Reply to Referee #1

We would like to thank Referee #1 for his/her time and effort in reviewing our manuscript, titled 'An improved Grassberger-Procaccia algorithm for analysis of climate system complexity' (ID: hess-2017-445). Your comments and suggestions are much appreciated. Please see our responses in the following section.

Comment 1. For readers to quickly catch your contribution, it would be better to highlight major difficulties and challenges, and your original achievements to overcome them, in a clearer way in abstract and introduction.

Response: Thank you for the comment here. In Introduction, we have stated some of the major problems associated with the current methods for computing correlation dimensions (e.g., page 1, lines 11-12 of the manuscript 'the use of this method is still not adaptive and relies heavily on subjective criteria', and lines 53-54, 'However, the G-P method relies on visual inspections for choosing scaling regions, which is subject to human errors (Sprott and Rowlands, 2001)'. To deal with this important problem, we tried to find 'more objective and adaptive algorithms for identifying scaling regions to obtain more accurate estimates of correlation dimensions' (lines 59-60). Nonetheless, based on the reviewer's comment here, we highlighted our contribution for computing correlation dimensions in Abstract and Introduction.

Changes in the manuscript

Page 1, lines 11-12: However, the use of this method is still not adaptive and the choice of scaling regions relies heavily on subjective criteria.

Page 2, lines 57-60: However, these existing methods for identifying scaling regions had the following problems: (1) the computing processes are still not adaptive and the choice of scaling regions relies heavily on subjective criteria, and (2) the use of the least squares method for fitting straight lines to determine correlation exponents can include outliers (Cantrell, 2008) and thus is not optimal.

A reference was added on page 13, lines 329-331:

Cantrell, C. A.: Technical Note: Review of methods for linear least squares fitting of data and application to atmospheric chemistry problems, Atmos. Chem. Phys., 8, 5477-5487, https://doi.org/10.5194/acp-8-5477-2008, 2008.

Comment 2. It is shown in the reference list that the authors have several publications in this field. This raises some concerns regarding the potential overlap with their previous works. The authors should explicitly state the novel contribution of this work, the similarities and the differences of this work with their previous publications.

Response: Thank you for the comment. First, the novelty of our current work as compared to previous studies is discussed in details in Abstract and Methodology sections. Secondly, the studies cited in the reference list (presumably with the last name of Wang) are done by others and not published by the current authors.

Comment 3. It is mentioned in p.1 that an improved Grassberger-Procaccia algorithm is adopted for analysis of climate system complexity. What are the other feasible alternatives? What are the advantages of adopting this particular algorithm over others

in this case? How will this affect the results? More details should be furnished.

Response: Thank you for the suggestion. In fact, we have given some alternative methods for studying climate system complexity, such as chaos theory, wavelet analysis, and dynamical analysis (see page 1, lines 34-35). In particular, for computing correlation dimensions, we also compared our newly proposed algorithm to two other commonly used algorithms, namely the intuitive judgment and the point-based *K*-means clustering methods (see page 5, lines 154-155), based on two classical chaotic systems.

Moreover, to address this comment as well as following comments made by the reviewer, we feel that in a single paper with limited space, it is not feasible and appropriate to include every aspect existing in the field of complexity analysis, which would deviate from the central theme of this study and make the manuscript unnecessarily excessive. In fact, there are several excellent books that are devoted to entirely discussing relevant problems (e.g., Bellie Sivakumar, 2017). We added this book to the reference list for the convenience of readers.

Changes in the manuscript

Page 15, lines 415-416:

Sivakumar, B.: Chaos in Hydrology: Bridging determinism and stochasticity. Sydney, Springer: Netherlands, https://doi.org/10.1007/978-90-481-2552-4, 2017.

Page 4, 114-116 and page 6, lines 167-171: We added more details to describe the advantage of our improved method in our revised manuscript.

Comment 4. It is mentioned in p.2 that Lorenz and Henon chaotic systems are adopted to test the effectiveness of the proposed algorithm for estimating correlation dimensions. What are the other feasible alternatives? What are the advantages of adopting these particular systems over others in this case? How will this affect the results? More details should be furnished.

Response: Thank you for this comment. Indeed, there are other chaotic systems (e.g., the Chen system, and the Rössler system. Among those chaotic systems, the Lorenz and Henon systems with existing theoretical correlation dimensions have been mostly studied in the past, and thus used to analyze the chaotic behavior in climate systems and to test the effectiveness of algorithms for computing climate system complexity (e.g., Grassberger and Procaccia, 1983a; Lai and Lerner, 1998; Ji et al., 2011). In our opinion, for the purpose of brevity and more importantly comparison among different studies and methods for computing climate system complexity, it is justified that standard systems, such as the Lorenz and Henon systems, should be adopted. Finally, the discussion on different chaotic systems is beyond the scope of this study. It would be unrealistic for us to compare all chaotic systems in one single paper. According to your suggestion, we added more details and the following references in our revised manuscript.

Changes in manuscript

The following sentences were added on page 5, lines 145-148:

The Lorenz and Henon systems with existing theoretical correlation dimensions have been mostly studied in the past, and thus used to analyze the chaotic behavior in

climate systems and to test the effectiveness of algorithms for computing climate system complexity (e.g., Grassberger and Procaccia, 1983a; Lai and Lerner, 1998; Ji et al., 2011).

A reference was added on page 14, lines 369-270:

Lai, Y. C., Lerner, D.: Effective scaling regime for computing the correlation dimension from chaotic time series, Physica D, 115, 1-18, https://doi.org/10.1016/S0167-2789(97)00230-3, 1998.

Comment 5. It is mentioned in p.2 that the Haihe River Basin is adopted as the case study. What are other feasible alternatives? What are the advantages of adopting this particular case study over others in this case? How will this affect the results? The authors should provide more details on this.

Response: Thank you for the comment here. The reasons that we took the Haihe River Basin (HRB) as a case study are both practical and theoretical: (1) The HRB has been facing serious water shortage due to climate change and increasing water demands. Although previous studies have investigated the climate variability (e.g., rainfall, air temperature, and evaporation) in the HRB from different perspectives, to our best knowledge, there are still no attempts to quantify nonlinear characteristics of climatic variables, especially regarding their chaotic behaviors in the HRB, which is essential for understanding the nonlinearity of the climate system in the region; and (2) The HRB is a diverse hydroclimatic region with many sub-watersheds of varying geographical and hydroclimatic conditions, which make the region ideal for understanding the climate system complexity. We added more details about the advantages of adopting this particular case in our revised manuscript.

Changes in manuscript

Page 2, lines 68-75: Although previous studies have investigated the climate variability (e.g., rainfall, air temperature, and evaporation) in the HRB from different perspectives, to our best knowledge, there are still no attempts to quantify nonlinear characteristics of climatic variables, especially regarding their chaotic behaviors in the HRB, which is essential for understanding the nonlinearity of the climate system in the region. Furthermore, the HRB is a diverse hydroclimatic region with many sub-watersheds of varying geographical and hydroclimatic conditions, which make the region ideal for understanding the climate system complexity.

Comment 6. It is mentioned in p.3 that the normal-based K-means clustering technique is adopted to partition all normals of the scatter points into K clusters with high similarity. What are other feasible alternatives? What are the advantages of adopting this particular technique over others in this case? How will this affect the results? The authors should provide more details on this.

Response: Thank you for the suggestion. We had provided some explanations in Section 2.2 and Section 3. The *K*-means clustering method is used to partition *n* observations into *K* clusters. For each cluster, each observation belongs to the cluster with the nearest mean. In this paper, in order to find a precise scaling region, we used the normal based *K*-means clustering algorithm to remove the points that were

obviously located outside of the real scaling region (see Section 2.2). Different from previous *K*-means methods (e.g., the point-based *K*-means clustering method), we measured the similarity of points using the normal-based *K*-means clustering technique (e.g., quantifying the diversity between normals of different points). This is because the normal directions of different points in Figure 4(a) are greatly different. By comparison, the distance between points is much less, due to the use of the logarithmic scale that makes the points more densely distributed as ln *r* goes backward (see Fig 3(a)). Therefore, we proposed to use the normal-based *K*-means clustering algorithm. As a comparison, taking the classical chaotic models of Lorenz and Henon as two examples, the results obtained by our proposed normal-based *K*-means method outperformed those from the point-based *K*-means method (see Table 1). To illustrate this more clearly, we added some sentences to show the advantages of normal-based *K*-means method.

Changes in manuscript

The following sentences were added on page 6, lines 167-171: Different from previous *K*-means methods (e.g., the point-based *K*-means clustering method), we measured the similarity of points using the diversity between normals of different points. The reason for using the normal-based method is that the directions of normals for different points may vary considerably (See Fig. 4b); whereas, for the point-based *K*-means method, the distance between different points might be small, making it difficult to separate the points into different clusters (Fig. 3a).

Comment 7. It is mentioned in p.4 that the Random Sample Consensus algorithm is adopted to fit a straight line through the log-transformed points. What are other feasible alternatives? What are the advantages of adopting this particular technique over others in this case? How will this affect the results? The authors should provide more details on this.

Response: Thank you for the comment. We have given the reasons for choosing the Random Sample Consensus algorithm (RANSAC) in Section 2.2 (page 4, 114-116). As shown in Section 2.2, the RANSAC algorithm outperformed the commonly used least squares method for linear fitting, based on a hypothetical example (Fig. 1).

Comment 8. It is mentioned in p.6 that the intuitive judgment method and the point-based K-means clustering method are adopted to compare the results obtained by the proposed method. What are the other feasible alternatives? What are the advantages of adopting these particular methods over others in this case? How will this affect the results? More details should be furnished.

Response: Thank you for the comment. The intuitive judgment method and the point-based *K*-means clustering method are two commonly used methods for identifying scaling region (e.g., Sprott and Rowlands, 2001; Ji et al., 2011). Although more comparisons can be done, additional comparisons may seem redundant. In addition, it is unrealistic to list all the comparisons in one single paper.

Comment 9. It is mentioned in p.6 that the normal-based K-means clustering

technique is adopted to determine the scaling regions of the curves in Fig. 3a. What are other feasible alternatives? What are the advantages of adopting this particular technique over others in this case? How will this affect the results? The authors should provide more details on this.

