# Analyzing and simulating spatial patterns of crop yield in Guizhou Province based on artificial neural networks

Boyi Liang[a,b], Hongyan Liu[a,*], Tim Quine[b], Xiaoqiu Chen[a], Paul D. Hallett[c], Elizabeth L. Cressey[b], Xinrong Zhu[a], Jing Cao[a], Shunhua Yang[c,f ,g], Lu Wu[a], Iain Hartley[b]

[a] College of Urban and Environmental Sciences, Peking University, Beijing 100871, China

[b] Department of Geography, University of Exeter, Amory Building, Rennes Drive, Exeter, EX 4 4RJ, UK

[c] Institute of Biological and Environmental Sciences, University of Aberdeen, Aberdeen, AB24 3UU, UK

[d] State Key Laboratory of Soil and Sustainable Agriculture, Institute of Soil Science, Chinese Academy of Sciences, Nanjing 210008, China

[e] University of the Chinese Academy of Sciences, Beijing 100049, China

*Corresponding author

## Abstract

The area of karst terrain in China covers 3.63×10$^6$ km$^2$, with more than 40% in the southwestern region over the Guizhou Plateau. Karst is comprised of exposed carbonate bedrock over approximately 1.30×10$^6$ km$^2$ of this area which suffers from soil degradation and poor crop yield. This paper aims at gaining a better understanding of the environmental controls on crop yield in order to enable more sustainable use of natural resources for food production and development. More precisely, four kinds of artificial neural network were used to analyze and simulate the spatial patterns of crop yield for 7 crop species grown in Guizhou Province, exploring the relationships with meteorological, soil, irrigation and fertilization factors. The results of spatial classification showed that most regions of high-level crop yield per area and total crop yield are located in the central-north area of Guizhou. Moreover, the three artificial neural networks used to simulate the spatial patterns of crop yield all demonstrated a good correlation coefficient between simulated and true yield. However, the Back Propagation network had the best performance based on both accuracy and runtime. Among the 13 influencing factors investigated: temperature (16.4%), radiation (15.3%), soil moisture (13.5%), fertilization of N (13.5%) and P (12.4%) had the largest contribution to crop yield spatial distribution. These results suggest that neural networks have potential application in identifying environmental controls on crop yield and in modelling spatial patterns of crop yield, which could enable local stakeholders to realize sustainable development and crop production goals.

Keywords: Karst; critical zone; crop yield; artificial neural network; crop model; Guizhou

## 1. Introduction

Karst landscape covers vast areas of the globe, including over 30% of China. They are characterized by exposed carbonate rocks that weather rapidly and are highly susceptible to environmental change and natural erosion. In China the karst landscape in the southwest region has experienced rapid and intensive alterations to land use and associated ecosystem degradation over the last 50 years (Moore et al., 2017; Chen et al., 2018; Li et al., 2018). The intensification of agriculture since the late 20th century has led to a rapid deterioration of the soil, reflected in reduced crop production and the rapid loss of soil (Green et al., 2019). Under the Grain for Green Program (GGP), millions of hectares of farmland have been turned into non-crop vegetation in order to combat "rocky desertification" (Cheng et al., 2015). Ensuring both ecological and food security is a top priority for all stakeholders in China.

The karst environment has unique characteristics, such as soluble rock, a calcium-rich and alkaline nature, soil scarcity, a double-layer structure, and water leakage through cavernous channels that rapidly link topsoil to groundwater. These environmental stresses impose an adverse influence on the growth of vegetations in the karst region (Yuan, 2001; Tong et al., 2017). Studying the ecosystem features of karst benefits from a comprehensive understanding of interactions among different element in critical zone (CZ) of this region. CZ observatories (CZOs) have thus been established in China's karst region to gain a holistic understanding of soil formation from bedrock, water transport to the groundwater below and beyond, and the interactions with vegetation (Anderson et al., 2008; Lin et al., 2011; Banwart et al., 2012; Grant and Dietrich, 2017). This research can not only promote the knowledge of CZ processes, but also provides fundamental information that could be applied in practice to help local people by applying, adapting and developing decision support tools (DSTs) and help to guide practices such as crop production (Davis et al., 2005; Banwart et al., 2013; Menon et al., 2014; Rose et al., 2016).

As agriculture is one of the largest drivers of land cover change (Scholes et al., 2018), it provides the focus of numerous CZOs across the globe (Guo and Lin, 2016; Kumar et al., 2018). In karst regions, food production is critically affected by the environment and is manifested as persistently poor and declining crop yield (Wang et al., 2004; Liu, 2006; Zhang et al., 2013). For example, karst rocky desertification and serious soil erosion from poor farming practice decreases land productivity (Nguyen et al., 1996; Tan et al., 2010; Yan and Cai, 2015). Up to now, the existing models for estimation of crop yield mainly contain statistical approaches and process-based models (including large scale global gridded crop models), which are not ideally parameterized for the unique and heterogeneous properties of karst landscapes (Zhao et al., 2016; Zhao et al., 2017). Process-based models have three limitations for their use in the complex karst landscapes. Firstly, the assumptions of relevant processes for crop growth varies greatly among different models, leading to different parameterization (Rötter et al., 2011). Secondly, the primary focus of most process-based models is on the aboveground crop biomass, whereas the belowground processes and soil parameters also play an important role in influencing or even controlling crop growth (Folberth et al., 2016), especially in karst systems where soil depth is often a limiting factor (Zhang et al., 2020). Lastly, the impact of climate change on the environmental factors affecting crop growth need to be included in the existing process-based models (Rosenzweig et al., 2014). Statistical approaches also have limitations in karst systems, with direct relationships between crop yield and meteorological data (or other environmental factors) underpinning predictions (Reynolds et al., 2000; Van Wart et al., 2013; Wu et al., 2015), however, traditional models cannot tackle groups of different factors and crop parameters with non-linear relationships (Prasad et al., 2006; Kogan et al., 2018). In addition, the complexity and heterogeneity of karst landscapes, the importance of both sub-surface and surface soil and water resources, and the prevalence of small-scale subsistence farming in Guizhou, all contribute to limiting the applicability of existing crop models. In recent years, new technology such as artificial neural networks (ANNs) have been fast developed, which may provide cost-effective and comprehensive solutions for better crop yield, environmental management and DSTs through their use of non-linear regressions and enabling interaction between different factors (Panda et al., 2010; Everingham et al., 2016; Chlingaryan et al., 2018).

In this paper, we address the current limitations of crop modelling for karst landscapes by assembling spatial data of crop yield per unit (hereinafter called crop yield or YPA) and its influencing factors into artificial neural networks, in order to analyze and simulate the spatial patterns of crop yield for 7 crop species in Guizhou Province. It is the first time that multi-factorial analysis has been undertaken to simulate and explain spatial patterns of crop yield in this environment. The approach is made possible by the application of powerful and novel machine learning technology to precisely simulate the spatial patterns of crop yield. Four kinds of artificial neural network were used to: (1) detect and classify the spatial patterns of the crop yield in Guizhou Province, and (2) to simulate the spatial patterns based on different influencing factors and evaluate the factor contribution for each. This research is valuable for further developing powerful DSTs to guide land management and farming decisions in karst regions. The approach has also potential to be expanded to the research on crops of other complex landscapes in the world.

## 2. Data and Methods

### 2.1 Study Region

Guizhou is located in southwest of China (Figure 1), with an area of 1.76×$10^5$ $km^2$, and has a population of 36 million (2018), ranking 19th of all 34 provinces, with GDP (Gross Domestic Product) for Guizhou Province in 2018 ranking 25th among Chinese provinces in 2018, according to statistical data (NSBC, 2019). Guizhou Province is located at the heart of the East Asia Karst, one of the three largest areas of almost unbroken karst in the world (Sweeting, 1993; He et al., 1997). About 73% of the total area is underlain by carbonate rocks, and karst landforms are widely distributed (Su, 2002). In terms of geomorphology, Guizhou Province contains 87% plateau-mountains, 10% hills and 3% basins (He et al., 1997).

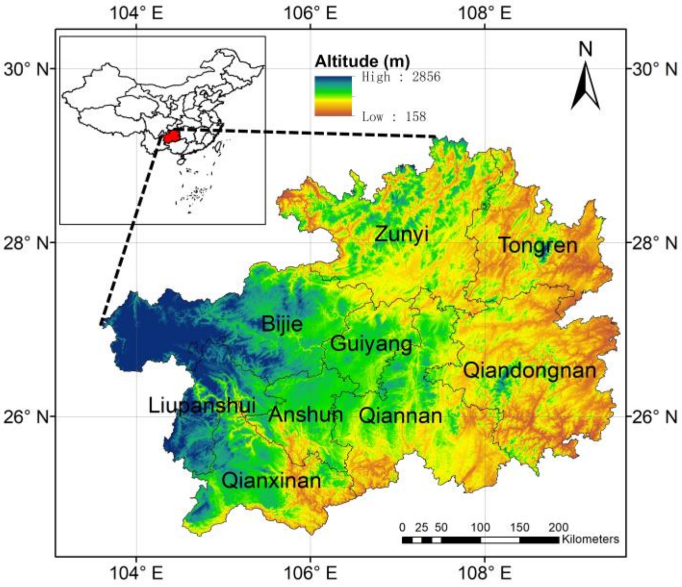

**Figure 1**. Location and administrative map and of Guizhou Province and the 9 prefectures

In the study area, the most widely grown food crops are paddy rice, maize, wheat, soybean, and potato. According to government statistics, in the past 60 years, the total crop yield of Guizhou has increased over threefold, while the crop yield per unit planting area (t/ha) is two times greater than 60 years ago, and therefore the economy has substantially grown. However, due to environmental limitations, the crop yield per unit planting area (t/ha) and income in this region is only 75.6% and 61.1% of the national average over 2005-2007, respectively (NSBC, 2019).

## 2.2 Data Resource

In this study, we selected 7 main crop species that are produced in Guizhou, including five kinds of food crop (maize, potato, rice, soybean and wheat) and two commercial crops (rapeseed and groundnut). The relative crop specific data for the 7 species were compiled from the datasets of Earthstat, which included crop yield data, total harvested area and fertilization rates, alongside irrigation data from the MIRCA2000 dataset (Table 1). We also imported nine additional crop yield influencing factors including meteorological, topographic (digital elevation model – DEM) and soil properties data (Table 1). Prior to analysis, we first unified the spatial resolution of all data resources into 5’ by aggregation and resampling (cubic method) and extracted all dataset for year of 2000.

