Journal of Hydrology 519 (2014) 3448–3465

Contents lists available at ScienceDirect

Journal of Hydrology

journal homepage: www.elsevier.com/locate/jhydrol

The Great Lakes Runoff Intercomparison Project Phase 1: Lake Michigan (GRIP-M)

Lauren M. Fry [a,b,*], Andrew D. Gronewold [b], Vincent Fortin [c], Steven Buan [d], Anne H. Clites [b], Carol Luukkonen [e], David Holtschlag [e], Laura Diamond [d], Timothy Hunter [b], Frank Seglenieks [f], Dorothy Durnford [g], Milena Dimitrijevic [c], Christopher Subich [c], Erika Klyszejko [h], Kandace Kea [i], Pedro Restrepo [d]

[a] Cooperative Institute for Limnology and Ecosystems Research, University of Michigan, 4840 S. State Rd., Ann Arbor, MI, USA
[b] NOAA, Great Lakes Environmental Research Laboratory, Ann Arbor, MI, USA
[c] Environmental Numerical Prediction Research Section, Environment Canada, Dorval, QC, Canada
[d] NOAA, National Weather Service, North Central River Forecast Center, Chanhassen, MN, USA
[e] U.S. Geological Survey, Michigan Water Science Center, Lansing, MI, USA
[f] Boundary Water Issues, Environment Canada, Burlington, ON, Canada
[g] Meteorological Service of Canada, Environment Canada, Dorval, QC, Canada
[h] Water Survey of Canada, Environment Canada, Ottawa, ON, Canada
[i] Department of Civil and Environmental Engineering, Howard University, Washington, DC, USA

ARTICLE INFO

Article history:
Received 11 December 2013
Received in revised form 3 July 2014
Accepted 12 July 2014
Available online 21 July 2014
This manuscript was handled by Konstantine P. Georgakakos, Editor-in-Chief, with the assistance of Alon Rimmer, Associate Editor

Keywords:
Large scale hydrologic modeling
Model comparison
Water budget
Laurentian Great Lakes

SUMMARY

We assembled and applied five models (one of which included three different configurations) to the Lake Michigan basin to improve our understanding of how differences in model skill at simulating total runoff to Lake Michigan relate to model structure, calibration protocol, model complexity, and assimilation (i.e. replacement of simulated discharge with discharge observations into historical simulations), and evaluate historical changes in runoff to Lake Michigan. We found that the performance among these models when simulating total runoff to the lake varied relatively little, despite variability in model structure, spatial representation, input data, and calibration protocol. Relatively simple empirical, assimilative models, including the National Oceanic and Atmospheric Administration (NOAA) Great Lakes Environmental Research Laboratory (GLERL) area ratio-based model (ARM) and the United States Geological Survey (USGS) Analysis of Flows in Networks of CHannels (AFINCH) model, represent efficient and effective approaches to propagating discharge observations into basin-wide (including gaged and ungaged areas) runoff estimates, and may offer an opportunity to improve predictive models for simulating runoff to the Great Lakes. Additionally, the intercomparison revealed that the median of the simulations from non-assimilative models agrees well with assimilative models, suggesting that using a combination of different methodologies may be an appropriate approach for estimating runoff into the Great Lakes. We then applied one assimilative model (ARM) to the Lake Michigan basin and found that there was persistent reduction in the amount of precipitation that becomes runoff following 1998, corresponding to a period of persistent low Lake Michigan water levels. The study was conducted as a first phase of the Great Lakes Runoff Intercomparison Project, a regional binational collaboration that aims to systematically and rigorously assess a variety of models currently used (or that could readily be adapted) to simulate basin-scale runoff to the North American Laurentian Great Lakes.

1. Introduction

Large-scale hydrologic models have historically been applied to a wide variety of freshwater resource and ecosystem management problems (e.g. Alcamo et al., 2003; Fekete et al., 2002; Tang et al., 2010). Specific examples range from quantifying interbasin water transfers (Hanasaki et al., 2010; Islam et al., 2007) and simulating impacts of human-induced stressors on global and regional water budgets (McCabe and Wolock, 2011; Nijssen et al., 2001a; Mao and Cherkauer, 2009), to modeling the atmospheric water balance (Xia et al., 2012), predicting soil moisture (Nijssen et al., 2001b) and crop irrigation demands (Wisser et al., 2008), simulating

* Corresponding author at: Cooperative Institute for Limnology and Ecosystems Research, University of Michigan, 4840 S. State Rd., Ann Arbor, MI, USA. Tel.: +1 734 741 2217.
E-mail address: lmfry@umich.edu (L.M. Fry).

http://dx.doi.org/10.1016/j.jhydrol.2014.07.021
0022-1694/

nutrient and sediment fluxes (Robertson and Saad, 2011; Seitzinger et al., 2010), and providing a streamflow boundary for coastal ocean models (Nakada et al., 2012). The geographic extent of these applications is equally broad. However, we find that the North American St. Lawrence River basin, while among the largest in the world (Fig. 1) and containing the world's largest system of lakes (i.e. the Laurentian Great Lakes), is the subject of relatively few basin-wide water budget modeling studies at a spatial resolution and across temporal scales suitable for addressing the challenges associated with managing this freshwater resource. In fact, the challenges facing the Great Lakes require an in-depth understanding, addressed in part by regional water budget modeling, of how the dynamics of the Great Lakes-St. Lawrence River basin's water budget impact Great Lakes water levels (Brinkmann, 2000), compliance with international water use agreements (i.e. the Great Lakes Compact), and the human, environmental, and economic well-being of North America (among other impacts, as described in Wilson and Carpenter (1999), Buttle et al. (2004), Millerd (2005)). For further discussion on Great Lakes basin water budget modeling and examples, see Coon et al. (2011), Lofgren et al. (2011).

1.1. Rationale for the Great Lakes Runoff Intercomparison Project (GRIP)

In recent years, the need to quantify each component of the lakes' net basin supply (i.e. runoff, over-lake precipitation, and over-lake evaporation) has gained significant attention due to recent persistent extreme low lake levels, especially in the Lake Michigan-Huron system (Gronewold and Stow, 2014). Quantifying the water balance and associated fluxes in the St. Lawrence basin (Fig. 1) is complicated by the Great Lakes themselves, which represent approximately 30% of the total basin area, and have a coastline of over 7000 km in the U.S. alone (Gronewold et al., 2013b). More specifically, modeling St. Lawrence River flows requires explicit computation of runoff, over-lake precipitation, and over-lake evaporation within the Great Lakes basin. Each of these three components is of roughly the same order of magnitude, and each is monitored at different (and, in the case of over-lake evaporation and over-lake precipitation, extremely coarse) spatial and temporal scales (for recent and historical perspectives, see Derecki (1976), Blanken et al. (2011), Holman et al. (2012)).

The uncertainty in Great Lakes basin runoff estimates and the corresponding uncertainty in the influence of runoff changes on water levels in the Great Lakes, the increasing number of runoff and water budget models being developed and applied within the Great Lakes basin, and the limited extent to which these models have been systematically evaluated and compared to one another, collectively underscore a need for Great Lakes basin runoff intercomparison studies (Gronewold et al., 2011; Coon et al., 2011; Lofgren et al., 2011). Gronewold and Fortin (2012) further emphasize the importance of improving Great Lakes basin-wide runoff estimates, and of maintaining the trajectory and momentum in regional collaborative research established during the recently-completed International Joint Commission (IJC) International Upper Great Lakes Study (IUGLS).

To address these needs, the Great Lakes Runoff Intercomparison Project (GRIP) was initiated to assess runoff models for simulation of runoff to the Great Lakes and advance the state-of-the-art in basin-scale hydrological modeling, beginning with assessing simulations of historical monthly runoff to Lake Michigan (GRIP-M). Because our study focuses on model skill for application to simulation of historical monthly runoff to Lake Michigan, the objectives differ from previous intercomparison studies, in that the skill of spatially and temporally aggregated monthly discharge is evaluated, in addition to performance at individual gages. Specifically, the objectives of GRIP-M are to compare historical runoff estimates from a group of models that are readily adaptable to Great Lakes basin-wide hydrological modeling, understand differences in data

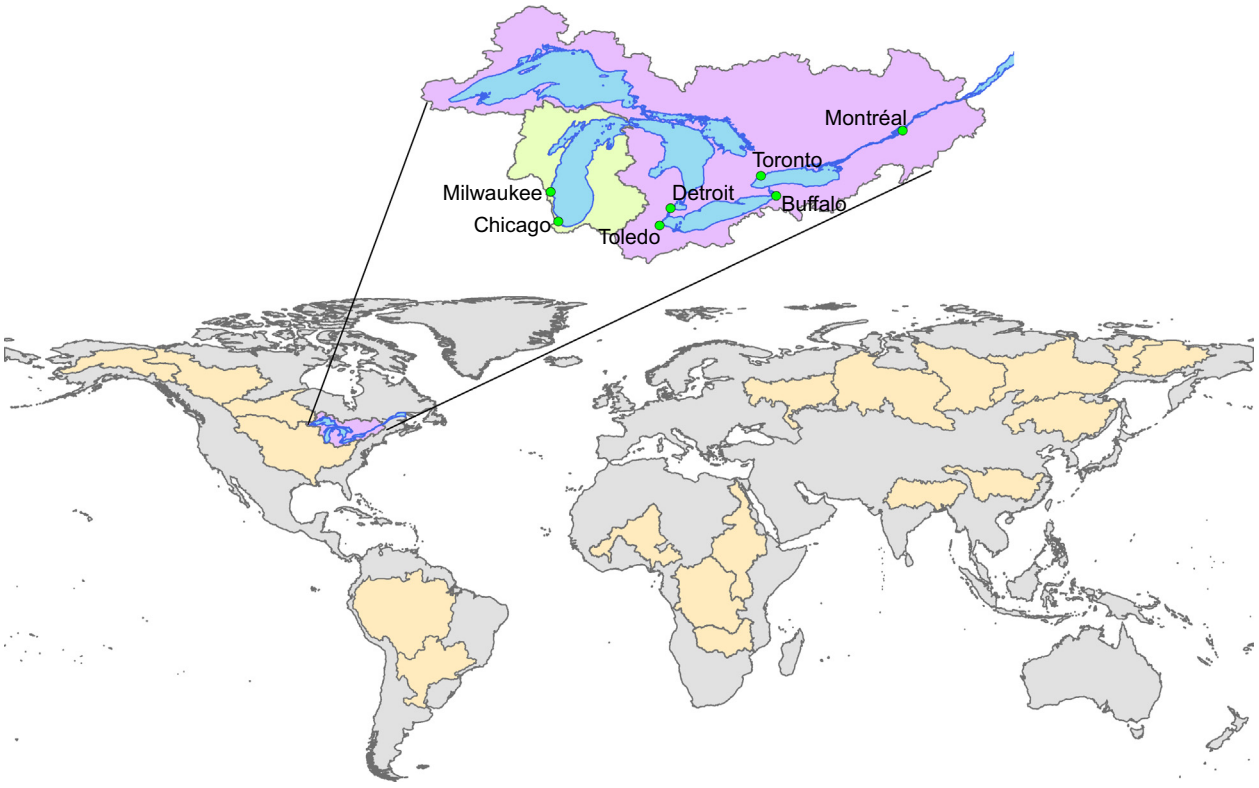

Fig. 1. Twenty largest (by basin surface area) river basins of the world, with detail of the Great Lakes – St. Lawrence River basin and the Lake Michigan basin. Derived from data provided by World Resources Institute (WRI) (2006).

requirements, simulation capabilities, and calibration procedures among those models, assess historical changes in the runoff component of the Lake Michigan water budget, and identify opportunities for improving simulations and forecasts of runoff to the Great Lakes.

An additional objective of GRIP-M (and future phases of GRIP) is to build on hydrological modeling studies from other regions (see, for example Smith et al., 2004; Smith et al., 2012; Reed et al., 2004) that have provided guidance to the hydrological modeling community on appropriate selection and application of runoff models across large spatial domains. In fact, the presence of the large lakes represents an opportunity for evaluating model skill at simulating continental-scale (i.e. gaged and ungaged area) runoff due to the fact that the net basin supply of a lake is, in effect, an observable response to runoff. Objectives of future phases of GRIP include expanding model simulations to other lake basins and verifying the runoff estimates using estimates of the other components of net basin supply and observed lake levels.

1.2. Evolution of Great Lakes basin-scale runoff modeling

Early advances in Great Lakes regional runoff modeling (Quinn, 1978; Derecki and Potok, 1979; Croley, 1983) were largely driven by a need to understand Great Lakes water level dynamics and how they impact regional commercial, industrial, and recreational activities (Millerd, 2011). More recent regional runoff modeling efforts (many of which are summarized in Coon et al. (2011)) are driven by a need to answer similar water resource management questions, yet we know of only one conceptual rainfall-runoff model, the National Oceanic and Atmospheric Administration (NOAA) Great Lakes Environmental Research Laboratory (GLERL) Large Basin Runoff Model (LBRM), that has been systematically applied to the *entire* Great Lakes basin. The NOAA-GLERL LBRM, while a critical component of research-oriented and operational Great Lakes water level forecasting systems (for details, see Gronewold et al., 2011), was developed and applied nearly two decades before the relatively recent surge in the advancement of methods for predicting flows in ungaged basins (i.e. as part of the PUB initiative, as described in Sivapalan et al. (2003), Wagener and Wheater (2006)). The LBRM is calibrated against synthetic discharge time series derived from a simple areal weighting scheme (hereafter referred to as the area ratio method, or ARM, and described in detail in Section 3.1).