Response: This comment is the same as the comment 6. We had provided some explanations in Section 2.2 and Section 3. To illustrate this more clearly, we added some sentences to show the advantages of normal-based *K*-means method.

Changes in manuscript

The following sentences were added on page 6, lines 167-171: Different from previous *K*-means methods (e.g., the point-based *K*-means clustering method), we measured the similarity of points using the diversity between normals of different points. The reason for using the normal-based method is that the directions of normals for different points may vary considerably (See Fig. 4b); whereas, for the point-based *K*-means method, the distance between different points might be small, making it difficult to separate the points into different clusters (Fig. 3a).

Comment 10-11. 10. Some key parameters are not mentioned. The rationale on the choice of the particular set of parameters should be explained with more details. Have the authors experimented with other sets of values? What are the sensitivities of these parameters on the results? 11. Some assumptions are stated in various sections. Justifications should be provided on these assumptions. Evaluation on how they will affect the results should be made.

Response: Thank you for this comment. We rechecked the paper and found that the ranges of *r* were missing. We added more details in our revised manuscript (see lines 94-96). Other parameters have been given in the paper. We must point out that some of the parameters in this study were determined by routinely used methods. For example, the time delay (see line 86) was determined by the autocorrelation function. Some other parameters (for example, T=5 ℃) were determined by testing the data. In terms of the assumption about the value *r*, we added it in our paper.

Changes in manuscript

The following sentences were added on page 3, lines 94-96: Set r_{min} and $r_{\max}$ as the minimum and maximum distances between points, respectively (Ji et al, 2011; Lai and Lerner, 1998). If $r \leqslant r_{min}$, none of the vector points falls within the volume element and $C(r, m)=0$. Otherwise, if $r \geqslant r_{\max}$, all vector points fall within the volume element and $C(r, m)=1$.

Comment 12. The discussion section in the present form is relatively weak and should be strengthened with more details and justifications.

Response: Considering that this is a technical paper, we limited our discussions for the purpose of brevity. We added more details and justifications in our revised version. Please see the following for details.

Changes in manuscript

Add sentences on page 10, lines 261-262:

The HRB is located in a monsoon-dominated region, where the EASM plays a leading

role in the regional meteorological system.

Add sentences on page 10, lines 264-268:

Wang et al. (2011) revealed that large-scale atmospheric circulations had close relationships with precipitation patterns in the HRB by analyzing the moisture flux derived from NCAR/NCEP reanalysis data. Influenced by the large-scale atmospheric circulation, precipitation in the middle and southeast parts of the HRB is more sensitive to climate variability due to their locations closer to the ocean.

Add sentences on page 11, lines 272-279:

Furthermore, at the north corner of the HRB, the westerlies primarily affect the hydrometeorological system and thus weaken the impact of the EASM on precipitation (Li et al., 2017). In addition, other factors (e.g., topography, vegetation distribution, and human activity) may also have impacts on regional patterns of climate variables. In particular, the Yan-Taihang mountain located in the northwest HRB obstructs the vapor transport driven by the EASM, resulting in lower spatiotemporal variability in precipitation in the north part of the HRB. As a result, precipitation had higher degrees of complexity in the southern HRB, while its complexity was lower in the mountainous area in the northwest HRB.

Add the following references on page 14, lines 375-377, and page 15, lines 431-433:

Li, F. X., Zhang, S. Y., Chen, D., He, L., and Gu, L. L.: Inter-decadal variability of the east Asian summer monsoon and its impact on hydrologic variables in the Haihe River Basin, China, J. Resour. Ecol., 8(2), 174-184, https://doi.org/10.5814/j.issn.1674-764X.2017.02.008, 2017.

Wang, W. G., Shao, Q. X., Peng, S. Z., Zhang, Z. X., Xing, W. Q., An, G. Y., and Yong, B.: Spatial and temporal characteristics of changes in precipitation during 1957-2007 in the Haihe River basin, China. Stoch. Environ. Res. Risk Assess., 25(7), 881-895, https://doi.org/10.1007/s00477-011-0469-5, 2011.

Comment 13. The manuscript could be substantially improved by relying and citing more on recent literatures about real-life case studies of contemporary soft computing techniques in hydrological engineering such as the followings: Gholami, V., Chau, K. W., Fadaee, F., Torkaman, J., and Ghaffari, A. (2015). "Modeling of groundwater level fluctuations using dendrochronology in alluvial aquifers." J. Hydrol., 529, 1060–1069. Taormina, R., Chau, K.W., Sivakumar, B.: Neural network river forecasting through baseflow separation and binary-coded swarm optimization", Journal of Hydrology 529 (3): 1788-1797 2015. Wu, C. L., Chau, K. W., Fan, C.: Prediction of rainfall time series using modular artificial neural networks coupled with data-preprocessing techniques, Journal of Hydrology 389(1-2): 146-167, 2010. Wang W. C., Chau, K. W., Xu, D. M., Chen, X., Y., Improving forecasting accuracy of annual runoff time series using ARIMA based on EEMD decomposition, Water Resources Management 29 (8): 2655-2675 2015. Chen, X. Y., Chau, K. W., Busari, A. O., A comparative study of population-based optimization algorithms for downstream river flow forecasting by a hybrid neural network model," Engineering Applications of Artificial Intelligence 46 (A): 258-268 2015. Chau, K. W., Wu, C. L., "A Hybrid Model Coupled with Singular Spectrum Analysis for Daily Rainfall Prediction,"

Journal of Hydroinformatics 12 (4): 458-473 2010.

Response: Thank you for providing the relevant references for further modification of our paper, and we have read them and cited one of them in the revised paper.

Changes in manuscript

The following reference was added on page 15, lines 439-441:

Wu, C. L., Chau, K. W., and Fan, C.: Prediction of rainfall time series using modular artificial neural networks coupled with data-preprocessing techniques, J. Hydrol., 389(1-2), 146-167, https://doi.org/10.1016/j.jhydrol.2010.05.040, 2010.

Comment 14. In the conclusion section, the limitations of this study, suggested improvements of this work and future directions should be highlighted.

Response: Thank you for this comment. We added the limitations and future work of this study in the conclusion section.

Changes in manuscript

Add a paragraph on page 12, lines 301-310:

The modified G-P algorithm proposed in this study can be used more objectively to characterize the complexity of climate systems (and other hydrological systems, such as streamflow, soil moisture, and groundwater), and thus provide a more reliable estimate of the number of dominant factors governing climate systems. Theoretically, it can provide valuable information for optimizing the number of parameters in climate models to reduce computational demands and model parameter uncertainties. Furthermore, the findings of this study can be used for the regionalization of hydrometeorological systems in the HRB, which has important significance in prediction in ungaged areas (Lebecherel et al., 2016). It should be noted that more studies are still required to verify the present results using other nonlinear techniques, such as the Lyapunov exponent (Wolf et al., 1985) and the approximate entropy (Pincus, 1995), which might provide additional insights into climate complexity analysis.

Add the following references on page 14, lines 373-374, 400-401, and page 15, lines 437-438:

Lebecherel, L., Andreassian, V., and Charles, P.: On evaluating the robustness of spatialproximity-based regionalization methods, J. Hydrol., 539, 196-203, https://doi.org/10.1016/j.jhydrol.2016.05.031, 2016.

Pincus, S.: Approximate entropy (ApEn) as a complexity measure, Chaos, 1995, 5(1), https://doi.org/10.1063/1.166092, 110.

Wolf, A., Swift, J. B., Swinney, H. L.: Determining Lyapunov exponents from a time series, Physica D Nonlinear Phenomena, 1985, 16(3), https://doi.org/10.1016/0167-2789(85)90011-9, 285-317.

Reply to Referee #2

We would like to thank the Referee #2 for his/her time and effort in reviewing our manuscript, titled 'An improved Grassberger-Procaccia algorithm for analysis of climate system complexity' (ID: hess-2017-445). Your comments and suggestions are much appreciated. Please see our responses in the following section.

Comment 1. Section 2.1 Algorithm for Computing Correlation Dimension may be reduced as correlation dimension is relatively old.

Response: Thank you for this comment. Section 2.1 introduces the original G-P algorithm. We can point out the problems existing in the traditional algorithm. Furthermore, Section 2.2 is based on Section 2.1.

Changes in manuscript

To shorten Section 2.1, the following were revised:

(1) Remove the sentence: The dimension of the time series of a variable is indicative of the number of factors governing the underlying dynamical processes;

(2) Page 3, lines 101-102 are modified as: According to the relationship between $D_2(m)$ and m, the saturation value of $D_2(m)$ is defined as the correlation dimension.

Comment 2. Lines 117-119 and Figure 1: Authors compared equations in terms of y =0.5x. What is R square value for both the equations and this also can be taken into consideration while judging superiority of methods.

Response: Thank you for this suggestion. Indeed, adding R square value is better for evaluating the fitting results. We added R square value in Fig. 1 and in the text.

Changes in manuscript

Page 4, lines 121-123: line 121 (y=0.60 x-0.068; R^2=0.854) , line 123 (y=0.49 x+0.007; R^2=0.990)

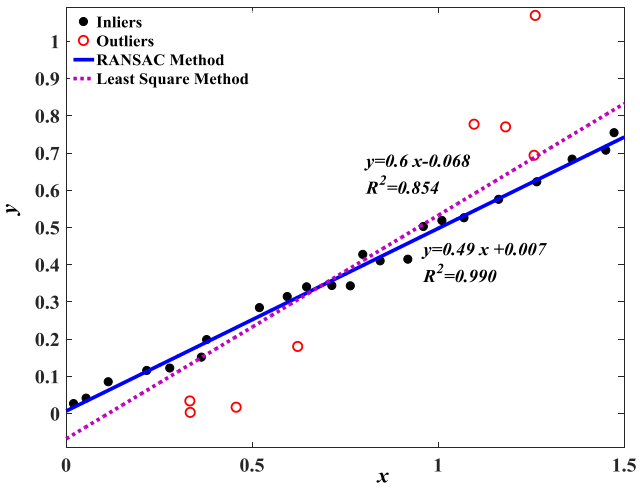

Figure 1: Comparison of the fitted lines obtained from the RANSAC algorithm and the least squares method.

Comment 3. Line 125-140: Detailed information is provided and flow chart is also presented.

