Great Lakes Runoff Inter-comparison Project for Lake Ontario (GRIP-O)

Final Report (updated version)

To:

The International Watershed Initiatives of the International Joint Commission

Environment Canada, Dorval, April 7th, 2016.

Étienne Gaborit[1,*], Vincent Fortin[1], Bryan Tolson[2]

[1] Environment Canada, Environmental Numerical Prediction Research (E-NPR), 2121 transcanadian highway, Dorval H9P1J3, QC, Canada.

[2] University of Waterloo, Civil and Environmental Engineering Dpt., Waterloo N2L3G1, ON, Canada.

contact: etienne.gaborit@canada.ca

Table of content

Foreword

This report aims at describing the work performed during the Great Lakes Runoff Inter-comparison Project for lake Ontario (GRIP-O); it follows four previous reports produced as part of the project, namely a data report (March 31st, 2014), a detailed experiment plan (September 30, 2014), a progress report (April 30, 2015), and a final report which was released on September 30th, 2015. This document consists of an updated version of the final report in order to include the last GRIP-O results obtained after the official end of the project (September 2015). It was released on April 21st, 2016. Compared to the official final report released in September 2015, this report includes many additional results, for example regarding the calibrations of the SVS surface scheme (embedded in GEM-Hydro or GEM-Surf + UH, see section 8.6), the comparison between calibrated semi-distributed models (Watflood, SA-MESH and GEM-Hydro, see section 8.7), the comparison between SVS and ISBA surface schemes (section 8.4), an improvement brought to SVS snowmelt formulation (section 8.3), or the methodology proposed to estimate runoff for the whole Lake Ontario watershed, including its ungauged parts (see section 9).

List of acronyms

ARM	Area-Ratio Method
CANSIS	Canadian Soil Information Service
CaPA	Canadian Precipitation Analysis
CLASS	Canadian LAnd Surface Scheme
DDS	Dynamically Dimensioned Search
EC	Environment Canada
ESA	European Space Agency
GEM-Hydro	Global Environmental Multi-scale weather model + hydrology - hydraulic model
GEM-Surf+UH	Global Environmental Multi-scale weather model + SURFace and Unit Hydrograph
GHCND	Global Historical Climatology Network - Daily
GLERL	Great Lakes Environmental Research Laboratory
GR4J	Génie Rural à 4 paramètres Journalier
GRIP-M	Great Lakes Runoff Inter-comparison Project for lake Michigan
GRIP-O	Great Lakes Runoff Inter-comparison Project for lake Ontario
GSDE	Global Soil Dataset for Earth System modeling
HWSD	Harmonized World Soil Dataset
ISBA	*Interaction Soil-Biosphère-Atmosphère*
LBRM	Large Basin Runoff Model
LSS	Land Surface Scheme
LULC	Land Use / Land Cover
LZS	Lower Zone Storage
MESH	*Modélisation Environnementale de la Surface et de l'Hydrologie*
NASA	National Aeronautics and Space Administration
NBS	Net Basin Supply
NCDC	National Climate Data Center
NGA	National Geospatial-Intelligence Agency
NOAA	National Oceanic and Atmospheric Administration
NRCS	Natural Resources Conservation Service
NWS	National Weather Service
NWIS	National Water Information System
NYSB	New-York State Barge
P. or precip.	Precipitation
RDPS	Regional Deterministic Prediction System
SA-MESH	Stand-Alone MESH
SPS	Surface Prediction System
SRTM	Shuttle Radar Topography Mission
STATSGO	State Soil Geographic

SVS	Soil, Vegetation and Snow
SWE	Snow Water Equivalent
T. or temp.	Temperature
UH	Unit Hydrograph
USGS	U.S. Geological Survey
WWF	World Wildlife Fund

1-Introduction

The project aims at comparing different hydrologic models in their ability to estimate lake Ontario's direct incoming runoff; it means that only the lake's direct watershed is modeled (and not the other upstream Great Lakes) as part of this project. The emphasis is therefore put on the land part of the trans-boundary lake Ontario watershed, which is considered to extend down to the St-Lawrence river flow regulation dam located near Cornwall, ON.

The Great Lakes Runoff Inter-Comparison Project for lake Ontario (GRIP-O) consists in the second phase of the GRIP project. The first phase, named GRIP-M (Fry *et al.* 2014), focused on lake Michigan and involved the comparison between several very different hydrologic models in their ability to simulate lake Michigan's tributary flows. The models used during GRIP-M include the National Oceanic and Atmospheric Administration's (NOAA) Large Basin Runoff Model (LBRM), the National Weather Service Sacramento soil moisture accounting model (NWS), and several configurations of Environment Canada's (EC) distributed MESH model (for *Modélisation Environnementale Communautaire* - Surface and Hydrology). These models were however embedded within hydrological prediction systems which used different meteorological inputs, geophysical data, and calibration procedures.

The main aim of GRIP-O is to compare different hydrologic models using the same forcings and calibration procedure, in order to allow a fair comparison between models and to identify the most promising one (for the area under study), depending on the targeted application. For example, lumped models can be useful to estimate runoff quantities from one area in an efficient manner (they are fast to implement and require small computation time), but cannot provide spatially detailed estimations of runoff, which are required for many environmental applications such as the generation of inundation maps. The lake Ontario watershed offers interesting challenges to hydrological modeling, essentially because most of its tributaries possess a regulated flow regime, and because as in GRIP-M, runoff has to be estimated for the ungauged areas of the watershed.

The ultimate aim of GRIP-O is to develop an efficient and reliable hydrologic simulation platform to estimate runoff from the whole lake Ontario Watershed, including its ungauged parts, and which could be used in many different environmental applications.

To do so, different sources of precipitation data are also compared. The improvements made in the lake's runoff simulation will benefit the International Great Lakes - St. Lawrence River Adaptive Management Task Team and the International St. Lawrence River Board of Control (through the Coordinating Committee for Great Lakes Basic Hydrologic and Hydraulic Data) by providing them with an efficient modelling tool having various potential environmental applications in the general field of water resources management, such as lake Ontario level and St. Lawrence streamflow forecasting, climate-change impact studies, dam management, etc.

Such applications could actually be synthesized by a general will to promote an integrated and optimized freshwater resource management in the Great Lakes region.

This report is organized as follows: section 2 describes the main hydrologic models that have been implemented in the context of the GRIP-O project, along with their specificities and required data. Section 3 presents the data selected to drive the models (the forcings), section 4 is devoted to the physiographic data used for the distributed models; section 5 presents the calibration details (spatial framework, objective function...); section 6 describes some sensitivity tests that were performed with GEM-Hydro, section 7 presents the results obtained with the lumped models, section 8 shows the results of the distributed models, section 9 presents the methodology followed to estimate runoff for the whole Lake Ontario watershed (including ungauged areas), section 10 performs a synthesis of the main results obtained during this work, while recommendations for future research directions are finally presented in section 11.

2- Hydrologic models

2.1) overview of the models used during GRIP-O

The three different main models considered in this study consist in the LBRM (see Croley and He 2002), the lumped conceptual "Génie Rural à 4 paramètres Journalier" (GR4J) model (Perrin et. al 2003), and the distributed physically-based Surface Prediction System (SPS) developed by Environment Canada, which was renamed GEM-Hydro recently (for the Global Environmental Multi-scale weather model + HYDROlogy - hydraulic model).

GR4J is a daily continuous lumped hydrologic model with four free parameters (Perrin *et al.* 2003), developed at IRSTEA (France). It basically relies on two tanks which represent the soil and routing reservoirs of a catchment, and on the Unit Hydrograph (UH) theory to produce streamflows. It was here coupled with the 2 free-parameter CémaNeige snow module (Valéry 2010; Nicolle *et al.* 2011).

The LBRM is also a daily continuous lumped hydrologic model; it includes a snow module and has a total of 9 free parameters by default. Originally developed by NOAA-Great Lakes Environmental Research Laboratory (GLERL), it simulates water transport through a series of cascading tanks (for a detailed description of the model, see Croley and He, 2002). It has been employed in a variety of research-based 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 maximum and minimum Temperature, as well as the catchment area (to convert mm to m^3/s). GR4J also requires the mean watershed latitude used in the Potential Evapo-Transpiration formulation, and the mean watershed elevation (up to 5 elevation classes can be defined in the *Cémaneige* snow module). It was here implemented using a unique elevation class for each GRIP-O sub-watershed, but with five elevation classes for the unique model implemented over the whole Lake Ontario area (see sections 7.3.4 and 9). See Table 1 for the data required by the different models.

The distributed SPS model developed at EC relies either on SVS (Soil, Vegetation and Snow scheme) or ISBA (*Interaction Soil-Biosphère-Atmosphère*) Land Surface Schemes (LSS) for the representation of surface processes, and on WATROUTE for the routing part. The whole ensemble was recently renamed "GEM-Hydro", which consists of SPS and WATROUTE. Throughout this report however, the term "SPS" will sometimes be used to designate the GEM-Hydro platform.

WATROUTE is a 1-D hydraulic model (Kouwen 2010) mainly relying on flow direction grids and elevation data. It allows routing surface runoff and recharge, produced by surface schemes, to the catchment outlet. In Watroute, runoff directly feeds the streams while recharge is first provided to a Lower Zone Storage (LZS) compartment, representing superficial aquifers, and which then releases the water to the streams.

Two other modeling platforms were implemented on some of the GRIP-O sub-catchments for a direct comparison to SPS results, namely the Stand-Alone MESH (SA-MESH, implemented by the University of Waterloo) and WATFLOOD models (Kouwen 2010). SA-MESH, Watflood and GEM-Hydro have in common a semi-distributed representation of some of the hydrological processes occurring in a watershed, and the fact that they rely on two main parts: a LSS for the representation of surface processes (Evapo-transpiration, infiltration, snow processes, water circulation in the soil), and the river routing scheme WATROUTE for simulating water transport in the streams.

SA-MESH is based on the CLASS (Canadian LAnd Surface Scheme) and consists in a research version of SPS available to the public, but it is generally not updated when new SPS versions arise. WATFLOOD can be seen as an intermediate model, lying somewhere between the lumped and fully distributed models. Both WATFLOOD and SA-MESH were here implemented with a 10 arcmin. spatial resolution (both for the LSS and the routing scheme), while GEM-Hydro was here implemented with a 2 or 10 arcmin. resolution for the LSS and 0.5 arcmin. (30 arcsec.) for the routing scheme (see section 6 for tests on the sensitivity of GEM-Hydro and WATROUTE spatial resolutions). Table 1 below synthesizes each model main characteristics and required data.

Table 1: Main GRIP-O models' characteristics; P: Precipitation, T: Temperature, H: Humidity, R: Radiation, W: Wind speed and direction, Ps: Pressure, LULC: Land Use/Land Cover, Topo: topographic data such as elevation or slope, Flow Dir: Flow Direction grid

Model name	Underlying theory	Spatial distribution	Time-step	Forcing data	Physiographic data
LBRM	Conceptual	lumped	daily	P, T	None
GR4J	Conceptual	lumped	daily	P, T	mean elevation and latitude
WATFLOOD	Physico-Conceptual	distributed	flexible	P, T	LULC, Topo, Flow Dir
GEM-Hydro (SPS)	Physical	distributed	flexible	P, T, H, R, W, Ps	LULC , Soil, Topo, Flow Dir
SA-MESH: CLASS+WATROUTE	Physical	distributed	flexible	P, T, H, R, W, Ps	LULC, Soil, Topo, Flow Dir

2.2) SVS versions

As the GRIP-O project started in January 2014 and officially finished in September 2015 (but this document was even updated in April 2016), the SVS LSS, which was mainly used with GEM-Hydro (the other option is the ISBA LSS), was the subject of many bug fixes, changes in the configuration options, improvements, etc. occurring during the work. Therefore, different versions of SVS were used during this work. Table **2** below synthesizes the main SVS versions used, the details associated to each version, and the section which contains the associated results. See also Table **12** which associates the different SVS versions with the different calibration experiments performed during this work.

Table 2: The different SVS versions used during this work along with their associated characteristics and the identification of sections containing the corresponding results.

version name	former	intermediate	new	new_2
canopy bug	present	fixed	fixed	fixed
baseflow bug	present	fixed	fixed	fixed
snow bug	present	present	partially fixed	partially fixed
soil freezing	on	off	off	off
vegetation class 26 error	present	fixed	fixed	fixed
max. soil depth	2.3m	2.3m	1.4m	1.4m
lateral flow removing excess water	on	on	off	off
snowmelt improvement	off	off	off	on
urban runoff improvement	off	off	off	on
name of platform used	GEM-Hydro	GEM-Hydro	GEM-Hydro	GEM-Surf + UH
section where used	8.1	8.1.3	8.2	8.3 to 8.7, 9

3- Forcing data

While the lumped models can be driven by daily watershed averages of Precipitation (P. or precip.) and Temperature (T. or temp.), the distributed ones require gridded information of several surface variables (see Table 1).

For GRIP-O, three different sources were used for gridded P. data (see Table **5** for the dataset sources). The first source of precipitation consists in the Canadian Precipitation Analysis (CaPA, see Mahfouf *et al.* 2007), a system relying on modeled precipitation fields derived from the Canadian Regional Deterministic Prediction System (RDPS, ≈15 km resolution), but corrected with ground-based precipitation observations. More precisely, we chose to use the 24h 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 available yet for CaPA 3.0 over the time period of interest to GRIP-O (see section 5.3 below). Both 6 and 24h accumulations are available from CaPA. It was decided to use the 24-hourly product because more ground stations are generally available in real-time at the 24-h interval than at the 6-hourly interval, making the 24-h CaPA product more tied down to observations than the 6-hourly 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 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 Canadian stations, US cooperative stations, etc. The data were interpolated on a 15 arcsec. grid (around 450m near lake Ontario) using a nearest neighbor (or Thiessen polygons) method. It was however noticed, when visualizing the data, that most US stations contain precipitation values corresponding to 24-h accumulations between around 12 UTC of the previous day and 12 UTC of the data's indicated day, whereas Canadian stations report data corresponding to accumulations between around 12 UTC of the data's indicated day and 12 UTC of the next day. Hence a one-day shift exists between the U.S. and Canadian data in the original GHCND precipitation database, which was corrected to harmonize the data when computing the daily basin averages for driving the lumped models. Note that the GHCND data also contain daily minimum and maximum 2-m temperature data.

The third set of gridded precipitation data consists in the precipitation of the RDPS (spatial resolution between 15 and 10 km depending on the time period) used to drive SPS. This source of gridded precipitation data was only used during the first implementation of SPS (see section 8.1 and Table **12**), and not with any of the other models (see Table 4). The model

precipitation contains generally more error than the CaPA or GHCND precipitation, but a test was however made to assess the simulation performances with SPS when driven with model precipitation.

See Table 4 for a summary of which precip. data were used with each model.

Table 3 shows the number of ground stations per GRIP-O sub-catchment (

Figure 1) for the CaPA and GHCND datasets. The density of the GHCND is higher than the one of the stations used in CaPA, because CaPA only uses stations for which data is available in real-time at the Canadian meteorological centre. 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 (see Table **3**).

P. and T. watershed averages were computed based on the sub-basin shapes, either for the whole sub-basin in case of calibration scheme 1 or for its gauged area for scheme 2 (see section 5.2 below), in order to force the lumped models. Lumped models were forced with temperature from the GHCND (same stations as the precipitation ones), even in the case of feeding the models with CaPA precipitation, because CaPA does not include T. data.

Table 3: Number of precipitation ground stations per GRIPO sub-catchment (see

Figure 1) for the GHCND and CaPA datasets.

			number of rain gauges			density (gauges/1000 km^2)		
Country	Sub-basin	Area (km^2)	CaPA	GHCND_2004	GHCND_2011	CaPA	GHCND_2004	GHCND_2011
CA	1	1087	5	7	3	5	6	3
USA	3	6455	15	17	22	2	3	3
USA	4bis	450	0	2	5	0	4	11
USA	5	13928	20	27	74	1	2	5
USA	6	2406	0	2	3	0	1	1
USA	7	5917	4	9	15	1	2	3
USA	8	4977	3	4	6	1	1	1
CA	10	2688	0	4	2	0	1	1
CA	10bis	2062	1	2	1	0	1	0
CA	11	2853	1	5	3	0	2	1
CA	12	12516	2	15	9	0	1	1
CA	13	1538	2	4	2	1	3	1
CA	14	2689	5	11	7	2	4	3
CA	15	2246	2	7	5	1	3	2

The SPS and MESH models both require, in addition to precipitation and temperature, gridded fields (at the surface) of pressure, relative humidity, long and shortwave solar radiations, and wind speed and direction. These data were derived from the RDPS which is run

at EC. The model having two updates per day over the GRIP-O studied period, the derived gridded surface fields should not see their quality significantly deteriorate due to the horizon increase. For temperature however, MESH and SPS require the variable at the elevation of 40 m instead of the surface (2-m) temperature needed by most hydrologic models. The 40-m temperature was also derived from the RDPS to drive distributed models.

While it is possible to force distributed models with GHCND observed precipitation, it is more tricky to do so with the observed surface (2-m) temperature because this variable is an output of the distributed surface schemes, which use 40-m temp. as input. Doing so would therefore probably result in inconsistencies between the model specific humidity and the 2-m temperature data, possibly leading to a model crash or numerical instabilities (V. Fortin, personal communication). Therefore, distributed models will, during GRIP-O, only be driven by the regional model 40-m temperature, and not by GHCND Temperature (see Table 4).

Finally, as MESH and SPS's internal time step is generally less than one hour, the models need to perform a temporal disaggregation of the 24-h CaPA fields into hourly ones. The other forcing fields described above are already available as hourly outputs from the RDPS. This temporal disaggregation is performed in SPS by making the 24-h CaPA precip. match the temporal pattern of the hourly model precip. SA-MESH, on the other side, uses its own temporal disaggregation method.

Table 4 below presents a summary of which P. and T. data were used for each model in the context of GRIP-O.

Table 4: summary of GRIP-O performed experiments; P.: precipitation, T.: Temperature, X: not performed.

Forcings	GHCND P.		CAPA P.		MODEL P.
Models	GHCND T.	MODEL T.	GHCND T.	MODEL T.	MODEL T.
GR4J	done	X	done	X	X
LBRM	done	X	done	X	X
GEM-Hydro (SPS)	X	X	X	done	done
SA-MESH + WATROUTE	X	X	X	done	X
WATFLOOD	X	X	X	done	X

In summary, it is possible to compare the models altogether (when using CaPA precip., even if not using the same temperature inputs), and to compare the GHCND and CaPA

precipitation sources. It will not actually be possible to compare CaPA and MODEL precip. the same way as other experiments (namely, by calibrating the models) because the SVS version has changed between the two experiments (see Table **12**). However, a sensitivity test was performed over the entire GRIP-O watershed, when using either CaPA or model precip., with some default Watroute parameters and the new SVS version (see section 8.2).

4- Physiographic data

All of the distributed models listed in Table 1 need some physiographic information in order to derive some parameter values used in the model equations. For example, all of these models (SPS, SA-MESH and WATFLOOD) need the Land Use/ Land Cover information (LULC) and some topographic data such as elevation, mean pixel land slope, mean pixel channel slope, and flow directions. Based on these data, the models will derive some parameter values and/or other intermediate valuable information such as each pixel's upstream drainage area. In addition, distributed models require soil data (percent sand and clay content from which they also derive the percent silt content, soil depth) to derive the soil hydraulic properties. We refer to previous GRIP-O reports for an extensive review of the physiographic data available over the lake Ontario (direct) watershed ("report on the available data to be used"), or for the justification of the choice regarding the physiographic data used in GRIP-O ("detailed experiment plan for the model comparison"). See the list of acronyms at the beginning of this report for the acronyms not explained below.

Here we simply give a brief overview of the datasets which were selected to implement SPS: they consist in GLOBCOVER (ESA 2009, 300m resolution) for the LULC, HydroSheds (USGS and WWF) at 30 arcsec. for the flow direction grid, the GTOPO30 (USGS 1996, 900m resolution) and SRTM (NGA and NASA 2000, 90m resolution) digital elevation data for the topographic information, and the soil dataset developed by Shangguan *et al.* (2014). This dataset incorporates recent and detailed soil data wherever available: for example, in the GRIP-O study area (see

Figure 1), it uses SLC 3.2 in Canada (CANSIS), and STATSGO 2 in the U.S. (USGS). It also uses coarser datasets such as the Harmonized World Soil Dataset (HWSD, see FAO *et al.* 2012) to fill any gaps, such as in the northern part of the Lake Ontario watershed of

Figure 1 (Yet more than 90% of the watershed is covered either by the SLC 3.2 or STATSGO 2 products). The maximum soil depth is assumed to be 2.8 m everywhere, and the vertical profile

is described by 8 different layers. The data consist in the percent sand and clay contents, from which the percent silt content and associated hydraulic properties are derived inside of SVS.

The datasets described above possess several characteristics that make them very suitable to the GRIP-O project, given the implementation of SPS that it involves: they are recent; they are available at a higher or same resolution than the model grid it is intended for (this is preferable to aggregate highly detailed physiographic data than to interpolate coarse-resolution data, when comparing the resulting values to real ones); they are harmonized over the U.S. / Canadian border. It has yet to be emphasized here that these data do not consist in the most up-to-date or the most detailed data for the area studied here, but as previously mentioned, the reader is referred to former reports for more details about the available physiographic data around lake Ontario.

Table **5** below synthesizes the datasets used to provide the physiographic and atmospheric inputs required by the models. Despite the three distributed models under study rely on the same dataset for LULC, 26 different land cover classes are defined in GEM-Hydro while Watflood and SA-MESH rely only on 7 different classes, which were derived from the aggregation of the 26 classes used by GEM-Hydro. The soil texture information exploited by the models is exactly the same, as the surface compartment of all models was here implemented at a 10 arcmin. spatial resolution and the same grid point coordinates. However, the maximum soil depth was defined to 1.4m by default in SVS (GEM-Hydro), while it is set to 4.1m by default in SA-MESH. Maximum soil depth is not defined in Watflood. Hence the maximum soil depth is calibrated in GEM-Hydro and SA-MESH, but not in Watflood (see Table 9 to **Table 11**). Sensitivity tests performed with GEM-Hydro indicated that the results have a limited sensitivity to the maximum soil depth value, given that it is above around 1m for the area under study. The GSDE dataset (see Table 5) contains soil texture information down to 2.8m. Therefore, in SA-MESH, the soil texture of the last GSDE layer was used for layers below 2.8m.

Table 5: Sources of data used to feed and calibrate the distributed models; NA: North America

Dataset	Type of data	Coverage	Resolution/scale	Source
GSDE	soil texture	World	~ 1km (30'')	Shangguan *et al.* 2014
GLOBCOVER 2009	land cover	World	300m (10'')	ESA 2009
HydroSheds	Flow directions	World	~ 1km (30'')	USGS and WWF 2006
SRTM	DEM	World	90m (3'')	NGA and NASA 2000
HyDAT	Gauge stations	CA	N/A	EC
NWIS	Gauge stations	US	N/A	USGS
CaPA v2.4b8	Precipitation	NA	~ 15 km	EC
GHCND	Precipitation	World	N/A	NOAA
Regeta	Atmospheric forcings	NA	15/10 km	EC

5- Calibration methodology

5.1) Spatial framework

The GRIP-O spatial framework is defined on

Figure 1 below, which presents the different sub-watershed delineations. This delineation is slightly different from the one made by Croley (1983) for the calibration of the LBRM model: some sub-watersheds were here subdivided (4 and 4bis, 10 and 10bis, 13 and 13bis, see

Figure 1) in order for the resulting areas to be more narrow around the flow gauges they contain. This was made in order for the data of the flow gauges to better represent the total sub-basin outflow in case of calibration scheme 1 (i.e. for sub-basins containing several flow stations, see section 5.2 below). However, this leaves some areas as ungauged, such as sub-basins 2, 4, 13bis and 9 (considered ungauged despite it contains a flow station).

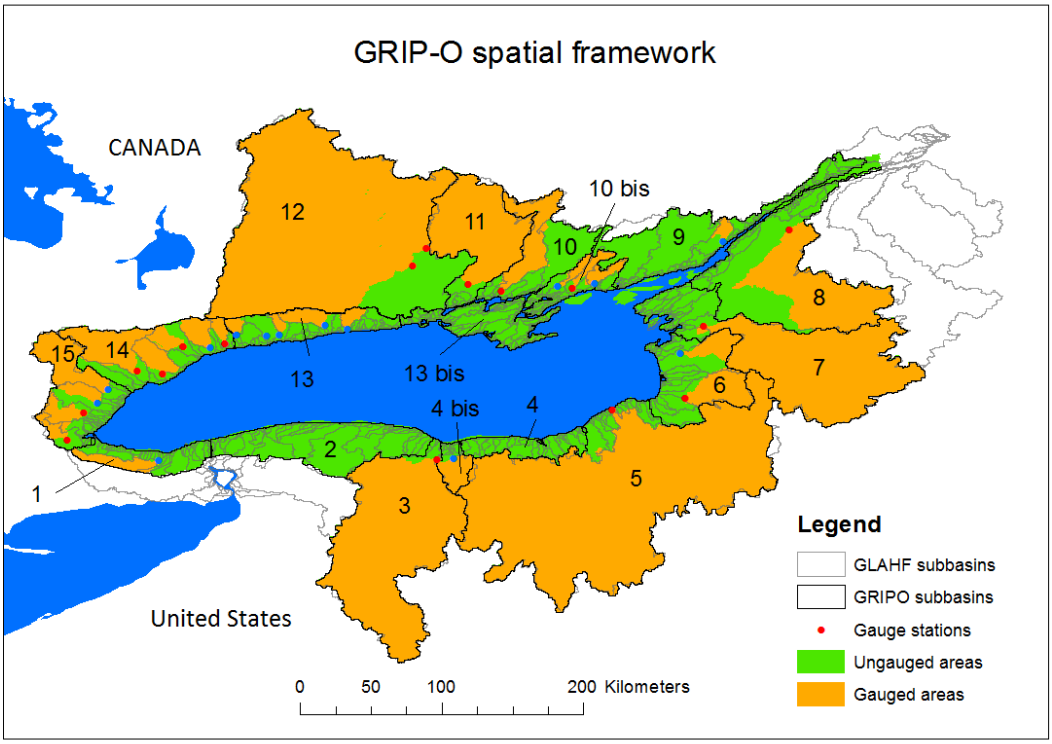

Figure 1: GRIP-O spatial framework: lake Ontario sub-watershed delineation (GRIP-O subbasins). GLAHF subbasins depict the delineation performed by the Great Lakes Aquatic Habitat Framework (U. of Michigan, see text). The main watershed outlet is located at Cornwall, ON. All dots correspond to most-downstream flow gauges selected for model calibrations; blue ones correspond to rivers with natural flow regimes while red ones are located on regulated

rivers. The orange area represents the gauged area (about 74% of the total - green - lake Ontario watershed). See Table 6 for the sub-basins' main characteristics.

The two lumped models studied here are subject to a specific calibration for each gauged GRIP-O sub-watershed identified on

Figure 1. This is more likely to lead to optimal hydrologic simulations for the different sub-watersheds than if performing global calibration over all the sub-basins at once (see for example Gaborit *et al.* 2015), in which case a unique parameter set would be used for all sub-catchments (see section 9). Moreover, the aim of GRIP-O being the inter-comparison of hydrologic models in their ability to simulate lake Ontario direct runoff, it was preferred to perform the calibrations independently for each sub-basin in order to have more case studies to assess the model performances but also to have a clear idea of which performances it is ultimately possible to reach over this area and the period and models under consideration, and how these performances vary spatially over the watershed, for example as a function of the flow regime, the country it is located in (the observation networks have not the same density on each side of the border), or the catchment physiographic properties (see section 7.3).

GLAHF sub-basins were delineated using 1 arcsec. resolution flow direction grids, while GRIP-O sub-basins rely on 30 arcsec. flow direction data. A few manual modifications to the 30 arcsec. data were required in order to correct a few discrepancies between the two resulting delineations, yet cases involving "missing areas" in the GRIP-O sub-basins were 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 streamgauges with minor missing records, regardless of the type of flow regime (i.e. natural of regulated).

Of the 30 streamgauges of

Figure 1, 27 have no missing data, 2 are complete at 94%, and the 20 mile creek gauge (sub-basin 1) is complete at 80% over the GRIP-O period (see section 5.3). No information could be found about the type of regulation performed for any of the sub-catchments of

Figure 1. Information found online suggests that regulation generally involve artificial reservoirs. A few are used for hydropower production, such as on the Trent river in sub-basin 12, but no information was found in regard of management policy or operating rules. The reader is referred to a previous GRIP-O report ("Detailed experiment plan for the model comparison") for a comprehensive review of the existing flow gauges inside the area depicted on

Figure 1, and for more details as to how the flow stations of

Figure 1 were finally selected to perform the GRIP-O calibrations.

Table 6: Selected most-downstream flow gauges' characteristics, along with their sub-basin affiliation (see

Figure **1**), the fraction of the sub-basin gauged by the station, (frac. GGD), the sub-basin's chosen calibration scheme, and the percent of available flow observations over the calibration period (% comp.).

River name	Station ID	Flow regime	Station drainage area (km^2)	Subbasin affiliation	Sub-basin area (km^2)	Frac. GGD	Calibration scheme	% comp.
20 mile	02HA006	Natural	293	1	1087	27.0	2	80.3
Genesee	4231600	Regulated	6407	3	6455	99.3	2	100.0
Irondequoit	423205010	Natural	368	4bis	449.5	81.9	2	100.0
Oswego	4249000	Regulated	13209	5	13928	94.8	2	100.0
Sandy	4250750	Natural	354	6	2406.3	14.7	1	100.0
Salmon (US)	4250200	Regulated	616	6		25.6		100.0
Black	4260500	Regulated	4827	7	5917.4	81.6	2	100.0
Oswegatchie	4263000	Regulated	2554	8	4977.3	51.3	2	100.0
Lyn	02MB006	Natural	111	9	2687.7	4.1	N/A	92.9
Salmon (CA)	02HM003	Regulated	891	10	2062	43.2	2	98.5
Moira	02HL001	Regulated	2620	11	2853.4	91.8	2	100.0
Crowe	02HK003	Regulated	1990	12	12515.5	15.9	1	99.7
Trent	02HK002	Regulated	9090	12		72.6		93.8
Bowmanville	02HD006	Natural	83	13	1537.5	5.4	1	100.0
Oshawa	02HD008	Natural	96	13		6.2		100.0
Wilmot	02HD009	Natural	83	13		5.4		100.0
Ganaraska	02HD012	Natural	232	13		15.1		100.0
Cobourg	02HD019	Natural	126.4	13		8.2		100.0
Lynde	02HC018	Natural	106	14	2689.4	3.9	1	100.0
Rouge	02HC022	Regulated	186	14		6.9		100.0

Duffins	02HC049	Natural	251	14		9.3		100.0
Don	02HC024	Regulated	316	14		11.7		100.0
Humber	02HC003	Regulated	800	14		29.7		100.0
East Oakville	02HB004	Natural	199	15	2245.8	8.9	1	100.0
Spencer	02HB007	Regulated	169	15		7.5		100.0
Bronte	02HB011	Regulated	235	15		10.5		100.0
Credit	02HB029	Natural	813	15		36.2		100.0
Wilton	02HM004	Natural	112	10bis	944	11.9	1	99.0
Collins	02HM005	Natural	155	10bis		16.4		100.0
Millhaven	02HM006	Regulated	150	10bis		15.9		100.0

The total watershed area (

Figure 1) is about 83 000 km^2, of which around 19 000 km^2 correspond to the lake surface. The U.S./CA. border follows the Niagara river (link between lake Erie and lake Ontario), the middle of lake Ontario, and the St.-Lawrence river. Major cities inside the catchment include Toronto (sub-catchment 14), Hamilton (sub-catchment 15), Rochester (sub-basin 3/4bis), Syracuse (Sub-basin 5), and Kingston (sub-basin 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).

Some aquifers do exist in the vicinity of Lake Ontario. The main ones are located between lake Simcoe (located west of sub-basin 12) and lake Ontario and consist in two captive aquifers separated by impervious layers and in an unconfined aquifer (the Oak Ridges Moraine), which are mainly located in sub-basins 14, 13, and 12. The deepest parts of the confined aquifers can be located down to 150m below surface. However, these aquifers do not seem to either contribute to river baseflows in a significant manner (about 3.5 m^3/s is generally released from the Oak Ridges Moraine, see Howard *et al.* 1996) or to import water from neighboring watersheds. Many other small aquifers do exist in the region (see Singer *et al.* 2003 for a comprehensive review), but again, they do not represent a significant part of the general water balance of Lake Ontario watershed.

A few derivations take place inside the watershed in order to fill the Welland and New-York State Barge (NYSB) canals. The Welland canal (located in sub-basin 1) diverts water from lake Erie into lake Ontario, and involves flows of the order of 250 $m^3.s^{-1}$. The NYSB canal diverts about 30 $m^3.s^{-1}$ in summer from the Niagara river into the Genessee river (sub-basin 3), and a

little less than 30 $m^3.s^{-1}$ is taken from the Genessee river to fill the NYSB canal between sub-basins 3 and 5, so this diversion is not envisioned to significantly affect the Genessee river overall balance. Finally, the amount of water derived from the Genessee river into the Oswego river through the NYSB canal is not taken into account by the Oswego streamflow gauge, which does not take the canal flow into account.

In summary, no canal, diversion, or aquifer is supposed to have any significant effect on the lake Ontario tributaries studied here, which is really helpful, but the flows involved in the diversions would still have to be taken into account when computing the overall lake Ontario's Net Basin Supply (NBS), which consists of runoff brought by the watershed land area, plus over-lake precip., and minus over-lake evaporation (see DeMarchi *et al.* 2009).

5.2) Calibration schemes

Two different calibration schemes are used for this project, which either consist in implementing the models over the gauged part of a sub-watershed (calibration scheme 2), or over the whole sub-watershed (calibration scheme 1), in which case the model is calibrated using an estimation of the whole catchment's runoff derived from the gauged part observed flow and a simple Area-based Ratio Method (ARM). At this point it is reminded that GRIP-O's ultimate aim is to develop efficient modeling tools able to simulate the runoff from the entire lake's watershed, including its ungauged areas. Calibration scheme 1 can be seen as a first step in this direction, because it allows simulating runoff for the ungauged fractions of the gauged sub-watersheds (see

Figure 1), but does not yet address the issue of the ungauged catchments. The ARM has proven reliable to estimate 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 sub-basin at once, thus taking into account rainfall over the ungauged part of the catchment. Other possibilities to estimate runoff from a complete sub-basin can be envisioned, but were not as attracting as calibration scheme 1. For example, one could simply 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 idea, but it requires two implementations of the same model for each sub-catchment. To exactly know which is the best practice among these three possibilities is beyond the scope of this work, but as mentioned, the preferred two are

used in GRIP-O. It was first envisioned to only use calibration scheme 1 for all gauged sub-basins for consistency with the former GRIP-M project (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: calibration scheme 2). As a consequence, an arbitrary subdivision was made between the sub-basins: when one contains several gauges, calibration scheme 1 is applied (5 cases); when it contains a unique most-downstream gauge (Figure 1, 8 cases), the second scheme is used. The performances will be carefully examined to investigate whether scheme 2 tends to systematically lead to better performances than scheme 1. If yes, then the relevance of scheme 1 could be questioned.

5.3) Calibration details

Due to limitations in the available RDPS outputs, the GRIP-O period extends from first of June 2004 until the 26th of September 2011. Because calibrating a hydrologic model over a set of 4 to 5 years is generally enough to achieve reasonable model robustness (Gaborit *et al.*, 2015), the calibration period was chosen to expand from first of June 2007 to the end (4.5 years). The validation period extends from first of June 2005 to first of June 2007 (2 years), with the first year of data (1st of June 2004 to 1st of June 2005) being used as the spin-up period.

The objective function used in calibration consists in the Nash-Sutcliffe criterion (see Nash and Sutcliffe 1970) but computed taking the square-root of the observed and simulated flow time series. This score will be further referred to as "Nash sqrt" in this report. Other quality criteria consist in the common Nash-Sutcliffe criteria ("Nash"), the Nash criteria calculated over the log (base 10) of the flows ("Nash Ln"), and a bias criteria (see PBIAS in Fry *et al.* 2013).

$$PBIAS = \frac{\sum_{i=1,n}(Qobs_i - Qsim_i)}{\sum_{i=1,n}(Qobs_i)} \qquad \text{Equation (1)}$$

The LSS of SPS, namely SVS, was at first not calibrated during GRIP-O. Only the routing scheme WATROUTE was. This is because on one hand, it was expected that with recent and detailed LULC and Soil datasets, SPS would perform good quality simulations, and on the other because of computational limitations: it takes a long time to run SPS over the 4.5-year GRIP-O calibration period and for the whole lake Ontario watershed domain. The following four WATROUTE parameters were hence first calibrated (section 8.1, 8.2 and Table **12**): the main channel roughness coefficient, the flood plain manning coefficient, and two parameters controlling the rate of discharge from the lower zone storage (representing the aquifer/deep soil storage) to the river (see Table 8).

A standalone version of SVS was however developed and used in conjunction with a simple UH methodology in order to save tremendous amounts of computational time while still estimating streamflows, in order to calibrate SVS parameters based on streamflow

performances (see further down). This stand-alone version is referred to as "GEM-Surf + UH" throughout this document, in order to differentiate with the GEM-Hydro platform which includes the routing scheme WATROUTE.

GR4J in conjunction with the "cemaneige" snow module together have a total of 6 free parameters to calibrate, and LBRM has 9 free parameters which are not detailed here. The reader is referred to the models' publications to know more about the lumped models' parameters (section 2).

The calibration algorithm consists in the Dynamically Dimensioned Search (DDS) algorithm (Tolson and Shoemaker 2007) for all models, and in the excel solver module (a gradient-based algorithm) for the first implementation of GR4J + the "Cemaneige" snow module (see section 7.1.1). GR4J has also been calibrated with the DDS algorithm. The comparison between using the DDS or the excel solver to calibrate the model is performed in section 7.1.2. Because the WATROUTE models using the SPS outputs take a lot of time to run, only 200 runs were made to calibrate them, while GR4J, LBRM, WATFLOOD, and SA-MESH will be calibrated using a maximum of model runs comprised between 1000 and 2000. 400 runs were used for GEM-surf + U.H. The maximum number of runs allowed during calibration varied depending on the model (or version in the case of GEM-Hydro) being used, but it was always assessed that the algorithm successfully converged to an optimal solution. See Table **12** for the maximum number of runs used in each calibration experiment.