While the ARM and LBRM have historically served as "golden standards" for Great Lakes regional hydrologic research and decision support systems, additional regional water budget models have been developed and applied recently, and collectively represent an opportunity to assess and advance the state-of-the-art in Great Lakes regional hydrological modeling (Gronewold and Fortin, 2012). One particular example is Environment Canada's (EC) *Modélisation Environmentale* – Surface Hydrology (MESH) model (Pietroniro et al., 2007; Deacu et al., 2012). MESH is an intriguing alternative to the relatively small suite of regional water budget models because its structure is different from LBRM (MESH is a distributed model, while LBRM is a lumped-parameter model), and because it explicitly propagates regional energy budget dynamics into the rainfall-runoff relationship. Similar regional models that have recently been applied to the Great Lakes basin include CHARM and REG-CM, as described in Lofgren (2004) and Holman et al. (2012).

Other modeling systems that have either been applied, or are expected to soon be applied, across relatively large spatial scales within the Great Lakes basin include the Precipitation-Runoff Modeling System (or PRMS, as described in Hunt, 2010), systems linking PRMS with the SPAtially Referenced Regressions On Watersheds model (or SPARROW; see Smith et al., 1997; McMahon et al., 2003, for details and recent applications), and similar systems linking SPARROW with the Water Availability Tool for Environmental Resources (WATER). The Analysis of Flows in Networks of CHannels (or AFINCH, as described in Holtschlag, 2009) is in a similar stage of development.

In consideration of the advancements in hydrological modeling over large spatial domains and the growing number of models that are applied in Great Lakes hydrology, there is a clear need to understand not only how the LBRM performs relative to alternative models on a subbasin scale, but also the implications (both in terms of potential skill as well as computational and operational effort) of expanding those alternative modeling systems (i.e. those currently applied at a sub-basin scale) across the entire basin of one or more of the Great Lakes.

In the following sections, we first (in Section 2) provide a description of the Lake Michigan basin, and then (in Section 3) identify and briefly describe the models selected for the GRIP-M study as well as our model confirmation and intercomparison procedures. We then (in Section 4) present and assess the study results, followed by concluding remarks including a synopsis of what we find constitutes the state-of-the-art in Great Lakes basin-wide hydrological modeling (Section 5).

2. The Lake Michigan basin

The Lake Michigan basin (shown in Fig. 1) was selected for the initial phase, in part, because it lies entirely within one country (the United States), and therefore most data sets required for GRIP-M (unlike data sets for basin-wide studies for the other Great Lakes) do not require coordination across the international border. The Lake Michigan basin covers 173,683 km^2, of which about 57,514 km^2 (33.10%) is covered by Lake Michigan. The runoff intercomparison considers only the land portion of the basin (i.e. runoff to the lake). Land use is dominated by urban areas in the southwest (including Milwaukee and Chicago metro areas), forest, wetlands, and open water in the north, and a mix of urban areas, agriculture, and natural areas throughout much of the southeast portion and central western portion. The land elevation ranges from 176 m at the coast to 600 m at the highest elevations in the northwest portion of the basin (north-central Wisconsin and the Upper Peninsula of Michigan). The mean slope of the basin's land area is about 2% (median 1.3%), with higher slopes in the northwest (northern Wisconsin) and northeast (northern lower Michigan) portions of the basin. Soil characteristics vary throughout the basin, with higher sand content in the northeast portion of the basin, and higher silt and clay content in the southern and western portions. Groundwater discharge comprises a large fraction of stream flow in the basin, with one estimate being that groundwater constitutes 79% of streamflow (USGS, 2011).

Mean annual air temperatures range from 3.4 °C to 11 °C, and annual precipitation ranges from about 700 mm to just over 1000 mm, estimated from 800 m PRISM data for the period of 1971–2000 (Oregon State University PRISM Group, 2008). The spatial variability in climate is dominated by north–south gradients, with generally decreasing temperatures from south to north. The basin is generally humid, with aridity index (mean annual precipitation divided by mean annual potential evapotranspiration) ranging from 0.8 to 1.1 (derived from data provided by the Consortium for Spatial Information and developed by Zomer et al. (2008)). The presence of Lake Michigan results in moderation of temperatures near the coastline and increased precipitation on the eastern side of Lake Michigan, with a larger percent of the precipitation falling as snow in the northern portions. The increased snowfall, both lake effect and non-lake effect snow, in the northern regions results in storage during the winter and large runoff ratios

(discharge/ precipitation) during the spring months. The seasonality of precipitation amount is moderate throughout the basin, with somewhat higher precipitation during summer months. However, this seasonality is less pronounced on the eastern and northern side of Lake Michigan. The median seasonality index, estimated using methods described by Dingman (2002) is 0.14 on the eastern side (0.11–0.18) and 0.20 on the western side (0.20–0.29) (as calculated by USGS (2011)). Changes in these drivers have the potential to influence each component of the net basin supply, pointing to a need to assess historical changes in runoff (in addition to over-lake evaporation and precipitation) relative to observed changes in lake levels.

3. Methods

Our evaluation of alternative models for simulating Lake Michigan basin runoff is divided into three steps. First, we identified a suite of models that are either currently implemented across the entire Lake Michigan basin or could easily be implemented across the entire basin with a reasonable amount of additional effort. Second, we assessed the relative skill of each model through a validation analysis at 20 streamflow gages distributed throughout the Lake Michigan basin. Third, we used each model to simulate runoff for the entire Lake Michigan basin. In the following sections, we describe each of these steps in greater detail.

3.1. Model package descriptions

We evaluated five model packages in GRIP-M, one of which was employed with three different configurations (for a total of seven separate model configurations, summarized in Table 1). Unlike model choices available for similar intercomparison studies, such as those described by Smith et al. (2004); Smith et al. (2012), Reed et al. (2004); and Breuer et al. (2009), relatively few candidate models were available for GRIP-M and, of those available, nearly all have a unique model structure, a unique spatial configuration (Fig. 2), and a unique set of forcing variables that need to be aggregated at a model-specific spatial and temporal resolution (Table 2). Furthermore, the candidate models for GRIP-M simulate flow over a range of time steps (e.g. sub-daily, daily, and monthly) and assimilate flow observations into flow simulation in different ways (if at all) (Table 3).

The models selected span a range of model structures and types, including (for details, see Dingman, 2002) "empirical" (models that determine parameter sets empirically and do not explicitly balance the water budget), "lumped conceptual" (models employing *a priori* relationships to simulate flows and storages and represent a watershed as a single unit), "distributed physical" (models that use physics-based equations to simulate flows and storages, and represent, to some degree, spatial variability within a region), and "assimilative" (models that use discharge measurements, when available, directly in historical flow simulations). It should be noted that, in this analysis, calibrated models do not account for the effects of nonstationarity (i.e. model parameters do not change over time), with the exception of ARM, for which the calibrated parameter runoff per unit area is estimated on a daily basis. Although we acknowledge the potentially significant impacts of this assumption of stationarity (Milly et al., 2008), our goal was to evaluate model configurations as they are traditionally applied in order to assess the current state-of-the-art in Great Lakes hydrological modeling.

In this phase of GRIP, we did not enforce controls on model structure, parameter estimation procedures, or forcings, as others have done (Smith et al., 2012; Clark et al., 2008; Duan et al., 2006). Recognizing the limitations that are associated with comparing models with different structures, calibration periods, forcings, and assimilative capability, we limit our intercomparison to the evaluation of model performance for simulating historical runoff to Lake Michigan. The intercomparison of simulated historical runoff provides a basis for choosing/using appropriate models for historical simulation in the context of understanding past changes in the basin-scale hydrological cycle and providing potentially useful synthetic records for use in development of predictive models. The systematic intercomparison of models that can be used to provide a historical monthly time series of runoff to the Great Lakes is a major contribution to advancing Great Lakes hydrological modeling, where ARM (a model that has not previously been systematically evaluated) has remained the only record of historical runoff to the Great Lakes, providing estimates of runoff to Lake Michigan from 1901 to present (the record goes as far back as 1898 for the Lake Erie basin).

Table 1
Summary of models evaluated in the GRIP-M project including model type, number of parameters that require calibration, number and average size of spatial units in the Lake Michigan basin, the time step of model simulations, and predictive capability. Note that for the NWS model, in ungaged basins, the estimated parameters were actually transferred from other, gaged, basins.

	Model type				Number of calibrated parameters *per spatial unit*	Number of spatial units in the Lake Michigan Basin	Average size of spatial unit (km^2)	Total number of parameters in the Lake Michigan Basin	Simulation timestep	Predictive capability
	Empirical	Lumped conceptual	Distributed physical	Assimilative						
ARM	X			X	1	27	4307	27	Daily	No
AFINCH	X			X	Number of independent variables + 1 = 9 in this analysis	32,497	3	292,833[b]	Monthly	Not known[d]
LBRM		X			9	27	4307	243	Daily	Yes
NWS		X			29	211	560	6119[c]	6-hourly	Yes
MESH-SA			X		64	480	250	30,720[b]	Hourly	Yes
MESH-5			X		64[a]	480	250	30,720[a]	Hourly	Yes
MESH-6			X	X	64[a]	480	250	30,720[a]	Hourly	Yes[e]

[a] Parameters for MESH-5 and MESH-6 are not calibrated.
[b] Parameters are identical across all spatial units in the basin for AFINCH and MESH.
[c] Although there are 211 NWS spatial units, only 75% of the basin area is calibrated, and this area includes subbasins contributing to other, calibrated subbasins. For these contributing subbasins, the parameters are the same as those of the calibrated subbasins. In the remaining 25% of the basin, parameters are transferred from other, calibrated subbasins.
[d] Skill would likely degrade, as there would be no assimilation of discharge observations.
[e] Note that without assimilation of discharge observations, MESH-6 is identical to MESH-5, so predictive modeling using MESH-6 would essentially be the same as predictive modeling using MESH-5.

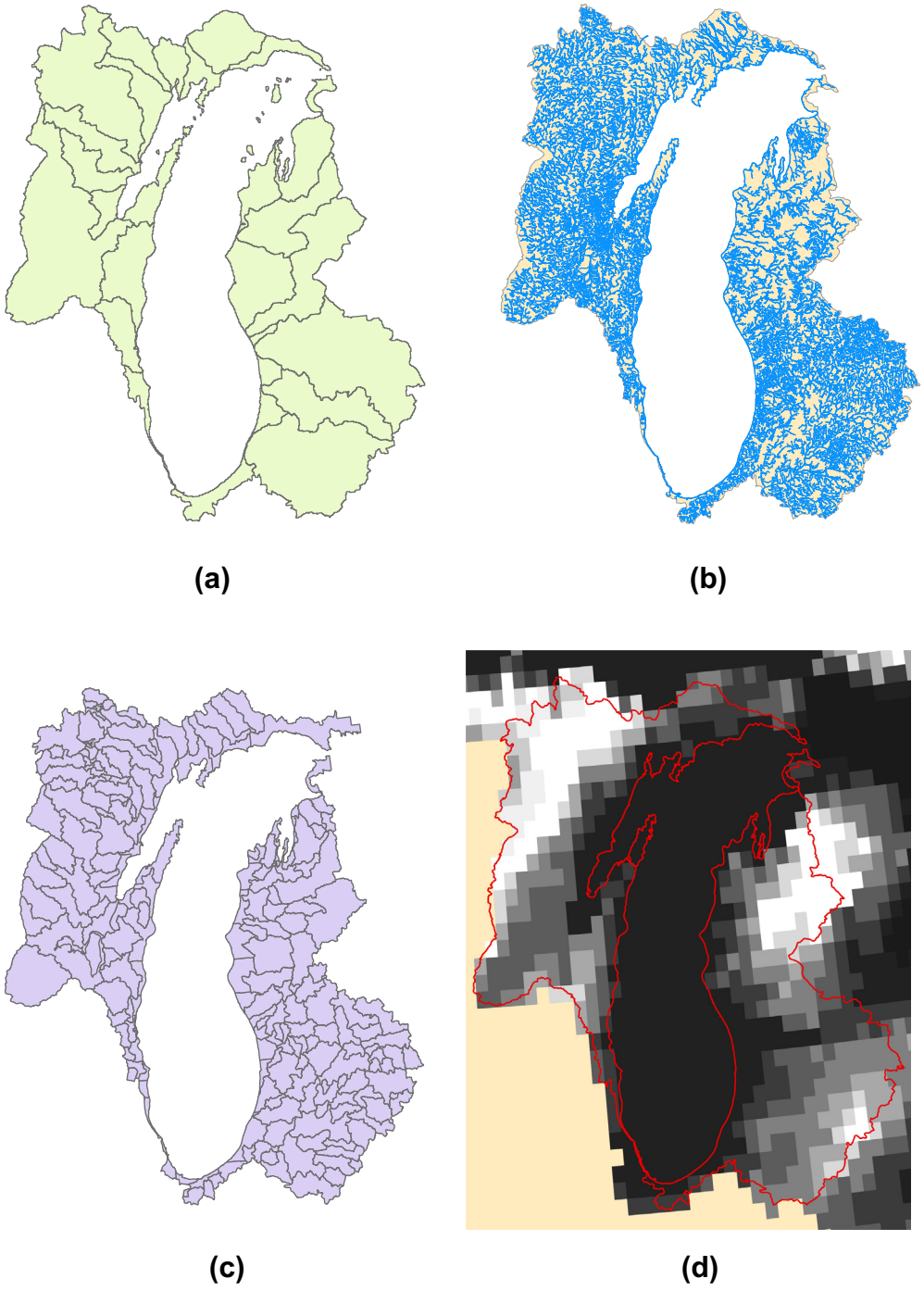

Fig. 2. Spatial frameworks for each of the models in GRIP-M; (a) ARM and LBRM subbasins, (b) NHDplus flowlines (McKay et al., 2012) used for AFINCH, (c) NWS hydrologic units, and (d) resolution of the WATROUTE grid used for all three MESH configurations (Kouwen, 2010), illustrated by the elevation field used to determine the flow direction with the 8 direction (D8) approach.