**Table 1.** Introduction of data resources (*Crop-specific data)

| Data resource | Category | Region | Temporal Coverage | Spatial resolution |
|---|---|---|---|---|
| WFDEI | Meteorological data | Global | 1981-2014 | 0.5° |
| GMTED2010 Global Grids | DEM | Global | - | 5’ |
| HWSD | Soil property | Global | 1995 | 5’ |
| NCEP CPC | Soil moisture | Global | 1948- | 0.5° |
| MIRCA2000* | Irrigation | Global | Circa 2000 | 5’ |

|  |  |  |  |  |
|---|---|---|---|---|
| Crop area* (Earthstat dataset) | Crop area | Global | Circa 2000 | 5’ |
| Crop production* (Earthstat dataset) | Crop yield | Global | Circa 2000 | 5’ |
| Fertilization rates* (Earthstat dataset) | Fertilization | Global | Circa 2000 | 5’ |

### 2.2.1 Earthstat Datasets

EarthStat provides geographic datasets that help solve the grand challenge of feeding a growing global population while reducing agriculture’s impact on the environment. EarthStat is a collaboration between the Global Landscapes Initiative at the University of Minnesota’s Institute on the Environment and the Land Use and Global Environment lab at the University of British Columbia. The datasets contain different kinds of agricultural data including harvested area, crop yield and fertilization rates (among them we selected the value of nitrogen-N, phosphorous-P, and potassium-K). The harvest area data was achieved by combining agricultural inventory data and satellite-derived land cover data (Ramankutty et al., 2008). The Earthstat data was produced by combining national, state, and county level census statistics with a recently updated global dataset of croplands on a 5’ by 5’ latitude/longitude grid. These two kinds of data depict, circa the year 2000, the area (harvested) and yield of 175 distinct crops of the world (Monfreda et al., 2008).

### 2.2.2 Soil Property Data

The Harmonized World Soil Database (HWSD, version 1.2) is a global soil database framed within a Geographic Information System (GIS) and contains up-to-date information on world soil resources (Nachtergaele et al., 2009, 2012; Shangguan et al., 2013). It provides a raster databases, with over 15,000 different soil mapping units, which combines existing regional and national updates of soil information worldwide (Batjes and Bridges, 1994; Shi et al., 2004, 2006). In this study, we analyzed 5 soil properties (soil bulk density, soil organic carbon, pH, soil cation exchange capacity and carbonate content), which were greatly proved influential on crop growth, to investigate the relationships between soil features and the spatial distribution of crop yield (Letey, 1958).

### 2.2.3 Meteorological Data

The European Union Water and Global Change project (http://www.eu-watch.org) provides a gridded European Union Water and Global Change-Forcing-Data-ERA-Interim (WFDEI) data product (Weedon et al., 2014; Ren et al., 2018). It contains 8 meteorological variables from 1979 with a spatial resolution of 0.5°. In this study, we selected and calculated the annual average temperature and shortwave radiation (for the year of 2000) as influencing factors on crop yield for further analysis.

### 2.2.4 Soil Moisture Data

For soil moisture, we employed the product released by NOAA’s National Center for Environmental Prediction (NCEP) - Climate Prediction Center (CPC), with global spatial coverage at 0.5° resolution from 1948 to present (Ibrahim et al., 2015). The monthly dataset consists of a file containing monthly averaged soil moisture water height equivalents for the globe from 1948 onwards. Values are model-calculated and not measured directly. Soil moisture is estimated by a one-layer hydrological model (Huang et al., 1996; Van den Dool et al., 2003). We extracted the data for 2000 and calculated the annual average of soil moisture in Guizhou Province.

### 2.2.5 Irrigation Information

MIRCA2000 (monthly irrigated and rainfed crop areas around 2000) global dataset shows us the monthly irrigated and rainfed crop areas around the year 2000 that distinguishes irrigated and rainfed areas for 26 crop classes, among them 21 major crops and the crop groups of pulses, citrus crops, fodder grasses, other perennial crops, and other annual crops (Portmann et al., 2010). The dataset refers to the period 1998-2002 and has a spatial resolution of 5' by 5' (Neumann et al., 2011).

### 2.2.6 DEM (Digital Elevation Model)

The U.S. Geological Survey (USGS) and the National Geospatial-Intelligence Agency (NGA) have collaborated on the development of a notably enhanced global elevation model named the Global Multi-resolution Terrain Elevation Data 2010 (GMTED2010) that replaces GTOPO30 as the elevation dataset of choice for global and continental scale applications (Danielson and Gesch, 2011). The GMTED2010 product suite contains 7 new raster elevation products for each of various spatial resolutions and incorporates the current best available global elevation data. The new elevation products have been produced using the following aggregation methods: minimum elevation, maximum elevation, mean elevation, median elevation, standard deviation of elevation, systematic subsample, and breakline emphasis (Carabajal et al., 2011; Athmania and Achour, 2014). Slope data was aggregated the variable to different spatial grains using several aggregation approaches (including the 5' resolution we utilized) (Amatulli et al., 2018).

## 2.3 Four Kinds of Artificial Neural Network

Herein, we adopted four kinds of ANN including Self-organization Feature Map (SOFM), Back Propagation (BP), General Regression Neural Network (GRNN) and Recurrent Neural Network (RNN). Among them, SOFM was used to realize unsupervised classification of Guizhou Province into high-, medium- and low-level crop yield regions for the 7 crop species. Meanwhile, BP, GRNN and RNN were employed to simulate the spatial patterns of crop yield for the 7 species, by inputting the different influencing factors introduced above. In addition, we compared the simulation of these three networks by evaluating different indices of accuracy and runtime. Details of the networks are included in the supplementary material.

Figure 2 shows the process of simulation of crop yield in Guizhou Province. Firstly, we input the four groups of influencing factors into the three kinds of ANN (BP, GRNN and RNN). Then we randomly assigned the pixels of crop yield into a training group (75% of the total number) and validation group (25% of the total number). Secondly, we trained the networks and simulated the crop yield for the 7 species, respectively. Lastly, we compared the simulation of the three networks by evaluating the indices of accuracy and the runtime of the networks. The indices chosen to evaluate the accuracy is R (correlation coefficient between true [observed] value and forecasted [simulated] value of the validation group), RMSE (Root Mean Square Error) and RME (Relative Mean Error), the latter two of which indicate absolute and relative deviation of the simulation, respectively (Equation 1, 2, where T and F represent true value and forecasted value, respectively; n is the total number of validation samples).

$$\mathrm{RMSE}=\sqrt{\frac{\sum_{i=1}^{n}(T_i-F_i)^2}{n}} \tag{1}$$

$$\mathrm{RSE}=\frac{\sum_{i=1}^{n}\left|T_i-F_i\right|}{n}\times 100\% \tag{2}$$

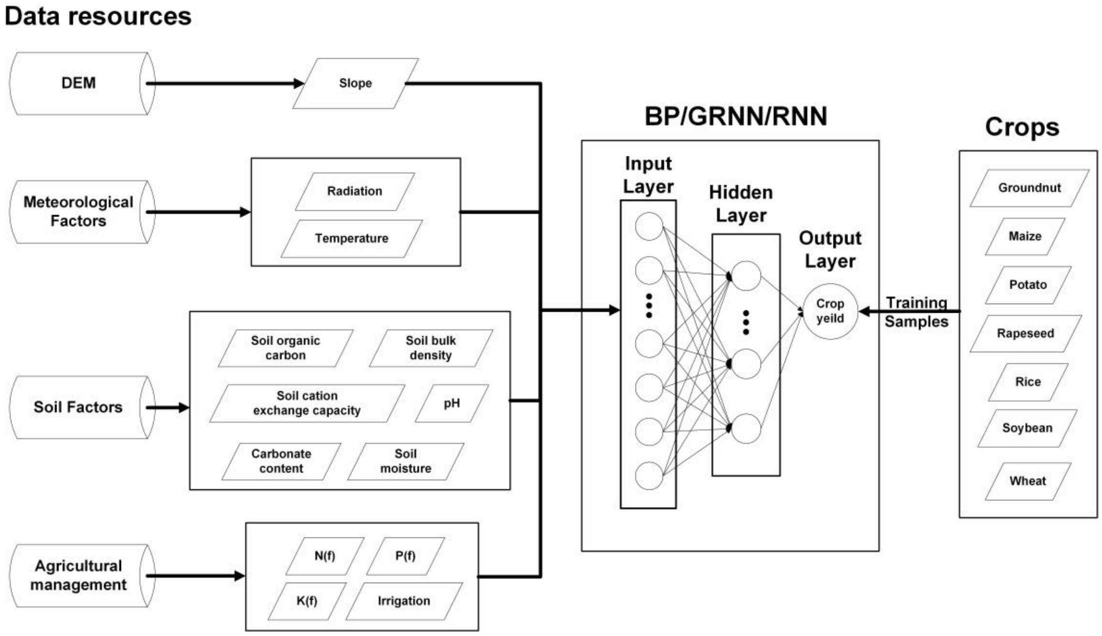

**Figure 2**. Process of simulation of crop yield using ANN

## 3. Results

### 3.1 Distribution of harvested area and agricultural management

Distribution of harvested area for the 7 species in Guizhou shows spatial variation in different prefectures (Figure 3a). Among them, Zunyi and Bijie have the largest total area of the selected crops with value of 641,807 (ha) and 586,506 (ha), respectively. Both of them are located in the northwestern Guizhou Province. On the contrary, Guiyang, Liupanshui and Anshun have the smallest total harvested area, with value of 193,178 (ha), 137,444 (ha) and 193,631 (ha), respectively. In terms of different species, maize, rice and wheat have a largest total harvest area in Guizhou, which are 653,805 (ha), 754,212 (ha) and 545,407 (ha), respectively, accounting for 22.9%, 26.5% and 19.1% of total crop area. However, the proportion of harvested area for each crop species differ greatly across different prefectures. For example, the maize area in Bijie is 166,121 (ha), which accounts for 28.3% of total area. In contrast, the maize area in Qiandongnan is 52,897 (ha), which only accounts for 9.8% of total area. All crop species received fertilizer (N, P and K) application, whereas the quantity of irrigation was species dependent (Figure 3b). Rice received the highest percentage of irrigated area (monthly mean value 52.2%), while there was no irrigation for rapeseed in the study region. From the result of fertilization, we can see the crops which are commonly cultivated (including maize, rice and wheat) tend to have higher rate of fertilization. The total amount of fertilization for these three species is 204.3 (kg/ha), 224.9 (kg/ha) and 198.4 (kg/ha), respectively. Overall, N fertilizer contributed 74.5% of fertilizer use across all crop species, followed by P fertilizer (16.5%) and K fertilizer (9.0%).

