Self-Learning Cellular Automata for Forecasting Precipitation from Radar Images

Hong Li[1]; Gerald A. Corzo Perez[2]; Carlos A. Martinez[3]; and Arthur E. Mynett, M.ASCE[4]

Abstract: This paper presents a new forecasting methodology that uses self-learning cellular automata (SLCA) for including variables that consider the spatial dynamics of the mass of precipitation in a radar forecast model. Because the meteorological conditions involve nonlinear dynamic behavior, an automatic learning model is used to aid the cellular automata rules (SLCA). The new methodology is applied to the western part of England (Brue river basin) using NIMROD data. The radar information from 1 month of hourly radar measurements is used. Two models, a regression model tree (MT) and an artificial neural network (ANN) model, are used to learn the dynamics of the spatially local effects within the cellular automation (CA) neighboring areas. A spatial correlation (tracking pattern) reference model is built for comparing the first hour of precipitation forecast. Model results show that the SLCA is more accurate than conventional tracking. Furthermore, it appears that this technique can be extended to include other important atmospheric variables in forecasting processes. **DOI: 10.1061/(ASCE)HE.1943-5584.0000646.** *© 2013 American Society of Civil Engineers.*

CE Database subject headings: Precipitation; Forecasting; Radar; Imaging techniques.

Author keywords: Precipitation forecasting; Radar images; Self-learning cellular automata; NIMROD data; Radar tracking.

Introduction

Weather radar systems are often used to obtain rainfall information with high spatial and temporal resolutions, and one of the main advantages is that they are able to scan large areas and take many measurements from a single location in real time.

Previous studies have been carried out using different methodologies for precipitation forecasting with radar rainfall data. Pessoa et al. (1993) showed that using weather radar and a distributed rainfall-runoff model can improve real-time flow forecasting in flood situations. This was concluded after analyzing the flood forecast with various properties of the spatial information of the radar-derived rainfall data. Sun et al. (2000) predicted floods using radar and rain gauge data and showed that the estimated rainfall was considerably improved when it combined both the rain gauge and radar data. Li et al. (2004) coupled weather radar rainfall data with a distributed hydrologic model for real-time forecasting. The results of the hydrologic models with weather radar data were better than those simulated by the rain gauge data. Cluckie et al. (2006) combined rainfall prediction from radar and high-resolution numerical weather prediction (NWP) model results at the mesoscale level for exploring the advantages of this combination. They obtained high-accuracy forecasts at field locations using radar and NWP models.

The use of radar for forecasting precipitation in fast-response basins is of particular importance. Some of the most used techniques, such as tracking on the basis of patterns and correlations, do not take into consideration the spatial dynamics and extension of the rainfall volume. Numerical weather prediction models do consider much more information, but in general, they lack spatial representation and their accuracy is less than the radar forecast for short lead times.

In this paper, the authors present a new forecasting methodology that integrates the use of cellular automata for exploring the spatial extension of the amount of precipitation in a radar forecast model. Because the meteorological conditions involve nonlinear dynamic behavior, an automatic learning model is used to aid the cellular automata rules, which the authors named self-learning cellular automata (SLCA). The SLCA is based on principles that have been presented in Li (2009) and Li et al. (2010). The new methodology is applied to the western part of England (Brue river basin), using NIMROD data (Golding 1998). The radar information for precipitation from 1 month of hourly radar measurements is used. Two models, a regression model tree (MT) and an artificial neural network (ANN), are used to derive the transition rules for the SLCA model. A spatial correlation (tracking pattern) reference model is built for comparing the first hour of the precipitation forecast.

Methodology

Cellular Automata

Cellular automation (CA) models deal with spatial variation and local interactions. They provide simple discrete deterministic mathematical models for physical, biological, and computational systems in which many simple components act together to produce

[1]International Water Association, Koningin Julianaplein 2, 2595 AA, The Hague, Netherlands (corresponding author). E-mail: hong.li@iwahq.org

[2]Hydrology and Quantitative Water Management Group, Centre for Water and Climate, Wageningen Univ., Droevendaalsesteeg 4, 6708 PB Wageningen, Netherlands; Technilogico de Monterrey, Centro del Agua para Amrica, Latina y el Caribe, Ave. Garza Sada 2501 Sur Col. Tecnolgico 64849 Monterrey, N.L., Mexico.