In the following sections, we briefly describe each model package (citing literature providing in-depth descriptions), including a general overview of model structure and spatial configuration, calibration routines, and historical application for Great Lakes hydrology. The first column in Tables 2 and 3 describes the meteorological forcings and streamflow observations, respectively, that are used in model packages for typical applications in Great Lakes basin hydrology.

3.1.1. The area ratio method (ARM)

The ARM is an empirical regionalization method with a single model parameter; runoff per unit area. Here, we employ NOAA-GLERL's implementation of ARM within the Great Lakes basin. The NOAA-GLERL implementation of ARM is based on dividing the entire Great Lakes basin into 121 subbasins, 27 of which constitute the Lake Michigan basin (Fig. 2(a)). Daily flow within each subbasin is estimated through a two-step procedure. First, daily flow is estimated in subbasins for which at least one daily flow measurement is available from among a set of pre-selected gages (373 gages throughout the Great Lakes basin; 81 of these are in the Lake Michigan basin). Daily flow in these "gaged" subbasins is estimated by calculating the average runoff-to-area ratio within the gaged portion of the subbasin and applying it to the entire subbasin. In effect, the process of calculating the average

Table 2
Summary of meteorological data employed for each model in the traditional application of the model package, the GRIP-M validation analysis, and the GRIP-M simulation of runoff from the Lake Michigan basin.

Model	Traditional package	Validation analysis	Simulation of basin runoff
ARM	None		
AFINCH	Monthly precipitation and temperature from PRISM (Oregon State University PRISM Group, 2008); 1970–2010		
LBRM	Daily observed precipitation and temperature interpolated to subbasins from NOAA GLERL (Hunter and Croley, 1993)	No changes to historical calibration, and simulation using daily precipitation and temperature from NOAA GLERL (Hunter and Croley, 1993) from 2002 to 2010, interpolated to validation gage drainage areas	Daily precipitation and temperature interpolated to subbasins from NOAA GLERL (Hunter and Croley, 1993) from 1948 to 2010
NWS	Daily precipitation and temperature interpolated to NWS subbasins from NOAA NWS (Anderson, 2002) from 1990 to 1999	No changes to historical calibration, and simulation using daily precipitation and temperature from NOAA NWS (Anderson, 2002) from 2002 to 2010, interpolated to validation gage drainage areas	Daily precipitation and temperature interpolated to NWS subbasins from NOAA NWS (Anderson, 2002) from 1948 to 2010
MESH-SA	6-hourly precipitation from CaPA (Mahfouf et al., 2007) disaggregated to hourly and interpolated to 10 arcmin horizontal resolution, and other variables from 15 km resolution GEM (Mailhot et al., 2006) interpolated to 10 arcmin resolution; 2005–2009		
MESH-5	NA	Multiple variables from 15 km resolution GEM (Mailhot et al., 2006) interpolated to 10 arcmin resolution; 2005–2009	
MESH-6	NA	Multiple variables from 15 km resolution GEM (Mailhot et al., 2006) interpolated to 10 arcmin resolution; 2005–2009	

Table 3
Summary of flow data employed for each model in the traditional application of the model package, the GRIP-M validation analysis, and the GRIP-M simulation of runoff from the Lake Michigan basin.

Model	Traditional package	Validation analysis	Simulation of basin runoff
ARM	Daily discharge at 81 preselected (by NOAA-GLERL) USGS gages	Repeat conventional calibration and simulation excluding 20 validation gages (and their downstream gages) for the period 2002–2010	Daily discharge at 81 preselected (by NOAA-GLERL) USGS gages
AFINCH	Daily discharge at all available USGS gages with continuous monthly records, aggregated to monthly discharge, from 1970 to 2010	Repeat calibration and simulation procedure excluding measurements from 20 validation gages from 2002 to 2010	Daily discharge at all available USGS gages with continuous monthly records, aggregated to monthly discharge, from 1970 to 2010
LBRM	Calibrated to synthetic subbasin daily runoff based on ARM. No flow data used in simulation	None (measurements from 20 validation gages between 2002 and 2010 were not used in the original calibration)	None
NWS	Calibration used all available USGS gages for 5-year period 1990–1999 (exact period varied by site). No flow data used in simulation	None (measurements from 20 validation gages between 2002 and 2010 were not used in the original calibration)	None
MESH-SA	Calibrated to daily flow observations from October 2004 to September 2005 at 11 locations across the Great Lakes watershed. No flow data used in simulation	None (measurements from 18 of the 20 validation gages between 2002 and 2010 were not used in the original calibration)	None
MESH-5	None		
MESH-6	No flow data used for calibration. All USGS gages having an area greater than 1000 km^2 are used in simulation	All USGS gages having an area greater than 1000 km^2 are used in simulation (excluding the 20 validation gages)	All USGS gages having an area greater than 1000 km^2 are used in simulation

runoff-to-area ratio for gaged portions constitutes model calibration, and application to ungaged portions constitutes simulation. In the second step, daily flow is simulated in subbasins for which (on a specific day) none of the pre-selected gages are "on-line" using the average runoff-to-area ratio from all gaged subbasins in the lake basin. Because streamflow observations are used during the simulation period, ARM is considered an assimilative model.

Because they provide a relatively long historical record, the NOAA-GLERL ARM runoff estimates often serve as a benchmark for comparing alternative runoff models, and have (as described in the next section) been used as a synthetic runoff time series for calibrating conceptual rainfall-runoff models (Croley and Hartmann, 1986; Croley and He, 2002). Although the NOAA-GLERL implementation of ARM serves as a cornerstone for Great Lakes basin hydrological studies and has recently been shown to provide reasonable simulations in individual subbasins when evaluated using several goodness-of-fit metrics (Nash–Sutcliffe efficiency, RMSE observations standard deviation ratio, and long term percent bias) (Fry et al., 2013), it has not yet been systematically evaluated for the entire Great Lakes basin (or any of the individual lake basins).

3.1.2. Analysis of Flows in Networks of CHannels (AFINCH)

AFINCH is an empirical modeling package in which catchment landscape characteristics and monthly climatic data are related to extended times series of monthly water yields computed from

measured flows at active streamgages (for details see Holtschlag, 2009). The package is configured to develop step-wise regression models for catchments defined in the National Hydrography Dataset (NHDPlus v.2) (McKay et al., 2012). The resulting water yields are multiplied by the drainage areas of the corresponding catchments to estimate monthly flows. Flows from catchments are accumulated downstream through the streamflow network described by the NHDPlus v.2 stream segments, called flowlines (shown in Fig. 2(b)). Although AFINCH is configured to calibrate stationary model parameters for GRIP-M, we note that AFINCH is able to accommodate effects of nonstationarity if parameter values are calibrated on an annual basis. AFINCH monthly flows can be constrained to match measured flows on stream segments with active streamgages, making it an assimilative model in this case. Specifically, ratios of measured flows to accumulated flows, based on regression estimates, are applied to upstream water yields to provide the constraint. It should be noted, however, that AFINCH does not require measured flow for simulation, and could therefore be used in a non-assimilative form. In GRIP-M, we consider the assimilative form.

The AFINCH package was developed as part of the Great Lakes Basin Pilot Project of the National Water Availability and Use Program (Holtschlag, 2009). Previous research confirms that AFINCH simulations are generally within 10% of measured monthly flows in networks with streamgage densities comparable to those operated in Ohio (Koltun and Holtschlag, 2010). AFINCH represents a potentially appealing alternative to similar regression-based hydrological models (see, for example Kokkonen et al., 2003; Oudin et al., 2008; Reichl et al., 2009), because of its unique capabilities to (1) integrate time varying climatological, streamflow, and water-use data, with user-specified land-use/land-cover information, (2) simulate nonstationary flow series over extended time intervals at regional scales for relatively low costs, and (3) produce monthly flows that are consistent with measured flow at active streamgages.

3.1.3. Large Basin Runoff Model (LBRM)

The LBRM is a lumped parameter conceptual rainfall-runoff model, originally developed by NOAA-GLERL, that simulates water transport through a series of cascading tanks (for a detailed description of the model, see Croley and He, 2002). It is applied across the same spatial subbasin configuration as the NOAA-GLERL implementation of ARM (Fig. 2(a)), and 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).

The NOAA-GLERL configuration of the LBRM uses parameter estimates from Croley (1983) that were determined using a grid search and minimizing the root mean square error based on the difference between LBRM-simulated subbasin runoff and ARM-derived runoff for each sub-basin (for details, see Croley and He, 2002). The model calibration and simulation both use daily climatological data interpolated to subbasin averages using the Thiessen polygon method.

3.1.4. NOAA National Weather Service (NWS) Model

The NOAA National Weather Service model selected for GRIP-M is a combination of two lumped conceptual models: the Sacramento Soil Moisture Accounting Model (SAC-SMA, described by Burnash, 1995) and Snow-17 model (described by Anderson, 1976) (hereafter collectively referred to as the "NWS" model). SAC-SMA estimates streamflow by representing water transfer among a set of storages linked by processes allowing approximation of soil moisture conditions controlling streamflow generation. SNOW-17 is an index model that uses temperature to determine the energy exchange across the snow–air interface. The NWS model evaluated in GRIP-M is implemented in a lumped configuration, with spatial units shown in Fig. 2(c), although a distributed version does exist (see Koren et al., 2004, 2006). NWS calibration procedures are described in detail by Anderson (2002).

The NWS model was implemented within the Community Hydrologic Prediction System (CHPS), a comprehensive modeling infrastructure employed in operational forecasting by (among other groups, and in other applications) NOAA's National Weather Service (NWS) as part of the Advanced Weather Interactive Processing System (AWIPS). CHPS provides a basis for sharing new and existing models with the broader hydrologic community (Roe et al., 2010). Data import, storage, and display are provided by the Delft Flood Early Warning System or FEWS (as described in Werner et al., 2013).

3.1.5. Modélisation Environmentale – Surface Hydrology (MESH) model

The *Modélisation Environmentale* – Surface Hydrology (MESH) model (Pietroniro et al., 2007; Deacu et al., 2012) is a distributed physical model developed by Environment Canada that combines a land surface model with a hydrologic routing module to simulate runoff to the Great Lakes from both the U.S. and Canadian portions of the basin. MESH land surface parameters are not specific to a watershed, but rather are based on landscape units (also known as Grouped Response Units, or GRUs), allowing them to be transferred to ungaged basins using a vegetation classification procedure.

Two different land surface models are available within MESH to represent the exchange of momentum, heat, and moisture between the atmosphere and the land surface: (1) a Canadian version of the ISBA (Interaction Sol-Biosphère-Atmosphère) model (Bélair et al., 2003), which was designed for numerical weather prediction, and has been embedded within the Global Environmental Multiscale (GEM) Regional model since 2001 (Bélair et al., 2003) and (2) the Canadian Land Surface Scheme (CLASS), which was designed for climate prediction and is available both within GEM and as a standalone model (Verseghy, 2000). All configurations of MESH use the WATROUTE routing model (Kouwen, 2010). A spatial resolution of 10 arcmin is used in this paper both for the land surface model and the routing component (Fig. 2(d)).

Previous work by Deacu et al. (2012) demonstrated the potential of various configurations of MESH for use in simulating Great Lakes net basin supply (over-lake precipitation minus over-lake evaporation plus runoff to the lakes). The two best performing configurations, which rely on the ISBA land surface model, are considered in the GRIP-M analysis (MESH-5 and MESH-6), in addition to the standalone version (MESH-SA) which relies on CLASS.

MESH-SA is calibrated using observed streamflow at eleven streamgages throughout the Great Lakes basin (five of which are located within the Lake Michigan basin). Calibration methods are described in Haghnegahdar et al. (in press). No calibration on streamflow is performed for MESH-5 or MESH-6; default model parameters are used, tuned against surface meteorological observations. However, MESH-6 replaces simulated streamflow with observed discharge when observations are available within a grid cell, making MESH-6 an assimilative model.

3.2. Validation of model packages

To assess each model package's skill at simulating discharge, we conducted a validation analysis at 20 gages which were removed from any calibration or simulation processes during the common validation time period (2004–2008) (gage locations are shown as red circles in Fig. 5). Calibration period, and whether or not the model was calibrated for this validation analysis, varied by model

package. For model packages whose original calibration routine did not use observations from the 20 validation gages during the validation period (2004–2008), re-calibration was not necessary for the validation analysis (i.e. LBRM, which was originally calibrated to 30 years of data prior to 1980, and NWS, which was originally calibrated to observations from 1990–1999). MESH-SA was calibrated using observed streamflow at 11 stations across the Great Lakes from October 2004 to September 2005 (Haghnegahdar et al., in press). The set of stations used for calibration includes two of the stations used for verification in this paper (04121970 and 05067500). Results from these two stations were not considered for validation of MESH-SA. MESH-5 and MESH-6 are not calibrated. ARM is calibrated at gaged locations during the simulation period, resulting in the estimated parameter of discharge per unit area (i.e. the model simulates discharge only in the ungaged areas). For the validation analysis, the 20 validation gages were not used in ARM calibration. For the validation analysis, AFINCH was calibrated for the period of 2001–2010, removing the 20 validation gages from the set of observations used to develop the regression equations.

Gages were selected according to the following criteria: (1) availability of a continuous daily record during the validation period, (2) relatively even distribution along the Lake Michigan coast, and (3) within 20 km of the Lake Michigan shore. Note that two gages (04121970 and 04084445) do not meet the 3rd criterion. These two gages were added in consideration of the need to represent larger watersheds within the basin. Observed discharge at these 20 gages varied from quite small maximum monthly discharge (less than 10 cms) to very large (nearly 450 cms) (examples are shown in Fig. 3). The strong seasonal cycle resulting from spring snowmelt is evident in each of the gages' time series. However the relative magnitude of these spring runoff events varies across the basin, with larger spring runoff (relative to average annual runoff) in the northern portions of the basin.