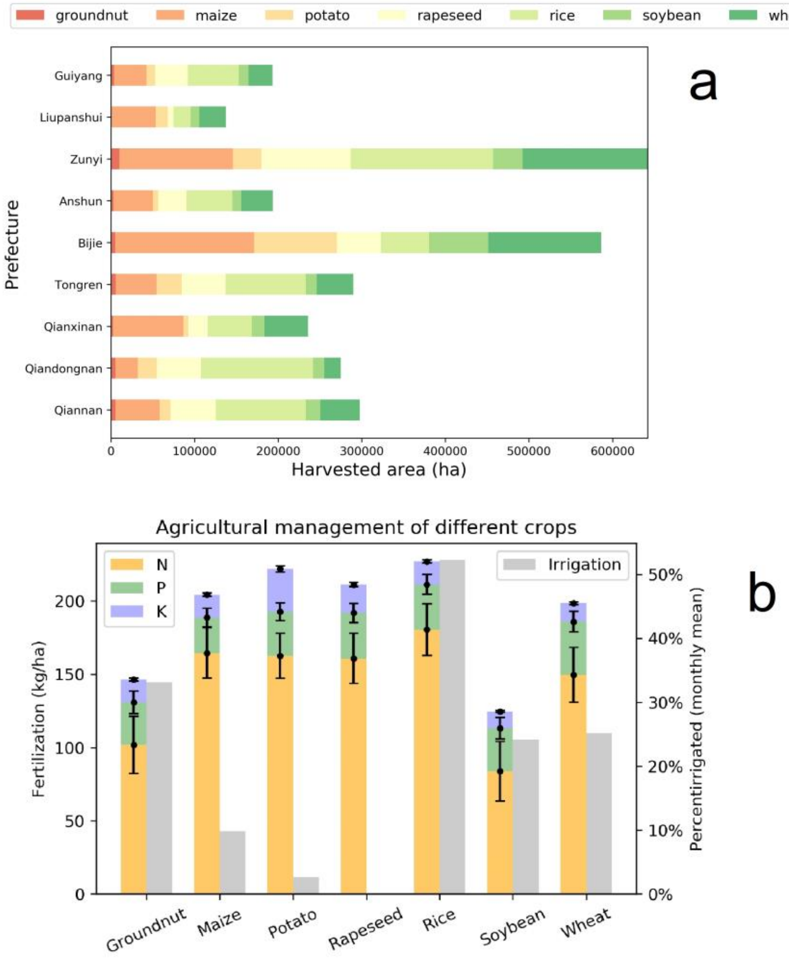

**Figure 3**. Harvested area (a), fertilization (b) and irrigation (b) of the 7 selected crop species in Guizhou province. Error bars indicates the standard deviation of value among all prefectures

### 3.2 Relationship between slope and harvested area/yield

Slope cropland is widely distributed in Guizhou Province. For all of the 7 crop species, most of the harvested area is concentrated in the slope region between 7.5° and 20°. Over 85% of harvested area is located on the slope larger than 7.5°. Among the 7 species, potato has the largest area with the slope larger than 7.5°, which accounts for 92.7% of the total harvest area. The slope of 15° is an important threshold for the implementation of the Grain for Green Program in many prefectures, with many local governments intending to remove slope cropland above 15° from production, to realize the goal of the project. From Figure 4 we can see that this management policy could impact on more than a quarter of all cropland, ranging from a minimum of 23.6% for rapeseed to a maximum of 30.1% for maize.

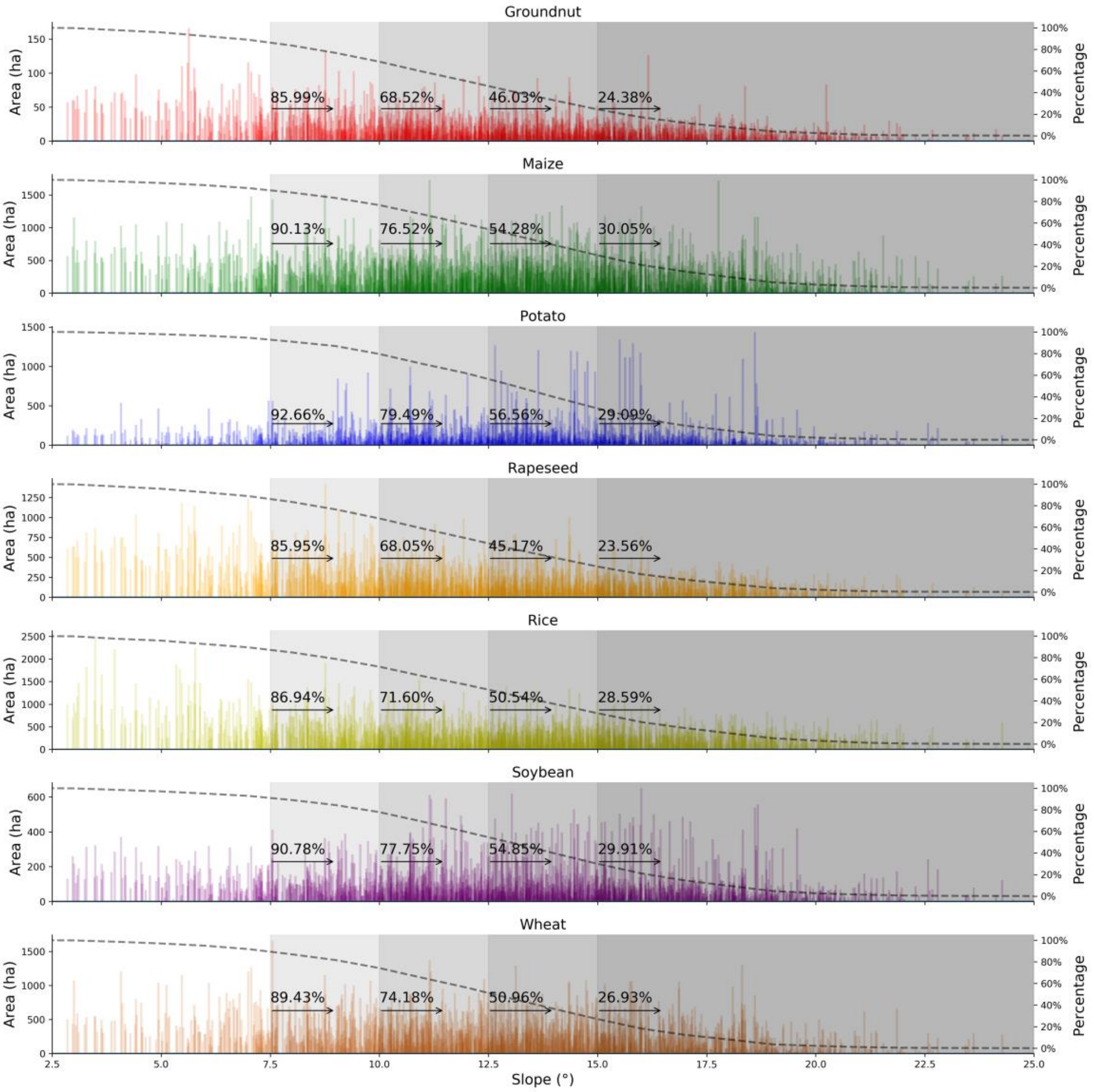

**Figure 4**. Distribution of harvested area along slope gradient (dashed line in each subplot indicates the cumulative percentage of harvested area for the area of slope greater than the x axis value).

Herein, we calculated the correlation coefficient between slope and yield per area/harvested area for the 7 crop species, based on pixel scale. As shown by figure 5, all cases show a significantly negative relationship, illustrating that with increasing slope, both yield and harvest area tend to decrease in the study region. Of the 7 species, slope has the greatest impact on the YPA of maize (with a significant R of -0.31) and the least impact on the YPA of rice (R = -0.05). However, for the total harvested area, the value of rapeseed decreases most distinctly with the increase of slope (corresponding R is -0.42), while the harvested area of potato has the least relationship with slope (corresponding R is -0.10) among all the crop species.

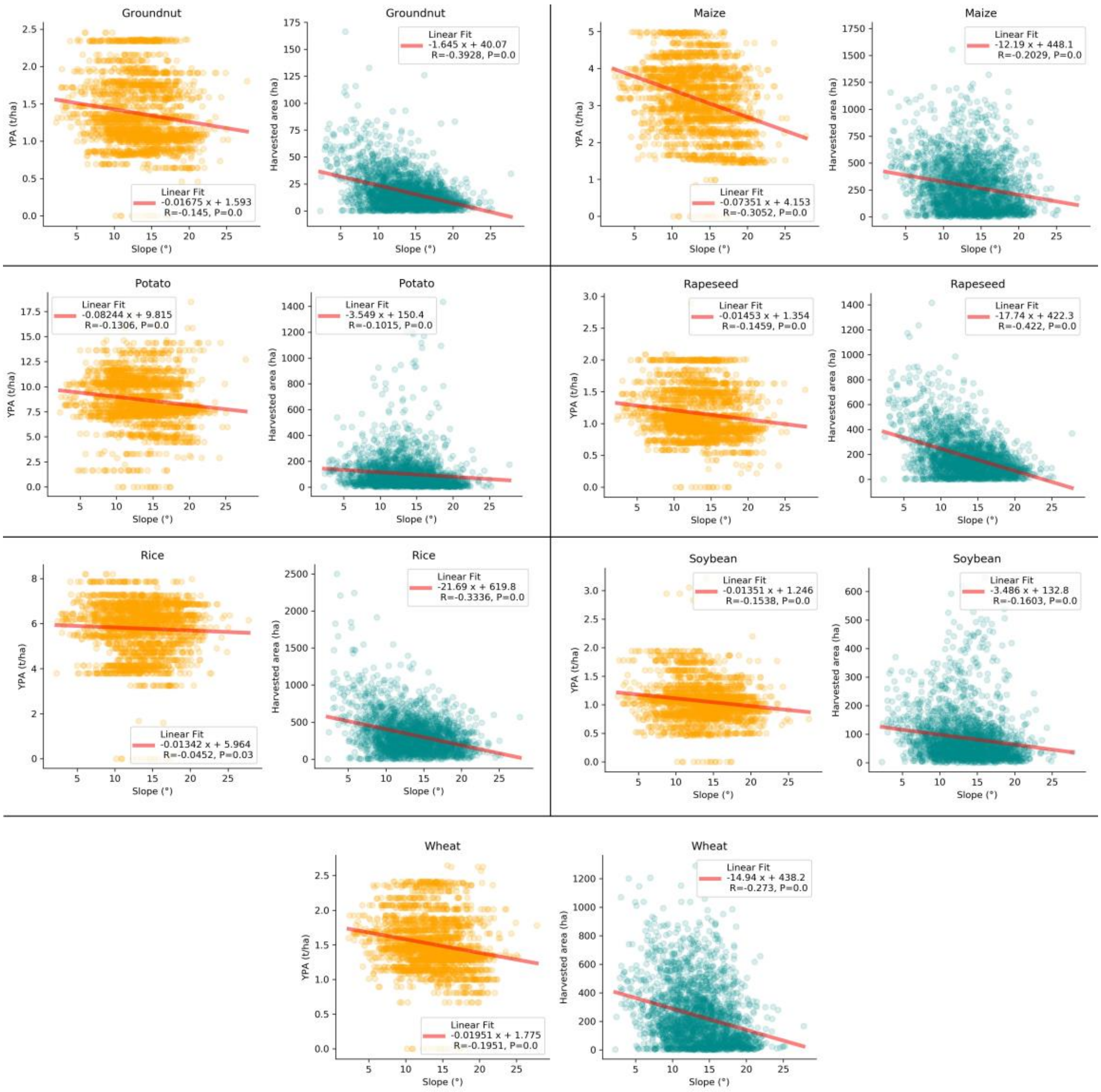

**Figure 5**. Linear regression between slope and yield per area (YPA; left panel) and total harvested area (right panel) for the 7 crop species (with all linear correlation passing significance test ($P < 0.001$), except for YPA and slope of rice ($P = 0.03$)).