The DDS algorithm is very efficient in the sense that it converges rapidly to optimal solutions for the parameter set (Tolson and Shoemaker 2007). Moreover, for GR4J and LBRM and when using the DDS algorithm, three calibration trials, each one with different initial parameter values, were used. Two were chosen randomly for GR4J while one corresponds to GR4J default parameter values found in literature (see Table 7 below). For LBRM, all three calibration trials were performed with random initial values. The parameter set leading to the best performances was finally kept. But for the distributed models, only one calibration trial was performed. When calibrating WATROUTE, the initial parameter values were obtained after manual tuning over the lake Ontario watershed. It takes a few minutes for GR4J to perform a thousand of runs, but two days for WATROUTE over most of the GRIP-O sub-watersheds. Calibrating GEM-Surf + UH in order to adjust SVS parameters generally required between 4 and 15 days (in the case of global calibration, see section 9). In the light of the results obtained, it is believed that 200 runs is enough for achieving the convergence to an optimal parameter set with WATROUTE, because only 4 free parameters were considered in this study. Following the same way of thinking, using a maximum of 400 runs with GEM-Surf + UH instead of 1000 with SA-MESH does make sense, because the former was calibrated using 16 free parameters while the latter had 60 free parameters (see Tables below).

Table 7 presents the GR4J and LBRM parameter ranges used in calibration, while Table 8 presents the Watroute parameter values used in calibration when only calibrating the routing part of GEM-Hydro (SPS). See Table **12** for a synthesis of the GEM-Hydro calibration experiments and associated SVS model versions.

Table 7: Parameter ranges used in calibration; one of the three GR4J trials was performed with GR4J default initial parameter values indicated in Table 8; other trials were made with a random initial parameter set; coeff: coefficient; UH: Unit Hydrograph; USZ: Upper Soil Zone

LBRM free parameters (10)	unit	low bound	upper bound	GR4J free parameters (6)	unit	initial	low bound	upper bound
Tbase	deg C.	1.00E-04	10	x1: Capacity of production	mm	191.5	10	2500
Snowmelt factor	cm/deg C./day	1.00E-04	5	x2: Water exchange coeff	mm	2.33	-15	10
linear reservoir coeff: percolation	day^{-1}	1.00E-04	100	x3: Capacity of routing store	mm	80.3	10	700
partial linear reservoir coeff: USZ	m^{-3}	1.00E-09	1.00E-06	x4: UH time base	day	3.1	0	7
linear reservoir coeff: interflow	day^{-1}	1.00E-04	10	x5: degree-day factor	-	3.7	1	30
linear reservoir coeff: deep percolation	day^{-1}	1.00E-05	10	x6: snowpack inertia factor	-	0.25	0	1
partial linear reservoir coeff: LSZ	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 8: WATROUTE parameter (param.) ranges used in calibration, along with the initial (init.) values obtained after manual tuning. coef.: coefficient.

param. Name	MAN2	MAN1	POW	FLZCOEF
description	main channel Manning roughness coef.	floodplain Manning roughness coef.	exponent in the baseflow equation	linear coef. In the baseflow equation
init. Value	0.15	0.25	2.2	1.00E-06
lower bound	0.01	0.04	0.4	1.00E-07
upper bound	2.00	8.00	5.0	1.00E-03

The distributed models could not be calibrated locally for all GRIP-O sub-catchments of **Figure 1** because they require a high computation time. Therefore, in order to compare the optimal performances of the distributed models under study, each of them was calibrated using the strategy described here but only for some of the GRIP-O tributaries, such as the Moira (catchment 11), the Black (7), Genessee (3), and Salmon (sub-basin 10) rivers. These tributaries were selected because results with the lumped models (see section 7) indicated that good performances could be reached for them (despite regulation effects), and because calibrating the models on catchments containing a unique gauge station is straightforward and does not involve any observed streamflow approximation, in opposition to calibration scheme 1 which was used for catchments with several streamflow gauges.

Table 9 to **Table 11** present the list of parameters used in calibration for the three distributed models considered, when using GEM-Surf + UH (the surface scheme of GEM-Hydro but associated to a UH instead of a routing model to save computation time, see above) instead of GEM-Hydro in order to calibrate SVS parameters. Different paradigms were used to calibrate distributed models: for example, GEM-Surf + UH was calibrated mainly using multiplicative coefficients that were then used to adjust each of the spatially-varying values of a given

parameter, leading to a reasonable number of free parameters (16) while preserving the spatial variability of parameter values. On the opposite, SA-MESH was implemented by calibrating its 5 different Grouped Response Units (GRU) considered (and defined according to soil texture and land use) in an independent manner, thus resulting in a higher number of free parameters (60). Watflood had the lowest number of free parameters during calibration, and involved calibrating parameter values which are valid for the entire catchment (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 9: list of parameters and associated ranges used to calibrate GEM-Surf + UH; LZS: Lower Zone Storage; coeff. : coefficient; mult. : multiplicative; precip. : precipitation. A total of 16 parameters were used.

parameter	description	initial	min.	max.	parameter	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 10: list of parameters and associated ranges used to calibrate Watflood; LZS: Lower Zone Storage; coeff. : coefficient; mult. : multiplicative. A total of 14 parameters were used.

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 11: list of parameters and associated ranges used to calibrate SA-MESH. A total of 60 parameters were used, as each parameter listed below was calibrated independently for each

of the five vegetation (crops, grass, and three forest types) or river classes considered in the model.

parameter	description	vegetation or river class (5)	minimum	maximum
ROOT	Annual maximum rooting depth of vegetation category [m]	crops and grass	0.2	1.0
		forests	1.0	3.5
RSMN	Minimum stomatal resistance of vegetation category [s m-1]	crops and grass	60.0	110.0
		grass	75.0	125.0
		forests	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/km2]	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

The initial parameter values were either set to default values which generally provide satisfactory results for the model (GEM-surf + UH, Table 9), or to random values (Watflood, SA-MESH). A maximum of 1000 model runs was used to calibrate SA-MESH, and of 1500 for Watflood, while the maximum was set to 400 for GEM-surf + UH (see Table **12**). This is because on one hand, GEM-surf + UH is more demanding than Watflood and SA-MESH in terms of computational time, and on the other because GEM-surf + UH is here calibrated with a reasonable number of free parameters (16, see Table 9). According to the tests performed in this study (see section 8), GEM-Surf + UH was always able to reach an optimal solution in less than 400 runs. The time required to perform 400 runs with GEM-Surf + UH depends on the catchment size, but was of the order of 5-6 days for the biggest GRIP-O sub-basins. Finally, the calibration strategy used here for SA-MESH consists in the best one (for this model and in terms of resulting performances) according to the work of Haghnegahdar *et al.* (submitted). Despite the random initial values used for SA-MESH and Watflood, only one calibration trial was performed for each of these models on a given catchment, because according to experiments performed with the lumped models (section 7), the optimal parameter set varies only slightly from one calibration trial to the other.

Despite the three distributed models studied here were not calibrated using the same number of free parameters or with the same maximum of model runs allowed, it is assumed that the calibration strategies employed allow to make each model come very close to its optimal performances for a given catchment and the time period considered. Indeed, the strategy used for each of the three models is the result of expert knowledge with each of them

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 choice to achieve a fair comparison between models consists, in our view, in using the same calibration algorithm, objective function, and of course the same physiographic and forcing data.

Table 12 below synthesize the information in regard of the maximum number of runs used with each calibration experiment, the section of this report containing the results of the associated experiment, and the version of the SVS model which was used with each of them. See Table **2** for a detailed description of the main SVS versions.

Table 12: Main calibration experiments associated with each SVS version and sections where associated results are described.

version name	former	intermediate	new	new_2
precipitation source used	Model P.	Model P.	CaPA P.	CaPA P.
calibration made?	yes	no	yes	yes
calibration on (Watroute / SVS), number of free parameters	Watroute (4)	N/A	Watroute (4)	SVS (16)
max. number of runs	200	N/A	200	400
name of platform used	GEM-Hydro	GEM-Hydro	GEM-Hydro	GEM-Surf + UH
Manning coefficients (see section 6.2.2)	fixed	fixed	fixed	variable
section where used	8.1	8.1.3	8.2	8.3 to 8.7, 9

6- SPS+SVS and WATROUTE sensitivity tests

This section mainly aims at describing some sensitivity tests which were performed with SPS+SVS (i.e. GEM-Hydro with the SVS LSS). These tests focused on the model spatial or temporal resolution, or other configuration settings. They allow to justify the choice which was finally made to implement the GEM-Hydro (or GEM-Surf + UH) model in section 8.

6.1) SVS sensitivity

SPS was firstly implemented for GRIP-O with a spatial resolution of 120 arc-seconds (about 4 km, see justification below), using the SVS land surface scheme and river routing scheme WATROUTE (see section 2) at a 30 arcsec. resolution (about 1 km, see text below). This consists in relatively high (detailed) resolutions for these models, which tended, during former experiments, to be implemented at lower (coarser) resolutions over larger areas (such as the entire Great Lakes domain), hence calling for coarser grids (such as 10 arcmin. or 20 km for SPS) mainly for memory and computing times purposes. During this first implementation, SPS+SVS+WATROUTE did use model (regional deterministic REGETA) precipitation and SVS "former" version (see Table 2 and Table **12**).

The SPS model here uses the recently released harmonized global soil dataset (30 arcsec. resolution) of Shangguan *et. al* (2014, see section 4). However, only 7 layers are used in SPS because the layers 2 and 3 were merged together. The spatial resolution of SPS was firstly chosen to be 2 arcmin. (or 120 arcsec. or 4 km) because this resolution allows to preserve the high level of spatial detail contained in this soil dataset while minimizing the computational time and disk space requirements, and because the resulting hydrographs were judged insensitive to SPS spatial resolution between 60 and 120 arcseconds (not shown here). Using a coarser resolution allows to save tremendous amounts of computational time and disk space requirements (mainly due to output files). However, a test was later made with implementing SPS+SVS at a 10 arcmin. spatial resolution (using the exact same sources of physiographic data, which were in this case aggregated to this 10 arcmin. / 20km resolution), and it turned out that the model resolution (either 2 or 10 arcmin.) had a minor influence on the resulting streamflow hydrographs, at least when using model precip. (see Figure 2 below).

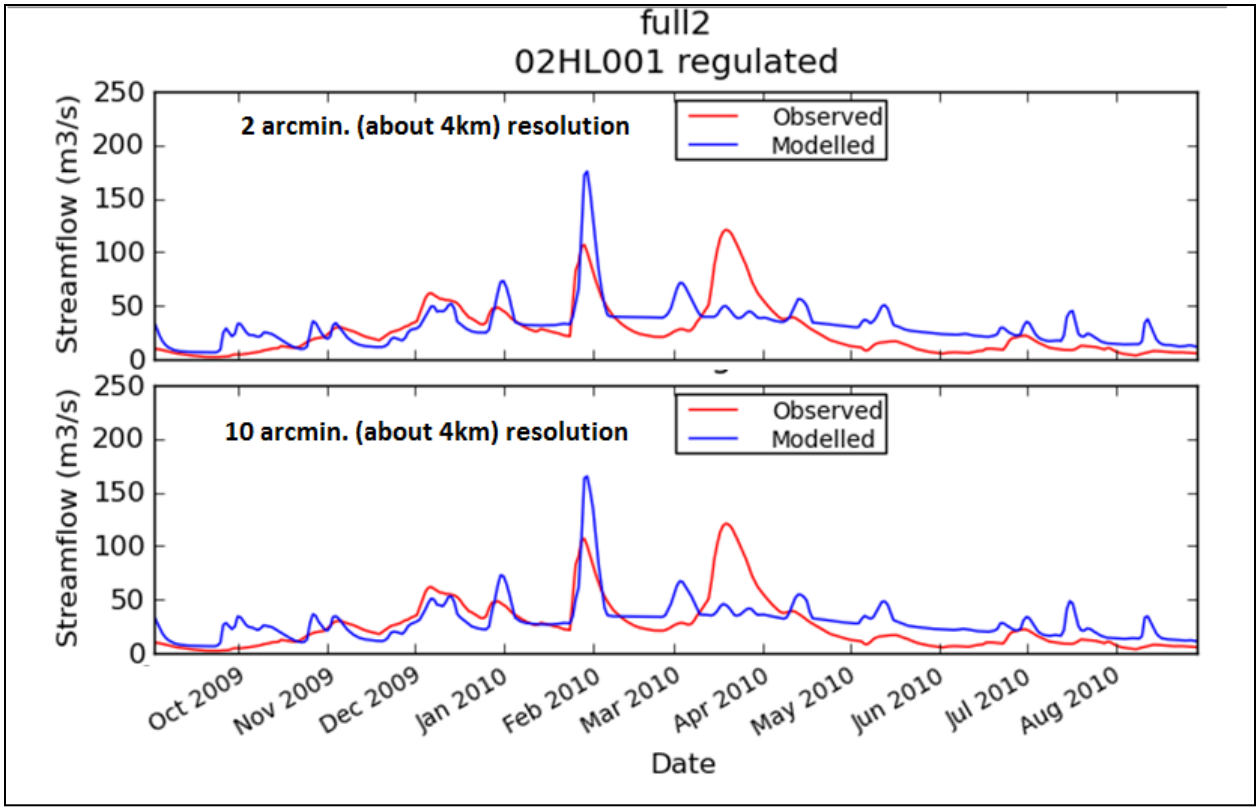

Figure 2: SPS+SVS sensitivity to the model spatial resolution (2 or 10 arcmin., or about 4 versus 20 km respectively); Hydrographs for the Moira river watershed.

The internal time step was also subject to a sensitivity test, and it turned out that the model was very sensitive to this internal configuration parameter (see Figure 3 below). Tests were made using internal time steps of 5, 10 and 30 min. No significant differences were seen in the resulting hydrographs between the 5 and 10 min. time steps, but major differences arose between the 10 and 30 min. time steps. Using a higher time step resolution (i.e. a lower time step interval) is best because it allows more precision when solving the differential equations, and when time series are linearly interpolated between two time steps. But again, as it has an influence on the computational time, the time step of 10 min. was preferred to a 5 min. one, given that they lead to similar results (not shown here).

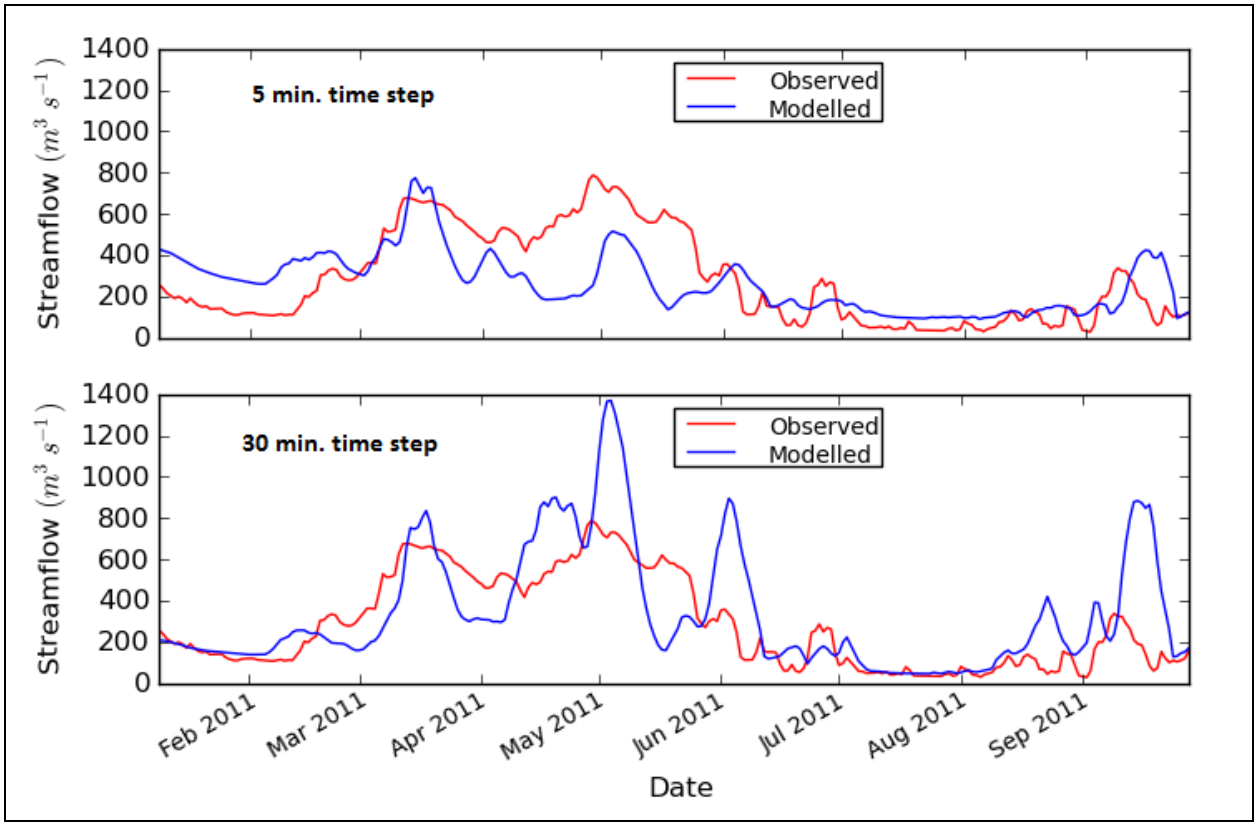

Figure 3: sensitivity of the Oswego sub-basin streamflow to the SPS internal time step.

6.2) WATROUTE sensitivity

6.2.1- Spatial and temporal resolution sensitivity

The WATROUTE river scheme was implemented at 30 arcseconds during all of the GRIP-O experiments performed with SPS. A test was made between implementing the model at 15 or 30 arcseconds. Some influence of the resolution was noticed on the hydrographs (not shown here), yet this consisted in small differences which could be neglected when considering the gain in terms of computational speed that running the model at 30 arcsec. actually represents, especially when considering that the model was to be calibrated in the context of GRIPO. Any modeler should however keep in mind that this could sometimes be needed to run the model at a higher (for example 15 arcsec.) resolution if someone is interested in simulating small or very small watersheds (i.e. less than about 300 km^2), because at 30 arcsecond resolution, one grid cell can actually contain several small rivers. This was not estimated to occur in GRIP-O however, as all of the selected flow stations consist in most-downstream gauges, close to the lake shoreline, and hence in stations on major rivers whose location can be well captured by the 30 arcsec. resolution. Given the size of some GRIP-O watersheds, it was not envisioned to

implement the model at a coarser resolution than 30 arcseconds (1 km). As the WATROUTE model makes use of a dynamic time step option, it was not needed to test the sensitivity to this parameter, the model continuously adjusting its time step in order to run as fast as possible, but decreasing it as soon as internal instabilities are detected. The water balance achieved with the model is generally almost perfect, meaning that the model only creates or loses insignificant amounts of water during the simulations. For GRIPO, the maximum time-step was set to 1200s (half an hour) and the minimum one to 10s.

6.2.2- Spatially-varying shed file parameters

The current version of WATROUTE makes use of constant parameter values over all pixels of a given watershed, for the parameters which are defined by the user during the shed file creation, namely the Manning coefficients (main channel and overland flow), the recession curve multiplicative coefficient and exponent (defining the rate at which water is released from the WATROUTE lower zone storage and represents base flow), the meandering factor, the number of channels per grid cell, and the three coefficients defining the main channel capacity as a function of the pixel drainage area. However, this is clear that in reality, these coefficients do vary in space. Moreover, with the default parameter values set to 1 for the number of channels and meandering factor, the simulated river length inside a pixel is always underestimated, and sometimes in a strong manner. This will result, when calibrating the model, in unrealistically high Manning values in order to compensate for the too small river lengths.

In order to circumvent this effect and try to get more realistic parameter values, a WATROUTE version with spatially-varying values of the meandering factor and Manning coefficients was implemented. These varying parameters are in this case derived from physiographic information such as drainage area, land-cover type, and channel slope. This original WATROUTE version was tested, and revealed promising (in the sense that it allows achieving streamflow simulations of the same quality than those resulting from using constant values, but with more realistic parameter values). See Table 40 (section 8.5) for the default range of values used for the spatially-varying Manning coefficients. During the first experiments (sections 8.1 to 8.4) of the GRIP-O project however (see Table 12), constant parameter values were used for the Manning coefficients, but varying values were used for the meandering factor. The last experiments involving the calibration of SVS parameters were made with spatially-varying values of Manning coefficients (sections 8.5 to 8.7, and 9) and of the meandering factor, which was however never calibrated.

It is possible to calibrate the parameter values even when spatially-varying fields are used, for example by tuning a constant value by which multiplying (or to which adding) the

variable values, which was done when calibrating SVS parameters (see Table 12 and section 8.6). Using spatially-varying values for Manning coefficients is not judged mandatory as it turns out that at least over small to medium watersheds (as is the case for GRIP-O), it is possible to achieve very satisfying simulations using constant Manning values over the whole sub-basin, given that the SPS outputs are of good quality: WATROUTE is only a routing scheme which, by tuning the Manning and recession curve coefficients, should be able to adjust the streamflow hydrographs to closely match the observed ones. Note that this is however true only if the watershed delineation is correct and the basin does not include too many wetlands, artificial reservoirs or other structures which are currently not simulated by the model and affect the streamflow pattern. Watroute can, at the moment, only simulate lakes or reservoirs for which the precise relationship between the volume and released streamflow is known, but this is rarely the case. In the case of natural lakes however, which behave almost like reservoirs controlled by a simple fixed weir at the outlet, their effect on streamflow can be accounted for by WATROUTE by increasing the manning coefficients. A small-lake model is however under development at the U. of Saskatchewan to be incorporated into WATROUTE in the future.

7- GRIP-O experiments with the lumped models

7.1) GR4J calibrations

7.1.1- calibration with the excel solver

As this lumped model's simulation quality has been proven very satisfactory over a very large set of worldwide catchments during past studies (Pagano *et al.* 2010), and because it is fast and simple to implement, it was included in the GRIP-O project. The model has a total of 6 free parameters because it was implemented here in conjunction with its "Cemaneige" snow module which has 2 free parameters. The model only needs the average (daily) catchment P., minimum, mean and maximum T., and mean elevation, latitude, and drainage area. The last three variables were computed using the ArcMap software when performing the GRIP-O sub-basins' delineations with the HydroSHEDS flow direction grids. The T. data were taken from the 15 arcsec. grid derived from GHCND data (see section 3), and the P. inputs either came from GHCND or from the 24-hourly CAPA product (interpolated on the same 15 arcsec. grid to compute the watershed averages). See section 5 for other details regarding the calibration methodology. See section 5.3 for the initial GR4J parameter values used for all sub-basins.

The resulting performances are very satisfying, especially given that most of the GRIP-O sub-basins have an influenced flow regime which is subject to anthropogenic regulations. Table 13 below synthesizes the results obtained with the GHCND precipitation. No correlation could be made between the model performances and the basin's size, country, elevation, flow regime type, or gauged fraction, the calibration scheme or final parameter set (not shown in Table 13).

The models generally display good performances in validation too, with performances not systematically better for calibration over validation.

Figure 4, 5 and 6 below respectively display the hydrographs (in calibration) of the Salmon, Oswego and sub-basin 13 sub-basins (ID 10, 5 and 13). The Salmon and sub-basin 13 respectively consist in the best and worst sub-basins, according to the Nash values of Table 13.

Table 13: GR4J performances over the different GRIP-O sub-basins when using GHCND P. and T. With calibration scheme 2, the model was implemented over the gauge watershed (100% gauged), while in calibration scheme 1 it was implemented over the whole sub-basin, which is only partly gauged (see section 5.2). CAL, VAL: calibration and validation period; Nash, Nash sqrt: Nash criterion in percent (traditional or with the square-root of the flows, respectively). Target is 100.

							CAL		VAL	
country	Subbasin	Cal.scheme	Station	%_gauged	Area(km2)	Flow regime	Nash	Nash sqrt	Nash	Nash sq
CA	1	2	20_mile	N/A	307	natural	75.6	79.6	81.4	87.3
USA	3	2	Genessee	N/A	6317	regulated	80.9	83.1	78.0	81.8
USA	4bis	2	Irondequoit	N/A	326	natural	66.7	71.2	52.1	63.5
USA	5	2	Oswego	N/A	13287	regulated	83.5	82.9	70.3	73.0
USA	6	1	N/A	40	2406	mixed	70.2	75.5	72.4	77.0
USA	7	2	Black river	N/A	4847	regulated	70.1	79.5	75.5	81.0
USA	8	2	Oswegatchie	N/A	2543	regulated	75.5	80.8	81.7	83.9
CA	10	2	Salmon_CA	N/A	912	regulated	85.9	88.1	81.8	85.9
CA	10bis	1	N/A	44.2	944	mixed	81.4	88.4	76.5	80.7
CA	11	2	Moira	N/A	2582	regulated	87.8	89.2	80.4	81.2
CA	12	1	N/A	88	12515.5	regulated	72.1	72.1	63.5	68.3
CA	13	1	N/A	40.3	1537.5	natural	59.9	63.9	62.0	68.9
CA	14	1	N/A	61.3	2689.4	mixed	78.9	78.1	70.6	77.0
CA	15	1	N/A	63	2245.8	mixed	86.2	83.5	85.1	86.8

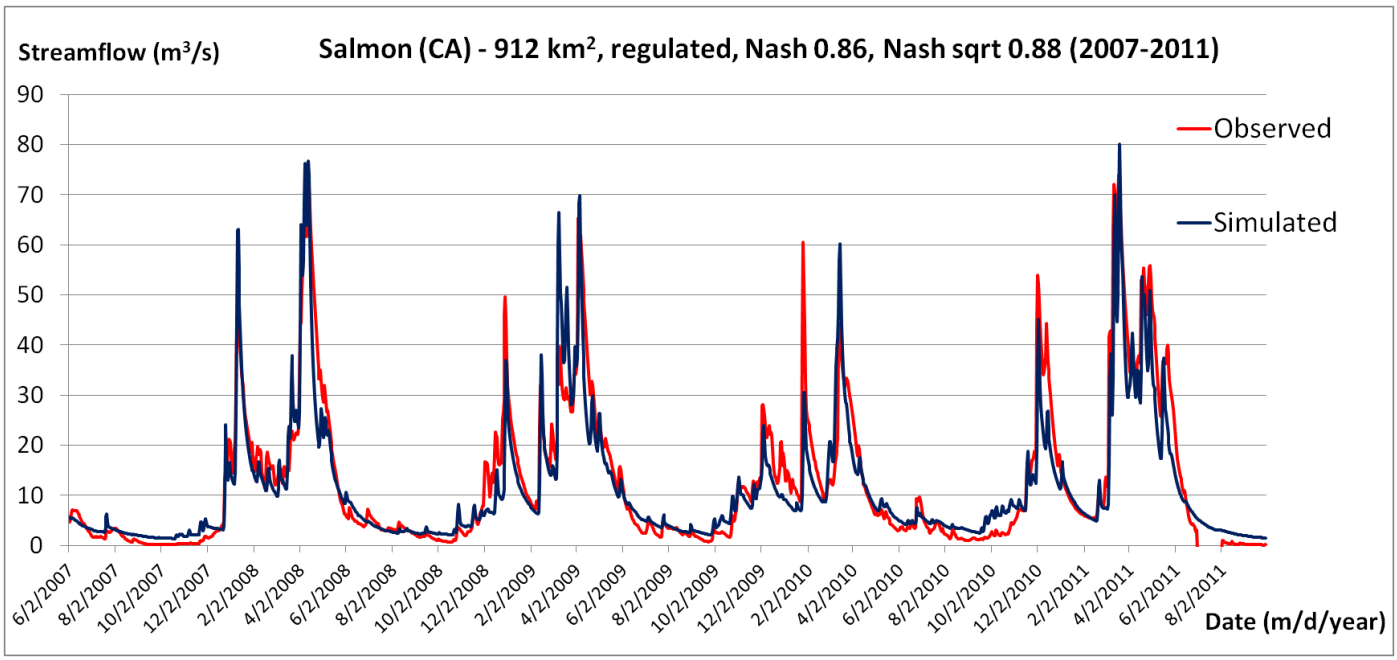

Figure 4: Salmon river hydrographs (sub-basin 10) in calibration obtained with GR4J and GHCND forcings.

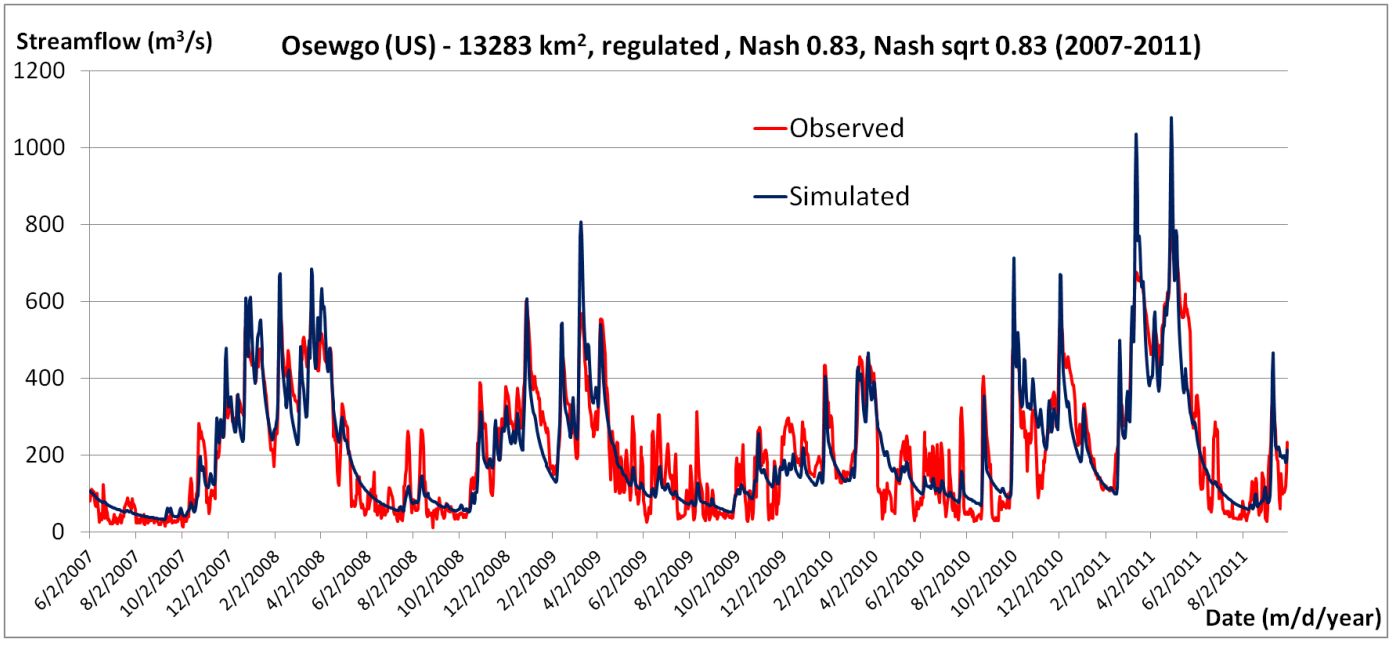

Figure 5: Oswego river hydrographs (sub-basin 5) in calibration obtained with GR4J and GHCND forcings.

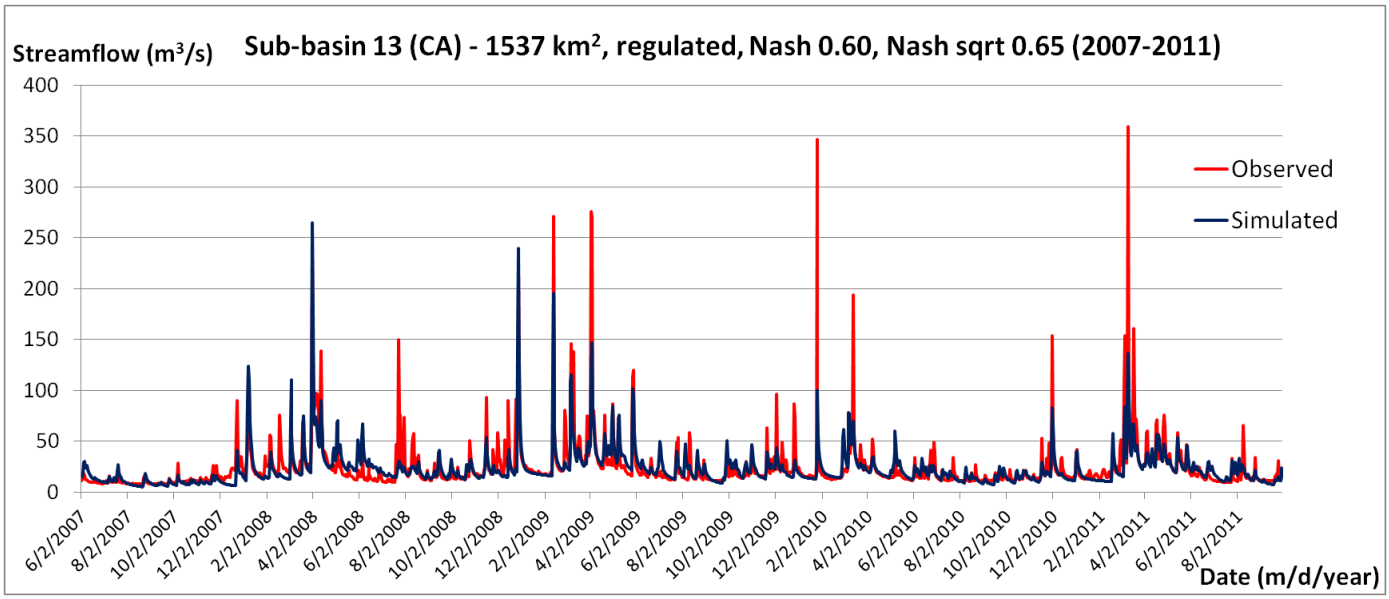

Figure 6: Sub-basin 13 hydrographs in calibration obtained with GR4J and GHCND forcings.

Note that the performances' difference between sub-basins 10bis and 13 (Table 13) is a good example of the fact that the model performances do not seem to be related to the flow regime, calibration scheme or basin size: despite being larger, with a natural flow regime and a similar gauged fraction, the sub-basin 13 performances are not as good as the sub-basin 10bis ones, while on a first guess it would have been logical to expect the opposite.

No information at all was found about the way regulation was performed (see section 5.1). Yet when looking at the Salmon and Oswego hydrographs for example (see Figure 4 and Figure 5), it seems that for the first regulation mainly consists in storing water during summer and releasing it during winter (as if the dams' regulation consisted in a simple smooth weir behavior), what the GR4J model seems to correctly simulate, and that for the second regulation involves more artificial flow releases all over the year (harder to simulate for a hydrologic model). The sub-basin 12 regulation (based on the observed versus simulated flow patterns, not shown here) seems to be similar to the Oswego regulation, whereas the Genessee basin displays observed flow patterns somehow falling in-between the two regulation types (not shown).

Another test consisted in showing that the results obtained in Table 13 were not very sensitive to the calibration scheme considered. For that, two GRIP-O sub-basins initially calibrated using the first calibration scheme (implementation over the whole sub-basin using extrapolated observed flows, see section 5.2), were then used to implement GR4J according to the second calibration scheme (implementation over the gauges' watersheds only, using the sum of the flows observed at the gauges' locations). The conclusion, as depicted in Table 14

below, is that extrapolating the observed flows to the whole sub-basin area using the ARM (see section 5.2) is not responsible for a significant loss of performance (except for sub-basin 6 in validation) compared to when calibrating at the gauge sites directly, even when the total gauged sub-basin fraction is of order of 50 %, as is the case for the two sub-basins considered in Table 14 below. This result hence comforts the choice of calibration scheme 1 for some GRIP-O sub-basins, because it allows implementing a model only once and over the whole sub-basin to estimate its total runoff contribution to lake Ontario.

Table 14: comparison between the two calibration schemes for two GRIP-O sub-basins; the term "gloc" in Subbasin column refers to the fact that the model was in this case implemented over the gauges' watershed instead of the total sub-basin.

						CAL		VAL	
country	Subbasin	Cal. scheme	% gauged	Area (km^2)	Flow regime	Nash	Nash sqrt	Nash	Nash sqrt
USA	6	1	40	2406	mixed	70.2	75.5	72.4	77.0
USA	6 gloc	2	100	970	mixed	66.9	74.4	76.3	80.6
CA	14	1	61.3	2689.4	mixed	78.9	78.1	70.6	77.0
CA	14 gloc	2	100	1659	mixed	79.4	78.2	70.7	76.7

After calibrating GR4J with the GHCND P. and T. forcings, the same set of experiments were done again but using the 24-h CaPA Precipitation instead (but still using GHCND T.). Table 15 below presents the difference between the final results obtained when using either CaPA or GHCND precipitation. The results indicate that the performances (both in calibration and validation) are almost always better when using GHCND precipitation, especially over the U.S. catchments. The GHCND observation network is more dense than the ground network used to bring corrections to the CaPA analysis, because the CaPA product only uses data from ground stations available in real-time. As the GHCND network is dense enough to capture most local precipitation events even in summer, the result is that the GHCND precipitation is closer to reality than the CaPA precipitation, over this watershed. Furthermore, as the density of the GHCND observation network is very dense on the U.S. side of the lake partly because it includes data from a cooperative network, the differences between using GHCND or CaPA precipitation are stronger over the U.S. catchments (see Table 15 below).

Table 15: difference in the performances when using GLERL or CaPA precipitation to calibrate GR4J, i.e: GLERL Nash - CaPA Nash using Nash values in percent. Nash sqrt, Nash : nash values either computed using the square-root of the flows, or the flow values themselves. Calibr, Valid: Calibration and validation periods, respectively.

		Difference in Nash percentage value: GLERL value - CaPA value			
		NASH SQRT		NASH	
Subbasin	Country	CALIBR	VALID	CALIBR	VALID
1	CA	2.07	10.19	2.91	7.92
3	US	12.65	16.85	13.67	18.88
4bis	US	14.16	4.84	19.47	-0.18
5	US	8.46	6.06	12.38	8.13
6	US	13.69	11.15	17.54	11.81
7	US	16.11	15.29	11.73	21.87
8	US	18.02	14.06	18.33	21.67
10	CA	4.08	2.45	6.48	-0.34
10bis	CA	6.66	-4.96	16.96	-7.81
11	CA	0.95	1.10	-0.36	9.89
12	CA	6.31	-4.10	4.71	-3.88
13	CA	-1.06	0.97	-1.60	1.93
14	CA	3.31	14.31	6.60	18.72
15	CA	0.97	8.06	5.54	16.94
	mean	7.6	6.9	9.6	9.0
	minimum	-1.1	-5.0	-1.6	-7.8
	maximum	18.0	16.9	19.5	21.9
	mean US	13.85	11.38	15.52	13.70
	mean CA	2.91	3.50	5.15	5.42

Because it was noticed that using the Excel solver and with CaPA precipitation, GR4J did not always converge to the optimal parameter set, tests were performed to calibrate GR4J with the DDS algorithm (and a maximum number of runs of one thousand), to see if it allows generally achieving better simulations than when using the Excel solver to calibrate the model (see next section).