For assimilative models (ARM, AFINCH, and MESH-6), no observations from the 20 validation gages were assimilated into the simulations for the validation analysis. Meteorological data used to produce the simulations for the validation analysis are shown in Table 2. The stream discharge data used for model calibration and assimilation (if applicable) for the validation analysis are shown in Table 3.

With the exceptions of ARM and LBRM, all model packages are configurable to provide estimates at gage locations. Because ARM and LBRM are configured to output discharge only at subbasin outlets, simulations at gages were calculated by interpolating subbasin-simulated flow to gage locations. In both cases, this was done by simply multiplying the subbasin-simulated flow by the ratio of the validation gage's catchment area to the subbasin area. This interpolation method is appropriate, because both ARM and LBRM operate under the assumption that parameters are homogeneous across each subbasin.

To assess model skill at simulating historical runoff, we evaluated Nash–Sutcliffe efficiency (NSE) Eq. (1) and Percent Bias (PBIAS) Eq. (2). The wide-spread use of NSE values throughout the hydrological modeling literature facilitates the comparison with other published studies (for example, as was done in the recent meta-analysis comparing methods of prediction in ungaged basins by Parajka et al. (2013)). We include PBIAS for model evaluation to determine whether the models generally over- or under-predict runoff, which is important for understanding implications on lake levels. Additionally, we evaluated the spearman rank correlation coefficient, ρ Eq. (3), to eliminate the influence of bias. To a

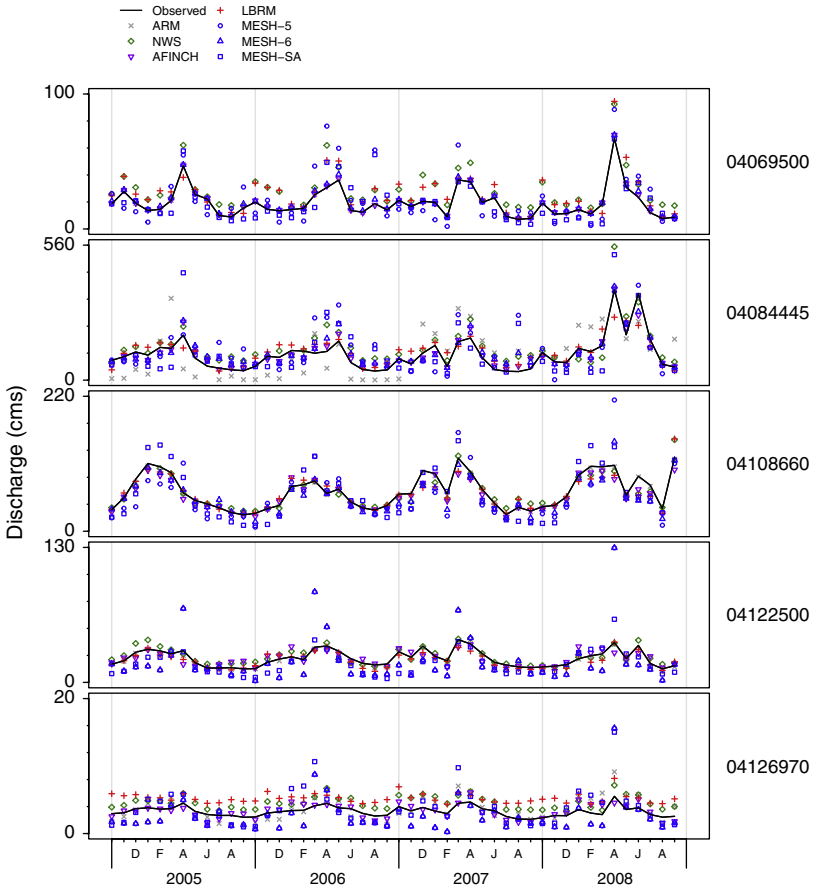

Fig. 3. Representative time series of simulated and observed monthly discharge for 5 of the 20 gaging stations shown in Fig. 5.

degree, ρ can indicate how well seasonality is represented by the simulated monthly flows, with ρ values close to one indicating that the modeled values are ranked similarly to the observed values, suggesting that modeled high flows occur during the same months as observed high flows, and modeled low flows occur during the same months as observed low flows. Additionally, we compared the total modeled discharge with total observed discharge from the 17 gages for which all models produced simulations to evaluate the effect of spatial aggregation and provide a comparison with between-model biases evident from the simulations of total Lake Michigan inflow.

$$NSE = 1 - \frac{\sum_{i=1}^{n}(O_i - P_i)^2}{\sum_{i=1}^{n}(O_i - \overline{O})^2} \quad (1)$$

$$PBIAS = \frac{\sum_{i=1}^{n}(O_i - P_i) * 100}{\sum_{i=1}^{n} O_i} \quad (2)$$

$$\rho = \frac{\sum_{i=1}^{n}(o_i - \bar{o})(p_i - \bar{p})}{\sqrt{\sum_{i=1}^{n}(o_i - \bar{o})^2 \sum_{i=1}^{n}(p_i - \bar{p})^2}} \quad (3)$$

In Eqs. (1) and (2) validation goodness-of-fit statistics, O_i is the observed value, P_i is the simulated value, and n is the number of simulations. In Eq. (3), o_i and p_i are the ranked observed and predicted values.

3.3. Simulation of Lake Michigan basin runoff

Contributors to GRIP-M then provided the best possible estimates of runoff to Lake Michigan with minimal changes to the existing model package (i.e. forcings, parameterizations, and assimilated streamflow observations, if applicable). Comparing model packages in their existing state provides a basis for evaluating the feasibility of adapting each package for use in Great Lakes hydrological modeling. Because ARM, LBRM, and all configurations of MESH were configured for the purpose of basin-wide runoff estimation, no changes were necessary to their configurations. AFINCH had not previously been applied to the entire Lake Michigan basin, so new regression models were developed for this analysis (independent variables incorporated into the AFINCH regression analysis in both cases are shown in Table 4). The NWS model had not previously been applied to ungaged portions of the Lake Michigan basin; for GRIP-M, parameters were assigned at ungaged hydrologic units based on landscape characteristics and proximity. Meteorological forcings and streamflow observations (if applicable) used by each model package in this basin-wide runoff simulation phase are shown in the third column of Tables 2 and 3.

Each model package simulated runoff from the entire Lake Michigan basin for as many years as possible (depending on the data requirements of the model package) for the period from 1950 to 2010. We then compared cumulative monthly basin-wide runoff, expressed as a depth (in meters) of water over Lake Michigan, from all models from October 2005 through September 2008 (the period for which continuous simulations were available for all models). While the cumulative inflow cannot be interpreted directly as a change in water level without knowledge of the flow through interconnecting channels (between the Great Lakes) and over-lake precipitation and evaporation, the cumulative inflow in depth over Lake Michigan does provide some indication as to potential relative differences in the simulated water balance among models. This comparison allows us to assess the relative impact of model bias over time and how it might impact regional water balance estimates and water level forecasts.

Table 4
Variables incorporated into AFINCH regression analysis.

Variable	Source
Monthly total precipitation	PRISM (Daly and Taylor, 1998b)
Preceding monthly total precipitation	PRISM (Daly and Taylor, 1998b)
Average monthly temperature	PRISM (Daly and Taylor, 1998a)
Percent woody wetlands	NLCD 1992 (Vogelmann et al., 2001)
Low/high intensity residential	NLCD 1992 (Vogelmann et al., 2001)
Mixed forest	NLCD 1992 (Vogelmann et al., 2001)
Pasture/hay/row crops	NLCD 1992 (Vogelmann et al., 2001)
Slope	NHDplus v2 (McKay et al., 2012)

4. Results and discussion

4.1. Validation of model packages

Evaluation of the skill of the models in representing monthly discharge at the validation gages considers how well the monthly hydrographs are represented (examples shown in Fig. 3), goodness of fit statistics in Eqs. (1)–(3) (Fig. 4 and Table 5), and the fit of modeled versus observed discharge to a one-to-one line (Fig. 5). The sum of the cumulative discharge across the validation gages (Fig. 6), as well as skill metrics for the aggregated discharge time series (Table 6), provide some insight into the aggregated impacts of the spatially varying skill of each model.

The median NSE, PBIAS, and ρ values for all model simulations at all validation watersheds was 0.45, 3.15, and 0.82, respectively. The NSE of 0.45 is within the range of NSE values reported in the meta-analysis described by Parajka et al. (2013), which compiled results from 34 studies covering 3,874 catchments worldwide, with NSE values ranging from 0.4 to 0.87. When spatially aggregated, the NSE values are better (median 0.76). Despite the small sample of validation gages and the limited common validation period, we attempt to rank models according to their performance in terms of NSE, PBIAS, and ρ, based on the results presented in Table 5. We propose a separate ranking for assimilative (ARM, AFINCH, MESH-6) and non-assimilative (NWS, LBRM, MESH-5, MESH-SA) models, as this constitutes an important difference in terms of model structure and application domain. Indeed, the degree of difficulty in predicting flow for assimilative methods depends on the percentage of the watershed which is gaged upstream of the validation point. This is not the case for non-assimilative methods. Hence, an overall ranking is not straightforward. Furthermore, non-assimilative methods, including a non-assimilative application of AFINCH which was not included in this analysis, can be used for applications such as forecasting, for which assimilative methods are not appropriate. It is therefore interesting to identify the best performing model in each category. For each of the three criteria presented in Table 5, we count the number of times that each model provides, for a given watershed, the best performance within its category. Only watersheds for which validation results are available for all models are considered (19 out of 20 gages for assimilative models, 17 out of 20 for non-assimilative models). Model rankings are summarized in Table 7.

Among assimilative models it can be seen that, for all three statistics, AFINCH outranks ARM and MESH-6 (Table 7). Among non-assimilative models, the NWS model outperforms both LBRM and MESH-5, but MESH-SA displays a lower bias than the NWS model. It is interesting to note that the two better performing models, AFINCH and NWS, are the only two models evaluated in this study which had never been considered before for estimating total runoff into Lake Michigan or into any of the Great Lakes.

Several gages stand out as problematic for all (or most) models in the validation analysis, most notably the three gages in northwest lower Michigan (04127800, 04126970, and 04126740). The annual runoff ratio (Runoff/ Precipitation) has been determined to be very near one (or greater) in this region, and the hydrographs show very little variability in monthly discharge (see, for example,

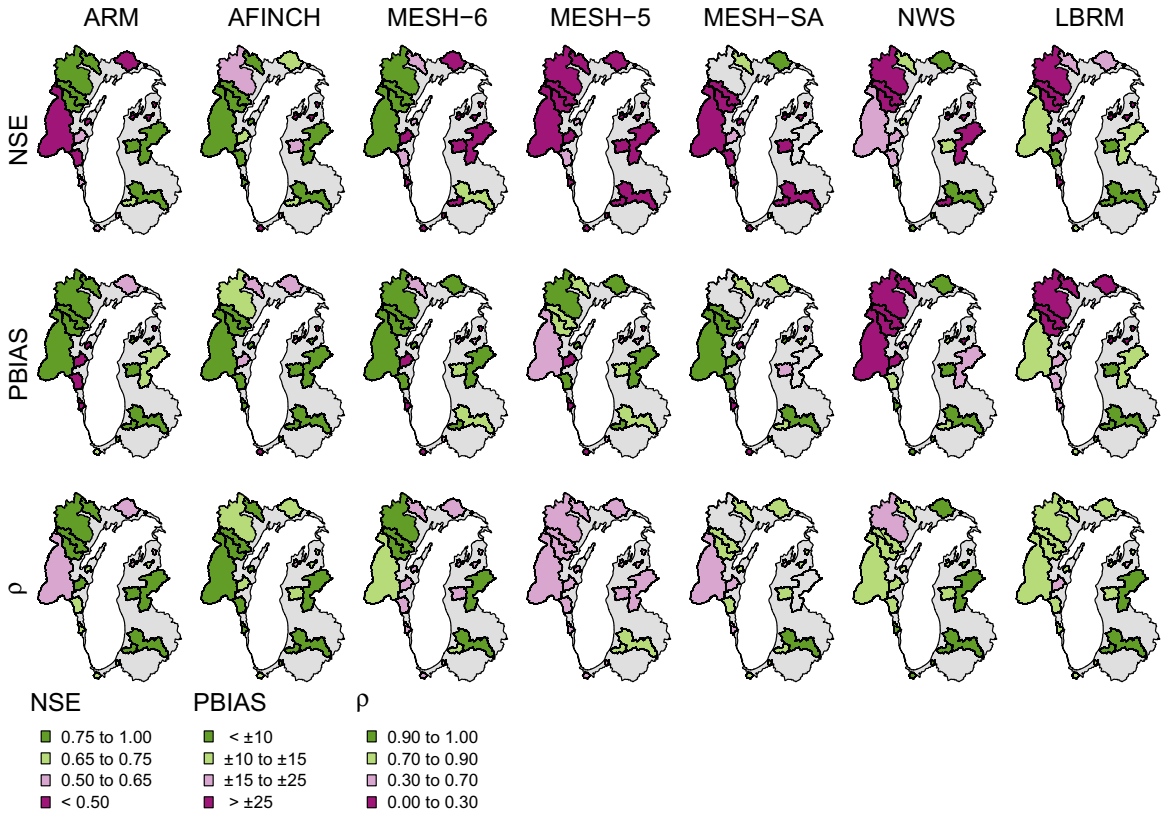

Fig. 4. Model validation goodness-of-fit statistics.

the monthly hydrograph of gage 04126970 in Fig. 3). In this region of Michigan, the hydrologic response is affected by highly permeable soils and low available water capacity, resulting in very high baseflow index (shown in Fig. 5) and little variability in monthly streamflow. In fact, these combined characteristics are distinct from the rest of the basin. This is a region with relatively few gages available for either calibration or assimilation into simulations, and therefore modeling in this region relies on information from other regions where the landscape characteristics result in very different hydrological response. In Fig. 5, we see that this model "ignorance" of locally unique hydrologic response at these three gages results in severe underrepresentation of low flows and overrepresentation of high flows, in a region where in reality there is relatively little variability in monthly flow throughout the year (as seen in the observations of flow at gage 04126970 in Fig. 3). The model that performs the best in this region, NWS, is the one model package for which parameterization incorporated substantial human judgement to improve results. In other regions of the basin, there was more variability in skill among the seven model packages. Sections 4.1.1 and 4.1.2 present results and discussion for validation of the assimilative and non-assimilative models, respectively.