Detailed information can be reduced as the flow chart is self explanatory.

Response: Thank you for this comment. According to your suggestion, we reduced the relevant section.

Changes in manuscript

Page 5, lines 129-139 were revised as follows:

The flow chart of the proposed procedures for calculating correlation dimensions is given in Fig. 2, which consists of five major steps. (1) For the time series $x(t)$, the time delay τ is computed by an autocorrelation function (Liebert and Schuster, 1989). Then set the minimum embedding dimension m_{min}=2 and reconstruct the phase space by increasing m to obtain the correlation exponent function $C(r, m)$; (2) The normals of the scatter points on the ln r~ln $C(r, m)$ line are estimated via principal component analysis (Mitra et al., 2004); (3) The K-means clustering technique is performed on the normal set N with K=2 to obtain two different clusters. Set a threshold value T to determine the angle α between the two clusters. If $\alpha>T$, the data set with larger differences in normals is discarded. Then, the K-means clustering technique is repeated on the remaining data set until $\alpha \leq T$; (4) The RANSAC algorithm is used to fit a straight line through the set of remaining scatter points; and (5) The slope of the line obtained from the RANSAC method is computed to acquire the correlation dimension $D_2(m)$ for each m. Finally, the saturation correlation dimension is determined using the plot $D_2(m)$ vs. m.

Comment 4. Figure 8: More discussion will help to understand the figure effectively.

Response: Thank you for this comment. Considering that this is a technical paper, we limited our discussions for the purpose of brevity. We added more discussion (i.e., the sentences in red color) about this figure.

Changes in manuscript

Add sentences on page 10, lines 261-262:

The HRB is located in a monsoon-dominated region, where the EASM plays a leading role in the regional meteorological system.

Add sentences on page 10, lines 264-268:

Wang et al. (2011) revealed that large-scale atmospheric circulations had close relationships with precipitation patterns in the HRB by analyzing the moisture flux derived from NCAR/NCEP reanalysis data. Influenced by the large-scale atmospheric circulation, precipitation in the middle and southeast parts of the HRB is more sensitive to climate variability due to their locations closer to the ocean.

Add sentences on page 11, lines 272-279:

Furthermore, at the north corner of the HRB, the westerlies primarily affect the hydrometeorological system and thus weaken the impact of the EASM on precipitation (Li et al., 2017). In addition, other factors (e.g., topography, vegetation distribution, and human activity) may also have impacts on regional patterns of climate variables. In particular, the Yan-Taihang mountain located in the northwest HRB obstructs the vapor transport driven by the EASM, resulting in lower spatiotemporal variability in precipitation in the north part of the HRB. As a result, precipitation had higher degrees of complexity in the southern HRB, while its

complexity was lower in the mountainous area in the northwest HRB.

Add the following references on page 14, lines 375-377, and page 15, lines 431-433:

Li, F. X., Zhang, S. Y., Chen, D., He, L., and Gu, L. L.: Inter-decadal variability of the east Asian summer monsoon and its impact on hydrologic variables in the Haihe River Basin, China, J. Resour. Ecol., 8(2), 174-184, https://doi.org/10.5814/j.issn.1674-764X.2017.02.008, 2017.

Wang, W. G., Shao, Q. X., Peng, S. Z., Zhang, Z. X., Xing, W. Q., An, G. Y., and Yong, B.: Spatial and temporal characteristics of changes in precipitation during 1957-2007 in the Haihe River basin, China. Stoch. Environ. Res. Risk Assess., 25(7), 881-895, https://doi.org/10.1007/s00477-011-0469-5, 2011.

Comment 5. Utility of estimation of correlation dimensions for the future work in HRB can be briefly mentioned.

Response: Thank you for this comment. We added the limitations and future work of this study in the conclusion section.

Changes in manuscript

Add a paragraph on page 12, lines 301-310:

The modified G-P algorithm proposed in this study can be used more objectively to characterize the complexity of climate systems (and other hydrological systems, such as streamflow, soil moisture, and groundwater), and thus provide a more reliable estimate of the number of dominant factors governing climate systems. Theoretically, it can provide valuable information for optimizing the number of parameters in climate models to reduce computational demands and model parameter uncertainties. Furthermore, the findings of this study can be used for the regionalization of hydrometeorological systems in the HRB, which has important significance in prediction in ungaged areas (Lebecherel et al., 2016). It should be noted that more studies are still required to verify the present results using other nonlinear techniques, such as the Lyapunov exponent (Wolf et al., 1985) and the approximate entropy (Pincus, 1995), which might provide additional insights into climate complexity analysis.

Add the following references on page 14, lines 373-374, 400-401, and page 15, lines 437-438:

Lebecherel, L., Andreassian, V., and Charles, P.: On evaluating the robustness of spatialproximity-based regionalization methods, J. Hydrol., 539, 196-203, https://doi.org/10.1016/j.jhydrol.2016.05.031, 2016.

Pincus, S.: Approximate entropy (ApEn) as a complexity measure, Chaos, 1995, 5(1), https://doi.org/10.1063/1.166092, 110.

Wolf, A., Swift, J. B., and Swinney, H. L.: Determining Lyapunov exponents from a time series, Physica D Nonlinear Phenomena, 1985, 16(3), https://doi.org/10.1016/0167-2789(85)90011-9, 285-317.

An improved Grassberger-Procaccia algorithm for analysis of climate system complexity

Chongli Di[1], Tiejun Wang[1], Xiaohua Yang[2], Siliang Li[1]

[1] Institute of Surface-Earth System Science, Tianjin University, Tianjin, 300072, P. R. China

[2] State Key Laboratory of Water Environment Simulation, School of Environment, Beijing Normal University, Beijing, 100875, P. R. China

Correspondence to: Tiejun Wang (tiejun.wang@tju.edu.cn)

Abstract. Understanding the complexity of natural systems, such as climate systems, is critical for various research and application purposes. A range of techniques have been developed to quantify system complexity, among which Grassberger-Procaccia (G-P) algorithm has been mostly used. However, the use of this method is still not adaptive and the choice of scaling regions relies heavily on subjective criteria. To this end, an improved G-P algorithm was proposed, which integrated the normal-based *K*-means clustering technique and Random Sample Consensus algorithm (RANSAC) for computing correlation dimensions. To test its effectiveness for computing correlation dimensions, the proposed algorithm was compared with traditional methods using the classical Lorenz and Henon chaotic systems. The results revealed that the new method outperformed traditional algorithms in computing correlation dimensions for both chaotic systems, demonstrating the improvement made by the new method. Based on the new algorithm, the complexity of precipitation and air temperature in the Haihe River basin (HRB) in northeast China was further evaluated. The results showed that there existed considerable regional differences in the complexity of both climatic variables across the HRB. Specifically, precipitation was shown to become progressively more complex from the mountainous area in the northwest to the plain area in the southeast; whereas, the complexity of air temperature exhibited an opposite trend with less complexity in the plain area. Overall, the spatial patterns of the complexity of precipitation and air temperature reflected the influence of the dominant climate system in the region.

1 Introduction

There are increasing interests in understanding system complexity, ranging from natural phenomena to social behaviors (Bras, 2015; Lin et al., 2015; Wang et al., 2016). As an open system with random external forcings and nonlinear dissipation, climate systems are highly complex (Nicolis and Nicolis, 1984; Jayawardena and Lai, 1994; Rind, 1999; Wang et al., 2015). Owing to nonlinear interactions among atmosphere, hydrosphere and biosphere, climatic variables exhibit highly nonlinear and dynamic characteristics, which reflect the complexity of climate systems (Palmer, 1999; Rial et al., 2004; Sivakumar, 2005; Wu et al., 2010). It is thus imperative to quantitatively measure the complexity of climatic variables for understanding underlying processes. However, no common definition of system complexity exists in scientific communities, particularly from a mathematical perspective (Carbone et al., 2016). To resolve this issue, numerous concepts and methods, including chaos theory, wavelet analysis and dynamical analysis, have been proposed to describe the complexity of climate

systems (Lorenz, 1963; Di et al., 2014; Feldhoff et al., 2014; Sivakumar, 2017; Meseguer-Ruiz et al., 2017). For instance, the chaos theory has been extensively used to characterize the chaotic and nonlinear features of climate systems (Sivakumar, 2001). Overall, previous studies based on the chaos theory revealed that the time series of air temperature and precipitation is non-stationary with abundant information. The complexity of rainfall and temperature dynamics has been widely used to indicate the extent of the complexity of climate systems (Dhanya and Kumar, 2010; Gan et al., 2014).

One of the important parameters in the chaos theory is correlation dimension, which can be used to measure the complexity and chaotic properties of variables, including precipitation and streamflow (Sivakumar et al., 2002; Dhanya and Kumar, 2011; Kyoung et al., 2011; Lana et al., 2016). Conceptually, the correlation dimension of a variable indicates the number of primary controls of the variable and thus determines the degree of freedom of the underlying process (Sivakumar and Singh, 2012). Despite the wide applications in various scientific fields, the use of the correlation dimension method is still hindered by certain limitations. For instance, the dimension method proposed by Grassberger and Procaccia (1983b) (denoted as the G-P method hereafter) is commonly used in the fields of hydrology and atmospheric science, however, its calculation procedures are still problematic (Ji et al., 2011). Specifically, the G-P method utilizes phase space reconstruction (Packard et al., 1980) and the embedding theorem (Takens, 1981) to compute correlation dimensions, which requires selection of an appropriate scaling region. The scaling region is a domain, over which an object exhibits self-similarity across a range of scales. However, the G-P method relies on visual inspections for choosing scaling regions, which is subject to human errors (Sprott and Rowlands, 2001). To tackle this problem, alternative methods have been developed to improve the original G-P method (Maragos and Sun, 1993). For example, Jothiprakash and Fathima (2013) utilized empirical equations to calculate the upper limit of scaling regions. Ji et al. (2011) applied the clustering analysis technique to determine scaling regions. However, these exsiting methods for identifying scaling regions are still not adaptive and the choice of scaling regions relies ~~rely~~ heavily on subjective criteria, and (2) the use of the least squares method for fitting straight lines to determine correlation exponents can include outliers (Cantrell, 2008) and thus is not optimal.Therefore, studies are still warranted to seek more objective and adaptive algorithms for identifying scaling regions to obtain more accurate estimates of correlation dimensions.