7.1.2- Calibration with the DDS algorithm

See section 5. for calibration details. Table 16 below presents the comparison of the performances obtained when calibrating GR4J with the DDS algorithm (three trials, see above) or the Excel solver (see 7.1.1).

Table 16: difference between the performances obtained when calibrating GR4J with the DDS algorithm and the Excel solver DDS - Excel); only values for the calibration (CAL) period are shown; GR4J was implemented in both cases using the GHCND precipitation data. The difference values shown here were derived from Nash values expressed in percentages (i.e., values between 0 and 100).

				CAL: DDS - Excel	
country	Subbasin #	Cal. scheme	Station	Nash	Nash sqrt
CA	1	2	20_mile	1.4	0.2
USA	3	2	Genessee	0.1	0.0
USA	4bis	2	Irondequoit	0.1	0.2
USA	5	2	Oswego	0.2	0.2
USA	6	1	N/A	1.2	0.7
USA	7	2	Black river	9.0	2.4
USA	8	2	Oswegatchie	3.0	1.3
CA	10	2	Salmon_CA	0.4	0.2
CA	10bis	1	N/A	0.8	0.3
CA	11	2	Moira	0.4	0.4
CA	12	1	N/A	0.9	0.7
CA	13	1	N/A	2.3	2.5
CA	14	1	N/A	-0.1	0.2
CA	15	1	N/A	0.2	0.3

This can be seen from Table 16 that performances obtained with the Excel solver and only one calibration trial are generally very close to those obtained with the DDS algorithm and three different trials, in which case the parameter values are also very close in both cases (not shown here). Cases which display significant better values with the DDS algorithm (i.e. sub-basins 7, 8 and 13) are however not always associated to very different final parameter values (not shown here). It is clear that the methodology used here to calibrate GR4J with the DDS algorithm however systematically leads to better performances than with the Excel solver, but that the latter does a very good job in almost all cases.

When looking at parameter values finally obtained after calibration (see Table 7 for a list of parameters used in calibration), no clear difference could be identified between using the CaPA or GHCND precipitation dataset, for any of the two models. Parameter values were, on the contrary, generally close with the two different precipitation sources. The values are not shown here for all of the catchments, because not judged much informative, yet 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. Table 17 and Table 18 are devoted to such purpose. For GR4J, two cases were distinguished, depending on if the production store maximum capacity was way lower than the routing store one or not. Indeed, the first case scenario was considered as not consisting in realistic values for the model, whose production store capacity is supposed by its creators (see Perrin *et al.* 2003) to represent the soil depth. In some cases however (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), i.e. of the order of less than 30 mm for the former, which is not realistic given land use and soil depths observed for the associated catchments. As a consequence, the statistics for GR4J parameters in Table 17 were subdivided in two cases as explained above.

Table 17: Statistics for the GR4J 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

The fact that calibration of GR4J sometimes converged to an unrealistic (yet optimal) parameter set is attributed to regulation, because it was only observed in the case of (partly) regulated sub-catchments. As regulation in these watersheds is mainly due to (artificial) reservoirs (dynamically controlled or not), the set of a priori unrealistic parameter values (see bottom part of Table 17) obtained for these catchments could be seen as an effort made by the model to optimally represent the effect of regulation on simulated streamflows, by trying to represent a big routing reservoir (X3). Because this reservoir then captures 90% of runoff generated by the production store (X1), in the case where X3 is very high, the value of X1 may become less important and set to very small values in order to still be able to produce peak

flow events via the 10% of runoff that does not flow into the routing store but turns directly into streamflow (see Perrin *et al.* 2003 for the concepts of the GR4J model).

The GR4J results obtained with DDS are described in the 7.3 section, because most tables include both LBRM and GR4J results.

7.2) LBRM calibrations

The same set of calibrations performed with GR4J have been performed with LBRM, using the exact same data and calibration details as explained in section 5. This work was done by Bryan Tolson and his team, from the university of Waterloo.

Table 18: Statistics for the LBRM 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.

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
	deg C	cm/˚C./d	d-1	m-3	d-1	d-1	m-3	d-1	d-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	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	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

7.3) Results obtained with LBRM and GR4J (DDS algorithm).

7.3.1 GR4J performances with GHCND precipitation

Among the four different possibilities studied here (LBRM or GR4J model, CaPA or GHCND precipitation dataset), the best simulation performances were generally achieved with GR4J and the GHCND dataset (see section 7.3.2). Results for this best combination are presented in Table 19, along with the sub-basins' main characteristics.

A few conclusions emanate from Table 19. Simulation performances (after calibration) do not seem to be related to any of the sub-basins' main properties (Country affiliation, area, elevation, flow regime nature) or the calibration scheme used (calibration scheme, percent gauged). In other words, very satisfying performances can be obtained even in the worst case scenario, i.e. a (partly) influenced flow regime, calibration scheme 1 (synthetic observed flows), and a relatively low gauged fraction (see for example sub-basin 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 (sub-basins 4bis and 6). This is also true for the three other experiments (LBRM / GR4J and GHCND/ CaPA precipitation).

This leads to two hypotheses in regard of regulation, at this point: either the degree of control is low in the lake Ontario area, letting the rivers with a behavior similar to a natural one, as if regulation simply consists in reservoirs with a weir at their outlet (static control, quite easy for the models to handle with calibration), or the lumped GR4J model is remarkably efficient in reproducing the effect of regulation on river flows, due to its "time-series fitting" ability (parameter X2, for example, can create or remove water in the model). In both cases, it is quite clear that this model is well adapted to hydrologic modeling in the lake Ontario watershed, and that obtaining satisfying simulation performances for regulated watersheds is promising both for the upcoming implementation of distributed models (see companion paper) and the estimation of the whole lake's runoff supply (section 7.3.4). Indeed, the regulated most-downstream flow gauges (see

Figure 1), altogether with the gauges on "natural" streams, cover about 74% of the total lake Ontario watershed land surface. If only natural flows could be correctly simulated in the watershed, then only about 20% of its total runoff could be simulated with confidence, preventing a reliable estimation of the whole lake's runoff (see Fry *et al.* 2014 for the reliability of a catchment's runoff estimation as a function of its gauged fraction).

The fact that no calibration scheme is always better than the other (also seen with the three other experiments) also supports the methodology used here to simulate runoff from the ungauged part of a gauged sub-catchment (calibration scheme 1, see section 5.2), because it means that using synthetic flows (derived from observed ones and the simple ARM process) to calibrate a model leads to simulation performances similar to when using real observed flows (calibration scheme 2), even in the case where only 40-60% of the sub-basin is gauged (see sub-catchments 6, 10bis, 13, 14 and 15 in Table 3).

Table 19: All evaluation metrics obtained with GR4J and the GHCND precipitation dataset. All values are in percent; optimal value is 100 except for PBIAS (0). Nash sqrt, Nash Ln: Nash values computed with either the square-root or log (base 10) of the flows, respectively. See section 2 for more details.

								CAL				VAL			
country	Subbasin #	Cal. scheme	Station	%_gauged	Area(km^2)	Flow regime	mean elev. (m)	Nash	Nash sqrt	Nash Ln	PBIAS	Nash	Nash sqrt	Nash Ln	PBIAS
CA	1	2	20_mile	100	307	natural	198	77	80	76	16	82	87	86	15
USA	3	2	Genessee	100	6317	regulated	418	81	83	79	2	79	81	82	3
USA	4bis	2	Irondequoit	100	326	natural	172	67	71	66	2	53	64	69	20
USA	5	2	Oswego	100	13287	regulated	259	84	83	79	2	71	73	72	12
USA	6	1	N/A	40	2406	mixed	264	71	76	76	2	73	77	79	13
USA	7	2	Black river	100	4847	regulated	471	79	82	82	2	75	75	72	9
USA	8	2	Oswegatchie	100	2543	regulated	250	79	82	84	2	82	84	85	9
CA	10	2	Salmon_CA	100	912	regulated	196	86	88	81	7	82	86	77	3
CA	10bis	1	N/A	44.2	944	mixed	115	82	89	85	8	78	81	79	10
CA	11	2	Moira	100	2582	regulated	228	88	90	84	4	82	81	75	0
CA	12	1	N/A	88	12515.5	regulated	282	73	73	66	4	66	69	62	-14
CA	13	1	N/A	40.3	1537.5	natural	178	62	66	63	3	65	70	70	-1
CA	14	1	N/A	61.3	2689.4	mixed	209	79	78	71	5	71	77	76	9
CA	15	1	N/A	63	2245.8	mixed	263	86	84	71	1	86	87	84	-8

This also tends to further confirm findings of the GRIP-M project (Fry *et al.*, 2014), namely that the ARM can provide reliable runoff estimations for a whole sub-basin even if less than half of it is gauged. To further confirm this, calibration scheme 2 was tested for sub-basins 6 and 14, so that both schemes can be truly compared for these catchments. The results (not shown here) confirm that performances are extremely close for both, sometimes even slightly better with calibration scheme 1.

Validation performances are often lower than in calibration (see sub-basins 4bis, 5, 7, 10bis, and 11 in Figure 13), especially when looking at the PBIAS criteria (Equation 1), which thus indicates that runoff is generally underestimated. When looking at Table 19, the higher drops in validation performances tend to be associated with regulated catchments (sub-basins 5, 7, 10bis and 11), even if other regulated catchments display similar, even better performances over validation (catchments 8 and 10). The link between the type of regulation and model robustness can however not be in-depth analyzed, because of a lack of information in regard of regulation practice (see section 5.1). However, the fact that this high drop of performances between calibration and validation, for the regulated catchments named above, is not seen when using CaPA precipitation (not shown here), does not support the hypothesis of a dynamic type of regulation which would be mimicked by the model in calibration but not as well in validation. This seems to have more to do with the precipitation source, and will be discussed in section 7.3.2.

Finally, the fact that with GR4J, the three different Nash criterion (Table 19) all display satisfactory values, generally close to the one used as objective function (both in calibration and validation), tends to indicate that the different components of streamflow are well reproduced by the model (baseflow, interflow, surface runoff, low-flow periods, floodings). Despite no test was performed here with using the traditional Nash criteria as the objective function, it is argued that using Nash sqrt (see section 5.3) as the objective function is a reasonable choice and a fair trade-off between putting the emphasis on peak or low flows, when using GR4J (but not LBRM, see section 7.3.2).

7.3.2 Model and precipitation dataset inter-comparison

It seems from Table 20 that GR4J performs generally slightly better than LBRM when looking at Nash sqrt values (the objective function), but results from Table 21 indicate that for this criteria, it cannot be stated with confidence that GR4J is better than LBRM. The superiority of GR4J over LBRM is however significant for the Nash criteria (Table 21). This suggests that calibrating LBRM using the Nash sqrt objective function may not be a good choice for this model, as it generally has difficulty simulating peak flow events in this case (illustrated on Figure 7 and Figure 8 below).

Table 20: Nash values (using the square root of flow values - NASH SQRT, or the flow values themselves - NASH) obtained with the GR4J model, minus Nash values obtained with LBRM, in calibration (CALIBR) or validation (VALID), and with the GHCND or CaPA precipitation (P.) datasets. Values are in percent, so that a value of 5 corresponds for example to a Nash of 75 (/100) for GR4J and of 70 for LBRM.

		GLERL_P.				CAPA_P.			
		NASH SQRT		NASH		NASH SQRT		NASH	
Subbasin	Country	CALIBR	VALID	CALIBR	VALID	CALIBR	VALID	CALIBR	VALID
1	CA	9.1	14.4	12.8	21.7	9.1	15.0	14.6	22.5
3	US	2.9	7.4	3.3	6.7	1.0	-0.2	1.4	1.2
4bis	US	-2.6	9.5	4.0	7.6	-1.6	4.2	3.5	11.4
5	US	-4.0	-3.8	-4.6	-3.5	-6.0	-8.4	-7.2	-15.6
6	US	8.1	5.5	10.3	9.2	2.3	0.1	4.0	1.8
7	US	4.1	1.3	5.4	7.2	-1.4	2.2	-5.5	-2.8
8	US	5.8	8.8	3.6	10.6	0.6	6.9	-0.8	9.8
10	CA	-1.9	2.6	-2.6	1.8	-1.5	1.5	-5.9	6.1
10bis	CA	2.8	0.4	4.0	-3.6	1.6	4.6	-2.0	8.1
11	CA	0.0	0.3	4.4	7.9	0.6	0.3	2.0	2.5
12	CA	-2.4	-4.3	0.4	-2.1	-2.8	-0.7	0.2	-0.4
13	CA	0.5	-1.2	9.4	-0.3	5.6	-8.4	10.5	-13.1
14	CA	0.5	0.5	4.8	3.3	2.7	-2.7	8.1	4.6
15	CA	-1.7	8.7	7.9	12.2	0.8	3.6	8.0	1.7
	mediane US	3.5	6.4	3.8	7.4	-0.4	1.1	0.3	1.5
	mediane CA	0.2	0.4	4.6	2.5	1.2	0.9	5.0	3.5
	mediane	0.5	1.9	4.2	6.9	0.7	0.9	1.7	2.2
number of positive cases		9/14	11/14	12/14	10/14	9/14	9/14	10/14	10/14

Table 21: number of positive cases associated to the inter-comparison study. In other words, number of cases with GR4J model or GHCND precipitation having 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 21 was derived from the mean of performances along the two precipitation datasets or models, respectively. The degree of confidence (significance) associated to main numbers is displayed. A number equal or greater than 11 out of 14 cases is considered significant (i.e. more than 95%).

	NASH SQRT		NASH	
	CALIBR	VALID	CALIBR	VALID
GR4J minus LBRM, mean GHNCD-CaPA	9/14	10/14	12/14	11/14
GHCND minus CaPA, mean GR4J-LBRM	12/14	9/14	13/14	9/14
number of cases	9/14	10/14	11/14	12/14
associated significance	80%	91%	97%	99%

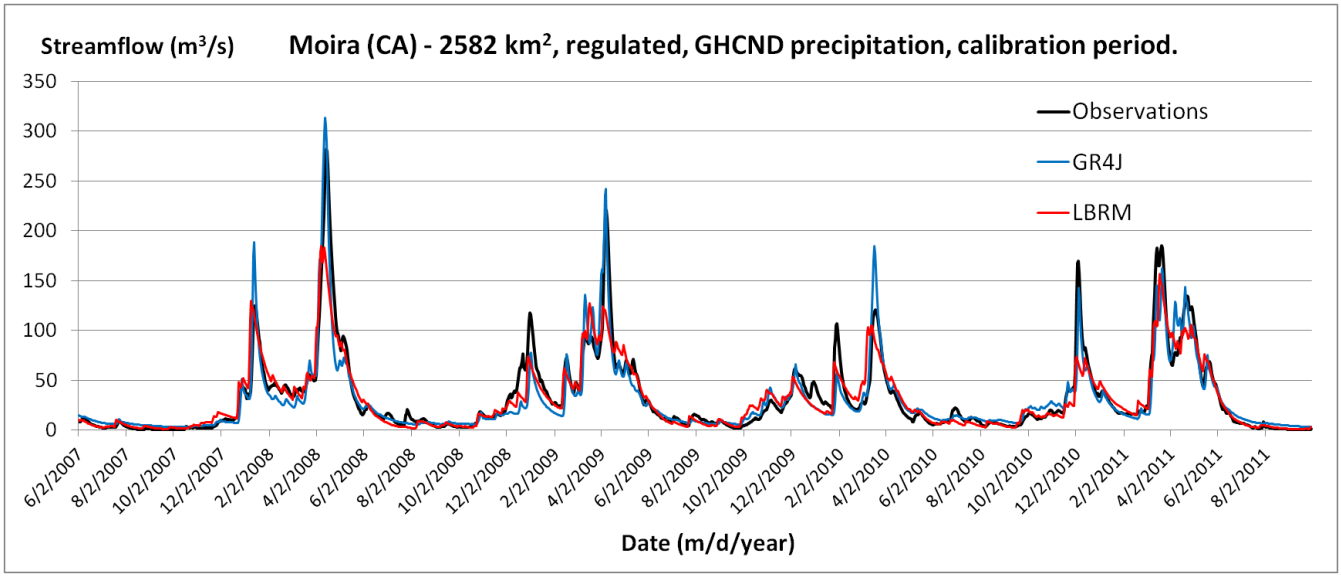

Figure 7: Sub-basin 11 (Moira river) Hydrographs in calibration, obtained with GHCND precipitation. Nash sqrt values for GR4J and LBRM are respectively 89.6 and 89.6 for the corresponding period. Nash values are 88.2 and 83.8, respectively.

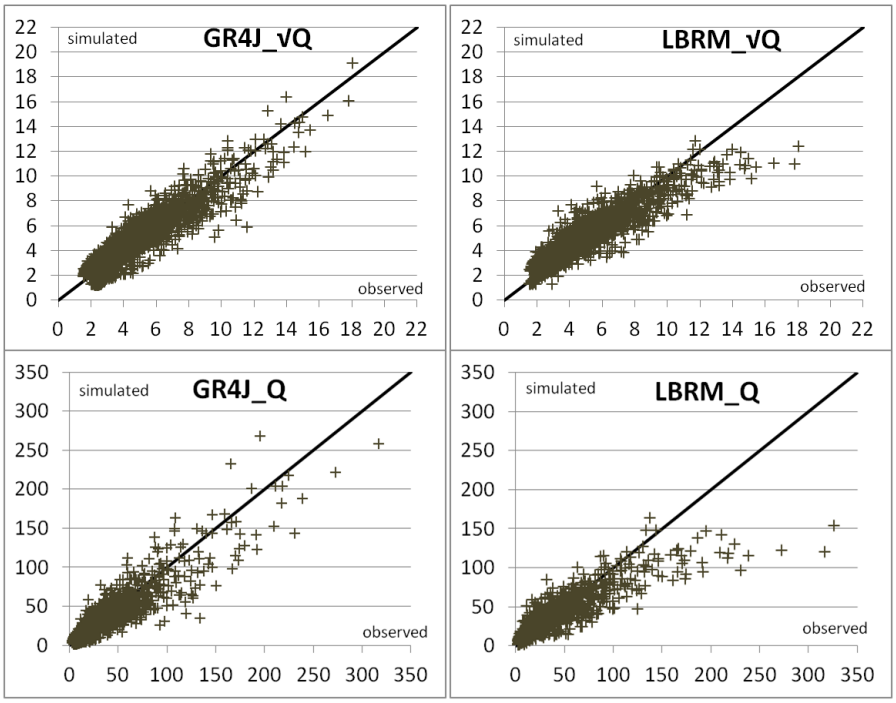

Figure 8: Sub-basin 15 plots of simulated (y axis) versus observed (x axis) flows (Q) or square-root of the flows (√Q) for the whole GRIP-O period (validation and calibration) and with the GHCND dataset, either with GR4J or LBRM models. See Table 21 for a quantitative assessment of the difference between model performances.

Table 22 contains some valuable information about the comparison between the GHCND and CaPA precipitation datasets. First of all, it is quite clear that using the GHCND precipitation dataset leads to better calibration performances than when using the CaPA dataset. This logically let us suppose 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 (see Table 3).

Table 22: Nash sqrt and Nash values obtained with the GHCND precipitation dataset, minus Nash 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 Nash of 70 (/100) with CaPA and of 75 with GHCND precipitation.

		LBRM				GR4J			
		NASH SQRT		NASH		NASH SQRT		NASH	
Subbasin	Country	CALIBR	VALID	CALIBR	VALID	CALIBR	VALID	CALIBR	VALID
1	CA	0.1	7.2	0.9	1.9	0.1	6.6	-0.9	1.1
3	US	1.1	5.4	0.6	5.4	3.0	13.0	2.5	10.9
4bis	US	3.9	-8.6	6.3	-5.2	2.9	-3.3	6.7	-8.9
5	US	1.8	2.5	1.9	3.1	3.8	7.0	4.5	15.2
6	US	-0.2	-5.5	1.2	-9.7	5.6	0.0	7.5	-2.4
7	US	3.9	2.6	3.5	1.5	9.4	1.7	14.4	11.4
8	US	5.8	0.4	6.4	3.4	11.0	2.3	10.7	4.2
10	CA	2.1	-0.3	2.3	2.2	1.7	0.8	5.5	-2.1
10bis	CA	1.0	-0.7	1.0	5.6	2.1	-4.9	7.0	-6.1
11	CA	2.2	-2.9	-0.2	-3.2	1.6	-2.8	2.1	2.2
12	CA	2.6	-4.1	2.1	-5.6	3.0	-7.7	2.3	-7.3
13	CA	1.6	-0.5	-1.3	-1.9	-3.5	6.7	-2.3	10.9
14	CA	1.9	6.2	2.4	11.8	-0.3	9.4	-1.0	10.5
15	CA	0.9	-1.3	1.7	-1.4	-1.6	3.7	1.5	9.1
	mediane US	2.8	1.4	2.7	2.3	4.7	2.0	7.1	7.6
	mediane CA	1.7	-0.6	1.3	0.2	0.9	2.3	1.8	1.7
	mediane	1.9	-0.4	1.8	1.7	2.5	2.0	3.5	3.2
number of positive cases		13/14	7/14	12/14	8/14	11/14	10/14	11/14	9/14

It can also be noticed that the comparison made in Table 22 is quite different when looking at the US or Canadian side of lake Ontario watershed. In Canada, the model performances resulting from GHCND and CaPA precipitation are close (see "mediane CA" in Table 22), while in the US ("mediane US") the difference resulting from using one source or the other is more pronounced. When having in mind the information contained in Table 3 at the same time (namely that in Canada, the density of the GHCND is much higher than the density of the ground stations used in CaPA), it can be formulated 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 (Table 23). In the US, the higher difference between the performances resulting from the two different precipitation datasets seems to be mainly due to sub-basins 7 and 8, for which the GHCND data lead to significantly better performances than CaPA, for both GR4J and LBRM.

Interestingly, these sub-basins contain or are located close to a small mountain on the eastern shore of lake Ontario (see

Figure 1), which is well-known for being the theater of local microclimate phenomenon involving higher annual rainfall amounts than in the surroundings. This is because of moisture brought by the frequent western winds traveling above the lake and being forced to uplifting due to the local positive topographic anomaly contained in or close to sub-catchments 7 and 8 (orographic precipitation). As a consequence, it is highly possible that these complex local effects are not well captured by the RDPS, which would logically affect CaPA rainfall amounts. In order to bring evidence to the assumptions formulated in this paragraph, the correlation was assessed between the daily average rainfall amounts of CaPA and the GHCND, and between the RDPS and the GHCND, for a few GRIP-O sub-basins in Table 23. It can be seen from Table 23 that correlation coefficients are indeed lower between CaPA and the GHCND, for sub-basins 7 and 8, in comparison to other GRIP-O sub-catchments.

Table 23: Correlation coefficients between CaPA/RDPS and the GHCND (observed) daily basin average precipitation for three GRIP-O sub-catchments (see

Figure **1**). RDPS: Canadian Regional Deterministic Prediction System (used in CaPA).

corelation with GHCND for:	CAPA	RDPS
Sub-basin 5	0.95	0.86
Sub-basin 6	0.89	0.80
Sub-basin 7	0.87	0.78
Sub-basin 8	0.85	0.76
Sub-basin 12	0.93	0.84
Sub-basin 15	0.91	0.77

A statement emanating from the interpretation of Table 22 remains to be highlighted. The difference between the GHCND and CaPA resulting performances, which is clear in calibration, cannot be stated with any confidence in validation (Table 21). In other words, CaPA precipitation does lead to better simulation performances for validation in several cases, such as for sub-catchments 4bis and 10 to 12. When looking at Table 3, one could argue that the lower robustness of the GHCND in comparison to CaPA is, at least for the U.S. basins, due to the fact that the density of the GHCND network decreases when moving from calibration to validation. However, the fact that this lower robustness can also be seen for the Canadian basins, for which the density of the GHCND network does increase in validation compared to the calibration period, forces us to propose another explanation.

This is supposed to be due to the fact that with GHCND precipitation, watershed averages are computed using the Thiessen interpolation method (i.e., nearest neighbors).

Despite the density of the GHCND network is quite high in the lake Ontario watershed, this interpolation method assumes that precipitation amounts are uniformly distributed in space inside each polygon delineated by the Thiessen method. Therefore, it can sometimes poorly reflect the spatial variability of precipitation, especially in summer where convective events can occur locally between two rain gauges, whereas the RDPS precipitation fields could capture, at some degree, some of these local events. Another possibility that could explain the sometimes better performances achieved with CaPA is that this analysis only uses rain gauges which are available in real-time and known to be generally reliable, i.e. they often possess a windshield in order to decrease the error associated to the solid precipitation "undercatch" issue (Adam and Lettenmaier 2003). The GHCND, on the opposite, contains a lot of stations which do not possess any windshield, and can then be more affected by a low precipitation bias especially over winter periods. As a consequence, it is not impossible that CaPA could sometimes better represent the winter precipitation amounts than the GHCND.

In summary, it is envisioned that the models compensate, in calibration, for some systematic errors contained in the GHCND precipitation, but which are not the same in validation (especially given that the density of the GHCND network varies in time) and would explain why model robustness is poorer with GHCND than with CaPA precipitation. Note that this behavior is not in contradiction with the fact that the GHCND-derived precipitation amounts are supposed to better reflect real values than CaPA amounts.

7.3.3 Temporal details of performances

Additional experiments were performed in order to assess the sensitivity of the evaluation metrics when looking at the observed and simulated time-series in a different way or in more details. For example, weekly performances were computed with the same metrics as for daily time-series, in order to see if doing so could help reduce the influence of regulation on performances. For example, it was noticed, when looking at hydrographs, that U.S. observed streamflows, especially for sub-basins 3 and 5 (Genessee and Oswego rivers, not shown here) tend to display a more instable and somehow cyclic behavior, when compared to simulated time-series. No such behavior could be seen for Canadian sub-catchments. Consequently, it is assumed that the U.S. and Canadian practices, in the lake Ontario watershed area, involve different types of regulation. When looking at the weekly performances however (not shown here), despite being always better on a weekly than a daily basis (Nash and Nash sqrt values are generally 2 to 6 % higher on a weekly basis), values for the regulated catchments did not show a stronger improvement than for natural ones, when using a seven-day time-step. No additional information could hence be gathered from this test, and despite improving the performances,

evaluating simulations on a 7-day basis does not allow to significantly reduce the effect of regulation.

Performances were also assessed on a seasonal basis, i.e. either for winter (November - May) or summer (June- October) periods, but still using a daily-time-step. The analysis was thus limited to the calibration period. Winter performances tend to be better than summer ones, in terms of all Nash criteria used in this study (Table 24). This is common in Hydrology, as winter processes are generally less chaotic and easier to reproduce than summer ones. For example, winter streamflows in the area studied are frequently governed by Temperature via snowmelt processes. In summer however, convective storms can be localized over very small areas and can be poorly captured by observation networks. Apart from this, there is a priori not much information which can be derived from this seasonal analysis. It is only highlighted that despite better than summer ones, winter Nash 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%.

Table 24: Seasonal performances (Nash sqrt) of LBRM and GR4J over the calibration period. WHOL: whole period; SUMM: summer period (1st June - 31st October); WINT: winter (1st November - 31st May).

			GHCND						CAPA					
			GR4J			LBRM			GR4J			LBRM		
Station	Area(km2)	Regime	WHOL	SUMM	WINT	WHOL	SUMM	WINT	WHOL	SUMM	WINT	WHOL	SUMM	WINT
20_mile	307.0	natural	79.8	64.1	78.3	70.7	34.5	71.0	79.7	57.4	79.6	70.6	24.8	69.5
Genessee	6317.0	regulated	83.1	75.4	80.1	80.2	68.6	77.5	80.1	68.1	77.6	79.1	56.7	77.2
Irondequoit	326.0	natural	71.4	55.7	70.0	74.0	73.8	68.4	68.6	55.4	65.6	70.2	1.1	63.1
Oswego	13287.0	regulated	83.2	70.8	76.6	87.2	77.2	82.6	79.4	62.1	72.3	85.4	71.7	80.0
N/A	2406.0	mixed	76.2	67.6	71.6	68.1	59.2	60.6	70.5	55.9	66.7	68.3	51.8	62.1
Black river	4847.0	regulated	81.9	80.0	77.7	77.8	80.5	70.2	72.5	69.6	66.3	73.9	70.6	66.8
Oswegatchie	2543.0	regulated	82.1	81.2	76.5	76.3	73.7	70.1	71.1	65.9	64.7	70.5	60.7	63.5
Salmon_CA	912.0	regulated	88.3	50.2	82.8	90.2	65.3	84.1	86.6	49.5	79.1	88.1	32.2	80.4
N/A	944.0	mixed	88.7	66.8	83.2	86.0	50.2	80.3	86.6	63.4	79.6	85.0	33.0	78.1
Moira	2582.0	regulated	89.6	61.4	83.8	89.6	71.0	82.2	87.9	41.2	83.8	87.3	38.4	80.1
N/A	12515.5	regulated	72.8	44.0	68.3	75.2	63.6	65.7	69.8	40.3	64.0	72.6	55.3	61.8
N/A	1537.5	natural	66.4	52.0	67.8	65.9	64.0	62.9	69.9	60.3	70.0	64.3	37.7	62.0
N/A	2689.4	mixed	78.3	74.7	79.1	77.8	80.3	75.8	78.6	73.9	79.9	75.9	71.6	74.0
N/A	2245.8	mixed	83.7	73.5	84.2	85.4	81.1	84.2	85.3	77.0	85.6	84.5	73.8	82.0

Finally, it can be seen from Table 25 below that winter streamflows are generally more strongly underestimated by the models than summer periods, especially for GR4J which even shows the tendency of overestimating streamflows over summer periods. In summer, this is

supposed to be due to the fact that as regulation over Canadian watersheds seems to consist in a simple weir behavior which dampens summer peak flow events but cannot prevent winter ones due to their higher magnitude, the GR4J model simulates more runoff in summer. In other words, regulation would tend to store water in summer and release it in winter (again, this remains an hypothesis and too few indications are available to confirm it), which models cannot truly handle (LBRM seems to better reproduce this behavior than GR4J, however, based on Table 25 results). This could also partly explain the underestimation of winter flows. The hydrographs of the Moira river, for example (Figure 7), tend to make sense with such an hypothesis.

Table 25: Seasonal performances (PBIAS) of LBRM and GR4J over the calibration period. WHOL: whole period; SUMM: summer period (1st June - 31st October); WINT: winter (1st November - 31st May).

			GHCND						CAPA					
			GR4J			LBRM			GR4J			LBRM		
Station	Area(km2)	Regime	WHOL	SUMM	WINT	WHOL	SUMM	WINT	WHOL	SUMM	WINT	WHOL	SUMM	WINT
20_mile	307.0	natural	16.3	9.3	17.0	16.7	14.5	16.9	13.4	35.2	10.2	16.3	27.9	15.8
Genessee	6317.0	regulated	2.4	4.6	1.7	5.8	1.5	7.2	2.7	3.8	2.3	3.7	0.3	4.9
Irondequoit	326.0	natural	2.4	-13.1	8.2	4.1	1.0	5.3	5.3	-13.1	12.2	2.0	7.8	2.0
Oswego	13287.0	regulated	2.2	-0.3	2.9	2.4	3.9	1.9	1.6	-2.8	2.8	1.3	4.6	0.2
N/A	2406.0	mixed	2.3	7.1	0.4	11.2	-0.8	15.3	4.0	7.1	2.8	6.6	5.8	6.8
Black river	4847.0	regulated	2.3	-5.8	5.5	6.1	-1.3	9.1	1.9	3.8	1.1	3.3	1.0	4.2
Oswegatchie	2543.0	regulated	1.6	1.0	1.8	3.3	-7.7	7.7	0.9	1.2	0.7	4.9	0.5	6.7
Salmon_CA	912.0	regulated	7.0	-37.3	12.1	4.6	-2.7	5.4	5.0	-26.5	8.7	3.7	0.0	4.6
N/A	944.0	mixed	7.9	-28.2	11.8	10.3	-5.1	11.9	5.3	-1.1	6.0	6.4	10.9	6.3
Moira	2582.0	regulated	4.3	-29.5	8.7	3.7	-1.2	4.3	3.4	-13.2	5.5	3.8	6.6	3.6
N/A	12515.5	regulated	4.3	-3.5	6.6	2.7	1.4	3.1	5.1	-3.3	7.5	4.9	8.4	3.8
N/A	1537.5	natural	2.8	-8.5	7.8	5.5	-0.8	8.2	2.4	-13.8	9.5	5.3	-5.4	10.2
N/A	2689.4	mixed	4.6	-0.5	6.8	3.4	0.5	4.6	5.0	4.2	5.3	5.2	-2.2	8.6
N/A	2245.8	mixed	0.7	-9.0	4.4	3.8	-2.0	6.0	1.2	-9.4	5.1	3.2	0.3	4.6

7.3.4 Runoff estimation for the whole GRIP-O area

In order to have a better overview of the performance of lake Ontario runoff simulations performed in this work, an evaluation was made by looking at the sum of runoff from the models implemented locally over each of the 14 GRIP-O gauged catchments (

Figure 1). The total area covered by the models is about 53460 km^2, i.e. more than the total gauged area of Figure 1 (47330 km^2), because of calibration scheme 1 (section 5.2) which involves implementing a model over the whole area of a sub-catchment, including its ungauged part. Therefore, runoff observations used for the 53460 km^2 area actually correspond to

estimations because they are derived from runoff truly observed for the gauged 47330 km^2 area (brown part of

Figure 1) and the ARM. However, these estimations are supposed to be close to reality as more than 88% of the studied area is gauged.

Moreover, a test was made with implementing a unique GR4J model over this 53460 km^2 area at once, using the GHCND. In this case, the GR4J model was calibrated using Precipitation and Temperature values corresponding to weighted averages of the 14 sub-catchments' time-series, and the sum of locally observed flows (extrapolated with the ARM) as reference. When using a unique model, the entire area is conceptually interpreted in the model as a unique watershed with one main river, which is far from reality (

Figure 1). The unique model was implemented here considering 5 different elevation classes of equal size, and still the Nash sqrt value as objective function.

This can be seen from Table 26 that the performances are very good (and better than performances obtained for any of the local sub-basins, see Table 19) when looking at the simulated watershed as a whole, whatever the quality criteria considered. This suggests that local model biases are compensated when grouping the catchments altogether, but this is also due to the fact that the Nash criteria is sensitive to streamflow magnitude, which is high in the case of Table 26 (see Figure 9). 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 criteria which stands somewhere in-between the extreme values obtained with the individual combinations.

Table 26: Daily runoff simulation performances when assessed for the 57460 km^2 lake Ontario area (see text). CAL, VAL: calibration, validation. 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 -		GR4J unique	
	CAL		VAL		CAL		VAL		GHCND - CAPA		GHCND	
	GR4J	LBRM	GR4J	LBRM	GR4J	LBRM	GR4J	LBRM	CAL	VAL	CAL	VAL
Nash sqrt	0.92	0.91	0.90	0.88	0.89	0.90	0.88	0.87	0.92	0.90	0.88	0.85
Nash	0.92	0.90	0.89	0.86	0.88	0.89	0.84	0.86	0.92	0.89	0.89	0.85
Nash ln	0.91	0.91	0.89	0.89	0.88	0.90	0.88	0.87	0.92	0.90	0.88	0.83
PBIAS	2.96	4.61	5.08	10.02	2.92	3.78	6.83	9.55	3.60	7.87	-0.31	0.34

All four individual combinations of Table 26 display very robust performances, which again highlights the compensation of local biases, which can especially be seen in validation (compare with Table 19). In all cases of Table 26 can it be noticed that observed runoff is generally underestimated by the models, which can be more precisely linked to the underestimation of peak flow events (Figure 9). It is nonetheless stated that any of the two global 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.

Finally, using a unique GR4J model to save computation time and effort for estimating runoff for the whole GRIP-O area described earlier in this section, revealed promising. Despite values for the different Nash criterion are slightly smaller for this unique model in comparison to performances obtained with summing all local models together (Table 26 and Figure 10), the opposite is true for the PBIAS criteria. As a consequence, 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 sub-basins. Tests were also performed by trying to calibrate the unique model using the conventional Nash criteria, but results were almost exactly the same (not shown here), which tends to suggest that for GR4J, calibrating on Nash sqrt or Nash does not change model performances. A calibration test with the Nash instead of Nash sqrt criteria performed on the Moira watershed (and still with GR4J and the GHCND precipitation) even revealed poorer simulation results with the former conventional criteria than the with the objective function used in this work. Using 5 elevation classes (as presented here) lead to slightly better performances and hydrographs than with a unique elevation class (not shown), but this difference may reveal more significant in case of watersheds with more pronounced topography.