[3]Grupo de Gestión Integrada del Recurso Hídrico—GIRH, Instituto Cinara, Universidad del Valle, Clle 13 No. 100-00, AA25157, Cali, Colombia.

[4]UNESCO-IHE Institute for Water Education, P.O. Box 3015, 2601DA Delft, Netherlands; Delft Univ. of Technology, Faculty CiTG, P.O. Box 5048, 2600 GA Delft, Netherlands.

Note. This manuscript was submitted on May 12, 2011; approved on April 30, 2012; published online on January 15, 2013. Discussion period open until July 1, 2013; separate discussions must be submitted for individual papers. This paper is part of the ***Journal of Hydrologic Engineering***, Vol. 18, No. 2, February 1, 2013. © ASCE, ISSN 1084-0699/2013/2-206-211/$25.00.

complicated patterns of behavior (Packard and Wolfram 1985). A cellular automaton consists of a regular uniform lattice of cells, usually infinite in extent, with a discrete variable at each cell (Wolfram 1986). A cellular automaton evolves in discrete time steps. The variables at each cell are updated simultaneously, on the basis of the values of the variables in their neighborhood at the preceding time step, and according to a predefined set of local rules. The neighborhood of a cell is typically assumed to be the cell itself plus all its immediately adjacent cells. There are a great number of particular CA models, each one being a specific system.

A simple CA can be defined by a lattice, L; a state space, Q; a neighborhood scheme, N; and a local transition function, f (Balzter et al. 1998)

$$CA = < L, Q, N, F > \quad (1)$$

Finding appropriate rules for a cellular automaton model is a crucial task (Balzter et al. 1998) because the rules require the most attention to obtain proper spatiotemporal dependence. One way to obtain the rules is to generate them automatically using the available data and machine learning techniques (Li et al. 2010), such as artificial neural networks.

Self-Learning Cellular Automata

The SLCA is defined in this paper as a cellular automata model developing its own transition function, f, using data-driven techniques like artificial neural networks and model trees.

For complex phenomena with limited understanding of mechanisms, traditional transition rules such as if-then rules are sometimes insufficient to represent nonlinear relations among neighboring cells. The use of ANN is hypothesized to supply an automatic learning mechanism for such nonlinear transition rules and can also be used spatially by training of known spatial data to forecast future spatial values on the basis of the neighboring information (Fig. 1). Therefore, the combination of artificial neural networks and CA can be a useful alternative modeling method to supply rapid prediction in situations in which the spatially local interactions are important. The combination of ANN and CA has previously been used in the modeling of urban development by Almeida et al. (2008), in the emulation of physically based water quality modeling results by Li et al. (2007), and for filling in missing data in remotely sensed images by Li (2009) and Li et al. (2010).

In this research, the learning of one type of common ANN, multilayer perceptron neural network (MLPs), is done through the minimization of the root-mean-square error (RMSE). This minimization is based on the Levenberg-Marquardt algorithm (Levenberg 1944; Marquardt 1963). Because this algorithm has fast convergence, the selected iteration (epoch) number was 150 for the entire experiment. The learning rate (0.1) and momentum term (0.7) in the algorithm were the same for all scenarios. More information and explanation relating to the parameters and algorithm can be found in Levenberg (1944) and Marquardt (1963). To select the most adequate ANN-MLP, a cross-validation technique was included on the basis of a 10-fold sampling of training and verification data. Each fold commonly contains 90% of the data for training and 10% for verification. As a result, the best model in this process is selected.

