

Statistical inference on specifying regression models for detecting dependence in autocorrelated series

Feng Chen¹, Yu Zhou^{2,3,*} and Holger Kantz³

¹*School of Mathematics and Statistics, Chongqing Jiaotong University, Chongqing 400074, China*

²*National Center for Applied Mathematics in Hunan and School of Mathematics and Computational Science, Xiangtan University, Xiangtan 411105, Hunan, China*

³*Max Planck Institute for the Physics of Complex Systems, Nöthnitzer Straße 38, 01187 Dresden, Germany*



(Received 2 July 2025; accepted 9 February 2026; published 5 March 2026)

The traditional regression model for dependence detection has been generalized by including the autocorrelated noise (ACN) or lagged dependent variable (LDV) to account for autocorrelation in time series. Existing studies on specifying the models with ACN or LDV are rare. We propose a statistical analysis framework of model specification and estimation, whose effectiveness is validated by simulations. As real-life cases we study the precipitation extreme dependence on temperature and temperature dependence on CO₂ and demonstrate how the traditional and misspecified models may overestimate dependence by up to three times. We provide a solid statistical basis for detecting dependence in autocorrelated series.

DOI: [10.1103/dv2v-q8tg](https://doi.org/10.1103/dv2v-q8tg)

I. INTRODUCTION

How a variable y depends on a variable x is a research topic of importance, exemplified by the dependence of precipitation on temperature and the dependence of temperature on greenhouse gases like CO₂ [1]. A traditional model for detecting dependence using linear regression is in a form of $y = \beta_1 x + \beta_0 + u$, with a regression coefficient β_1 to quantify dependence, the intercept β_0 , and noise u [2]. In regression, y and x are usually called the dependent and explanatory variables, respectively. The traditional regression analysis based on this traditional model commonly uses ordinary least squares (OLS) for the parameter estimation. OLS assumes that the data sample is independent, which forbids autocorrelations in the time series. This means that the traditional model is driven by independent and identically distributed (i.i.d.) noise u . The ubiquitousness of autocorrelation in real-life series motivates generalizations of the traditional model. In this study, our discussion is on the first order autoregressive [AR(1)] process, which is thought to be an appropriate model to account for autocorrelation in many real-life series [1].

To tackle autocorrelation, one commonly used generalized model replaces the i.i.d. noise of the traditional model by an AR(1) process [3–6], named the autocorrelated noise (ACN) model in this study. Alternatively, inclusion of a lagged dependent variable (LDV) as one more explanatory variable is

capable of accounting for autocorrelation. It is referred to the LDV model in some research fields, such as politics and econometrics [7–10]. It is also common to consider LDV in many climate dynamical models [11–14]. These two models cannot be transformed into each other and therefore offer two different explanations for observed autocorrelations in the dependent variable y , and the determination of the dependence parameter β_1 via regression is done by different expressions and hence generally yields different values. Existing methodology studies mainly focus on estimation of these two models [3–5,7–10] and the consequences of incorrect inclusion or exclusion of LDV in or from regression [8–10], i.e., the model misspecification effect. Autocorrelation in the noise would increase the uncertainty in estimation, and inclusion of LDV usually leads to a traditional OLS-estimated regression coefficient smaller than those estimated by the model without LDV. Although the ACN and LDV models may result in very different estimates, to our best knowledge, studies on the statistical inference on specifying these two models for given series are rare.

To detect dependence in autocorrelated series, we propose here a feasible statistical framework, including model specification and parameter estimation. In spatial regression, the spatial error model and the spatial lag model [15–18] are two common models to offer different explanations for the observed spatial autocorrelations in data. Therefore, we need to specify these two models to determine an appropriate model according to the given data for implementation of spatial regression. Combination of the Lagrange multiplier (LM) tests and their robust scenarios is considered a standard technique to specify these two spatial regression models [17]. As shown below, the ACN and LDV models can be reexpressed as matrices, which are respectively identical to those of the spatial error model and the spatial lag model with a specific adjacency matrix. In this sense, they

*Contact author: zhouyu@pks.mpg.de

Published by the American Physical Society under the terms of the [Creative Commons Attribution 4.0 International](https://creativecommons.org/licenses/by/4.0/) license. Further distribution of this work must maintain attribution to the author(s) and the published article's title, journal citation, and DOI. Open access publication funded by Max Planck Society.

are special cases of the ACN and LDV models. Therefore, the LM tests and their robust scenarios are also applicable to specification of the ACN and LDV models. We use simulated series to examine their capability of specification and estimate dependence based on the specified model. Afterwards, we demonstrate their applicability in empirical analysis of two real-life cases, namely the dependence of precipitation extreme on temperature and the temperature dependence on CO₂. At the end, the dependence of precipitation extreme on temperature has been identified by our proposed framework as an ACN system, while the temperature dependence on CO₂ is more likely to be governed by an LDV system. Furthermore, these case studies highlight the effect of model misspecification, consistent with that shown in the corresponding simulation experiments.

II. METHODOLOGY

For a given pair of variables $\{x_t, y_t\}_{t=1}^n$ we can respectively express the ACN model as

$$y_t = \beta_1 x_t + \beta_0 + u_t, \quad (1)$$

with correlated $u_t = \lambda u_{t-1} + e_t$, and the LDV model as

$$y_t = \alpha y_{t-1} + \beta_1 x_t + \beta_0 + e_t. \quad (2)$$

In both models, e_t denotes white noise and the coefficient β_1 is an indicator to quantify dependence. The traditional model ignoring autocorrelations of y_t is obtained from Eq. (1) by replacing u_t by e_t , the independent noise.

Using the transpose notation “ $\top$ ” and matrix forms of $\tilde{X} = (\mathbf{1}_{n \times 1}, X)$ with $X = (x_1, x_2, \dots, x_n)^\top$ and a ones matrix $\mathbf{1}$, $Y = (y_1, y_2, \dots, y_n)^\top$, $\tilde{\beta} = (\beta_0, \beta_1)^\top$, $\tilde{u} = (u_1, u_2, \dots, u_n)^\top$, and $W = (\mathbf{0}_{1 \times (n-1)}, 0; I_W, \mathbf{0}_{(n-1) \times 1})$ with a $(n-1) \times (n-1)$ unit matrix I_W and two zeros matrices $\mathbf{0}$, the ACN and LDV models can be respectively reexpressed in matrix form as

$$Y = \tilde{X} \tilde{\beta} + \tilde{u}, \quad (3)$$

with $\tilde{u} = \lambda W \tilde{u} + \tilde{e}_I$, and

$$Y = \alpha W Y + \tilde{X} \tilde{\beta} + \tilde{e}_{II}. \quad (4)$$

Here $\tilde{e}_I$ and $\tilde{e}_{II}$ are independent and consist of white noises. It is noteworthy that the matrix forms of the ACN and LDV models are actually identical to those of the spatial error model and the spatial lag model [15–18], with a specific adjacency matrix W .

