CALIBR	VALID	CALIBR	VALID	CALIBR	VALID	CALIBR	VALID
1	CA	9.1	9.7	12.8	18.6	9.1	13.6	14.6	21.7
3	US	2.9	8.5	3.3	8.6	1.0	0.5	1.4	2.1
4bis	US	-2.6	5.9	4.0	9.8	-1.6	10.2	3.5	19.0
5	US	-4.0	0.0	-4.6	1.1	-6.0	0.1	-7.2	-0.8
6	US	8.1	9.8	10.3	12.0	2.3	5.1	4.0	6.9
7	US	4.1	4.9	5.4	9.7	-1.4	6.5	-5.5	5.9
8	US	5.8	13.7	3.6	7.4	0.6	10.8	-0.8	12.6
10	CA	-1.9	3.3	-2.6	5.2	-1.5	2.8	-5.9	2.6
10bis	CA	2.8	2.6	4.0	1.1	1.6	6.7	-2.0	7.6
11	CA	0.0	4.7	4.4	14.5	0.6	5.1	2.0	6.0
12	CA	-2.4	2.4	0.4	7.8	-2.8	4.9	0.2	10.1
13	CA	0.5	-1.8	9.4	1.4	5.6	4.2	10.5	2.9
14	CA	0.5	3.0	4.8	5.5	2.7	5.5	8.1	8.8
15	CA	-3.7	-8.7	2.0	-4.4	0.8	6.0	8.0	8.1
	median US	3.5	7.2	3.8	9.1	-0.4	5.8	0.3	6.4
	median CA	0.2	2.8	4.2	5.4	1.2	5.3	5.0	7.8
	median	0.5	4.0	4.0	7.6	0.7	5.3	1.7	7.3
number of positive cases		9/14	12/14	12/14	13/14	9/14	14/14	9/14	13/14

Table 5. The number of positive cases associated to the inter-comparison study, i.e., the number of cases with the GR4J model or GHCND precipitation showing better performances than LBRM or CaPA, respectively. For the comparison of GR4J vs LBRM or GHCND vs CaPA, the number of cases displayed in Table 6 was derived from the mean of performances along the two precipitation datasets or models, respectively. The p-value of the two-sided sign test is displayed in parenthesis. A number equal or greater than 12 out of 14 cases is considered significant (i.e., a p-value of less than 5%). CALIBR=calibration, VALID=validation, GHCND=Global Historical Climatology Network- Daily, CaPA=Canadian Precipitation Analysis.

	NSE √		NSE	
	CALIBR	VALID	CALIBR	VALID
GR4J vs LBRM, mean GHNCD-CaPA	9/14 (p=0.42)	13/14 (p<0.01)	12/14 (p=0.01)	14/14 (p<0.01)
GHCND vs CaPA, mean GR4J-LBRM	12/14 (p=0.01)	5/14 (p=0.42)	13/14 (p<0.01)	7/14 (p=1)

Table 6: NSE √ and NSE (Nash-Sutcliffe Efficiency index) values obtained with the GHCND precipitation dataset, minus NSE values obtained with CaPA precipitation, in calibration (CALIBR) or validation (VALID), and with the LBRM or GR4J models. Values are in percent, so that a value of 5 corresponds, for example, to a NSE of 70 %with CaPA and of 75 % with GHCND precipitation.

		LBRM				GR4J			
		NSE √		NSE		NSE √		NSE	
Subbasin	Country	CALIBR	VALID	CALIBR	VALID	CALIBR	VALID	CALIBR	VALID
1	CA	0.1	6.4	0.9	4.7	0.1	2.6	-0.9	1.6
3	US	1.1	2.4	0.6	2.3	3.0	10.5	2.5	8.8
4bis	US	3.9	-3.4	6.3	-0.7	2.9	-7.7	6.7	-9.8
5	US	1.8	3.2	1.9	3.7	3.8	3.1	4.5	5.7
6	US	-0.2	-2.5	1.2	-5.9	5.6	2.2	7.5	-0.9
7	US	3.9	-0.2	3.5	-1.7	9.4	-1.8	14.4	2.1
8	US	5.8	-3.5	6.4	0.0	11.0	-0.6	10.7	-5.2
10	CA	2.1	0.6	2.3	0.8	1.7	1.0	5.5	3.4
10bis	CA	1.0	1.7	1.0	3.0	2.1	-2.5	7.0	-3.5
11	CA	2.2	-0.4	-0.2	-3.7	1.6	-0.8	2.1	4.8
12	CA	2.6	-0.6	2.1	-4.1	3.0	-3.2	2.3	-6.5
13	CA	1.6	0.5	-1.3	-2.3	-3.5	-5.4	-2.3	-3.8
14	CA	1.9	5.5	2.4	6.6	-0.3	3.0	-1.0	3.3
15	CA	0.9	-1.8	1.7	-3.8	-1.6	-16.5	1.5	-16.2
	median U.S.	2.8	-1.4	2.7	-0.3	4.7	0.8	7.1	0.6
	median CA	1.7	0.6	1.3	-0.7	0.9	-1.7	1.8	-1.0
	median	1.9	0.1	1.8	-0.3	2.5	-0.7	3.5	0.4
	number of positive cases	13/14	7/14	12/14	7/14	11/14	6/14	11/14	7/14