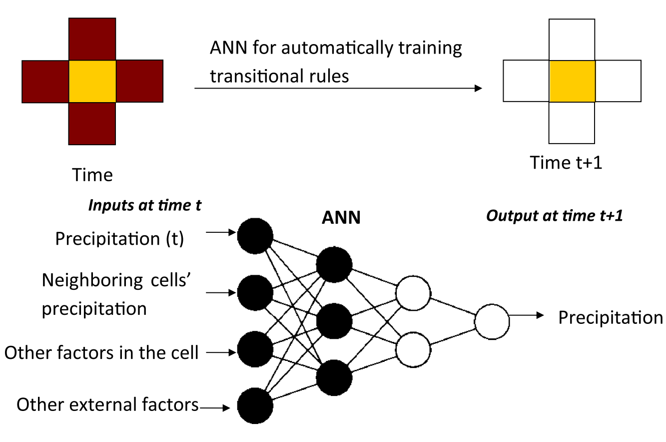

Fig. 1. SLCA using ANN for spatial forecasting

Alternatively, other data-driven techniques such as MTs can be used for automatically training CA transition rules, which have also been applied in this research. A model tree is a classification method in which the data are split into classes or into a number of simple tasks. A model tree belongs to a class of modular models that uses hard (i.e., yes-no) splits of input space to create regions, progressively narrowing the regions of the input space. Thus, the model tree is a hierarchical (or tree-like) modular model that has splitting rules in nonterminal nodes and the expert models at the leaves of the tree. In M5 model trees, the expert models are simple linear regression equations derived by fitting to the nonintersecting data subsets. Once these models are formed recursively in the leaves of the hierarchical tree, then prediction with the new input vector would consist of following two steps: (1) attributing the input vector to a subspace by following the tree and (2) running the corresponding model. More information about model trees can be found in studies such as those conducted by Quinlan (1986, 1992), Witten and Frank (2005), and Corzo et al. (2009).

Radar Tracking

Radar tracking (RT) is used as a reference model in this study. It is based on the persistence of the rainfall pattern (Lagrangian persistence). Radar tracking predicts new precipitation patterns on the basis of displacement in space and time of the rainfall in existing cells by means of a cross-correlation between consecutive radar scans. Its maximum correlation shows the displacement of the rainfall event. The radar-tracking echo is successful when it has been predicted 1 or 2 h ahead, but the error increases with longer lead times. Previous studies using this method have shown that an improved quantitative precipitation forecast (QPF) with a lead time of 3 h can be reached (Cluckie et al. 2006).

The spatial correlation is determined by a MATLAB (2010) function called xcorr2, which is applied to return the cross-correlation of matrices **A** and **B**. If matrix **A** has dimensions ($m1$, $n1$) and matrix **B** has dimensions ($m2$, $n2$), then the equation for the two-dimensional discrete cross-correlation is

$$C(i,j) = \sum_{m=0}^{(m1-1)} \sum_{n=0}^{(n1-1)} A(m,n) \times \mathrm{conj}[B(m+i, n+j)] \quad (2)$$

where $0 \le i < m1 + m2 - 1$ and $0 \le j < n1 + n2 - 1$.

Methods for Error Analysis

Aside from commonly used RMSEs, the authors have also analyzed the quantitative differences between spatial patterns exhibited in radar maps and modeled maps. The measures for the accuracy of modeling results for spatial pattern dynamics are the probability of detection (PoD), the false alarm rate (FAR) (Stanski et al. 1989; Werner et al. 2005), and the critical success index (CSI) (Schaefer 1990). For all these methods, the rain threshold is set to be 0.1 mm

J. Hydrol. Eng. 2013.18:206-211.

of rain. The calculations of PoD, CSI, and FAR are shown in Eqs. (3)–(5), respectively. The PoD should answer, What fraction of the observed yes events (higher than the threshold) was correctly forecasted? This measure has a perfect forecast when the value is 1. However, as is shown, it is only sensitive to hits and ignores false alarms (FAs) compared with radar maps (Hits). The PoD needs to be combined with the opposite situation of failure. The FAR answers, What fraction of the predicted yes events actually did not occur? Similarly to the PoD, the FAR goes from 0 to 1, and it is complementary in the sense that it ignores the missed cells. The CSI indicates how well the modeled precipitation spatial pattern corresponds to that observed in radar maps, which considers the influences from both false alarm and missed cells

$$\text{Probability of detection: PoD} = \text{Hits}/(\text{Hits} + \text{Missed}) \quad (3)$$

$$\text{Critical success index: CSI} = \text{Hits}/(\text{Hits} + \text{Missed} + \text{FA}) \quad (4)$$

$$\text{False alarm rate: FAR} = \text{FA}/(\text{Hits} + \text{FA}) \quad (5)$$

Model Setup

Radar Data Description

The information used in this study comes from the UK Meteorological Office, NIMROD system (Golding 1998), which is a now-casting system that analyzes and forecasts weather variables including precipitation amount in a very short period of time. The processes carried out hourly include processing of remotely sensed imagery, data fusion, and extrapolation forecast merging with numerical weather prediction model output. The radar processing includes image quality control, echo removal, and vertical profile correction. The imagery is calibrated against the radar composite and combined with lightning data. The NIMROD technique was used for extrapolating the precipitation distribution by selecting the optimum vector for each precipitation object.