The primary aims of this study were two-fold. First, a new algorithm was proposed to improve the original G-P method, which integrated the methods of normal estimation, *K*-means clustering (Lloyd, 1982) and Random Sample Consensus (RANSAC; Fischler and Bolles, 1981). The classical Lorenz and Henon chaotic systems were chosen to test the effectiveness of the proposed algorithm for estimating correlation dimensions. Afterwards, the newly developed algorithm was utilized to investigate the nonlinear characteristics of precipitation and air temperature across the Haihe River basin (HRB) in northeast China. The HRB has been facing serious water shortage issues due to climate change and increasing water demand. Although previous studies have investigated climate variability (e.g., precipitation, air temperature and evaporation) in the HRB from different perspectives (Bao et al., 2012; Sang et al., 2012; Chu et al., 2010a), to our best knowledge, there are still no attempts made to quantify the nonlinear characteristics of climatic variables, especially regarding their chaotic behaviors in the HRB, which is essential for understanding the nonlinearity of the climate system in the region. Furthermore, the HRB is a diverse hydroclimatic region with many sub-watersheds of varying geographical and hydroclimatic conditions, which makes the region ideal for understanding the climate system

complexity. ~~and chaotic behaviors of those climatic variables in the HRB, which is essential for understanding the complexity of the climate system in this region.~~ The rest of this paper is organized as follows: Section 2 describes the calculation procedures of the proposed algorithm, which is then tested using classical mathematical models in Section 3. Section 4 describes the data obtained from the HRB and presents the results and analysis. Conclusions are made in the last part of this paper.

2 Methodology

2.1 Algorithm for Computing Correlation Dimension

~~The dimension of the time series of a variable is indicative of the number of factors governing the underlying dynamical processes.~~ Correlation dimensions can be used to identify the complexity of dynamical systems with varying complexity degrees (e.g., low-dimensional vs. high-dimensional systems). A wealth of algorithms have been developed for computing correlation dimensions, among which the G-P algorithm has been mostly used and is also adopted in this study. The G-P algorithm uses the concept of phase space reconstruction (Packard et al., 1980) from a single-variable time series. Here, the method of delays (Takens, 1981) was employed for reconstructing phase space. Given a time series $\mathbf{X}_i$ (i=1, 2,…, N), a multi-dimensional phase space can be reconstructed as:

$$\mathbf{Y}_j = (\mathbf{X}_j, \mathbf{X}_{j+\tau}, \mathbf{X}_{j+2\tau}, \cdots, \mathbf{X}_{j+(m-1)\tau}), \quad (1)$$

where $j = 1, 2, \cdots, N-(m-1)\tau$, m is the dimension of $\mathbf{Y}_j$ called embedding dimension, τ is delay time, and $\mathbf{X}_j$ is the reconstructed phase space vector.

For the m-dimensional reconstructed phase space, the correlation function $C(r, m)$ is defined as:

$$C(r,m) = \lim_{N\to\infty} \frac{2}{N(N-1)} \sum_{i,j=1}^{N} H(r - \left\|\mathbf{Y}_i - \mathbf{Y}_j\right\|), \quad 1 \le i \le j \le N, \quad (2)$$

where $\left\|\mathbf{Y}_i - \mathbf{Y}_j\right\|$ is the Euclidean distance between the vectors $\mathbf{Y}_i$ and $\mathbf{Y}_j$. $H(x)$ is the Heaviside function with $H(x)$=1 for x>0 and $H(x)$=0 for x≤0, where $x = r - \left\|\mathbf{Y}_i - \mathbf{Y}_j\right\|$ and r is the vector norm (i.e., radius of a sphere) centered on $\mathbf{Y}_i$ or $\mathbf{Y}_j$. Set r_{min} and r_{max} as the minimum and maximum distances between points, respectively (Ji et al, 2011; Lai and Lerner, 1998). If $r \leqslant r_{min}$, none of the vector points falls within the volume element and $C(r, m)$=0. Otherwise, if $r \geqslant r_{max}$, all vector points fall within the volume element and $C(r, m)$=1. If there exists an attractor in the reconstructed system, $C(r, m)$ and r are related through the following relationship:

$$C(r,m) \underset{\substack{r\to 0 \\ N\to\infty}}{\approx} \alpha r^{D_2(m)}, \quad (3)$$

where α is a constant and $D_2(m)$ is the correlation exponent.

$D_2(m)$ is usually estimated using the least squares method by fitting a straight line through ln r vs. ln $C(r, m)$. According to the relationship between $D_2(m)$ and m, the saturation value of $D_2(m)$ is defined as the correlation dimension. ~~By varying the embedding dimension m, a relationship can be derived between $D_2(m)$ and m. The saturation value of $D_2(m)$ is defined as the correlation dimension.~~ If the saturation value is low (e.g., a low correlation dimension), the system is considered to exhibit low-dimensional deterministic dynamics (i.e., a chaotic system); otherwise, the system is a stochastic one. The range, over which the straight line is fitted

through ln r vs. ln C(r, m), is called the scaling region, where the slope is defined. Clearly, choosing an appropriate scaling region is critical for computing correlation dimensions. In previous studies, scaling regions are usually determined by visual inspections, and this will be prone to individual preferences and thus not objective. Therefore, an objective method with adaptive procedures for computing correlation dimensions is still desired.

2.2 Scaling Region Identification

To overcome the limitation of the original G-P algorithm for selecting scaling regions, we propose an adaptive identification algorithm of scaling regions, which utilizes the normal-based *K*-means clustering technique and the RANSAC algorithm. The use of the normal-based *K*-means clustering technique is to partition all normals of the scatter points into *K* clusters with high similarity and to remove the points that are outside of the range of the scaling region. The RANSAC algorithm was introduced to fit a straight line through the log-transformed points obtained by the normal-based *K*-means clustering technique, which had been shown to outperform the traditional least squares method for fitting straight lines (Kyoung, 2011; Ji et al., 2011). To illustrate the advantages of using the RANSAC algorithm for linear fitting, a hypothetical example is shown in Fig. 1, which compares the fitting results obtained from the RANSAC algorithm and the traditional least squares method. The input data are sampled from a line y=0.5 x, with added noises and outliers. Here, for the RANSAC algorithm, the inliers are the points used to fit the line; whereas, the outliers are removed from the line fitting. It can be seen from Fig. 1 that the fitting line (y=0.60 x-0.068; R^2=0.854) obtained from the least squares method is seriously affected by outliers and deviated from the original line y=0.5 x. By contrast, the RANSAC method is able to distinguish the inliers from outliers effectively and results in a satisfactory fitting line (y=0.49 x+0.007; R^2=0.990), demonstrating the advantage of using the RANSAC algorithm for linear fitting.

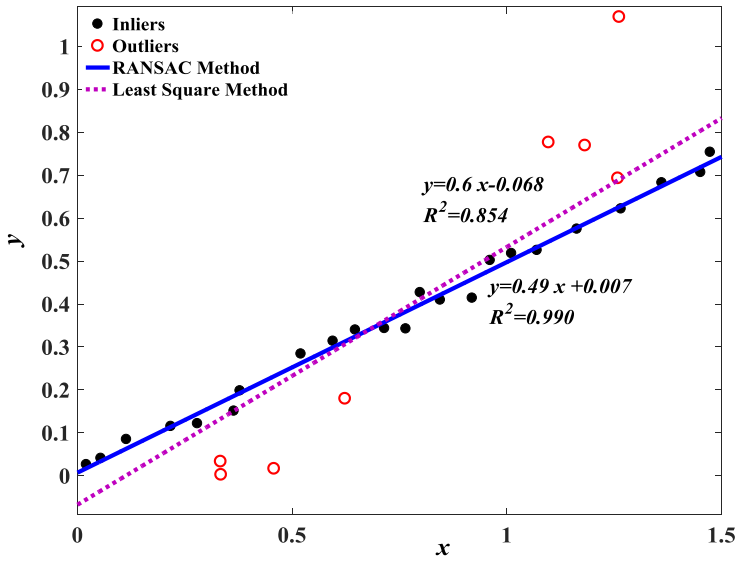

Figure 1: Comparison of the fitted lines obtained from the RANSAC algorithm and the least squares method.

The flow chart of the proposed procedures for calculating correlation dimensions is given in Fig. 2, which consists of five major steps. (1) For the time series $x(t)$, the time delay τ is computed by an autocorrelation function (Liebert and Schuster, 1989). Then set the minimum embedding dimension m_{min}=2 and reconstruct the

phase space by increasing m to obtain the correlation exponent function $C(r, m)$; (2) The normals of the scatter points on the ln r~ln $C(r, m)$ line are estimated via principal component analysis (Mitra et al., 2004); (3) The K-means clustering technique is performed on the normal set N with K=2 to obtain two different clusters. Set a threshold value T to determine the angle α between the two clusters. If $\alpha > T$, the data set with larger differences in normals is discarded. Then, the K-means clustering technique is repeated on the remaining data set until $\alpha \leq T$; (4) The RANSAC algorithm is used to fit a straight line through the set of remaining scatter points; and (5) The slope of the line obtained from the RANSAC method is computed to acquire the correlation dimension $D_2(m)$ for each m. Finally, the saturation correlation dimension is determined using the plot $D_2(m)$ vs. m. ~~and the details are listed as follows:~~

~~**Step 1: Computation of Correlation Exponent Function**. For the time series $x(t)$, the time delay τ is computed by an autocorrelation function (Liebert and Schuster, 1989). By setting the minimum embedding dimension m_{min}=2 and reconstructing the phase space by increasing m, the correlation exponent function $C(r, m)$ is then obtained. The correlation dimension is determined when saturation occurs for $C(r, m)$;~~

~~**Step 2: Normal Estimation**. To obtain the set of normals N, the normals of the scatter points on the ln r~ln $C(r, m)$ line are estimated via principal component analysis (Mitra et al., 2004);~~

~~**Step 3: Determination of the Scaling Region**. The K-means clustering technique is performed on the normal set N with K=2 to obtain two different clusters. Set a threshold value T to determine the angle α between the two clusters. If $\alpha > T$, the data set with larger differences in normals is discarded. Then, the K-means clustering technique is performed again on the remaining data set until $\alpha \leq T$;~~

