Modeling water quality to determine a safe distance between cities: a case study in China
Ning Jia, Xin Dong and Yi Liu

ABSTRACT
Both the distance between the upstream and downstream cities and the capacity of urban water infrastructure would affect the water safety of the cities in the catchment. In this work, the concept of safe distance for urban growth was proposed. If the water quality between the upstream and the downstream cities can meet the functional requirements of the water environment, then the distance between the upstream and the downstream cities is safe. Taking two neighboring cities in the Yangtze River catchment as a case study, a distributed Cellular Automata (CA) model and a backpropagation (BP) neural network water quality model were used to discuss the safety distance between the two cities. The results provided some decision-making on urban sprawl control and rational urban development.

Key words | BP neural network, CA model, safe distance between cities, urban growth boundary, water quality model

INTRODUCTION
As China experiences rapid urbanization (Gregory 2012), the scale of construction land is gradually expanding, and its cities are growing more and more contiguous. Along with this comes the increasingly prominent pressure on natural resources and the environment (Tang et al. 2005; Duh et al. 2008). Studies show that the transformation of non-construction land into construction land will increase the demand of water resources (Liu et al. 2014), storm runoff (Fletcher & Burns 2012), and sewage emission (Dong et al. 2014), and the higher proportion of built-up land has been proven to be a negative impact on water quality (Cao & Huang 2011). At the same time, the expansion of the cities shrinks the space between the built-up areas of adjacent cities, while the impact of pollution in the upstream cities on the drinking water of the downstream cities cannot be minimized or eliminated by the self-purification ability of rivers alone. In this context, water security in the upstream cities exercises a major influence on the downstream cities. The time has now come to evaluate the effects of the expansion of two or more cities on the water environment.

In recent research, setting up the Urban Growth Boundary (UGB) has usually been considered an efficient way to control urban sprawl and achieve rational urban development (Benfield et al. 2005). However, there still remains gaps in the study of interactions between cities. Existing studies about UGB mostly focus on the quantization of a single city’s boundary (Bhatta 2009; Tayyebi et al. 2011; Xu et al. 2013), where the local ecological security pattern (Li 2011; Zhou et al. 2014; Zhang et al. 2015) and the carrying capacity of resources (Wang et al. 2015) and the environment are regarded as constraints, while the interactions between cities are rarely involved. As for the technologies on simulating urban expansion, the Cellular Automata (CA) is widely used (Liu et al. 2012). In the current CA models, the concentrated parameter model is adopted, which means that the parameters are homogeneous on the spatial scale. However, the regional characteristics and the differences in the

land changes on the spatial scale are ignored in these models (Li & Ye 2001), which leaves space for making improvements.

This study introduces the concept of safety distance into UGB research. Considering both the effect of city expansion on its own emissions and the water quality response between upstream and downstream cities, and taking the water environmental function of the downstream cities as the constraint, we integrated the CA and backpropagation (BP) neural network to develop a comprehensive model (Zhao et al. 2007; Song 2009; Wang et al. 2012), which is used to determine the reasonable distance between the upstream and downstream cities and then the reasonable boundary of urban spatial growth. The urban expansion part that we use in this research is the distributed CA model, which is originally improved from the traditional concentrated CA model (Liu et al. 2012).

METHODS

An Integrated Water Quality Simulation of Land Use Change Model is developed to quantitatively assess the environmental impacts of urbanization via a comprehensive, analytical process that includes forecasting urban land-use changes and simulating water quality response. The model consists of two subsystems: an urban expansion module based on a Distributed Cellular Automata and a water quality simulation module based on a BP neural network. These subsystems are discussed in the following sections.

An urban expansion module based on the distributed Cellular Automata

According to the transition rule, which is crucial for a CA model, all the factors that influence land use change can be divided into two categories: driving factors and restricting factors. The driving factors mainly include the natural growth effect and the agglomeration effect. The former includes the genetic effect and the neighborhood effect, denoting the transforming trend of the land cell itself and the influence of the surrounding cells, respectively. The agglomeration effect denotes the promoting force of development centers that transform a certain cell to the construction land. This study considers rail/roadways, rivers, and city centers as development centers, where a distance decay function in the double logarithmic mode (Eldridge & Jones 1991) is used to quantify the process of pollutant degradation.