## 3.3 Classification using SOFM

We used SOFM to classify Guizhou province into regions with different levels (high-, medium- and low- level) of crop production (including crop YPA, and total crop yield multiplied by corresponding harvested area; Figure 6) of the 7 species. The regions of high level YPA are mainly located in the central area of Guizhou Province, which occupies a large proportion of Guizhou and Tongren prefectures, and some of Bijie and Anshun. The prefectures of Qiannan and Qiandongnan have relatively large regions of low-level YPA, especially in the southeastern area. Some western and southwestern areas of Guizhou also have low-level characteristics. The result of total crop yield also shows similar spatial traits with crop YPA. Firstly, most regions of high-level are widely concentrated in the middle and northern area of Guizhou. Secondly, some prefectures like Qiannan and Qiandongnan also have large proportion of low-level regions, which are mainly located in the southern and southeastern area, as well as some other part located in the very eastern and northern area. However, different from that of crop YPA which has relatively large clustering of spatial distribution, the pixels with one level of total crop yield tend to be more heterogeneously distributed, resulting in the fragmentation of different levels in the whole region.

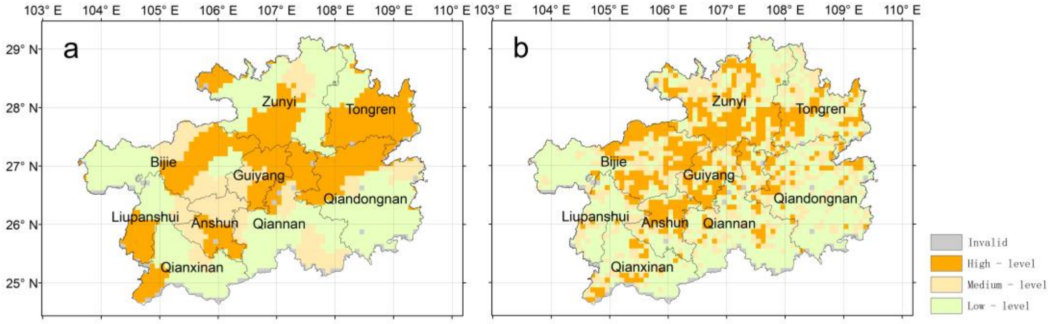

**Figure 6**. Classifying Guizhou Province spatially into different metrics of crop production: (a) crop YPA; (b) total crop yield.

## 3.4 Simulation of crop yield using three artificial neural networks

Table 2 exhibits different indices to evaluate the result of three ANNs for simulating crop yield of the 7 selected species. We randomly divided the thousands of pixels within Guizhou into two groups of training (75%) and validation (25%) and subsequently calculated the indices separately. Overall, the three kinds of networks performed well, with the correlation coefficient of R exceeding 0.40 and passing the significance test ($P < 0.001$). However, there are differences among the three networks. BP always performs the best, with R ranging from 0.87 (groundnut) to 0.65 (soybean), while GRNN and RNN have lower accuracy of the simulation. Specifically, the MRE of GRNN and RNN is relatively large, indicating a greater deviation between forecasted value and true value. For example, MRE of GRNN and RNN in simulating crop yield of rapeseed is 25.6% and 28.1%, compared to 17.8% for BP. Meanwhile, the accuracy of simulation for all the three networks is “crop-specific”, which means it tends to be easier to simulate the crop yield for some specific species. For example, the result of groundnut has the highest R of simulation within each network. On the other hand, if we compare the result of validation group and training group, it is obvious that the training group always have better accuracy in terms of R, RMSE and MRE. This is because during each iteration, the parameters of each network are adjusted based on the performance of simulation in the training group, instead of validation group. Lastly, the value of runtime for each network shows the efficiency of each simulation. GRNN has the smallest value (less than 1 second in most cases; Table 2) of runtime while RNN has the largest (all are longer than 1 minute). Although the runtime for BP was longer for each simulation than GRNN, the difference was smaller than with RNN (on average less than 7 seconds; Table 2). Therefore, based on all the indices of simulation, BP made the best balance between accuracy and temporal efficiency.

**Table 2.** Results of simulation of crop yield by three artificial neural networks

| | | Validation group | | | Training group | | | Time (s) | Total pixel included |
|---|---|---|---|---|---|---|---|---|---|
| | | R | RMSE | MRE | R | RMSE | MRE | | |
| **BP** | groundnut | 0.87 | 0.23 | 11.5% | 0.90 | 0.19 | 10.1% | 12.30 | 2287 |
| | maize | 0.75 | 0.61 | 14.2% | 0.83 | 0.52 | 11.7% | 5.84 | 2301 |
| | potato | 0.67 | 1.90 | 16.0% | 0.72 | 1.69 | 14.0% | 6.20 | 2301 |
| | rapeseed | 0.80 | 0.28 | 17.8% | 0.87 | 0.19 | 12.3% | 5.58 | 2301 |

|  |  |  |  |  |  |  |  |  |  |
|---|---|---|---|---|---|---|---|---|---|
|  | rice | 0.81 | 0.70 | 8.9% | 0.89 | 0.49 | 6.1% | 5.38 | 1250 |
|  | soybean | 0.65 | 0.27 | 16.5% | 0.76 | 0.23 | 14.4% | 5.88 | 2301 |
|  | wheat | 0.74 | 0.26 | 12.4% | 0.79 | 0.23 | 11.0% | 5.55 | 2291 |
| **Average** |  | **0.76** | **0.61** | **13.9%** | **0.82** | **0.51** | **11.4%** | **6.67** | **-** |
| **GRNN** | groundnut | 0.71 | 0.39 | 23.3% | 0.72 | 0.38 | 23.4% | 1.08 | 2287 |
|  | maize | 0.64 | 0.83 | 21.3% | 0.63 | 0.83 | 21.2% | 0.36 | 2301 |
|  | potato | 0.40 | 2.52 | 21.6% | 0.38 | 2.39 | 21.7% | 0.41 | 2301 |
|  | rapeseed | 0.65 | 0.36 | 25.6% | 0.67 | 0.33 | 23.8% | 0.39 | 2301 |
|  | rice | 0.59 | 1.13 | 14.6% | 0.58 | 1.01 | 13.3% | 0.19 | 1250 |
|  | soybean | 0.46 | 0.31 | 21.3% | 0.49 | 0.33 | 21.1% | 0.25 | 2301 |
|  | wheat | 0.46 | 0.35 | 17.8% | 0.46 | 0.36 | 19.1% | 0.30 | 2291 |
| **Average** |  | **0.56** | **0.84** | **20.8%** | **0.56** | **0.81** | **20.5%** | **0.42** | **-** |
| **RNN** | groundnut | 0.73 | 0.32 | 17.6% | 0.89 | 0.20 | 11.1% | 96.95 | 2287 |
|  | maize | 0.64 | 0.79 | 19.1% | 0.86 | 0.47 | 13.9% | 106.08 | 2301 |
|  | potato | 0.42 | 2.40 | 20.8% | 0.74 | 1.66 | 14.4% | 100.75 | 2301 |
|  | rapeseed | 0.80 | 0.39 | 28.1% | 0.89 | 0.17 | 27.0% | 93.70 | 2301 |
|  | rice | 0.43 | 1.28 | 15.3% | 0.93 | 0.41 | 5.4% | 60.70 | 1250 |
|  | soybean | 0.43 | 0.35 | 24.1% | 0.80 | 0.21 | 14.7% | 101.92 | 2301 |
|  | wheat | 0.50 | 0.35 | 17.3% | 0.80 | 0.23 | 11.4% | 113.98 | 2291 |
| **Average** |  | **0.56** | **0.84** | **20.3%** | **0.84** | **0.48** | **14.0%** | **96.30** | **-** |
| *All the value of R (correlation coefficient) with significant test result (P<0.001) |  |  |  |  |  |  |  |  |  |

### 3.5 Factor contribution

Two methods (see supplementary material) of factor contribution in BP network were analyzed to assess the relative weighting of each variable on overall crop yield, both methods reveal similar results for the 7 selected crop species (Figure 7). Among the 13 factors, temperature (16.4%), radiation (15.3%), soil moisture (13.5%), fertilization of N (13.5%) and P (12.4%) had the largest contribution to crop yield, based on the average proportion of the two methods. In contrast, slope, irrigation and other soil properties have lower mean proportions of factor contribution, ranging from 2.1% (slope) to 6.1% (pH). Compared with N and P fertilizer, K fertilizer has a relatively small impact on crop yield, with an average proportion of 3.3%. From Figure 7, we can also see there is some inter-species difference in terms of crop influencing factors. For example, for rice irrigation has the mean contribution of 12.2% on crop yield, compared to 0% contribution for rapeseed yield where no irrigation was recorded.

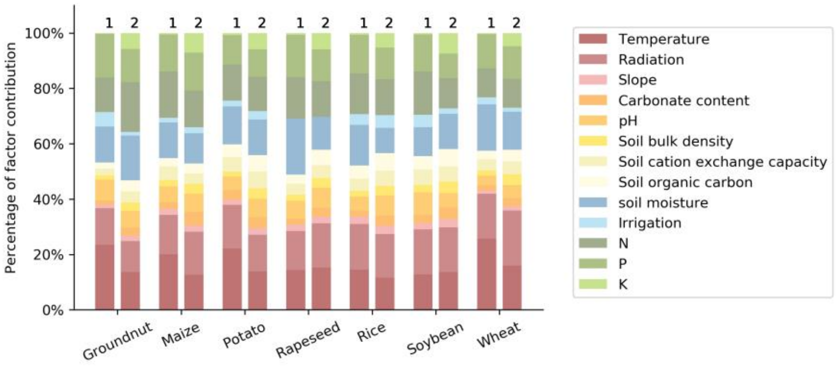

**Figure 7**. Factor contribution on crop yield of 7 species (by using two methods shown in Equation 4 and 5 in supplementary material, annotated as (1) and (2)).

