2 Methods

2.1 Dataset and VWF estimation

We produced provinces-crops associations by using a rich dataset on the water footprint of crop production and consumption in China, which is openly accessible on (waterfootprint.org). The dataset contains annual statistics for the period 1978-2008 for 22 individual crops (wheat, maize, rice, sorghum, barley, millet, potato, sweet potato, soybean, groundnuts, sunflower, rapeseed, sugar beet, sugar cane, cotton, spinach, tomato, cabbage, apple, grapes, tea, tobacco). For each crop, The annual water footprint of both production and consumption are available at the province level of China ($n = 31$, no data in Taiwan, Hong Kong and Macau) are available. The dataset was widely used when evaluating Chinese virtual water flows (Xie et al., 2020; Sun et al., 2021; Zhuo et al., 2016b). Using the difference between the water footprint of production and consumption, the surplus is considered to be Virtual Water out-Flows (VFW):

$$\begin{cases} D_{ij} = F_{i,j}^{production} - F_{i,j}^{consuption}, & \text{if } F_{i,j}^{production} - F_{i,j}^{consuption} > 0 \\ D_{ij} = 0, & \text{if } F_{i,j}^{production} - F_{i,j}^{consuption} \geq 0 \end{cases} \tag{1}$$

In this way, we obtained the annual volume matrix of VFW D_{ij} in year y, where i and j indicate a certain province and a certain crop, $F_{production}$ and $F_{consumption}$ are virtual water embedded in crops production and consumption directly from the original dataset. Only virtual water outflows were considered, thus virtual water inputs (i.e., if consumption is larger than its consumption) were set to zero.

2.2 Construction of network

To construct a bipartite network, we reduce the VWF volume matrix D_{ij}^{y} to a binary matrix to indicate whether there is linking edge between a province and a crop (for capturing the network topology). The Relative Comparative Advantage (RCA) procedure was used to construct the bipartite network between provinces and crops. The RCA of production referred to a region's (provinces in this study) export of a particular product (crops, here) in terms of its proportion of the total trade of that product. An analogy can be made to the VWF (D_{ij}) of province i as a proportion of the national total volume of VWF ($\sum_i D_{ij}$), embedded in a specific crop j. A higher RCA indicates a larger part of shares among the national total VWF, i.e., more virtual water exported from the province through a specific crop. Firstly, for a specific year y, RCA matrix of province i and crop j are calculated according to (Balassa, 1965; Dolan et al., 2021; Sciarra et al., 2020):

$$RCA_{ij} = \frac{D_{ij} / \sum_j D_{ij}}{\sum_i D_{ij} / \sum_{ij} D_{ij}} \tag{2}$$

Then, for $y \in [1978, 2008]$, we constructed such bipartite network RCA_{ij} year by year to capture topology changes of Chinese VWF by time. Then, the network matrix M is given by $M_{ij} = 1$ if $RCA_{ij} \geq 1$, and 0 if $RCA_{ij} < 1$.

2.3 Quantitative metrics of complexity

We used a GENEPY index to distil information on the networks in reference to economic complexity (Sciarra et al., 2020):

$$GENEPY_i = \left(\sum_{x=1}^{2} \lambda_x X_{i,x}^2 \right)^2 + 2 \sum_{x=1}^{2} \lambda_x^2 X_{i,x}^2 \tag{3}$$

where $X_{i,1}$ and $X_{i,2}$ are the normalized eigenvectors of province i corresponding to the first two largest eigenvalues λ_1 and λ_2 of the proximity matrix N_{ii^*}:

$$\begin{cases} N_{ii^*} = \sum_j \dfrac{M_{ij} M_{i^* j}}{k_i k_{i^*} \left(k'_j\right)^2}, & \text{if } i \neq i^* \text{ where, } k_i = \sum_j M_{ij}, k'_j = \sum_i M_{ij}/k_i \\ N_{ii^*} = 0, & \text{if } i = i^* \end{cases} \tag{4}$$