The restricting factors include construction suitability, land use change cycle, and policy intervention. The effects of these factors on land use change are global, so the discrete valuation is always adopted, and it quantifies the constraints at the global level.

The urban land use change is affected by all the factors above. The conversion rule (the probability function P) of a construction cell is

\[
P = C_{\text{Landsuit}} \times C_{\text{Time}} \times C_{\text{Plan}} \\
\times (w_1 W_{\text{Inherit}} + w_2 W_{\text{Neighbor}} + w_3 W_{\text{Traffic}} + w_4 W_{\text{River}} \\
+ w_5 W_{\text{Urban}}) \times RI
\]

\[
\sum_{i=1}^{5} w_i = 1, \quad w_i \geq 0
\]  

where \(C_{\text{Landsuit}}, C_{\text{Time}}, \) and \(C_{\text{Plan}}\) denote the impact of construction suitability, current land-use duration restriction, and planning intervention, respectively; \(W_{\text{Inherit}}\) and \(W_{\text{Neighbor}}\) are the genetic effect and neighborhood effect, respectively; \(W_{\text{Traffic}}, W_{\text{River}},\) and \(W_{\text{Urban}}\) are the roadway and railway transportation effect, river effect, and the city center effect, respectively; \(w_i (i = 1, 2, ..., 5)\) is the weight coefficient, the sum of which must equal 1 for each kind of construction land; and RI, within \([0.8, 1.2]\), is the random perturbation.

In particular, if the current type of the target cell is non-construction land, then only the genetic effect and the neighborhood effect should be considered. The function of the conversion rule is

\[
P = w_1 W_{\text{Inherit}} + (1 - w_1) W_{\text{Neighbor}}
\]

This study classifies land use into eight types, three of which are construction lands and five are non-construction lands. After CA calculation, a cell will transform to the type with the maximum probability in the next period.

Current CA models possess homogeneous parameters on the spatial scale, so they can be called ‘concentrated parameter model’. However, in these models, the
characteristics and differences on the spatial scale during land use change process are always ignored (Li & Ye 2001). To address this limitation, this study proposes a distributed CA model. The core of the distributed CA model is: the study area is divided into several sub-regions, and the parameters of each cell within one sub-region are the same, and the parameters in different sub-regions are independent of each other. The parameter identification process is also independent among different sub-regions. Study area division is based on administrative division and catchment division according to the Digital Elevation Model, with the assistance of artificial modification. Different ways of division can be set and calculated for comparison.

Since the distributed model multiplies the number of parameters, the parameter identification methods widely used, like the Monte Carlo method (García et al. 2011), the Latin hypercube method (Xu et al. 2005), and the HSY algorithm (Xu et al. 2012), are not applicable. So, the Genetic Algorithm is chosen for this process, which has the parallel computing power and the gradual optimizing process, and these can not only solve the problem of multi-parameters but also avoid generous invalid operations in the region far away from the optimal value (Aytug et al. 2003). In this study, the cross probability used in the genetic algorithm is set to 0.8, and the variation probability is set to 0.005, 0.01, 0.05, 0.1, 0.2, and 0.3, respectively, in each round of operation for trials.

As for the simulation accuracy of a CA model, the Matching Rate (MR) and the Kappa Coefficient are often used to measure the matching degree of the simulated results to the actual distribution of land use types (Duan et al. 2012). MR is defined as the proportion of the cells in which the simulated result equals the actual land use type. The Kappa coefficient was developed based on a confusion matrix to assess the accuracy level of land-use classification. In this study, not only are the two indexes of the whole study area used, but also those of only the construction lands (MR_Cst and Kappa_Cst) are adopted.

A water quality module based on the BP neural network

Artificial Neural Network (ANN) is a type of information technology that imitates biological neural networks, which calculates via constantly adjusting the responses of a series of interconnected neurons (Duan et al. 2012). Among different types of ANNs, the BP neural network is the one that is commonly used, which possesses the core of the error back-propagation algorithm and the capacity of nonlinear reflecting (Duan et al. 2012). This study uses this network to simulate water quality. First, it is assumed that there exists a certain functional relationship between the pollution concentration $c$ of the river cross section and the position of the cross section $y$, and the upstream pollutant discharge position $x$ and emissions $P_x$, shown in formula (3). The aim of the BP neural network is to look for the function $f$ through the training of monitoring data samples and then simulate the pollutant concentrations of the cross section. Based on the land use prediction results from the distributed CA module above, future water quality can be predicted and analyzed using the well-trained BP model.