A. Lagrange multiplier tests and their robust scenarios

As aforementioned, existing studies mainly focus on estimation and the model misspecification effect but scarcely on specification of the ACN and LDV models. Estimates of β_1 obtained by the two models are usually obviously different. The model specification is therefore of paramount importance for appropriately quantifying dependence.

Essentially, specification of the ACN and LDV models is to test the significance of autocorrelation in noise of the ACN model and the significance of the lagged dependence of the LDV model. The same expression in matrix form implies the applicability of the LM tests and their robust scenarios for the spatial error and lag models to the ACN and LDV models.

For the ACN model, the null and alternative hypotheses are respectively set as

$$H_0 : \lambda = 0 \quad (5)$$

and

$$H_1 : \lambda \neq 0. \quad (6)$$

The corresponding LM test and its robust scenario are respectively based on the statistics LM_λ and LM_λ^* [16] in the following forms:

$$LM_\lambda = \frac{1}{D} \left(\frac{\tilde{e}_{OLS}^\top W Y}{\sigma^2} \right)^2 \sim \chi^2(1), \quad (7)$$

$$LM_\lambda^* = \frac{1}{D-L} \left(\frac{\tilde{e}_{OLS}^\top W Y - \tilde{e}_{OLS}^\top W \tilde{e}_{OLS}}{\sigma^2} \right)^2 \sim \chi^2(1), \quad (8)$$

where $D = (\tilde{\beta}^\top X W^\top H W X \tilde{\beta} + L \sigma^2) / \sigma^2$, $H = I - X(X^\top X)^{-1} X^\top$, $L = \text{tr}((W + W^\top)W)$, and $\sigma^2 = \tilde{e}_{OLS}^\top \tilde{e}_{OLS} / n$. Here, tr denotes the trace of a matrix. $\tilde{\beta}$ and $\tilde{e}_{OLS}$ are the OLS estimators in the null model $Y = X \tilde{\beta} + \tilde{e}$.

The LM test and its robust scenario for the LDV model use the hypothesis pair

$$H_0 : \alpha = 0 \quad (9)$$

and

$$H_1 : \alpha \neq 0. \quad (10)$$

The corresponding statistics LM_α and LM_α^* respectively are

$$LM_\alpha = \frac{n^2}{\sqrt{\text{tr}(L)}} \left(\frac{\tilde{e}_{OLS}^\top W \tilde{e}_{OLS}}{\tilde{e}_{OLS}^\top \tilde{e}_{OLS}} \right)^2 \sim \chi^2(1), \quad (11)$$

$$LM_\alpha^* = \frac{1}{L(1-LD^{-1})} \left(\frac{\tilde{e}_{OLS}^\top W \tilde{e}_{OLS} - LD^{-1} \tilde{e}_{OLS}^\top W Y}{\sigma^2} \right)^2 \sim \chi^2(1). \quad (12)$$

Based on these test statistics, Anselin [19] suggests the standard procedure of model specification in spatial regression as follows: If only LM_α (LM_λ) is significant, then we reject the null hypothesis and use the spatial lag (error) model. If both LM statistics are significant, then we use the spatial lag (error) model with the p value of the corresponding robust statistic LM_α^* (LM_λ^*) less than the given significance level. If both robust LM statistics are significant, then the more complicated model with both the lagged term and the autocorrelated error may be considered. However, this complicated model is usually not recommended, especially in applications, because of the high marginal costs for modeling [20]. In this study, only the ACN and LDV models are specified by using these testing statistics based on their correspondence to the spatial error model and the spatial lag model.

B. Estimation

Estimating the specified model serves for quantification of the dependence of y on x . For the ACN model, OLS is inapplicable and its estimates would be problematic, given the existence of auto-correlation in the noise u corresponding to $H_1 : \lambda \neq 0$ [6,8]. In this situation, the maximum likelihood estimation (MLE) is suggested to estimate the ACN model.

The log-likelihood function is

$$\begin{aligned} L(\vec{\beta}, \lambda, \sigma^2) &= -\frac{n}{2} \log(2\pi) - \frac{n}{2} \log(\sigma^2) + \log |I - \lambda W| \\ &\quad - \frac{1}{2\sigma^2} (Y - \tilde{X}\vec{\beta})^\top (I - \lambda W)^\top (I - \lambda W) (Y - \tilde{X}\vec{\beta}). \end{aligned}$$

Then, taking partial derivatives of $\vec{\beta}$ and σ^2 in the likelihood function $L(\vec{\beta}, \lambda, \sigma^2)$, respectively, and setting these partial derivatives to 0, we can obtain

$$\begin{aligned} \vec{\beta}_{ACN} &= (\tilde{X}^\top (I - \lambda W)^\top (I - \lambda W) \tilde{X})^{-1} \tilde{X}^\top \\ &\quad \times (I - \lambda W)^\top (I - \lambda W) Y \end{aligned}$$

and

$$\tilde{\sigma}^2 = \frac{1}{n} \tilde{e}^\top (I - \lambda W)^\top (I - \lambda W) \tilde{e}$$

with $\tilde{e} = Y - \tilde{X}\vec{\beta}_{ACN}$. Substituting $\vec{\beta}_{ACN}$ and $\tilde{\sigma}^2$ into the log-likelihood function, we have

$$\begin{aligned} L(\lambda) &= c - \frac{n}{2} \log \left(\frac{1}{n} \tilde{e}^\top (I - \lambda W)^\top (I - \lambda W) \tilde{e} \right) \\ &\quad + \log |I - \lambda W|, \end{aligned}$$

where c represents a quantity independent of λ . Maximizing $L(\lambda)$ by some optimization procedures such as the grid point and the golden section methods, we can derive an estimate of λ . Substituting the estimate of λ into the expressions of $\vec{\beta}_{ACN}$, the MLE estimate of $\vec{\beta}$ for the ACN model is obtained.