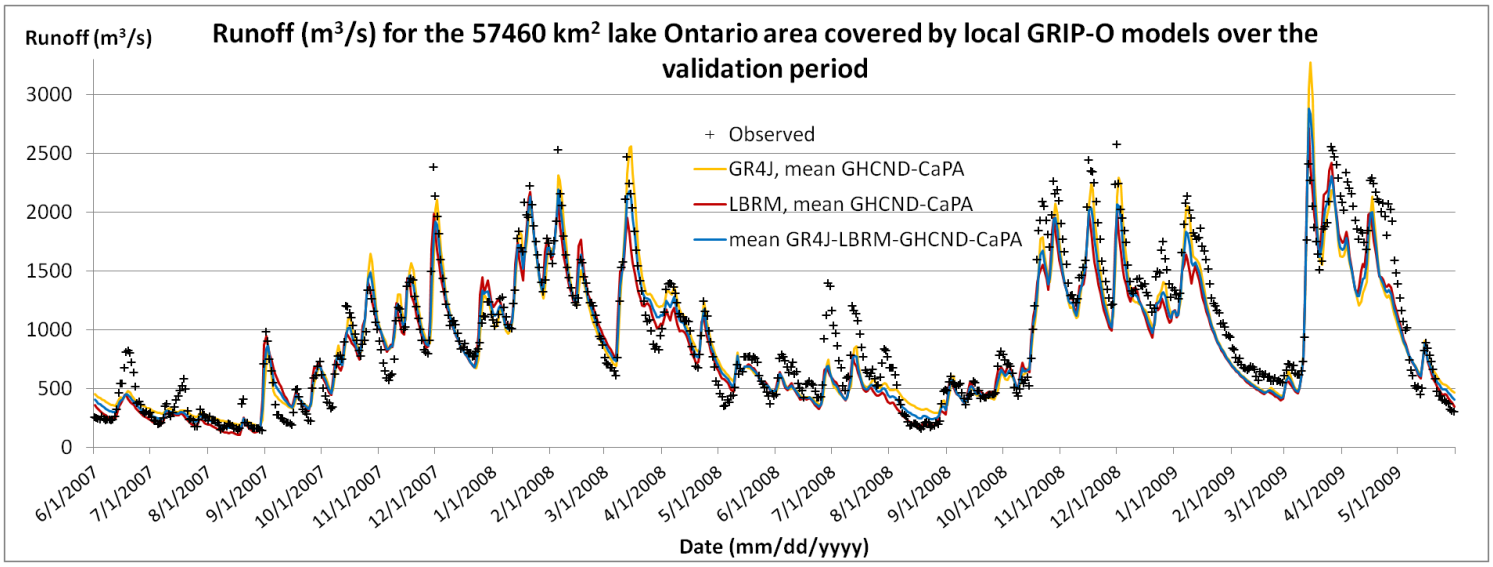

Figure 9: comparison between observed (estimated, see text) and simulated runoff for the whole lake Ontario area covered by the local models implemented during the GRIP-O experiments (see text). Results are presented for the GRIP-O validation period; simulated values correspond either to the mean of values simulated with the four different combination possibilities (two models, two precipitation sources), or the mean of simulations with the two precipitation datasets (mean GHCND-CaPA), for each of the two models.

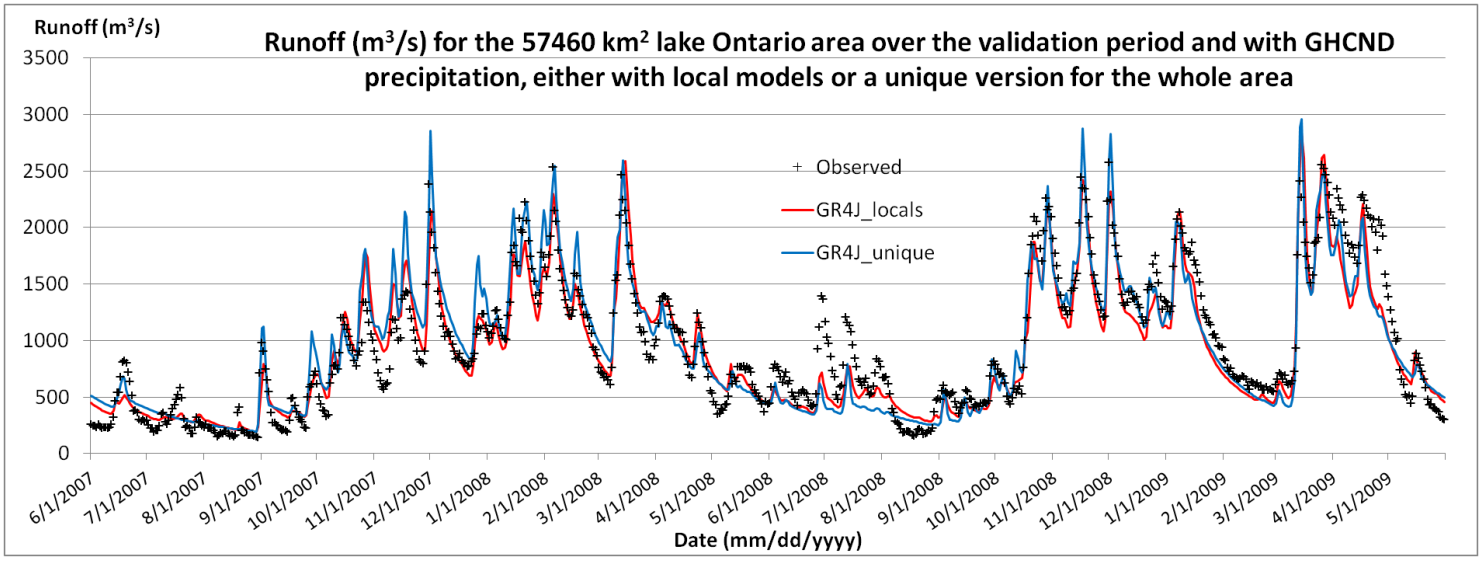

Figure 10: comparison between hydrographs for the 57460 km^2 lake Ontario area obtained either with summing runoff from the 14 local GR4J models (GR4J_locals) or with the unique GR4J model (GR4J_unique) calibrated over the whole area at once. In both cases, runoff was produced with using the GHCND precipitation. See text for details about the unique model implementation.

7.3.5 Correlation between parameter values and catchment characteristics

An attempt was made to correlate parameter values to catchment properties, which, if successful, could serve to transfer parameters to ungauged areas, or at least allow to better define initial values for calibration. To this aim, some catchment characteristics, which do not consist in quantitative values such as the country it is located in, the presence or absence of regulation, and the precipitation dataset used, were transformed into numeric values in a manner defined in Table 27 and Table 28 below. If a particular parameter tends to be strongly related to one or the other possibility, then it should translate into the correlation coefficient values. At the same time, it was checked that the performances obtained for the objective function were truly not related to any of the catchment properties, as argued in section 7.3.1.

Table 27: correlation coefficients between LBRM parameters and performances, and some catchment properties. Correlation coefficient values above 0.5 are highlighted. Lin.: linear, res.: reservoir, coeff.: coefficient, evap.: evaporation, perco.: percolation, USZ: Upper Soil Zone. Nash sqrt.: Nash computed with the square-root of the flows (objective function); cal., val.: calibration and validation; precip.: precipitation, sub.: sub-catchment.

Parameters - criteria / correlation with:	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	Nash sqrt cal	Nash sqrt val
	deg C	cm/˚C./d	d-1	m-3	d-1	d-1	m-3	d-1	d-1	cm		
area	-0.10	-0.13	0.13	-0.39	0.03	0.08	0.07	-0.14	-0.35	-0.05	0.15	0.09
mean altitude	-0.05	-0.08	0.37	-0.45	0.00	-0.01	-0.06	-0.04	-0.36	-0.24	-0.03	-0.08
regulation (0 no, 0.5 mixed, 1 yes)	-0.46	-0.13	0.18	-0.11	-0.12	-0.11	0.22	-0.28	-0.71	-0.27	0.55	0.54
cal. Scheme (1 whole sub., 2 gauge location)	-0.38	-0.01	0.15	0.10	-0.36	-0.36	0.44	0.18	-0.50	-0.45	0.24	-0.18
precip (0 GHCND, 1 CAPA)	0.31	0.34	0.20	0.18	0.12	0.05	-0.02	-0.24	-0.09	0.19	-0.13	0.00
Country (0 CA., 1 US)	-0.21	-0.02	0.22	-0.44	-0.27	-0.27	0.05	-0.13	-0.37	-0.48	-0.22	-0.44
Nash sqrt cal.	-0.43	-0.34	0.03	0.34	0.01	-0.02	0.41	-0.18	-0.48	-0.16	N/A	N/A
Nash sqrt val.	-0.23	0.05	-0.18	0.31	0.10	0.05	0.16	-0.06	-0.31	0.21	N/A	N/A

As can be seen from Table 27 (correlation with LBRM), most of LBRM parameters do not seem to be significantly related to any of the GRIP-O sub-catchment properties, except for the linear reservoir surface flow coefficient, which seems to be quite strongly correlated with the regulation attribute of the catchments, and to a lower degree to the calibration scheme used. The fact that this parameter is negatively correlated to regulation makes sense as it has a strong influence on the simulated hydrographs by affecting the reactivity of the watershed. If this

parameter has a higher value, then more peak flows will be simulated. However, as regulation tends to decrease the number and magnitude of peak flow events in the downstream rivers, it will generally be accompanied with a lower value for this LBRM parameter. The fact that it is also slightly related to the type of calibration scheme also makes sense here, because calibration scheme 1 involves simulation for a whole catchment with several most-downstream gauges near the lake shoreline (see section 5.1). As a consequence, peak flows resulting from a rain event for such catchments is supposed to occur more rapidly and with a higher magnitude than in the case of only one gauge for the catchment, because of the shape associated to these different watersheds (see

Figure 1): with several gauges, upstream areas will together contribute more rapidly and in a more synchronized manner to streamflow, while with one gauge, the upstream area is generally closer to a more elongated shape which involves a longer distribution of runoff in time, at the gauge location, because not all parts of the upstream area will contribute to the gauge streamflow at the same time. To finish with the case of LBRM, it seems that the performances of the model do actually show some kind of a link with the type of regulation performed: when the catchment's streamflow is regulated, performances generally tend to be higher than when it is not. This can be a priori surprising, but it is argued here that this is because regulation not always results in a "random" downstream streamflow, but, on the contrary, that it can somehow facilitates the job of the model by tampering peak flows and producing streamflows with smoother seasonal trends, an effect that is probably due to the fact that regulation involves high storage capacities. Even in the case of U.S. regulated catchments that tend to be accompanied by a more "nervous" behavior for the resulting flows does the model correctly reproduce the observed trends (see Figure 5 for the Oswego basin in calibration). However, the correlation between regulation and model performances remains low.

Table 28 presents the same type of information as Table 27, but for GR4J this time. However, cases where X3 was way below X1 were not taken into account to produce Table 28, as these "uncommon" values tend to affect this exercise of linking parameter values to catchment properties. As for LBRM, some GR4J parameters seem to present some correlation with regulation status. This does make sense because as stated previously, regulation is accompanied with high reservoir capacities. Hence in such cases, the model tends to decrease the production store capacity and increase the one of the routing store. It is thus reminded that despite not taken into account to produce Table 28 below, cases with high routing store capacity values compared to production store capacities were always associated to regulated catchments, confirming the explanation given above.

Table 28: correlation coefficients between GR4J parameters and performances, and some catchment properties. Correlation coefficient values above 0.5 are highlighted. The cases where

X3 was way below X1 were not taken into account to produce this table (see Table 9). Nash sqrt.: Nash computed with the square-root of the flows (objective function); cal., val.: calibration and validation; precip.: precipitation, sub.: sub-catchment, UH: Unit Hydrograph, loc.: location

Parameters - criteria / correlation with:	Capacity of production store	Water exchange coefficient	Capacity of routing store	UH time base	degree-day factor	snowpack inertia factor	Nash sqrt cal	Nash sqrt val
	X1, mm	X2, mm	X3, mm	X4, days	X5, []	X6, []		
area	-0.01	0.25	0.05	0.33	0.18	-0.11	-0.12	-0.33
mean altitude	-0.08	0.28	0.05	0.32	0.04	0.07	-0.07	-0.11
regulation (0 no, 0.5 mixed, 1 yes)	-0.52	-0.30	0.74	0.71	-0.32	-0.39	0.40	0.23
cal. Scheme (1 whole sub., 2 gauge loc.)	-0.46	-0.41	0.39	0.65	-0.42	0.33	0.22	0.07
precip (0 GHCND, 1 CAPA)	0.26	-0.10	-0.01	-0.09	-0.31	-0.34	-0.20	-0.16
Country (0 CA., 1 US)	0.02	0.61	-0.17	0.11	0.03	0.11	-0.29	-0.34
Nash sqrt cal.	-0.74	-0.58	0.63	0.54	-0.55	-0.12	N/A	N/A
Nash sqrt val.	-0.84	-0.43	0.53	0.46	-0.42	0.08	N/A	N/A

Regulation is also logically generally accompanied by a longer residence time, justifying the positive correlation with the Unit Hydrograph base time. This latter parameter is also logically correlated to the shape of the catchments, which translated here into the calibration scheme used, as explained above for LBRM (see text associated to Table 27). The link between X2 (water exchange coefficient, adding water when positive) and the country is not clearly known but assumed to be also somehow related to regulation. For Canadian GRIP-O basins, regulation almost removes all peak flow events from the downstream flows during summer (see for example Figure 7) while it is not the case in the U.S. where regulation involves a strong streamflow variability even in summer (see Figure 5). The fact that this exchange coefficient is set to low (negative) values for Canadian regulated watersheds (not shown here) is supposed to allow GR4J to prevent simulating peak flows in summer. Finally, it can be seen that most of GR4J parameters do show some correlation with the model performances. This is supposed that it could not be seen with LBRM parameters mainly because of the fact the LBRM has more free parameters and that there is more inter-dependency between the parameters and their resulting effect on streamflow, which results in lower correlations between LBRM parameters and the model performances. With GR4J however, each parameter seems to have a significant effect on streamflows, which translated into a stronger correlation between the parameters and model performances.

7.4) Conclusion regarding the lumped models' results

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 streamflows. Both models are robust and perform well whatever the precipitation dataset used as input.

However, GR4J performs generally better than LBRM when looking at the Nash criteria. This suggests that using the Nash sqrt value as the objective function (Nash value computed using the square-root of the flows) may not be the best choice for implementing LBRM, which would probably achieve better hydrologic simulations using the conventional Nash criteria.

The GHCND dataset leads to better performances than CaPA precipitation in calibration, especially when sub-catchments involve complex local meteorological phenomenon such as orographic precipitation. However, using CaPA to drive the experiments can sometimes result in the models showing more robustness than with the GHCND, suggesting that the latter source of precipitation is not devoid from drawbacks. CaPA thus leads to hydrologic performances very close to the ones obtained with the GHCND for some GRIP-O sub-watersheds, even in the case where the network of ground stations used in CaPA has a significantly lower density than the GHCND stations (such as for Canadian sub-basins in the north-west part of the lake Ontario watershed, see Table 3 and

Figure 1), suggesting that CaPA is a very useful source of (daily) precipitation data.

The two different calibration schemes tried during this study both result in very satisfactory performances. Using synthetic streamflow time series for a whole sub-catchment, derived from observed data and a simple ARM, hence consist in a promising and efficient way to estimate the contribution of ungauged parts of a gauged sub-catchment.

Despite regulation affects most of the GRIP-O sub-basins, it did not prevent the models from achieving excellent streamflow simulations. The models, through calibration, generally find a way to mimic the effect of regulation.

In the light of the performances obtained in this study, promising results are expected in regard of the estimation of runoff for the whole lake Ontario area (64000 km^2,

Figure 1) by applying the ARM to runoff simulated for the 53460 km^2 area (74% of the whole area) over which models were implemented during this work. Indeed, performances are even better when looking at the total 53460 km^2 GRIP-O area as a whole. Moreover, even a unique GR4J model implemented over the total area results in very satisfying performances, especially for the PBIAS criteria, which could save a lot of time when interested in runoff for the whole lake Ontario area.

In the case of the 64000 km^2 area could simulated runoff be compared to runoff derived from the component or residual NBS estimation procedures, but only for monthly or yearly time-steps. This is however interesting for a better understanding of long-term water supply trends (The GRIP-O period can be extended for example using CaPA reanalyzes).

Finally, it is of course not argued that the performances obtained during this work consist in the best performances that are possible to obtain for the area studied. Numerous possibilities remain to be tested to further improve the hydrologic simulations, for example by calibrating the LBRM model with the conventional Nash-Sutcliffe values (the Nash sqrt criteria seems well suited for GR4J calibrations) or using multi-objective functions that would focus at the same time on the streamflow and snow water equivalent simulations.

Considering the promising results obtained during this work with global lumped conceptual models, the second phase of the project will focus on the implementation of a distributed, physically-based model developed at EC and named GEM-Hydro. Distributed models are interesting because they offer more possible applications than lumped models, by representing in details the physical processes occurring everywhere inside the watershed. The area of Lake Ontario is thus relevant to distributed modeling, as detailed datasets exist for LULC and soil texture. However, it is quite clear that the performances of the global models will be hard to reach, but this was expected and these lumped models could moreover generally be seen as a mean to set the hydrological simulation target for more sophisticated models.

8- GRIP-O experiments with distributed models

Despite experiments with the lumped models (section 7) highlighted that better results were generally obtained in calibration with the GHNCN-D data than with CaPA, the latter revealed very satisfactory and even often led to better robustness than with GHCN-D precipitation. Moreover, the superiority of the GHCN-D over CaPA identified during GRIP-O is expected to be due to the high density of the GHCN-D network in this area, while its density is very scarce in most areas of Canada, and to the fact that the version of CaPA used in this work does not make use of rainfall fields observed by radars. It was however preferred to use CaPA in most of the experiments with distributed models as it is expected to represent the dataset which will generally be used to drive hydrologic models in Canada. Therefore, unless otherwise specified, the precipitation source used in the sub-sections below consists of the CaPA dataset. Only section 8.1 makes use of model precipitation in GEM-Hydro (see **Table 29** for the specificities of the different SVS experiments and the sections where associated results are described).

GEM-Hydro (formerly SPS) is used at EC to produce operational lake-level and streamflow forecasts in the Great Lakes domain. It currently relies on ISBA for the LSS, but SVS is planned to replace it in the near future as it is a more sophisticated surface scheme (see section 2). It was used during GRIP-O with SVS, however, unless specified differently. Section 8.3 describes a SVS improvement in regard of the snowmelt process, and shows the effect in terms of resulting streamflow performances. The comparison between this improved SVS version and the ISBA surface scheme is shown in section 8.4. For sections 8.3 and 8.4, GEM-Hydro was not calibrated and makes use of the "default" Watroute parameter values presented in Table 40. Indeed, calibrating GEM-Hydro with each of the SVS versions or each of the LSS (ISBA or SVS) and for several GRIP-O watersheds was beyond the scope of this paper and would only have a limited interest to the hydrologic community. Moreover, using non-calibrated parameter values for the comparisons of section 8.3 and 8.4 allows to better highlights the processes' differences, as a calibration would tend to compensate for the inherent drawbacks of a model. For the whole section 8 however, results are generally shown over the GRIP-O calibration period (2007-2011, see section 5.3) even for uncalibrated models. A one-year spin-up period (2006-2007) is always used before the period for which results are displayed and scores are computed.

Sections 8.1 and 8.2 present results from the first experiments and sensitivity tests made with GEM-Hydro and SVS (see **Table 29** for the different SVS versions) and involve calibrations of WATROUTE parameters only, while section 8.6 describes results obtained when calibrating SVS parameters, in which case the GEM-Hydro version used for calibration is called

GEM-Surf + UH (see section 8.5), because a UH is used instead of Watroute to produce streamflows (the UH parameter is also calibrated, see Table **9**). Section 8.7 compares the performances of the calibrated GEM-surf + UH to the two other distributed models considered in this study, namely WATFLOOD and SA-MESH.

Table 29: The different SVS versions used during this work along with their associated characteristics and the identification of sections containing the corresponding results.

version name	former	intermediate	new	new_2
canopy bug	present	fixed	fixed	fixed
baseflow bug	present	fixed	fixed	fixed
snow bug	present	present	partially fixed	partially fixed
soil freezing	on	off	off	off
vegetation class 26 error	present	fixed	fixed	fixed
max. soil depth	2.3m	2.3m	1.4m	1.4m
lateral flow removing excess water	on	on	off	off
snowmelt improvement	off	off	off	on
urban runoff improvement	off	off	off	on
precipitation source used	Model P.	Model P.	CaPA P.	CaPA P.
calibration made?	yes	no	yes	yes
calibration on (Watroute / SVS + UH), numer of free parameters	Watroute (4)	N/A	Watroute (4)	SVS + UH (16)
name of platform used	GEM-Hydro	GEM-Hydro	GEM-Hydro	GEM-Surf + UH
section where used	8.1	8.1.3	8.2	8.3 to 8.8, 9

8.1) WATROUTE first calibrations (former SPS model, model precipitation)

The first set of calibrations performed with WATROUTE made use of what is referred here as the "former" SPS model fed with model precipitation (Table 30), with SPS implemented at a 2 arcmin. spatial resolution (see section 6), using a 10 min. internal time-step, and with WATROUTE at a 30 arcsec. resolution and fixed parameter values over the whole sub-basin, except the MNDR parameter which varies spatially (see section 6.2.2) but was not calibrated (Table **9**). The "former" version of the SPS+SVS model actually consists in a version containing a few bugs affecting the water balance (a significant amount of water was lost from the system, see results below). These bugs were not known when the model was implemented and were discovered thanks to the results presented below, after a careful investigation of some of the model components (see section 8.1.2). Section 8.1.3 below contains more details about these bugs and their effect on the model water balance. This "former" SPS+SVS version which was used to produce the results of section 8.1 below thus involves a fixed maximum soil depth of 2.3m (which was due to the soil dataset used), which seems to be an overestimated value when considering the results obtained and an estimated map of the actual maximum soil depth over the Lake Ontario watershed (see text below). Finally, in the former SPS version used during the first calibrations, some soil freezing was simulated during winter, which had the effect of making the third soil layer become almost impervious during winter. See **Table 29** for a synthesis of the differences between the former (used in this 8.1 section) and new (used in section 8.2) SVS versions.

8.1.1 - calibration and results

The WATROUTE calibrations, as stated in section 5, were performed using the DDS calibration algorithm which was included in the Ostrich (Matott 2005) calibration platform. The platform allows to couple WATROUTE to the DDS algorithm in a quite straightforward way. A maximum of 200 model runs (generally more than two days) was used during the calibrations, and Table 30 below gives the parameter initial values and ranges. The initial values correspond to values obtained after manual tuning of WATROUTE over the whole lake Ontario watershed and by looking at several flow station simulations at the same time.

Table 30: WATROUTE parameter ranges and initial values used in calibration; the transformation indicates the type of transformation the parameter was submitted to before calibrating it. Manning 1, 2: Manning coefficients for the overbank flow and the main channel; Flzcoef, Power: multiplicative and power coefficients used in the recession curve equation.

	initial	minimum	maximum	transformation
Manning 1	0.15	0.1	9	none
Manning 2	0.25	0.02	2.5	none
Flzcoef	1.00E-06	1.00E-08	1.00E-04	log10
Power	2.2	0.5	5	none

The very high maximum values for the Manning coefficients are due to the fact that WATROUTE, in order to simulate the numerous lakes' effect on the streamflow hydrographs, has to increase these Manning values, and also to the sometimes unrealistic model river length, as explained in 6.2.2.

Table 31 below shows the final values obtained after calibration over the GRIP-O sub-basins using SPS+SVS+WATROUTE schemes, and the model precipitation.

Table 31: WATROUTE Parameter values obtained after calibration with model precipitation; see Table 8 for parameter definition.

Subbasin	Station	Area(km^2)	Manning 1	Manning 2	Flzcoef	Power
1	20_mile	307	6.60E-02	4.99E-01	1.31E-06	1.90E+00
3	Genessee	6317	1.34E+00	1.55E-01	1.00E-06	2.12E+00
4bis	Irondequoit	326	5.60E+00	6.09E-01	9.60E-07	2.10E+00
5	Oswego	13287	3.50E-01	1.11E-01	4.60E-07	2.22E+00
6	N/A	2406	4.97E+00	2.41E-01	2.00E-07	2.20E+00
7	Black river	4847	7.96E+00	6.37E-02	2.52E-08	2.22E+00
8	Oswegatchie	2543	1.89E+00	7.00E-02	1.02E-05	1.50E+00
10	Salmon_CA	912	7.18E-01	2.58E-01	2.09E-06	2.00E+00
10bis	N/A	944	7.52E+00	9.83E-01	2.44E-06	2.06E+00
11	Moira	2582	3.81E-01	2.54E-01	8.57E-07	2.05E+00
12	N/A	12515.5	5.00E+00	2.11E-01	4.54E-07	2.25E+00
13	N/A	1537.5	4.38E+00	7.60E-01	9.54E-06	1.64E+00
14	N/A	2689.4	4.99E+00	2.22E+00	6.31E-06	1.65E+00
15	N/A	2245.8	7.88E+00	1.58E+00	1.31E-06	2.12E+00

Figure 11 to Figure 14 below present some hydrographs and cumulative hydrographs for the Oswego and Moira sub-basins, over the calibration period, after Watroute calibration with model precip. Figure 15 shows the Hydrographs in validation for the Oswego sub-basin, because a significant drop of performance is observed between the calibration and validation periods, for this basin as for others (see Figure 13 and Figure 15), when using SPS + WATROUTE. It seems like the last winter of the validation period was difficult for SPS + WATROUTE (i.e. GEM-Hydro) simulations. When looking at the different hydrographs, it looks also like SPS is not generating enough runoff during snowmelt events, and that some of the peaks which occur during summer in the simulated flows but do not exist in the observed ones may be due to regulation, which GR4J was quite able to handle, whereas WATROUTE does not seem to be able to. This is logical however that WATROUTE has less flexibility than GR4J in terms of "time-series fitting", i.e. less flexibility in its capacity to adjust the simulated hydrographs with its four free parameters, which in WATROUTE only control the base flow and peak flows' shape, but cannot

do more than that; GR4J, on the opposite, can create a virtual routing reservoir to mimic a regulation similar to the behavior of a weir at a lake's outlet. Moreover, GR4J has an "exchange parameter", which allows to remove or add water to the routing reservoir and runoff hydrograph. To have the same flexibility than with GR4J, SPS would have to be calibrated (for example to adjust the soil depth and hydraulic properties), which is beyond the current capacities of this modeling platform given the significant computational time it requires, especially when implemented at a 2 arcmin. resolution. Finally, it has also to be reminded that GR4J was here not using the same precipitation and temperature inputs as SPS. SPS forcings come, for this first set of calibrations, from the RDPS. The second set of calibrations (section 8.2) will make use of CaPA precipitation, which will make the comparison with GR4J actually possible in that case.

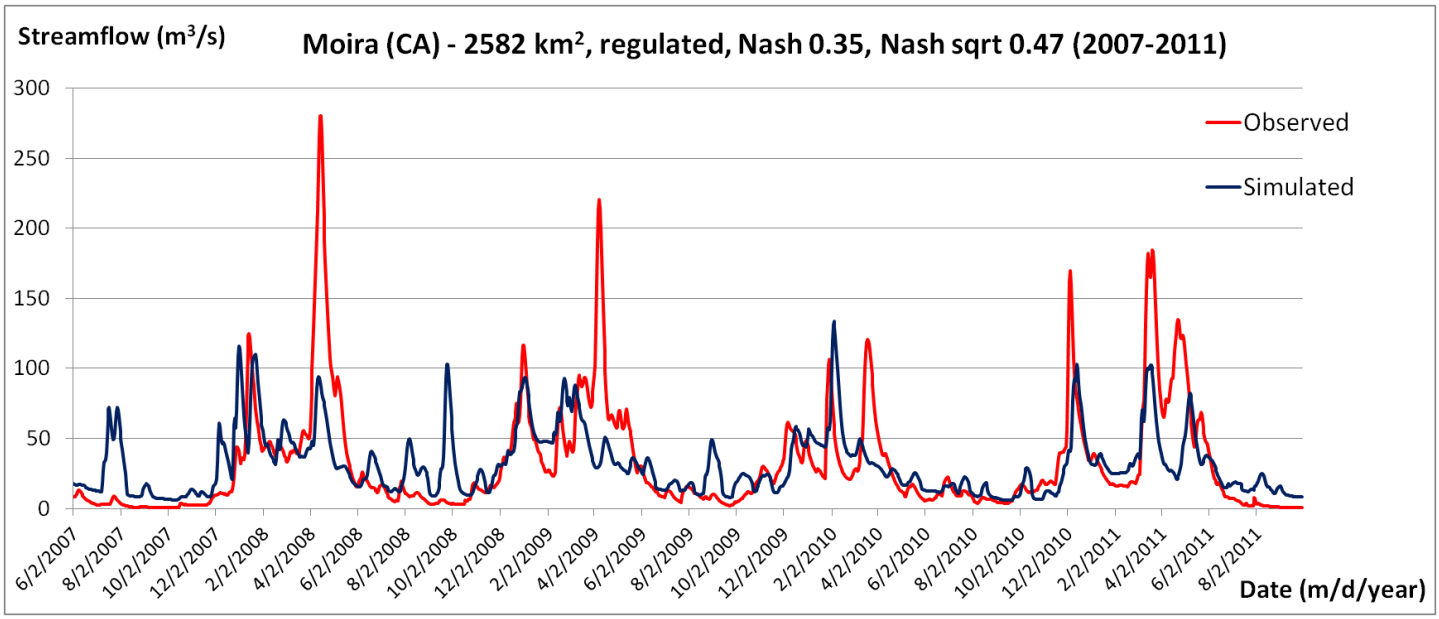

Figure 11: WATROUTE hydrographs for the Moira watershed over calibration period, with the best parameter set. Results obtained with SVS "former" version (Table 29) and model precip.

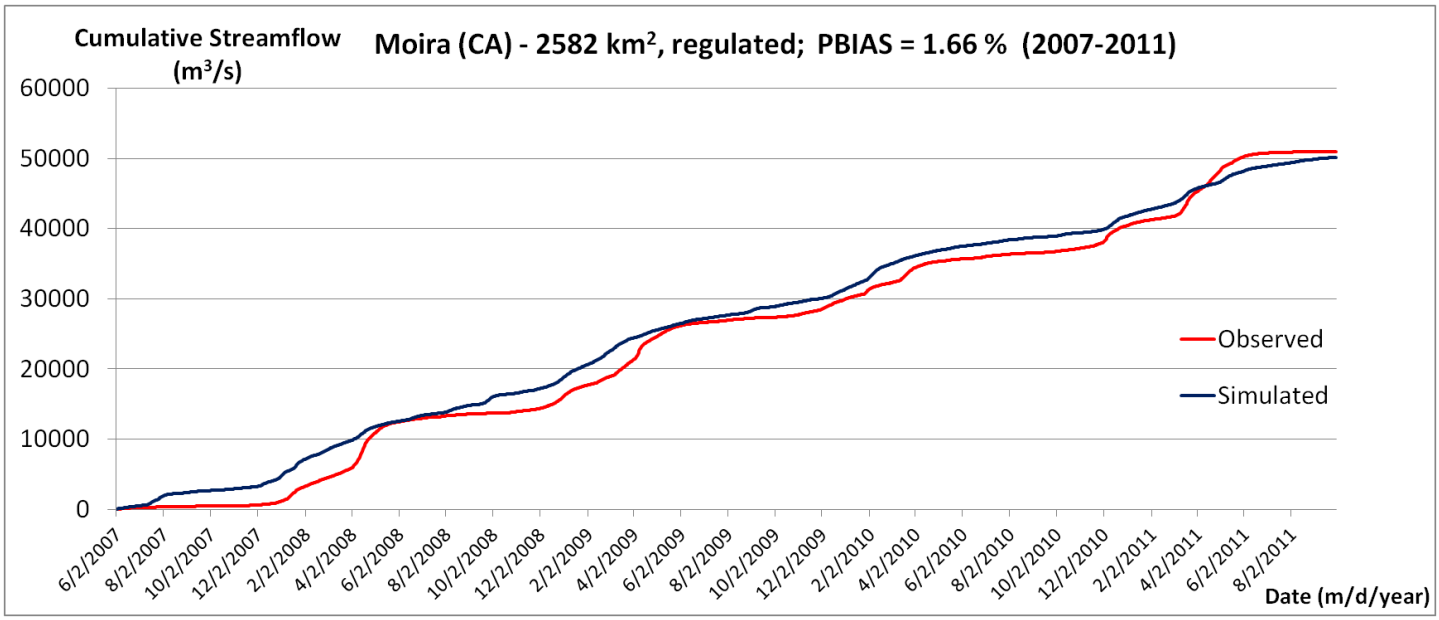

Figure 12: Cumulative streamflows in calibration obtained after Watroute calibration, over the Moira watershed. Results obtained with SVS "former" version (Table 29) and model precip.

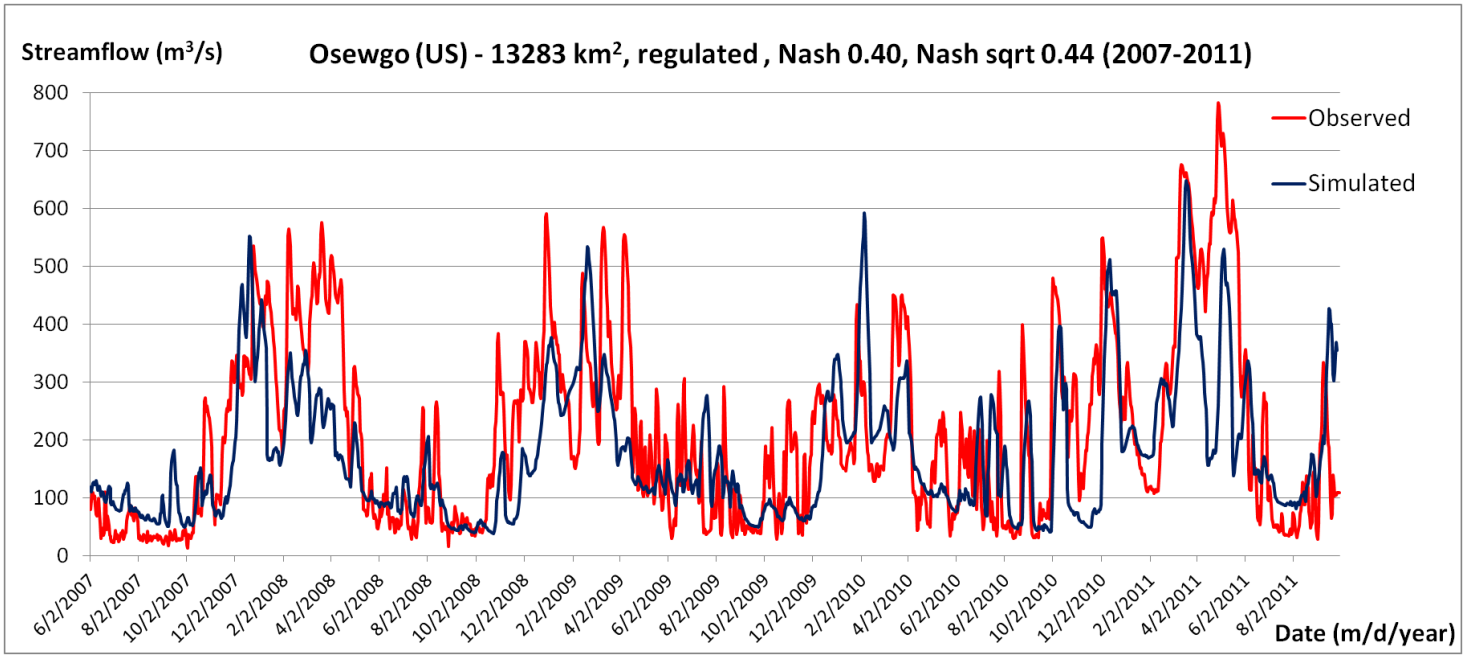

Figure 13: WATROUTE hydrographs for the Oswego watershed over calibration period, with the best parameter set. Results obtained with SVS "former" version (Table 29) and model precip.

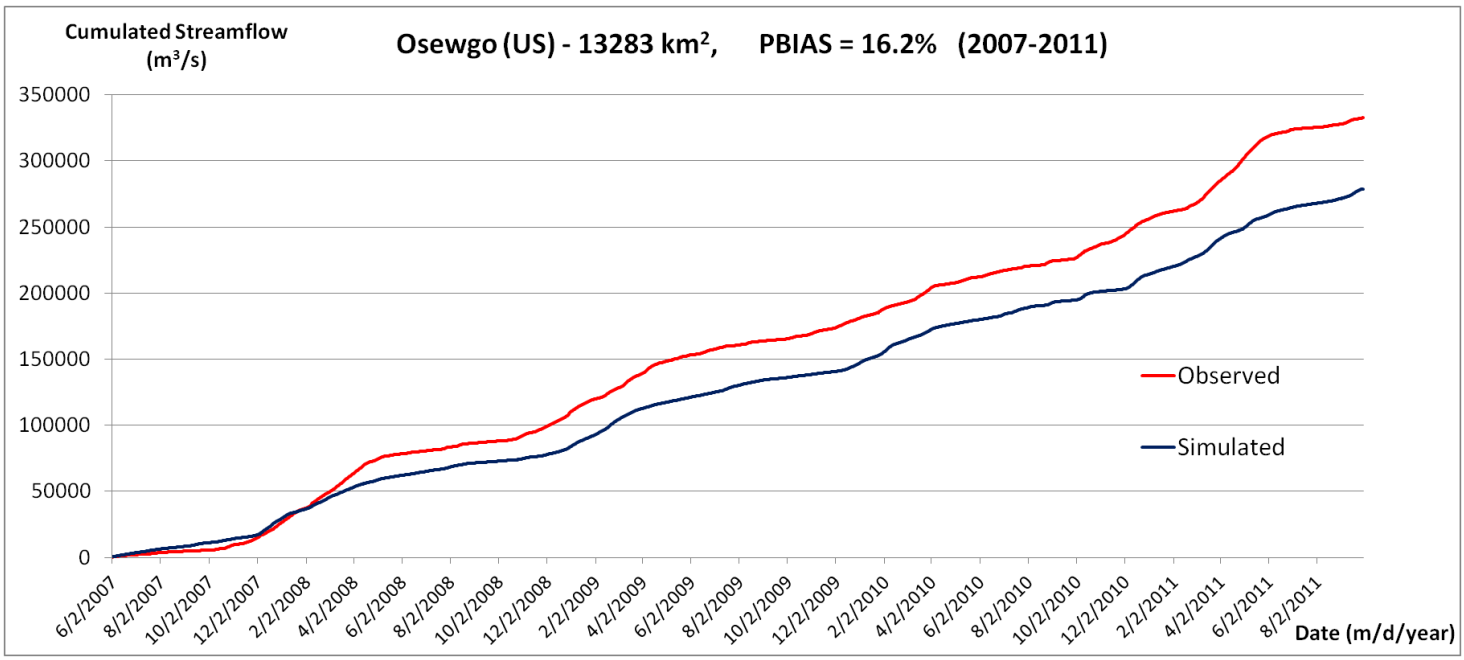

Figure 14: Cumulative streamflows in calibration obtained after Watroute calibration, over the Oswego watershed. Results obtained with SVS "former" version (Table 29) and model precip.