$$c = f(x, P_x, y) \quad (3)$$

A BP neural network consists of three parts: an input layer, an output layer, and usually one hidden layer between them. In this study, the BP neural network water quality model contains five kinds of variables as inputs: pollutant emissions from wastewater treatment plants (WWTPs) ($P$), non-point source emission intensities of built-up areas or non-built-up areas ($q$), the geographic coordinates of each monitoring section ($x$), the upstream flow ($Q_0$) and pollutant concentration ($C_0$), and the position of one cross section needed ($\chi' \text{ or } y$). When training and validating the model, the cross section needed is one of the other monitoring cross sections ($y$). When using this model to make predictions, the cross section needed is the control cross section ($y'$). The output variable of this model is the pollutant concentration of the selected cross section above. When training and validating the model, the concentration is of one of the monitoring cross sections ($C$); when making predictions, it is of the controlling cross section ($C'$). In this study, this BP model simulates two water pollutants, COD and NH$_3$-N. The number of neurons in the hidden layer is usually gained through trials with the reference of the neuron numbers of the other two layers. Figure 1 shows the structure of this BP neural network with three nodes in the hidden layer.
This study adopts a serial training method to train all the training samples of the BP model at the same time. If the overall error value meets the convergence requirements, the training is terminated; if not, the weight correction process is reversed. This process is shown in Figure 2.

In each round of training, the division of training and validating samples is randomly produced, with 80% samples for training and 20% for validating. After trials, this study sets two termination conditions for ending the training process: the Mean Squared Error (MSE) for the training samples is less than 0.015 (for COD) or 0.001 (for NH₃-N). What is more, the BP network itself has a termination condition: the maximum constant rising numbers of the epochs of the MSE of the validating samples. Rather than looking for the ‘best’ network, we regard the networks that meet these two termination conditions as acceptable networks: the MSE of the training samples reaches its goal, and the relative error between the simulated output and the actual output is less than ±20% for all the corresponding validation samples.

Use the BP neural network water quality model to predict the pollutant concentrations of the control cross section of the downstream city. If the concentration of the control cross section meets a certain water quality standard of surface water, then the distance between the upstream and the downstream cities is safe, which means that the river’s carrying capacity and its self-cleaning capacity can release the pressure caused by the urban space expansion of the upstream and downstream cities and the constant shrinking of the distance between these cities. If the concentration of the control cross section cannot meet the certain water quality standard, the urban space expansion forms of the upstream and downstream cities need to be reselected and the UGB needs to be redesigned until the standard is met. Then, it gets the safe distance.

Before running the BP neural network, the study area and the emissions from all sources need generalization. Based on the land use change simulation result from the CA module, the water pollution emissions from point and non-point sources both in the built-up and in the non-built-up areas can be calculated. The point sources include residential and industrial sources, whose emissions are quantified through the water use and discharge characteristics of residential and industrial sectors. The non-point sources mainly include the pollutants from surface runoff into water bodies caused by rainfall in the built-up and non-built-up areas. This study assumes that the wastewater from point sources is collected and treated by the sewer system of the built-up areas and then discharged into the natural water body, while the non-point sources discharge into the water body along the way in the form of a line source.

The calculation of the residential emissions is based on the population and that of the industrial emissions is based on the area of industrial land. The non-point source emissions in the built-up areas are quantified at the administrative unit, while the non-point emissions in the
non-built-up areas (mainly farmlands) are quantified according to the catchments divided via hydrological analysis on ArcGIS, and both of them are also based on the emission intensity per unit of area. The emission quantities divided by the distance along the river of one region is regarded as the linear emission intensity of the non-point sources. The distance along the river is calculated based on the geographic coordinate along the river of the region.