For the LDV model, the correlation between y_{t-1} and x_t may bias the OLS estimation for the dependence detection, even if there is no autocorrelation in the noise [9]. Referring to the similar case in spatial regression analysis, namely the spatial lag model [21,22], the MLE would be a more appropriate estimation for the LDV model. In this situation, the log-likelihood function can be expressed as

$$\begin{aligned} L(\vec{\beta}, \alpha, \sigma^2) &= -\frac{n}{2} \log(2\pi) - \frac{n}{2} \log(\sigma^2) + \log |I - \alpha W| \\ &\quad - \frac{1}{2\sigma^2} ((I - \alpha W)Y - \tilde{X}\vec{\beta})^\top ((I - \alpha W)Y - \tilde{X}\vec{\beta}). \end{aligned}$$

In addition, we have

$$\vec{\beta}_{LDV} = (\tilde{X}^\top \tilde{X})^{-1} \tilde{X}^\top (I - \alpha W)Y$$

and

$$\tilde{\sigma}^2 = \frac{1}{n} ((I - \alpha W)Y - \tilde{X}\vec{\beta}_{LDV})^\top ((I - \alpha W)Y - \tilde{X}\vec{\beta}_{LDV}).$$

Similar with estimation of the ACN model, substituting the $\vec{\beta}_{LDV}$ and $\tilde{\sigma}^2$ into the log-likelihood function and maximizing the function with respect to α by the grid point and the golden section methods, the estimate of α can be obtained. Substituting the estimate of α into the expressions of $\vec{\beta}_{LDV}$, the MLE estimate of $\vec{\beta}$ for the LDV model is calculated.

C. Implementation of our proposed statistical framework of model specification and estimation

For a given pair of time series, the proposed statistical framework of dependence detection consisting of the model specification and estimation can be used as follows:

(1) Obtain the statistics LM_λ , LM_λ^* , LM_α , and LM_α^* for the given pair of series using Eqs. (7), (8), (11), and (12), respectively, for performing the LM tests and their robust scenarios.

(2) Examine significance of these obtained statistics at the given significance level by examining whether their p values are less than the given significance level.

(3) Specify the model using the standard procedure as suggested by Anselin [19] in spatial statistics: If only LM_λ or LM_α is significant, then we accordingly determine the ACN or LDV model as an appropriate model. If both LM statistics are significant, then we further consider the significance of their robust statistics LM_λ^* and LM_α^* . The ACN (LDV) model can be determined if LM_λ^* (LM_α^*) is significant.

(4) Estimate the regression coefficient β_1 to quantify dependence using the MLE for the determined ACN (LDV) model.

The effectiveness and validity of our proposed framework can be justified using simulated series as shown in the next section.

III. RESULTS AND DISCUSSION

A. Simulated series

As mentioned, given the same expression in matrix form, the LM tests and their robust scenarios should be able to specify an appropriate model for detecting dependence in time series. Therefore, we propose a statistical analysis framework based on the LM tests and their robust scenarios for model specification and estimation. Actually, the LM tests and their robust scenarios were developed for the situation with autocorrelation only in the dependent variable. However, if we explore the dependence on real-life series, autocorrelation usually also exists in the explanatory variable. Therefore, we first examine the validity of the LM tests and their robust scenarios to specify models by simulated series with autocorrelation existing in both the dependent and explanatory variables. In our simulations, we therefore use an AR(1) process with the first order AR parameter ρ to model the correlated explanatory time series X and then feed this into the ACN and LDV models to generate the dependent time series Y . We then apply the LM tests and their robust scenarios for model specification. The parameters for generating data using the ACN (LDV) model are set as $n = 36, 412$; $\beta_1 = 5, 30, 50$; and λ (α) = 0.25, 0.55, 0.75, 0.95. The parameter ρ for generating the AR(1) explanatory variable x is 0.75 and 0.95, corresponding to the autocorrelation time $s_c = -1/\ln(\rho)$ equaling 3.48 and 19.50, respectively. As shown in what follows, these parameter settings are determined according to the real-life cases, to be studied later. The significance level is at 0.05. For model specification, we follow the standard procedure in spatial regression [19] with the correspondence of the ACN (LDV) model to the spatial error (lag) model, as summarized in Sec. II C.

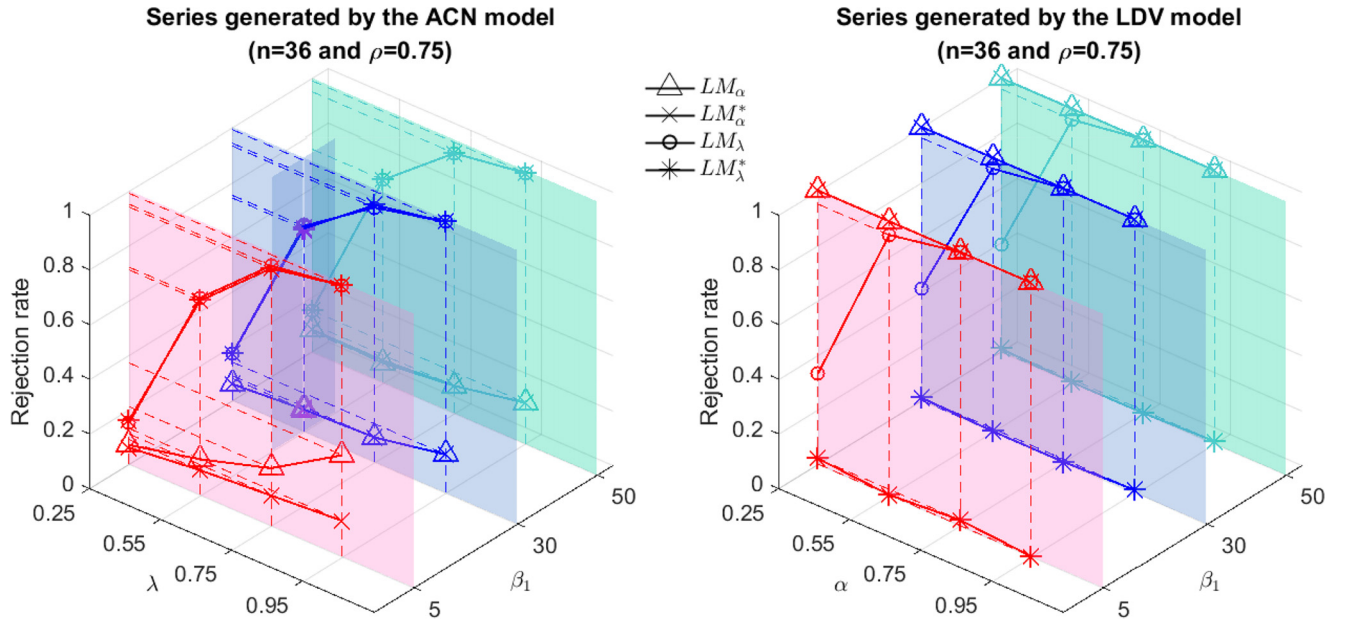


FIG. 1. The rejection rate of the null hypothesis that $\lambda = 0$ or $\alpha = 0$ using the LM tests and their robust scenarios for the series generated by the ACN (left panel) and LDV (right panel) models with $n = 36$ and $\rho = 0.75$. Referring to Sec. II C, if the null hypothesis that $\lambda = 0$ can be rejected, then the ACN model is preferred; if the null hypothesis that $\alpha = 0$ can be rejected, then the LDV model is more appropriate. The symbols in bold indicate the results obtained by the parameter settings corresponding to Case I focusing on the dependence of the top 10% precipitation on global mean temperature in 1979–2014.