## 4 Discussion

In late 1990s, the Grain for Green Program was first introduced in China (Song et al., 2015). The focus of the project has been on the potential restoration of ecosystem integrity by allowing low-yielding cropland on slopes greater than 15° to revert to natural vegetation where synthetic nutrient input has been withdrawn (Zhang et al., 2015; Wang et al., 2017). However, there has been conflict between conservation and food security, with people blaming the policy as one of the main causes for the recent surge in grain prices and rising food imports (Xu et al., 2006). Therefore, how to put this program into practice rationally is vital important for both environment and stakeholders. In this study, the distribution of cropland along elevation gradient, relationship between slope and crop yield/area, as well as the spatial region of different yield levels, can all provide an important reference in terms of the practice of Grain for Green Program and other land use policy aspects. When we carry out the program and other land-use policy, we should consider 1) distribution of slope cropland, 2) difference distribution of cropland among species, 3) potential crop yield per unit in different regions. Firstly, as the spatial distribution of harvested area is not even across different prefectures in Guizhou, some of them like Zunyi and Bijie will be mostly affected by the implementation of the policy. Secondly, the percentage of slope cropland larger than 15° is greatest for maize, which is also one of the mostly widely cultivated crops in Guizhou. Specifically, 30.1% area of maize will be impacted due to the set goal of the policy. Thirdly, from the result of SOFM, the distribution of high-level region of crop YPA and total crop yield are not strictly consistent. Thus, replacing some croplands with low potential of crop yield (like southern Qianxinan) and developing additional croplands with high potential of crop yield (like eastern Tongren) may have more benefits in terms of total crop yield.

In the past, some researchers have tried to use statistical approaches to simulate the spatial distribution of crop YPA (Drummond et al., 1995; Buchholz et al., 2004). Most of these studies were based on field-scale data. Many have relied on vegetation parameters such as NDVI (Normalized Difference Vegetation Index) or LAI (Leaf Area Index) as input factors, without adequately considering the influence of environmental factors (Doraiswamy et al.. 2004). Compared with previous research, this study included more environmental factors to simulate the spatial patterns of crop yield and examine their effect by evaluating the results of ANN. Actually, this work imported the new idea of critical zone into the study of yield crop from an angle of system science, considering multiple elements from underground (soil moisture and soil properties) to vegetation (crops of 7 species) and atmosphere (meteorological factors), to research the interaction among different elements. In our study, we employed three artificial neural networks (BP, GRNN and RNN) to conduct the simulation and relevant analyses through the power of machine learning, and the 21 networks (7*3) combined were built to finish the work. The interrelationship between crop yield and the environmental factors can be very complicated, as meteorological, lithological, soil and land management factors can all have an impact, most in nonlinear ways (Cassman, 1999; Godfray et al., 2010). Therefore, ANN can bring their superiority into full play, improving the performance of simulation as well as the credibility of factor contribution analysis. Performance varied among the networks, with BP having the best accuracy while GRNN having the least time cost. Although RNN also had acceptable accuracy, it took much longer to finish the training process. Therefore, although the usage of ANN can greatly improve the simulation, consideration in choosing the most appropriate network to balance accuracy and time cost is still needed.

In this study, we focused on the spatial distribution of crop yield and their relationship with other environmental factors, rather than research on the temporal features of these parameters. Indeed, from a temporal perspective, the change in meteorological conditions, or climate, can affect crop production through different pathways (Zhang et al., 2004; Poulter et al., 2009; Liang et al., 2019, 2020). For example, warming during the day can increase or decrease net photosynthesis (photosynthesis-respiration), depending on the measured temperature relative to the optimum temperature. A warmer temperature at night, however, can raise respiration costs without any potential benefit for photosynthesis (Lobell and Gourdji, 2012). Furthermore, a rising temperature, along with greater atmospheric $CO_2$, may favour the growth and survival of pests and diseases that target agricultural crops (Ziska et al., 2011). In addition, the response of crop yield to climate change varies with the spatial distribution pattern of the crop (Leng and Huang, 2017). From a spatial perspective, factor contribution indicated that in total 31.7% of crop yield variation was dependent on annual average temperature and radiation in the study region. Meanwhile, we also imported soil moisture (accounting for 13.5%) instead of rainfall for analysis, as rainfall may not be a direct driven factor on vegetation growth (Singh and Sasahara, 1981; Leuschner and Lendzion, 2009). On the whole, the climatic conditions provide a basic environmental background for the crop growth, which was shown by the significant influence on the spatial patterns of crop yield.

Crops have two special features that are different from natural vegetation. Firstly, most of the crops grow in the topsoil, having no direct contact with the rock below. In contrast, natural vegetation, particularly in karst regions, can grow in thin soils that would not typically be cultivated for agriculture, and sometimes even in thicker soils their roots may penetrate into fissures in weathered rock (Kosmas et al., 2000; Stehfest and Bouwman, 2006). Previous research also revealed the importance of bedrock on natural vegetation growth (Zhang et al., 2013; Jiang et al., 2020). Besides, with natural vegetation, climate is considered the most important determinant of vegetation species and distribution at the global scale. In a given region, with no obvious differentiation of climatic conditions, geomorphic features and geological substrates may influence the spatial heterogeneity of natural vegetation at smaller scales, and this influence has been verified worldwide, especially for some lithophytes (Moore and Attwell, 1999; Yetemen et al., 2010; Dasti et al., 2013). Secondly, crop growth is greatly influenced by human activities, such as fertilization, irrigation and ploughing. All of these management practices have direct impacts on soil, changing its physical and chemical properties, potentially affecting processes from deep in the critical zone that are reflected in surface vegetation (Sanchez et al., 2002; Tugel et al., 2005; García-Orenes et al., 2010). For example, irrigation strategy may be manipulated to offset the impact of insufficient precipitation in a specific time period or to address climate change impacts, thus reducing the influence from meteorological factors (Schütze and Schmitz, 2010; Da Cunha et al., 2015). This impact of agricultural management (including irrigation and fertilization) was also verified by their proportion of factor contribution (31.6% combined). As the most applied fertilizer, nitrogen and phosphatic fertilizer (N and P) had the biggest impact, accounting for 13.5% and 12.4%, respectively. This function is suggested by obvious increase of total phosphorus, potassium and other elements in the soil (Zhang et al., 2007).

In contrast, soil properties have less total impact on the spatial variation of crop yield (averaging 21.0% for the 5 factors). Amongst them, pH had the greatest influence (averaging 6.1%), as it can impact the edaphic environment by 1) controlling the activity of microorganisms and 2) changing the solubility of metals (e.g., the potentially toxicity of Al, Mn, and Cd in soils), as well as the base saturation of soil that further restricts the growth of roots (Tyler et al., 1987; Falkengren-Grerup et al., 1987; Falkengren-Grerup, 1989).

In natural vegetation, topographical variation has been shown to influence the spatial distribution of species in the karst region of southwest China (Zhang et al., 2010). Several studies have suggested that soil erosion was very severe in karst areas in southwest China due to the low soil formation rate from the carbonate bedrocks, steep sloping topography, high annual precipitation and poor vegetation cover (Lin and Zhu, 1999; Yan and Cai, 2015). Sloping cropland is widely distributed in the karst region, because of its climatic and geological features. As well as steep slopes, tillage practice can also accelerate nutrient and soil loss in the study region, causing the water, nutrient and productive capability to be reduced for crop growth (He et al., 1997; Peng and Wang, 2012). Therefore, tillage erosion and water erosion are two main factors in the reduction of crop yield on slopes, by transfer of soil materials from the upper to lower slope positions, increasing the soil depth and nutrients there (Su et al., 2010). However, this influence of topography is constrained by local meteorology. For example, with wet weather conditions, the difference of yield in low slope and high slope is distinctly larger than that with dry weather conditions (Kravchenko et al., 2000). This conclusion was also indicated by our study as karst region captures both humid climate and steep topography (Chen et al., 2009). From the results of linear regression, we observed a linear decrease in crop yield for all 7 species with increasing slope ($P < 0.001$). However, the factor contribution analysis following ANN suggested that the proportion of influence of slope is only 2.1% on average. This phenomenon shows that the impact of slope does not directly act on vegetation, but through changing soil water condition, nutrient content or some other elements inside soil. Therefore, when considering land use, we should not only focus on slope as the only index, but also include other soil variables which have more direct influence on crop growth, in order to achieve the best management practices.

In future work using ANN with remote sensing data, additional optimization could be undertaken. Firstly, during the training process of the networks, we used default values in most circumstances for optimizing hyper-parameters for the number of iterations and hidden nodes, which may affect the accuracy of simulation. However, as the difference is relatively small we felt the default settings were appropriate with this dataset. Secondly, there are constraints with the available remote sensing data. For instance, the time attribute (when the data was collected) and spatial resolution of the data sources were different, which may also cause uncertainties in comparative analyses. For instance, soil property data was available for 1995, whereas other datasets used in this study were more recent. In addition, the unrestricted use of inorganic fertilization and the unique environmental conditions of karst soils have induced a great change to the mineralogy of the soil during the past 30 years (Richardson and Kumar, 2017), which could impact on the comparability of the soil property data. Lastly, the crop yield data could also be influenced by breed improvement. The introduction of hybrid maize has improved the yield distinctly over the last decades (Ping et al., 2007; Bai et al., 2007). However, use of improved crop species varies greatly from county to county inside the province and, due to the data availability, this influence by change of breed was unavoidable.

## 5 Conclusions

The karst region of southwest China experienced rapid population and economic growth, producing many competing demands on the available soil and water resources that supports livelihoods and ensures food security. In this study, we utilized four kinds of ANNs to analyze and simulate the spatial patterns of crop yield and the relationships with meteorological factors, soil properties, irrigation and fertilization in the landscape of Guizhou Province. According to relevant analyses and results, we drew the following conclusions in this study:

1) The negative relationship between crop yield and slope of cropland is distinct. Among all species, the yield of maize decreases the fastest with the increase of slope. Meanwhile, maize (as a staple crop) has the largest percentage of cropland over 15°, this should be considered with the application of Grain for Green Program.
2) The spatial distribution of crop yield in Guizhou Province is uneven. Most high-level yield regions are located in the central-north area of Guizhou, despite some regions with high-level yield per area not being spatially consistent with those of total crop yield.
3) All crop specific artificial neural networks have significant correlation between the forecasted crop yield and true value. Among them, BP has the best performance, balancing both accuracy and time cost. From the results of factor contribution analysis, temperature, radiation, soil moisture, N and P fertilizers have the most impact on crop yield of the selected 7 species.