4.1.1. Assimilative models

For all assimilative models, there was substantial improvement in validation watersheds for which at least one gage was assimilated into simulations. In these "partially gaged" locations, ARM simulations ranked highest, followed by AFINCH, then MESH-6; however, all models performed well. NSE values in partially gaged locations ranged from −0.26 to 0.99 (median 0.87), PBIAS values ranged from −19.5 to 47.8 (median 1.9), and ρ values ranged from 0.61 to 1.0 (median 0.93). Note that the large range in skill metrics is a result of very poor metrics of a single validation watershed, that of gage 04084445, Fox River at Appleton, WI. The poor performance at this watershed was a result of the application of the area ratio from a single gage (04085068, Ashwaubenon Creek near Little Rapids, WI) to the subbasin until September 30, 2006, at which point it was discontinued. The Aswaubenon Creek gage has a very small catchment area (52 km^2) relative to the Fox River gage (15,410 km^2). Model skill was much more variable in ungaged validation catchments. In these catchments, where no observations were assimilated into model simulations (indicated by superscripts a, b, and c, for ARM, MESH6, and AFINCH, respectively in Table 5), the NSE ranged from −107 to 0.80 (median 0.26), PBIAS ranged from −73.7 to 64.9 (median −5.5), and ρ ranged from 0.41 to 0.94 (median (0.66). All models performed very poorly (e.g. large negative NSE values) at the three gages in the northeast portion of the basin identified as problematic in the previous discussion (04126740, 04126970, and 04127800).

Among the five validation gages for which all models were considered to have ungaged catchments, AFINCH performance was higher for nearly all metrics, outperforming both ARM and MESH-6 at five, two, and five locations in terms of NSE, PBIAS, and ρ values, respectively (Table 5). The much larger gage network used to develop empirical relationships for AFINCH simulations (294 gages, compared to only 58 ARM gages, shown in Fig. 5) partially explains the significant improvement of AFINCH simulations over ARM. Additional improvements over ARM likely result from the inclusion of drivers of hydrologic response beyond simply watershed area (e.g. landscape variables and meteorological drivers) in the development of the regression equations. Despite better skill than other models, AFINCH performance is still poor in the northeast portion of the basin, reflecting the challenge of applying empirically based models developed over a large area to a specific region in which the hydrologic response is significantly different from the region in which the empirical relationships are derived. The overall skill of the physically based MESH-6 in ungaged watersheds was generally not better than that of the empirical models in the same ungaged watersheds. This may be a result of poorly

Table 5
Validation goodness-of-fit statistics.

Gage	NSE							PBIAS							ρ						
	Assimilative			Non-assimilative				Assimilative			Non-assimilative				Assimilative			Non-assimilative			
	ARM	AFINCH	MESH-6	MESH-5	MESH-SA	NWS	LBRM	ARM	AFINCH	MESH-6	MESH-5	MESH-SA	NWS	LBRM	ARM	AFINCH	MESH-6	MESH-5	MESH-SA	NWS	LBRM
04046000[a,b,c]	0.34	0.62	NA	NA	NA	NA	−0.09	13.30	36.10	NA	NA	NA	NA	78.20	0.55	0.91	NA	NA	NA	NA	0.91
04056500[a,b]	0.28	0.73	0.35	0.35	0.79	0.91	0.53	−21.90	−19.50	−7.60	−7.60	−13.90	0.20	28.10	0.56	0.88	0.41	0.41	0.77	0.91	0.76
04059000	0.91	0.80	0.55	0.29	0.75	0.72	0.51	4.90	22.50	15.20	13.80	10.30	29.60	32.70	0.92	0.91	0.62	0.58	0.80	0.78	0.90
04067500	0.99	0.60	0.93	−1.70	NA	−0.67	0.40	1.60	12.10	−9.70	7.20	NA	43.60	29.40	1.00	0.82	0.99	0.45	NA	0.62	0.83
04069500	0.99	0.99	0.97	−0.50	0.36	−0.02	0.17	3.10	3.60	6.80	13.80	−1.00	48.10	36.50	1.00	0.99	0.99	0.62	0.71	0.85	0.85
04071765	0.94	0.94	0.91	−0.11	0.13	0.34	0.39	9.10	8.30	−1.00	12.60	5.30	39.60	38.30	0.98	0.97	0.95	0.53	0.75	0.83	0.88
04084445	−0.26	0.88	0.78	0.00	−0.11	0.51	0.68	1.90	1.50	8.40	16.30	5.90	27.20	11.30	0.61	0.93	0.79	0.40	0.41	0.73	0.89
04085200[a,b,c]	0.31	0.80	0.51	0.51	0.38	0.68	0.60	64.90	8.60	3.20	3.20	−26.40	18.80	10.20	0.80	0.91	0.61	0.61	0.71	0.88	0.77
04085427[b]	0.51	0.74	0.37	0.37	0.58	0.59	0.50	47.80	15.20	39.70	39.70	6.80	32.10	22.50	0.93	0.89	0.46	0.46	0.63	0.88	0.70
04086600[b]	0.31	0.97	0.54	0.54	0.42	0.53	0.76	46.20	6.50	8.90	8.90	2.50	13.00	19.00	0.82	0.95	0.67	0.67	0.82	0.85	0.83
04087240[b]	0.63	0.92	0.21	0.21	0.24	0.86	−0.06	25.90	−7.40	−36.50	−36.50	−43.10	−1.20	24.90	0.87	0.99	0.66	0.66	0.69	0.92	0.76
04093000[b,c]	0.53	0.28	0.23	0.23	0.39	0.82	0.75	−11.30	−25.30	−34.40	−34.40	−30.80	−9.30	−12.10	0.80	0.83	0.66	0.66	0.74	0.92	0.82
04096015[b,c]	0.62	0.64	0.43	0.43	0.66	0.90	0.74	8.40	−4.90	−17.90	−17.90	−7.10	−0.50	7.50	0.88	0.94	0.69	0.69	0.82	0.95	0.86
04102500[b]	0.66	0.69	−0.38	−0.38	−0.13	0.08	0.78	−3.10	−9.80	−1.10	−1.10	10.90	13.10	4.30	0.90	0.90	0.71	0.71	0.86	0.93	0.93
04108660	0.99	0.87	0.71	0.05	0.46	0.88	0.83	−3.10	−9.10	−13.10	−11.40	−5.00	−2.40	−3.80	1.00	0.97	0.92	0.84	0.92	0.96	0.95
04121970	0.91	0.98	0.47	−3.86	NA	0.26	0.71	−10.00	1.80	−3.90	2.20	NA	24.50	12.50	0.99	0.98	0.95	0.61	NA	0.94	0.92
04122500[b,c]	0.92	0.57	−5.65	−5.65	−0.24	0.75	0.77	−6.10	−0.80	−11.60	−11.60	−21.70	7.50	−7.10	0.98	0.77	0.62	0.62	0.82	0.88	0.89
04126740[a,b,c]	−16.14	−8.02	−107.11	−107.11	−41.13	−2.11	−23.94	−6.10	16.30	−18.10	−18.10	−7.90	−10.80	42.10	0.51	0.65	0.59	0.59	0.54	0.82	0.79
04126970[a,b,c]	−2.06	0.07	−7.91	−7.91	−8.89	−2.60	−6.85	7.70	2.30	−20.30	−20.30	14.50	39.90	59.40	0.74	0.79	0.65	0.65	0.80	0.91	0.68
04127800[a,b,c]	−54.87	−9.62	−70.75	−70.75	−55.23	−0.37	−30.32	−66.50	−26.60	−73.70	−73.70	−65.30	8.80	−50.40	0.57	0.71	0.49	0.49	0.73	0.85	0.84
Median	0.58	0.74	0.43	0.00	0.36	0.53	0.52	2.50	2.05	−7.60	−1.10	−5.00	13.10	20.75	0.88	0.91	0.66	0.61	0.75	0.88	0.85

Note that no simulation was provided for gage 04046000 by NWS or MESH models, because its catchment was too small to model.

[a] Designates validation locations for which no upstream gages were assimilated into ARM simulations.

[b] Designates validation locations for which no upstream gages were assimilated into MESH-6 simulations.

[c] Designates validation locations for which no upstream gages were assimilated into AFINCH simulations.

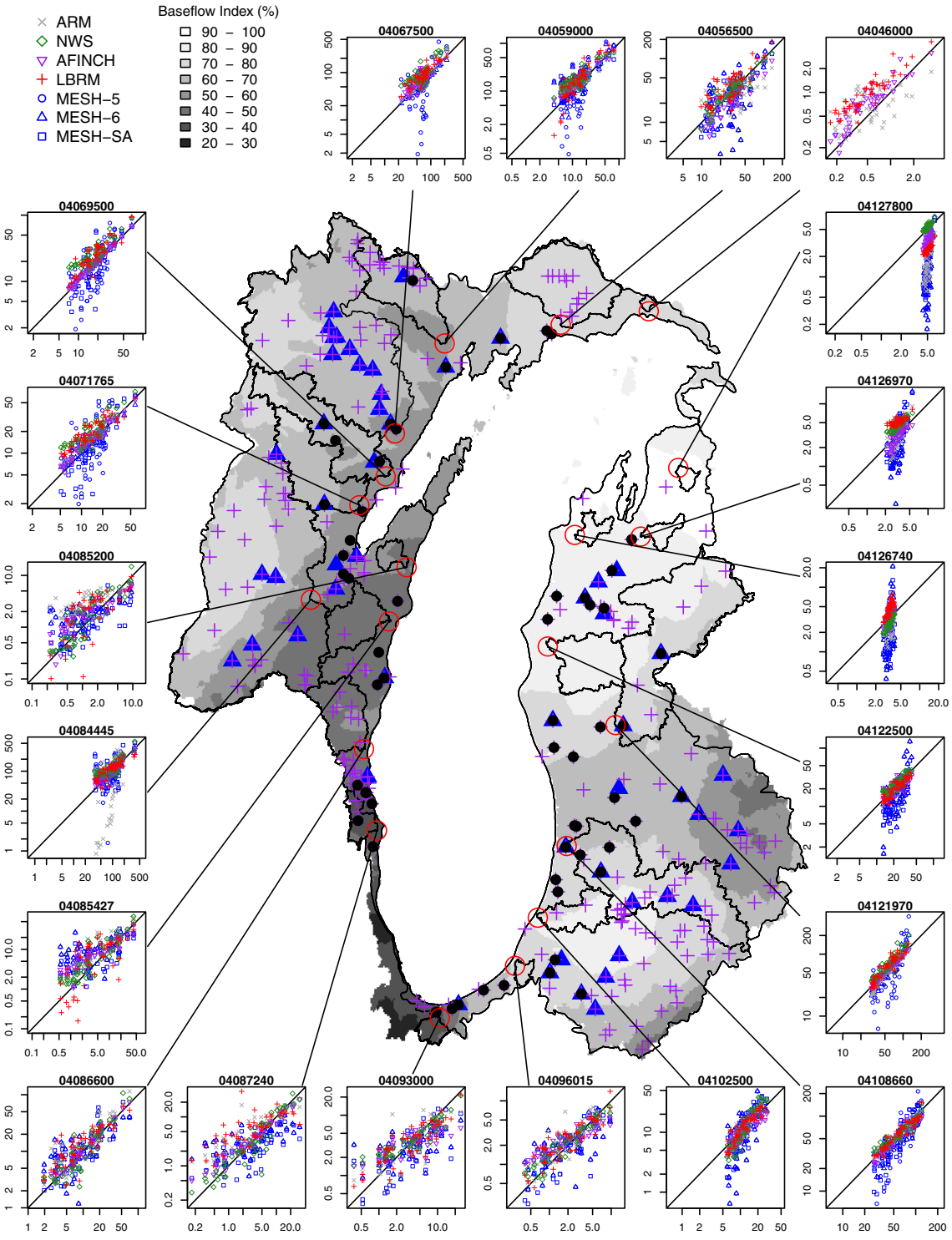

Fig. 5. Modeled (vertical axes) vs. observed (horizontal axes) discharge (in cms) at the 20 validation gages. Both axes in each panel are on a logarithmic scale. Black points represent gages assimilated by ARM and MESH-6 during the validation simulation, purple crosses represent gages assimilated by AFINCH into the validation simulation, and red circles represent the validation locations. Baseflow index (data source: Wolock, 2003) is displayed to demonstrate the significantly different hydrologic response in the northeast portion of the basin, where limited gage information is available for calibration or assimilation.

mapped hydrogeological characteristics or use of default model parameters.