For each combination of parameters, we generate 1000 pairs of series using each of two data-generation models and then calculate the rejection rate of the LM tests and their robust scenarios, and the mean of $\hat{\beta}_1$ estimated by the MLE to fit the ACN model and the LDV model. All simulation results are presented in Figs. 1–4 and Table I. In general, the results

for $\rho = 0.75$ (Figs. 1 and 3, Table I) are similar to those for $\rho = 0.95$ (Figs. 2 and 4, Table I).

For the series generated by the ACN model with short length (see Figs. 1 and 2), the LM tests and their robust scenarios have similar performance. Under this circumstance, we only look into the rejection rate of the LM tests. For small λ , both tests have low chance to correctly determine the ACN

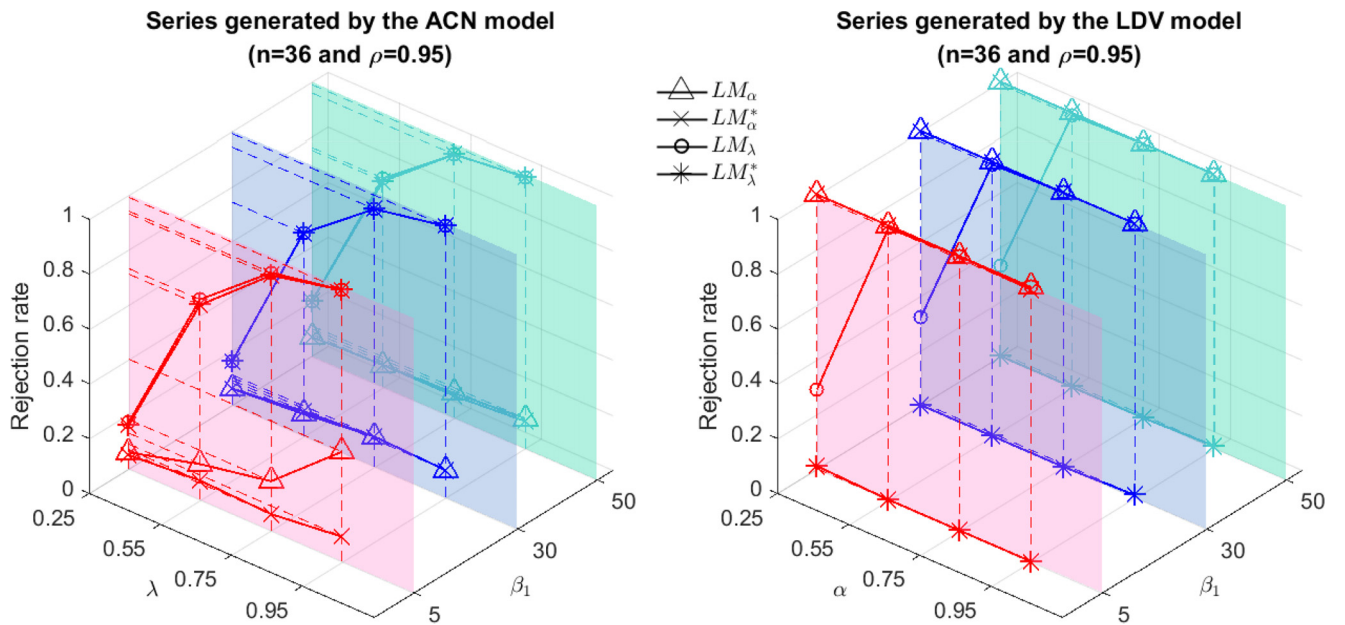


FIG. 2. The rejection rate of the null hypothesis that $\lambda = 0$ or $\alpha = 0$ using the LM tests and their robust scenarios for the series generated by the ACN (left panel) and LDV (right panel) models with $n = 36$ and $\rho = 0.95$. Referring to Sec. II C, if the null hypothesis that $\lambda = 0$ can be rejected, then the ACN model is preferred; if the null hypothesis that $\alpha = 0$ can be rejected, then the LDV model is more appropriate.