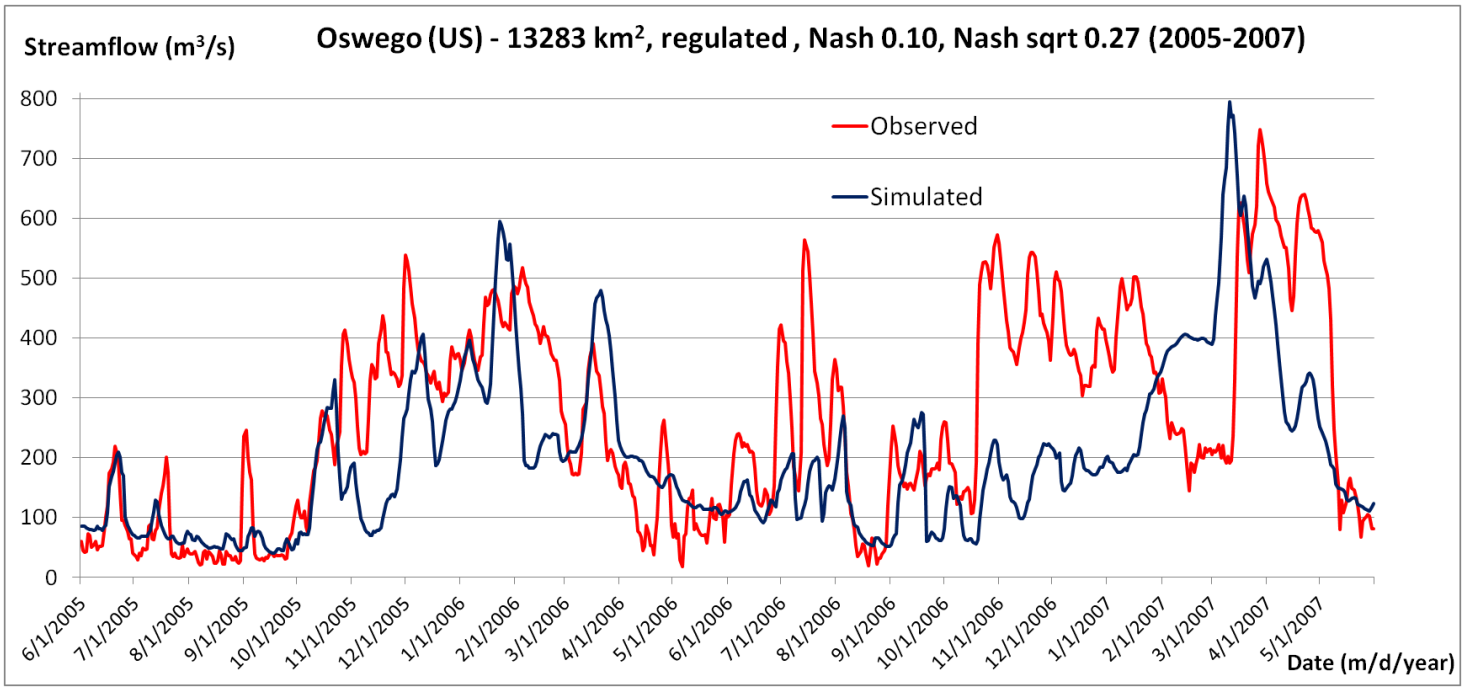

Figure 15: WATROUTE hydrographs for the Oswego watershed over validation period, with the best parameter set. Results obtained with SVS "former" version (Table 29) and model precip.

One can see on Figure 14 that SPS tends to underestimate (see Table 34) the fraction of precipitation that will eventually become streamflow, for most of the sub-basins (the Moira sub-basin on Figure 11 is actually an exception). This is attributed to SPS because WATROUTE does not accumulate water in its LZS on a pluri-annual basis and has a perfect water balance, i.e. does not loose water. This was envisioned to be possibly due to an overestimation of the evapotranspiration component, bugs in the model, etc. This issue was investigated, and a first test was made in order to assess the snowpack's Snow Water Equivalent simulation quality (following section).

8.1.2- Evaluation of SPS snowpack simulation

When looking at the streamflow simulations resulting from SPS + WATROUTE (see Figure 11 to Figure 15), it is clear that at least the peak flow events occurring during snowmelt periods are underestimated. Therefore, a comparison between SPS simulated SWE (the liquid water quantity in mm that the whole snowpack represents, including snow and liquid water of the snowpack), GR4J SWE, and the SWE emanating from the SNODAS observations (NOHRSC 2004) was made in order to investigate potential simulation errors related to the snow processes (Figure 16 below).

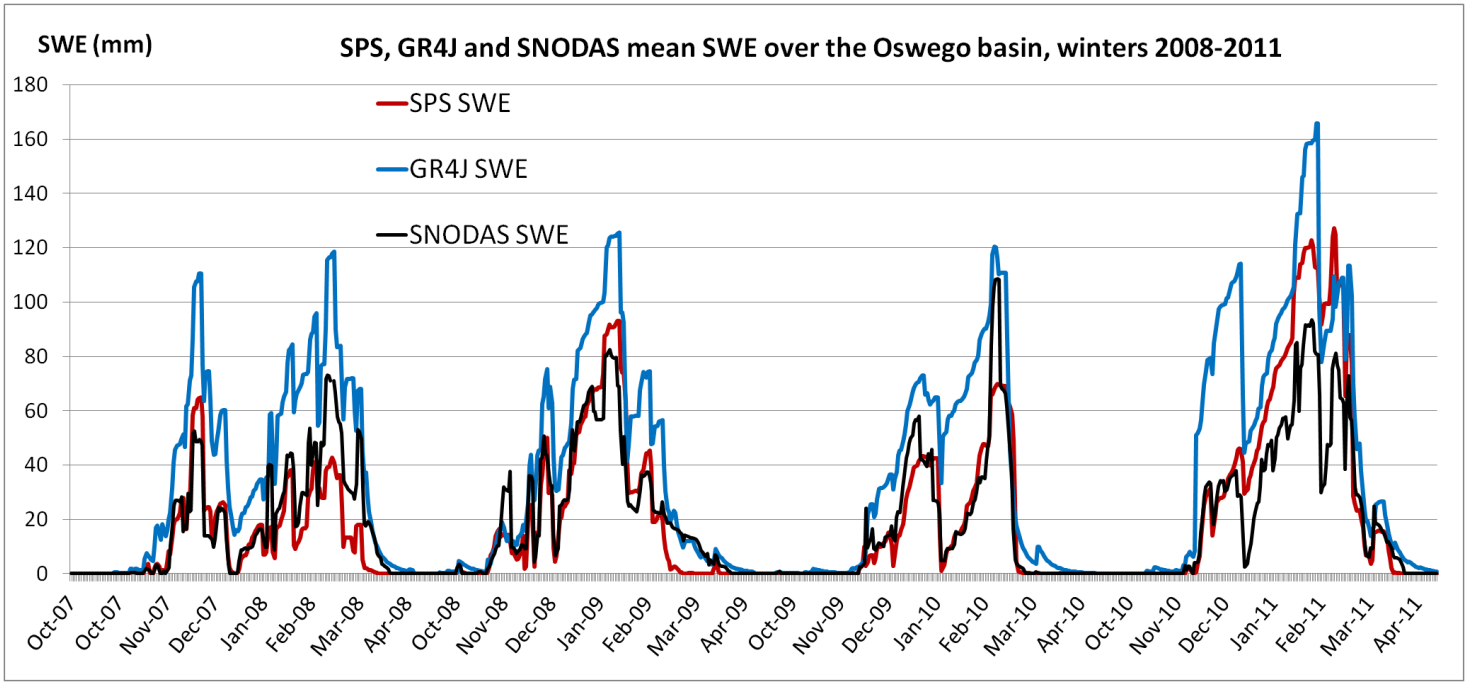

Figure 16: SPS+SVS, GR4J and SNODAS (observations) mean SWE over the Oswego basin. GR4J was fed with CaPA precipitation while SPS used model precipitation.

It seems however, when looking at Figure 16, that the SPS model does a good job in simulating the snowpack's SWE, and a better one than the GR4J model (which does not simulate sublimation from the snowpack). Note that the two models do not make use of the same precipitation or temperature forcings: GR4J relies on CaPA P. and GHCND T. while SPS uses the model 40-m temperature forcings to compute its own 2-m temperature, and model precip.

A similar trend was observed in the SWE time-series over the Genessee basin; only U.S. basins are covered by the SNODAS data.

The main conclusion that is derived from the above Figure 16 is that it is not the snowpack's SWE simulation which is responsible for the underestimated streamflow peaks in SPS, during snowmelt periods (see Figure 11 for example).

8.1.3- Water balance, bug fixes, and model improvements

The fact that the snowmelt peak flows are generally underestimated by the model, with this first implementation settings (Figure 11, Figure 13 and Figure 15), was supposed to be attributed to an overestimated soil depth (currently fixed to 2.3 m everywhere over the domain), in conjunction to the fact that soil freezing was not properly addressed in the model.

As a consequence, a first test was performed by forcing all of the snowmelt runoff to leave the model as surface runoff, instead of having the possibility to infiltrate, and by deactivating the unreliable simulation of soil freezing during winter. This test did not lead to promising results at all (not shown), because some huge peak flows did occur in this case during Winter, whereas none could be seen in the observed data at all. Moreover, the baseflow was in this case strongly underestimated during summer seasons, because the soil water storage was then deprived of its snowmelt recharge.

However, none of these model settings (soil depth, soil freezing de-activation) were thought to be responsible for the general low bias highlighted for several GRIP-O sub-basins and displayed for example on Figure 14. This low-bias phenomenon was more likely to be due to bugs in the LSS, an hypothesis which was enforced by a water balance assessment hereafter detailed. Indeed, even if the model did not put water in the right place at the right moment, which is supposed to be the case due to overestimated maximum soil depth as well as unreliable soil freezing processes, water should be conserved in the model: it should be released sooner or later in the river as part of the baseflow component. In other words, when assessed over several years, the cumulated streamflows as those of Figure **12** should somehow match the observed ones, unless a strong bias does exist in the precipitation forcings. But the thing is that Figure **12** was produced with using model P., however the RDPS precip. is known

for having a positive bias (i.e., precipitation is generally overestimated by the model), which cannot explain the negative bias of Figure 14.

As a consequence, the SPS model water balance was assessed. This water balance is based on the following requirement for any physically-based model: water should be conserved (none created or removed), i.e. the sum of the water inputs should be equal to the sum of water outputs, minus the change in storage inside of the model, which can be described by the following equation:

$P - E - R - BF = \Delta S$, with: Eq. (2)

- P: sum of precipitation inputs over simulation period (mm)
- E: sum of evaporation fluxes during simulation (mm)
- R: sum of surface runoff plus lateral flow from the 7 soil layers over simulation (mm)
- BF: sum of baseflow during simulation (mm)
- ΔS = final volume of water stored in the model at the end of the simulation period, minus the initial volume at the beginning of the simulation period (mm).

The water balance is computed as BALANCE (mm) = P - E - R - BF - ΔS (Eq. 3). A positive value indicates that some water is lost from the model, while a negative value indicates that water is created during simulation. None of these two cases should occur, however it can be hard to distinguish one case from the other, as they compensate for each other. Figure 17 below presents the water balance which was achieved when using the former version of the model (i.e. max. depth of 2.3m, soil freezing occurring, but spatial resolution of 10 arcmin.). As can be seen on the picture, most of Lake Ontario watershed depicts positive water balance anomalies of the order of 100mm, meaning that on average, the same amount of water is lost from the system during only one summer. In this area, this can represent up to 20% of the total input precipitation which is lost from the system! When performing this type of assessment with distributed physical models however, one must carefully check that the water balance computation is error-free, and that all of internal water reservoirs were taken into account, as well as all output water fluxes. Things are a bit more complicated when several land-use types are present inside the same grid cell (high/low vegetation, glaciers, sea, etc.), because one has to make sure using the appropriate weights for each land-use class. In the scope of this water balance exercise, only the pixels covered at 100% with land were assessed. This means that one should not pay attention to the water balance anomalies inside of lakes, on Figure 17 to Figure 20 below. With the version of SPS used during the GRIPO project however (and whether be it the former or new version presented in this report, see Table **29**), the subgrid-scale lakes were filtered out (no one is considered) in order to avoid overestimating evapo-transpiration during simulation. This actually means that the water balance, as computed during this balance

exercise, is valid over most of the pixels displayed on Figure 17 to Figure 20 below, except the pixels located inside of the great lakes.

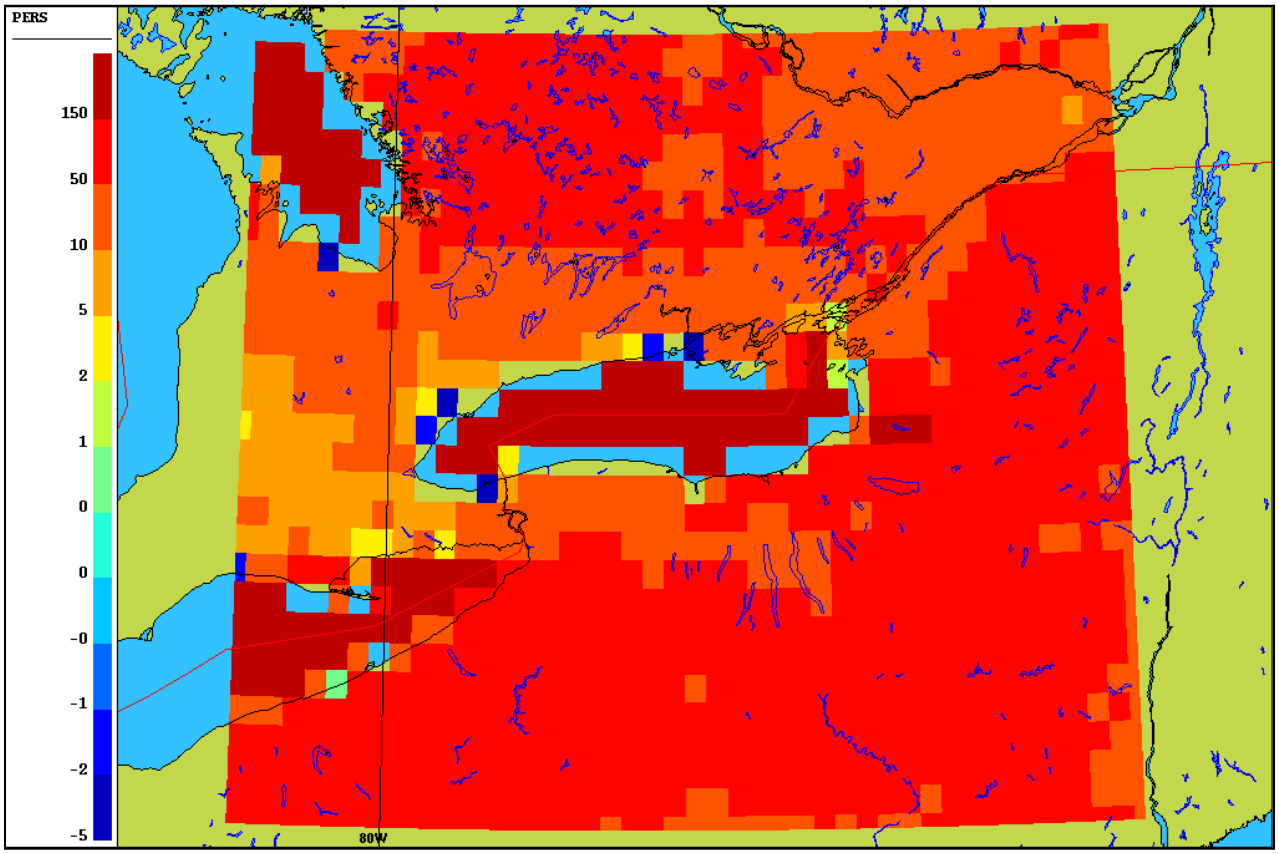

Figure 17: SPS+SVS water balance (PERS, mm) with the former version of the model (same as the one used in section 8.1 but with a spatial resolution of 10 arcmin.). This balance was computed over the period spanning from May to October 2010 and is valid for land pixels.

After many investigations to find the source of these balance anomalies, it was noticed, when printing the water stored in the vegetation canopy for one land pixel during a simulation, that the water stored in the canopy reservoir was not conserved from simulation day to the other (it was reset to 0). A maximum of 1 mm of water could hence be lost this way for each pixel and each rainy day of the simulation period. After having fixed this bug, the water balance was again assessed. It then turned out that some significant positive anomalies were remaining over the same period as the one used in Figure 17 above. This time, it was when confronting maps of the water balance anomalies to maps of the cumulated baseflow over the same period that the bug could be found, because the anomalies were closely related to the amount of water released as baseflow. After checking the script where baseflow is computed, it was found

that baseflow was removed twice from the last soil layer. After fixing this bug (leading to the "intermediate" SPS version, see **Table 29**), the water balance seemed almost perfect over summer (between -0.1 + 0.1mm, see Figure 18 below) but some anomaly remained during winter (Figure 19 below). After investigating the snow-related processes occurring in SPS, it could be found that some water was sometimes created during winter, because more water than actually available was allowed to escape the snowpack as snowmelt. This was also fixed, leading to the water balance anomaly of Figure 20 below, which still displays some positive anomalies of the order of a maximum of 4/5 mm lost during winter. The source of this positive anomaly has not yet been found at this time, but the overall anomaly was judged acceptable to create the "new" SPS version (see **Table 29**) currently being used for the second set of GRIPO calibrations with WATROUTE (see section 8.2).

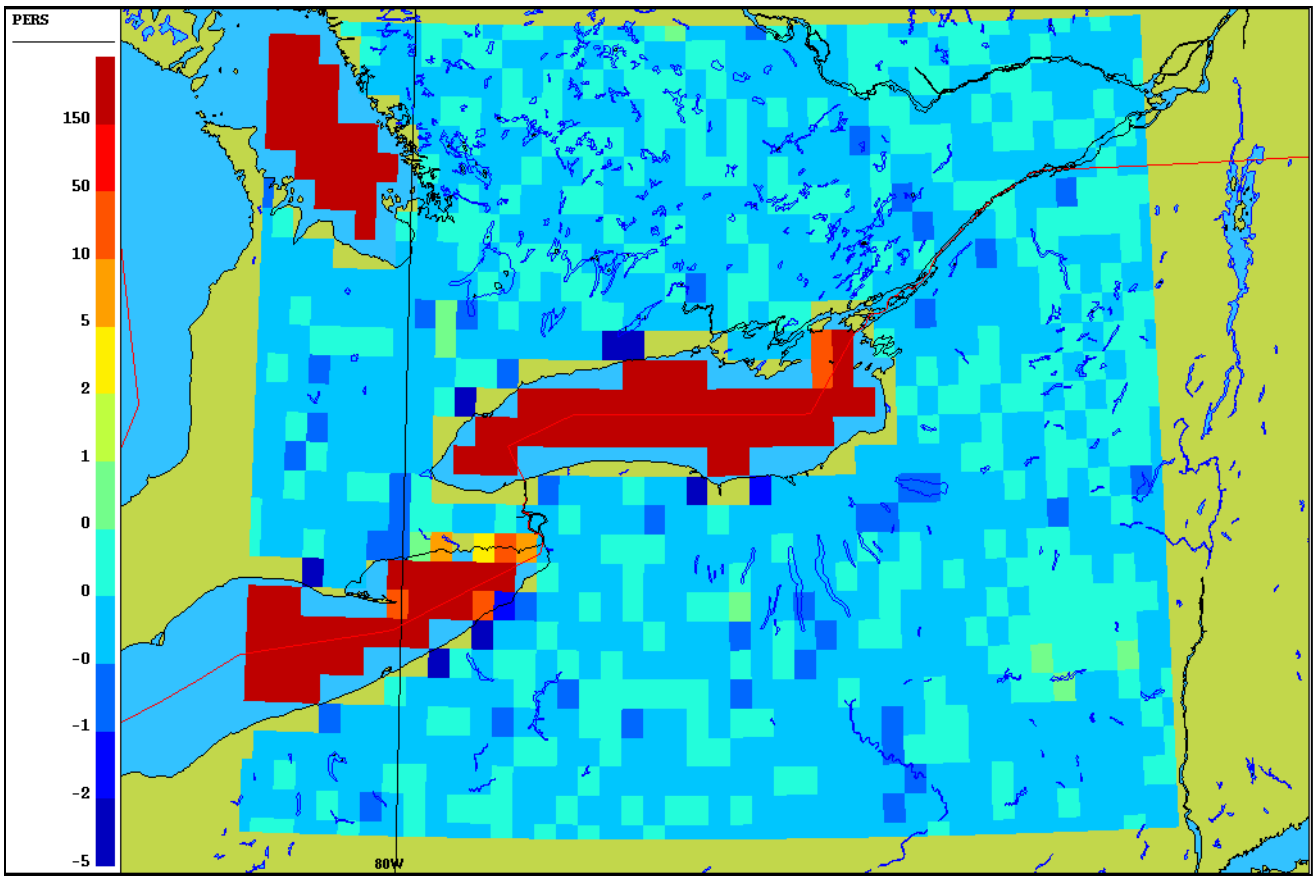

Figure 18: SPS+SVS water balance (PERS, mm) with the canopy and baseflow bugs fixed (consists in the intermediate SPS version, see Table 29), at a 10 arcmin. resolution. This balance was computed over the period spanning from May to October 2010) and is valid for land pixels.

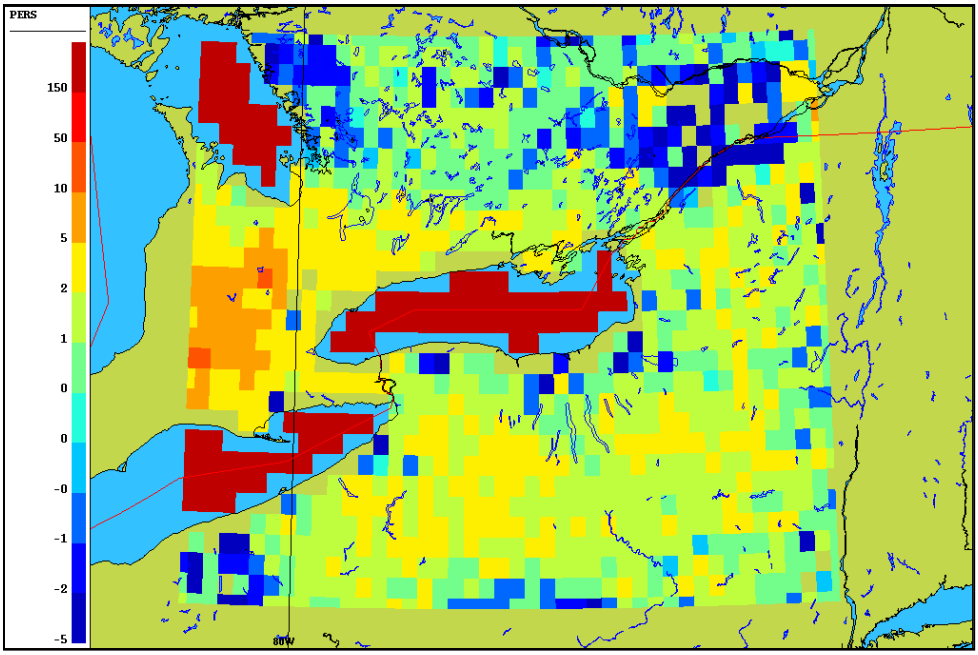

Figure 19: SPS+SVS water balance (PERS, mm) with the canopy and baseflow bugs fixed (consists in the intermediate SPS version, see Table 29), at a 10 arcmin. resolution. This balance was computed over the period spanning from November 2009 to April 2010) and is valid for land pixels.

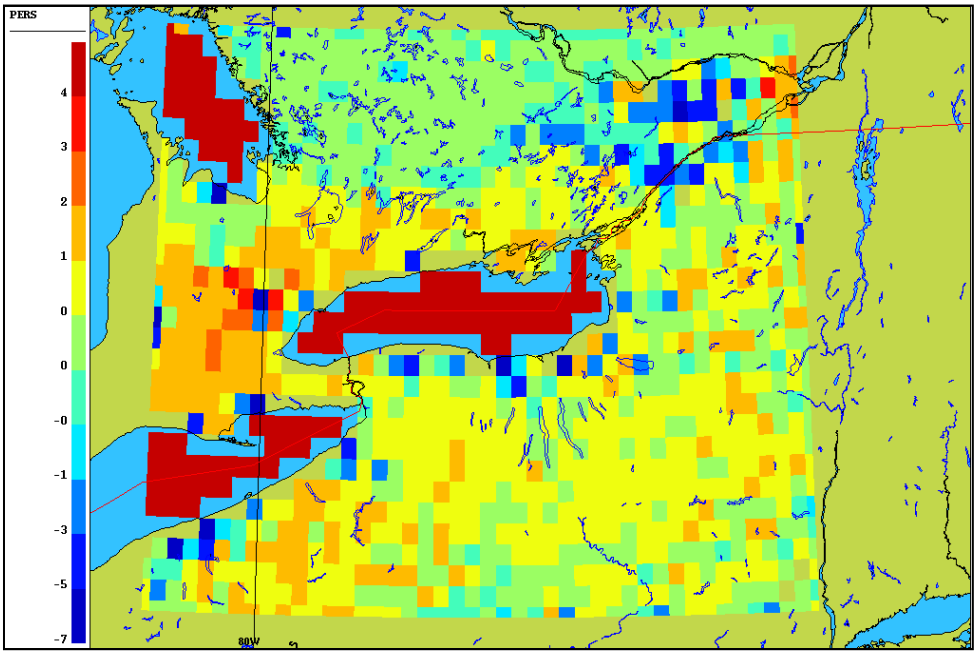

Figure 20: SPS+SVS water balance (PERS, mm) with the canopy, baseflow and snow bugs fixed (partially for snow; 10 arcmin. resolution; consists in the new SPS version). This balance was computed over the period spanning from November 2009 to April 2010) and is valid for land pixels.

Figure 21 below presents the comparison between the former and intermediate SPS versions, from the view point of the resulting hydrographs produced over one GRIPO sub-basin whose hydrograph behavior is representative of the general change, between the former and fixed SPS versions, over most of GRIPO sub-basins. See **Table 29** for a summary of the differences between the different SVS versions used during this work.

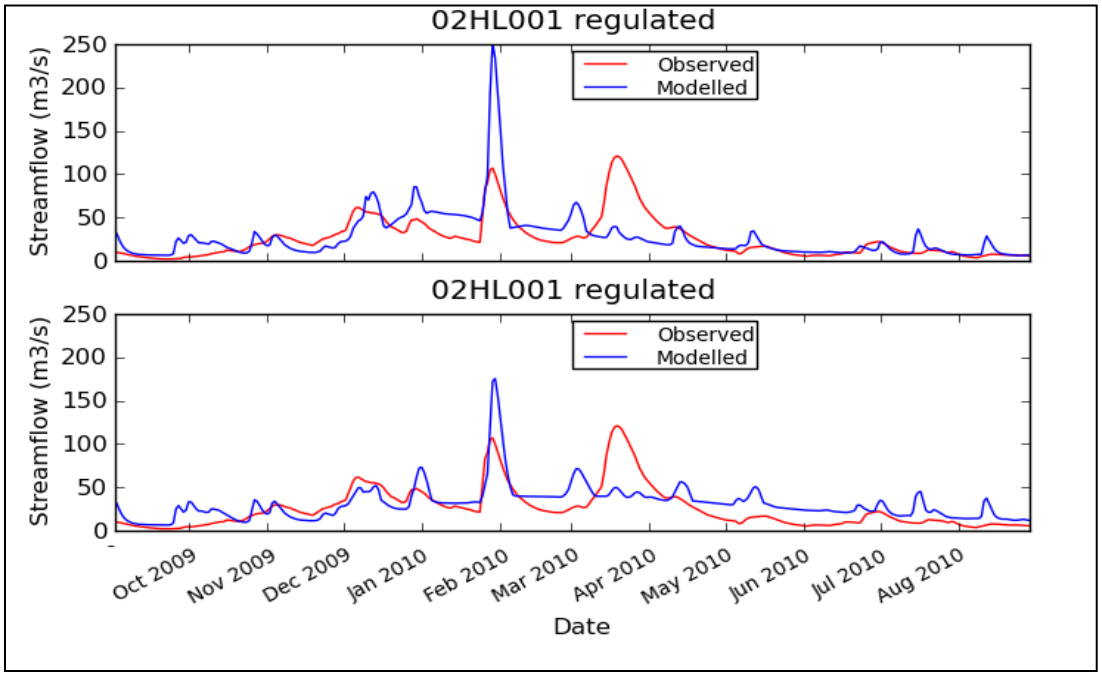

Figure 21: comparison between the former and intermediate SPS+SVS versions (implemented at a 2 arcmin. resolution). See Table 29 for details about the versions. The version differences are assessed with the resulting Moira river Hydrographs between 2009 and 2010.

This can be seen on Figure 21 above that when moving from the former to the intermediate SPS+SVS version, the winter peak flow occurring at the beginning of February 2010 decreases, and the summer baseflow increases. This makes sense, as turning the soil freezing off in the intermediate version (see **Table 29**) prevents the upper soil layers from becoming saturated during the snowmelt event of February 2010, therefore resulting in a smaller peak flow for this event, which is thus closer to the observations with the intermediate version. The baseflow increase during summer is simply due to the baseflow bug fix.

A test was then made with using the intermediate SPS+SVS version, without the soil-freezing related processes being simulated, and a maximum soil depth of either 1.4m or 2.3m. This test was promising as it had the direct consequence of increasing the simulated baseflow (see Figure 22 below) , because less water could then be stored in the thick (90 cm) former 7th

layer. In the case of fixing the maximum soil depth to 1.4m, 7 layers were still used in the model, but not with the same thicknesses as when a total depth of 2.3m was used.

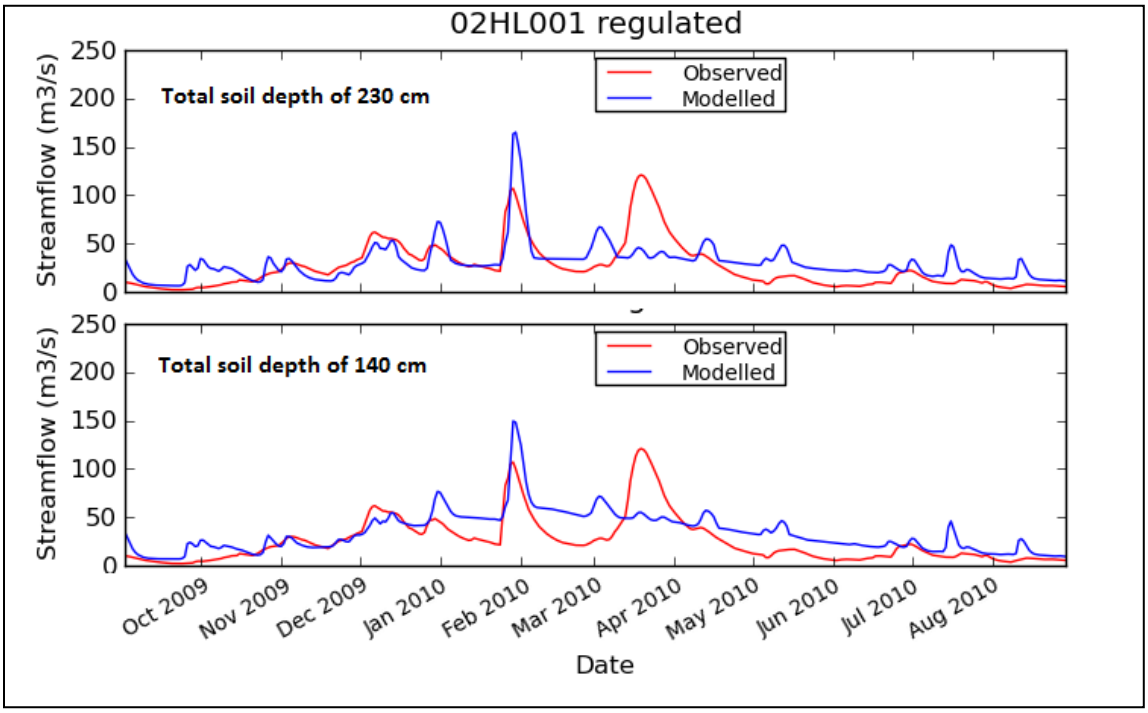

Figure 22: Example of the SPS+SVS sensitivity to the maximum soil depth, over the Moira river watershed. This figure was produced with the intermediate SPS version, a spatial resolution of 10 arcmin., and model precipitation.

However, it is clear, when looking at Figure 22 above, that baseflow is now strongly overestimated, in comparison to the hydrographs produced with the former SPS version (see top graph of Figure 21 above). This is due both to the baseflow bug fix and the smaller maximum soil depth of 1.4m. Instead of concluding that a total soil depth of 1.4m is less appropriate than a total depth of 2.3m however, it was supposed that the watroute parameters related to baseflow were not appropriate any more to the SPS outputs, because of the major model modifications brought to the model. After manually tunning the watroute baseflow-related parameters a little bit, and making them switch from 1.10^{-6} to 1.10^{-5} for the FLZ and from 2.2 to 2.8 for the PWR coefficients (see Table 30), Figure 23 and Figure 24 below were produced, which compare the hydrographs produced with the former 2 arcmin. SPS+SVS version and former watroute baseflow parameters, to the new 10 arcmin. SPS+SVS version and new watroute baseflow parameters. These figures represent the difference, from the resulting streamflow viewpoint, between the former and new SVS versions (**Table 29**), which are

respectively used in the first calibrations of section 8.1 and in the second Watroute calibrations of section 8.2. The differences come from the SVS version, the spatial resolution (which was however proven to have only a minor influence, see Figure 2), and the Watroute baseflow parameters, which were not appropriate to both SPS implementations. These figures also represent the model streamflow simulations before any automatic calibration of the four Watroute parameters. However, both SVS versions of Figure 23 and Figure 24 used model precipitation, in opposition to the runs made in the context of section 8.2 below.

It seems like all changes made to SVS in order to create the "new" version are generally beneficial, i.e. they lead to better streamflow simulations (see Figure 23 below). However, it was noticed in some cases that turning the soil freezing processes off in the new version is not always beneficial, as some peak flows occurring for example during rainy winter events are now underestimated in the new compared to the former SVS version (not shown here). It seems then needed to simulate these freezing processes, and work is under way at CMC to develop a decent representation of such processes. This is not envisioned however, to wait for such changes to be ready in order to proceed with the second set of Watroute calibrations, which will make use of the new SVS version as explained earlier in this section. This is dedicated to future work and will be reminded in the perspectives of this report (section 11).

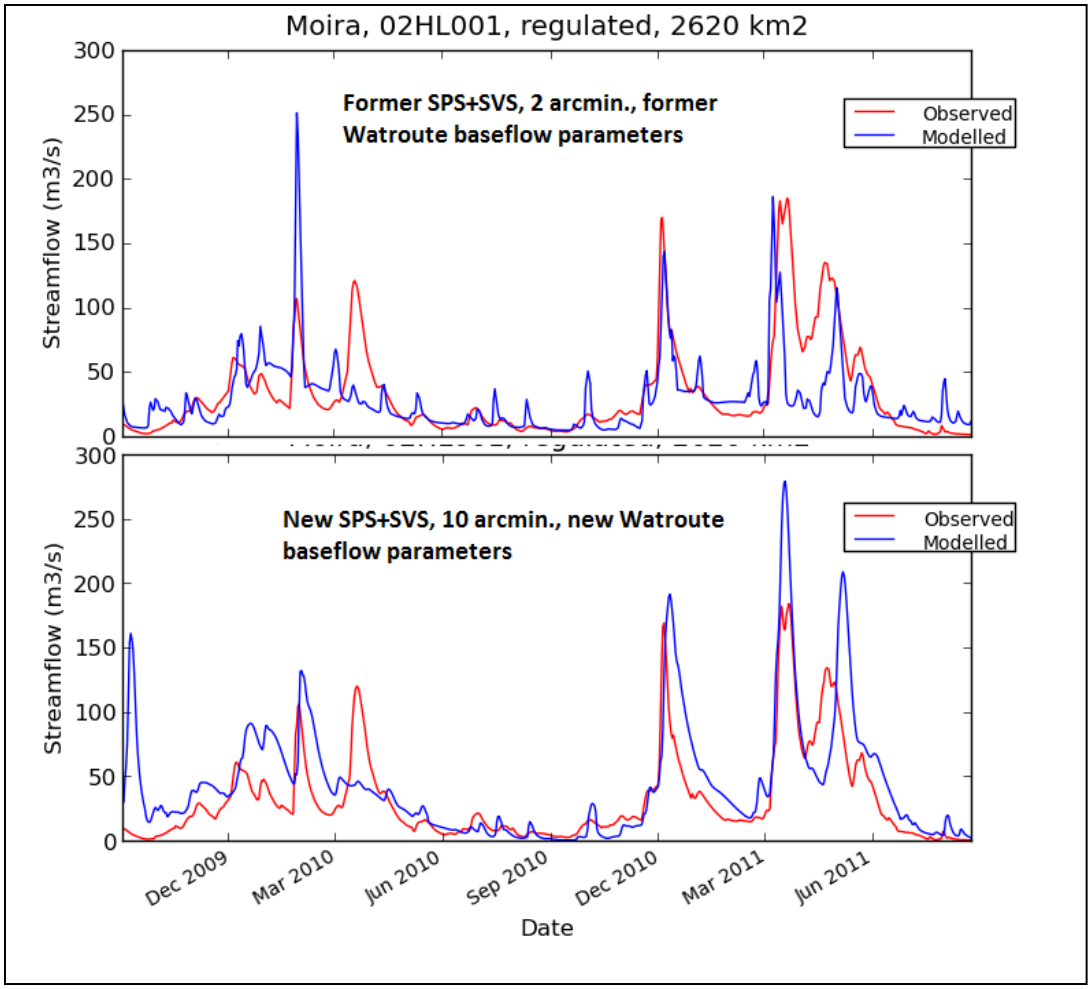

Figure 23: comparison between the former and new SPS+SVS model (they were implemented at different spatial resolutions, and the Watroute baseflow parameters which are appropriate to the model outputs are not the same, see text). Hydrographs for the Moira river watershed.