By combining analysis of processes occurring in the critical zone (from belowground environment to vegetation and atmosphere) with ANN modelling, the study has advanced the potential to improve and parameterize other models to simulate crop growth in karst region with high accuracy and credibility. Meanwhile, it can help to develop informed decision support tools that could be used to guide both regional land-use decisions and local farming practices to enhance crop productivity and further deliver societal good through farming practices that are more efficient, less polluting and more sustainable for food, land and water.

**Conflicts of Interest:** The authors declare no conflict of interest.

**Funding:** The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This research was funded by the National Natural Science Foundation of China under grant no. 41571130044, and the scholarship of the China Scholarship Council. The UK team were supported by the Natural Environmental Research Council MIDST-CZO project (NE/S009167/1, NE/S009175/1).

### References

Amatulli G, Domisch S, Tuanmu M-N, et al. (2018) A suite of global, cross-scale topographic variables for environmental and biodiversity modeling. *Scientific data* 5: 180040.

Anderson SP, Bales RC and Duffy CJ. (2008) Critical Zone Observatories: Building a network to advance interdisciplinary study of Earth surface processes. *Mineralogical Magazine* 72: 7-10.

Athmania D and Achour H. (2014) External validation of the ASTER GDEM2, GMTED2010 and CGIAR-CSI-SRTM v4. 1 free access digital elevation models (DEMs) in Tunisia and Algeria. *Remote Sensing* 6: 4600-4620.

BAI G-x, ZHAO Z and QIU H-b. (2007) Studies on the germplasm resource diversity of drought-resistant maize in Guizhou [J]. *Agricultural Research in the Arid Areas* 3.

Banwart S, Chorover J, Gaillardet J, et al. (2013) Sustaining Earth's critical zone basic science and interdisciplinary solutions for global challenges. *University of Sheffield, Sheffield* 48.

Banwart S, Menon M, Bernasconi SM, et al. (2012) Soil processes and functions across an international network of Critical Zone Observatories: Introduction to experimental methods and initial results. *Comptes Rendus Geoscience* 344: 758-772.

Batjes NH and Bridges E. (1994) Potential emissions of radiatively active gases from soil to atmosphere with special reference to methane: development of a global database (WISE). *Journal of Geophysical Research: Atmospheres* 99: 16479-16489.

Buchholz DD, Brown JR, Garret J, et al. (2004) *Soil test interpretations and recommendations handbook*: University of Missouri-College of Agriculture, Division of Plant Sciences ⋯.

Carabajal CC, Harding DJ, Boy J-P, et al. (2011) Evaluation of the global multi-resolution terrain elevation data 2010 (GMTED2010) using ICESat geodetic control. *International Symposium on Lidar and Radar Mapping 2011: Technologies and Applications.* International Society for Optics and Photonics, 82861Y.

Cassman KG. (1999) Ecological intensification of cereal production systems: yield potential, soil quality, and precision agriculture. *Proceedings of the national academy of sciences* 96: 5952-5959.

Chen X, Zhang Z, Chen X, et al. (2009) The impact of land use and land cover changes on soil moisture and hydraulic conductivity along the karst hillslopes of southwest China. *Environmental Earth Sciences* 59: 811-820.

Chen X, Zhang Z, Soulsby C, et al. (2018) Characterizing the heterogeneity of karst critical zone and its hydrological function: an integrated approach. *Hydrological processes* 32: 2932-2946.

Cheng J, Lee X, Theng BK, et al. (2015) Biomass accumulation and carbon sequestration in an age-sequence of Zanthoxylum bungeanum plantations under the Grain for Green Program in karst regions, Guizhou province. *Agricultural and Forest Meteorology* 203: 88-95.

Chlingaryan A, Sukkarieh S and Whelan B. (2018) Machine learning approaches for crop yield prediction and nitrogen status estimation in precision agriculture: A review. *Computers and electronics in agriculture* 151: 61-69.

Da Cunha DA, Coelho AB and Féres JG. (2015) Irrigation as an adaptive strategy to climate change: an economic perspective on Brazilian agriculture. *Environment and Development Economics* 20: 57-79.

Danielson JJ and Gesch DB. (2011) *Global multi-resolution terrain elevation data 2010 (GMTED2010)*: US Department of the Interior, US Geological Survey.

Daoxian Y. (2001) On the karst ecosystem. *Acta Geologica Sinica-English Edition* 75: 336-338.

Dasti AA, Saima S, Mahmood Z, et al. (2013) Vegetation zonation along the geological and geomorphological gradient at Eastern slope of Sulaiman range, Pakistan. 9: 6105-6115.

Davis RK, Hamilton S and Brahana JV. (2005) ESCHERICHIA COLI SURVIVAL IN MANTLED KARST SPRINGS AND STREAMS, NORTHWEST ARKANSAS OZARKS, USA 1. *JAWRA Journal of the American Water Resources Association* 41: 1279-1287.

Doraiswamy, P. C., Hatfield, J. L., Jackson, T. J., Akhmedov, B., Prueger, J., & Stern, A. (2004). Crop condition and yield simulations using Landsat and MODIS. *Remote sensing of* *environment*, 92(4), 548-559.

Drummond S, Birrell S and Sudduth KA. (1995) *Analysis and correlation methods for spatial data*: ASAE.

Everingham Y, Sexton J, Skocaj D, et al. (2016) Accurate prediction of sugarcane yield using a random forest algorithm. *Agronomy for sustainable development* 36: 27.

Falkengren-Grerup U. (1989) Soil acidification and its impact on ground vegetation. *Ambio*: 179-183.

Falkengren-Grerup U, Linnermark N and Tyler G. (1987) Changes in acidity and cation pools of south Swedish soils between 1949 and 1985. *Chemosphere* 16: 2239-2248.

Folberth C, Skalský R, Moltchanova E, et al. (2016) Uncertainty in soil data can outweigh climate impact signals in global crop yield simulations. *Nature communications* 7: 1-13.

García-Orenes F, Guerrero C, Roldán A, et al. (2010) Soil microbial biomass and activity under different agricultural management systems in a semiarid Mediterranean agroecosystem. *Soil and Tillage Research* 109: 110-115.

Godfray HCJ, Beddington JR, Crute IR, et al. (2010) Food security: the challenge of feeding 9 billion people. *science* 327: 812-818.

Grant GE and Dietrich WE. (2017) The frontier beneath our feet. *Water Resources Research* 53: 2605-2609.

Green SM, Dungait JA, Tu C, et al. (2019) Soil functions and ecosystem services research in the Chinese karst Critical Zone. *Chemical Geology*: 119107.

Guo L and Lin H. (2016) Critical zone research and observatories: Current status and future perspectives. *Vadose Zone Journal* 15.

He C, Xiong K, Li X, et al. (1997) Karst geomorphology and its agricultural implications in Guizhou, China. *Fourth International Conference On Geomorphology–Bologna, Italy.*

Huang J, van den Dool HM and Georgarakos KP. (1996) Analysis of model-calculated soil moisture over the United States (1931–1993) and applications to long-range temperature forecasts. *Journal of Climate* 9: 1350-1362.

Ibrahim YZ, Balzter H, Kaduk J, et al. (2015) Land degradation assessment using residual trend analysis of GIMMS NDVI3g, soil moisture and rainfall in Sub-Saharan West Africa from 1982 to 2012. *Remote Sensing* 7: 5471-5494.

Jiang Z, Liu H, Wang H, et al. (2020) Bedrock geochemistry influences vegetation growth by regulating the regolith water holding capacity. *Nature communications* 11: 1-9.

Kogan F, Guo W, Yang W, et al. (2018) Space-based vegetation health for wheat yield modeling and prediction in Australia. *Journal of Applied Remote Sensing* 12: 026002.

Kosmas C, Gerontidis S and Marathianou M. (2000) The effect of land use change on soils and vegetation over various lithological formations on Lesvos (Greece). *Catena* 40: 51-68.

Kravchenko AN, Bullock DG and Boast CW. (2000) Joint multifractal analysis of crop yield and terrain slope. *Agronomy journal* 92: 1279-1290.

Kumar P, Le PV, Papanicolaou AT, et al. (2018) Critical transition in critical zone of intensively managed landscapes. *Anthropocene* 22: 10-19.

Leng G and Huang M. (2017) Crop yield response to climate change varies with crop spatial distribution pattern. *Scientific Reports* 7: 1463.

Letey J. (1958) Relationship between soil physical properties and crop production. *Advances in soil science.* Springer, 277-294.

Leuschner C and Lendzion J. (2009) Air humidity, soil moisture and soil chemistry as determinants of the herb layer composition in European beech forests. *Journal of Vegetation Science* 20: 288-298.

Li D, Zhang X, Green SM, et al. (2018) Nitrogen functional gene activity in soil profiles under progressive vegetative recovery after abandonment of agriculture at the Puding Karst Critical Zone Observatory, SW China. *Soil Biology and Biochemistry* 125: 93-102.

Liang B, Dahlsjö CA, Maguire-Rajpaul V, et al. (2019) Modelling error evaluation of ground observed vegetation parameters. *IEEE Transactions on Instrumentation and Measurement*.

Liang B, Liu H, Chen X, et al. (2020) Periodic Relations between Terrestrial Vegetation and Climate Factors across the Globe. *Remote Sensing* 12: 1805.

Lin C and Zhu A. (1999) Study on soil erosion and prevention in karst mountainous region of Guizhou. *Research of Soil and Water Conservation* 6: 109-113.

Lin H, Hopmans JW and Richter Dd. (2011) Interdisciplinary sciences in a global network of critical zone observatories. *Vadose Zone Journal* 10: 781-785.

Liu F. (2006) Vegetation succession with karst rocky desertification and its impact on water chemistry of runoff. *Acta Pedologica Sin* 143: 27-32.

Lobell DB and Gourdji SM. (2012) The influence of climate change on global crop productivity. *Plant Physiology* 160: 1686-1697.

Menon M, Rousseva S, Nikolaidis NP, et al. (2014) SoilTrEC: a global initiative on critical zone research and integration. *Environmental Science and Pollution Research* 21: 3191-3195.

Monfreda C, Ramankutty N and Foley JA. (2008) Farming the planet: 2. Geographic distribution of crop areas, yields, physiological types, and net primary production in the year 2000. *Global biogeochemical cycles* 22.

Moore A and Attwell C. (1999) Geological controls on the distribution of woody vegetation in the central Kalahari, Botswana. *South African Journal of Geology* 102: 350-362.

Moore OW, Buss HL, Green SM, et al. (2017) The importance of non-carbonate mineral weathering as a soil formation mechanism within a karst weathering profile in the SPECTRA Critical Zone Observatory, Guizhou Province, China. *Acta Geochimica* 36: 566-571.

Nachtergaele F, van Velthuizen H, Verelst L, et al. (2010) The harmonized world soil database. *Proceedings of the 19th World Congress of Soil Science, Soil Solutions for a Changing World, Brisbane, Australia, 1-6 August 2010.* 34-37.