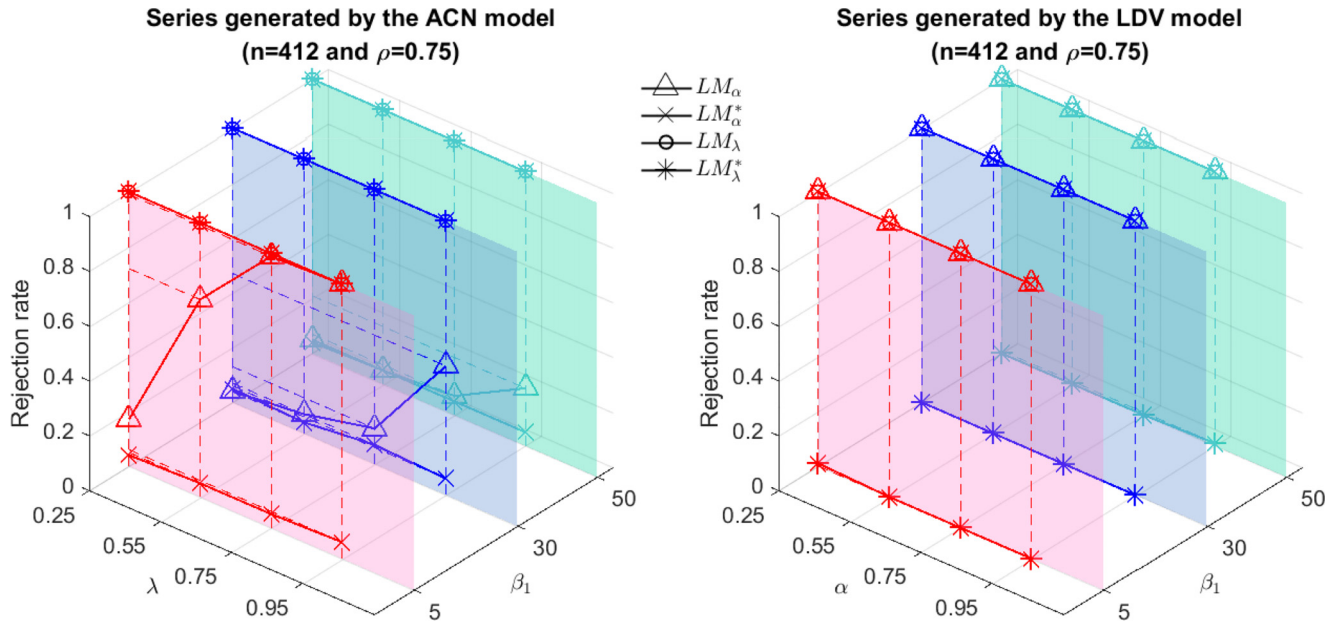


FIG. 3. The rejection rate of the null hypothesis that $\lambda = 0$ or $\alpha = 0$ using the LM tests and their robust scenarios for the series generated by the ACN (left panel) and LDV (right panel) models with $n = 412$ and $\rho = 0.75$. Referring to Sec. II C, if the null hypothesis that $\lambda = 0$ can be rejected, then the ACN model is preferred; if the null hypothesis that $\alpha = 0$ can be rejected, then the LDV model is more appropriate.

model (rejection rates less than 0.2). If λ has a mediate value 0.55, the rejection rate of LM_α is small (below 0.15) but that of LM_λ is over 0.7. Therefore, it is most likely to correctly determine the ACN model. For large λ , the rejection rate of LM_α is relatively low (less than 0.2) whereas that of LM_λ is very high (over 0.9), which indicate the high probability to correctly determinate the ACN model by the LM tests. For

the long ACN model-generated series (see Figs. 3 and 4), the LM tests only work well for $\lambda < 0.55$. By contrast, the robust LM tests can work well for almost all series. In terms of the series generated by the LDV model, the LM tests hardly work, whereas their robust scenarios work perfectly for nearly all series regardless of their length being short or long; see Figs. 1–4.

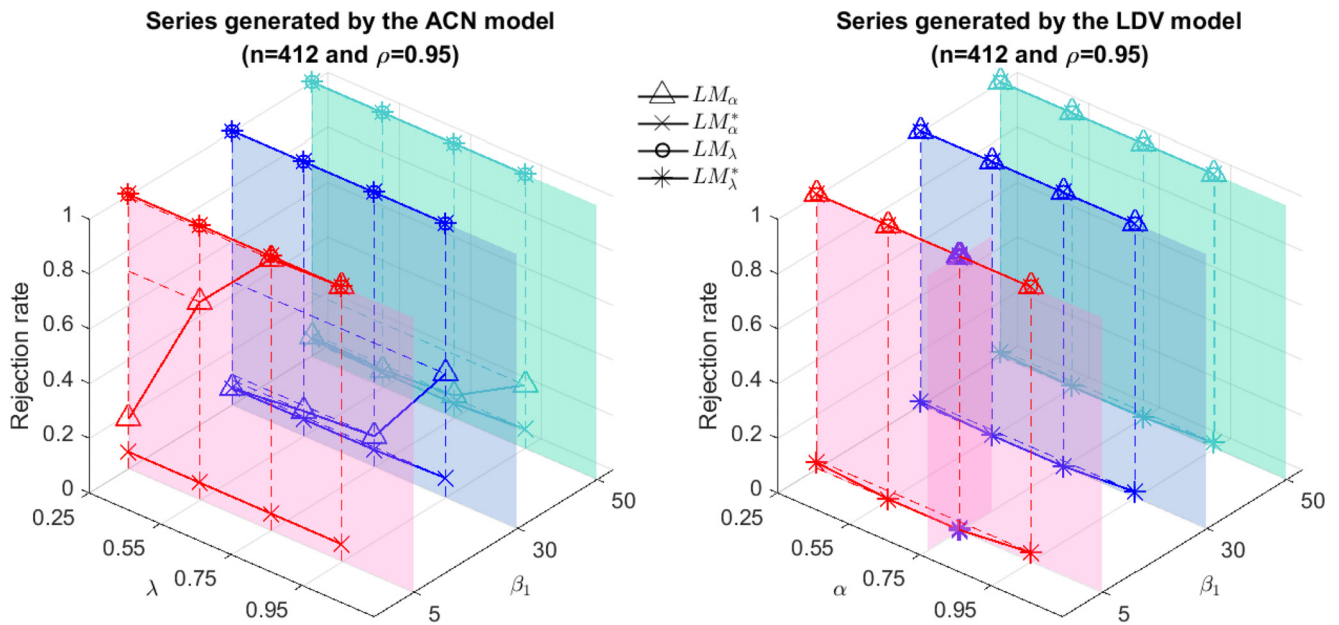


FIG. 4. The rejection rate of the null hypothesis that $\lambda = 0$ or $\alpha = 0$ using the LM tests and their robust scenarios for the series generated by the ACN (left panel) and LDV (right panel) models with $n = 412$ and $\rho = 0.95$. Referring to Sec. II C, if the null hypothesis that $\lambda = 0$ can be rejected, then the ACN model is preferred; if the null hypothesis that $\alpha = 0$ can be rejected, then the LDV model is more appropriate. The symbols in bold indicate the results obtained by the parameter settings corresponding to Case II exploring the temperature dependence on CO_2 for 412 kyr.

TABLE I. The mean of the MLE estimated dependence $\hat{\beta}_1$ obtained by respectively applying the ACN and LDV models to the series generated by the ACN and LDV models. The results in bold are with the parameter settings corresponding to Case I focusing on the dependence of the top 10% precipitation on global mean temperature in 1979–2014 (the ACN data-generation model) and Case II exploring the temperature dependence on CO₂ for 412 kyr (the LDV data-generation model).