When spatially aggregated, nearly all skill metrics of the assimilative models improve (Fig. 6 and Table 6). Accumulated over time, the aggregated discharge simulated by the assimilative models varied by only about 0.02 m depth at the end of the simulation period (a small difference compared to the observed cumulative depth of 0.75 m). Even the ARM simulations resulted in good model skill when accumulated over space and time, with the exception that during the first two years of the validation simulation, ARM underestimates flow dramatically. This a result of the inclusion of the Ashwaubenon Creek gage (04085068) for the ARM simulations until 2006. Because the subbasin containing this gage is a very large portion of the Lake Michigan basin, the

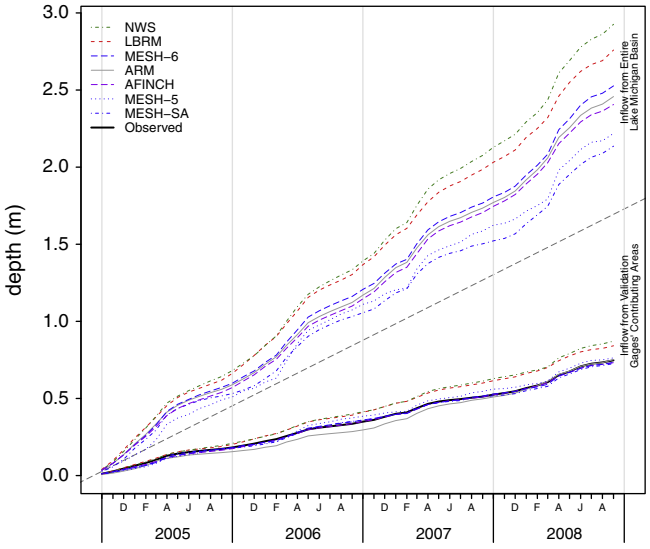

Fig. 6. Cumulative discharge for the entire Lake Michigan basin (upper cluster of lines) and for 17 of the 20 validation gages (lower cluster of lines, separated from upper cluster by a straight gray dashed line) expressed as depth (in meters) over the surface area of Lake Michigan. Observations are shown only for the lower set of lines (representing aggregated flow from the validation watersheds), but not for the upper group (representing total inflow from all land area), because no observations exist for runoff from the entire land area.

Table 6
Skill metrics for the time series of simulated discharge aggregated over the 17 gages for which all models provided simulations. The best model performance is shown in bold.

Model	NSE	PBIAS	ρ
Assimilative			
ARM	0.71	**0.60**	0.86
AFINCH	**0.95**	−2.10	**0.98**
MESH-6	0.76	−1.90	0.85
Non-assimilative			
NWS	0.78	16.60	0.95
LBRM	**0.82**	12.60	**0.96**
MESH-5	0.28	**1.90**	0.60
MESH-SA	0.66	−2.90	0.81

Table 7
Number of watersheds for which each model ranks in first place when comparing goodness-of-fit statistics over the validation period. Highest number shown in bold. Since scores are compared to two decimal places, this results in ties, hence the total number of first places can exceed the number of watersheds.

Model	NSE	PBIAS	ρ
Assimilative			
ARM	7/19	6/19	9/19
AFINCH	**14/19**	**9/19**	**11/19**
MESH-6	0/19	4/19	0/19
Non-assimilative			
NWS	**10/17**	6/17	**13/17**
LBRM	5/17	1/17	6/17
MESH-5	0/17	2/17	0/17
MESH-SA	2/17	**8/17**	0/17

Ashwaubenon Creek watershed had a large influence on basin-wide ARM simulations for the validation analysis. In normal ARM operation, the Ashwaubenon Creek gage does not have nearly as much influence, because gage 04084445 provides consistent observations. The good skill metrics despite the poor performance during the early part of the validation period, suggest that, for application to simulation of the historical monthly water balance over a large, partially gaged area, ARM can provide adequate (and, in fact quite good) results. This finding is important, considering the efficiency with which ARM simulations can be run, the historical importance of ARM in Great Lakes water balance modeling, and the need for high quality simulated historical runoff time series for use in conditioning predictive hydrological models like LBRM.

It should be noted that while ARM is strictly an assimilative model and can only be used to simulate periods for which observations at some gages exist, AFINCH and MESH-6 can be configured to operate in a predictive mode. In a predictive mode (i.e. no assimilation of discharge observations), MESH-6 is equivalent to MESH-5 (results described in Section 4.1.2). Although AFINCH is an empirical approach (in the same category as ARM), monthly temperature and precipitation are included in the regression. Therefore, in theory, it would be possible to extend AFINCH beyond historical simulation to forecasting monthly flows in a non-assimilative application, using forecasts of monthly temperature and precipitation as drivers. However, skill may be degraded in forecast mode, as there would be no assimilation of discharge observations. To consider AFINCH for forecasting, it would be important to investigate the ability of AFINCH in simulating flow without assimilation of data during the simulation period. Alternatively, evaluation of forecasting skill could involve an assessment of the stability of the regression parameters over time.

4.1.2. Non-assimilative models

Perhaps not surprisingly, there was considerably more variability in model skill among the non-assimilative models (LBRM, NWS, MESH5, and MESH-SA) than assimilative models, with the spread in cumulative, spatially aggregated discharge among the models totalling about 0.15 m, compared to 0.02 m for the assimilative models (Fig. 6). In fact, the improvement of MESH-6 (the assimilative MESH configuration) over MESH-SA and MESH-5 suggests that the assimilation of observed streamflows has a stronger impact than the combined benefits of calibration and a more complex representation of land surface processes. LBRM and NWS both tended to overestimate discharge, while MESH-5 and MESH-SA tended to underestimated discharge somewhat. The standalone version of MESH (MESH-SA) simulations resulted in good NSE values ($\geqslant$0.50) for only four validation locations. The MESH-SA PBIAS values were generally good, however, with a median PBIAS of −5%. For 12 of the 17 locations, the PBIAS was less than ±15%. In contrast to MESH-SA, MESH-5 was not calibrated to any streamflow records, and not surprisingly, NSE and ρ statistics were not as good for MESH-5 (median values of 0 and 0.61, respectively). The improved performance of MESH-SA (which uses the CLASS land surface model and the CaPA precipitation forcing) over MESH-5 (which uses the ISBA land surface model and improved GEM precipitation forcings) also demonstrates the potential for greatly improved model performance resulting from calibration, even to a small set of spatially disperse streamgages and/or a more complex land surface scheme.

MESH-5 and MESH-SA are the only non-assimilative model configurations considered here that represent hydrological processes using physics-based equations, without the need to calibrate the model to observed discharge. MESH is also the only model that operates on a distributed (gridded) spatial framework, allowing for more detailed representation of spatial heterogeneity across the basin. Although the performance of MESH-5, in terms of representing variability of flow at gages (low NSE) was less than other non-assimilative models considered in this study, the long-term bias (PBIAS) was such that the MESH simulations did result in reasonable cumulative discharge (Fig. 6), suggesting that these configurations have value for large scale water balance modeling. Future configurations using higher resolution routing such as that provided by HydroSHEDS (Lehner et al., 2006) and more complex

land surface schemes will likely result in further improvements with the assimilation of more discharge observations.

Interestingly, both LBRM and NWS resulted in large positive bias and low NSE in the northwest region of the basin (i.e. northern Wisconsin and the Upper Peninsula of Michigan, including at gages 04046000, 04056500, 04059000, 04067500, 04069500, and 04071765). An assessment of LBRM and NWS skill in this region over time revealed that their skill has declined dramatically in recent years, suggesting that the hydrologic response has changed since the periods used to calibrate LBRM and NWS. In this region, annual discharge appears to have been declining since the 1980s. While an assessment of drivers of change is outside the scope of this manuscript, there is evidence of land use changes (e.g. changes in forest cover) in this region that may contribute to altered hydrologic response (CCAP, 2013). LBRM and NWS both performed much better when bias effects were removed, with most ρ values above 0.7 for both models.

Over time, the accumulated error in LBRM and NWS simulations is potentially significant (Fig. 6). While LBRM and NWS do provide good forecasting capability in the context of flood forecasting (NWS) and basin-wide water budget forecasting (LBRM), improved calibration may be worthwhile for both models, and the models may be better suited for near-term forecasting to limit accumulation of bias. Despite these recommended updates for basin-wide modeling, application of NWS to operational forecasting within the basin remains particularly attractive, as the infrastructure exists (within the U.S.) for forcing the model with meteorological forecasts within the National Weather Service's river forecasting system, and considering the good performance relative to other non-assimilative models. Although LBRM has some shortcomings in representing the physical processes, its simplicity, the data requirements that can be met relatively easily on both sides of the border, and its current existence within the framework for forecasting water levels of the Great Lakes mean that it will remain attractive for short-term operational basin-wide forecasting until a viable alternative is put forth.

4.2. Simulations of monthly Lake Michigan basin runoff

Comparison of the time series of total runoff to Lake Michigan indicates promising results demonstrating the relative comparability of all models. In Fig. 7, time series are displayed for the total length of simulation from each model under current configurations. From the time series, it appears that while there are some differences among models, the timings of high and low flows are generally similar. In general, NWS and LBRM tend to overestimate discharge compared with the other models, and MESH tends to underestimate discharge compared with the other models.

The consistent overestimation of LBRM and NWS is especially evident in the plot showing cumulative discharge to Lake Michigan (upper group of lines in Fig. 6). When compared with the cumulative inflow to Lake Michigan simulated at the validation gages (lower group of lines in Fig. 6), the total simulated discharge to Lake Michigan (from the entire land area of the basin) exhibits similar relative errors, although spatially aggregated ARM estimates at the validation watersheds underestimate flow during the first two years of the validation period (explained by the inclusion of the Ashwaubenon Creek gage for estimation at validation watershed 04084445). The Aswhaubenon Creek does not cause such a large effect in the basin-wide GLERL simulation, however, because 04084445 consistently contributes discharge observations. NWS and LBRM overestimate discharge relative to the other models, with some confirmation provided by the cumulative discharge simulated at the gages, for which NWS and LBRM overestimate discharge compared to observations.

Interestingly, the three models that assimilate observations of discharge (MESH-6, AFINCH, and ARM) are similar in magnitude, with ARM providing the central prediction among the three (Fig. 6). Furthermore, the ensemble median of the non-assimilative models (NWS, LBRM, MESH-SA, and MESH-5) also agrees well with the assimilative model simulations, suggesting that an ensemble approach may result in good model skill. This finding contributes to evidence that combining multiple models in

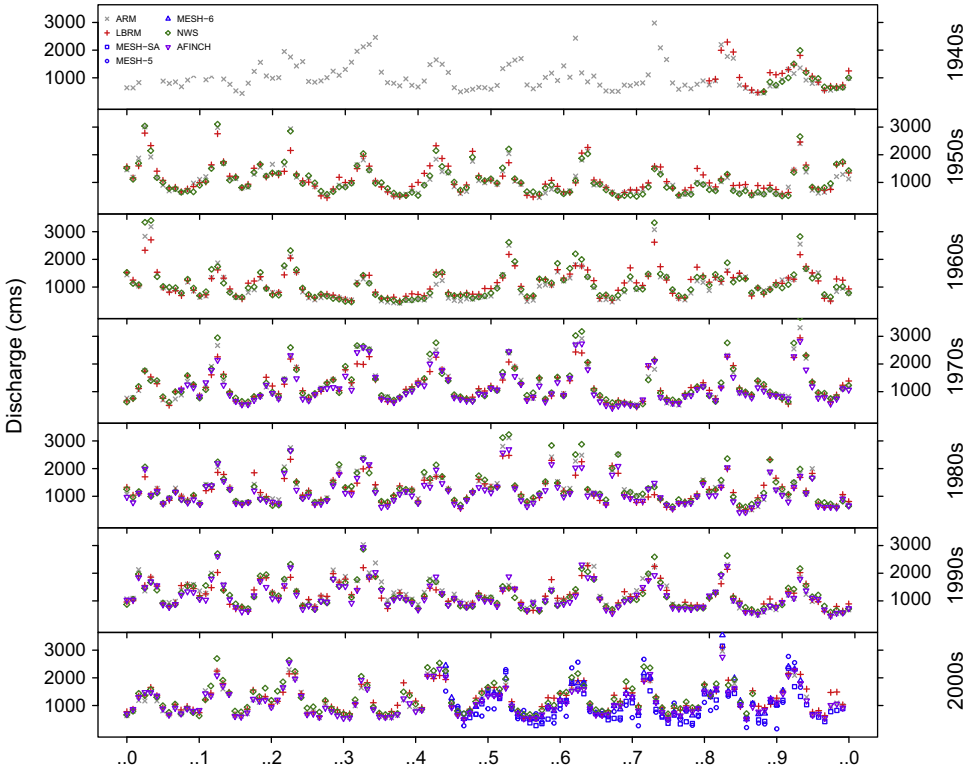

Fig. 7. Monthly inflow to Lake Michigan, by decade, for each model. No observational record is shown, as there is no observational record covering the entire land area.

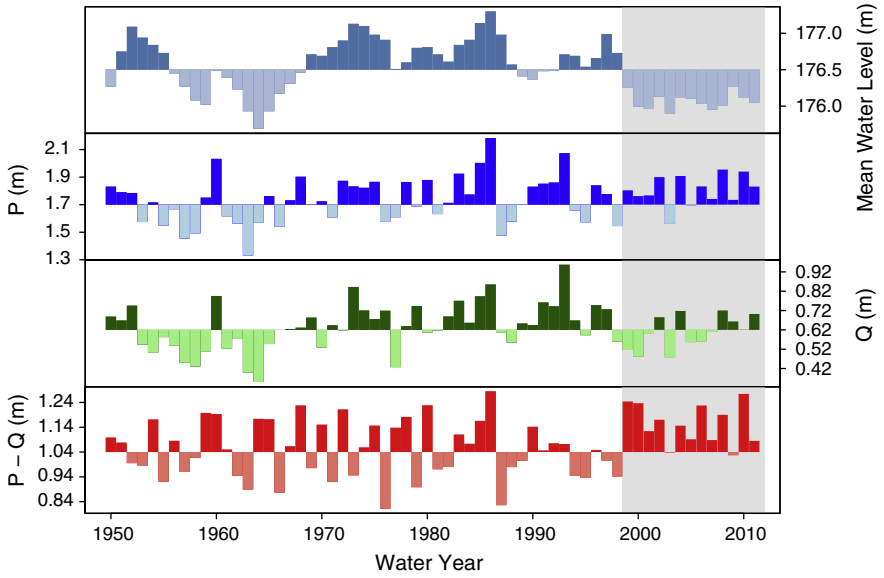

Fig. 8. Mean annual lake level, overland precipitation, runoff to the lake, and difference between overland precipitation and runoff for water years 1950–2011. Precipitation, runoff, and precipitation minus runoff are annual totals and are expressed in units of depth over Lake Michigan. Lake levels were provided by Gronewold et al. (2013a); precipitation estimates were estimated using the Thiessen polygon method; and runoff estimates are from ARM simulations of runoff to Lake Michigan.

ensemble approaches presents a method for dealing with uncertainty resulting from model structure (Refsgaard et al., 2006; Vrugt et al., 2005; Duan et al., 2007; Georgakakos et al., 2004; McIntyre et al., 2005; Viney et al., 2009). The fact that there is more spread amongst predictions from non-assimilative models is expected, and reflects the additional uncertainty related to the prediction of runoff in ungaged basins.