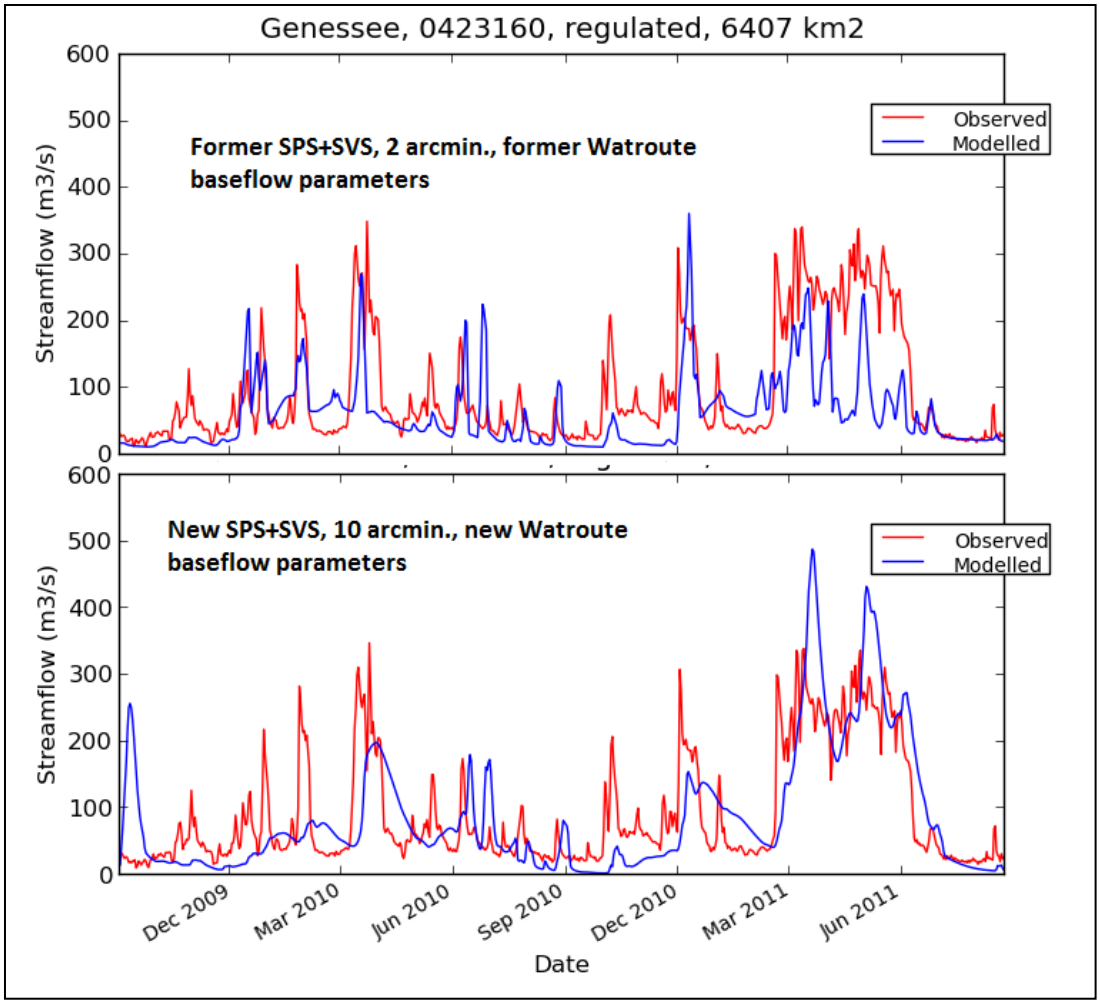

Figure 24: comparison between the former and new SPS+SVS model (they were implemented at different spatial resolutions, and the Watroute baseflow parameters which are appropriate to the model outputs are not the same, see text). Hydrographs for the Genessee river watershed.

8.2) Watroute second calibrations (new GEM-Hydro model, CAPA precipitation)

This set of second calibrations, which make use of the new SVS version as detailed in section 8.1.3 above (see **Table 29** for a synthesis of the changes) and use a 10 arcmin. spatial resolution, will most importantly use the CaPA source for precipitation inputs. As the reader will have understood at this point, it was not expected to implement so many changes to SVS during this work, so that the changes described in section 8.1.3 consist in a new SVS version.

As a consequence, it is not possible to compare the streamflows obtained in sections 8.1.1 and 8.2, because two different versions of the model, as well as two different datasets for precipitation, are used in these experiments. This therefore means that the two precipitation datasets cannot be compared in the case of GEM-Hydro. Indeed, due to the computational time required, it is not envisioned that two distinct Watroute calibrations can be performed with the new Gem-Hydro version and each of the precip. sources. This was preferred to make the run with CaPA precipitation for the new GEM-Hydro version, because this dataset is supposed to be closer to reality than model precipitation (see Table 35). One expectation with using CaPA precipitation instead of model precipitation (see Figure 23 and Figure 24 above), is that it will allow correcting at least partly for the positive bias which can be observed in streamflow simulations with the new SVS version, and which is due to the positive model precipitation bias.

Table 32: Final Watroute parameter values after the second calibrations (CaPA precipitation, new GEM-Hydro model). See **Table 30** for the range and initial values used for calibration.

Subbasin #	Station	Area(km2)	MAN2	MAN1	POW	FLZCOEF
1	20_mile	307	1.21E-01	7.04E-02	3.74E+00	7.20E-06
3	Genessee	6317	3.62E-02	7.43E+00	2.41E+00	4.32E-06
4bis	Irondequoit	326	7.00E-02	7.64E+00	1.00E+00	1.93E-04
5	Oswego	13287	0.04112	7.83750	2.43350	0.00000
6	N/A	2406	8.46E-02	2.37E-01	2.22E+00	4.70E-06
7	Black river	4847	0.04722	0.19501	1.51787	0.00007
8	Oswegatchie	2543	5.00E-02	7.74E+00	1.10E+00	3.80E-04
10	Salmon_CA	912	5.63E-02	7.97E+00	2.26E+00	4.89E-06
10bis	N/A	944	1.99E+00	4.97E-02	5.00E+00	8.79E-04
11	Moira	2582	8.06E-02	3.66E-01	2.89E+00	1.86E-07
12	N/A	12515.5	6.98E-02	2.29E+00	2.05E+00	2.38E-06
13	N/A	1537.5	4.76E-01	7.97E+00	7.46E-01	9.83E-04
14	N/A	2689.4	8.61E-01	2.98E-02	1.14E+00	4.12E-04
15	N/A	2245.8	3.01E-01	9.26E-01	1.70E+00	1.93E-05

Table 33: comparison of performances obtained with Watroute default (no_cal) or calibrated (Cal) parameter values, with the second set of experiments (CaPA precipitation, new GEM-Hydro model). Values are in percent (best is 100 for Nash values and 0 for PBIAS).

			no_cal	Cal	no_cal	Cal	no_cal	Cal	no_cal	Cal
Subbasin #	Station	Area(km2)	Nash		Nash sqrt		Nash Ln		PBIAS	
1	20_mile	307	0.2	4.0	14.2	20.0	3.5	30.0	56.2	60.0
3	Genessee	6317	20.2	32.0	3.4	23.0	-9.4	-23.0	26.9	27.0
4bis	Irondequoit	326	29.3	33.0	-21.8	37.0	10.4	36.0	12.8	21.6
5	Oswego	13287	30.6	44.0	22.5	50.0	7.7	31.0	19.9	20.0
7	Black river	4847	13.9	35.0	-1.3	41.0	-23.9	35.0	21.6	22.4
8	Oswegatchie	2543	9.2	33.0	10.9	41.0	12.5	48.0	17.3	19.2
10	Salmon_CA	912	57.5	68.0	68.4	71.0	68.0	54.0	-2.5	-2.4
11	Moira	2582	29.7	57.0	63.3	67.0	67.3	54.0	-24.6	-24.5

Table 34: comparison of performances obtained with Watroute first or second set of experiments, i.e. respectively with model precip. and old GEM-Hydro model (old_cal), or CaPA precip and new GEM-Hydro model (new_cal). See Table 29 for differences between model versions. In both cases the performances shown were obtained after the Watroute calibrations. Values are in percent (best is 100 for Nash values and 0 for PBIAS).

			old_cal	new_cal	old_cal	new_cal	old_cal	new_cal	old_cal	new_cal
Subbasin #	Station	Area(km2)	Nash		Nash sqrt		Nash Ln		PBIAS	
1	20_mile	307	22.0	10.0	18.0	24.0	1.0	21.0	18.1	53.0
3	Genessee	6317	21.0	46.0	24.0	24.0	12.0	-78.0	33.5	34.7
4bis	Irondequoit	326	11.0	23.0	12.0	27.0	-2.0	20.0	23.0	21.6
5	Oswego	13287	40.0	63.0	44.0	54.0	44.0	31.0	16.2	17.0
6	N/A	2406	17.0	41.0	25.0	40.0	29.0	18.0	31.0	16.7
7	Black river	4847	15.0	40.0	20.0	41.0	24.0	30.0	32.4	19.1
8	Oswegatchie	2543	23.0	39.0	31.0	43.0	40.0	43.0	22.9	11.0
10	Salmon_CA	912	31.0	62.0	45.0	72.0	39.0	70.0	18.1	7.6
10bis	N/A	944	25.0	52.0	46.0	63.0	42.0	56.0	28.6	29.5
11	Moira	2582	34.0	55.0	42.0	66.0	32.0	61.0	1.4	-14.9
12	N/A	12515.5	37.0	54.0	44.0	55.0	41.0	47.0	18.0	6.4
13	N/A	1537.5	6.0	26.0	-10.0	33.0	-71.0	21.0	46.0	29.3
14	N/A	2689.4	6.0	10.0	9.0	-1.0	8.0	-72.0	35.0	57.3
15	N/A	2245.8	12.0	7.0	16.0	-5.0	5.0	-74.0	36.0	53.0

Despite the new GEM-Hydro model involved the correction of bugs which were responsible for the model losing water (see section 8.1.3), it can be seen from Table 34 that the values for PBIAS are sometimes higher for the new set of calibrations than for the older one. This is due to the fact that the new set of experiments make use of CaPA precipitation, which consist of lower annual rainfall amounts than model precipitation (used in the former set of experiments). In other words, the difference between model and CaPA precipitation is sometimes strong enough (see Table 35 below) to dominate the change induced by the water balance bug fixes, which tend to decrease these PBIAS values in comparison to the first set of experiments.

Table 35: Mean annual precipitation (mm) between June 2004 and April 2011, for the different precipitation sources considered in this work and a few of GRIP-O sub-basins.

	GHCND	CaPA	Model
Oswego	1051	1094	1388
Sub-basin 12	964	953	1102
Sub-basin 8	1075	1197	1395
Sub-basin 6	1285	1411	1737
Black	1249	1305	1740
Sub-basin 15	873	869	1052

The fact that the model (i.e. new GEM-Hydro model) still underestimates runoff (positive PBIAS values in Table 34) is attributed to two main factors, such as an overestimated evapo-transpiration (which is much higher than evaporation simulated with GR4J even over summer periods, see Table 36 below) and overestimated infiltration, at least for watersheds including urban areas: currently, the SVS module does not make use of the Urban scheme developed at EC, and which is named TEB. It would hence be possible to either make SVS use TEB over urban areas, or to implement a simple correction factor for urban land-uses, that would increase runoff due to the impervious surfaces of urban areas. This was noticed when looking at sub-basins 14 and 15 especially, because these sub-basins include a high fraction of urban areas and are the ones for which runoff is generally very strongly underestimated (see Table 34 and Figure 25 below). This simple modification to better account for urban areas in SVS was brought in SVS "new_2" version (Table **29**). Results are shown in sections 8.3 to 9.

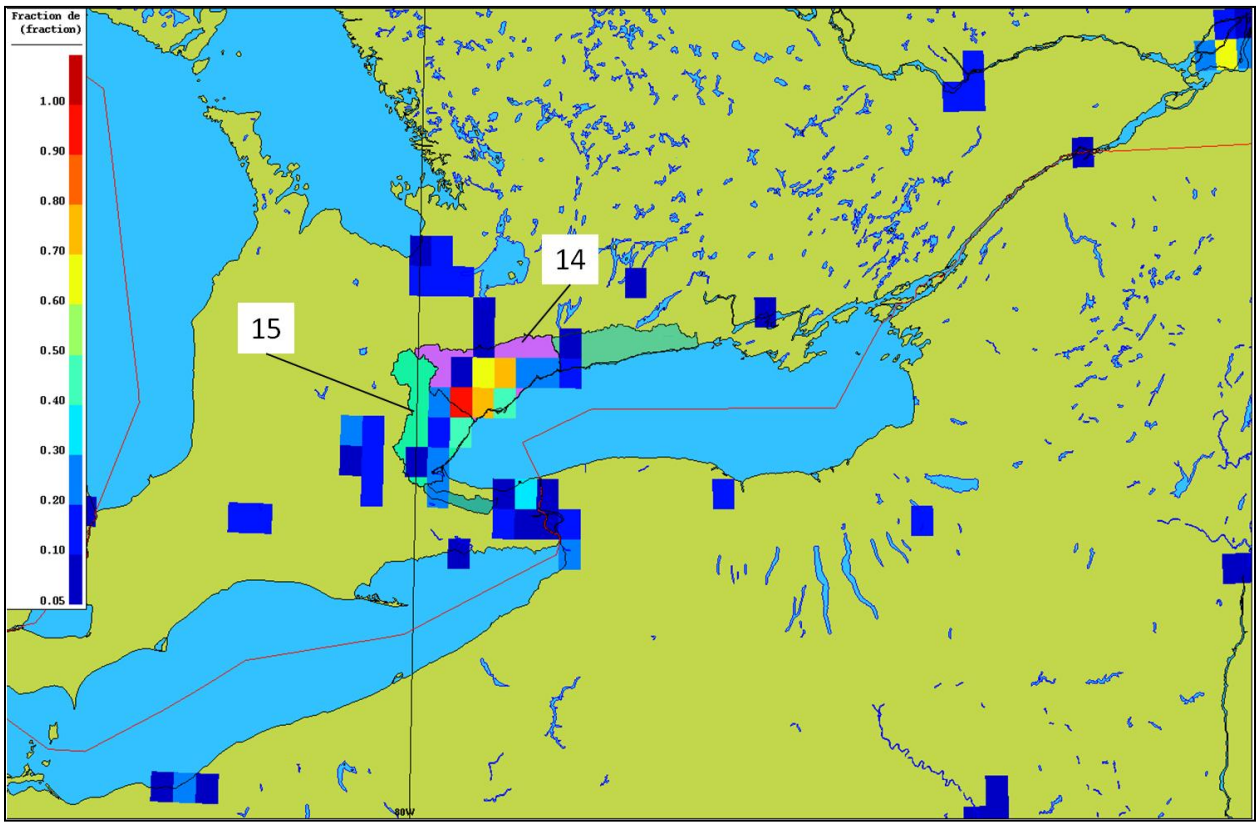

Figure 25: Fraction of the SVS pixels occupied by urban land-cover. No pixels are shown for pixels not containing any urban land-cover type.

Table 36: Mean Summer evapo-transpiration (mm) for the GR4J and GEM-Hydro models, when using CaPA precipitation. The summers (1st June - 31st October, 152 days) between 2004 and 2011 (included) were used to derive these statistics.

	GR4J	GEM-Hydro
Oswego	116	404
Sub-basin 12	114	378
Sub-basin 15	241	371
Sub-basin 14	241	361

Despite the underestimation issue, the new set of GEM-Hydro simulations lead to decent simulation results (Table 34). This is also highlighted on the two hydrographs below (Figure 26 and Figure 27). Some of the GRIP-O watersheds seem to have their outflow quite strongly influenced by artificial modifications (i.e., regulation), such as for the Oswego river (see observed flow of Figure 28).

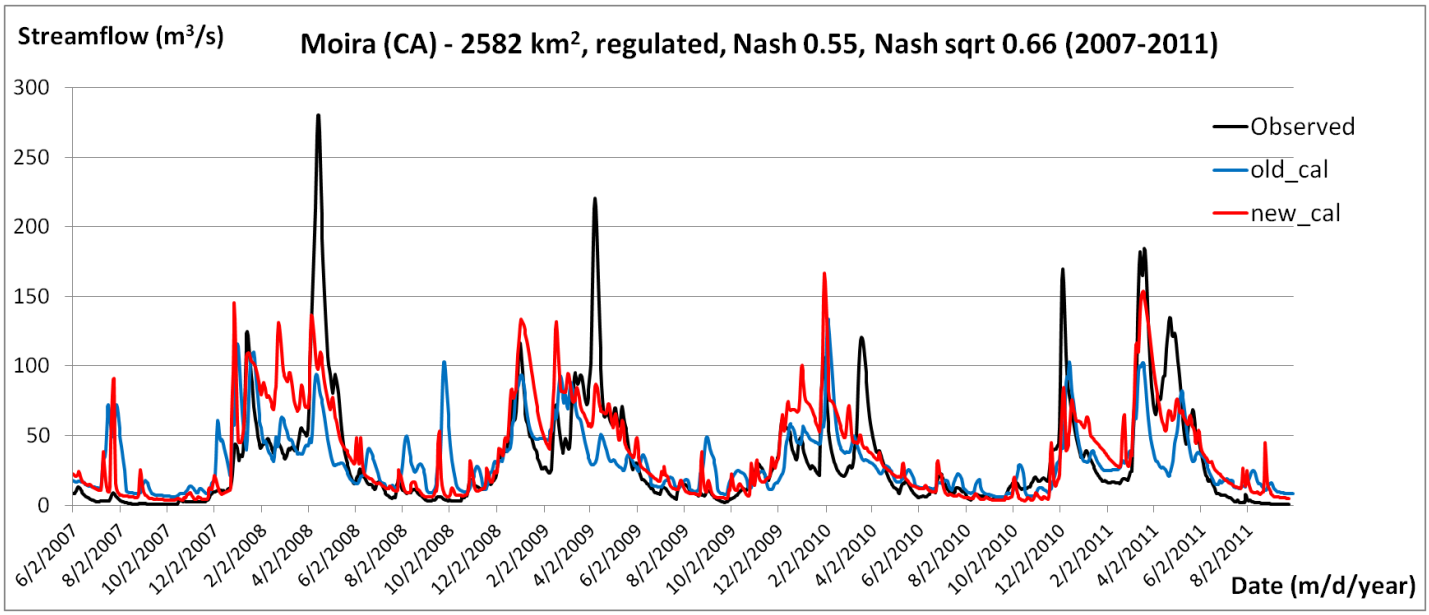

Figure 26: comparison of performances obtained with Watroute first or second set of experiments, i.e. respectively with model precip. and old GEM-Hydro model (old_cal), or CaPA precip and new GEM-Hydro model (new_cal), for the Moira watershed and the calibration period. In both cases the performances shown were obtained after the Watroute calibrations.

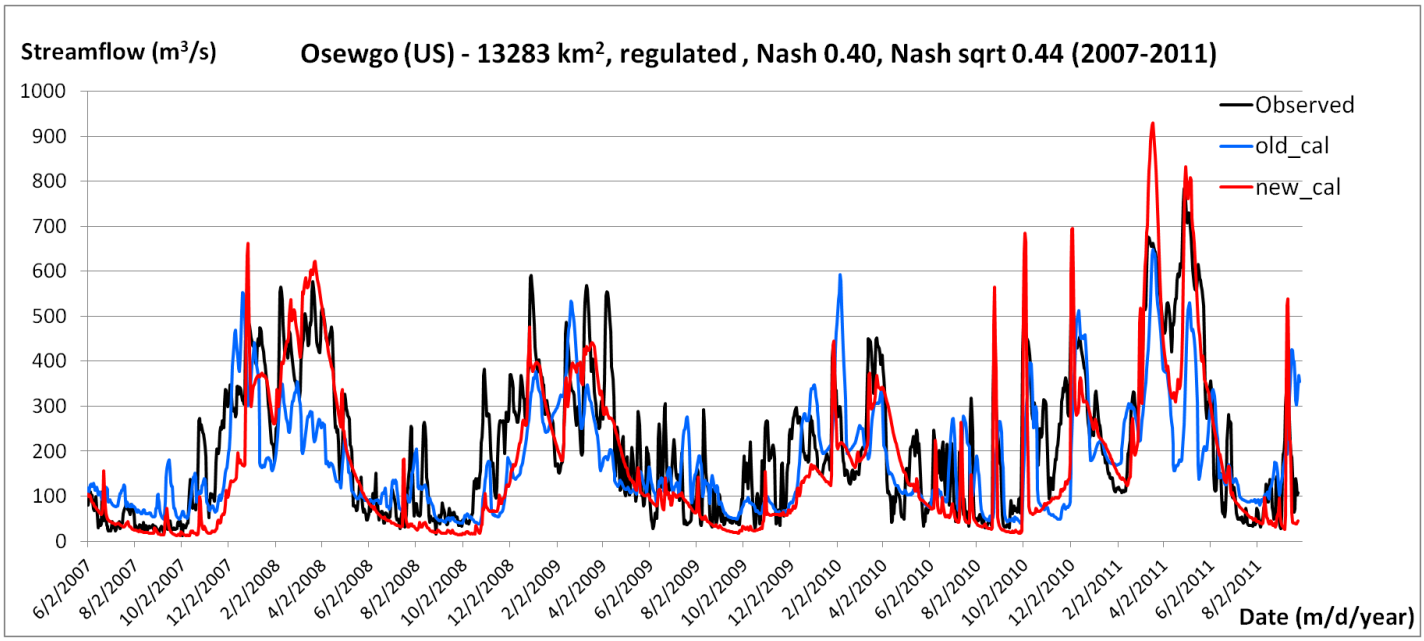

Figure 27: comparison of performances obtained with Watroute first or second set of experiments, i.e. respectively with model precip. and old GEM-Hydro model (old_cal), or CaPA precip and new GEM-Hydro model (new_cal), for the Oswego river and the calibration period. In both cases the performances shown were obtained after the Watroute calibrations.

In this case, where actual streamflow involves significant variations all along the year which are not due to natural causes, the model is not able to reproduce such behavior. As a consequence, model performances were also assessed on a weekly basis, to assess whether increasing the evaluation time-step would allow or not to partly circumvent the effect of regulation on the observed flows. Results are presented in Table 37 below, and despite it is clear that assessing the performances on a weekly basis is generally accompanied with better performances than when using a daily basis, the improvement is generally quite slight and not stronger for regulated catchments (such as Oswego, Oswegatchie, and the Black rivers), suggesting that the weekly time-step cannot allow to truly get rid of the effect of regulation on the observed values. This is also displayed on Figure 28 and Figure 29 below.

Table 37: comparison between daily and weekly performances over calibration period, for the new set of Watroute calibrations (new GEM-Hydro model, CaPA precipitation). Values are in percent (best is 100 for Nash criteria).

Subbasin #	Station	Flow regime	Area(km2)	daily Nash	weekly Nash	daily Nash sqrt	weekly Nash sqrt	daily Nash Ln	weekly Nash Ln	daily PBIAS	weekly PBIAS
1	20_mile	natural	307	10.0	9.0	24.0	23.9	21.0	23.3	53.0	51.6
3	Genessee	regulated	6317	46.0	50.4	24.0	27.1	-78.0	-75.0	34.7	34.7
4bis	Irondequoit	natural	326	23.0	21.5	27.0	24.1	20.0	19.6	21.6	21.9
5	Oswego	regulated	13287	63.0	68.8	54.0	58.5	31.0	32.9	17.0	17.6
6	N/A	mixed	2406	41.0	43.5	40.0	42.4	18.0	23.3	16.7	16.7
7	Black river	regulated	4847	40.0	40.6	41.0	41.0	30.0	29.8	19.1	19.2
8	Oswegatchie	regulated	2543	39.0	41.4	43.0	44.4	43.0	45.8	11.0	11.1
10	Salmon_CA	regulated	912	62.0	65.2	72.0	74.8	70.0	70.8	7.6	7.7
10bis	N/A	mixed	944	52.0	58.2	63.0	67.9	56.0	61.5	29.5	29.6
11	Moira	regulated	2582	55.0	57.9	66.0	67.8	61.0	61.0	-14.9	-14.7
12	N/A	regulated	12515.5	54.0	56.3	55.0	57.1	47.0	50.0	6.4	6.6
13	N/A	natural	1537.5	26.0	25.3	33.0	28.3	21.0	22.8	29.3	29.4
14	N/A	mixed	2689.4	10.0	-3.1	-1.0	-20.8	-72.0	-90.0	57.3	57.3
15	N/A	mixed	2245.8	7.0	3.1	-5.0	-10.9	-74.0	-74.2	53.0	53.1

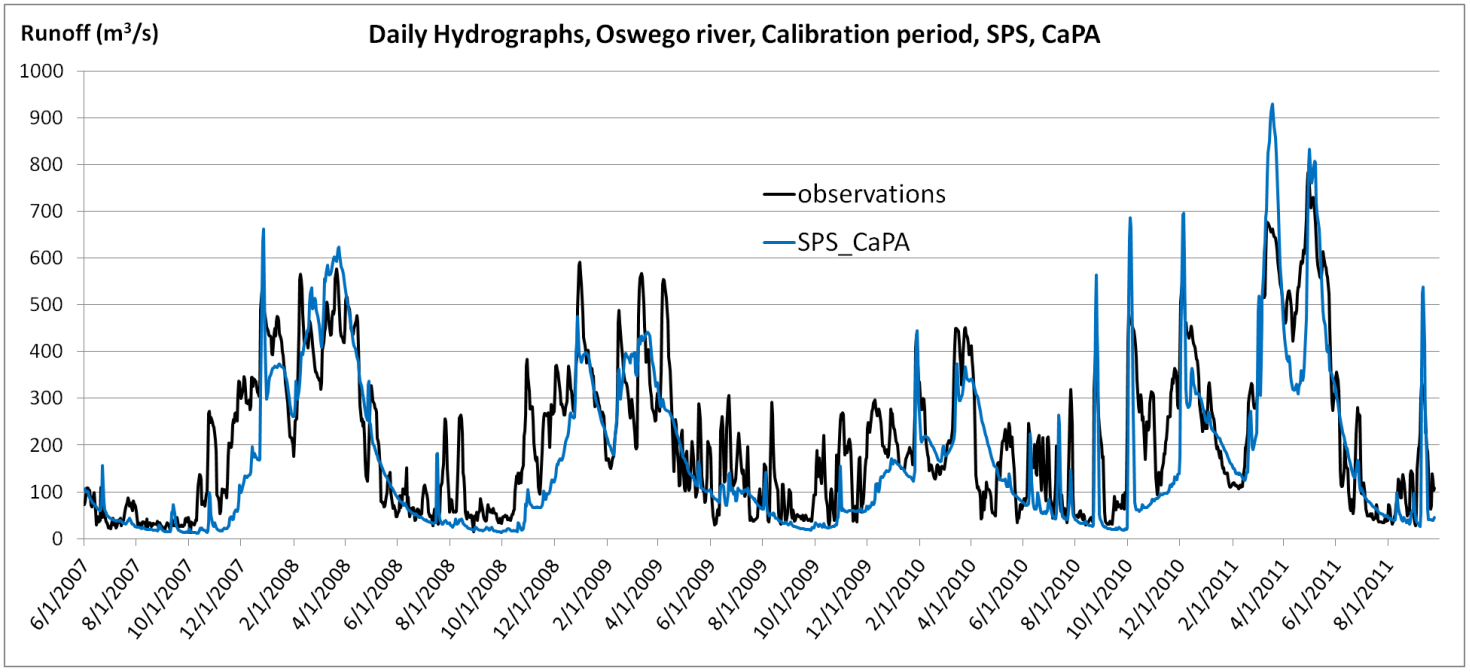

Figure 28: daily hydrographs in calibration period for the Oswego river, with the new set of Watroute calibrations.

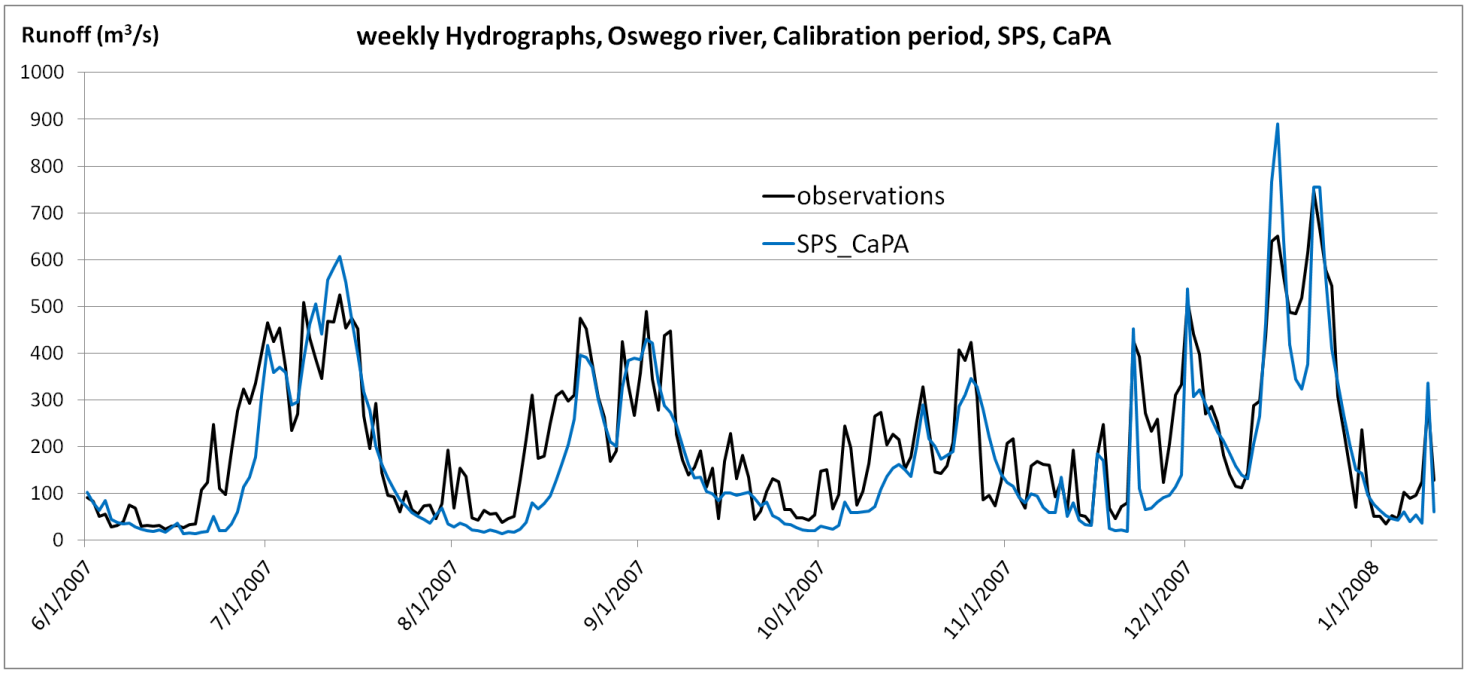

Figure 29: weekly hydrographs in calibration period for the Oswego river, with the new set of Watroute calibrations.

Table 38: Seasonal performances for the new Watroute calibrations, over the calibration period and for the Nash sqrt criterion; ALL: overall performances, WINT: winter (1st November - 31st of May), SUMM: summer (1st June - 31st October). Values are in percent.

					NASH SQRT		
Subbasin	Scheme	Station	Area(km2)	Regime	SPS_ALL	SPS_WINT	SPS_SUMM
1	2.0	20_mile	307.0	natural	24	21	-39
3	2.0	Genessee	6317.0	regulated	24	25	-171
4bis	2.0	Irondequoit	326.0	natural	27	16	-22
5	2.0	Oswego	13287.0	regulated	54	41	-15
6	1.0	N/A	2406.0	mixed	40	36	-15
7	2.0	Black river	4847.0	regulated	41	33	4
8	2.0	Oswegatchie	2543.0	regulated	43	25	30
10	2.0	Salmon_CA	912.0	regulated	72	52	-9
10bis	1.0	N/A	944.0	mixed	63	42	-20
11	2.0	Moira	2582.0	regulated	66	48	-52
12	1.0	N/A	12515.5	regulated	55	43	-19
13	1.0	N/A	1537.5	natural	33	16	32
14	1.0	N/A	2689.4	mixed	-1	-20	22
15	1.0	N/A	2245.8	mixed	-5	-32	-23

It can be finally noticed that winter performances are better than summer ones (Table 38), for the new GEM-Hydro simulations. This can be explained by two main factors, namely that in summer, the effect of regulation on streamflow is more severe, and flows are generally smaller, which make the same magnitude of errors between observed and simulated flows having more effect on the Nash sqrt criteria in summer than in winter. Furthermore, winter processes mainly consist in a large amount of runoff resulting from the spring snowmelt, which is generally better captured by the models than other processes related for example to infiltration, runoff, etc.

8.3) SVS snowmelt improvement

In SVS, the snowmelt formulation relies on a simple force-restore model and a degree-day formulation. Initially, the amount of water resulting from snowmelt was computed in SVS as a function of T_s, the snow temperature at the surface of the snowpack, which is allowed to increase above zero because of air temperature and solar radiations' influence. Relying solely on T_s hence results in overestimating snowmelt, and therefore in underestimating the SWE (the amount of snow and liquid water stored in the snowpack but expressed in mm of liquid water). But instead of implementing a more sophisticated, physically-based model to represent snow processes in SVS, a simple modification was first brought in order to improve snowmelt simulation. Instead of relying solely on T_s, an average temperature was estimated for the entire snowpack and then used to compute snowmelt quantities. This average temperature is computed as a function of T_s and T_m, with T_m consisting in a temporal average of T_s but without taking the air and solar radiation influence into account. As such, T_m is more representative of the deep snowpack temperature, which generally has a very limited daily fluctuation. The hypothesis was made that the temperature profile in the snowpack is an exponential function between the surface temperature T_s and the deep temperature T_m, which is reached at the damping depth, d, below which snow temperature is constant and equal to T_m. This temperature profile assumption is based on You *et al.* (2013). Furthermore, d is computed as a function of the snow thermal conductivity, still following the same study. The integral is then used to compute, at each model internal time-step, the average snowpack temperature, T_{avg}, which is then used to compute snowmelt in the "new_2" SVS version (Table **29**). Figure 30 below synthesizes the information explained above. Note that the physically-based snow model CROCUS (see Boone and Etchevers 2001)is currently being implemented in SVS in order to further improve the representation of snow processes in this LSS.

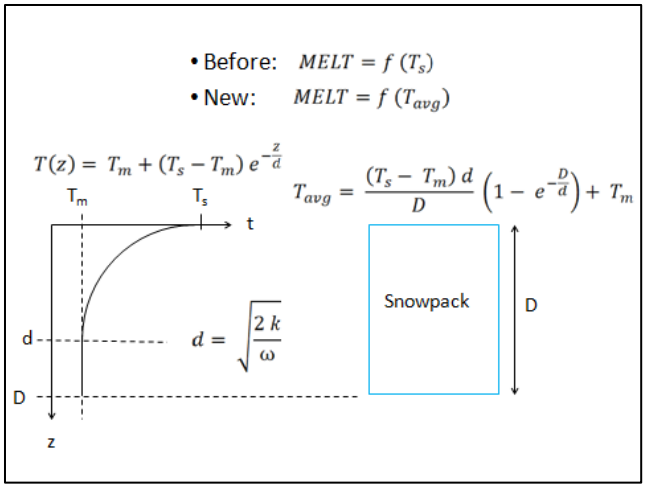

Figure 30: Description of the snowpack vertical temperature profile used to compute the snowpack average temperature, Tavg, used to compute snowmelt in SVS. This modification leads to the "new_2" SVS version presented in section 2.

Figure 31 below presents the comparison between the two versions of SVS in regard of snowmelt formulation, as described above, and in terms of resulting streamflow performances obtained with the default Watroute parameter set used with GEM-Hydro during this work (see Table **8**). In other words, the performances shown on Figure 31 are not the result of any calibration, but correspond to performances obtained with GEM-Hydro default parameters, for a few of the GRIP-O sub-basins (see Figure **31** and Table 39 below), as stated in the introduction of this section.

It is clear that relying on a more representative snowpack temperature leads to better streamflow performances, as snowmelt consists in one of the major hydrologic processes under this type of climate.

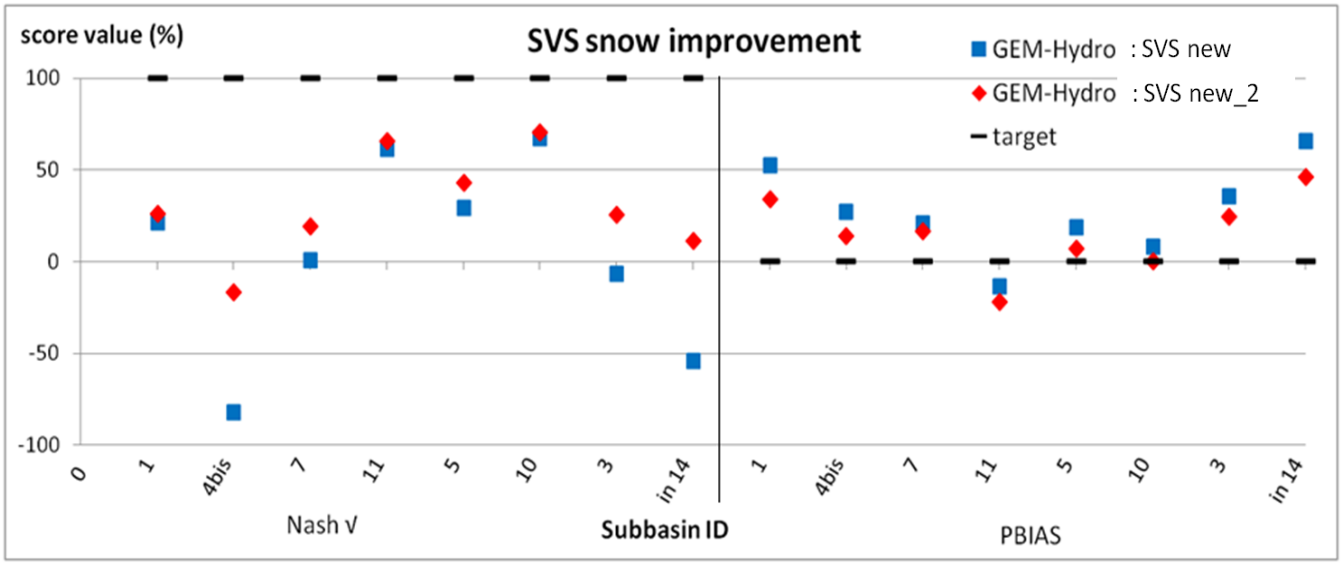

Figure 31: comparison between the new and new_2 SVS versions (Table 29). The new_2 version contains the new snowmelt formulation (see Text). Results are presented for the Nash √ (left) and PBIAS (right panel) criteria (see section 5.3) over the 2007-2011 period, and for a few GRIP-O sub-basins (see Table 39 below for catchment characteristics). Optimal values (target) for the scores are represented with the black markers.