This study uses the precipitation maps retrieved from 5-km-grid hourly radar Cobbacombe data covering Devon, United Kingdom, reaching a 250-km range, 31 days in total in October 2004, which were obtained from the British Atmospheric Data Centre (http://www.badc.nerc.ac.uk). The site has been selected because of its wide availability of data, and the period of time selected was considered to contain an important number of events.

A radar image is composed of 100×100 pixels (10,000 cells). Fig. 2 shows only radar images in which there are nonzero cells (rain), plotted against time (hours). In Fig. 2, it is shown that within the month of the radar data (hourly data) used by the authors, the maximum coverage of precipitation in one radar image was 6,000 cells. In several images, the precipitation covered more than 5,000 cells. Because of the mass conservation in the radar tracking and the possible creation of convective rain, the radar images with less than 1,000 cells (10% of the total radar images) covered by precipitation were not considered in the model setup.

Model Setup

Two models, a regression MT and an ANN are used to learn the dynamics of the CA transition rules with a selected neighboring scheme. A spatial correlation (tracking pattern) reference model is built for comparing the first hour of the precipitation forecast. The model flowchart is shown in Fig. 3. Using the CA concept, the inputs were composed of the values of the cell and of the four adjacent cells. The transition functions are derived by using MT and ANN, and the output of the model is 1-h lead time of rain values for the cell. As shown in Fig. 3, the modeling process includes data analysis, data splitting, and random sampling of the data sets; data learning to train the model; and then testing and validating the trained model. For this, a number of primary steps were used as follows:

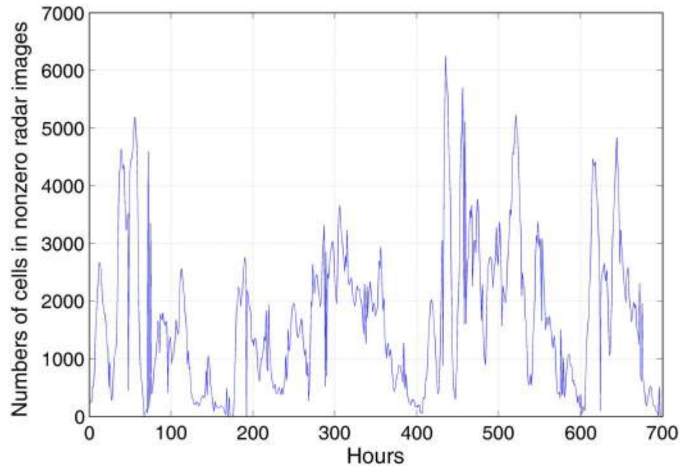

Fig. 2. Total numbers of wet cells in each nonzero radar image

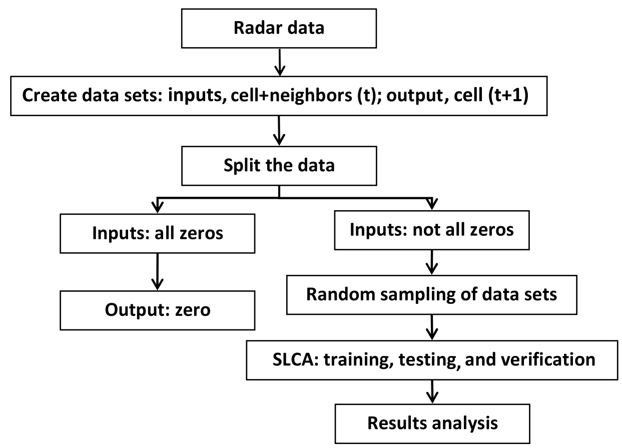

Fig. 3. Modeling process

1. The statistical and graphical representation of the samples was done to identify the radar images to be removed and to understand the relative variability of the cell patterns. Statistical analyses of radar images on mean, standard derivation, and maximum were measured from each radar image. The selected nonzero rain radar images were extracted, and a graph of the accumulated radar image precipitation was calculated.
2. Cells were arranged in a matrix of columns that contain neighbors and the actual value of the cell. The last column contains the 1-h-ahead cell values. Each row is composed of one combination of consecutive cells of radar images. Cell samples (rows) that only contain zero values (input zero—output zero) were removed. After this, a random sampling was performed. The process of random sampling is based on iteratively comparing samples and their statistical deviations between different training and validation data sets. The description of the random-sampling process can be found in Elshorbagy et al. (2010).
3. The SLCA algorithm was set up by first defining the number of rules required for the radar image cells. The training was performed by identifying models (rules) for the neighbors that

will apply to the whole radar image, so all rules found will apply to every cell in the region. This implicitly indicates that there is a spatial physical homogeneity between the hydrology and meteorology that creates the rainfall phenomena. This is captured in the equations that are incorporated into the ANN model or the MT.