The findings that ARM provides the central tendency among the assimilative models and that the assimilative models perform well at the validation gages suggests that ARM (or AFINCH or MESH6) simulations can provide a useful historical record of runoff to Lake Michigan. A practical application of this result is demonstrated in Fig. 8, showing the difference between annual overland precipitation (Thiessen polygon estimates) and overland runoff (ARM simulations) for the Lake Michigan basin over the period of 1950 to 2011. During the period of 1998–2011, an important shift in this difference is evident, with all but two years having a higher than average difference between precipitation and runoff. This period corresponds to a period of low Great Lakes water levels, especially in the Lake Michigan-Huron system. The early stage of this shift was evaluated by Assel et al. (2004), who suggested that increased over-lake evaporation and reduced basin runoff were the main causes. Since the initial observation of this shift, research has resulted in evidence of shifts in lake temperature, ice cover, and water temperature following 1998 (Van Cleave et al., in press; Gronewold and Stow, 2014), however little focus has been paid to basin runoff. The ARM simulations suggest that there does not appear to be a dramatic shift in total annual runoff, despite higher than average precipitation, providing evidence of water loss to either evapotranspiration or storage. The increased water loss since 1998 provides further evidence that the changes in Lake Michigan levels are a result of water budget shift in the entire Lake Michigan basin.

5. Conclusions

Despite the wide range in model structure, forcings, calibration routines, and calibration periods, the analysis provided new and potentially significant insight toward improving regional hydrologic modeling. For example, the side-by-side comparison provided the opportunity to evaluate the importance of incorporation of spatially varying drivers of hydrologic response, the value of data assimilation for historical runoff simulation, and opportunities for improving the historical record of runoff to the Great Lakes.

The physical characteristics driving hydrologic response vary widely across the Great Lakes basin (demonstrated by one hydrologic response, baseflow index, in Fig. 5). Higher spatial resolution models (e.g. AFINCH, MESH, and NWS) allow for more detailed representation of these spatially varying drivers, compared to ARM and LBRM which operate on large subbasins. Although the models are quite different in their structure, comparison of sets of models offers some insight into effects of spatial representation of hydrologic response (and the drivers of hydrologic response) on model skill. For example, ARM, AFINCH (as it is applied for GRIP-M), and MESH-6 are similar in that they are all assimilative. They are different in the spatial unit (ARM subbasin, NHDplus flowlines, and a 10 arcmin grid, respectively) and representation of heterogeneity of hydrologic response. All three have very good model skill where observations are available for assimilation, suggesting some degree of spatial correlation of the runoff field. However, the spatial representation of drivers of hydrologic response within AFINCH results in better skill at validation gages, and validation simulation at gage 04084445 revealed that very poor ARM model performance can result even when nearby gages are assimilated if the watershed characteristics are not similar in the gaged and ungaged areas (for further discussion on selection of appropriate gages, see Archfield and Vogel, 2010; Fry et al., 2013). When aggregated over the entire basin, however, the simulated runoff from MESH-6, AFINCH, and ARM are quite similar, suggesting that the representation of spatial variability in drivers of hydrologic response is less important if the goal is basin-wide water balance simulation.

In addition to differences in spatial representation of varying drivers of hydrologic response, the comparison among empirical, lumped conceptual, and distributed physical models contributes to the ongoing debate over the philosophical question as to the significance of representing physical processes with physics-based models. In reality, it seems that simulation problems are on a continuum from purely deterministic to purely stochastic. The art of estimation involves finding the point on that continuum that is consistent with the data available to describe the system, the data

available to drive the system (inputs), the measured outputs from the system, and our needs and resources. It could be argued that the term "physically based" models is already a hedge toward statistical estimation. Data layers are seldom available for a physically based model that are good enough to use with theoretically-derived parameters to estimate the flow response. If observations are available, the parameters can generally be adjusted by calibration. This inverse modeling approach results in parameters that have a well-defined uncertainty and sensitivity, if the number of calibrated parameters is appropriate for the information content of the data.

The intercomparison revealed the potential for significantly more accurate historical discharge estimates among models that assimilate discharge observations (ARM, AFINCH, and MESH-6). In fact, in watersheds where no observations were available for assimilation, ARM and MESH-6 performance was dramatically degraded, providing a strong case for maintaining a consistent gage network for basin-wide water budget modeling. While assimilation is not possible in forecasting mode, the finding that assimilation improves historical simulation is significant, as accurate representation of the historical runoff to the lakes is critical for describing past changes in the hydrologic budget, as well as development of predictive models such as LBRM, NWS, and MESH.

To conclude, the model intercomparison revealed the following regarding what constitutes the current state-of-the-art in Great Lakes basin hydrological modeling:

- Few models exist that are readily adaptable to Great Lakes basin-wide (i.e. binational) water budget modeling. When simulating discharge at specific locations, there is variable model skill (both among models and among watersheds). When aggregated over space, however, model skill improved for all models, suggesting that differences among model structure, spatial representation, input data, and calibration procedures are somewhat less important if the goal is basin-scale water budget modeling.
- Relatively simple empirical models (i.e. ARM and AFINCH) simulate historical monthly runoff to Lake Michigan with good model skill. This finding is especially important when considering that no record of actual runoff to the lakes is available, and that ARM has traditionally filled this gap by providing a synthetic historical record.
- The median of the simulations from non-assimilative models agrees well with assimilative models, suggesting that using a combination of different methodologies might be the best approach for estimating runoff into the Great Lakes.
- ARM simulations add to the evidence that recent extreme low levels of Lake Michigan-Huron are a manifestation of an important shift in multiple components of the basin-scale water budget.

The variability in performance among non-assimilative models reveals the need for significant advancement in Great Lakes hydrological modeling for predictive purposes. However, the finding that, given the existing gage network, ARM and AFINCH provide very good historical records of runoff to Lake Michigan allows for improved understanding of historical changes in the water budget (as demonstrated in this study) and can provide the basis for future development of predictive models (like LBRM, NWS, and MESH). The opportunity to improve predictive modeling in the basin should be of interest not just to the more than 33 million residents within the Great Lakes basin, but also to the hydrologic and water resources management community at large, considering that the Great Lakes constitute the world's largest source of surface fresh water. The GRIP-M project lays the foundation for future research in subsequent phases of the GRIP initiative, which will expand model intercomparison to other Great Lakes basins, all of which span the international boundary. In some cases, this international spatial domain will likely pose significant challenges in the form of data compilation and harmonization and establishing new parameter sets (e.g. NWS and AFINCH). Future research should include development of homogenized input data sets (including meteorological, landscape, and hydrographic data) that cross the international border and can provide a basis for comparing the influence of model structure, calibration procedures, spatial frameworks, and spatiotemporal heterogeneity in drivers of hydrologic response.

Acknowledgments

This work was partially completed under a post-doctoral fellowship with the Cooperative Institute for Limnology and Ecosystems Research, awarded under a Cooperative Agreement between the University of Michigan and the NOAA Great Lakes Environmental Research Laboratory. This publication is NOAA-GLERL Contribution No. 1723.

References

Alcamo, J., Doll, P., Henrichs, T., Kaspar, F., Lehner, B., Rosch, T., Siebert, S., 2003. Development and testing of the WaterGAP 2 global model of water use and availability 48, 317–337.

Anderson, E.A., 1976. A Point Energy and Mass Balance Model of a Snow Cover. NOAA Technical Report NWS 19. Technical Report. NOAA National Weather Service.

Anderson, E., 2002. Calibration of Conceptual Hydrologic Models for Use in River Forecasting: NOAA Technical Report. NWS 45. Technical Report August. NWS Hydrology Laboratory. Silver Spring, MD.

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

Archfield, S.A., Vogel, R.M., 2010. Map correlation method: selection of a reference streamgage to estimate daily streamflow at ungaged catchments. Water Resources Res. 46, W10513.

Assel, R.A., Quinn, F.H., Sellinger, C.E., 2004. Hydroclimatic factors of the recent record drop in Laurentian Great Lakes water levels. Bull. Am. Meteorol. Soc. 85, 1143–1151.

Bélair, S., Crevier, L.-P., Mailhot, J., Bilodeau, B., Delage, Y., 2003. Operational implementation of the ISBA land surface scheme in the Canadian regional weather forecast model. Part I: Warm season results. J. Hydrometeorol. 4, 352–370.

Blanken, P.D., Spence, C., Hedstrom, N., Lenters, J.D., 2011. Evaporation from Lake Superior: 1. Physical controls and processes. J. Great Lakes Res. 37, 707–716.

Breuer, L., Huisman, J., Willems, P., Bormann, H., Bronstert, A., Croke, B., Frede, H.G., Gräff, T., Hubrechts, L., Jakeman, A., Kite, G., Lanini, J., Leavesley, G., Lettenmaier, D., Lindström, G., Seibert, J., Sivapalan, M., Viney, N., 2009. Assessing the impact of land use change on hydrology by ensemble modeling (LUCHEM). I: Model intercomparison with current land use. Adv. Water Resources 32, 129–146.

Brinkmann, W.A.R., 2000. Causes of variability in monthly Great Lakes water supplies and lake levels. Clim. Res. 15, 151–160.

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

Buttle, J., Muir, T., Frain, J., 2004. Economic impacts of climate change on the Canadian Great Lakes hydro-electric power producers: a supply analysis. Can. Water Resources J. 29, 89–110.

NOAA Coastal Services Center (CCAP), 2013. Lake Michigan Basin Land Cover Change Report. Technical Report. NOAA Coastal Services Center.

Clark, M.P., Slater, A.G., Rupp, D.E., Woods, R.A., Vrugt, J.A., Gupta, H.V., Wagener, T., Hay, L.E., 2008. Framework for understanding structural errors (FUSE): a modular framework to diagnose differences between hydrological models. Water Resources Res. 44, W00B02.

Coon, W.F., Murphy, E.A., Soong, D.T., Sharpe, J.B., 2011. Compilation of Watershed Models for Tributaries to the Great Lakes. United States, as of 2010, and Identification of Watersheds for Future Modeling for the Great Lakes Restoration Initiative: U.S. Geological Survey Open-File Report 2011-1202. Technical Report. U.S. Geological Survey. Reston, VA.

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

Croley, T.E., Hartmann, H.C., 1986. NOAA Technical Memorandum ERL GLERL-61: Near-Real-Time Forecasting of Large-Lake Water Supplies; A User's Manual. Technical Report. U.S. Dept. of Commerce, National Oceanic and Atmospheric Administration, Great Lakes Environmental Research Laboratories. Ann Arbor, MI.

Croley, T.E., He, C., 2002. Great Lakes large basin runoff modeling. In: Second Federal Interagency Hydrologic Modeling Conference, Subcommittee on Hydrology of the Interagency Advisory Committee on Water Data, Las Vegas, NV.

Daly, C., Taylor, G., 1998a. United States average monthly or annual mean temperature 1961–1990.

Daly, C., Taylor, G., 1998b. United States average monthly or annual precipitation 1961–1990.

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

Derecki, J.A., 1976. Multiple estimates of Lake Erie evaporation. J. Great Lakes Res. 2, 124–149.

Derecki, J.A., Potok, L.A.J., 1979. Regional runoff simulation for Southeast Michigan. J. Am. Water Resources Assoc. 15, 1418–1429.

Dingman, S.L., 2002. Physical Hydrology, second ed. Prentice-Hall, Inc., Upper Saddle River, New Jersey.

Duan, Q., Schaake, J., Andréassian, V., Franks, S., Goteti, G., Gupta, H., Gusev, Y., Habets, F., Hall, a., Hay, L., Hogue, T., Huang, M., Leavesley, G., Liang, X., Nasonova, O., Noilhan, J., Oudin, L., Sorooshian, S., Wagener, T., Wood, E., 2006. Model parameter estimation experiment (MOPEX): an overview of science strategy and major results from the second and third workshops. J. Hydrol. 320, 3–17.

Duan, Q., Ajami, N.K., Gao, X., Sorooshian, S., 2007. Multi-model ensemble hydrologic prediction using Bayesian model averaging. Adv. Water Resources 30, 1371–1386.

Fekete, B.M., Vörösmarty, C.J., Grabs, W., 2002. High-resolution fields of global runoff combining observed river discharge and simulated water balances. Glob. Biogeochem. Cycl. 16, 10–15.

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 Res. 49, 2689–2700.

Georgakakos, K.P., Seo, D.J., Gupta, H., Schaake, J., Butts, M.B., 2004. Towards the characterization of streamflow simulation uncertainty through multimodel ensembles. J. Hydrol. 298, 222–241.

Gronewold, A.D., Fortin, V., 2012. Advancing Great Lakes hydrological science through targeted binational collaborative research. Bull. Am. Meteorol. Soc. 93, 1921–1925.

Gronewold, A.D., Stow, C.A., 2014. Water loss from the Great Lakes. Science 343, 1084–1085.

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

Gronewold, A.D., Clites, A.H., Smith, J.P., Hunter, T.S., 2013a. A dynamic graphical interface for visualizing projected, measured, and reconstructed surface water elevations on the earth's largest lakes. Environ. Model. Softw. 49, 34–39.

Gronewold, A.D., Fortin, V., Lofgren, B., Clites, A., Stow, C.A., Quinn, F., 2013b. Coasts, water levels, and climate change: a Great Lakes perspective. Clim. Change 120, 697–711.

Haghnegahdar, A., Tolson, B., Davison, B., Seglenieks, F.R., Klyszejko, E., Soulis, E.D., Fortin, V., Matott, L.S., 2014. Calibrating Environment Canada's MESH modeling system over the Great Lakes basin. Atmosphere-Ocean (in press).

Hanasaki, N., Inuzuka, T., Kanae, S., Oki, T., 2010. An estimation of global virtual water flow and sources of water withdrawal for major crops and livestock products using a global hydrological model. J. Hydrol. 384, 232–244.

Holman, K.D., Gronewold, A.D., Notaro, M., Zarrin, A., 2012. Improving historical precipitation estimates over the Lake Superior basin. Geophys. Res. Lett. 39, L03405.

Holtschlag, D., 2009. Scientific Investigations Report 2009-5188: Application Guide for AFINCH (Analysis of Flows in Networks of Channels) Described by NHDPlus. Scientific Investigations Report 2009-5188. Technical Report. U.S. Geological Survey. Reston, VA.

Hunt, R.J., 2010. Forecasting Great Lakes Basin Response to Future Change: U.S. Geological Survey. Great Lakes Restoration Initiative. <http://cida.usgs.gov/glri/projects/accountability/responses_future_change.html>.

Hunter, T.S., Croley, T.E., 1993. Great Lakes Monthly Hydrological Data, NOAA Data Report ERL GLERL. Technical Report. National Technical Information Service. Springfield, VA.

Islam, M.S., Oki, T., Kanae, S., Hanasaki, N., Agata, Y., Yoshimura, K., 2007. A grid-based assessment of global water scarcity including virtual water trading. Water Resources Manage. 21, 19–33.

Kokkonen, T.S., Jakeman, A.J., Young, P.C., Koivusalo, H.J., 2003. Predicting daily flows in ungauged catchments: model regionalization from catchment descriptors at the Coweeta Hydrologic Laboratory, North Carolina. Hydrol. Process. 17, 2219–2238.

Koltun, G., Holtschlag, D., 2010. Application of AFINCH as a Tool for Evaluating the Effects of Streamflow-Gaging-Network Size and Composition on the Accuracy and Precision of Streamflow Estimates at Ungaged Locations in the Southeast Lake Michigan Hydrologic Subregion. U.S. Geolological Survey. Technical Report.

Koren, V., Reed, S., Smith, M., Zhang, Z., Seo, D.J., 2004. Hydrology laboratory research modeling system (HL-RMS) of the US national weather service. J. Hydrol. 291, 297–318.

Koren, V., Moreda, F., Reed, S., Smith, M., Zhang, Z., 2006. Evaluation of a grid-based distributed hydrological model over a large area. In: Predictions in Ungauged Basins: Promise and Progress. Proceedings of Symposium S7 Held during the Seventh IAHS Scientific Assembly. IAHS Publ, Foz do Iguaçu, Brazil, pp. 47–56.

Kouwen, N., 2010. WATFLOOD/WATROUTE Hydrological Model Routing & Flow Forecasting System. Department of Civil Engineering, University of Waterloo, Waterloo, ON.

Lehner, B., Verdin, K., Jarvis, A., 2006. HydroSHEDS Technical Documentation Version 1.0.

Lofgren, B.M., 2004. A model for simulation of the climate and hydrology of the Great Lakes basin. J. Geophys. Res. 109, 1–20.

Lofgren, B.M., Hunter, T.S., Wilbarger, J., 2011. Effects of using air temperature as a proxy for potential evapotranspiration in climate change scenarios of Great Lakes basin hydrology. J. Great Lakes Res. 37, 744–752.

Mahfouf, J.f., Brasnett, B., Gagnon, S., 2007. A Canadian precipitation analysis (CaPA) project: description and preliminary results. Atmosp. – Ocean 45, 1–17.

Mailhot, J., Bélair, S., Lefaivre, L., Bilodeau, B., Girard, C., Glazer, A., Leduc, A.M., Méthot, A., Plante, A., Rahill, A., Robinson, T., Talbot, D., Tremblay, A., Desgagné, M., Patoine, A., 2006. The 15-km version of the Canadian regional forecast system. Atmosp. – Ocean 44, 133–149.

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

McCabe, G.J., Wolock, D.M., 2011. Century-scale variability in global annual runoff examined using a water balance model. Int. J. Climatol. 31, 1739–1748.

McIntyre, N., Lee, H., Wheater, H., Young, A., Wagener, T., 2005. Ensemble predictions of runoff in ungauged catchments. Water Resources Res. 41, 1–14.

McKay, L., Bondelid, T., Dewald, T., 2012. NHDPlus Version 2: User Guide. Technical Report. U.S. Environmental Protection Agency and U.S. Geological Survey.

McMahon, G., Alexander, R.B., Qian, S., 2003. Support of total maximum daily load programs using spatially referenced regression models. J. Water Resources Plan. Manage. 129, 315–329.

Millerd, F., 2005. The economic impact of climate change on Canadian commercial navigation on the Great Lakes. Can. Water Resources J. 30, 269–280.

Millerd, F., 2011. The potential impacts of climate change on Great Lakes international shipping. Clim. Change 104, 629–652.

Milly, P.C.D., Betancourt, J., Falkenmark, M., Hirsch, R.M., Zbigniew, W., Lettenmaier, D.P., Stouffer, R.J., 2008. Stationarity is dead: whither water management? Science 319, 573–574.

Nakada, S., Ishikawa, Y., Awaji, T., In, T., Shima, S., Nakayama, T., Isada, T., Saitoh, S.i., 2012. Modeling runoff into a region of freshwater influence for improved ocean prediction: application to Funka Bay. Hydrol. Res. Lett. 6, 47–52.

Nijssen, B., O'Donnell, G.M., Hamlet, A.F., Lettenmaier, D.P., 2001a. Hydrologic sensitivity of global rivers to climate change. Clim. Change 50, 143–175.

Nijssen, B., Schnur, R., Lettenmaier, D.P., 2001b. Global retrospective estimation of soil moisture using the variable infiltration capacity land surface model, 1980–93. J. Clim. 14, 1790–1808.

Oregon State University PRISM Group, 2008. Parameter-Elevation Regressions on Independent Slopes Model. <http://www.prism.oregonstate.edu/>.

Oudin, L., Andreassian, V., Perrin, C., Michel, C., Le Moine, N., 2008. Spatial proximity, physical similarity, regression and ungaged catchments: a comparison of regionalization approaches based on 913 French catchments. Water Resources Res. 44, W03413.

Parajka, J., Viglione, A., Rogger, M., Salinas, J.L., Sivapalan, M., Blöschl, G., 2013. Comparative assessment of predictions in ungauged basins Part 1: Runoff-hydrograph studies. Hydrol. Earth Syst. Sci. 17, 1783–1795.

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

Quinn, F.H., 1978. Hydrologic response model of the North American Great Lakes. J. Hydrol. 37, 295–307.

Reed, S., Koren, V., Smith, M., Zhang, Z., Moreda, F., Seo, D.J., 2004. Overall distributed model intercomparison project results. J. Hydrol. 298, 27–60.

Refsgaard, J.C., van der Sluijs, J.P., Brown, J., van der Keur, P., 2006. A framework for dealing with uncertainty due to model structure error. Adv. Water Resources 29, 1586–1597.

Reichl, J.P.C., Western, A.W., McIntyre, N.R., Chiew, F.H.S., 2009. Optimization of a similarity measure for estimating ungauged streamflow. Water Resources Res. 45, W10423.

Robertson, D.M., Saad, D.A., 2011. Nutrient inputs to the Laurentian Great Lakes by source and watershed estimated using SPARROW watershed models. J. Am. Water Resources Assoc./AWRA 47, 1011–1033.

Roe, J., Dietz, C., Resetrepo, P., Halquist, J., Hartman, R., Horwood, R., Olsen, B., Opitz, H., Shedd, R., Welles, E., 2010. Paper 7B.3: Introduction of NOAA's Community Hydrologic Prediction System. In: 26th Conference on Interactive Information and Processing Systems (IIPS) for Meteorology, Oceanography, and Hydrology, Atlanta, GA.

Seitzinger, S.P., Mayorga, E., Bouwman, A.F., Kroeze, C., Beusen, A.H.W., Billen, G., Van Drecht, G., Dumont, E., Fekete, B.M., Garnier, J., Harrison, J.A., 2010. Global river nutrient export: a scenario analysis of past and future trends. Glob. Biogeochem. Cycl. 24, GB0A08.

Sivapalan, M., Takeuchi, K., Franks, S.W., Gupta, V.K., Karambiri, H., Lakshmi, V., 2003. IAHS decade on predictions in ungauged basins (PUB), 2003–2012: shaping an exciting future for the hydrological sciences. Hydrol. Sci. - J. Sci. Hydrol. 48, 857–880.

Smith, R.A., Schwarz, G.E., Alexander, R.B., 1997. Regional interpretation of water-quality monitoring data. Water Resources Res. 33, 2781–2798.

Smith, M.B., Seo, D.J., Koren, V.I., Reed, S.M., Zhang, Z., Duan, Q., Moreda, F., Cong, S., 2004. The distributed model intercomparison project (DMIP): motivation and experiment design. J. Hydrol. 298, 4–26.

Smith, M.B., Koren, V., Zhang, Z., Zhang, Y., Reed, S.M., Cui, Z., Moreda, F., Cosgrove, B.A., Mizukami, N., Anderson, E.A., 2012. Results of the DMIP 2 Oklahoma experiments. J. Hydrol., 17–48.

Tang, Q., Gao, H., Yeh, P., Oki, T., Su, F., Lettenmaier, D.P., 2010. Dynamics of terrestrial water storage change from satellite and surface observations and modeling. J. Hydrometeorol. 11, 156–170.

USGS, 2011. GAGES-II: Geospatial Attributes of Gages for Evaluating Streamflow. <http://water.usgs.gov/lookup/getspatial?gagesII_Sept2011>.

Van Cleave, K., Lenters, J.D., Wang, J., Verhamme, E.M., 2014. A regime shift in Lake Superior ice cover, evaporation, and water temperature following the warm El Niño winter of 1997-98. Limnol. Oceanogr. (in press).

Verseghy, D., 2000. The Canadian land surface scheme (CLASS): its history and future. Atmosp. – Ocean 38, 1–13.

Viney, N.R., Bormann, H., Breuer, L., Bronstert, a., Croke, B., Frede, H., Gräff, T., Hubrechts, L., Huisman, J., Jakeman, a.J., Kite, G., Lanini, J., Leavesley, G., Lettenmaier, D., Lindström, G., Seibert, J., Sivapalan, M., Willems, P., 2009. Assessing the impact of land use change on hydrology by ensemble modelling (LUCHEM) II: Ensemble combinations and predictions. Adv. Water Resources 32, 147–158.

Vogelmann, J., Howard, S., Yang, L., Larson, C.R., Wylie, B.K., Driel, J.N.V., 2001. Completion of the 1990's national land cover data set for the conterminous United States. Photogramm. Eng. Rem. Sens. 67, 650–662.

Vrugt, J.A., Diks, C.G.H., Gupta, H.V., Bouten, W., Verstraten, J.M., 2005. Improved treatment of uncertainty in hydrologic modeling: combining the strengths of global optimization and data assimilation. Water Resources Res. 41, W01017.

Wagener, T., Wheater, H.S., 2006. Parameter estimation and regionalization for continuous rainfall-runoff models including uncertainty. J. Hydrol. 320, 132–154.

Werner, M., Schellekens, J., Gijsbers, P., van Dijk, M., van den Akker, O., Heynert, K., 2013. The delft-FEWS flow forecasting system. Environ. Model. Softw. 40, 65–77.

Wilson, M.A., Carpenter, S.R., 1999. Economic valuation of freshwater ecosystem services in the United States: 1971–1997. Ecol. Appl. 9, 772–783.

Wisser, D., Frolking, S., Douglas, E.M., Fekete, B.M., Vörösmarty, C.J., Schumann, A.H., 2008. Global irrigation water demand: variability and uncertainties arising from agricultural and climate data sets. Geophys. Res. Lett. 35, 1–5.

Wolock, D., 2003. Base-flow index grid for the conterminous United States. <http://water.usgs.gov/GIS/metadata/usgswrd/XML/bfi48grd.xml>.

World Resources Institute (WRI), 2006. WRI Major Watersheds of the World Delineation. <http://www.wri.org/publication/watersheds-of-the-world>.

Xia, Y., Mitchell, K., Ek, M., Cosgrove, B., Sheffield, J., Luo, L., Alonge, C., Wei, H., Meng, J., Livneh, B., Duan, Q., Lohmann, D., 2012. Continental-scale water and energy flux analysis and validation for North American land data assimilation system project phase 2 (NLDAS-2): 2. Validation of model-simulated streamflow. J. Geophys. Res. 117, 1–23.

Zomer, R., Trabucco, A., Bossio, D., van Straaten, O., Verchot, L., 2008. Climate change mitigation: a spatial analysis of global land suitability for clean development mechanism afforestation and reforestation. Agric. Ecosyst. Environ. 126, 67–80.