~~**Step 4: Straight Line Fitting**. The RANSAC algorithm is used to fit a straight line through the set of remaining scatter points obtained by the K-means clustering technique to determine the linear region;~~

~~**Step 5: Dynamic Characteristic Analysis**. The slope of the line obtained from the RANSAC method is computed to acquire the correlation dimension $D_2(m)$ for each m. Finally, the saturation correlation dimension is determined using the plot $D_2(m)$ vs. m.~~

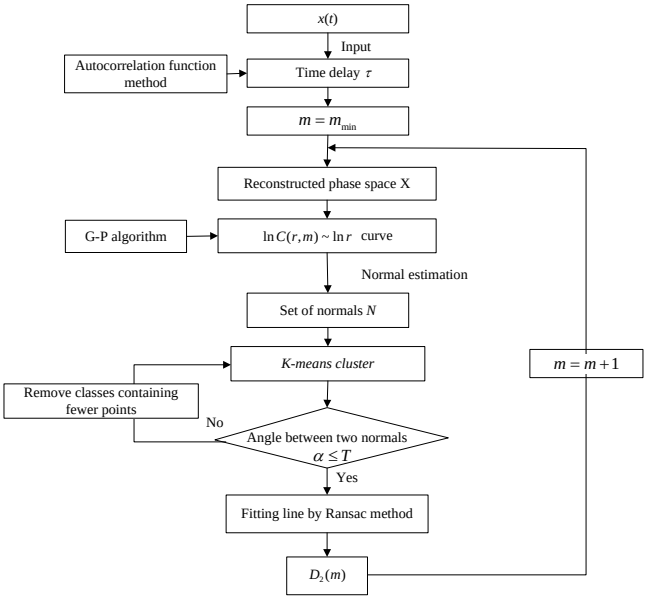

Figure 2: Flow chart of the proposed algorithm for computing correlation dimensions (The details are listed in the text).

3 Verification of the Proposed Algorithm

To test the effectiveness of the proposed algorithm, the classical chaotic models of Lorenz (1963) in Eq. (4) and Henon (1976) in Eq. (5) were used. The Lorenz and Henon systems with existing theoretical correlation dimensions have been mostly studied in the past, and thus widely used to analyze the chaotic behavior in climate systems and to test the effectiveness of algorithms for computing climate system complexity (e.g., Grassberger and Procaccia, 1983a; Lai and Lerner, 1998; Ji et al., 2011).

$$\dot{x} = \sigma(-x + y),\ \dot{y} = -xz + rx - y,\ \dot{z} = xy - bz\ , \tag{4}$$

$$x_{n+1} = y_n + 1 - ax_n^2,\ y_{n+1} = bx_n\ , \tag{5}$$

where σ=10, b=28, r=8/3, in Eq. (4), and a=1.4, b=0.3 in Eq. (5). The theoretical dimensions of the Lorenz and the Henon systems are 2.05±0.01 and 1.25±0.02, respectively (Grassberger and Procaccia, 1983a). As a comparison, the results obtained by our proposed method were compared with the theoretical dimensions and the values obtained by another two commonly used algorithms, including the intuitive judgment method (IJM) and the point-based *K*-means clustering method (PKC).

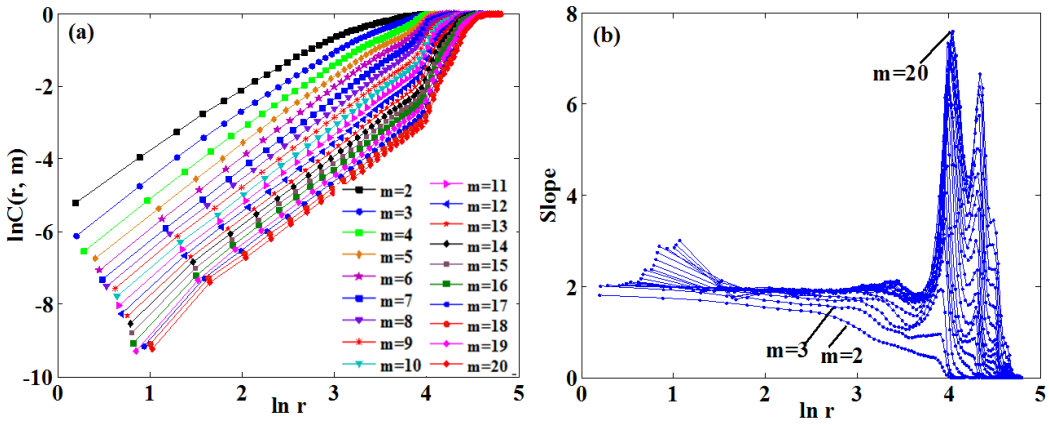

Figure 3: Correlation integral as a function of *r* with embedding dimension *m* ranging from 2 to 20 for the Lorenz attractor: (a) ln *C*(*r*, *m*) versus ln *r*, (b) the slopes of ln *C*(*r*, *m*) versus ln *r*.

According to the autocorrelation function, the time delay τ was determined to be 10 for the Lorenz system, with *m* varying from 2 to 20. Figure 3a shows the relationship between ln *C*(*r*, *m*) and ln *r* with *m* ranging from 2 to 20. Figure 3b shows the slopes of ln *C*(*r*, *m*) against ln *r* by increasing the embedding dimension *m* (i.e., the bottom curves are associated with smaller *m* values in Fig. 3b). The threshold value *T* was set as 5° and *K* was set as 2. The scaling regions of the curves in Fig. 3a were determined using the normal-based *K*-means clustering technique. As an example, an arbitrary curve was first selected from Fig. 3a, and the results are presented in Fig. 4. It can be seen from Fig. 4 that the process of the proposed method for determining the scaling region is adaptive. Specifically, for the selected curve shown in Fig. 4a, the normals of the curve were first computed based on Step 2 and the results are plotted in Fig. 4b. Different from previous *K*-means methods (e.g., the point-based *K*-means clustering method), we measured the similarity of points using the diversity between normals of different points. The reason for using the normal-based method is that the directions of normals for different points may vary considerably (see Fig. 4b); whereas, for the point-based *K*-means method, the distance between

different points might be small, making it difficult to separate the points into different clusters (Fig. 3a). The obtained two separate clusters of the normals (in red and blue) are shown in Fig. 4c. If the angle α between the two clusters was larger than T, the one with larger differences in normals was discarded. Then, the K-means clustering technique was performed again on the remaining data set. This process was usually repeated for 2-3 times until $\alpha \leq T$ (e.g., Figs. 4c to 4e). The final scaling region was determined as shown in Fig. 4f.

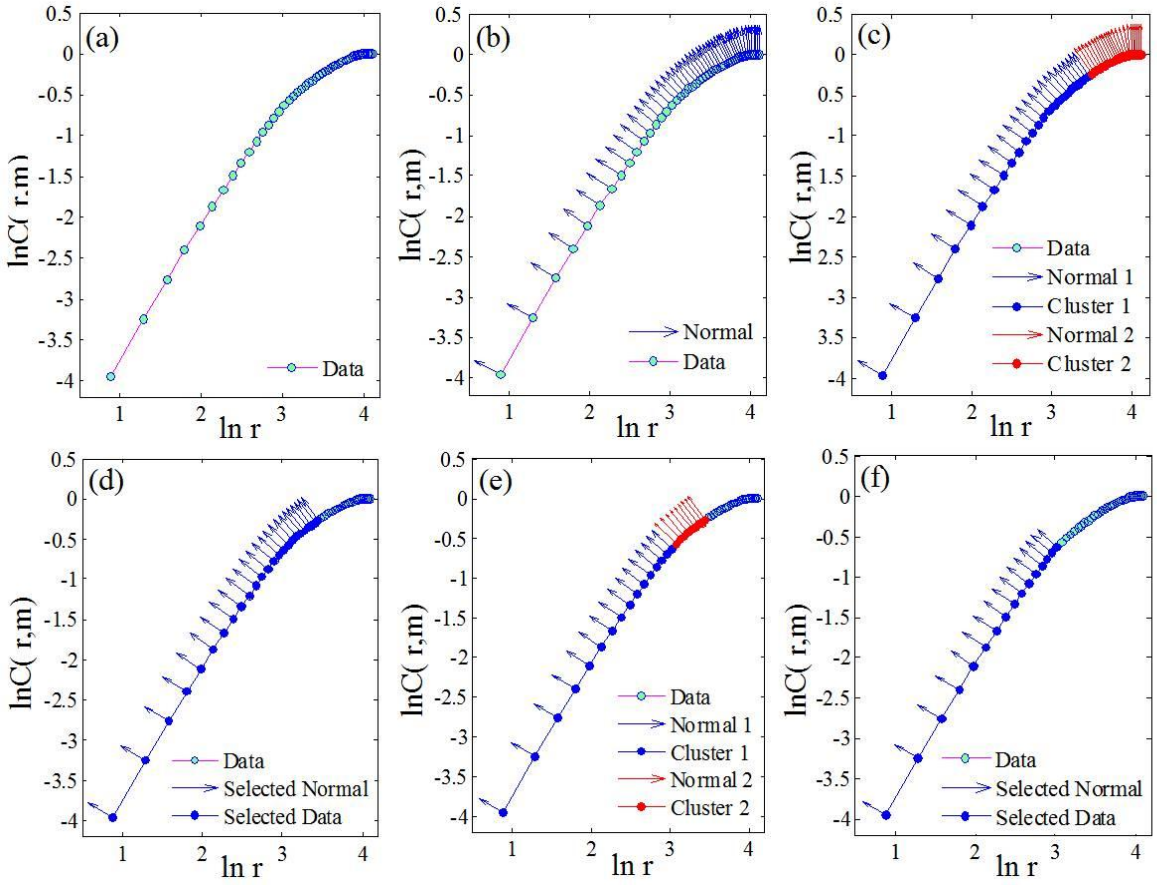

Figure 4: Illustration of using the Normal-based K-means clustering technique for determining the scaling region. The curve shown here was randomly selected from Fig. 3.