Table 7: Statistics for the LBRM (Large-Basin Runoff Model) parameters after calibration taking all cases into account (28: 14 catchments and 2 precipitation datasets). Lin= linear, res= reservoir, coeff= coefficient, evap= evaporation, perco= percolation, USZ= Upper Soil Zone, std= standard deviation, LSZ= Lower Soil Zone, Tbase = base temperature, val. = value.

Parameters / statistics	Tbase	Snowmelt factor	Lin. res. coeff: perc.	Partial lin. res. coeff: USZ evap	Lin. res. coeff: interflow	Lin. res. coeff: deep perco.	Partial lin. res. coeff: LSZ evap	Lin. res. coeff: groundwater	Lin. res. coeff: surface flow	USZ thickness
	˚C	cm/˚C/d	day^{-1}	m^{-3}	day^{-1}	day^{-1}	m^{-3}	day^{-1}	day^{-1}	cm
mean	6.2	4.4E-01	8.0E-02	1.4E-07	6.5E-01	6.6E-01	1.0E-07	2.2E-02	2.3E-01	15.2
Lowest val.	3.8	7.1E-04	1.5E-03	1.0E-09	1.6E-03	1.0E-05	1.0E-12	1.1E-06	5.7E-02	2.2
Highest val.	9.7	1.0E+00	1.7E+00	9.9E-07	8.6E+00	7.8E+00	5.0E-07	4.3E-01	9.9E-01	77.0
std	2.0	2.0E-01	3.1E-01	2.6E-07	1.7E+00	1.7E+00	1.9E-07	8.2E-02	2.3E-01	15.3

Table 8: Statistics for the GR4J (*modèle du Génie Rural à 4 paramètres Journalier*) parameters after calibration separating the 28 final parameter sets in two cases depending on the relative values of X1 and X3, either of similar magnitude (or X1 > X3), or with X1 way lower than X3 (X1 << X3).

	Parameters / statistics	X1: Capacity of production store (mm)	X2: Water exchange coefficient (mm)	X3: Capacity of routing store (mm)	X4: UH time base (days)	X5: degree-day factor	X6: snowpack inertia factor
X1 higher or similar to X3, 19 cases	mean	561.12	-0.36	97.87	2.63	7.36	0.06
	lowest	69.06	-3.93	13.03	1.22	3.35	0.00
	highest	2089.42	2.93	288.86	5.50	18.12	0.31
	std	566.67	1.76	95.15	1.26	3.51	0.09
X1 << X3, 9 cases	mean	21.11	-4.23	506.06	3.64	8.47	0.09
	lowest	12.42	-14.22	362.33	1.31	6.61	0.00
	highest	49.37	2.59	693.86	6.55	10.53	0.35
	std	11.89	5.21	118.06	1.81	1.29	0.11

Table 9: Daily runoff simulation performances when assessed for the 57,460-km^2 Lake Ontario area (see text). CAL= calibration, VAL= validation, NSE= Nash-Sutcliffe Efficiency index. The mean of runoff simulated with the four different combination possibilities (mean GR4J-LBRM-CAPA-GHCND) was assessed. The last column shows performances obtained with a unique GR4J model applied to this large area using GHCND precipitation.

	GHCND				CaPA				mean GR4J - LBRM - GHCND - CaPA		GR4J unique GHCND	
	CAL		VAL		CAL		VAL		CAL	VAL	CAL	VAL
	GR4J	LBRM	GR4J	LBRM	GR4J	LBRM	GR4J	LBRM				
NSE √	0.92	0.90	0.87	0.82	0.87	0.89	0.86	0.81	0.92	0.86	0.88	0.80
NSE	0.92	0.91	0.88	0.84	0.88	0.90	0.86	0.83	0.92	0.88	0.89	0.81
NSE ln	0.91	0.91	0.87	0.85	0.87	0.90	0.85	0.83	0.91	0.87	0.88	0.80
PBIAS	2.95	4.49	2.38	6.80	1.96	1.64	2.58	4.03	2.76	3.95	-0.32	-1.57