Table 39: Catchment main characteristics for a few GRIP-O sub-basins (**Figure** 1) used to perform the comparisons of section 3.

Sub-basin	1	4bis	7	11	5	10	3	In 14	8	12
Station (river)	20 mile	Irondequoit	Black river	Moira	Oswego	Salmon	Genessee	Humber	Oswegatchie	Trent and Crowe
Area(km^2)	307	326	4847	2582	13287	912	6317	830	2544	12515
Flow regime	natural	natural	regulated	regulated	regulated	regulated	regulated	regulated	regulated	regulated

Finally, the effect of the snowmelt formulation improvement is shown on Figure 32 below for the mean SWE of sub-basin 7 (Black river). SNODAS is a modeling and data assimilation system which integrates snow data from satellite, airborne platforms and ground stations with model estimates of snow cover and associated variables (see NOHRSC 2004). Again, it is quite clear that SWE simulations are improved due to the SVS modification described in this section. However, it is highly possible that the snowmelt formulation used here may reveal less efficient in different areas, such as in the Rocky Mountains for example. This is why a more sophisticated snow model, CROCUS, is currently being implemented in SVS. Despite not shown, the new SVS version developed here also led to better SWE simulations for the other GRIP-O U.S. sub-basins (for which SNODAS data were available). The improvement brought to SVS in regard of snowmelt processes and described in this section led to another SVS version, referred to as the "new_2" version (see **Table 29**) used hereafter in sections 8.4 to 9.

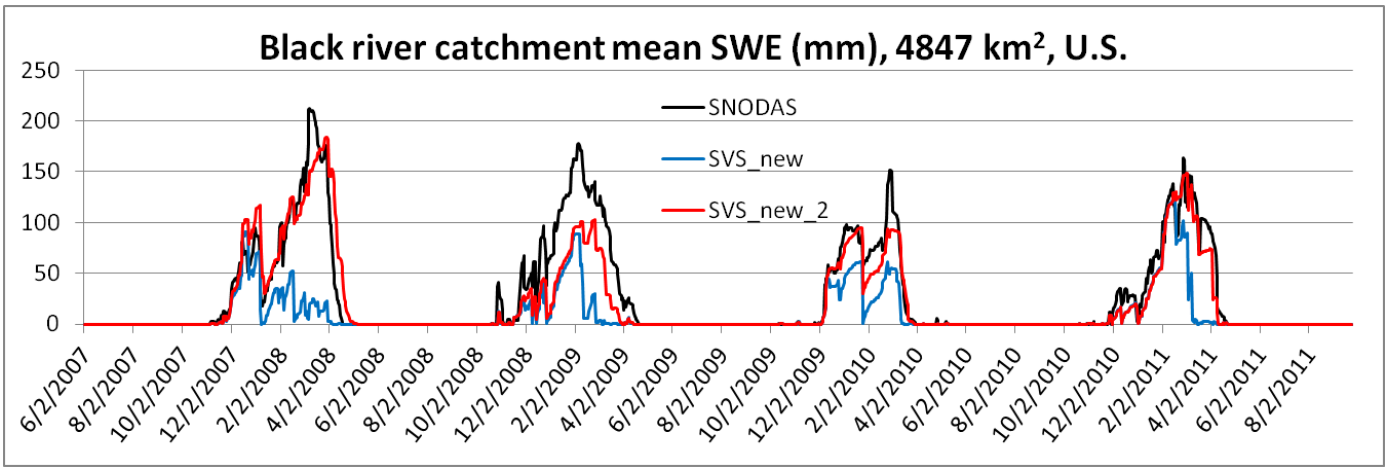

Figure 32: SWE simulations with the SVS versions under study. The new_2 version contains the new snowmelt formulation (see Text). The SWE quantities shown here represent mean catchment daily values for the Black river sub-basin (catchment 7, see Table **39** and **Figure 1**). SNODAS values are supposed to accurately represent real values.

8.4) Comparison between SVS and ISBA surface schemes

A test was finally made to compare the SVS and ISBA LSSs. ISBA is the LSS which is used in the atmospheric GEM model to produce the meteorological forcings required to drive the distributed models used in this work. All the work performed here to improve SVS is made in view of replacing ISBA with a more sophisticated and efficient LSS. Hence a run was made with GEM-Hydro and the ISBA LSS instead of SVS, using the exact same forcings as those used to drive SVS for the second watroute calibrations (see section 8.2), i.e. with CaPA precipitation. Watroute was then run using the outputs produced by ISBA (runoff and drainage, ISBA does not produce any lateral flow in opposition to SVS), and the same default parameter set that was used as the initial one when performing Watroute calibrations (see Table 8). In other words, Watroute was not calibrated with ISBA outputs. For a more fair comparison when comparing both LSSs, streamflows resulting from using ISBA in GEM-Hydro were compared to those resulting from using SVS with the same default Watroute parameter set in both cases (see Table **40** for ranges of the default values). The SVS version used in this section consists in the "new_2" SVS version (see **Table 29**). This is the latest SVS version as of April 4th, 2016.

Results are presented on Figures Figure 33 to Figure 37 below. It can be seen on these Figures that ISBA has the general tendency to produce more surface runoff than SVS, the latter producing more drainage than ISBA, which can be seen in Watroute storing more water in the LZS with SVS than with ISBA (see Figures Figure 33 to Figure 37 below). This results in streamflows produced with ISBA having a more flashy behavior, whereas SVS leads to smooth streamflows, which are generally too smooth with the default Watroute parameter set. It can

also be seen that despite ISBA seems to sometimes have a too flashy behavior and to overestimate peak flows, the ISBA simulated streamflows are however often better than SVS ones, especially for peak flows. As a consequence, it cannot be stated yet that SVS is superior to ISBA. However it seems that ISBA is lacking baseflow, so in regard of simulating soil processes, SVS seems better than ISBA. Furthermore, when looking for example at Figure 34, it seems that ISBA may not be conserving water. Work is under way to assess ISBA water balance. Work is also under way to compare SVS and ISBA LSSs in more details, for example by comparing the surface temperature and SWE simulated by both models.

It is reminded that results on Figures Figure 33 to Figure 37 are not the result of any calibration, and to better assess the relative performance of both LSSs would call for Watroute calibrations with ISBA outputs, the same way they were performed with SVS outputs in section 8.2. However this first comparison gives a good overview of the main differences between the SVS and ISBA surface schemes, in terms of resulting streamflows.

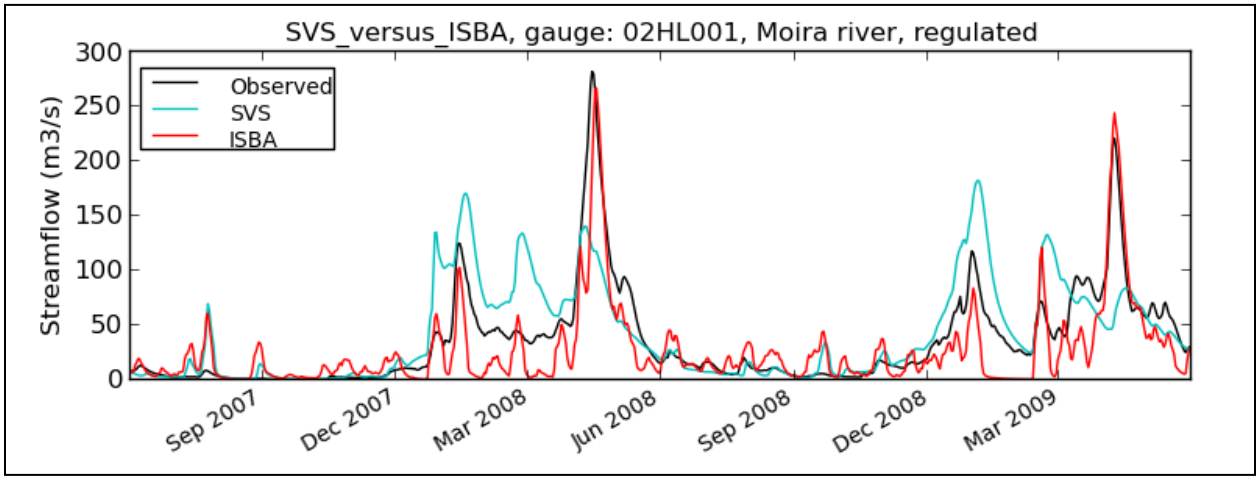

Figure 33: SVS (new_2 version) versus ISBA hydrographs obtained for the Moira river with CaPA precipitation forcings and a default Watroute parameter set (see initial values in Table 8).

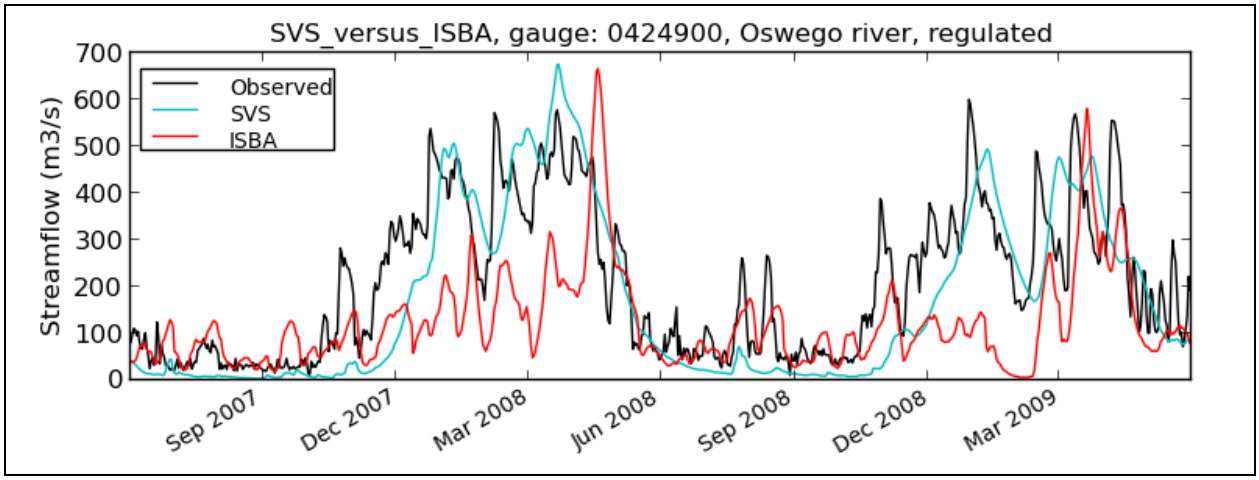

Figure 34: SVS (new_2 version) versus ISBA hydrographs obtained for the Oswego river with CaPA precipitation forcings and a default Watroute parameter set (see initial values in Table 8).

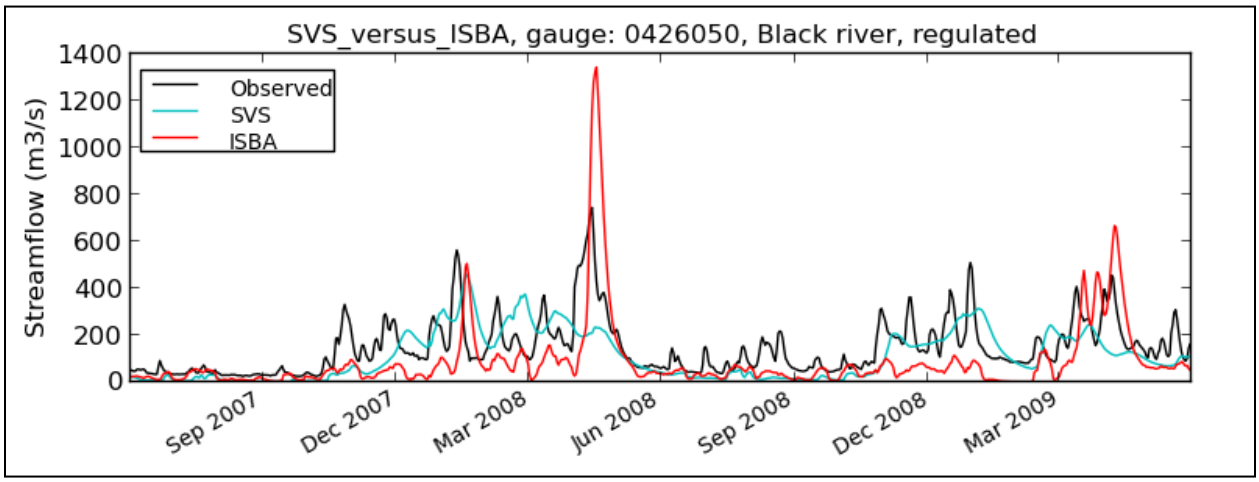

Figure 35: SVS (new_2 version) versus ISBA hydrographs obtained for the Black river with CaPA precipitation forcings and a default Watroute parameter set (see initial values in Table 8).

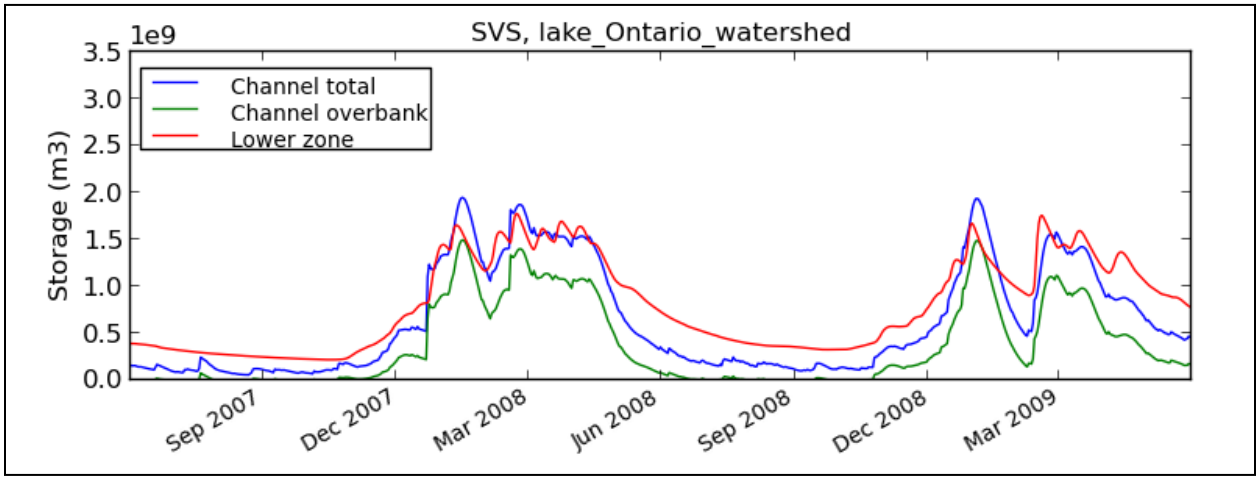

Figure 36: Watroute water storage (channel and lower zone) resulting from SVS (new_2 version) forcings, for the whole lake Ontario watershed.

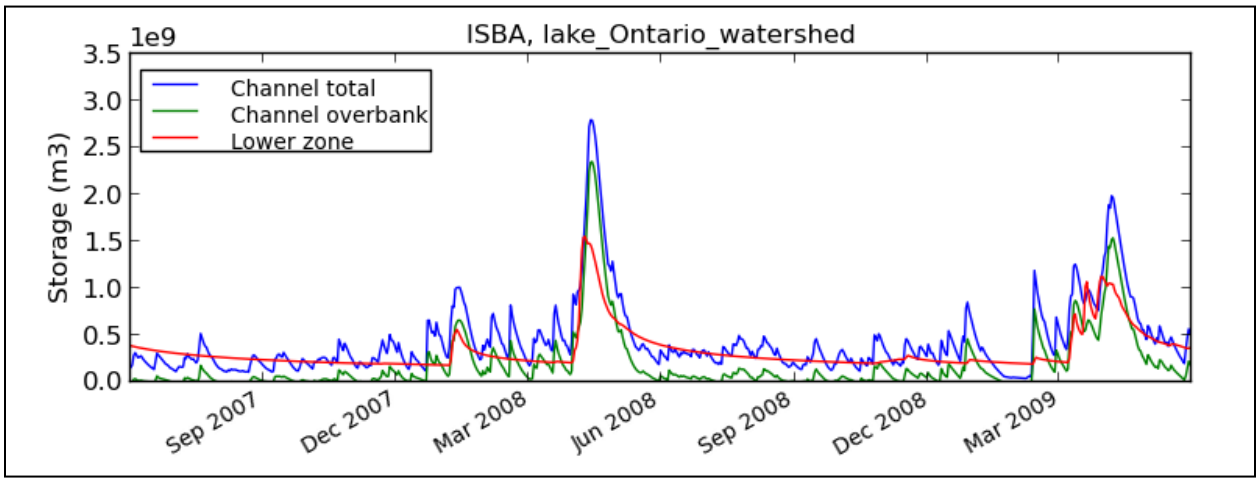

Figure 37: Watroute water storage (channel and lower zone) resulting from ISBA forcings, for the whole lake Ontario watershed.

After the improvement brought to SVS in regard of snowmelt, a comparison was performed between the two LSSs available in GEM-Hydro. ISBA is currently used at EC in the Great Lakes prediction system. Again, this comparison was performed without calibrating the LSSs parameters and using the default Watroute parameter values of Table 40. The comparison was performed over the same sub-basins as those of section 8.3. Figure 38 below presents the comparison between the new SVS version and the ISBA LSSs, when embedded in GEM-Hydro.

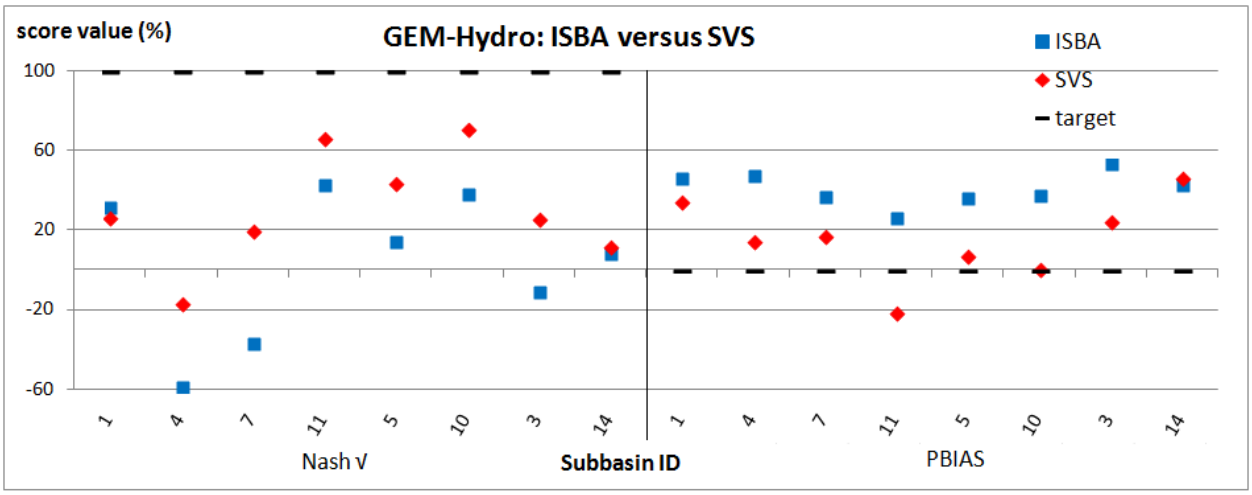

Figure 38: Comparison between ISBA and SVS (new_2 version) resulting streamflow performances, with default Watroute parameter values (Table 40). Results are presented for the Nash √ (left) and PBIAS (right panel) criteria (see section 2.3) over the 2007-2011 period, and for a few GRIP-O sub-basins (see Table 39 above for catchment characteristics). Optimal values (target) for the scores are represented with the black markers.

Figure 39 below shows the hydrographs corresponding to Figure 38's sub-basin 7 performances. It is obvious that according to the metrics used here, SVS leads to better streamflow simulations than ISBA. Moreover, ISBA tends to overestimate peak flows and does not accurately represent the hydrological processes occurring in the soil compartment, which thus results in poor baseflow simulations (Figure 39). This behavior was also observed for all other sub-basins (not shown here). Also not shown here are the SWE simulations obtained with both schemes. They are however quite close to each other, with no LSS being systematically better than the other when looking at mean simulated catchment SWE versus SNODAS (as in Figure 38). The spring flow peak overestimation with ISBA is supposed to be due to frozen soil water melting too quickly, because SWE simulations are however very similar between the new SVS version and ISBA. SVS is currently not representing soil water freezing and melting phenomena, which is supposed to be the reason why it generally does not succeed to represent spring flow peaks, in comparison to SA-MESH for example (see section 8.7). The implementation of a soil freezing and melting representation is currently in progress for SVS, at EC.

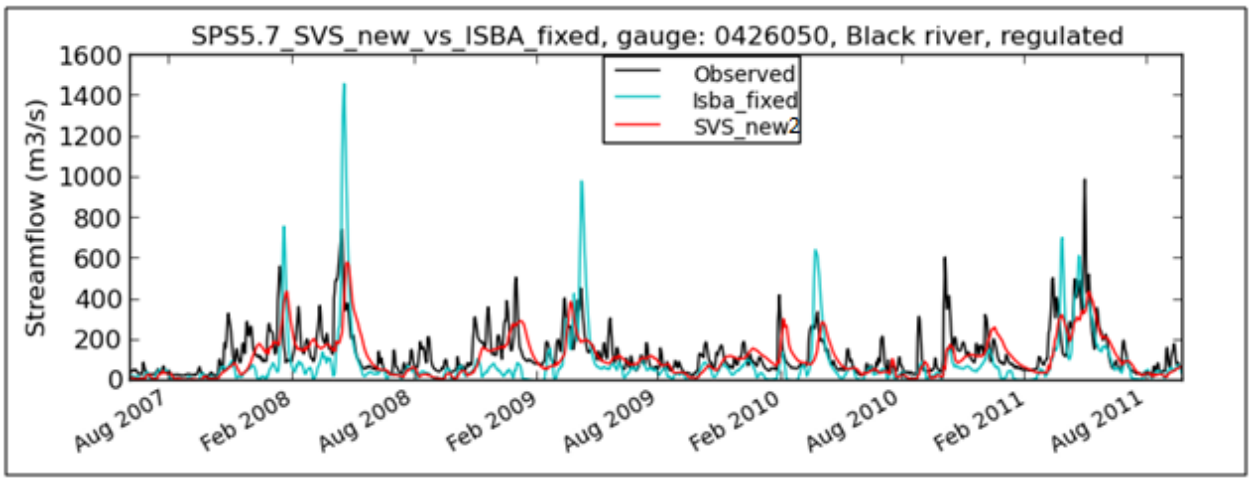

Figure 39: Comparison of ISBA and SVS (new_2 version) resulting streamflows for the Black river (catchment 7 on **Figure 1** and Table **39**). The Hydrographs were obtained with default Watroute parameter values (Table **40**).

8.5) GEM-Surf + UH implementation

As the whole GEM-Hydro suite (including Watroute) is quite highly demanding in terms of computational time, another version was developed to avoid having to run Watroute. For that, a simple UH was used to simulate streamflows at a gauge location, instead of using the routing scheme Watroute. This allows to save a significant amount of computational time and hence allows to calibrate the model, which requires a high number of runs (see section 5.3). The version using a UH to simulate streamflows is referred to as "GEM-surf + UH", in opposition to GEM-Hydro which includes Watroute. The UH allows to estimate streamflow at a basin outlet by partitioning basin averages of runoff and recharge in time. The same LZS formulation as in Watroute was used in GEM-surf + UH. The UH only requires a decay parameter corresponding to the lag time or response time of the considered catchment, which controls the gap between the rainfall event and the resulting streamflow peak. . It is derived from the UH used with the routing store of GR4J (Perrin et al., 2003), but was employed here with an hourly time-step. This way, hydrographs from GEM-Hydro and GEM-surf + UH can be very similar given that the UH lag-time value is well approximated (see Figure 40). The method used to approximate it is the Epsey method (see Almeida *et al.* 2014), which requires the catchment area, perimeter, and the max. and minimum elevations along the catchment main river. The two parameters actually used by the Epsey method are the main river length and mean slope, which are derived from the aforementioned catchment characteristics. The UH lag-time is also used as a free parameter during calibrations (see Table **9**). All calibrated parameters obtained with GEM-surf + UH can afterwards be provided to GEM-Hydro, which then only needs

Watroute Manning coefficients to be adjusted in order to mimic the optimal hydrographs obtained with GEM-surf + UH.

GEM-Hydro was however used during this work to assess a snowmelt improvement in SVS (section 8.3), and to compare SVS and ISBA (section 8.4), in which cases parameter values were not calibrated but set to "default" values. The version of Watroute used in this section relies on spatially-varying Manning values derived from physiographic information (i.e., land use), and on fixed values (i.e. the same everywhere inside a given watershed) for the two LZS coefficients (see section 6.2.2). See **Table 29** for a synthesis of the different SVS versions used during this work. The SVS version used with GEM-Surf + UH in this section consists of the "new_2" version (see same Table). The Routing parameter values were actually manually tuned in order to end up with default values generally suited to the whole GRIP-O area (see Table **40**). In SA-MESH and Watflood however, Watroute relies on fixed values for both the Manning and the LZS coefficients. Table 40 below summarizes the range of Watroute default parameter values used in the case where the spatially-varying Manning coefficients were used.

Table 40: Default Watroute parameter values used in this section; for Manning coefficients (coeff.), a range is displayed because values vary in space over the domain as a function of land use (see text). These values result from manual tuning over the area under study.

Parameter	range / value
Manning roughness coeff. for main channels, N2	0.06 - 0.1
Manning roughness coeff. for overbank flow, N1	0.07 - 0.2
LZS coefficient (LZS)	1.0 E-5
LZS exponent (PWR)	2.8

Figure 40 below presents the hydrographs simulated for the Moira river (sub-basin 11, see **Figure 1**), with default Watroute parameter values (see Table 40) in the case of GEM-Hydro and a UH lag time estimated with the Epsey method in the case of GEM-Surf + UH. As can be seen from this figure, GEM-Surf + 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 computational time. It is reminded that the SVS version used in this section consists of the last one at the time of writing, namely the "new_2" version (see **Table 29**).

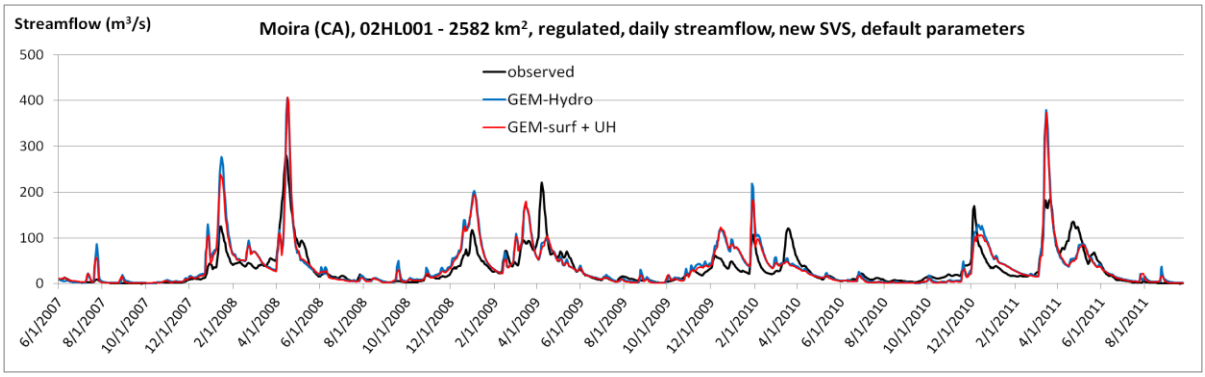

Figure 40: Hydrographs produced with GEM-Hydro and GEM-Surf + UH for the Moira river, over the GRIP-O calibration period (section 5.3). These simulations were made with the models' default parameter values (see Table 40 and text of this section), and SVS "new_2" version.

8.6) GEM-Surf + UH calibration

This section presents the performances of GEM-Surf + UH, either with its default parameter values (as derived from physiographic information and Epsey method, see section 8.5) or after calibration, for the main sub-basins of Table 39. Results are presented on Figure 41. The SVS version used in GEM-Surf +UH to calibrate SVS parameters consists of the "new_2" version (the last one at the time of writing), see **Table 29**.

Sub-basin 12 (**Figure 1**) includes 2 gauge stations. Therefore, for this catchment, the two observed flow time series were summed, and the resulting time-serie was extrapolated to the whole sub-basin area using the ARM (see Fry *et al.* 2014). This resulting "synthetic" flow time series is considered as the reference target to calibrate the model, and GEM-Surf + UH was implemented in this case over the whole sub-basin 12 area. This methodology was used with sub-basins with more than one gauge for consistency with the calibration experiments performed in the first part of the work (see sections 5.2 and 7).

As can be seen from Figure 41, the calibration allows to significantly improve the Nash √ performances of the streamflow simulations, but also the conventional Nash and Nash Ln (Nash computed with the logarithm of the flows), despite not shown here for the last two. The PBIAS criteria also generally improves after calibration, but to a lower degree, and sometimes the sign changes after calibration (see sub-basins 12 and 11 on Figure 41), switching from an over- to under-estimation of the observed flows in these cases. Interestingly, the Nash √ performances associated to the default parameter values are already above zero for all sub-basins of Figure 41, suggesting that these default simulations already possess some skill before the calibrations.

The default simulations are even satisfactory in some cases, such as for catchments 10 and 11. This is encouraging for example in the context of ungauged catchments, where no

calibration is possible. Note that with default parameter values (Table 40), performances obtained with GEM-Hydro (not shown here), are most of the time better than those obtained with GEM-Surf + UH, but are generally very close. This is in favor of using the whole GEM-Hydro suite instead of the simpler GEM-Surf + UH in case no calibration is performed, again for example in the case of ungauged catchments.

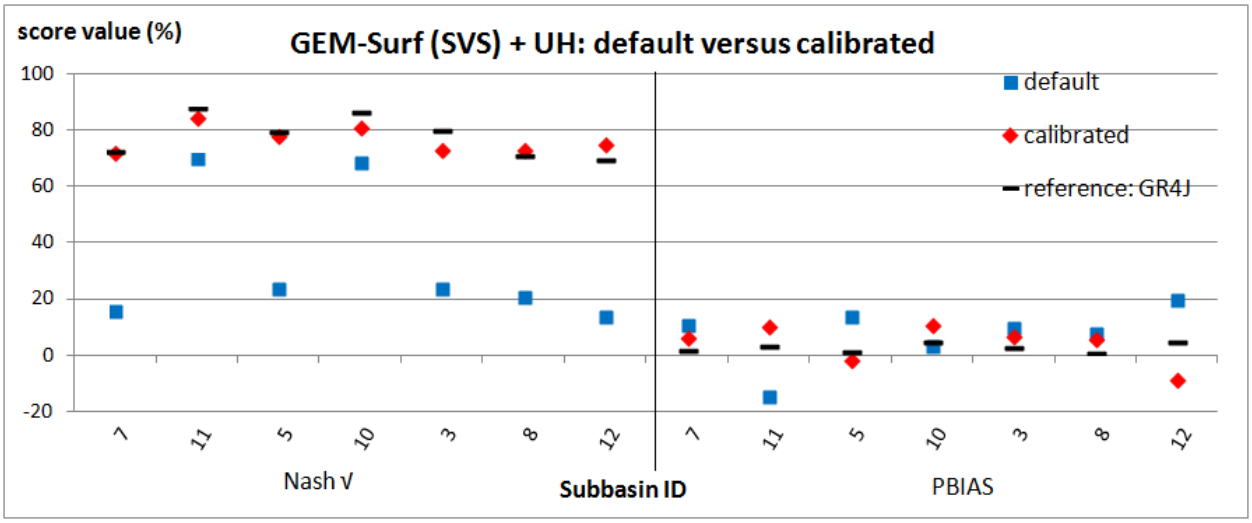

Figure 41: Comparison between GEM-Surf (with SVS new version) + UH default and calibrated performances. Results are presented for the Nash √ (left) and PBIAS (right panel) criteria over the calibration period, and for a few GRIP-O sub-basins (see Table 39 for catchment characteristics). The black markers represent the calibrated GR4J performances obtained with CaPA precipitation and in calibration, and consist in the target to be able to reach or surpass for GEM-Surf + UH.

Beyond the performances associated to streamflow simulations, which clearly demonstrate the superiority of, and therefore the mandatory need for calibration, GEM-Surf + UH calibrations also allow to improving SWE simulations. These results are not presented here for the sake of brevity and because the gain in SWE simulations, after calibration, was much smaller than the gain achieved for streamflow, from which the objective function is computed. The SVS calibrations also affect evapo-transpiration (see Table 9), but no observed values were available for this variable in a large enough extent for the area under study. For example, for the Moira river, the mean catchment annual evapo-transpiration (over the calibration period) is equal to 527 mm and to 647 mm, respectively before and after calibration. The robustness of the model is generally very good (i.e., performances do not strongly deteriorate between the calibration and validation performances (not shown).

Final GEM-Surf + UH calibrated parameter values are not shown because they highly depend on the sub-basin considered. Even for neighbor catchments such as the 10 and 11, the

3 and 5, or the 7 and 8 sub-basins (**Figure 1**), final calibrated parameter values are very different from one catchment to the other, which may due to equifinality (different parameter sets can lead to similar simulations) but also to the anthropogenic streamflow regulations of these catchments. Moreover, the GEM-Surf + UH calibration highlighted the equifinality issue, where different parameter sets can lead to similar performances for streamflow simulations, and which is even clearer when the number of free parameters used in calibration increases. In summary, final parameter values did significantly vary from one catchment to the other for most of the 16 parameters used here, and this strongly limits the potential for parameter transferability to other catchments. One way of overcoming this drawback would be to perform a global calibration over all the catchments considered here, which would result in a unique, spatially coherent parameter set for all of them, which generally leads to robust simulations and could therefore be used for the remaining ungauged parts of the lake Ontario watershed. This is done in section 9.

But maybe the main key finding of Figure 41 is the fact that the calibrated GEM-Surf + UH performances are generally very close to those obtained with GR4J and CaPA precipitation. This means that GEM-Surf + UH, via calibration, is able to reach the performances of GR4J, a very simple, efficient and robust model. This is very encouraging as the performance benchmark set by GR4J simulations is most of the time quite high and hard to reach for other models. Moreover, as new improvements are in progress for SVS (see sections 8.3 and 8.4), it is probable that GEM-Surf + UH (and hence GEM-Hydro too but maybe only after tuning Watroute parameters) will even be able to surpass GR4J in terms of performances, in the near future.

8.7) inter-comparison of calibrated semi-distributed models

This section aims at comparing the performances obtained with the calibrated versions of SA-MESH, Watflood and GEM-Surf + UH (with SVS new_2 version, see **Table 29**) for some GRIP-O sub-basins, in order to highlight behavior differences between these models. The calibration strategy used for each of them is described in section 5.3. The SA-MESH and Watflood models were only calibrated for the Moira, Black and Salmon rivers due to the strong effort required. All models were calibrated with the exact same forcings, i.e. CaPA precipitation and model temperature (see Table 4). As mentioned in section 2, WATFLOOD and SA-MESH were implemented with a 10 arcmin. spatial resolution (both for the LSS and the routing scheme), while GEM-Hydro was implemented with a 10 arcmin. resolution for the LSS and 30 arcsec. for the routing scheme (WATROUTE). Watroute is able to produce streamflow simulations which are of similar quality, be it implemented at a low (10 arcmin. for SA-MESH and Watflood) or high (0.5 arcmin. with GEM-Hydro) resolution (and given that the catchment

size is not too small), but it was still preferred the high-resolution version in GEM-Hydro for consistency with the version being used in the recent Great Lakes levels' forecasting system developed at EC. Furthermore, this difference in the routing scheme resolution will not give GEM-Hydro any advantage in comparison to the other models, given that the GEM-Hydro version used in calibration (i.e., GEM-Surf +UH) does not involve any routing scheme (see section 8.6).

The internal time-step used for SA-MESH and Watflood is respectively equal to 30 and 60 min, while it is equal to 10 min. for GEM-Surf + UH. All models were calibrated using parameters affecting a wide range of the hydrological processes involved in the models (see Table **9** to Table **11**).

Results are presented on Figure 42 and **Figure 43** below for the calibration period. It was not judged informative to present the final calibrated parameter values for each model, especially considering that these values are highly location dependant and subject to the equifinality issue (see section 5.3). In other words, parameter values obtained here for the small sub-basins considered may a priori not be transferable to other regions of the world. Despite the number of study cases is limited to only two or three sub-basins of the same area, and results about the model relative performances cannot be generalized, it can be seen that GEM-Surf + UH generally outperforms SA-MESH and Watflood, both in calibration and validation (Figure 42). This was verified also for the two other quality criteria considered in this study and not shown here. GEM-Surf + UH thus sometimes outperforms or comes very close to GR4J performances (Figure 41). The robustness of the models is generally quite good (i.e., performances do not strongly deteriorate between the calibration and validation performances, except for SA-MESH on the Black river sub-basin (catchment 7 on Figure 42), which seems to have difficulties at this location.