Results and Analysis

Several examples of modeling results are shown in this section, including the results of an SLCA model using ANN or MT and the results of a radar-tracking model (Fig. 4), as well as the performance analyses for SLCA models compared with radar data and radar-tracking models only (Figs. 5 and 6).

In Fig. 4, it is shown that the SLCA gives better spatial coverage than RT (RT has a pattern shift; this is also shown in Figs. 5 and 6). In Figs. 4–6, it is shown that the SLCA model results are capturing, in general, the spatial patterns of the rainfall dynamics quite well, but with relatively lower peaks compared with the radar data. This problem might be improved by adjusting the CA model structure and neighborhood schemes to reduce the diffusive effect.

The model results from SLCA using both ANN and MT are compared quantitatively with the results using RT by using the RMSE, PoD, CSI, and FAR measures (Table 1). The precipitation forecast using SLCA-MT and SLCA-ANN is able to reduce the RMSE values for a 1-h forecast. The results in Table 1 show values of approximately 25% reductions compared with RT. The rainfall detection capabilities by SLCA using both ANN and MT are better than RT based on the PoD of a threshold of 0.1 mm. For instance,

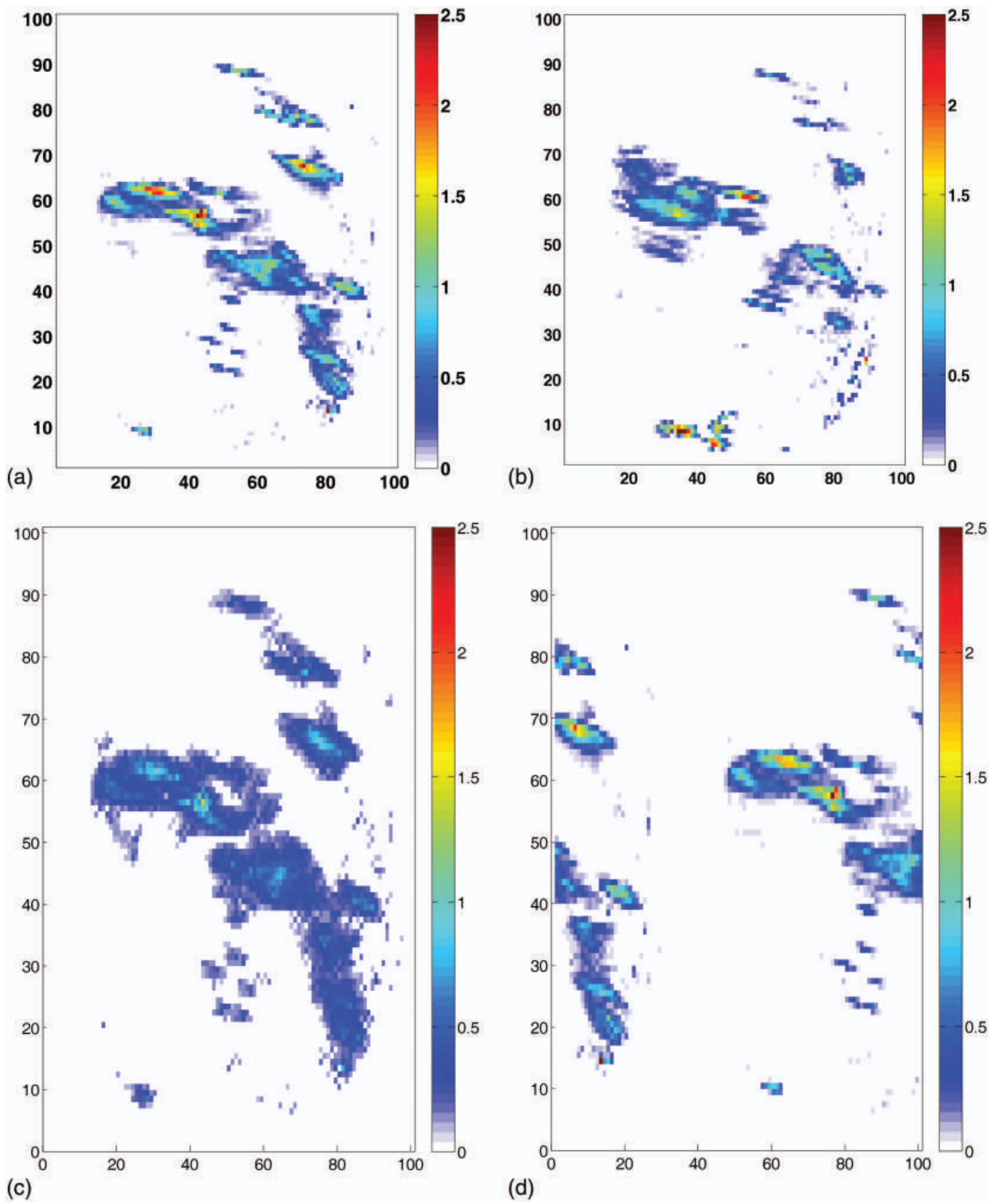

Fig. 4. (Color) Examples of precipitation maps from radar: (a) at 15:00; (b) at 16:00; (c) forecasted by SLCA (MT); (d) radar tracking on 16:00, Oct. 2, 2004 (unit: cm)

J. Hydrol. Eng. 2013.18:206-211.

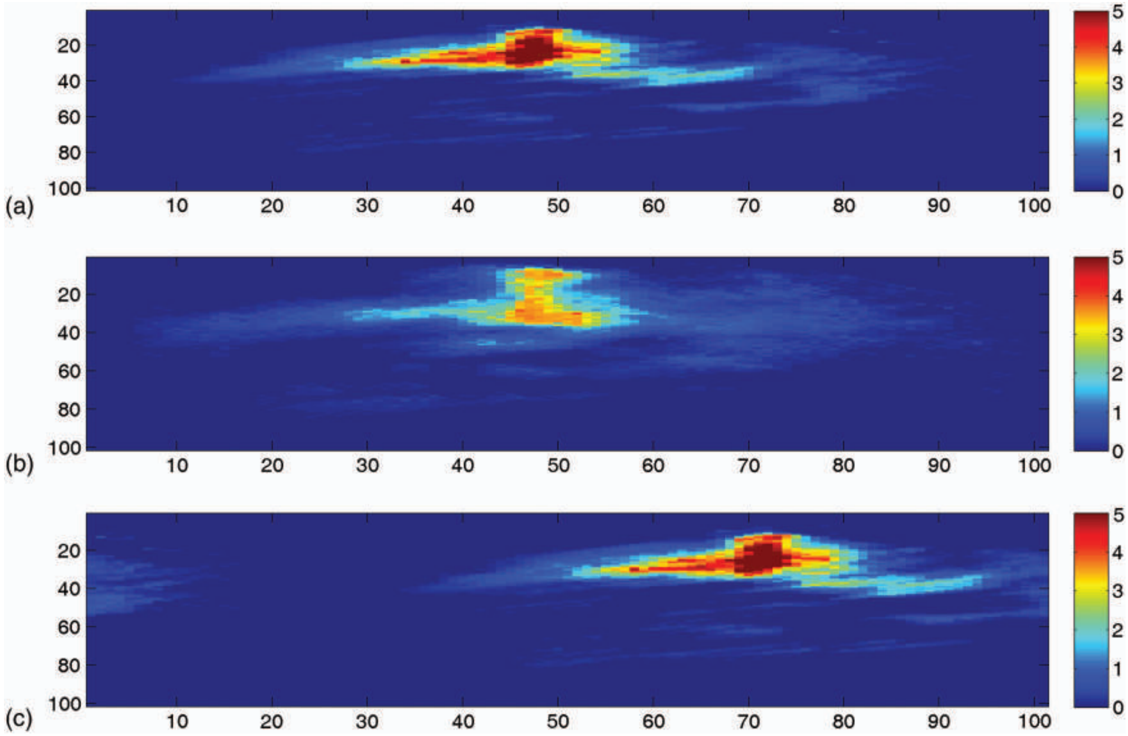

Fig. 5. (Color) (a) Radar data versus (b) SLCA model result (with ANN) versus (c) tracking model result after 46 time steps (hourly)