Nachtergaele F, Velthuizen H, Verelst L, et al. (2009) Harmonized World Soil Database (HWSD). *Food and Agriculture Organization of the United Nations, Rome*.

Neumann K, Stehfest E, Verburg PH, et al. (2011) Exploring global irrigation patterns: A multilevel modelling approach. *Agricultural systems* 104: 703-713.

Nguyen T, Cheng E and Findlay C. (1996) Land fragmentation and farm productivity in China in the 1990s. *China Economic Review* 7: 169-180.

Panda SS, Ames DP and Panigrahi S. (2010) Application of vegetation indices for agricultural crop yield prediction using neural network techniques. *Remote Sensing* 2: 673-696.

Peng T and Wang S-j. (2012) Effects of land use, land cover and rainfall regimes on the surface runoff and soil loss on karst slopes in southwest China. *Catena* 90: 53-62.

Ping W, LI UC, Zhongmi J, et al. (2007) The Compare Test of New Hybrid Maize Varieties [J]. *Guizhou Agricultural Sciences* 5.

Portmann FT, Siebert S and Döll P. (2010) MIRCA2000—Global monthly irrigated and rainfed crop areas around the year 2000: A new high-resolution data set for agricultural and hydrological modeling. *Global biogeochemical cycles* 24.

Poulter B, Heyder U and Cramer W. (2009) Modeling the sensitivity of the seasonal cycle of GPP to dynamic LAI and soil depths in tropical rainforests. *Ecosystems* 12: 517-533.

Prasad AK, Chai L, Singh RP, et al. (2006) Crop yield estimation model for Iowa using remote sensing and surface parameters. *International Journal of Applied Earth Observation and Geoinformation* 8: 26-33.

Ramankutty N, Evan AT, Monfreda C, et al. (2008) Farming the planet: 1. Geographic distribution of global agricultural lands in the year 2000. *Global biogeochemical cycles* 22.

Ren S, Chen X, Lang W, et al. (2018) Climatic controls of the spatial patterns of vegetation phenology in mid-latitude grasslands of the Northern Hemisphere. *Journal of Geophysical Research: Biogeosciences*.

Reynolds C, Yitayew M, Slack DC, et al. (2000) Estimating crop yields and production by integrating the FAO Crop Specific Water Balance model with real-time satellite data and ground-based ancillary data. *International Journal of Remote Sensing* 21: 3487-3508.

Richardson M and Kumar P. (2017) Critical Zone services as environmental assessment criteria in intensively managed landscapes. *Earth's Future* 5: 617-632.

Rose DC, Sutherland WJ, Parker C, et al. (2016) Decision support tools for agriculture: Towards effective design and delivery. *Agricultural systems* 149: 165-174.

Rosenzweig C, Elliott J, Deryng D, et al. (2014) Assessing agricultural risks of climate change in the 21st century in a global gridded crop model intercomparison. *Proceedings of the national academy of sciences* 111: 3268-3273.

Rötter RP, Carter TR, Olesen JE, et al. (2011) Crop–climate models need an overhaul. *Nature Climate Change* 1: 175-177.

Sanchez LA, Ataroff M and Lopez R. (2002) Soil erosion under different vegetation covers in the Venezuelan Andes. *Environmentalist* 22: 161-172.

Scholes R, Montanarella L, Brainich E, et al. (2018) IPBES (2018): Summary for policymakers of the assessment report on land degradation and restoration of the Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services.

Schütze N and Schmitz GH. (2010) OCCASION: new planning tool for optimal climate change adaption strategies in irrigation. *Journal of Irrigation and Drainage Engineering* 136: 836-846.

Shangguan W, Dai Y, Liu B, et al. (2013) A China data set of soil properties for land surface modeling. *Journal of Advances in Modeling Earth Systems* 5: 212-224.

Shi X, Yu D, Warner E, et al. (2004) Soil database of 1: 1,000,000 digital soil survey and reference system of the Chinese genetic soil classification system. *Soil Horizons* 45: 129-136.

Shi X, Yu D, Warner ED, et al. (2006) Cross-reference system for translating between genetic soil classification of China and soil taxonomy. *Soil Science Society of America Journal* 70: 78-83.

Singh M and Sasahara T. (1981) Photosynthesis and transpiration in rice as influenced by soil moisture and air humidity. *Annals of Botany* 48: 513-518.

Song W, Deng X, Liu B, et al. (2015) Impacts of grain-for-green and grain-for-blue policies on valued ecosystem services in Shandong Province, China. *Advances in Meteorology* 2015.

Stehfest E and Bouwman L. (2006) N 2 O and NO emission from agricultural fields and soils under natural vegetation: summarizing available measurement data and modeling of global annual emissions. *Nutrient Cycling in Agroecosystems* 74: 207-228.

Su W. (2002) Rare and endangered plants in guizhou karst regions with the consideration of their conservation. *Resources & Enuironment in the Yangtza Basin* 11: 111-116.

Sweeting M. (1993) Reflections on the development of Karst geomorphology in Europe and a comparison with its development in China. *Z Geomoph* 37: 127-136.

Tan S, Heerink N, Kuyvenhoven A, et al. (2010) Impact of land fragmentation on rice producers' technical efficiency in South-East China. *NJAS-Wageningen Journal of Life Sciences* 57: 117-123.

Tong X, Wang K, Yue Y, et al. (2017) Quantifying the effectiveness of ecological restoration projects on long-term vegetation dynamics in the karst regions of Southwest China. *International Journal of Applied Earth Observation and Geoinformation* 54: 105-113.

Tugel A, Herrick J, Brown J, et al. (2005) Soil change, soil survey, and natural resources decision making. *Soil Science Society of America Journal* 69: 738-747.

Tyler G, Berggren D, Bergkvist B, et al. (1987) Soil acidification and metal solubility in forests of southern Sweden. *Effects of atmospheric pollutants on forests, wetlands and agricultural ecosystems.* Springer, 347-359.

Van den Dool H, Huang J and Fan Y. (2003) Performance and analysis of the constructed analogue method applied to US soil moisture over 1981–2001. *Journal of Geophysical Research: Atmospheres* 108.

Van Wart J, Kersebaum KC, Peng S, et al. (2013) Estimating crop yield potential at regional to national scales. *Field Crops Research* 143: 34-43.

Wang B, Gao P, Niu X, et al. (2017) Policy-driven China's Grain to Green Program: Implications for ecosystem services. *Ecosystem services* 27: 38-47.

Wang SJ, Liu QM and Zhang DF. (2004) Karst rocky desertification in southwestern China: geomorphology, landuse, impact and rehabilitation. *Land degradation & development* 15: 115-121.

Weedon GP, Balsamo G, Bellouin N, et al. (2014) The WFDEI meteorological forcing data set: WATCH Forcing Data methodology applied to ERA-Interim reanalysis data. *Water Resources Research* 50: 7505-7514.

Wu B, Gommes R, Zhang M, et al. (2015) Global crop monitoring: a satellite-based hierarchical approach. *Remote Sensing* 7: 3907-3933.

Xu Z, Xu J, Deng X, et al. (2006) Grain for green versus grain: conflict between food security and conservation set-aside in China. *World Development* 34: 130-148.

Yan X and Cai Y. (2015) Multi-scale anthropogenic driving forces of karst rocky desertification in Southwest China. *Land degradation & development* 26: 193-200.

Yetemen, O., Istanbulluoglu, E., & Vivoni, E. R. (2010). The implications of geology, soils, and vegetation on landscape morphology: Inferences from semi-arid basins with complex vegetation patterns in Central New Mexico, USA. Geomorphology, 116(3-4), 246-263.

Zhang W, Chen H-S, Wang K-L, et al. (2007) The heterogeneity and its influencing factors of soil nutrients in peak-cluster depression areas of karst region. *Agricultural Sciences in China* 6: 322-329.

Zhang W, Zhao J, Pan F, et al. (2015) Changes in nitrogen and phosphorus limitation during secondary succession in a karst region in southwest China. *Plant and soil* 391: 77-91.

Zhang X, Friedl MA, Schaaf CB, et al. (2004) Climate controls on vegetation phenological patterns in northern mid-and high latitudes inferred from MODIS data. *Global change biology* 10: 1133-1145.

Zhang Z-h, Hu G and Ni J. (2013) Effects of topographical and edaphic factors on the distribution of plant communities in two subtropical karst forests, southwestern China. *Journal of Mountain Science* 10: 95-104.

Zhang Z, Huang X and Zhang J. (2020) Soil thickness and affecting factors in forestland in a karst basin in Southwest China. *Tropical Ecology*: 1-11.

Zhang ZH, Hu G, Zhu JD, et al. (2010) Spatial patterns and interspecific associations of dominant tree species in two old-growth karst forests, SW China. *Ecological research* 25: 1151-1160.

Zhao C, Liu B, Piao S, et al. (2017) Temperature increase reduces global yields of major crops in four independent estimates. *Proceedings of the national academy of sciences* 114: 9326-9331.

Zhao C, Piao S, Wang X, et al. (2016) Plausible rice yield losses under future climate warming. *Nature plants* 3: 1-5.

Zheng-An S, Zhang J-H and Xiao-Jun N. (2010) Effect of soil erosion on soil properties and crop yields on slopes in the Sichuan Basin, China. *Pedosphere* 20: 736-746.

Ziska LH, Blumenthal DM, Runion GB, et al. (2011) Invasive species and climate change: an agronomic perspective. *Climatic Change* 105: 13-42.

## Supplementary Materials: Analyzing and simulating spatial patterns of crop yield in Guizhou Province based on artificial neural networks

The four artificial neural networks (ANNs) utilized in our paper are introduced below.

### 1. SOFM

A Self-organization Feature Map is a major branch of artificial neural networks, which has self-organizing and self-learning features (Chen et al., 2014). It is trained by unsupervised learning to produce a low-dimensional (typically two-dimensional), discretized representation of the input space of the training samples, called a map, and is therefore a method for performing dimensionality reduction. The advantage of SOFM is that it can preserve the topological properties of the input space by using a neighborhood function (Liu and Song, 2005; Tian et al., 2012). It has been widely used in classification and clustering analysis (Lin and Lin, 2006; Zhang et al., 2001).