		$n = 36$		$\rho = 0.75$	$n = 36$	$\rho = 0.95$	$n = 412$	$\rho = 0.75$	$n = 412$	$\rho = 0.95$	
Data-generation model	α or λ	β_1	ACN	LDV	ACN	LDV	ACN	LDV	ACN	LDV	
ACN model	0.25	5	5.00	4.95	5.01	4.92	5.00	4.96	5.00	4.95	
		30	30.00	29.95	30.00	29.94	30.00	30.00	30.00	30.00	30.00
		50	49.99	49.97	50.00	49.96	50.00	50.00	50.00	50.00	50.00
	0.55	5	4.99	4.89	5.01	4.88	5.00	4.89	5.00	4.86	
		30	30.00	29.96	30.00	29.93	30.00	29.99	30.00	30.00	30.00
		50	50.01	49.96	50.00	49.94	50.00	50.00	50.00	50.00	50.00
	0.75	5	5.00	4.86	5.00	4.83	5.00	4.78	5.00	4.72	
		30	29.99	29.92	30.00	29.92	30.00	29.97	30.00	29.97	29.97
		50	49.99	49.91	50.00	49.92	50.00	49.99	50.00	50.00	50.00
	0.95	5	5.00	4.69	5.00	4.63	5.00	4.07	5.00	3.88	
		30	30.01	29.87	30.01	29.82	30.00	29.79	30.00	29.74	
		50	50.00	49.93	50.00	49.85	50.00	49.88	50.00	49.86	
LDV model	0.25	5	5.68	5.01	6.09	5.01	5.89	5.00	6.45	5.00	
		30	33.93	30.01	36.24	30.00	35.30	30.00	38.68	30.00	
		50	56.73	49.99	60.50	50.00	58.84	50.00	64.50	50.00	
	0.55	5	5.68	5.01	7.14	5.01	6.09	5.00	8.96	5.00	
		30	34.16	30.00	42.41	30.01	36.6	30.00	53.67	30.00	
		50	56.32	50.00	71.43	50.00	61.11	50.00	89.18	50.00	
	0.75	5	4.80	5.01	6.98	5.01	4.98	5.00	9.75	5.00	
		30	28.63	29.99	42.05	29.99	29.86	30.00	58.85	30.00	
		50	47.73	50.00	70.66	50.00	49.66	50.00	97.11	50.00	
	0.95	5	3.34	4.99	6.70	5.00	3.02	5.00	4.84	5.00	
		30	20.05	30.00	40.22	30.00	18.13	30.00	29.34	30.00	
		50	33.09	49.99	64.17	50.00	30.21	50.00	50.00	50.00	

As mentioned, the regression coefficient β_1 quantifies dependence. All estimates are summarized in Table I. The accuracy of estimate of β_1 using the correctly specified model is always very high. The misspecification for the ACN model generated series generally results in little underestimation. By contrast, for the LDV model generated series, the obvious deviation from the expected value can be observed in the estimates by the misspecified model. The misspecification effect appeared in simulations actually depends on the parameters ρ , α or λ , n , and β_1 in a very complicated way. Therefore, to reliably and accurately obtain the true dependence it is of importance to correctly specify the model before estimation.

As summarized in Figs. 1–4, the performance of the LM tests and their robust scenarios for the model specification is satisfactory, even if there is autocorrelation in both dependent and explanatory variables, except for the situation with series generated by the ACN model with the weakly autocorrelated noise and short length (see Figs. 1 and 2). Fortunately, in such situation, the misspecification effect is very little (see Table I).

As shown, the effectiveness of our proposed statistical framework of dependence detection can be supported by the simulation experiments conducted in this subsection, over a wide range of parameters. In brief, this framework first uses the LM tests and their robust scenarios for model specification; afterwards, the determined ACN or LDV model is estimated by the MLE.

B. Real-life series

The evident global warming may have many serious consequences. In empirical analysis, we explore the dependence of precipitation on temperature and the temperature dependence on CO₂, both of which are important research topics in climate under global warming [1]. Particularly, the variation of precipitation extreme under global warming is closely related to the human welfare, ecosystems, and occurrence of floods [23]. Usually, the traditional model ignoring autocorrelations with the OLS estimation is employed to detect the dependence of precipitation on temperature, e.g., the analyses in [24–26]. For example, the traditional OLS-estimated dependence uncovers that heavy precipitation (the top 10%) normalized by the long-term mean annual record can increase about 50% for 1 K growth of the global mean temperature in the period 1979–2014 [26], which is an impressive effect. Correctly quantifying the effect of global warming is fundamental for understanding the underlying process and the further policy making to mitigate its severe consequences.

However, it should be emphasized that autocorrelation in the precipitation and temperature series may invalidate estimation of the traditional linear regression using OLS and was often not appropriately handled in analysis. To tackle autocorrelation in regression for the dependence detection, the ACN model is usually adopted in geophysical studies [3–5]. However, by looking at a physically motivated energy

TABLE II. The p values of the statistics of the LM tests and their robust scenarios and $\hat{\beta}_1$ estimated by the MLE for the ACN model and the LDV model in two case studies. Case I focuses on the dependence of the top 10% precipitation on global mean temperature in 1979–2014. Case II explores the temperature dependence on CO₂ for 412 kyr. The values of $\hat{\beta}_1$ in bold are results obtained by the preferred model and the corresponding estimation using our proposed analysis framework. The units of estimates in Cases I and II are %K⁻¹ and °C/ppmv, respectively. For systematic comparisons, we also list $\hat{\beta}_1$ obtained by the traditional model ignoring autocorrelation in series.

Case	Length	p value				$\hat{\beta}_1$ (traditional model)	$\hat{\beta}_1$ (ACN model)	$\hat{\beta}_1$ (LDV model)
		LM _{α}	LM _{λ}	LM _{α} [*]	LM _{λ} [*]			
I	36	0.0084	0.0007	0.3841	0.0223	49.75	32.28	31.05
II	412	0.0000	0.0000	0.0000	0.6283	18.94	14.81	5.42

balance equation [11,14], it seems more reasonable to explicitly express autocorrelation in the dynamics instead of being hidden in the noise, especially for detecting the dependence of temperature on some factors, e.g., CO₂, which would be an LDV model. Therefore, effectively specifying the regression model is of great importance in empirical analysis.

Using the proposed statistical framework, we study dependence in two real-life cases and examine how much difference can be caused if autocorrelation is not appropriately handled. The first case (Case I) is the dependence of the top 10% precipitation normalized by the long-term annual mean on the global mean temperature, which is the focus in [23,26,27]. The other case (Case II) is whether and how temperature depends on CO₂, which is studied in [28].

In Case I, the top 10% precipitation is annual data and was first extracted from the Global Precipitation Climatology Project (GPCP) [29] using the procedure proposed by Shiu *et al.* [27] with normalization relative to the long-term mean annual record. Therefore, this variable is without unit. The annual global temperature data are from NOAA Merged Land-Ocean Surface Temperature Analysis (MLOST) V3.5.4 [30] with the unit K. These two annual series have length of 36 covering the period 1979–2014 with similar ρ about 0.75 (actual values 0.78 and 0.74, respectively for the top 10% precipitation and global mean temperature).

In Case II, the temperature and CO₂ series are reconstructed from the Vostok ice core [31] with time interval 1 kyr and length 412 [28,32]. The units of temperature and CO₂ are °C and ppmv, respectively. It is suggested that one should associate temperature with the logarithm of CO₂ rather than directly with CO₂ [13,14]. Therefore, we actually use the logarithm of CO₂ in all following analyses and discussion but still denote it as CO₂ for simplification. The value of ρ of the ice core temperature and CO₂ is about 0.95 (exactly 0.94 and 0.97, respectively).