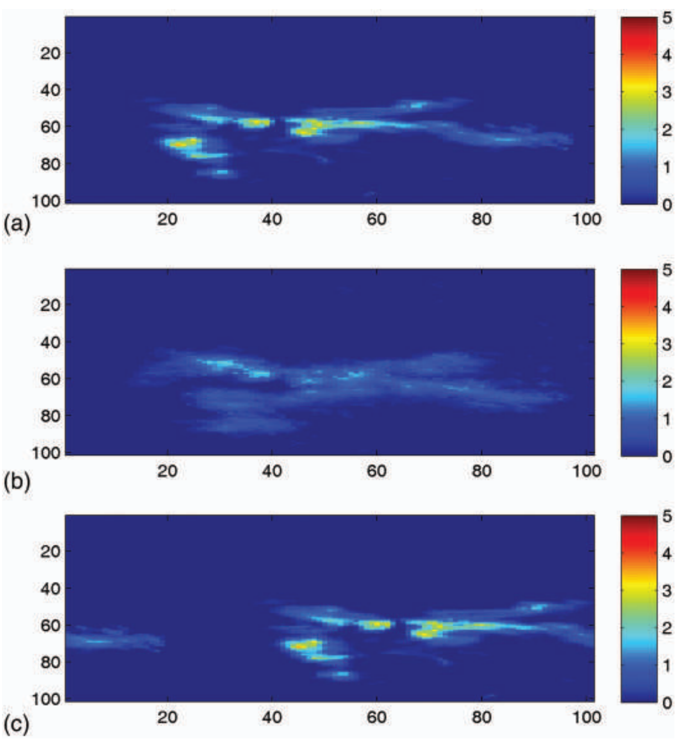

Fig. 6. (Color) (a) Radar data versus (b) SLCA model result (with ANN) versus (c) tracking model result after 195 time steps (hourly)

for a 1-h forecast, the PoD of the radar tracking is 33%; whereas, with the ANN model, the value obtained was 100%. The FAR values using SLCA are similar to those obtained from radar tracking. Nevertheless, relatively high values are observed attributable to the fact that the MLP in the normalization procedure assigned a small amount of rain value to those cells with zero rainfall value, so the probability of obtaining a false alarm increased. The values of the CSI score are better for an MT model and can achieve 13% improvement for a 1-h forecast.

Table 1. Performance Comparison between RT and SLCA Models Using MT and ANN in a 1-h Forecast

Error analysis method	SLCA-ANN	SLCA-MT	RT
RMSE (mm)	0.11	0.11	0.36
PoD	1.00	0.96	0.33
FAR	0.68	0.67	0.66
CSI	0.31	0.32	0.19

Conclusions and Discussion

The use of an SLCA has shown the capability to complement the information content in weather radar outputs for rainfall prediction. Although the cellular automata developed in this paper did not account for other variables such as wind in addition to radar precipitation data, it has shown the capability to forecast, in general, the precipitation with higher accuracy than a radar-tracking technique. To be able to include these variables, further analysis of the relationship between the spatial pattern of neighbors and the SLCA structure is needed to link precipitation dynamics with the information given by the wind velocity vectors.

In this research, the use of data sampling was an important step to build the data-driven models (ANN, MT). After some trial and error, it was possible to recognize that the sampling of data in the data-driven models was vital for the SLCA to capture the precipitation dynamics. However, the sampling might be vulnerable to errors because of the random selection, which may result in overrepresentation or underrepresentation of low or high rainfall values. This selection can be improved through optimization methods.

J. Hydrol. Eng. 2013.18:206-211.

The SLCA in this study was applied to one whole radar image, and therefore the rules identified were applied to the whole region. This can be improved by the use of separate SLCAs that can be applied in regions with relatively higher elevation changes. Future investigation of the use of SLCA can be combined with modular models when needed.

Acknowledgments

The authors greatly acknowledge the data supplier: the British Atmospheric Data Centre.

References

Almeida, C. M., Gleriani, J. M., Castejon, E. F., and Soares, B. S. (2008). "Using neural networks and cellular automata for modelling intra-urban land-use dynamics." *Int. J. Geogr. Inf. Sci.*, 22(9), 943–963.

Balzter, H., Braun, P. W., and Kohler, W. (1998). "Cellular automata models for vegetation dynamics." *Ecol. Model.*, 107(2–3), 113–125.