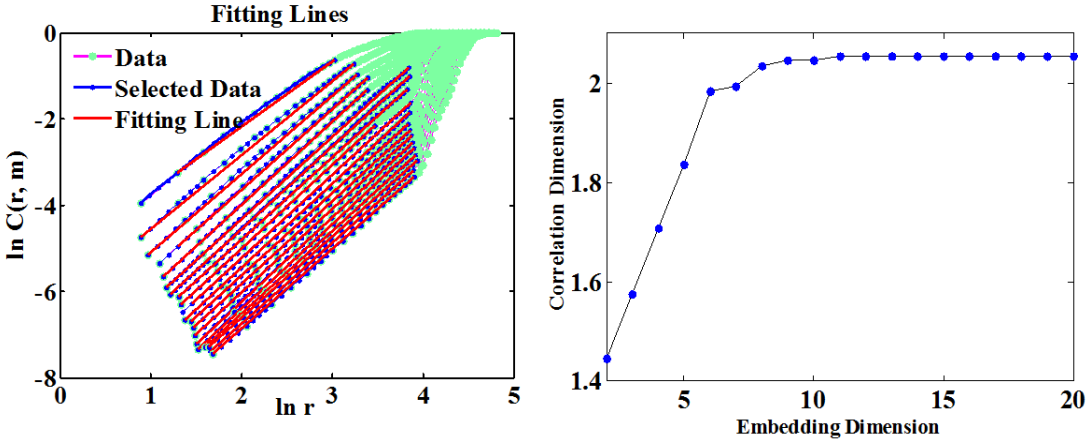

Figure 5: The final fitted lines and the correlation dimension of the Lorenz system: (a) the final fitted lines through the scaling regions and (b) correlation dimensions as a function of embedding dimension.

Figure 5a shows the final fitted lines through the scaling regions using the RANSAC method. The slope of the fitted line is the correlation dimension for each corresponding m. Figure 5b presents the graph of $D_2(m)$ against m with the value of m varying from 2 to 20. From Fig. 5b, we can see that $D_2(m)$ was saturated when $m>5$ with the saturation value approximately equal to 2.054, which was comparable to the theoretical value of the correlation dimension for the Lorenz attractor (i.e., 2.05±0.01). Following the same procedures, the obtained correlation dimension for the Henon attractor was 1.243, which was also close to its theoretical value (i.e., 1.25±0.02).

To verify the accuracy of our algorithm for computing correlation dimensions, the results derived from the proposed algorithm were compared with the ones obtained from the IJM and PKC methods. The IJM method was based on visual inspections to determine scaling regions (Jothiprakash and Fathima, 2013), while the PKC method integrated the *K*-means algorithm and the point-slope-error technique to determine scaling regions (Ji et al., 2011). The obtained correlation dimensions are reported in Table 1. For the Lorenz system, the differences in the correlation dimensions between the theoretical value and the ones obtained from IJM and PKC were 0.18±0.01 and 0.014±0.01, respectively; whereas, the difference was much smaller for the newly proposed algorithm (i.e. 0.004±0.01). Similar conclusions can be also made for the Henon system, demonstrating the improved performance of the proposed algorithm for determining correlation dimensions. It should be stressed that despite the improvement made by our proposed algorithm, further studies are still needed to address the issues on the computation of correlation dimensions. For example, estimation of correlation dimensions is partly dependent on the proper selection of time delay and embedding dimension; therefore, the impacts of their uncertainties should be further assessed.

Table 1. Comparison of the correlation dimensions derived from different methods.

Attractor	TCD	IJM	PKC	NPA
Lorenz	2.05±0.01	2.23±0.02	2.064	2.054
Henon	1.25±0.02	1.354±0.02	1.240	1.243

Note-TCD: Theoretical correlation dimension; IJM: Intuitive judgment method; PKC: Point-based K-means clustering; NPA: Newly proposed algorithm.

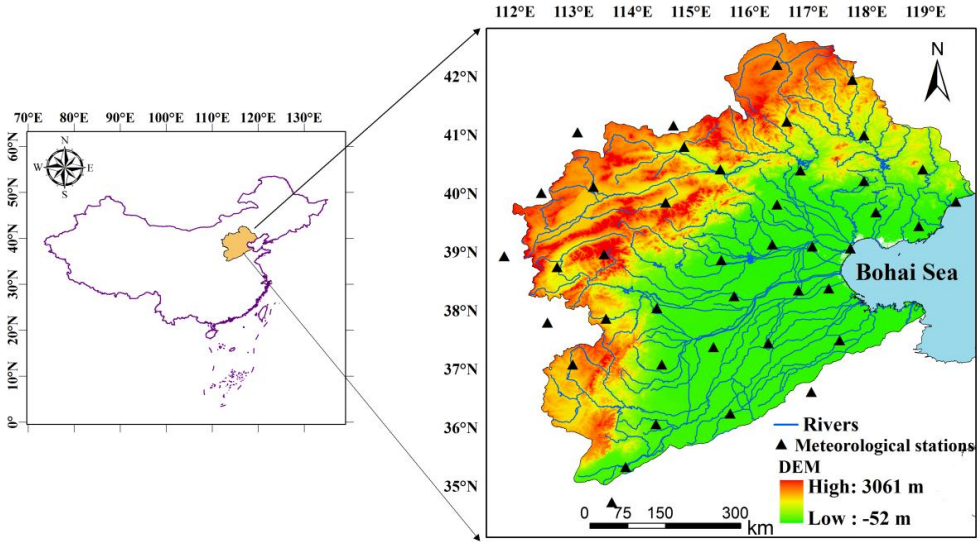

Figure 6: Locations of meteorological stations in the Haihe River basin.

4 Application, Results and Analysis

The correlation dimension method is an important diagnostic tool for understanding the complexity of natural systems with chaotic characteristics. In this section, a case study is presented to illustrate the use of the newly developed algorithm for studying the complexity of climate systems. Specifically, the algorithm was first utilized to compute the correlation dimensions of precipitation and air temperature using time series obtained from the HRB. Afterwards, the regional patterns of correlation dimensions for precipitation and air temperature in the HRB were analyzed.

4.1 Study area and Data

The HRB is located in northeast China (112°-120°E, 35°-43°N; Fig. 6), which hosts one of the most important economic zones in China (White et al., 2015). Under the influences of climate change and human activities, complex water issues have become increasingly prominent in the HRB (Liu and Xia, 2004). Topography varies considerably across the area, with 22% of the total area for mountains in the western and northern parts, 40% for plains in the eastern and southern parts, and 38% for hilly areas in the central part. The regional climate in the HRB is of a semiarid or subhumid type, with mean annual precipitation of 539.0 mm/year and mean annual temperature of 10.2˚C. Mean annual precipitation increases from the mountainous areas in the west to the plains in the east, while mean annual temperature decreases along the direction from south to north. In addition, precipitation in the HRB exhibits significant interdecadal and interannual variations. To apply the proposed algorithm for computing correlation dimensions, monthly precipitation and air temperature data spanning from 1951 to 2016 were retrieved from 40 meteorological stations in the HRB and nearby areas (Fig. 6), which were operated by the China Meteorological Administration (http://data.cma.cn/site/index.html).

4.2 Results and Analysis

The correlation dimensions of precipitation and air temperature at all 40 meteorological stations were computed using the algorithm proposed in this study. Figure 7 shows the relationships between correlation dimension and embedding dimension for precipitation and air temperature at five representative stations across the HRB (i.e., Beijing, Fengning, Shijiazhuang, Xinxiang and Zhangbei). The embedding dimensions of precipitation and air temperature for the five stations varied between 10 and 12. It is evident that the relationship between correlation dimension and embedding dimension for precipitation and air temperature differed among the selected stations. In genearal, correlation dimensions for precipitation showed gradual saturation processes with respective saturation values of 2.378, 2.407, 3.055 and 2.550 for Beijing, Fengning, Shijiazhuang and Zhangbei stations (Fig. 7a), indicating chaotic dynamical characteristics of precipitation. By comparison, the correlation dimension for precipitation at the Xinxiang station increased with increasing embedding dimensions, suggesting random characteristics of precipitation. For air temperature, the correlation dimensions at the five stations also showed gradual saturation processes (Fig. 7b), suggesting low dimensional chaotic characteristics for air temperature.

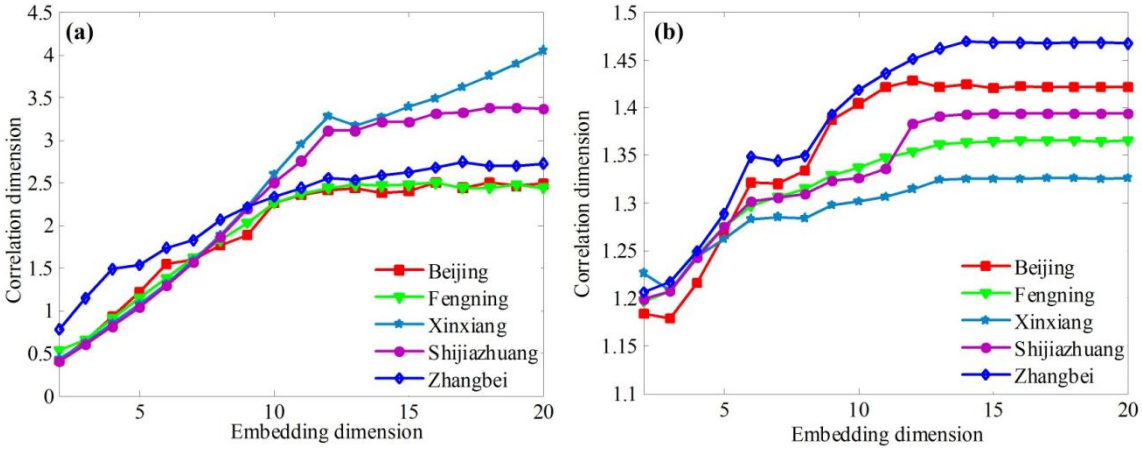

Figure 7: Variation of correlation dimension versus embedding dimension of climate variables: (a) precipitation and (b) air temperature.