For the purpose of systematic comparisons, we also apply OLS to estimate dependence based on the traditional linear regression model and use the misspecified model to estimate dependence. Table II gives the p values of the LM tests and their robust scenarios and the estimated dependence $\hat{\beta}_1$ in these two case studies.

For Case I, the p values of LM _{α} , LM _{λ} , and LM _{λ} ^{*} are less than 0.05; whereas that of LM _{α} ^{*} is 0.3841, indicative of the preference for the ACN model. The traditional OLS-estimated regression coefficient gives $\hat{\beta}_1$ about 50% K⁻¹ (exactly, 49.75% K⁻¹ with 95% confidence interval, CI, [28.52% K⁻¹, 70.97% K⁻¹]), which turns out to overestimate

the dependence. Our proposed framework suggests that the more likely dependence should be only about 65% of this traditional estimate, specifically 32.28% K⁻¹ with the 95% CI [6.27% K⁻¹, 58.28% K⁻¹], which is obtained by the MLE using the ACN model. Denoting the top 10% precipitation and temperature as p_t and temp_t at the annual resolution, the preferable model for Case I specified by our proposed framework is an ACN model:

$$p_t = \beta_1 \text{temp}_t + \beta_0 + u_t, \quad (13)$$

with $u_t = \lambda u_{t-1} + e_t$. The misspecified model, i.e., the LDV model, only causes little underestimation, indicating the slight effect if we misspecify an ACN model to an LDV model in this case. Actually, for Case I, the estimated ρ , λ , β_1 of the correctly specified ACN model are about 0.75, 0.55, and 30. It corresponds to one simulation situation, whose results are indicated in bold for the results of the series generated by the ACN model in Table I and marked by the bold symbols in Fig. 1. The performance of estimation and the little misspecification effect are perfectly consistent with the simulation results, which supports the results obtained by our proposed framework in this case study.

With regard to Case II, only LM _{λ} ^{*} has a p value larger than 0.05, implying that the LDV model is a more appropriate model, as expected from the recent finding in [13,14]. Denoting temperature and CO₂ as temp_t and c_t at the 1 kyr temporal resolution, the determined model for Case II is an LDV model in the following form:

$$\text{temp}_t = \alpha \text{temp}_{t-1} + \beta_1 c_t + \beta_0 + e_t. \quad (14)$$

The estimated β_1 of the above LDV model by the MLE is less than one-third of the traditional OLS-estimated β_1 , i.e., 5.42°C/ppmv with the 95% CI [4.38°C/ppmv, 6.46°C/ppmv] versus 18.94°C/ppmv with the 95% CI [17.86°C/ppmv, 20.03°C/ppmv]. Such difference in estimates is enormous. The misspecified ACN model produces the estimate 14.81°C/ppmv with the 95% CI [12.57°C/ppmv, 17.05°C/ppmv]. Therefore, the model misspecification effect is evident in Case II. For Case II, the estimated ρ , α , β_1 of the correctly specified LDV model are about 0.95, 0.75, and 5, which corresponds to one simulation situation with its results in bold when the LDV model is used as the data-generation model presented in Table I and represented by the bold symbols in Fig. 4. It is noteworthy that the estimation performance and overestimation by the misspecified ACN model are also perfectly consistent

with simulations, justifying the findings of our proposed framework in studying Case II.

In summary, the validity of analysis of these two real-life case studies can be justified by simulations with similar parameter settings. All dependences obtained by the correctly specified models are significant and have values less than those obtained by the traditional OLS estimation, which omits autocorrelation in series. In addition, it is demonstrated that misspecified models may result in highly biased estimates.

IV. CONCLUSIONS AND REMARKS

Autocorrelation is ubiquitous in real-life series. However, as a common method for dependence detection, the traditional linear regression fitted by OLS assumes the absence of autocorrelation in series. Existing studies have demonstrated that the omission of autocorrelation in variables may mislead the statistical conclusion regarding the estimated dependence. To tackle autocorrelation in series, two generalizations, namely the regression model with inclusion of the ACN and the LDV, have been proposed. Existing studies on these two models mainly focus on estimation and the misspecification effect, and have shown the difference in estimate between these two generalized models. However, little attention has been paid to statistically specify the appropriate model, although its necessity is obvious.

Using simulations, we have demonstrated the effectiveness of the LM tests and their robust scenarios to correctly determine model even if autocorrelation exists in both of the dependent and explanatory variables, unless autocorrelation is too weak in short series. Taking advantage of the LM tests and their robust scenarios to specify models for dependence detection when series are autocorrelated, we propose a statistical framework of model specification and estimation for the autocorrelated series as follows: First, the LM tests and their robust scenarios are used to determine the ACN or LDV model. Second, the determined ACN or LDV model is estimated by the MLE.

After its effectiveness is justified by simulation experiments, the proposed statistical framework is employed to study two real-life cases to examine its applicability in empirical analysis. With respect to Case I exploring heavy-precipitation dependence on global warming, the specified ACN model gives the estimated dependence only about 65% of that obtained by the traditional OLS estimation omitting autocorrelation. For Case II focusing on the dependence of temperature on CO₂, the LDV model is preferred. Its estimate is less than 30% of the traditional OLS estimate and is about

35% of that obtained using the misspecified ACN model. All results of case studies are consistent with those obtained by simulation experiments with similar parameter settings, which makes the results of these two case studies obtained by our proposed framework convincing.

As justified by analysis of simulated and real-life data, our proposed framework can provide a solid statistical basis to significantly advance the frontiers of research exploring the true dependence when series are autocorrelated. It is greatly helpful for further investigation of the underlying mechanism of one process impacting another one. It should be emphasized that the proposed framework is actually applicable to detect dependence for not only the geophysical series but the autocorrelated series in other fields. Although we only consider autocorrelation in a form of the AR(1) process in this study, it does not mean that we rule out the higher order process and long-range correlation modeling and simulations for geophysical series. In principle, our proposed framework is based on the framework developed in spatial statistics. Taking longer correlated processes into consideration, one may need to consider the lag term with higher orders. Especially, if the long-range correlated processes need to be considered, then we may need to consider the lag term with infinite orders. In this sense, the corresponding tests may become very complicated, which is not easy to handle by extending our framework proposed in this study. However, by assuming the AR(1) processes in the series, the AR(1) process obtained by fitting to the long-range correlation can still account for part of the autocorrelation effect. Therefore, our proposed framework can still provide more useful information on dependence detection, compared with those that do not consider the autocorrelation like the traditional model. How to generalize this study to tackle long-range correlation would be our future research topic.