Here, M is the constructed network matrix, k_i is the degree (how many types of crops connected in the network) of the province i and k'_j represents the degree of a crop corrected by how easily it is found within the network. The redundant information of the self-proximity (i.e., when $i = i^*$) is deleted by setting related values to zero. In addition, the symmetric square matrix N is interpreted as the mathematical description of the weighted topology of the network, -such that the provinces are the nodes and the similarities between the VWF-supporting crops are the links connecting them. Then, eigenvector centrality of the nodes (refered by the eigenvectors of matrix N) can be a useful tool to interpret complexity of the network. In practical sense, two eigenvalues vectors $X_{i,1}$ and $X_{i,2}$ of each province i can be combined into a unique metrics to distill its ability in producing highly competitive crops. Higher GENEPY index indicates a potential superiority (VWF embedded are more accessible and irreplaceable) in the current VWF networks. For more details in math we refer the reader to Sciarra et al. (2020). In this way, based on the idea of dimensionality reduction, we could use the average GENEPY index of the YRB to simply assess its importance to the virtual water of China:

$$\begin{cases} GENEPY_{YRB} = \frac{1}{n} * \sum_{i \in YRB}^{n} (GENEPY_i) \\ GENEPY_{China} = \frac{1}{N} * \sum_{i \in China}^{N} (GENEPY_i) \end{cases} \tag{5}$$

Here, $i \in YRB$ indicates that i is one of the major province that heavily rely on water resource in the YRB (Figure 1 **C.**), $GENEPY_i$ is the complexity index of the province i according to the **equation (3)**.

2.4 Decomposition of complexity

The main factors affecting the outflow capacity of regional water resources needed to be decomposed to explain the reasons for the changes of the complexity index. The Reflection Method can describe a structure of bipartite network (Hidalgo and Hausmann, 2009; Hidalgo, 2021). To a certain province i, number of existed connections $k_{i,N}$ are vary according to different reflecting times N:

$$\begin{cases} k_{i,N} = \frac{1}{k_{i,0}} \sum_j M_{ij} k_{j,N-1}, \\ k_{j,N} = \frac{1}{k_{j,0}} \sum_i M_{ij} k_{i,N-1} \end{cases} \tag{6}$$

where j represents a certain crop and i a certain province, M_{ij} defines the network, and $k_{\mathrm{i},0}$ represents the observed levels of diversification of a province (the number of products exported by that province). When reflecting times N takes different integer values, it adds further potential connections to the total number in the previous layer (i.e., $N-1$) of the network. We therefore characterized each province p through the vector $\boldsymbol{k_i}(\boldsymbol{k_{i,1}}, \boldsymbol{k_{i,2}}, \boldsymbol{k_{i,3}})$ in its different dimensions. For example, $N=1$, with initial conditions given by the degree (i.e., the number of links of provinces and crops): $k_{i,0}=k_i$, same as $k_i=\sum_j M_{ij}$ in **equation 4**.

Table 1. Interpretations of the first three pairs of variables describing the province-crop network through the method of reflections.

	Definition	Working name	Description
N=1	$k_{i,1}$	Diversification	How many products are exported by province i?
N=2	$k_{i,2}$	Uniqueness	How common are the crops exported by province i?
N=3	$k_{i,3}$	Competitiveness	How diversified are provinces exporting crops similar to those of province i?

With reference to existing complexity studies, the first three major dimensions can therefore be intuitively explained when $N=1$, $N=2$ and $N=3$ as summarized in Table 1. When the reflection method is used in this way, it reflects the crop diversification, crop uniqueness and regional competitiveness of the water footprint outflow of the YRB. Although we could have used the reflection approach to continue iterating for more complex explanations, the decomposition of complexity into three steps helped explain the changes in complexity more clearly and intuitively.

2.5 Null models and sensitivity tests

As a sensitivity test, we randomly created provincial-crop bipartite networks, and we calculated the same metrics as a comparable reference values to decide wheather the structure of networks was trivial. We imagined three scenarios that randomly generated (executed by Python 3.9 and numpy 1.2) comparable dichotomies based on progressively stricter assumptions. They were consistent with the network based on empirical data (M_{ij}) of the number of edges and the sequence of edges on a side (province i or crop j) respectively (Table 2):

Table 2. How different null models were generated.

	Null model 1	Null model 2	Null model 3
Number of links	$=M_{ij}$	$=M_{ij}$	$=M_{ij}$
$k_{i,0}$ sequence	$\neq M_{ij}$	$\neq M_{ij}$	$=M_{ij}$
$k_{j,0}$ sequence	$\neq M_{ij}$	$=M_{ij}$	$\neq M_{ij}$