Cluckie, I. D., Rico-Ramirez, M. A., and Xuan, Y. (2006). "An experiment of rainfall prediction over the Odra catchment by combining weather radar and numerical weather model." *Proc., 7th Int. Conf. on Hydroinformatics HIC 2006*, Nice, France.

Corzo, G., Xuan, Y., Martinez, C. A., and Solomatine, D. (2009). "Ensemble of radar and MM5 precipitation forecast models with M5 model trees." *IAHS Redbook of Proc. of Symp. JS.4 at the Joint Convention of the Int. Association of Hydrological Sciences (IAHS) and Int. Association of Hydrogeologists (IAH)*, IAHS Publication 331, IAHS, Wallingford, UK.

Elshorbagy, A., Corzo, G., Srinivasulu, S., and Solomatine, D. P. (2010). "Experimental investigation of the predictive capabilities of data driven modeling techniques in hydrology—Part 1: Concepts and methodology." *Hydrol. Earth Syst. Sci.*, 14(10), 1931–1941.

Golding, B. (1998). "Nimrod: A system for generating automated very short range forecast." *Meteorol. Appl.*, 5(1), 1–16.

Levenberg, K. (1944). "A method for the solution of certain problems in least squares." *Q. Appl. Math.*, 2, 164–168.

Li, H. (2009). "Spatial pattern dynamics in aquatic ecosystem modelling." Ph.D. thesis, CRC Press/Balkema, Leiden, the Netherlands.

Li, H., Arias, M., Blauw, A., Los, H., Mynett, A. E., and Peters, S. (2010). "Enhancing generic ecological model for short-term prediction of southern North Sea algal dynamics with remote sensing images." *J. Ecol. Model.*, 221(20), 2435–2446.

Li, H., Mynett, A. E., and Corzo, G. (2007). "Model-based training of artificial neural networks and cellular automata for rapid prediction of potential algae blooms." *Proc., 6th Int. Symp. on Ecohydraulics*, Christchurch, New Zealand.

Li, Z., Ge, W., Liu, J., and Zhao, K. (2004). "Coupling between weather radar rainfall data and a distributed hydrological model for real-time flood forecasting." *Hydrol. Sci. J.*, 49(6), 945–958.

Marquardt, D. (1963). "An algorithm for least-squares estimation of nonlinear parameters." *SIAM J. Appl. Math.*, 11(2), 431–441.

MATLAB version 7.11.0584 [Computer software]. (2010). MathWorks, Natick, MA.

Packard, N. H., and Wolfram, S. (1985). "Two-dimensional cellular automata." *J. Stat. Phys.*, 38, 901–946.

Pessoa, M., Bras, R., and Williams, E. (1993). "Use of weather radar for flood forecasting in the Sieve river basin: A sensitivity analysis." *J. Appl. Meteorol.*, 32(3), 462–475.

Quinlan, J. R. (1986). "Induction of decision trees." *Mach. Learn.*, 1(1), 181–186.

Quinlan, J. R. (1992). "Learning with continuous classes." *Proc. of Australian Joint Conf. on AI*, World Scientific, Singapore, 343–348.

Schaefer, J. T. (1990). "The critical success index as an indicator of warning skill." *Weather Forecasting*, 5(4), 570–575.

Stanski, H. R., Wilson, L. J., and Burrows, W. R. (1989). "Survey of common verification methods in meteorology." *WMO World Weather Watch Technical Rep. No. 8, WMO/TD No. 358*, Atmospheric Environment Service, Forecast Research Division, Geneva.

Sun, X., Mein, R. G., Keenan, T. D., and Elliot, J. F. (2000). "Flood estimation using radar and rain gauge data." *J. Hydrol. (Amsterdam)*, 239(1–4), 4–18.

Werner, M., Blazkova, S., and Petr, J. (2005). "Spatially distributed observations in constraining inundation modelling uncertainties." *Hydrol. Processes*, 19(16), 3081–3096.

Witten, I., and Frank, E. (2005). *Data mining: Practical machine learning tools and techniques*, 2nd Ed., Morgan Kaufmann, San Francisco.

Wolfram, S. (1986). *Theory and applications of cellular automata: Including selected papers 1983–1986*, World Scientific, Singapore.

J. Hydrol. Eng. 2013.18:206-211.