ACKNOWLEDGMENTS

This work is supported by the Key Program of Natural Science Foundation Joint Fund of Hunan Province of China (Grant No. 2026JJ30055), the Key Project of the Education Department of Hunan Province of China (Grant No. 25A0133), and the Chongqing Social Science Planning Project (Cultivation Project) (Grant No. 2025PY43).

DATA AVAILABILITY

The data that support the findings of this article are openly available [29–31].

-
- [1] T. Stocker, D. Qin, G.-K. Plattner, M. Tignor, S. K. Allen, J. Boschung, A. Nauels, Y. Xia, V. Bex, and P. M. Midgley, *Climate Change 2013: The Physical Science Basis* (Cambridge University Press, Cambridge, UK, New York, 2014).
 - [2] P. L. Meyer, *Introductory Probability and Statistical Applications* (Addison-Wesley, Reading, MA, 1965).
 - [3] H. von Storch, Misuses of statistical analysis in climate research, *Analysis of Climate Variability: Applications*

of Statistical Techniques (Springer-Verlag, Berlin, 1995), pp. 11–26.

- [4] J. Imbers, A. Lopez, C. Huntingford, and M. Allen, Testing the robustness of the anthropogenic climate change detection statements using different empirical models, *J. Geophys. Res.* **118**, 3192 (2013).
- [5] J. Imbers, A. Lopez, C. Huntingford, and M. Allen, Sensitivity of climate change detection and attribution to the characterization of internal climate variability, *J. Clim.* **27**, 3477 (2014).

- [6] G. E. Box, G. M. Jenkins, G. C. Reinsel, and G. M. Ljung, *Time Series Analysis: Forecasting and Control* (John Wiley & Sons, New York, 2015).
- [7] D. F. Hendry, *Dynamic Econometrics* (Oxford University Press, Oxford, 1995).
- [8] C. H. Achen, Why lagged dependent variables can suppress the explanatory power of other independent variables, in *The Annual Meeting of the Political Methodology Section of the American Political Science Association*, Los Angeles (2000).
- [9] L. Keele and N. J. Kelly, Dynamic models for dynamic theories: The ins and outs of lagged dependent variables, *Polit. Anal.* **14**, 186 (2006).
- [10] A. S. Wilkins, To lag or not to lag?: Re-evaluating the use of lagged dependent variables in regression analysis, *Polit. Sci. Res. Methods* **6**, 393 (2018).
- [11] J. Hansen, M. Sato, P. Kharecha, and K. von Schuckmann, Earth's energy imbalance and implications, *Atmos. Chem. Phys.* **11**, 13421 (2011).
- [12] M. Mudelsee, *Climate Time Series Analysis* (Springer, Dordrecht, 2013).
- [13] S. Hu and A. V. Fedorov, The extreme El Niño of 2015–2016 and the end of global warming hiatus, *Geophys. Res. Lett.* **44**, 3816 (2017).
- [14] P. Meyer, M. Hoell, and H. Kantz, Reproducing long-range correlations in global mean temperatures in simple energy balance models, *J. Geophys. Res.* **123**, 4413 (2018).
- [15] L. Anselin, Lagrange multiplier test diagnostics for spatial dependence and spatial heterogeneity, *Geogr. Anal.* **20**, 1 (1988).
- [16] L. Anselin, A. K. Bera, R. Florax, and M. J. Yoon, Simple diagnostic tests for spatial dependence, *Reg. Sci. Urban Econ.* **26**, 77 (1996).
- [17] L. Anselin and R. J. Florax, Small sample properties of tests for spatial dependence in regression models: Some further results, in *New Directions in Spatial Econometrics* (Springer, Berlin, Heidelberg, 1995), pp. 21–74.
- [18] P. Burridge, On the Cliff-Ord test for spatial correlation, *J. R. Stat. Soc. Ser. B, Stat. Methodol.* **42**, 107 (1980).
- [19] L. Anselin, *Modern Spatial Econometrics in Practice: A guide to GeoDa, GeoDaSpace and PySAL* (GeoDa, Urbana, IL, 2014).
- [20] J. LeSage, What regional scientists need to know about spatial econometrics, *Rev. Reg. Stud.* **44**, 13 (2014).
- [21] L. Anselin, *Spatial Econometrics: Methods and Models*, Studies in Operational Regional Science Vol. 4 (Springer Science & Business Media, Dordrecht, 1988).
- [22] L. Anselin and A. K. Bera, Spatial dependence in linear regression models with an introduction to spatial econometrics, in *Handbook of Applied Economic Statistics*, edited by A. Ullah and D. E. A. Giles (Marcel Dekker, New York, 1998), pp. 237–289.
- [23] S. C. Liu, C. Fu, C.-J. Shiu, J.-P. Chen, and F. Wu, Temperature dependence of global precipitation extremes, *Geophys. Res. Lett.* **36**, L17702 (2009).
- [24] K. E. Trenberth, J. Fasullo, and L. Smith, Trends and variability in column-integrated atmospheric water vapor, *Clim. Dyn.* **24**, 741 (2005).
- [25] Y. Sun, S. Solomon, A. Dai, and R. W. Portmann, How often will it rain? *J. Clim.* **20**, 4801 (2007).
- [26] Y. Zhou, M. Luo, and Y. Leung, On the detection of precipitation dependence on temperature, *Geophys. Res. Lett.* **43**, 4555 (2016).
- [27] C.-J. Shiu, S. C. Liu, C. Fu, A. Dai, and Y. Sun, How much do precipitation extremes change in a warming climate? *Geophys. Res. Lett.* **39**, 2012GL052762 (2012).
- [28] E. H. Van Nes, M. Scheffer, V. Brovkin, T. M. Lenton, H. Ye, E. Deyle, and G. Sugihara, Causal feedbacks in climate change, *Nat. Clim. Change* **5**, 445 (2015).
- [29] Global precipitation climatology project, <https://www.ncei.noaa.gov/data/global-precipitation-climatology-project-gpcp-pentad/access/>.
- [30] NOAA Merged Land-Ocean Surface Temperature Analysis, <https://climatedataguide.ucar.edu/climate-data/global-surface-temperature-data-mlost-noaa-merged-land-ocean-surface-temperature>.
- [31] Vostok ice core, <https://www.usap-dc.org/view/dataset/609242>.
- [32] J.-R. Petit, J. Jouzel, D. Raynaud, N. I. Barkov, J.-M. Barnola, I. Basile, M. Bender, J. Chappellaz, M. Davis, G. Delaygue, *et al.*, Climate and atmospheric history of the past 420,000 years from the Vostok ice core, Antarctica, *Nature (London)* **399**, 429 (1999).