Figure 8 presents the spatial distributions of the saturated correlation dimensions at the 40 meteorological stations for precipitation and air temperature in the HRB. For both precipitation and air temperature, the correlation dimensions varied markedly across the area. The correlation dimension for precipitation ranged from less than 3 to more than 6, while the correlation dimension was much lower for temperature (i.e., less than 2). Overall, the ranges of the correlation dimensions for precipitation and air temperature were comparable to previously reported values in other regions with similar climatic conditions (Kyoung et al., 2011; Sivakumar and Singh, 2012; Sivakumar et al., 2014). More importantly, the considerable spatial variations in the dimensionality for both climatic variables suggest the regional differences in the complexity of the climate system in the HRB. Specifically, the correlation dimension for precipitation tended to be smaller in the northwestern mountainous area with the values less than 2.5. In the central area, the correlation dimension for precipitation became larger with the values greater than 3, while precipitation in the southeastern plain area showed very high correlation dimensions with the values larger than 6. Given that correlation dimensions indicate the number of controls on the underlying process (Sivakumar and Singh, 2012), Fig. 8a suggests that precipitation processes become progressively more complex from the mountainous area to the plain area in the HRB. Interestingly, the regional pattern of the correlation dimension for air temperature showed an opposite trend with smaller values mainly located in the northern HRB, indicating more complex temporal dynamics of air temperature in the area.

The spatial pattern of the correlation dimension for precipitation in the HRB may be largely attributed to the regional flow pathway of moisture flux, which is mainly controlled by the East Asian Summer Monsoon (EASM). The HRB is located in a monsoon-dominated region, where the EASM plays a leading role in the regional meteorological system. Chen et al. (2013) showed that EASM had significant impacts on the spatiotemporal distribution of precipitation in East China. Li et al. (2017) further suggested that there was a significant correlation between precipitation and the EASM index in the HRB. Wang et al. (2011) revealed that large-scale atmospheric circulations had close relationships with precipitation patterns in the HRB by analyzing the moisture flux derived from NCAR/NCEP reanalysis data. Influenced by the large-scale atmospheric circulation, precipitation in the middle and southeast parts of the HRB is more sensitive to climate variability due to their locations closer to the ocean. This leads to the decreasing trend of precipitation from southeast to

northwest in the HRB, suggesting that the supply of moisture for precipitation in the region mainly comes from the ocean.

Partly owing to the closer geographical proximity to the ocean (Fig. 8), the EASM has a stronger impact on precipitation in the southern and central areas than in the northern part of the HRB. Furthermore, at the north corner of the HRB, the westerlies primarily affect the hydrometeorological system and thus weaken the impact of the EASM on precipitation (Li et al., 2017). In addition, other factors (e.g., topography, vegetation distribution, and human activity) may also have impacts on regional patterns of climate variables. In particular, the Yan-Taihang mountain located in the northwest HRB obstructs the vapor transport driven by the EASM, resulting in lower spatiotemporal variability in precipitation in the north part of the HRB. As a result, precipitation had higher degrees of complexity in the southern HRB, while its complexity was lower in the mountainous area in the northwest HRB. ~~, which probably led to the lower complexity of precipitation in the mountainous area in the northwest.~~ As to air temperature, the orographic effect in the mountainous area on air temperature might be stronger (Chu et al., 2010b), resulting in the higher complexity of temperature in this area. However, it should be noted that the range of the correlation dimension for air temperature from 1.0 to 2.0 suggests that two primary controls on temperature exist at all stations across the region.

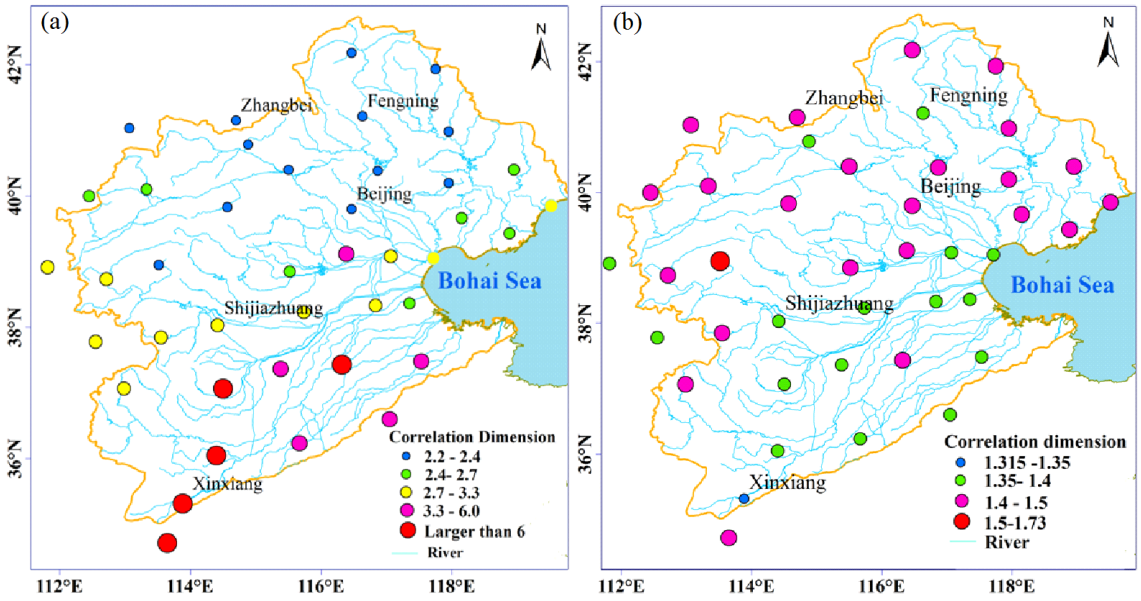

Figure 8: The spatial distribution of the correlation dimension values for all the 40 stations: (a) precipitation and (b) temperature.

5 Conclusions

In this study, the original G-P algorithm for calculating correlation dimensions was modified by incorporating the normal-based *K*-means clustering technique and the RANSAC algorithm. Using the proposed method, the spatial patterns of the complexity of precipitation and air temperature in the HRB were analyzed. The following conclusions were reached:

(1) The effectiveness of the proposed method for calculating correlation dimensions was illustrated using the classical Lorenz and Henon chaotic systems. The results showed that the new method outperformed the traditional intuitive judgment and point-based *K*-means clustering method for computing correlation dimensions.

(2) Except for few stations in the northern region, precipitation at most of the meteorological stations in the HRB showed chaotic behaviors. Specifically, the correlation dimension for precipitation showed an increasing trend from the mountainous region in the northwest to the plain area in the southeast, indicating that precipitation processes became progressively more complex from the mountainous area to the plain area. The spatial pattern of the complexity of precipitation reflected the influence of the dominant climate system in the region. Meanwhile, air temperature at all meteorological stations showed chaotic characteristics. In contrast to precipitation, the complexity of air temperature exhibited an opposite trend with less complexity in the plain area.

The modified G-P algorithm proposed in this study can be used more objectively to characterize the complexity of climate systems (and other hydrological systems, such as streamflow, soil moisture, and groundwater), and thus provide a more reliable estimate of the number of dominant factors governing climate systems. Theoretically, it can provide valuable information for optimizing the number of parameters in climate models to reduce computational demands and model parameter uncertainties. Furthermore, the findings of this study can be used for the regionalization of hydrometeorological systems in the HRB, which has important significance in prediction in ungaged areas (Lebecherel et al., 2016). It should be noted that more studies are still required to verify the present results using other nonlinear techniques, such as the Lyapunov exponent (Wolf et al., 1985) and the approximate entropy (Pincus, 1995), which might provide additional insights into climate complexity analysis.

Data Availability

The datasets used in this study are publicly available. The monthly precipitation and temperature data can be downloaded from the China Meteorological Administration Network (http://data.cma.cn/site/index.html). The code for computing correlation dimension can be acquired from the first author C. Di.

Competing interests

The authors declare that they have no conflict of interest.

Acknowledgements

The work was supported by the National Key R&D Program of China (No. 2016YFA0601002, No.2016YFC0401305, and No. 2017YFC0506603), and by the National Natural Scientific Foundation of China (No. 51679007 and No. U1612441). The authors would also like to acknowledge the financial support from the Tianjin University and T. Wang also acknowledges the financial support from the Thousand Talent Program for Young Outstanding Scientists for this study.

References

Bao, Z. X., Zhang, J. Y., Wang, G. Q., Fu, G. B, He, R. M., Yan, X. L., Jin, J. L., Liu, Y. L., and Zhang, A. J.: Attribution for decreasing streamflow of the Haihe River basin, northern China: Climate variability or human activities? J. Hydrol., 460, 117-129, https: //doi.org/10.1016/j.jhydrol.2012.06.054, 2012.

Bras, R. L.: Complexity and organization in hydrology: A personal view, Water Resour. Res., 51, 6532-6548, https: //doi.org/10.1002/2015WR016958, 2015.

Cantrell, C. A.: Technical Note: Review of methods for linear least squares fitting of data and application to atmospheric chemistry problems, Atmos. Chem. Phys., 8, 5477-5487, https://doi.org/10.5194/acp-8-5477-2008, 2008.

Carbone, A., Jensen, M., and Sato, A. H.: Challenges in data science: a complex systems perspective, Chaos Soliton. Fract., 90, 1-7, https://doi.org/10.1016/j.chaos.2016.04.020, 2016.

Chu, J. T., Xia, J., Xu, C. Y., and Singh, V. P.: Statistical downscaling of daily mean temperature, pan evaporation and precipitation for climate change scenarios in Haihe River, China. Theor. Appl. Climatol., 99, 149-161, https://doi.org/10.1007/s00704-009-0129-6, 2010b.

Chu, J. T., Xia, J., Xu, C. Y., Li, L., and Wang, Z. G.: Spatial and temporal variability of daily precipitation in Haihe River basin, 1958-2007. J. Geogr. Sci., 20, 248-260, https://doi.org/10.1007/s11442-010-0248-0, 2010a.

Dhanya, C.T., and Kumar, D.N.: Multivariate nonlinear ensemble prediction of daily chaotic rainfall with climate inputs. J. Hydrol., 2011, 403, 292-306, https://doi.org/10.1016/j.jhydrol.2011.04.009, 2011.

Dhanya, C.T., and Kumar, D.N.: Nonlinear ensemble prediction of chaotic daily rainfall. Adv. Water Resour., 33, 327-347, https://doi.org/10.1016/j.advwatres.2010.01.001, 2010.

Di, C. L., Yang, X. H., and Wang, X. C.: A four-stage hybrid model for hydrological time series forecasting, Plos One, 9, e104663, https://doi.org/10.1371/journal.pone.0104663, 2014.