**CASE AREA**

Wuhu and Ma’anshan, two adjacent cities along the Yangtze River, serve as important economic centers of Anhui Province, China. Since the year 2000, the construction lands of the two cities have expanded dramatically, with 68–135 km² in 2010 of Wuhu and 36–135 km² of Ma’anshan. Rapid urbanization made the two built-up areas expand oppositely along the Yangtze River, and the distance between them is shrinking. In 2010, the National Development and Reform Commission approved a planning (National Development and Reform Commission of the People’s Republic of China 2010) proposal ‘to accelerate the urban integration of Wuhu and Ma’anshan’ and ‘build up modern group-mode large cities’. It can be foreseen that these two built-up areas will further expand oppositely in the future, and the distance will further reduce. Since the main drinking water sources and wastewater whereabouts of the two cities are the main streams of the Yangtze River, the constant shrinking of the distance between the two cities will lead to a potential water security threat to Ma’anshan, the downstream city caused by the emissions of upstream Wuhu.

This study is mainly concerned with the change of urban construction lands. Considering the distribution of the two municipal districts and the 15 km buffer along the east coast of the Yangtze River, we extract about 1,063 km² of the study area, as shown in Figure 3. The vector data of land use and transportation network in 2000, 2005, and 2010 are collected, and the first two are for parameter calibration and the last two are for validation. The data that the BP network needed, including the water quality, rainfall, facts about WWTPs, agricultural pollution emission data, etc., are collected from the local environmental department and related yearbooks.

For the first module, all the influencing factors mentioned under section ‘An urban expansion module based on the distributed Cellular Automata’ are used in this research. In particular, for the development centers of the agglomeration effect, the gravity center of the built-up area of Wuhu and Ma’anshan, the mainstream of the Yangtze River, and the railway mainlines are selected. For the restricting factors, we select the slope gradient as the criterion of the construction suitability and the planning and policy about the construction land as the intervention information; the land use change cycle is limited to 5 years. According to this, the parameters that the model needs to identify for each sub-region consist of 12 weight parameters, 10 controlling parameters, 3 intervention parameters, and 1 random perturbation. See Supplementary material, Table S1 for details on all parameters.

The study area covers six water quality monitoring cross sections and five WWTPs. According to the catchment area division prepared earlier and the distribution of the construction lands of the two cities, the study area can be abstracted into three adjacent rectangles along the river, and from south to north, there are the built-up area of Wuhu (W for short), the non-built-up area (N for short), and the built-up area of Ma’anshan (M for short). It is assumed that within each rectangle, the non-point sources are discharged into the Yangtze River along the bank uniformly. Figure 4 shows the generalization of the spatial characteristics of the regional effluent emissions. The upstream point of W along the river is the start point. The upstream boundary of M is selected as the control cross section, the water quality of which reflects the influence on the water environmental quality from the distance change between the two cities. For any point along the riverbank, the distance along the river from the start point to the bank is regarded as its coordinate. This study assumes that the water quality near the bank is affected only by the pollution from the same side of the riverbank and the pollution from the other side is ignored because the Yangtze River is relatively wide.

It is assumed that the residential and industrial wastewater is treated up to the standard of Level 1B (GB 18918-2002) and then discharged. Since 2006, the original positions of the five WWTPs have been extended, and their positions have not changed. It is assumed that this exercise will...
continue until 2020 so as to meet the demand of urban expansion; therefore, the emissions of point sources will change, but their positions will not. Based on this, in the following BP neural network model, only the five emissions other than the positions of point sources are regarded as input variables (see Supplementary material, Table S2).

RESULTS AND DISCUSSION

Urban land expansion

The study area is divided into 2, 4, 8, and 16 sub-regions and one remains undivided for comparison, as shown in
Supplementary material, Figure S1. The parameter identification results through the genetic algorithm show that when the study area is divided into 8 sub-regions and the variation probability is 0.1, the distributed CA model obtains the maximum fitness value of 0.553, which is approximately 10% higher than that of the concentrated model. In other words, the distributed parameter CA model has better simulation capability and potential than the traditional concentrated model in this case.

Fit the optimal parameters identified above into the CA CoDel. The simulation results suggest moderately good approximations to reality for the period 2005–2010: MR is 0.879, Kappa is 0.801, MR_Cst is 0.664, and Kappa_Cst is 0.537. This means that the distributed parameter CA model that we build is reasonable and effective.

Use these parameters to predict the land use of the region in 2020. In 2020, the area of construction lands is expected to reach 490 km², increasing by 45% compared with 2010. But the area of farmlands is expected to reduce to 151 km², with a decrease of 27%. These results confirm a continuous distance decrease between the two built-up areas in the future.

Water quality simulation

Collecting the annual COD and NH₃-N concentration data of the seven monitoring cross sections on the mainstream of the Yangtze River Wuhu-Ma’anshan section from 2007 to 2012, excluding the part that was undetected or outlier, we obtained 30 groups of ‘year-position of cross section-concentration’ data of COD and 29 groups of NH₃-N.

We repeated 10,000 training rounds of the BP model and recorded the final MSEs of the training samples. For COD, 7,012 of 10,000 rounds reached its MSE goal, while the remaining rounds were terminated because of the validation check. For NH₃-N, all of the 10,000 rounds
reached its MSE goal. According to the acceptable network criteria, after 10,000 rounds of training, we got 2,439 acceptable networks for COD and 2,749 for NH$_3$-N.

The distance between the end point of the built-up area of Wuhu and the start point of the built-up area of Ma’an-shan is regarded as the distance between the two cities, which was 10.7 km in 2010 and is predicted to be 4.6 km in 2020. Use the coordinate of the control cross section and other 14 estimated variables of 2020 in the acceptable networks above and then get the predicted pollutant concentrations of the control cross section for 2020. Excluding the results that are below 0 for COD (then, 2,194 results are accepted), the frequency histograms of the predicted values are shown in Figure 5.

According to the Environmental Quality Standards for Surface Water (GB 3838-2002), for Class II waters that the control cross section adapts to, the COD standard is 15 mg/L and NH$_3$-N 0.5 mg/L. As a result, there is 98.8% of the probability for the control cross section to meet the COD standard and 99.8% for NH$_3$-N. It could be deduced that if these two cities expand according to the existing speed and mode, the water quality of the control section is likely to meet the local water function requirement in 2020.

For the same control cross section, we used the acceptable neural networks above to simulate COD and NH$_3$-N concentrations in 2010, respectively. Both the results of 2010 and 2020 are shown in Figure 6, and the results of the paired-samples $T$ test between 2010 and 2020 are
shown in Table 1. The concentrations of the same control cross section between 2010 and 2020 are significantly different at the level of 5%, for both of these pollutants. Compared with the year 2010, the mean concentration of COD in 2020 improves significantly with a percentage of 32.2, while NH$_3$-N mean deteriorates significantly with a percentage of 45.3. It could be deduced that NH$_3$-N is more possible to face a risk of exceeding the standard in the future.

These results reveal that the distance between the two built-up areas in 2020 according to the current developing trend is safe enough to assure that both these pollutants meet their standards. But in the future, it may not be safe anymore for the deterioration trend of NH$_3$-N.

Mitigation strategies analysis

Since NH$_3$-N may deteriorate in the future, mitigation strategies need to be considered. Here, we set one control group and two potential mitigation strategies to reduce NH$_3$-N concentration.

- **Control Group:** Reduce the inflow concentration. In the current situation, the inflow NH$_3$-N concentration of 2020 is estimated as the mean values from 2006 to 2012, which is 0.256 mg/L. Reduce it by 15%, 30%, and to the surface water quality standard Class I 0.15 mg/L.

- **S1:** Upgrade the emission standard of the WWTPs of the upstream city. In the current situation, the wastewater from both Wuhu and Ma’anshan is discharged according to Level 1B standard (8 mg/L) (GB 18918-2002). If the WWTPs in upstream Wuhu are required to lift their discharge standards to Level 1A (5 mg/L), then the two-point source emissions P1 and P2 could be reduced to 37.5%.

- **S2:** Reduce the length along the river of Wuhu built-up area in 2020 to that in 2010, which equals the distance between the two built-up areas that increases by 134% compared with that in 2020, while the newly appeared construction lands expand vertical to the river.

The BP water quality model runs for the three scenarios above, and here are the results and analysis.

Figure 7 shows that as the inflow concentration reduces gradually, the NH$_3$-N concentration decreases significantly (at 5% level). It means that merely reducing the inflow concentration has a significant effect on downstream NH$_3$-N reduction, which reveals a linear correlation with $R^2$ very close to 1 (see Figure 7). The downstream reduction could be offset partially by emissions from the upstream city, resulting in smaller reducing percentages of NH$_3$-N on the control cross section. It could be deduced in such a way that when the inflow water contains no NH$_3$-N, the NH$_3$-N concentration of the control section would be 0.2015 mg/L after the emissions from upstream Wuhu.

NH$_3$-N concentrations under different mitigation strategies and the corresponding inflow reduction rates are shown in Figure 8. The paired-samples T test results show that the NH$_3$-N concentration between S1 and Baseline is not significantly different (at 5% level), though the average concentrations are slightly below those of the baseline. It means that merely reducing the discharge of the WWTPs

---

**Table 1** Results of the paired-samples T test of the simulated concentrations between 2010 and 2020

<table>
<thead>
<tr>
<th>Pollutants</th>
<th>Mean concentration (mg/L)</th>
<th>Paired differences (year 2020-2010)</th>
<th>Significance (two-tailed)</th>
</tr>
</thead>
<tbody>
<tr>
<td>COD</td>
<td>8.6</td>
<td>5.8</td>
<td>-2.8</td>
</tr>
<tr>
<td>NH$_3$-N</td>
<td>0.19</td>
<td>0.27</td>
<td>0.08</td>
</tr>
</tbody>
</table>

---

Downloaded from http://waponline.com/aqua/article-pdf/69/8/833/624062/jws0890833.pdf by guest
of the upstream city makes little contribution to \( \text{NH}_3\text{-N} \) reduction in the adjacent downstream cross section in this case.

The \( T \) test shows that reducing the length of the Wuhu built-up area and thus expanding the length of the non-built-up area is statistically significant (at 5% level) to bring about a decrease in the \( \text{NH}_3\text{-N} \) concentration in the control section. The average concentration is 4.4% lower than that of the baseline, and it is equivalent to the scenario when the inflow concentration reduces by 17%.

If both point sources are reduced and the length of the built-up area is restricted in Wuhu, the average concentration in that control section will be 0.249 mg/L, 8.29% lower than that of the baseline. This strategy combination has the same effect as that of the 32% reduction of the inflow concentration.

Traditional strategies that merely control the point sources have little to do with this case. On the contrary, it will take huge economic costs to upgrade and reconstruct the WWTPs, which makes the use of this strategy less advisable. However, restricting the length of the built-up areas alone proves to have a statistically significant effect on downstream \( \text{NH}_3\text{-N} \) reduction, which reveals that the distances within and between the built-up areas are negligible factors affecting the downstream water quality in this case. Since it is hard to decrease the inflow concentrations and barely effective to control point sources, the restriction of urban expansion along rivers seems to be a more feasible strategy to improve the downstream water quality. Smart urban planning needs to be made before urban expansion.

It is true that there are many uncertain factors that affect the safety distance of cities, such as the decay function of pollutant degradation and the hydrological conditions of the catchment. These are not covered in this article and can be discussed in future work.

**CONCLUSION**

In this paper, we construct a distributed parameter CA model for land use change simulation. The case results show that the distributed model has a better simulating performance than the traditional concentrated model. When the study area is divided into 8 sub-regions and the variation probability is 0.1, the maximum fitness value is obtained and it is 10% higher than that of the concentrated model. Besides, the application of the genetic algorithm is satisfactory for the identification of numerous parameters.

This study uses a combination of the BP neural network and the CA model to simulate the water quality between two adjacent cities. According to the current speed and mode for urban expansion in the case area, it is likely for the control section on Ma’anshan in 2020 to meet the surface water quality standard Class II. Compared with 2010, the COD concentration will drop significantly, but a marked increase will occur in \( \text{NH}_3\text{-N} \), which may make \( \text{NH}_3\text{-N} \) out of standard in the future.

This study explores the safe distance between the two adjacent cities based on water quality. According to the current trend of urban expansion, the distance between the two built-up areas will be safe because both pollutants will meet the quality standard. But in order to mitigate \( \text{NH}_3\text{-N} \)'s deterioration in the future, Restricting the length of the built-up area is advisable because it is more effective than reducing point source emissions in this case. A longer distance between the two cities will bring a great deal of benefit to the downstream city’s water quality.

**REFERENCES**


Song, G. 2009 Application Research of Artificial Neural Network Method on Water Quality Simulation and Assessment. Chongqing University, Chongqing (in Chinese).


First received 21 June 2020; accepted in revised form 2 December 2020. Available online 7 December 2020