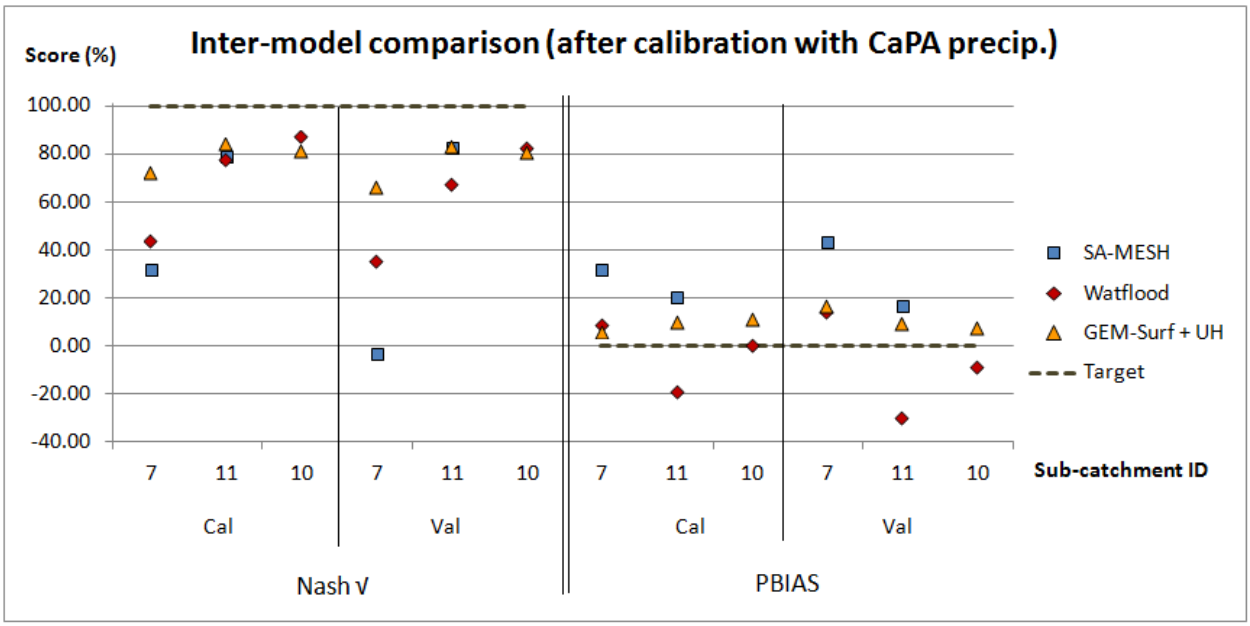

Figure 42: Inter-model comparison for calibrated distributed models on three GRIP-O sub-catchments (see Table 39 for catchment properties). SA-MESH was not implemented on sub-basin 10. Cal, Val: calibration and validation periods, respectively. Perfect scores (target) are displayed.

When looking more closely at the hydrographs produced in calibration and for the Moira river (Figure 43), strong differences arise between the three semi-distributed models studied here. For example, Watflood seems to have a more flashy behavior and tends to overestimate peak flow events, while SA-MESH generally underestimates observed flows, with GEM-Surf + UH lying somewhere in between. The peak flow events resulting from spring snowmelt are generally better represented by SA-MESH and reflects through the traditional Nash criterion. This is supposed to be due to the fact that SA-MESH, relying on the CLASS LSS, does represent freezing and melting processes occurring in the soil. As a consequence, even if GEM-Surf + UH accurately represents the timing of snowmelt, the additional decay caused by soil freezing on resulting hydrographs may remain unseen by the model (see the last spring peak flow events on Figure 43 below). As stated in section 8.4, work is under way to fix this drawback. This was confirmed when looking at the two other sub-basins (the Salmon and Black rivers).

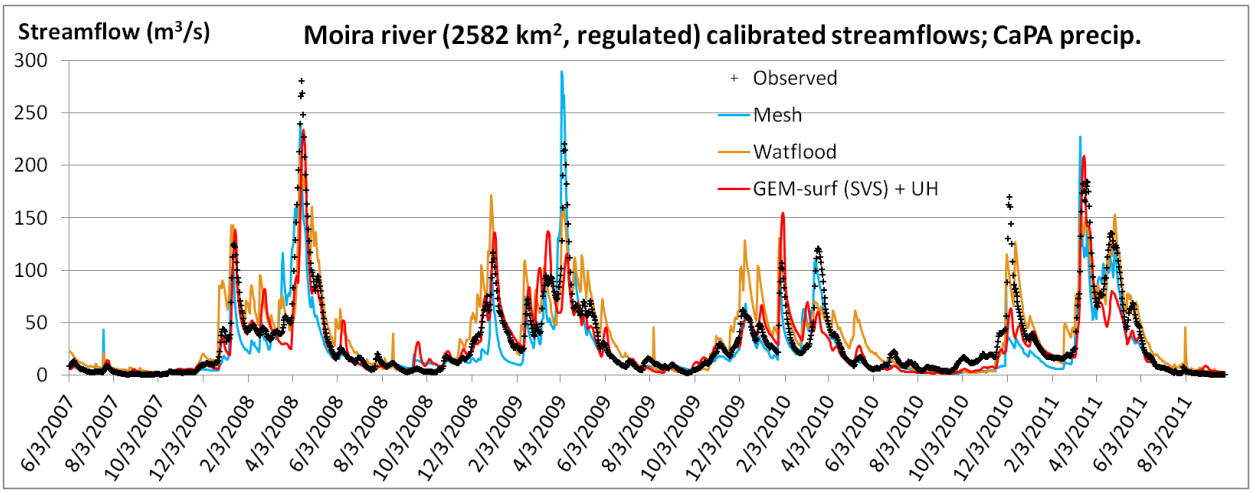

Figure 43: inter-model comparison for the calibrated distributed models. Hydrographs of the calibration period for the Moira river (sub-basin 11).

At this point, the differences between the strategies used to calibrate the models (see section 5.3) are reminded. First, the comparison between SA-MESH and GEM-Surf + UH tend not only to confirm the relevance of the UH choice to estimate streamflows by saving a tremendous amount of computational time, but also to comfort the choice of mainly using a small number of multiplicative coefficients to calibrate GEM-Surf + UH (see Table 9). Indeed, it seems that the methodology followed for GEM-Surf + UH allows to achieve very satisfying performances after calibration, while limiting the number of free parameters (16 versus 60 for SA-MESH), and therefore allowing to use a smaller number of model runs (400 versus 1000) for the model during calibration. Therefore, it could be interesting to try to apply this calibration framework to SA-MESH too, which would result in using 12 instead of 60 free parameters, which may therefore allow to use additional free parameters. The less satisfying hydrographs obtained for Watflood in comparison to the other two models (Figure 43) also logically arise the interrogation whether another calibration strategy should be used for this model, for example by using additional free parameters and also using the multiplicative coefficients framework in order to use only one given calibration coefficient for the three main land covers considered.

These questions about the best model calibration framework however remain completely open, as the best framework depends on the model but also on the area and data under consideration. Moreover, multi-objective calibrations could be investigated, for example by trying to simultaneously optimize the SWE and streamflow simulations, or by using more than one criteria to assess performances (for example the PBIAS criteria). This work is beyond the scope of this paper, which aimed at comparing the three semi-distributed models studied here in the most fair manner as possible. Therefore, in this case we chose to rely on expert knowledge with each of the models to calibrate them, which led to differences in calibration strategies (see section 5.3) and therefore arise the question whether each strategy was actually

the best or not for each of the models. However, perfect models would achieve almost perfect simulations even with a poor calibration framework, which leads us to the obvious conclusion that none of the models used here is perfect.

This inter-comparison work allowed us to compare these models in a quite fair manner but for a limited number of test cases, yet it allows to highlight the mandatory need of calibrating hydrologic models, that models are always unique, which translates in significant differences for the Hydrographs, and that each of the models under study could benefit from some strengths of its competitors. For example, SVS would benefit from the implementation of soil freezing and melting processes. Of course, this inter-comparison also clearly demonstrates the interest and potential of GEM-Surf + UH, and hence of the SVS LSS, but also of the GEM-Hydro platform used at EC for example for flood and lake level forecasts of the Great Lakes area.

8.8) conclusion of the work performed with semi-distributed models

Results have shown that the SVS LSS, as embedded in the GEM-Hydro or GEM-Surf + UH platform used here, demonstrates a strong potential in regard of runoff simulations for the lake Ontario watershed. This SVS version is named "new_2" (see Table **29**) and is the result of several improvements, bug fixes, and configuration refinements in comparison to the first version available at the beginning of this work. This "new_2" version of SVS is better than the ISBA LSS, and is expected to replace it in EC operational models in the near future. However, there is still much room for improvement for SVS, as demonstrated by the fact that it significantly benefited from a very simple modification to its current snowmelt representation. For example and as illustrated by the comparison to Watflood and SA-MESH, the two other distributed models studied here, SVS may benefit from the implementation of soil freeze-thaw processes, which are supposed to be responsible for SVS missing some of the runoff peaks in spring. According to the inter-comparison experiments conducted during this work on three sub-catchments of the lake Ontario watershed, GEM-Surf + UH seems to be a strong competitor for SA-MESH and Watflood. However, as a very limited number of catchments were used for the inter-comparison due computational time limitations, no reliable general model ranking can be made in this study. Calibration has of course proven that it is mandatory to optimize model performances, which thus allows to make GEM-Surf + UH performances get very close to GR4J ones.

The main conclusion of the experiments with distributed models is that they can compete with lumped ones when both are calibrated, yet of course not in terms of implementation efficiency or required computational time. In order to calibrate the GEM-Hydro model, its routing part was replaced by a simple unit hydrograph during calibration, which

allows to save a tremendous amount of computational time. The routing model of GEM-Hydro can be run after calibration, by adjusting the default Manning values to obtain performances similar to GEM-Surf + UH, if needed. Lumped models have limited applications, while distributed ones can be used for a priori any environmental study. Many distributed models 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 study quite challenging, if not unrealistic. Several improvements are currently under development for SVS, such as a physically-based snow model (ISBA-ES), and freeze-thaw soil processes, which are being implemented at EC and whose resulting benefits on streamflow simulations are dedicated to future work.

9- Runoff estimation for the whole lake Ontario watershed

The ultimate aim of the GRIP-O project consisting in the runoff estimation for the whole lake Ontario watershed, including its ungauged areas, the following methodology was chosen in order to fulfill this objective. Calibration was performed using the GR4J and GEM-Surf + UH models on the GRIP-O gauged area, and the resulting parameter set was transferred to the models when implemented on the total lake Ontario area, including its ungauged parts (Figure 44). The GRIP-O gauged area consists in the true gauged area (Figure **44**), plus the ungauged areas of the gauged sub-catchments including multiple gauge stations. This is because with local models (as with GEM-Surf + UH in this case) and in the case of catchments with multiple gauges, this was preferred to consider a unique reference flow time-series for the whole sub-catchment. To do so, the same methodology described in section 8.6 for sub-catchment 12 was used for sub-catchments 6, 10bis, 13, 14, and 15. Therefore, the gauged area considered in this section and referred to as the "GRIP-O gauged area" is actually gauged at 88.5%.

Despite several different possibilities exist for performing the runoff estimation on the total GRIP-O area, the one explained above (calibration on the whole gauged area and parameter transfer to models on the total area) was preferred mainly for these two reasons: firstly, doing so allows to calibrate the models using close approximations of observed flows (the area used for calibration is gauged at 88.5%), instead of less reliable flow estimations for the whole area (gauged at 69%). And secondly, using this methodology allows to take 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 over the total area. Yet this methodology involves two different implementations of each model: one for the gauged part of the watershed, and one for the total area which includes the ungauged parts (Figure 44).

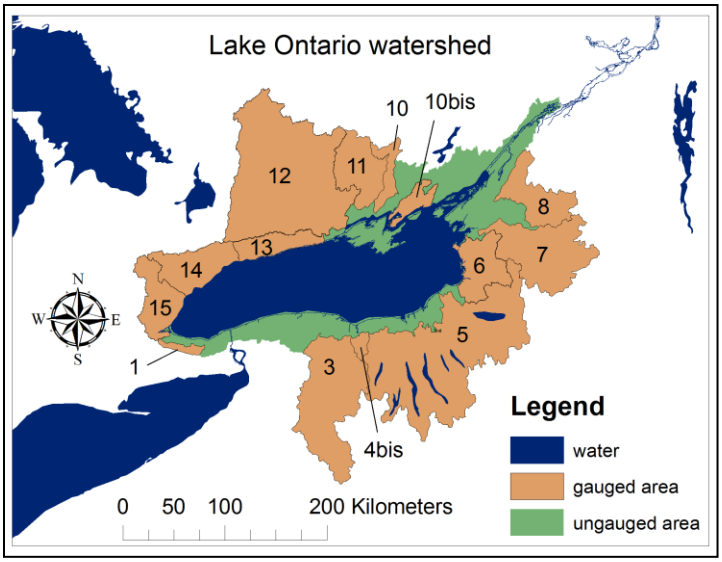

Figure 44: GRIP-O area considered as gauged for model (global) calibration, and ungauged area of the total lake Ontario watershed.

Moreover, the aim of this section is to propose an efficient and reliable methodology with the purpose of estimating runoff from large areas. Therefore, the methodology proposed involves a minimal number of steps and may be transferred to different regions of the world.

It was demonstrated in section 7.3.4 that a unique GR4J model calibrated over a large area could lead to runoff estimations of similar quality than multiple models implemented over sub-catchments of the total area, in terms of runoff estimations for the total area. A unique model has the advantage of requiring only a unique implementation and a unique calibration, whereas local models require at least multiple implementations (if global calibration is used), and in the worst case multiple calibrations (if local calibration is performed). Global calibration consists in finding a unique, trade-off parameter set which allows to simultaneously improve performances for all sub-catchments, whereas local calibration consists in finding each sub-catchment's optimal parameter set. Local calibration logically leads to optimal performances for a given catchment, but global calibration generally still allows to achieve satisfactory simulations while presenting interesting advantages. First, global calibration can lead to models with a better temporal robustness than local calibration (see Gaborit et al., 2015). Secondly, global calibration allows to achieve a spatial consistency of the parameter values, because they are either fixed or adjusted the same way over the total area under study. Local calibration, on the opposite, because of equifinality and experiment imperfections (model processes, forcing data, observed flows, etc.), compensates for simulation errors and leads to parameter sets than can hardly be transferred to other (even neighbor) catchments, as tends to suggest the fact that very different parameter sets were obtained here (not shown) with local calibrations of GEM-Surf + UH, even for neighbor catchments.

For GR4J, local calibration was used but with a unique model for the complete gauged area. For GEM-Surf + UH, the model was implemented for each of the GRIP-O gauged sub-catchments, and then calibrated with global calibration and the following objective function:

$$OF = \sum_{i=1}^{N}(1 - \frac{Nloc_i}{Nglob_i}) \qquad \text{Equation (2)}$$

with $Nloc_i$ the Nash criteria (using the square-root of the flows) obtained with local calibration on catchment i, and $Nglob_i$ the Nash criteria (using the square-root of the flows) with global calibration on catchment i. The objective function used during global calibration aims at minimizing the difference between performances obtained with the global parameter set, and those obtained with local calibration. However, as GEM-Surf + UH was not locally calibrated for all of the 14 GRIP-O sub-basins used during global calibration, the performances obtained with the local calibration of GR4J was used instead of the GEM-Surf + UH one when needed.

The models were then implemented over the total lake Ontario watershed (Figure 44), and runoff simulations were performed with the parameter sets calibrated over the gauged area. However, it is GEM-Hydro which was used over the total area, since it was more straightforward and a priori more realistic to use Watroute instead of the simple UH for the ungauged areas of the lake Ontario watershed. In this case however, the same default Manning coefficients of Table 40 were used in Watroute, while the lag-time of GEM-Surf + UH was adjusted during calibration. But it was assessed that simulations with GEM-Hydro (with calibrated SVS parameters, calibrated LZS parameters and default Manning coefficients) were very close, both in terms of hydrographs and performances at the gauged sites, to those from the calibrated GEM-Surf + UH. Performances are actually generally even slightly better with GEM-Hydro (despite the default Manning values) than with GEM-Surf + UH.

Results are shown first for the gauged sub-catchments individually, in order to compare global calibration, local calibration and default parameters (with GEM-Surf + UH, Figure 45), then for the whole GRIP-O gauged area as considered here (see text above) to compare GR4J and GEM-Surf + UH performances (Table 41 and Figure 46). It is better to assess model performances over the gauged part of the lake Ontario watershed, because as previously stated, runoff observations are closer to reality in this case than when extrapolated to the entire lake Ontario watershed using the ARM. Finally, results are shown for the whole lake Ontario watershed (Table 41 and Figure 47), which is the final aim of this work.

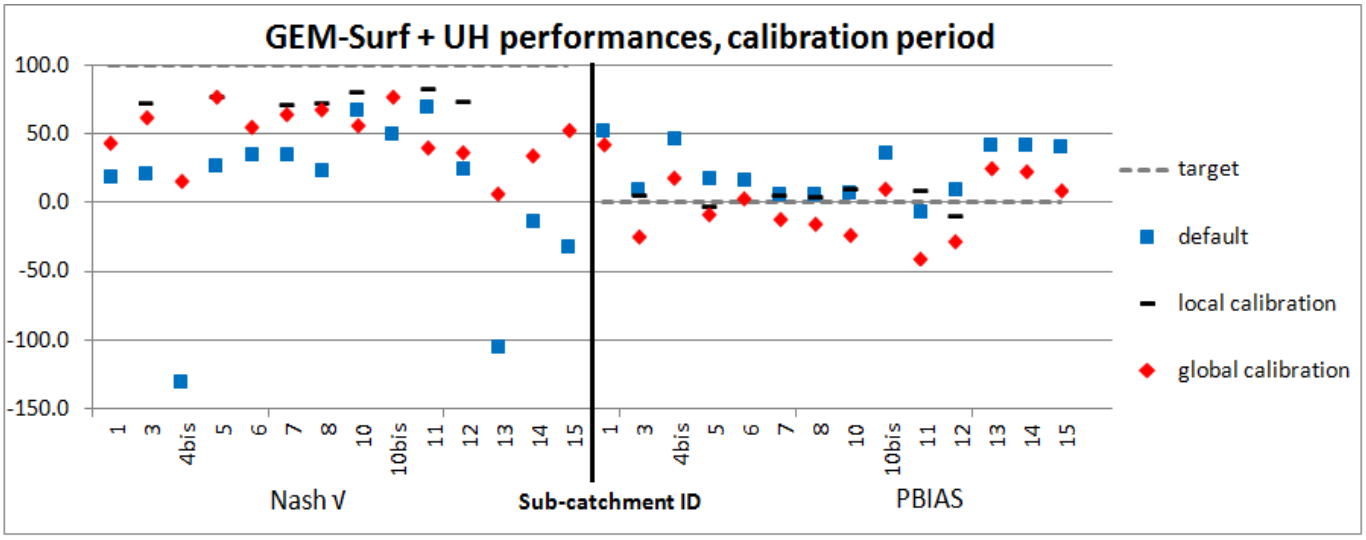

Figure 45: GEM-Surf + UH performances in calibration for the 14 GRIP-O gauged sub-basins (Figure 44), and for the default, locally and globally-calibrated parameters. The target (perfect) score is shown.

Table 41: performances assessed for the whole gauged area or total lake Ontario watershed (Figure 44), with the unique GR4J, the globally-calibrated GEM-Surf + UH, or the GEM-Hydro models (see text for more details). Cal., val.: calibration and validation periods, respectively.

	GRIPO gauged area: 53459.2 km^2				lake Ontario watershed: 68214.8 km^2			
	GR4J		GEM-Surf + UH		GR4J		GEM-Hydro	
Scores (%)	cal	val	cal	val	cal	val	cal	val
Nash	82.4	84.6	78.9	83.8	82.9	85.5	83.5	85.3
Nash √	84.7	85.5	83.1	88.3	84.4	85.0	83.0	86.8
Nash Ln	83.3	84.0	83.2	89.2	82.4	82.8	81.1	86.5
Pbias	-0.3	1.5	-10.6	-6.2	-2.2	-1.2	-10.6	-4.6

Performances of GEM-Surf + UH are lower with global calibration than with local calibration, as expected. Performances are sometimes even lower after global calibration than with default parameters, for some catchments (see sub-catchments 10 and 11 on Figure 45).

However, performances are satisfactory for most of the 14 GRIP-O sub-catchments with a unique parameter set, which confirms that global calibration fulfilled its role. Given that it takes between 2 to 5 days to achieve a local calibration, global calibration, which was completed in 10 days, allows to save a significant amount of computational time (there are 14 GRIP-O sub-catchments for which local calibration would have to be performed). Furthermore and as previously stated, global calibration thus leads to spatial consistency for the parameter values, and hence allows a straightforward parameter transfer to ungauged areas, whereas there is no a priori best manner to transfer parameter values obtained with local calibration (Parajka et al. 2013).

Performances for the whole GRIP-O gauged area (Figure 46 and Table 41) show that GR4J is better than GEM-Surf + UH in calibration, but worse in validation, in this case. GEM-Surf + UH leads to very satisfactory performances, but most importantly to often better streamflow simulations than GR4J in terms of dynamics: GR4J for example sometimes only represents a smooth streamflow recession, while GEM-Surf + UH better follows the small peaks and drops occurring during the recession (see for example the 2006 summer of Figure 46). Note that thus smooth GR4J behavior is not due to the unique model approach for the whole area, as the same behavior occurred when aggregating the simulations from local GR4J models calibrated with local calibration (not shown). This smooth behavior is inherent to the lumped attribute and the concepts of the GR4J model.

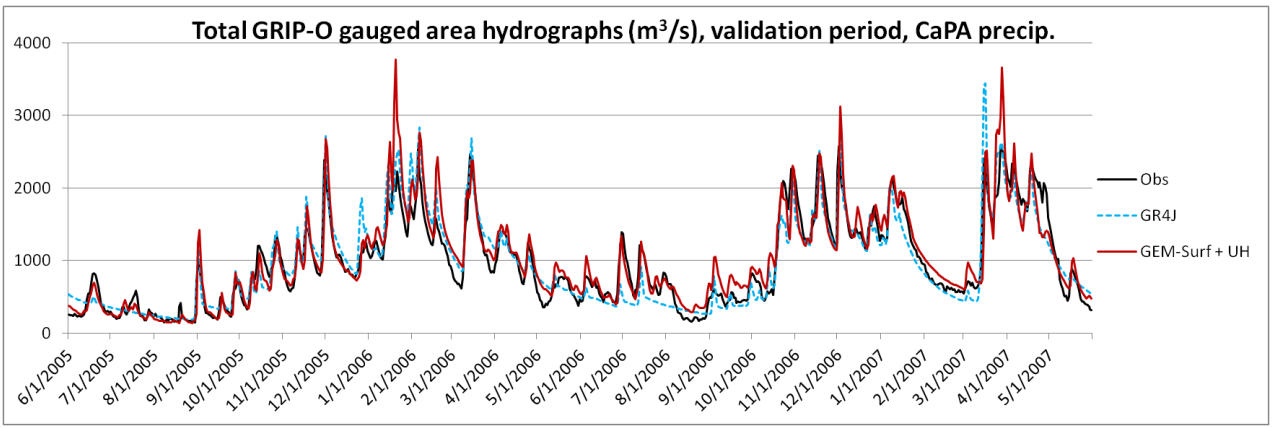

Figure 46: Hydrographs for the total GRIP-O gauged area (Figure 44) in validation, with the unique GR4J or local GEM-Surf + UH models calibrated using global calibration.

Finally, runoff simulations for the total lake Ontario watershed, including its ungauged areas, are very promising in the sense that performances are very close to those obtained over the gauged area only. Even if runoff observations actually consist in this case in estimations based on the ARM, they are a priori reliable given that the true gauged fraction of the total area is equal to about 69%. GEM-Hydro (or GEM-Surf + UH) tends to overestimate streamflow total volumes, which is shown by the PBIAS criteria (Table 41), while GR4J achieves a better

estimation of total runoff volumes. The overestimation of GEM-Surf + UH is supposed to be (at least partly) due to the catchments including urban areas, for which performances are very poor and runoff volumes are underestimated with default parameter values (see for example catchments 13 to 15 on Figure 45). During global calibration, parameter values are hence influenced by these problematic catchments for which runoff has to be increased. The fact that GR4J is better than GEM-Surf + UH in terms of the overall runoff balance (PBIAS criteria) is expected to derive from the fact that GR4J here consists in a unique (global) model for the whole area considered, because performances with local GR4J models (see section 7) were poorer for the PBIAS criteria.

Therefore, the interest of global calibration has been demonstrated, because it allows to achieve satisfactory performances for a large area with a unique calibration, and leads to a strong model temporal robustness and a spatially-consistent, reliably transferable parameter set. This global calibration procedure of the SVS LSS is envisioned to be used in future versions of EC's Great lakes forecasting system. For very large areas such as the whole great lakes watershed however, it may be better to use several global calibrations, for example one for each of the great lakes' watersheds. Finally, an efficient and transferable methodology has been proposed to estimate runoff for the ungauged parts of a (partly) gauged watershed. It is true, however, that it does not solve the problem of runoff estimation for almost completely ungauged large areas. For this however, GEM-Hydro has proven able to produce decent, generally satisfactory runoff simulations with default parameter values, except for areas containing a fraction of urban cover. This drawback is under investigation and expected to derive from an underestimation of urban areas in the LULC dataset used (Globcover 2009, see Table 5) as well as from a too simple representation of urban areas in SVS (33% of urban areas are estimated as being completely impervious).

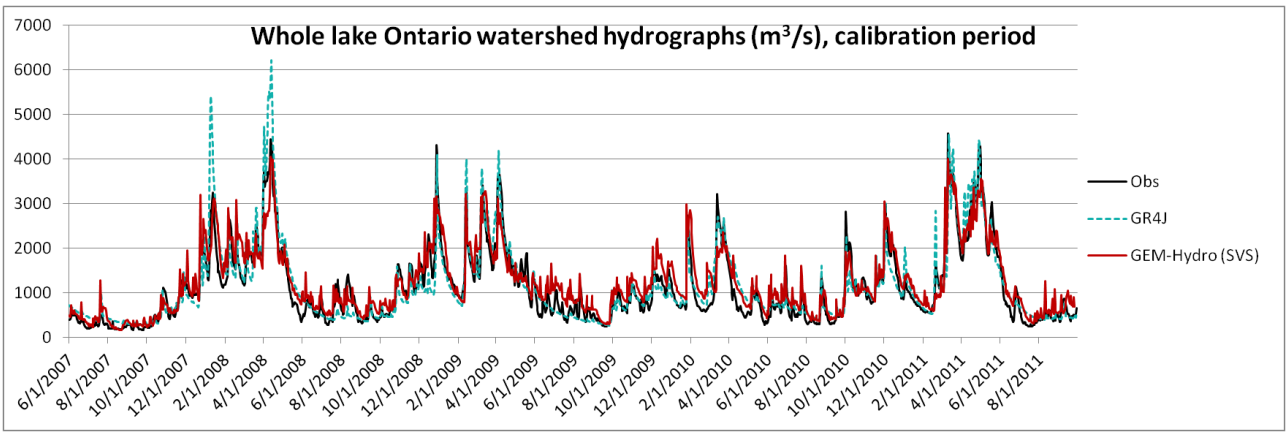

Figure 47: Hydrographs for the total lake Ontario watershed including the ungauged areas (Figure 44) over the calibration period, with the unique GR4J or the GEM-Hydro model. Parameters were obtained from calibration over the gauged area, and GEM-Hydro makes use of the Watroute default Manning coefficients defined in Table 40.

It is important to emphasize that for the total area including the ungauged parts, runoff was estimated with GEM-Hydro instead than with GEM-Surf + UH, which means that streamflow simulations are available everywhere inside the domain, whereas GR4J only delivers estimations for the total watershed area. Based on these results, it is therefore argued that the methodology proposed here (global calibration of GEM-Surf + UH and parameter transfer to GEM-Hydro) is relevant, efficient and reliable to perform runoff estimations for the ungauged parts of a large watershed, provided that a large enough fraction of the total area is gauged and that global calibration is employed. It could moreover be applied in different climatic contexts, regions, and with different models.

This work successfully led to the implementation of an efficient distributed hydrological modelling platform for the land part of the lake Ontario watershed, which has therefore become a readily testing ground for distributed models.

10- Performance synthesis

Table 42 and Table 43 below give a good overview of the performances of GR4J, LBRM and GEM-Surf + UH models, when calibrated with CaPA precipitation. The version of SVS used when calibrating GEM-Surf + UH is the last one at the time of writing, i.e. version " new_2" (see Table **29**). GEM-Surf + UH consists of a particular version of GEM-Hydro where the WATROUTE routing component was replaced by a very simple and time-efficient UH, which allowed to calibrate SVS parameters because of the strong amount of simulation time saved this way. Calibrated SVS parameters can then be transferred into GEM-Hydro afterwards, with using, in WATROUTE, the Manning coefficients obtained after manual tuning and which perform well for the Lake Ontario area (see Table **40**).

One has to keep in mind that the distributed models have more applications than the lumped ones. GEM-Hydro can for example predict lake levels, snow depth, water temperature, etc. In a general manner, the lumped models used during this work could be seen as useful tools allowing to set the target to try to reach with distributed models, in terms of hydrologic performances. Each of this family of models possesses its own advantages and drawbacks, and one cannot replace another. Despite its regulated regime, the lake Ontario watershed revealed that it was possible to have its runoff simulated with very satisfying performances. The distributed models consist of powerful tools with many potential applications, yet they cannot be exempted from calibration (yet) when it comes to daily hydrologic simulations. The work presented in this report has allowed to significantly improve the SVS LSS, yet much work remains to be done to further improve it. This work has also performed a comprehensive exploration of achievable hydrologic performances for the lake Ontario watershed. Finally, the interest of global calibration was demonstrated with GEM-Surf + UH, and was associated to a simple and reliable methodology (which can be transferred to any area of the world) in order to estimate runoff from ungauged areas (see section 9) of the whole Lake Ontario watershed.

Therefore, the authors are convinced that the main of the GRIP-O was fulfilled during this work, with the development of an efficient way to implement and optimize the GEM-Hydro platform to large areas, in view of many environmental applications, one of which will be its use with the Great Lakes' forecasting system currently developed at Environment Canada. This system consists of a two-way coupled Atmospheric, Surface and Hydrodynamic model of the whole Great Lakes system. The work achieved during GRIP-O is envisioned to significantly contribute to the amelioration of the performances of this system, and many more improvements are in progress for the SVS Land-Surface Scheme (see promising research directions in the next section).

Table 42: Inter-comparison of main model performances (Nash sqrt criteria) with CaPA precipitation. CAL, VAL: calibration and validation.

			CAL			VAL		
Subbasin	Station	Area(km2)	GR4J	LBRM	GEM-Surf+UH	GR4J	LBRM	GEM-Surf+UH
1.0	20_mile	307.0	79.7	70.6		80.8	65.8	
3.0	Genessee	6317.0	80.1	79.1	73.0	68.5	68.7	65.0
4bis	Irondequoit	326.0	68.6	70.2		67.4	63.2	
5.0	Oswego	13287.0	79.4	85.4	78.0	66.4	74.7	73.0
6.0	N/A	2406.0	70.5	68.3		77.4	77.3	
7.0	Black river	4847.0	72.5	73.9	72.0	73.7	71.5	66.0
8.0	Oswegatchie	2543.0	71.1	70.5	73.0	81.7	74.8	75.0
10.0	Salmon_CA	912.0	86.6	88.1	81.0	85.2	83.7	80.0
10bis	N/A	944.0	86.6	85.0		86.2	81.6	
11.0	Moira	2582.0	87.9	87.3	84.0	84.3	84.0	83.0
12.0	N/A	12515.5	69.8	72.6	75.0	77.1	77.8	67.0
13.0	N/A	1537.5	69.9	64.3		63.4	71.8	
14.0	N/A	2689.4	78.6	75.9		67.7	70.4	
15.0	N/A	2245.8	85.3	84.5		83.6	79.9	

Table 43: Inter-comparison of main model performances (PBIAS criteria) with CaPA precipitation. CAL, VAL: calibration and validation.

			CAL			VAL		
Subbasin	Station	Area(km2)	GR4J	LBRM	GEM-Surf+UH	GR4J	LBRM	GEM-Surf+UH
1.0	20_mile	307.0	13.4	16.3		4.3	13.3	
3.0	Genessee	6317.0	2.7	3.7	6.4	0.8	6.1	5.8
4bis	Irondequoit	326.0	5.3	2.0		-3.3	10.7	
5.0	Oswego	13287.0	1.6	1.3	-2.0	11.1	13.7	5.5
6.0	N/A	2406.0	4.0	6.6		7.5	11.3	
7.0	Black river	4847.0	1.9	3.3	6.0	10.5	11.5	16.8
8.0	Oswegatchie	2543.0	0.9	4.9	5.5	10.4	16.3	15.9
10.0	Salmon_CA	912.0	5.0	3.7	10.8	2.5	5.8	7.3
10bis	N/A	944.0	5.3	6.4		-1.3	6.2	
11.0	Moira	2582.0	3.4	3.8	10.0	1.3	8.9	9.4
12.0	N/A	12515.5	5.1	4.9	-9.0	-2.6	0.1	17.2
13.0	N/A	1537.5	2.4	5.3		14.6	5.1	
14.0	N/A	2689.4	5.0	5.2		19.4	14.3	
15.0	N/A	2245.8	1.2	3.2		-2.5	-1.8	

11- Future research directions

This work allowed to highlight some of the SVS weaknesses, and therefore interesting directions for further improving the model and, hopefully, render it even more competitive against the lumped models used during this study. Some of the main potential improvements to the SVS LSS are listed below.

1- using a gridded soil depth dataset in order to have a more realistic representation of the soil maximum depth in SVS, which is currently fixed to a given value for the area under study. The soil maximum depth plays a key role on the water balance, because a deeper soil will have a higher storage capacity and will typically generally be accompanied by less surface runoff. An experiment was made with such a SVS version in the Lake Ontario watershed, but did not significantly improve the results as the soil depth is generally very close to 1.4m in this area, and therefore the simulations were very similar between using spatially-varying or fixed soil depth values in SVS for this watershed, but may reveal interesting in different areas, such as those with thin soils which favor high runoff/rainfall ratios.

2- implementing a soil-freeze process allowing the frozen water to appear and develop in the soil during winter. This process is responsible for decreasing the soil hydraulic conductivity and making it behave more like an impervious layer, which also has a significant effect on surface runoff generation, especially during snowmelt periods. This is envisioned to let SVS be able to capture runoff peaks in spring which it currently misses (see Figure **43**).

3- implementing a subgrid-scale lake model in the WATROUTE routing scheme, which would allow representing the lake effect on hydrographs, but also a potential better coupling between the LSS and routing schemes, which would allow for a more realistic representation of evaporation for example, by linking the physical flux estimated with thermal equations, to physical quantities of water available to the evaporation process.

4- using the most detailed soil data in the region (about 1 arcsec. resolution, see former GRIPO report on available data) in conjunction with a (very) high resolution implementation of SPS+SVS (for example at 60 or even 30 arcsec.) and WATROUTE (1 sec), using GLAHF flow direction grids and land cover data (1 arcsec. resolution). This consists in a good opportunity to compare the performances as a function of the spatial resolution, with two very different resolutions (30/60 arcsec. vs 10 arcmin for SPS) for which soil data are available for both with a similar spatial resolution.

5- switching the SVS simulation paradigm from 1D to 2D, by making pixels able to interact with their neighbors, to transfer water horizontally between pixels in the LSS and not

only in the routing scheme. Many processes of the SVS LSS could benefit from a 2D representation.

6- Implementing a way to better represent urban areas in SVS, which is suited to vegetation covers but less so for urban areas. A simple test was made by increasing the amount of runoff produced in urban areas by associating a constant precipitation/runoff ratio over this type of land cover: generally, high density urban areas are attributed an average percentage of 30% of impervious surfaces, which was specified in SVS in the "new_2" version (see Table **2**). This simple modification however did not allow to significantly better simulations for urban catchments. Therefore, a more complex representation of urban areas may be beneficial to the SVS Land-Surface Scheme.

12- references

Adam, J. C., and Lettenmaier, D.P. 2003. Adjustment of global gridded precipitation for systematic bias. *J. Geophys. Res.*, 108, 4257, doi:10.1029/2002JD002499, D9.

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

Boone, A., and Etchevers, P. 2001. An intercomparison of three snow schemes of varying complexity coupled to the same land surface model: local-scale evaluation at an Alpine site. *J. Hydrometeor*, **2**: 374–394. Doi: http://dx.doi.org/10.1175/1525-7541(2001)002<0374:AIOTSS>2.0.CO;2

Croley, T. 1983. Great Lakes basins (U.S.A.-Canada) runoff modeling. *Journal of Hydrology*, 64: 135-158.

Croley, T.E., and 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.

DeMarchi, C., Dai, Q., Mello, M. E., and Hunter, T. S. 2009. Estimation of Overlake Precipitation and Basin Runoff Uncertainty. *International Upper Great lakes Study*.
FAO/IIASA/ISRIC/ISS-CAS/JRC, 2012. *Harmonized World Soil Database (version 1.2)*. FAO, Rome, Italy and IIASA, Laxenburg, Austria.

Fortin, V., Roy, G., and Donaldson, N. 2014. Assimilation of radar QPE in the Canadian Precipitation Analysis (CaPA). In *ASCE International Symposium on Weather Radar and Hydrology*.

Fry, L.M., Hunter, T.S., Phanikumar, M.S., Fortin, V., Gronewold, A.D. 2013. Identifying streamgage networks for maximizing the effectiveness of regional water balance modeling. *Water Resources Research* **49**, 2689-2700.
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). *Journal of Hydrology*, **519**(D), 3448-3465, ISSN 0022-1694, http://dx.doi.org/10.1016/j.jhydrol.2014.07.021.

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. *Canadian Journal of Earth Sciences*, **52**(11): 990-999. DOI: 10.1139/cjes-2015-0015

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

Haghnegahdar A., Tolson B.A., Davison B., Seglenieks F.R., Klyszejko E., Soulis E.D., Fortin V., Matott L.S. Submitted. Calibrating Environment Canada's MESH Modelling System over the Great Lakes Basin. *Submitted to Atmosphere-Ocean*.

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

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

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

Mahfouf, J.-F., Brasnett, B., and Gagnon, S. 2007. A Canadian precipitation analysis (CaPA) project: Description and preliminary results. *Atmosphere-Ocean*, **45**(1): 1-17. DOI: 10.3137/ao.v450101

Matott, L. S. 2005. *OSTRICH: An optimization software tool: Documentation and users guide.* University at Buffalo, Buffalo, NY.

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

National Operational Hydrologic Remote Sensing Center (NOHRSC). 2004. *Snow Data Assimilation System (SNODAS) Data Products at NSIDC*, Boulder, Colorado, USA: National Snow and Ice Data Center. http://dx.doi.org/10.7265/N5TB14TC

Nicolle, P., Ramos, M.H., Andréassian, V., and 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.

Pagano, T., Hapuarachchi, P., and 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., and Andréassian, V. 2003. Improvement of a parsimonious model for streamflow simulation. *Journal of Hydrology*, **279** (1-4): 275–289.

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

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

Tolson, B. A., and Shoemaker, C. A. 2007. Dynamically dimensioned search algorithm for computationally efficient watershed model calibration. *Water Resources Research,* **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.

You, J., Tarboton, D.G., and Luce, C.H. 2013. Modeling the snow surface temperature with a one-layer energy balance snow model. *Hydrol. Earth Syst. Sci. Discuss.*, **10**: 15071-15118

List of Tables:

List of Figures: