WATER RESOURCES RESEARCH, VOL. 49, 2689–2700, doi:10.1002/wrcr.20233, 2013

Identifying streamgage networks for maximizing the effectiveness of regional water balance modeling

L. M. Fry,[1] T. S. Hunter,[2] M. S. Phanikumar,[3] V. Fortin,[4] and A. D. Gronewold[2]

Received 21 December 2012; revised 21 March 2013; accepted 28 March 2013; published 28 May 2013.

[1] One approach to regional water balance modeling is to constrain rainfall-runoff models with a synthetic regionalized hydrologic response. For example, the Large Basin Runoff Model (LBRM), a cornerstone of hydrologic forecasting in the Laurentian Great Lakes basin, was calibrated to a synthetic discharge record resulting from a drainage area ratio method (ARM) for extrapolating beyond gaged areas. A challenge of such approaches is the declining availability of observations for development of synthetic records. To advance efficient use of the declining gage network in the context of regional water balance modeling, we present results from an assessment of ARM. All possible combinations of "most-downstream" gages were used to simulate runoff at the gaged outlet of Michigan's Clinton River watershed in order to determine the influence of gages' drainage area and other physical characteristics on model skill. For nearly all gage combinations, ARM simulations resulted in good model skill. However, the gages' catchment area relative to that of the outlet's catchment is not an unquestionable predictor of model performance. Results indicate that combinations representing less than 30% of the total catchment area (less than 10% in some cases) can provide very good discharge simulations, but that similarity of the gaged catchments' developed and cultivated area, stream density, and permeability relative to the outlet's catchment is also important for successful simulations. Recognition of thresholds on the relationship between the number of gages and their relative value in simulating flow over large area provides an opportunity for improving historical records for regional hydrologic modeling.

Citation: Fry, L. M., T. S. Hunter, M. S. Phanikumar, V. Fortin, and A. D. Gronewold (2013), Identifying streamgage networks for maximizing the effectiveness of regional water balance modeling, *Water Resour. Res.*, *49*, 2689–2700, doi:10.1002/wrcr.20233.

1. Introduction

[2] Over the past two decades, several approaches for simulating flow over partially gaged regions have emerged, driven in large part by the Prediction in Ungaged Basins (PUB) Initiative [*Sivapalan et al.*, 2003]. This large and growing body of research aims to reduce the uncertainty of hydrologic prediction in partially or totally ungaged basins by employing knowledge of relationships between hydrologic function and physical characteristics of gaged basins.

[3] Approaches for predicting flow in ungaged basins have traditionally been to (1) calibrate the chosen rainfall-runoff model in basins where observations are available and develop regression equations for parameter estimation at ungaged locations, (2) estimate model parameters using only physical watershed characteristics, or (3) transfer complete parameter sets from other hydrologically similar watersheds [*Wagener and Montanari*, 2011]. An additional emerging approach has been identified by *Wagener and Montanari* [2011] as one which mimics calibration in ungaged basins by quantifying expected hydrologic signatures that can be assimilated into any hydrologic model to inform a calibration process. Some examples of hydrologic signatures that can be used for regionalization include magnitudes of high, low and average flows [e.g., *Yadav et al.*, 2007], runoff ratio (fraction of rainfall which becomes runoff) [e.g., *Carrillo et al.*, 2011; *Sawicz et al.*, 2011; *Singh et al.*, 2011; *Yadav et al.*, 2007; *Zhang et al.*, 2008], slope of the flow duration curve [e.g., *Carrillo et al.*, 2011; *Sawicz et al.*, 2011; *Yadav et al.*, 2007; *Zhang et al.*, 2008], and base-flow index [e.g., *Bulygina et al.*, 2009; *Carrillo et al.*, 2011; *Sawicz et al.*, 2011; *Singh et al.*, 2011]. Streamflow volume itself, in fact, can be considered a hydrologic signature [e.g., *Yadav et al.*, 2007].

[4] Interestingly, the method employed by the National Oceanic and Atmospheric Administration's Great Lakes Environmental Research Laboratory (GLERL) to forecast runoff to the Great Lakes nearly 25 years ago (and still used today) is similar to the seemingly novel approach of

[1]Cooperative Institute for Limnology and Ecosystems Research, University of Michigan, Ann Arbor, Michigan, USA.
[2]Great Lakes Environmental Research Laboratory, National Oceanic and Atmospheric Administration, Ann Arbor, Michigan, USA.
[3]Department of Civil and Environmental Engineering, Michigan State University, East Lansing, Michigan, USA.
[4]Environmental Numerical Prediction Research Section, Environment Canada, Dorval, Québec, Canada.

Corresponding author: L. M. Fry, Cooperative Institute for Limnology and Ecosystems Research, University of Michigan, Ann Arbor, MI 48108, USA. (lmfry@umich.edu)

0043-1397/13/10.1002/wrcr.20233

conditioning a rainfall-runoff model on a synthetic hydrologic response probability distribution for ungaged basins. The Large Basin Runoff Model (LBRM) is a lumped parameter conceptual model that is used as the principal hydrological model in a variety of research-based and operational forecasting frameworks, ranging from the Great Lakes Advanced Hydrologic Prediction System (AHPS) [*Gronewold et al*., 2011] to the Huron-Erie Connecting Waterways Forecasting System (HECWFS) [*Anderson et al*., 2010]. LBRM, which has nine calibrated parameters, is the only conceptual rainfall-runoff model to be systematically applied to the entire Great Lakes basin [*Coon et al*., 2011]. These nine parameters are historically conditioned on a synthetic discharge time series derived from a simple drainage area ratio method (ARM) that extrapolates discharge from gaged portions of each Great Lakes subbasin to the downstream ungaged portions of the subbasins (equation (1)),

$$Q_u = Q_g\left(\frac{A_u}{A_g}\right), \tag{1}$$

where Q_u is the streamflow estimated for the ungaged area, Q_g is the observed flow in the gaged area, A_u is the ungaged area, and A_g is the gaged area.

[5] While more complex and modern models have been developed to simulate and forecast runoff over large portions of the basin [see *Coon et al*., 2011], more large-scale runoff modeling is done for individual subbasins or portions of the basin that are entirely on one side of the international border [e.g., *Holtschlag*, 2009; *Mao and Cherkauer*, 2009; *Robertson and Saad*, 2011] as a result of transborder data coordination issues. In addition to the LBRM (forecasts) and GLERL ARM (simulated historical records) runoff estimates, Environment Canada uses its Modélisation Environmentale Communautaire-Surface Hydrology (MESH) model [*Pietroniro et al*., 2007], a distributed model combining land surface models with land surface parameterization and hydrologic routing, to forecast runoff to the lakes from both the U.S. and Canadian portions of the basin. The ARM, however, to the best of our knowledge, provides the longest simulation of historical record incorporating daily observations of discharge over the entire Great Lakes basin.

[6] The ARM, as it is used by GLERL, is a two-step process that operates on 121 subbasins in the Great Lakes basin [*Croley and Hartmann*, 1986; *Croley and He*, 2002]. The first step is to apply the ARM in partially gaged subbasins (on each day), and the second step is to extrapolate from these partially gaged subbasins to totally ungaged basins by applying ARM a second time. Any gage providing observations on a given day could potentially contribute to the model if there are no additional gages downstream. Depending on the operational status of the gages used in the GLERL ARM calculations, the daily gaged portion of the Great Lakes basin ranged from about 62% to 70% between 1960 and 2008 with an overall decline in gaged area since the 1980s, as shown in Figure 1. In very recent years, this gaged portion is likely increasing somewhat, however, as some streamgages have been added to major tributaries as part of the Great Lakes Restoration Initiative. The subbasin runoff estimates are aggregated and made available online as

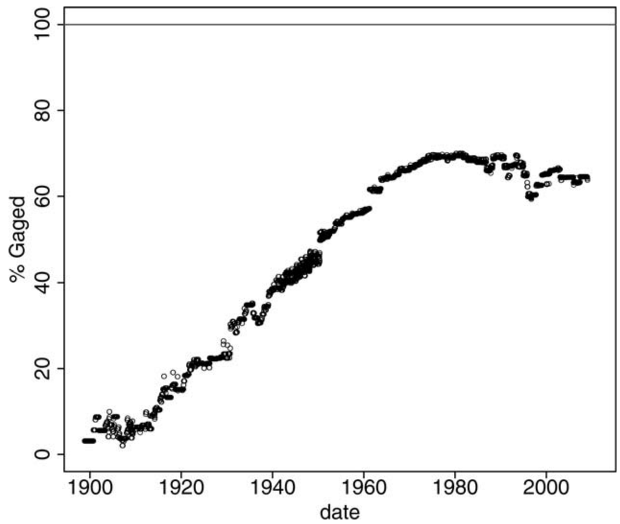

Figure 1. Daily percent of the Great Lakes basin that gaged. Percent gage was estimated from GLERL's Great Lakes Monthly Hydrologic Data, available at http://www.glerl.noaa.gov.

time series of monthly runoff to each of the Great Lakes with GLERL's Great Lakes Monthly Hydrologic Data, and span the period of 1898–2010 (http://www.glerl.noaa.gov/).

[7] The ability to simulate runoff over the entire Great Lakes basin (U.S. and Canadian sides) by incorporating all available discharge observations is a significant achievement of the GLERL ARM, because complete spatial coverage of runoff estimates from land is critical for understanding and predicting lake level changes. Incorporation of stream gage data may reduce uncertainty by providing observations wherever possible. The simplicity of ARM data requirements in intermittently gaged basins (available daily streamgage discharge and contributing area) has remained an advantage in this international basin.

[8] While in the remainder of this article, we use "ARM" to describe the general approach of extrapolation of discharge from gaged to ungaged areas by multiplying the ratio of ungaged to gaged area by the gaged discharge (as in equation (1)), the preceding description of GLERL's employment of ARM provides an important contextual background for the significance of this research. Although there have been considerable research investments and advancements in modeling software (for example Hydromad; *Andrews et al*. [2011]), computational power [*Beven*, 2007; *Wood et al*., 2011], rainfall-runoff modeling incorporating state-of-the-art GIS-based terrain and land use analysis [*Grimaldi et al*., 2010; *Noto and La Loggia*, 2007], and *approaches* to acknowledging and quantifying uncertainty in recent decades (e.g., GLUE; *Beven and Freer* [2001], Parameter ESTimation Software (PEST); *Doherty and Johnston* [2003], and DREAM; *Vrugt and Ter Braak* [2011]), estimation of runoff in ungaged basins remains one of the most pressing challenges to the hydrological science community. This is exemplified by the continued reliance on the simple ARM estimator. Implicit in these ARM-derived synthetic runoff time series are two important and potentially problematic assumptions: (1) homogeneity of hydrologic response across an entire partially gaged basin,

and (2) homogeneity of precipitation across the same area. However, no systematic evaluation of these assumptions or the effectiveness of ARM for simulating historical runoff from the subbasins has been applied to date.

[9] Previous research has resulted in various recommendations for implementation of ARM simulations. For example, *Hortness* [2006] cites several recommendations for suitable drainage area ratios ranging from a wide interval of 0.3–1.5 to a narrow interval of 0.85–1.15, ultimately recommending that a suitable area ratio should be between 0.5 and 1.5. *Emerson et al*. [2005] evaluated the drainage area ratio method for simulating runoff in the Red River of the North Basin in North Dakota and Minnesota and found that the gaged drainage area was the only significant variable contributing to model skill. However, more recently, *Archfield and Vogel* [2010] found that ARM simulations in gaged areas could be improved by selecting the nearest donor gages with highest correlation in streamflow series, leading to a "map-correlation" method in which a correlation coefficient (to existing gages' time series) at an ungaged location is determined via kriging to spatially interpolate correlation coefficients between gages. *Juckem et al*. [2012] evaluated an adjusted area ratio method, in which they developed a regression relating the observed ratio of discharge to watershed area at short-term gaging stations to the water yield from nearby long-term gaging stations in order to extend the flow records of the short-term stations, and found that the adjusted area ratio method resulted in smaller residuals than simulations produced by a water balance model.

[10] The increasing number and variety of approaches to simulating and forecasting runoff over large spatial domains collectively underscore the importance of continually assessing the relationship between model complexity and predictive skill [*Beven*, 2007; *Doherty*, 2011]. However, complex and novel approaches to simulating and predicting flows in ungaged basins are often applied across a broad range of problems, with an implicit degree of confidence that is disproportionately greater than the time invested in assessing and communicating model skill [e.g., *Deacu et al*., 2012; *Lofgren*, 2004]. To help close the gap between research-oriented and "real-world" hydrological modeling, we provide a critical assessment of an empirical runoff model (ARM) that, despite its simplicity, has served as a basis for operational hydrological modeling and forecasting in the Great Lakes for the past 30 years.

[11] This assessment is motivated in part by recommendations from the 2011 Workshop on Improving Hydrological Modeling Predictions in the Great Lakes, including (1) improving runoff predictions for the entire Great Lakes basin, starting with hindcast mode, assimilating streamflow observations; (2) assembling lake ice area and thickness data; (3) improving the representation of the lakes' thermocline structure; and (4) improving flow projections from Lake Ontario for propagating into hydrodynamic models linking the St. Lawrence River with the Gulf of St. Lawrence [*Gronewold and Fortin*, 2012]. Accordingly, one objective of this assessment is to assess ARM for continued use in providing a synthetic record of historical runoff to the Great Lakes. Beyond Great Lakes hydrological modeling, a second objective is to use this assessment as a basis for gaining insight not only into the relationship between model skill and complexity but also between the model skill and both the quantity and relative value of empirical evidence and answer the compelling question, "how many and which gages are needed to characterize the hydrological response of a particular area?" This question is particularly relevant in recent decades when the network of gaging stations has been declining (see Figure 1 for an example in the Great Lakes basin), despite the continuing need to simulate and forecast streamflow and water budgets over regions with large ungaged areas. The analysis presented here builds on previous research by investigating the value of individual gages for extrapolation to downstream areas using ARM simulations, and relating model skill to catchment similarity, presenting an opportunity for further improving ARM simulations by selecting appropriate gages. The analysis represents a novel approach to determining thresholds on the relationship between number of observations and their relative value for regional hydrological modeling. While the Clinton River watershed in Michigan provides the test bed for this research, the methods developed can be applied to the remaining subbasins of the Great Lakes and any region where extrapolation beyond gaged areas is necessary. The evaluation of gages for use in extrapolation to ungaged areas is a significant contribution in the context of U.S. water resources management, as the need for simulation over large regions is explicitly or implicitly stated in three of the five federal needs forming the basis of the USGS National Streamflow Information Program (described online at http://water.usgs.gov/nsip/federalneeds.html and shown in Table 1).

2. Data and Methods

2.1. Location and Data Sources

[12] This analysis evaluates ARM for simulating historical runoff from one Great Lakes subbasin: the Clinton River watershed. The Clinton River watershed is a subbasin of the Lake Saint Clair basin (Figure 2). The Clinton River watershed contains 11 streamgages in the United States Geological Survey (USGS) GAGESII database (available online at http://waterdata.usgs.gov/) that have operated nearly continuously in the recent period, and therefore offers opportunities for cross validation and investigation of the impact of varying gage operation on model fit. The watershed is diverse in land use, with planted, forested, and developed areas (Table 2). Nearly complete time series of daily mean discharge for each gage are available from the

Table 1. Five Federal Needs Determining the Locations of the USGS "Backbone" Network of Gages in Its National Streamflow Information Program (NSIP)[a]

Meeting interstate and international legal and treaty obligations	*
Forecasting streamflow	*
Measuring river basin outflows	*
Monitoring sentinal watersheds to determine long-term trends across the country	
Measuring flow to support the USGS water quality networks	

[a]Starred goals are met in part by incorporation of gage data into regional scale hydrological modeling. These goals are described online at http://water.usgs.gov/nsip/federalneeds.html.

Figure 2. Location of the Clinton River watershed within the Great Lakes Basin.

USGS for the period of analysis 1950–2009 (available online at http://waterdata.usgs.gov/).

2.2. Determination of the Impact of Gaged Area on Model Skill

[13] ARM allows for use of any operational gage on a given day, and therefore the skill of the model likely varies due to the portion of the basin that is gaged on a given day. To evaluate the impact of gage operation on model skill for an intermittently gaged area, ARM estimates at the Cinton River's most-downstream gage (at gage 04165500) were produced for every possible combination of operating gages. For the 11 selected streamgages in the Clinton River Watershed, there are 2047 possible scenarios of streamgage operation for this network, considering all networks of size 1–11. These streamgage combinations were evaluated to determine the most downstream gages in each combination, and unique most-downstream gage combinations were kept for analysis. For example, in a hypothetical three-station network of streamgages [1,2,3], in which streamgage 3 is the most downstream, there are three two-station combinations [1,2],[1,3],[2,3], for which only [1,2], and [1,3] OR [2,3] would be considered because streamgage 3 appears as the most downstream gage in two of the combinations. Applying this process to the Clinton River leaves 185 possible gage combinations (including outlet gage 04165500). ARM-modeled discharge at the outlet (gage 04165500) was estimated for each of the 185 combinations for each day during the period of water years 1950–2009.

[14] In addition to evaluating residuals between modeled and observed discharge, we evaluated a series of skill metrics for ARM simulations at simulating discharge at gage 04165500, including Nash-Sutcliffe efficiency (NSE) (equation (2)), Percent Bias (PBIAS) (equation (3)), and RMSE observations standard deviation ratio (RSR) (equation (4)) for daily and monthly simulations resulting from each combination of operating gages using R package hydroGOF [*Zambrano-Bigiarini*, 2011]. *Moriasi et al*. [2007] recommend these three goodness-of-fit statistics for inclusion in calibration and validation of hydrological models. Although r^2 (the squared Pearson's correlation coefficient) (equation (5)) is often used for evaluating model skill, *Moriasi et al*. [2007] do not recommend including r^2 because it is overly sensitive to large outliers and under-sensitive to additive and proportional differences between observations and predictions. NSE is a measure of the noise in residuals compared to the data variation, and indicates how well a plot of observed versus simulated discharge fits a 1:1 line. PBIAS is a measure of the average tendency of modeled data to be larger or smaller than the observations and is appropriate for quantifying water balance errors because of its similarity to percent deviation of streamflow volume. RSR is a standardized version of the RMSE, which combines both an error index and the standard deviation of the observed data, which is recommended by *Legates and McCabe* [1999] and *Moriasi et al*. [2007]. In equations (2)–(5), O_i is the observed value, P_i is the simulated value, and n is the number of simulations. Goodness-of-fit statistics were calculated for log-transformed simulated and observed discharge values.

$$\mathrm{NSE} = 1 - \frac{\sum_{i=1}^{n} (O_i - P_i)^2}{\sum_{i=1}^{n} \left(O_i - \overline{O}\right)^2}, \tag{2}$$

$$\mathrm{PBIAS} = \frac{\sum_{i=1}^{n} (O_i - P_i) \times 100}{\sum_{i=1}^{n} O_i}, \tag{3}$$

Table 2. Selected Catchment Characteristics[a] and Operational Status (% of Days During Analysis Period for Which the Gage Provides an Operation) for Each Gage in the Clinton River Watershed[b]

	Catchment Area (km^2)	Stream Density (km/km^2)	Slope (%)	Permeability (in/h)	Available Water Capacity (in/in)	% Open Water	% Forest	% Planted	% Developed	Average Annual Precipitation (cm)	Average Annual Temperature (^{0})	Operational Status (%)
04160800	55	0.5	1.5	7.0	0.14	6.6	19.7	3.1	45.6	79.2	9.1	83
04160900	203	0.5	1.6	6.6	0.13	8.8	20.1	4.7	51.4	79.7	9.0	83
04161540	186	0.7	1.8	5.6	0.14	3.9	23.3	14.4	41.6	79.7	9.1	84
04161580	64	0.5	1.9	3.2	0.16	4.0	37.1	23.3	11.8	80.4	9.1	75
04161800	168	0.5	1.8	4.6	0.15	5.5	34.6	20.9	18.4	81.0	9.1	86
04163400	47	0.8	0.6	2.8	0.15	1.1	3.2	0.1	92.2	81.5	9.3	72
04164000	1143	0.6	1.3	5.1	0.14	4.5	15.0	6.5	63.3	80.8	9.2	99
04164100	55	0.5	2.1	3.0	0.16	1.7	32.5	26.4	17.6	81.0	9.0	86
04164300	33	0.8	0.1	1.1	0.15	0.0	8.7	76.3	7.1	80.7	8.6	85
04164500	512	0.9	0.6	1.7	0.15	0.4	17.7	55.0	13.8	82.0	8.8	99
04165500[c]	1893	0.7	1.0	4.0	0.14	2.8	15.0	20.3	51.3	81.5	9.1	99

[a]Available from the USGS GAGESII database online at http://waterdata.usgs.gov/./.
[b]Gage data available at http://waterdata.usgs.gov/.
[c]Gage 04165500 is the outlet gage.

$$\mathrm{RSR} = \frac{\sqrt{\sum_{i=1}^{n}\left(O_i - P_i\right)^2}}{\sqrt{\sum_{i=1}^{n}\left(O_i - \overline{O}\right)^2}}, \tag{4}$$

$$r^2 = \frac{\sum_{i=1}^{n}\left(O_i - \overline{O}\right)\left(P_i - \overline{P}\right)}{\sqrt{\sum_{i=1}^{n}\left(O_i - \overline{O}\right)^2}\sqrt{\sum_{i=1}^{n}\left(P_i - \overline{P}\right)^2}}. \tag{5}$$

[15] To evaluate the contribution of each gage to model skill, we evaluated goodness-of-fit statistics and plots of modeled versus observed discharge for the 11 simulations of daily discharge resulting from inclusion of only one gage. Additionally, we explored the relationships between goodness-of-fit statistics and the selected watershed characteristics (Table 2) using a dissimilarity measure. An overall measure of dissimilarity between the gaged area, a, and the outlet gage's total drainage area, b, was estimated, following *Kay et al*. [2007], as the Euclidean distance in the catchment property space with the catchment properties normalized by their standard deviations (equation (6)). Additionally, we investigated the importance of each catchment characteristic by evaluating the term contained within the parentheses in equation (6) for each characteristic.

$$\mathrm{dist}_{a,b} = \sqrt{\sum_{j=1}^{J}\left(\frac{X_{a,j} - X_{b,j}}{\sigma_{X,j}}\right)^2} \tag{6}$$

[16] In equation (6), j is one of J catchment characteristics, $X_{a,j}$ is the value of that catchment characteristic for the ath gage combination's catchment area, $X_{b,j}$ is the value of that characteristic for the outlet gage's total drainage area, and $\sigma_{X,j}$ is the standard deviation of that catchment characteristic across the entire set of gage combinations' drainage areas.

3. Results

3.1. Overall ARM Skill

[17] Boxplots showing NSE, PBIAS, and RSR for simulations of discharge at 04165500 resulting from all potential most-downstream gage combinations are shown in Figure 3. Both monthly and daily simulations perform reasonably well for most combinations of most-downstream gages. Of the 185 possible combinations, 143 (136) of the combinations resulted in at least satisfactory daily (monthly) NSE values ($0.50 < \mathrm{NSE} \leq 1.00$), 81 (83) of which rated very good ($0.75 < \mathrm{NSE} \leq 1.00$). Likewise, 145 (137) combinations resulted in at least satisfactory daily (monthly) RSR values ($0.00 \leq \mathrm{RSR} \leq 0.70$), 85 (86) of which rated very good ($0.00 \leq \mathrm{RSR} \leq 0.50$). While nearly all combinations result in negative bias, the bias is within the recommended range for percent bias (PBIAS $\leq$25%) for nearly all combinations' simulated daily and monthly discharge. In fact, 120 (110) combinations resulted in very good PBIAS values (PBIAS $\leq \pm 10$). Although undocumented, flow augmentation is known to occur in the

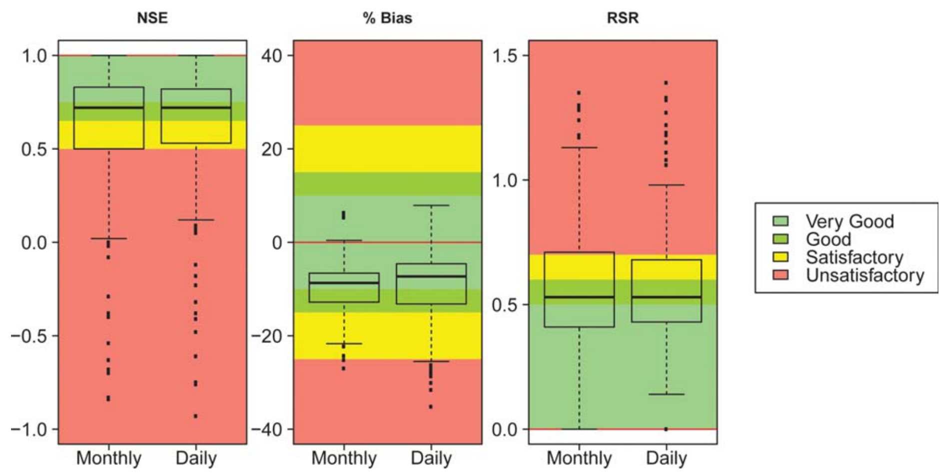

Figure 3. Nash-Sutcliffe Efficiency coefficient (NSE), Percent Bias (PBIAS), and RMSE observations standard deviation ratio (RSR) resulting from all potential combinations of most-downstream gages in the Clinton River basin. ARM output was compared with observed discharge at gage 04165500 at monthly and daily time steps. Plot backgrounds are colored by performance ratings recommended by *Moriasi et al*. [2007]. Note that one outlier is not shown. This outlier is produced by daily and monthly simulations using only gage 04164300. Goodness-of-fit statistics for the daily and monthly simulations produced by this gage are NSE = −6.03, PBIAS = −71.9%, RSR = 2.65 and −6.17, −52.3%, 2.68, respectively.

Clinton River due to effluent from wastewater treatment plants that divert water from the Detroit River, perhaps explaining some underestimation of ARM-simulated discharge at the outlet.

3.2. Influence of Area Ratio on ARM Skill

[18] As the fraction of the watershed that is gaged increases from zero to almost 0.3, the model skill appears to improve as certain gages are added into the most-downstream gage combinations, evident by increasing NSE, reduced absolute PBIAS, and reduced RSR (Figure 4). However, some gage combinations with very small catchment areas (even less than 10% of the outlet gage's catchment area) do result in very good model skill, with NSE near one, absolute PBIAS less than 10%, and RSR less than 0.5. An obvious reduction in skill occurs when the gaged fraction reaches about 0.3, evident by a sudden dramatic decrease in NSE, increase in absolute bias, and increase in RSR in Figure 4. This sudden reduction in skill occurs simultaneously with the addition of gage 04164500 into the combinations of most-downstream gages, suggesting that although the catchment area of gage 04164500 is large, the catchment's hydrologic response is not representative of the outlet's catchment. Beyond about 60% gaged, the addition of gage 04164000 into further combinations results in much improved simulations, with NSE approaching 1.0, average percent bias less than ±10%, and RSR less than 0.5 for daily and monthly simulations resulting from all combinations including gage 04164000. These results suggest that increasing the proportion of the outlet catchment that is gaged does not always improve ARM simulations and the success of a gage combination at simulating discharge at the outlet is related to more than just the fraction of the outlet's catchment that gages' catchments represents. Figure 4 also demonstrates the effect of temporal aggregation (daily vs. aggregated monthly) on model skill. While the NSE, PBIAS, and RSR statistics do not necessarily suggest differing performance at different time steps, the simulations of monthly discharge resulted in a notable reduction in the magnitude of residuals as a result of temporal aggregation.

3.3. Contribution of Individual Gages to ARM Skill

[19] To investigate the contribution of each individual gage to model skill and assess the assumption of spatial homogeneity of hydrologic response, we show log-log plots of modeled versus observed daily discharge at each gage when discharge is simulated using only one gage and find that hydrologic response is not homogeneous across the outlet's catchment (Figure 5). When considered individually, it is clear that gage 04164000 provides the best simulation of discharge at gage 04165500. Several gages stand out as problematic for simulating discharge at 04165500. Among these are 04163400, with negative NSE, 04164500 with large negative bias and negative NSE, 04160800 with large negative bias and low NSE, 04161580 with low NSE, and 04164300 with very large negative bias and very large negative NSE. These gages have catchments that are, for the most part, small relative to that of the outlet gage 04165500, especially 04163400, 04160800, 04161580, and 04164300 (all less than 5% gaged). However, gage 04164500 is a notable exception, with its catchment representing almost 30% of the total watershed area, and other gages with much smaller catchment areas have considerably better goodness-of-fit statistics (e.g., 04161800 and 04164100, whose catchment areas each represent less than 10% of that of the outlet gage).

3.4. Other Factors Influencing ARM Skill

[20] Inspection of the log-log plots of modeled versus observed discharge in relation to the gages' delineated catchments (Figure 5) provides some indication of

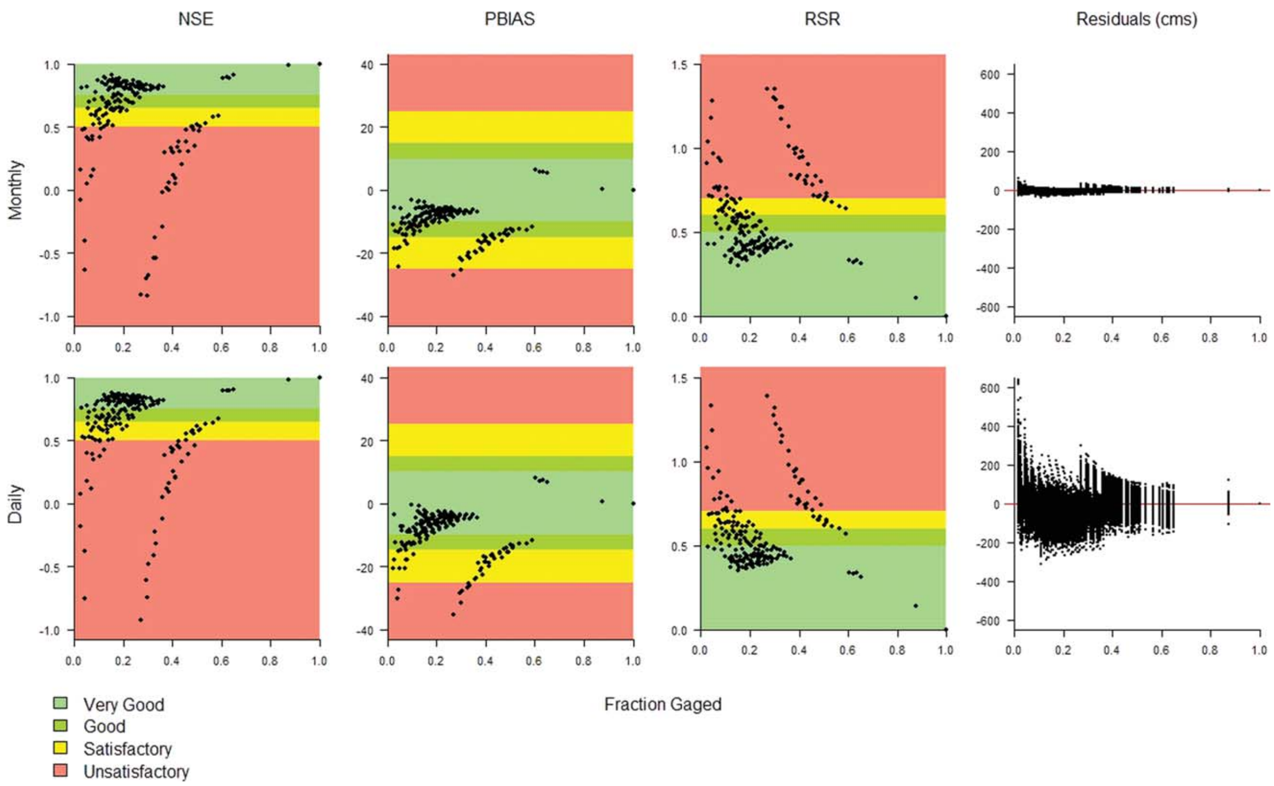

Figure 4. Nash-Sutcliffe Efficiency coefficient (NSE), Percent Bias (PBIAS), RMSE observations standard deviation ratio (RSR), and residuals versus the fraction of the basin that is gaged for all possible combinations of most-downstream gages for monthly and daily discharge simulations. Note that one outlier is not shown. This outlier is produced by daily and monthly simulations using only gage 04164300. Goodness-of-fit statistics for the daily and monthly simulations produced by this gage are NSE = −6.03, PBIAS = −71.9%, RSR = 2.65 and NSE = −6.17, PBIAS = −52.3%, RSR = 2.68, respectively.

underlying factors complicating the relationship between gaged area and model skill. For example, while the catchment area of gage 04164500 represents nearly 30% of the total watershed area, when it is used alone to simulate discharge at the outlet, it is among the poorest performers, both in terms of uncertainty (spread of the residuals) and goodness-of-fit statistics. A closer look at the catchment characteristics reveals that, while the catchment area of gage 04164500 is large, it is quite different from that of outlet gage 04165500 (Figure 5). For example, the catchment contains the largest portion of cultivated land (55%, compared to 20% for the outlet gage's catchment). Conversely, the relatively small catchment of gage 04161540 (10% of the outlet gage's catchment) is somewhat representative of the watershed of the outlet, with cultivated land making up 14%, developed land comprising 42%, and a comparable average permeability (5.6 in/h compared with 4 in/h for the outlet gage's catchment). Consequently, model skill of gage 04161540 is good, despite its small size. Additionally, the plots of modeled versus observed discharge in Figure 5 also confirm findings by *Archfield and Vogel* [2010] that distance between gages does not serve as a proxy for correlation of streamflow.

[21] Goodness-of-fit statistics are plotted against a measure of dissimilarity of the gaged catchments to the outlet gage's catchment area for the simulations resulting from each combination of most-downstream gage in order to evaluate the impact of similarity of catchment characteristics on model skill (Figure 6). Dissimilarity is shown for each physical characteristic considered, as well as an overall dissimilarity measure. In Figure 6, the horizontal red lines represent a perfect goodness-of-fit value, and the vertical blue lines are drawn where the gaged catchment characteristics are exactly the same as those of the outlet gage's catchment (the red circle is drawn around the point resulting from the simulation using gage 04165500). We found that each of these characteristics appears to be somewhat related to model skill and that similarity in cultivated area, developed area, stream density, and permeability appears particularly important for ARM skill (Figure 6).

4. Discussion

[22] In this analysis, we not only identify some key drivers of hydrologic response within the Clinton River watershed but also show that runoff for a moderately sized basin can be estimated from a small network of gages covering less than 20%–30% of the watershed, and that additional information from gages whose catchment area characteristics are dissimilar to those of the outlet's total catchment area may, in fact actually reduce model skill. This finding provides important insights into the rationale behind current ensemble modeling techniques, many of which implicitly assume that any additional information has value, but

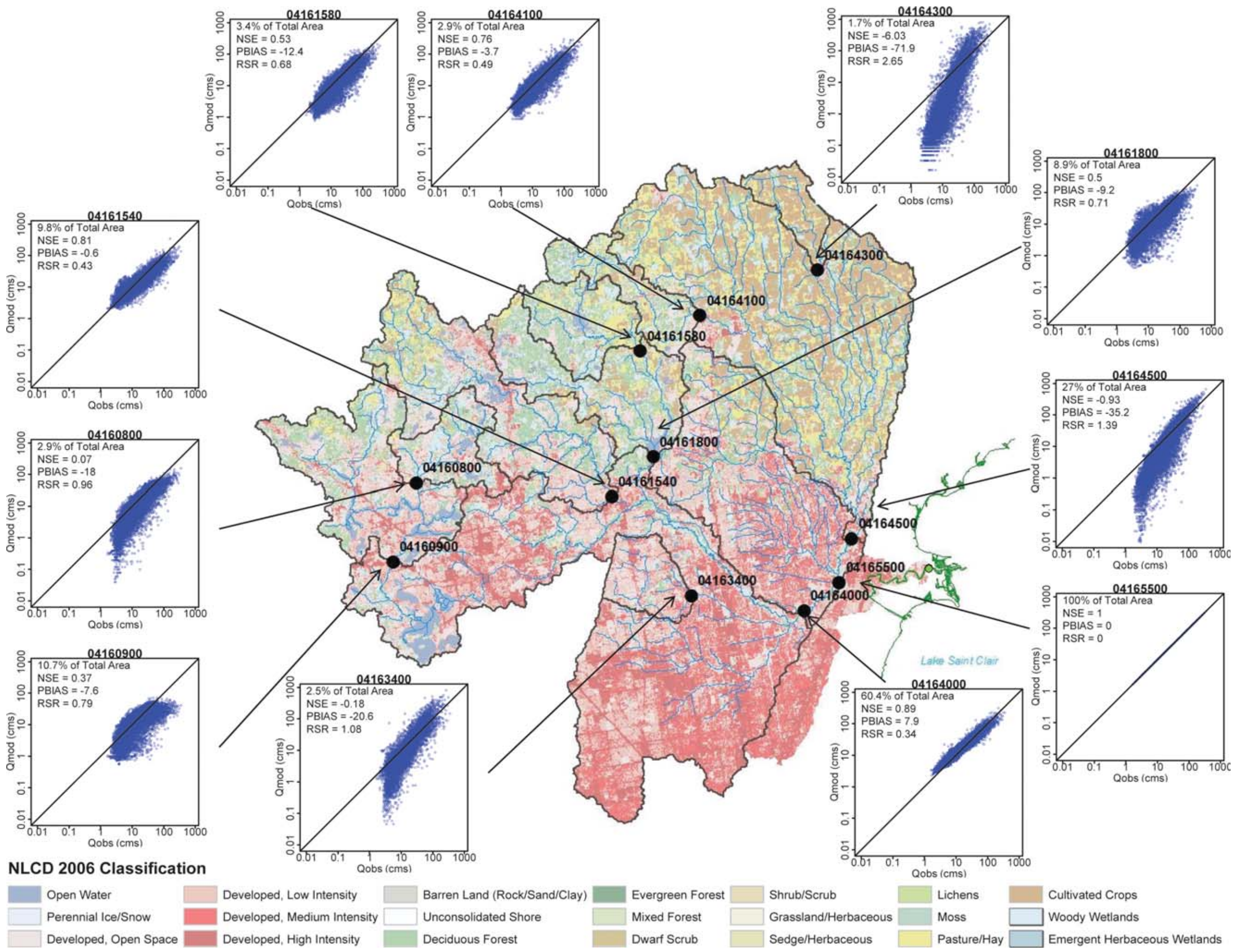

Figure 5. Goodness-of-fit statistics and log-log plots of modeled discharge versus observed daily discharge at gage 04165500 when only one gage is operating. Land cover data are from the National Land Cover Dataset [*Fry et al*., 2011].

that the value may be weighted based on catchment dissimilarity [e.g., *Reichl et al*., 2009]. This analysis shows that, in the case of large scale water budget estimation, adding information from some gages, even when weighted appropriately, may result in diminished model skill if the additional gages' catchment characteristics do not adequately represent the outlet's catchment.

[23] While an in-depth assessment of the contribution of each characteristic to model skill is outside the scope of this analysis, the similarity of the gaged area to the outlet gage's catchment is clearly an important consideration in selection of gages for inclusion in an ARM simulation. In particular, it appears that gages whose catchments are characterized by similar cultivated area, developed area, stream density, and permeability are likely to perform much better. Application of state-of-the-art GIS analysis and remote sensing may offer opportunities for selecting gages whose catchment areas are most similar in land use, morphology, and underlying hydrogeological characteristics.

[24] Note also that differences in runoff generation mechanisms likely contribute to differing model skill, and that many of the catchment characteristics in Figure 6 are factors determining whether runoff is generated via infiltration excess (i.e., Hortonian runoff) or saturation excess (i.e., Dunne runoff). If, as concurrent work [*Shen et al.*, 2013] has found, catchments of some gages are dominated by saturation excess, then the spatial extent of the area contributing to runoff is not the total subwatershed area but the areas adjacent to streams participating in saturation excess at any instant of time.

[25] It is also important to note that the groundwater shed cannot necessarily be considered to be equivalent to the surface watershed [e.g., *Hunt et al*., 1998], and this may be one explanation for poor model skill in some smaller watersheds. For example, the considerable underestimation of low flow when gage 04164300 is used to simulate discharge (see Figure 5) is likely due to a smaller proportional contribution of base flow in the hydrograph of 04164300. The under-representation of low flows (and likewise the overrepresentation of high flows) at 04164300 may also result from its different land use (76% cultivated). Similarly, the underestimation of low flows at gage 04163400 may be due to misrepresentation of base-flow characteristics resulting from its small size, as well as local impacts of

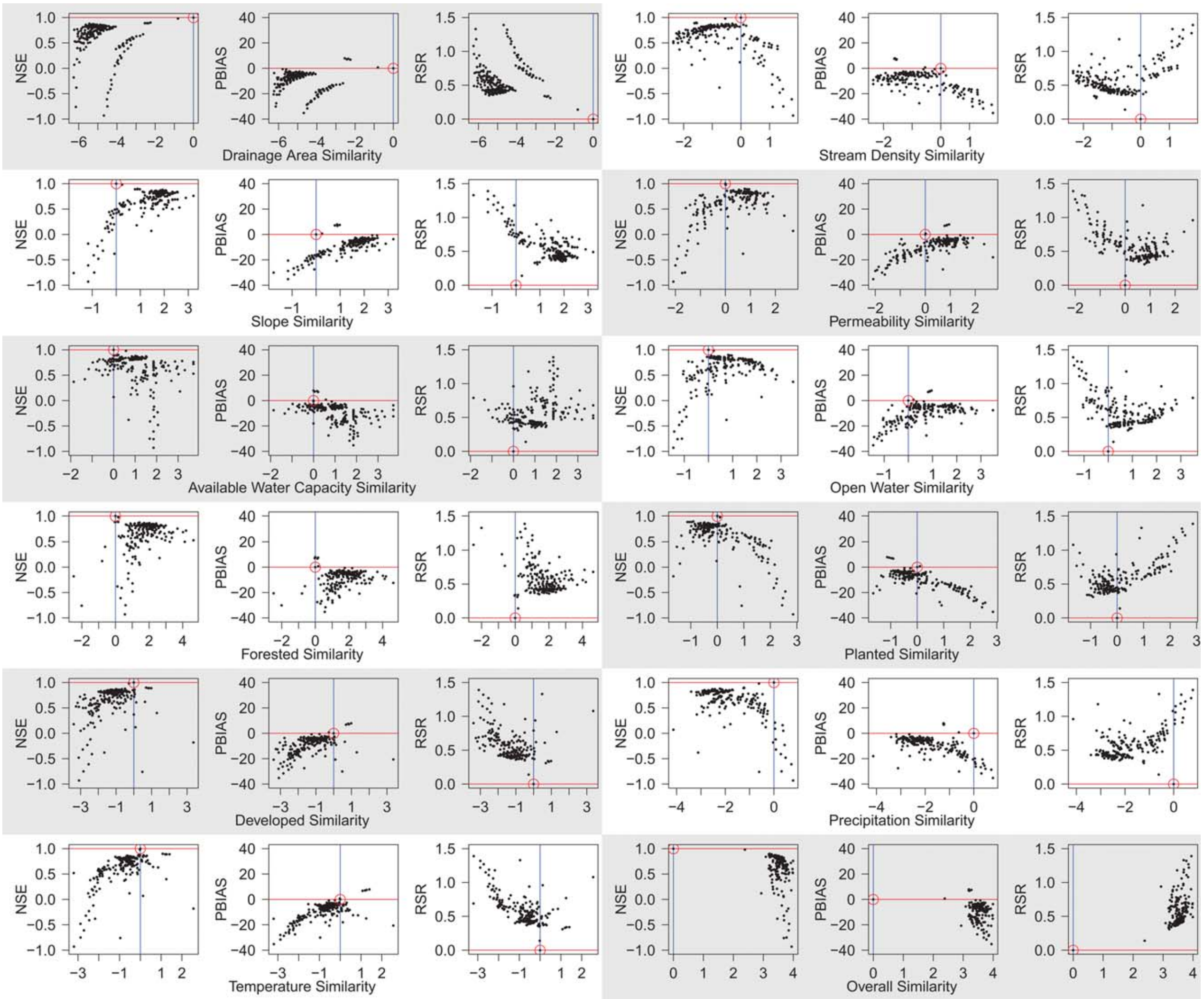

Figure 6. Goodness-of-fit statistics versus a measure of the dissimilarity (unitless) of the gaged area and the outlet watershed area. Overall dissimilarity was estimated using equation (6) and dissimilarity for individual characteristics was determined by evaluating the term inside the parentheses in equation (6). The vertical blue line is drawn at zero, representing exact similarity between the gaged catchment and the outlet's catchment. The horizontal red line is drawn at the optimal value for the goodness-of-fit statistic, for reference. If physically similar gage combinations result in better area-ratio predictions, then we expect the best goodness-of-fit result to intersect the blue vertical line (dissimilarity = 0).

urbanization (92% developed), which include decreasing low flows [*Martin et al*., 2012].

[26] Figures 5 and 6 demonstrate an inherent problem with extrapolation approaches to regionalization: for many physical characteristics, very few gage combinations' catchment areas are representative of the outlet gage's catchment area. For example, in the Clinton River watershed 169 of the 185 gage combinations' catchment areas have a lower fraction of developed area than the developed fraction of the outlet gage's catchment. In the Great Lakes basin, the largest urban areas are near the coasts, including Chicago, Toronto, Detroit, Montreal, Cleveland, Milwaukee, Buffalo, and Rochester. This is likely generally true worldwide, considering the challenges of monitoring coastal stream segments and the fact that 72% of the world's largest cities are located on or near the coast [*United Nations Department of Economic and Social Affairs*, 2012]. The heterogeneity in developed area may be especially important, as research has shown consistent impacts of urbanization across streamflow metrics including peak flows, low flows, flow durations, and flow variability [*Martin et al*., 2012]. The identification of spatial heterogeneity of urban area, as well as other physical characteristics determining ARM skill, facilitated by remote sensing and GIS-based terrain analysis, may allow for improvement of ARM simulation results, allowing for the selection of more appropriate gages to be included for extrapolating beyond intermittently gaged areas.

[27] In consideration of literature describing drivers of hydrologic response [e.g., *Jencso and McGlynn*, 2011; *Nippgen et al*., 2011; *Sawicz et al*., 2011; *Wagener et al*., 2007;

Wagener and Montanari, 2011], the influence of watershed characteristics on the ability of each gage to simulate flow at the outlet is not surprising. The significance in the analysis described here is in the potential for describing model skill and uncertainty of the area ratio method for simulating runoff in intermittently gaged basins using only catchment area and available discharge observations. Additionally, the analysis provides insight into selection of gages that should be prioritized for estimating discharge from coastal watersheds. For example, although gage 04164500 is likely important for other hydrologic analyses, the inclusion of this gage within ARM simulations actually worsens the overall model skill in some cases, because its catchment is physically different from the outlet's catchment (Table 2). While the combination of 04164500 and 04164000 does provide the best simulation of discharge, the addition of 04164500 offers only marginal improvement over 04164000, and in simulations not including 04164000, the inclusion of 04164500 actually reduces model skill and increases the uncertainty. The ability to select gages contributing to the best ARM discharge estimates could enhance ARM simulations, which is an attractive approach for simulating historical runoff for application to large-scale water balance modeling because of its minimal data requirements (complete or incomplete daily discharge records at gages and gage catchment area) and simple computational methods.

[28] This analysis provides insight into model skill and uncertainty of ARM within intermittently gaged areas. While we recognize that this does not fully describe ARM skill in always ungaged areas, it provides an important step in improving our understanding of the appropriateness of using ARM for simulating discharge in water balance studies. In the Lake Michigan basin, for example, the intermittently gaged portion represents nearly 80% of the total basin area, so enhancing our understanding of ARM skill in the intermittently gaged portions is a significant advancement for estimating discharge to the Great Lakes. While the ability of ARM to simulate discharge from the portion of the basin that is never gaged remains unclear, insight from the research investigating relationships between watershed characteristics and hydrologic response may contribute to a new, combined approach to estimate total flow from partially gaged basins, with ARM providing estimates for intermittently gaged portions, and a form of regionalization contributing estimates in totally ungaged portions.

[29] The decline in gaged portion of the Great Lakes basin in the last decade (Figure 1) is an example of the pervasive problem of a declining gage network. In fact, *Mishra and Coulibaly* [2010] found that a large part of the Canadian portion of the basin can be classified as either "deficit" or "highly deficit" in their stream gage networks. The recent downward trend in gaged area is especially troubling, considering concurrent dramatic changes in the Great Lakes water budget, evidenced by changing water levels. For example, Lake Erie experienced an unprecedented shift in seasonal water level cycles during water year 2012, with the second highest increase in runoff during November and December (0.2 m, occurring also in 1927) followed by the longest period of decreasing water levels during the spring runoff period, which would normally coincide with rising water levels (A. D. Gronewold and C. A. Stow, Unprecedented seasonal water level dynamics on one of the earths largest lakes, submitted to *Bulletin of the American Meteorological Society*, 2012). In light of the need for both maximizing the beneficial use of existing gage observations and planning for inclusion of new gage information into efforts to understand the water budget of the Great Lakes, analyses similar to that performed on the Clinton River should be carried out across all intermittently gaged Great Lakes subbasins.

[30] Results of this analysis provide an opportunity for development of a protocol for selecting gages to include in ARM simulations of runoff to the Great Lakes. In GLERL's current implementation of ARM, the protocol would be as follows. If gage 04165500 is operating, it provides the "most-downstream" observation of discharge to Lake St. Clair. If gage 04165500 is offline, however, then gage 04164000 and 04164500 provide the most-downstream observations if they are operational. If in the case that 04165500 and 04164000 are both offline, then ARM would select gages 04163400, 04160900, 04161540, 04161800, and 04164500 if they are all operational (59% of the outlet's catchment area). Our analysis shows, however, that this five-gage combination actually performs worse (NSE $=$ 0.67, PBIAS $= -12\%$, RSR $=$ 0.57) than the four-gage combination including all of those gages *except* gage 04164500 (32% of the outlet's catchment area) (NSE $=$ 0.81, PBIAS $= -3\%$, RSR $=$ 0.44), and that this reduction in performance is likely due to the dissimilarity in catchment characteristics. An improved protocol, informed by the analysis presented here, would therefore not necessarily choose all next-most-downstream gages if gage 04165500 is offline, but instead select the next available most-downstream gage combination *providing the best model skill*, i.e., 04163400, 04160900, 0416540, and 04161800. If this analysis is carried out on all Great Lakes subbasins, modern computing power would allow for the selection of the best gage combination for each Great Lakes subbasin on a given day, with implications for both improving our estimates of historical runoff and providing a basis for recalibration of the LBRM (and potentially other rainfall-runoff models).

5. Conclusions

[31] ARM is a simple method for extrapolating to ungaged areas to create synthetic time series of historical runoff. Although no catchment characteristics are considered for the implementation of ARM, the method performs reasonably well for simulating outflow from the Clinton River watershed, based on evaluation of the NSE, PBIAS, and RSR resulting from simulations involving 185 different combinations of gages. While most combinations performed at least satisfactorily based on these measures, it was clear that some combinations of gages performed better than others, and the inclusion of some gages into certain gage combinations actually had the effect of reducing model skill, owing to dissimilarity in their catchment characteristics compared with those of the outlet. By eliminating "bad" gages from the group of available gages from which to draw daily observations, ARM estimates can be improved within intermittently gaged areas. This is somewhat contrary to weighted ensemble approaches [e.g., *Beven and Freer*, 2001; *Reichl et al.*, 2009], which do not reflect the

possibility that additional gage information may in fact decrease model skill. Weighted ensemble approaches often employ any gage information available, under the assumption that more information is better, but weigh the relative value of that information according to some similarity or skill metric. Here we have provided a clear basis for a different perspective acknowledging thresholds on the relationship between the number of gages and their relative value in supporting flow estimates over large areas. *McIntyre et al.* [2005] present an approach that accommodates such thresholds; they build on established Bayesian methods described by *Beven and Freer* [2001], by updating prior likelihoods of a large sample of models based on the similarity of the gaged catchments to the ungaged catchments, eliminating donor catchments with a dissimilarity measure above a threshold value. While our assessment does suggest that some gages are better than others for the purpose of simulating flow to the Great Lakes, we are not advocating the elimination of gages, but instead presenting an analysis that allows for making the best use of available data in a time when available observations are becoming more scarce.

[32] Model skill did appear to be somewhat related to the simulations' fraction of gaged area. However, a noncontinuous relationship between model skill and gaged area suggested that additional catchment characteristics are important. Accordingly, the generalized assumption of spatial homogeneity in hydrologic response, while serving as a computationally efficient basis for Great Lakes regional basin-scale hydrological modeling for the past several decades, may not be particularly suitable for simulating flows in relatively small basins. We also find, however, that this same type of simple, rainfall-independent approach that relies exclusively on gage data can provide insight into patterns of spatial heterogeneity of hydrologic response and serve as a basis for both improving flow estimates in ungaged basins and for calibrating predictive runoff models. Exploration of the patterns of heterogeneity of hydrologic response with heterogeneity of catchment characteristics related to land cover, soils, morphology, and climate adds to the body of research confirming that catchment similarity influences the ability to simulate flow in ungaged basins [e.g., *Kay et al.*, 2007; *McIntyre et al.*, 2005; *Reichl et al.*, 2009; *Wagener et al.*, 2007]. The analysis also demonstrates that ARM may still offer an efficient means for simulating flow over ungaged areas, and that the method can be improved by first eliminating observations from gages that do not effectively simulate flow at the outlet of the intermittently gaged area. Assessment of the relationships between ARM skill and catchment dissimilarity may also contribute to improved estimates of discharge in the totally ungaged areas by further eliminating inclusion of observations from gages whose catchment characteristics are not representative of the combined intermittently gaged and totally ungaged area of a subbasin (e.g., contributing area of the Clinton River mouth in our example). Catchment characteristics derived from remote sensing data and GIS-based terrain analysis provide opportunities for selecting the most appropriate gages for inclusion in ARM simulations.

[33] Our analysis provides the basis for selection of appropriate combinations of gages for ARM simulations, which could have implications for both improving historical runoff simulations and calibrating rainfall-runoff models over large regions. It could be argued that the approach of selecting a set of gages that provide the best ARM simulations contrasts with recommendations that model parameters be linked to landscape attributes to simulate effects of future changes in those attributes. While we acknowledge the potential value of physically based models which are linked directly to landscape attributes, we also recognize that many rainfall-runoff model applications span time scales in which these types of changes are unlikely to have an impact (i.e., hours, days, and months). Certainly, over large time scales, these changes are likely to be important [*Milly et al.*, 2008], and considerable research is being invested in understanding the importance of these changes. In future research, and as part of the ongoing binational Great Lakes Runoff Intercomparison Project (GRIP), we will test this assumption over broader spatial and temporal scales and compare the skill of this simple approach with other more complex models, including (but not limited to) Analysis of Flows in Networks of Channels (AFINCH) [described by *Holtschlag*, 2009], LBRM, the Sacramento Soil Moisture Accounting model and Snow-17 models encoded within National Weather Service's Community Hydrologic Prediction System (NWS), and several configurations of MESH. Evaluation of the ARM regionalization method in a side-by-side comparison with a regionalization model incorporating physical characteristics of catchments at a semidistributed level (AFINCH), as well as physically based lumped conceptual models (e.g., NWS and LBRM) and distributed models (e.g., MESH) will provide significant further insight into the value of individual daily discharge observations for model parameterization and/or assimilation into large scale water balance modeling efforts.

[34] **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. Additionally, the authors thank three anonymous reviewers who provided constructive comments, as well as Thorsten Wagener for his thoughtful early review of this manuscript. This publication is GLERL Contribution 1659.

References

Anderson, E. J., D. J. Schwab, and G. A. Lang (2010), Real-time hydraulic and hydrodynamic model of the St. Clair River, Lake St. Clair, Detroit River system, *J. Hydraul. Eng.*, *136*(8), 507–518, doi:10.1061/(ASCE)HY.1943-7900.0000203.

Andrews, F., B. Croke, and A. J. Jakeman (2011), An open software environment for hydrological model assessment and development, *Environ. Model. Software*, *26*(10), 1171–1185, doi:10.1016/j.envsoft.2011.04.006.

Archfield, S. A., and R. M. Vogel (2010), Map correlation method: Selection of a reference streamgage to estimate daily streamflow at ungaged catchments, *Water Resour. Res.*, *46*, W10513, doi:10.1029/2009WR008481.

Beven, K. (2007), Towards integrated environmental models of everywhere: uncertainty, data and modelling as a learning process, *Hydrol. Earth Syst. Sci.*, *11*(1), 460–467, doi:10.5194/hess-11-460-2007.

Beven, K., and J. Freer (2001), Equifinality, data assimilation, and uncertainty estimation in mechanistic modelling of complex environmental systems using the GLUE methodology, *J. Hydrol.*, *249*(1–4), 11–29, doi:10.1016/S0022-1694(01)00421-8.

Bulygina, N., N. Mcintyre, and H. Wheater (2009), Conditioning rainfall-runoff model parameters for ungauged catchments and land management impacts analysis, *Hydrol. Earth Syst. Sci.*, *13*(6), 893–904.

Carrillo, G., P. A. Troch, M. Sivapalan, T. Wagener, C. Harman, and K. Sawicz (2011), Catchment classification: hydrological analysis of catchment behavior through process-based modeling along a climate gradient, *Hydrol. Earth Syst. Sci.*, *15*(11), 3411–3430, doi:10.5194/hess-15-3411-2011.

Coon, W. F., E. A. Murphy, D. T. Soong, and J. B. Sharpe (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. Geol. Surv. Open-File Rep. 2011-1202*, U.S. Geol. Surv., Reston, Va.

Croley, II, T. E., and H. C. Hartmann (1986), NOAA technical memorandum ERL GLERL-61: Near-real-time forecasting of large-lake water supplies; a user's manual, *Tech. Rep.*, ERL GLERL-61, U.S. Dept. of Commer., Nat. Oceanic and Atmos. Admin., Great Lakes Environ. Res. Lab., Ann Arbor, Mich.

Croley, II, T. E., and C. He (2002), Great Lakes Large Basin runoff modeling, paper presented at Second Federal Interagency Hydrologic Modeling Conference, Subcommittee on Hydrology of the Interagency Advisory Committee on Water Data, Las Vegas, Nev.

Deacu, D., V. Fortin, E. Klyszejko, C. Spence, and P. D. Blanken (2012), Predicting the Net Basin Supply to the Great Lakes with a hydrometeorological model, *J. Hydrometeorol.*, *13*, 1739–1759, doi :10.1175/JHM-D-11-0151.1.

Doherty, J. (2011), Modeling: Picture perfect or abstract art? *Ground Water*, *49*(4), 455, doi:10.1111/j.1745-6584.2011.00812.x.

Doherty, J., and J. M. Johnston (2003), Methodologies for calibration and predictive analysis of a watershed model, *J. Am. Water Works Assoc.*, *39*(2), 251–265, doi:10.1111/j.1752-1688.2003.tb04381.x.

Emerson, D. G., A. V. Vecchia, and A. L. Dahl (2005), Evaluation of drainage-area ratio method used to estimate streamflow for the Red River of the North Basin, North Dakota and Minnesota, *Scientific Investigations Rep. 20055017*, U.S. Geol. Surv. (USGS), Reston, Va.

Fry, J., G. Xian, S. Jin, J. Dewitz, C. Homer, L. Yang, C. Barnes, N. Herold, and J. Wickham (2011), Completion of the 2006 National Land Cover Database for the conterminous United States, *Photogramm. Eng. Remote Sensing*, *77*(9), 858–864.

Grimaldi, S., A. Petroselli, G. Alonso, and F. Nardi (2010), Flow time estimation with spatially variable hillslope velocity in ungauged basins, *Adv. Water Resour.*, *33*(10), 1216–1223, doi:10.1016/j.advwatres.2010.06.003.

Gronewold, A. D., and V. Fortin (2012), Advancing Great Lakes hydrological science through targeted binational collaborative research, *Bull. Am. Meteorol. Soc.*, *93*(12), 1921–1925, doi:10.1175/BAMS-D-12-00006.1.

Gronewold, A. D., A. H. Clites, T. S. Hunter, and C. A. Stow (2011), An appraisal of the Great Lakes advanced hydrologic prediction system, *J. Great Lakes Res.*, *37*(3), 577–583, doi:10.1016/j.jglr.2011.06.010.

Holtschlag, D. (2009), Application guide for AFINCH (Analysis of Flows in Networks of Channels) described by NHDPlus, Sci. Invest. Rep. 2009–5188, U.S. Geol. Surv., Reston, Va.

Hortness, J. E. (2006), Estimating low-flow frequency statistics for unregulated streams in Idaho, *Sci. Invest. Rep. 20065035*, U.S. Geol. Surv. (USGS), Reston, Va.

Hunt, R., M. Anderson, and V. Kelson (1998), Improving a complex finite-difference ground water flow model through the use of an analytic element screening model, *Ground Water*, *36*(6), 1011–1017.

Jencso, K. G., and B. L. McGlynn (2011), Hierarchical controls on runoff generation: Topographically driven hydrologic connectivity, geology, and vegetation, *Water Resour. Res.*, *47*, W11527, doi:10.1029/2011WR010666.

Juckem, P. F., P. C. Reneau, and D. M. Robertson (2012), Estimation of natural historical flows for the Manitowish River near Manitowish Waters, Wisconsin, Sci. Invest. Rep. 20125135, U.S. Geol. Surv. (USGS), Reston, Va.

Kay, A. L., D. A. Jones, S. M. Crooks, T. R. Kjeldsen, and C. F. Fung (2007), An investigation of site-similarity approaches to generalisation of a rainfall-runoff model, *Hydrol. Earth Syst. Sci.*, *11*(1), 500–515, doi:10.5194/hess-11-500-2007.

Legates, D. R., and G. McCabe Jr. (1999), Evaluating the use of ''goodness-of-fit'' measures in hydrologic and hydroclimatic model validation, *Water Resour. Res.*, *35*(1), 233–241, doi:10.1029/1998WR900018.

Lofgren, B. M. (2004), A model for simulation of the climate and hydrology of the Great Lakes basin, *J. Geophys. Res.*, *109*, 1–20, doi:10.1029/2004JD004602.

Mao, D., and K. A. Cherkauer (2009), Impacts of land-use change on hydrologic responses in the Great Lakes region, *J. Hydrol.*, *374*(1–2), 71–82, doi:10.1016/j.jhydrol.2009.06.016.

Martin, E. H., C. Kelleher, and T. Wagener (2012), Has urbanization changed ecological streamflow characteristics in Maine (USA)? *Hydrol. Sci.*, *57*(7), 1337–1354.

McIntyre, N., H. Lee, H. Wheater, A. Young, and T. Wagener (2005), Ensemble predictions of runoff in ungauged catchments, *Water Resour. Res.*, *41*, W12434, doi:10.1029/2005WR004289.

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

Mishra, A., and P. Coulibaly (2010), Hydrometric network evaluation for Canadian watersheds, *J. Hydrol.*, *380*(3–4), 420–437, doi:10.1016/j.jhydrol.2009.11.015.

Moriasi, D. N., J. G. Arnold, M. W. Van Liew, R. L. Bingner, R. D. Harmel, and T. L. Veith (2007), Model evaluation guidelines for systematic quantification of accuracy in watershed simulations, *Trans. Am. Soc. Agric. Biol. Eng.*, *50*(3), 885–900.

Nippgen, F., B. L. McGlynn, L. A. Marshall, and R. E. Emanuel (2011), Landscape structure and climate influences on hydrologic response, *Water Resour. Res.*, *47*, W12528, doi:10.1029/2011WR011161.

Noto, L., and G. La Loggia (2007), Derivation of a distributed unit hydrograph integrating GIS and remote sensing, *J. Hydrol. Eng.*, *12*(6), 639–650, doi:10.1061/?ASCE?1084-0699?2007?12:6?639?.

Pietroniro, A., et al. (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*(4), 1279–1294, doi:10.5194/hess-11-1279-2007.

Reichl, J. P. C., A. W. Western, N. R. McIntyre, and F. H. S. Chiew (2009), Optimization of a similarity measure for estimating ungauged streamflow, *Water Resour. Res.*, *45*, W10423, doi:10.1029/2008WR‘007248.

Robertson, D. M., and D. A. Saad (2011), Nutrient inputs to the Laurentian Great Lakes by source and watershed estimated using SPARROW watershed models., *J. Am. Water Resour. Assoc.*, *47*(5), 1011–1033, doi:10.1111/j.1752-1688.2011.00574.x.

Sawicz, K., T. Wagener, M. Sivapalan, P. A. Troch, and G. Carrillo (2011), Catchment classification: empirical analysis of hydrologic similarity based on catchment function in the eastern USA, *Hydrol. Earth Syst. Sci.*, *15*(9), 2895–2911, doi:10.5194/hess-15-2895-2011.

Shen, C., J. Niu, and M.S. Phanikumar (2013), Evaluating controls on coupled hydrologic and vegetation dynamics in a humid continental climate watershed using a subsurface–land surface processes model, *Water Resour. Res.*, doi: 10.1002/wrcr.20189.

Singh, R., T. Wagener, K. van Werkhoven, M. E. Mann, and R. Crane (2011), A trading-space-for-time approach to probabilistic continuous streamflow predictions in a changing climate accounting for changing watershed behavior, *Hydrol. Earth Syst. Sci.*, *15*(11), 3591–3603, doi:10.5194/hess-15-3591-2011.

Sivapalan, M., K. Takeuchi, S. W. Franks, V. K. Gupta, H. Karambiri, and V. Lakshmi (2003), IAHS Decade on Predictions in Ungauged Basins (PUB), 2003–2012: Shaping an Exciting Future for the Hydrological Sciences, *Hydrol. Sci.*, *48*(6), 857–880, doi:10.1623/hysj.48.6.857.51421.

United Nations Department of Economic and Social Affairs (2012), *World urbanization prospects:* The 2011 revision, *Tech. Rep.*, ST/ESA/SER.A/322, United Nations, New York.

Vrugt, J. A., and C. J. F. Ter Braak (2011), DREAM(D): an adaptive Markov Chain Monte Carlo simulation algorithm to solve discrete, noncontinuous, and combinatorial posterior parameter estimation problems, *Hydrol. Earth Syst. Sci.*, *15*(12), 3701–3713, doi:10.5194/hess-15-3701-2011.

Wagener, T., and A. Montanari (2011), Convergence of approaches toward reducing uncertainty in predictions in ungauged basins, *Water Resour. Res.*, *47*, W06301, doi:10.1029/2010WR009469.

Wagener, T., M. Sivapalan, P. Troch, and R. Woods (2007), Catchment classification and hydrologic similarity, *Geogr. Compass*, *1*(4), 901–931, doi:10.1111/j.1749-8198.2007.00039.x.

Wood, E. F., et al. (2011), Hyperresolution global land surface modeling: Meeting a grand challenge for monitoring Earth's terrestrial water, *Water Resour. Res.*, *47*, W05301, doi:10.1029/2010WR010090.

Yadav, M., T. Wagener, and H. Gupta (2007), Regionalization of constraints on expected watershed response behavior for improved predictions in ungauged basins, *Adv. Water Resour.*, *30*(8), 1756–1774, doi:10.1016/j.advwatres.2007.01.005.

Zambrano-Bigiarini, M. (2011), *hydroGOF: Goodness-Of-Fit Functions for Comparison of Simulated and Observed Hydrological Time Series*, R package version 0.3-2. [Available at http://CRAN.R-project.org/package=hydroGOF.]

Zhang, Z., T. Wagener, P. Reed, and R. Bhushan (2008), Reducing uncertainty in predictions in ungauged basins by combining hydrologic indices regionalization and multiobjective optimization, *Water Resour. Res.*, *44*, W00B04, doi:10.1029/2008WR006833.