Feldhoff, J. H., Lange, S., Volkholz, J., Donges, J. F., Kurths, J., and Gerstengarbe, F.: Complex networks for climate model evaluation with application to statistical versus dynamical modeling of South American climate, Clim. Dynam., 44, 1567-1581, https://doi.org/10.1007/s00382-014-2182-9, 2015.

Fischler, M. A., and Bolles, R. C.: Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. Communications of the Acm, 24, 381-395, 1981.

Gan, T. Y., Wang, Q., and Seneka, M.: Correlation dimensions of climate sub-systems and their geographic variability, Advances in Hydraulics and Water Engineering, 494-498,https://doi.org/10.1029/2001JD001268, 2014.

Grassberger, P., and Procaccia, I.: Characterization of strange attractors, Phys. Rev. Lett., 50, 346-349, https://doi.org/10.1103/PhysRevLett.50.346, 1983a.

Grassberger, P., and Procaccia, I.: Measuring the strangeness of strange attractors, Physica D, 9, 189-208, https://doi.org/10.1016/0167-2789(83)90298-1, 1983b.

Henon, M.: A two-dimensional mapping with a strange attractor, Comm. Math. Phys., 50, 69-77, https://doi.org/10.1007/BF01608556, 1976.

Jayawardena, A. W., and Lai, F.: Analysis and prediction of chaos in rainfall and stream flow time series, J. Hydrol., 153, 23-52, https://doi.org/10.1016/0022-1694(94)90185-6, 1994.

Ji, C. C., Zhu, H., and Jiang, W.: A novel method to identify the scaling region for chaotic time series correlation

dimension calculation, Chinese Sci. Bull., 56, 925-932, https://doi.org/10.1007/s11434-010-4180-6, 2011.

Jothiprakash, V., and Fathima, T. A.: Chaotic analysis of daily rainfall series in Koyna reservoir catchment area, India, Stoch. Env. Res. Risk a., 27, 1371-1381, https://doi.org/10.1007/s00477-012-0673-y, 2013.

Kyoung, M. S., Kim, H. S., Sivakumar, B., Singh, V. P., and Ahn, K. S.: Dynamic characteristics of monthly rainfall in the Korean Peninsula under climate change, Stoch. Env. Res. Risk a., 25, 613-625, https://doi.org/10.1007/s00477-010-0425-9, 2011.

Lai, Y. C., and Lerner, D.: Effective scaling regime for computing the correlation dimension from chaotic time series, Physica D, 115, 1-18, https://doi.org/10.1016/S0167-2789(97)00230-3, 1998.

Lana, X., Burgueno, A., Martinez, M. D., and Serra, C.: Complexity and predictability of the monthly Western Mediterranean Oscillation index, Int. J. Climatol., 36, 2435-2450, https://doi.org/10.1002/joc.4503, 2016.

Lebecherel, L., Andreassian, V., and Charles, P.: On evaluating the robustness of spatialproximity-based regionalization methods, J. Hydrol., 539, 196-203, https://doi.org/10.1016/j.jhydrol.2016.05.031, 2016.

Li, F. X., Zhang, S. Y., Chen, D., He, L., and Gu, L. L.: Inter-decadal variability of the east Asian summer monsoon and its impact on hydrologic variables in the Haihe River Basin, China, J. Resour. Ecol., 8(2), 174-184, https://doi.org/10.5814/j.issn.1674-764X.2017.02.008, 2017.

Liebert, W., and Schuster, H. G.: Proper choice of the time delay for the analysis of chaotic time series, Physics Letters A, 142, 107-111, https://doi.org/10.1016/0375-9601(89)90169-2, 1989.

Lin, H., Vogel, H., Phillips, J., and Fath, B. D.: Complexity of soils and hydrology in ecosystems, Ecol. Model., 298, 1-3, https://doi.org/10.1016/j.ecolmodel.2014.11.016, 2015.

Liu, C., and Xia, J.: Water problems and hydrological research in the Yellow River and the Huai and Hai River basins of China, Hydrol. Process., 18, 2197-2210, https://doi.org/10.1002/hyp.5524, 2004.

Lloyd, S. P.: Least squares quantization in PCM, 28, 129-137, https://doi.org/10.1109/TIT.1982.1056489, 1982.

Lorenz, E. N.: Deterministic nonperiodic flow, Journal of the Atmospheric Sciences, 20, 130-141, https://doi.org/10.1175/1520-0469(1963)020<0448:TMOV<2.0.CO;2, 1963.

Maragos, P., and Sun, F.: Measuring the fractal dimension of signals: Morphological covers and iterative optimization, 41, 108, https://doi.org/10.1109/TSP.1993.193131, 1993.

Meseguer-Ruiz, O., Olcina Cantos, J., Sarricolea, P., and Martin-Vide, J.: The temporal fractality of precipitation in mainland Spain and the Balearic Islands and its relation to other precipitation variability indices, Int. J. Climatol., 37, 849-860, https://doi.org/10.1002/joc.4744, 2017.

Mitra, N. J., Nguyen A. N., and Guibas, L.: Estimating surface normals in noisy point cloud data, J. Comput. Geom. Appl., 14, 261-276, https://doi.org/10.1145/777792.777840, 2004.

Nicolis, C., and Nicolis, G.: Is there a climatic attractor? Nature, 311, 529-532, https://doi.org/10.1038/311529a0, 1984.

Packard, N. H., Crutchfield, J. P., Farmer, J. D., and Shaw, R. S.: Geometry from a time series, 45, 712-716, https://doi.org/10.1103/PhysRevLett.45.712, 1980.

Palmer, T. N.: A Nonlinear Dynamical Perspective on Climate Prediction., Journal of Climate, 12, 575-591, https://doi.org/10.1175/1520-0442(1999)012, 1999.

Pincus, S.: Approximate entropy (ApEn) as a complexity measure, Chaos, 1995, 5(1), https://doi.org/10.1063/1.166092 110.

Rial, J. A., Pielke, R. A., Beniston, M., Claussen, M., Canadell, J., Cox, P., Held, H., De Noblet-Ducoudre, N.,

Prinn, R., Reynolds, J. F., and Salas, J. D.: Nonlinearities, feedbacks and critical thresholds within the Earth's climate system, Climate Change, 65, 11-38, https://doi.org/10.1023/B:CLIM.0000037493.89489.3f, 2004.

Rind, D.: Complexity and climate, Science, 284, 105-107, https://doi.org/10.1126/science.284.5411.105, 1999.

Sang, Y., Wang, Z., and Li, Z.: Discrete wavelet entropy aided detection of abrupt change: A case study in the haihe river basin, china, Entropy-Switz, 14, 1274-1284, https://doi.org/10.3390/e14071274, 2012.

Sivakumar, B., and Singh, V. P.: Hydrologic system complexity and nonlinear dynamic concepts for a catchment classification framework, Hydrol. Earth Syst. Sc., 16, 4119-4131, doi: 10.5194/hess-16-4119-2012, 2012.

Sivakumar, B., Persson, M., Berndtsson, R., and Uvo, C.B.: Is correlation dimension a reliable indicator of low-dimensional chaos in short hydrological time series? Water Resour. Res., 38, 3-1, https://doi.org/10.1029/2001WR000333, 2002.

Sivakumar, B., Woldemeskel, F. M., and Puente, C. E.: Nonlinear analysis of rainfall variability in Australia, Stoch. Env. Res. Risk a., 28, 17-27, https://doi.org/10.1007/s00477-013-0689-y, 2014.

Sivakumar, B.: Chaos in Hydrology: Bridging determinism and stochasticity. Sydney, Springer: Netherlands, https://doi.org/10.1007/978-90-481-2552-4, 2017.

Sivakumar, B.: Chaos in rainfall: variability, temporal scale and zeros, J. Hydroinform., 7, 175-184, https://doi.org/10.2166/hydro.2005.0015, 2005.

Sivakumar, B.: Rainfall dynamics at different temporal scales: A chaotic perspective, Hydrol. Earth Syst. Sc., 5, 645-651, https://doi.org/10.5194/hess-5-645-2001, 2001.

Sprott, J. C., and Rowlands, G.: Improved correlation dimension calculation, Int. J. Bifurcat. Chaos, 11, 1865-1880, https://doi.org/10.1142/S021812740100305X, 2001.

Takens, F.: Detecting strange attractors in turbulence. Dynamical Systems and Turbulence, Warwick 1980. Springer Berlin Heidelberg, 366-381, 1981.

Wang, W. G., Shao, Q. X., Peng, S. Z., Zhang, Z. X., Xing, W. Q., An, G. Y., and Yong, B.: Spatial and temporal characteristics of changes in precipitation during 1957-2007 in the Haihe River basin, China. Stoch. Environ. Res. Risk Assess., 25(7), 881-895, https://doi.org/10.1007/s00477-011-0469-5, 2011.

Wang, W., Lai, Y., and Grebogi, C.: Data based identification and prediction of nonlinear and complex dynamical systems, Physics Reports, 644, 1-76, https://doi.org/10.1016/j.physrep.2016.06.004, 2016.

Wang, W., Wei, J., Shao, Q., Xing, W., Yong, B., Yu, Z., and Jiao, X.: Spatial and temporal variations in hydro-climatic variables and runoff in response to climate change in the Luanhe River basin, China, Stoch. Env. Res. Risk a., 29, 1117-1133, https://doi.org/10.1007/s00477-014-1003-3, 2015.

White, D. J., Feng, K. S., Sun, L. X., and Hubacek, K.: A hydro-economic MRIO analysis of the Haihe River Basin's water footprint and water stress, Ecol. Model., 318, 157-167, https://doi.org/10.1016/j.ecolmodel.2015.01.017, 2015.

Wolf, A., Swift, J. B., and Swinney, H. L.: Determining Lyapunov exponents from a time series, Physica D, 16(3), 285-317, https://doi.org/10.1016/0167-2789(85)90011-9, 1985.

Wu, C. L., Chau, K. W., and Fan, C.: Prediction of rainfall time series using modular artificial neural networks coupled with data-preprocessing techniques, J. Hydrol., 389(1-2), 146-167, https://doi.org/10.1016/j.jhydrol.2010.05.040, 2010.