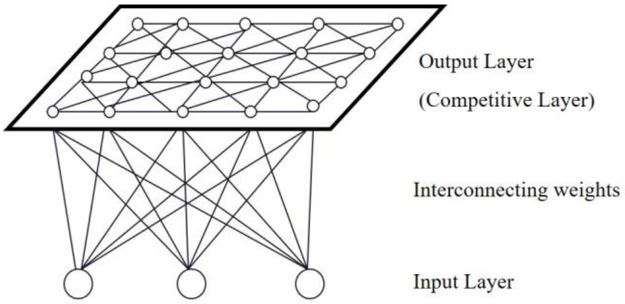

**Figure S1**. Structure of SOFM

The SOFM network (Figure 1) consists of a fully interconnected array of neurons with a topology of only two layers; the input layer and the competition layer (Kohonen, 1982). All inputs are connected to each node on the network grid, and each grid node is an output node that is only connected to adjacent nodes. That is, the input received by each neuron is the same, and each node has two weights: 1) weight of the neuron's response to the external input; and 2) weight of the connection between the neurons (controlling the magnitude of the interaction between neurons; this can be zero). The training process is completed to adjust the weight of each node in the output layer until it meets the fixed terminal condition, in order to reflect inputs in lower dimensional space and to complete the work of classification. The specific steps of the SOFM are as follows (Kohonen and Honkela, 2007):

(1) The weights in the network are initialized; each weight vector is given an initial value of a small random number and each node weight should take a different value.

(2) A sample x is randomly selected as an input in the sample dataset.

(3) The best matching unit is selected (completing the process). The weight vector that has the greatest similarity to the input vector x is selected as the winning unit, and similarity is judged by Euclidean distance, as shown below:

$$\left\| x - W_c \right\| = \min \left\| x_i - W_i \right\| \qquad (1)$$

where c represents the winning unit and i is the sequence number for x and the weight vector.

(4) Weights are updated until the terminal conditions are met, and the function for updating is as below:

$$F(i) = \exp(\frac{-\left\| p_i - p_c \right\|^2}{2\sigma^2}) \qquad (2)$$

where $p_i$ and $p_c$ are the positions of the output units of i and c, respectively, while σ is the width of the neighborhood function.

## 2. BP

Back Propagation (BP) is an algorithm widely used in the training of feedforward neural networks for supervised learning (Rumelhan and Hinton, 1986; Rumelhart et al., 1986). It is one of the most widely applied neural network models across different research disciplines (Wu et al., 2005). The algorithm uses the mean square error and gradient descent algorithm to achieve the correction of the network connection weights, and its goal is to minimize the difference between the mean square error of the actual output and the regulations output (Zhang and Lu, 2015).

The structure of the BP network includes an input layer, hidden layer and output layer. The computational methodology consists of two parts: 1) the forward propagation of information and 2) the back propagation of error. In forward propagation, the input information is transmitted from the input to the output layer through the hidden layer. Through this process, the state of each layer of neurons only affects the state of the next layer of neurons. If the desired output is not obtained at the output layer, the error change value of the output layer is calculated and then returned to the propagation process. The error signal is transmitted back along the original connection path, through the network, to modify the weights of neurons until the desired target is reached (Wu et al., 2005; Wu et al., 2011). The different nodes in the network are connected by weights, the activation function and bias. The learning algorithm is running until accuracy of the model reaches the target level. Each run of the algorithm is described as an epoch and the resultant model can be characterized by the number of epochs needed to reach a solution. In this study, we imported one hidden layer with 10 nodes and used a Tan-Sigmoid (Equation 3) function and linear function to transfer values between the different layers.

$$\mathrm{tansig}(\mathrm{x}) = \frac{2}{(1+\exp(-2\mathrm{x}))}\text{-1} \qquad (3)$$

Up to now, several methods have been raised to estimate the influence of each input variable and its contribution to the output in ANNs (Gevrey et al., 2003; Olden et al., 2004). Herein, two methods are used to calculate the factor contribution (FC) of each input factor. However, the process of normalization needs to be completed beforehand in case of disturbance from the different units. Firstly, factor contribution analysis is implemented by using the weights in each node:

$$\mathrm{FC}(i) = \frac{\sum_{j=1}^{10} |w_{ij} \times v_j|}{\sum_{i=1}^{n}\sum_{j=1}^{10} |w_{ij} \times v_j|}, i = 1, ..., n \qquad (4)$$

where n is the total number of input factors, and w is the weight of the input layer, while v is the weight of the hidden layer. Secondly, both the weights between each node and the variation of the input parameters is considered in the factor contribution using the equation:

$$\mathrm{FC(i)} = \sigma_{\mathrm{i}} \sum_{\mathrm{j=1}}^{10} |\mathrm{w_{ij}}| \qquad (5)$$

Where $\sigma_i$ is the standard deviation of the $X_i$.

## 3. GRNN

General Regression Neural Network was first proposed by Donald F. Specht in 1991, which is one kind of Radial Basis Function Neural Network (RBFNN) (Specht, 1991). GRNN has strong nonlinear mapping ability, flexible network structure, high fault tolerance and robustness, which is suitable for solving nonlinear problems. It also has more advantages in approximation ability and learning speed compared with RBFNN, especially when the quantity of sample data is small. In addition, GRNN networks can also handle unstable data (Polat and Yıldırım, 2008; Tomandl and Schober, 2001). GRNN has similar structure with RBFNN, containing four different layers (input layer, pattern layer, summation layer and output layer).

The number of neurons in the input layer is equal to the dimension of the input vector in the learning sample. Each neuron is a unit with simple distribution, and the input variables are directly passed to the pattern layer. The number of neurons in the pattern layer is equal to the number of learning samples n. Each neuron corresponds to a different sample. The transfer function of the neurons in the pattern layer is:

$$\mathrm{p_i} = \exp[-\frac{(\mathrm{X}-\mathrm{X_i})^{\mathrm{T}}(\mathrm{X}-\mathrm{X_i})}{2\sigma^2}],\ \mathrm{i=1,2,...,n} \qquad (6)$$

Accordingly, the output of neuron i is the exponential square of the squared Euclidean distance between the input variable and its corresponding sample X. The summation layer calculates the weighted sum of all neurons in the pattern layer. The weight between neuron i in the pattern layer and the neuron j is element j in output sample Yi. And the transfer function is:

$$S_{Nj}=\sum_{i=1}^{n} y_{ij}P_i,\ j=1,2,...,k \qquad (7)$$

The number of neurons in the output layer is equal to the dimension k of the output vector in the learning sample. Each neuron divides the output of the summation layer. The output of neuron j corresponds to the element j of the estimation result Y (X).

## 4. RNN

A Recurrent Neural Network is a class of artificial neural networks which also consist of three layers including input layer, hidden layer and output layer (Figure 2). In contrast to traditional feedforward neural networks, RNN have self-connected recurrent connections which model the temporal evolution (Mikolov et al., 2010; Li et al., 2016). The decision a recurrent net reaches at time step t-1 affects the decision it will reach one moment later at time step t. Therefore, recurrent networks have two sources of input, the present and the recent past, which combine to determine how they respond to new data, much as we do in life. And the output response $h_t$ of a recurrent hidden layer can be formulated as follows:

$$h_t = \theta_h(W_{xh}x_t + W_{hh}h_{t-1} + b_h) \qquad (8)$$

where $W_{xh}$ and $W_{hh}$ are mapping matrices from the current inputs $x_t$ to the hidden layer h and the hidden layer to itself. $b_h$ denotes the bias vector. $\theta_h$ is the activation function in the hidden layer.

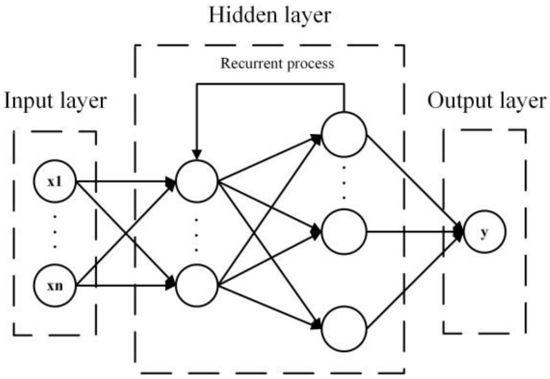

**Figure S2**. Structure of RNN

Chen Z, Li S and Yue W. (2014) SOFM neural network based hierarchical topology control for wireless sensor networks. *Journal of Sensors* 2014.

Gevrey M, Dimopoulos I and Lek S. (2003) Review and comparison of methods to study the contribution of variables in artificial neural network models. *Ecological modelling* 160: 249-264.

Kohonen T. (1982) Self-organized formation of topologically correct feature maps. *Biological cybernetics* 43: 59-69.

Kohonen T and Honkela T. (2007) Kohonen network. *Scholarpedia* 2: 1568.

Li Y, Lan C, Xing J, et al. (2016) Online human action detection using joint classification-regression recurrent neural networks. *European Conference on Computer Vision.* Springer, 203-220.

LIN J and LIN M. (2006) Research in Clustering of SOFM Neural Network [J]. *Modern Electronics Technique* 24.

LIU Y-b and SONG X-f. (2005) FUNCTION CLASSIFICATION OF SEVERAL CITIES IN THE YANGTZE DELTA BASED ON SOFM NEURAL NETWORK [J]. *Yunnan Geographic Environment Research* 6.

Mikolov T, Karafiát M, Burget L, et al. (2010) Recurrent neural network based language model. *Eleventh annual conference of the international speech communication association.*

Olden JD, Joy MK and Death RG. (2004) An accurate comparison of methods for quantifying variable importance in artificial neural networks using simulated data. *Ecological modelling* 178: 389-397.

Polat Ö and Yıldırım T. (2008) Genetic optimization of GRNN for pattern recognition without feature extraction. *Expert Systems with Applications* 34: 2444-2448.

Rumelhan D and Hinton G. (1986) R]. Williams, Learning representations by back-propagation errors. *Nature* 323: 533436.

Rumelhart DE, Hinton GE and Williams RJ. (1986) Learning internal representations by back-propagating errors in Parallel Distributed Processing: Explorations in the Microstructure of Cognition. Eds. Cambridge, MA: MIT Press.

Specht DF. (1991) A general regression neural network. *IEEE Transactions on Neural Networks* 2: 568-576.

Tian J, Tang G-a and Zhou Y. (2012) Points Group of Topographical Feature Based On SOFM. *2012 2nd International Conference on Remote Sensing, Environment and Transportation Engineering.* IEEE, 1-4.

Tomandl D and Schober A. (2001) A modified general regression neural network (MGRNN) with new, efficient training algorithms as a robust 'black box'-tool for data analysis. *Neural Networks* 14: 1023-1034.

Wu W, Feng G, Li Z, et al. (2005) Deterministic convergence of an online gradient method for BP neural networks. *IEEE Transactions on Neural Networks* 16: 533-540.

Wu W, Wang J, Cheng M, et al. (2011) Convergence analysis of online gradient method for BP neural networks. *Neural Networks* 24: 91-98.

Zhang M and Lu Y. (2015) Adaptive network traffic prediction algorithm based on BP neural network. *International Journal of Future Generation Communication and Networking* 8: 195-206.

Zhang Z-l, Sun S-h and Zheng F-c. (2001) Image fusion based on median filters and SOFM neural networks:: a three-step scheme. *Signal Processing* 81: 1325-1330.