



ARTICLE



<https://doi.org/10.1057/s41599-025-05107-0>

OPEN

Big data and firm greenwashing: evidence from China's national big data comprehensive pilot zone

Jianjun Sun¹, Nannan Zu¹✉ & Chao Zhang¹

Big data is playing an increasingly important role in addressing environmental challenges. Using the launch of China's national big data comprehensive pilot zone in 2016 as an exogenous event, this paper employs the difference-in-differences method to examine the causal effect of big data on firm greenwashing. The regression result indicates that big data has an inhibitory effect on firm greenwashing. A plausible underlying mechanism is that big data enhances environmental performance and alleviates information asymmetry between the firm and stakeholders. The inhibitory effect is more pronounced in state-owned enterprises and those with high market competition. The findings better explain how the digital factor drives green development at the firm level.

¹ International Business School, Hainan University, Haikou, China. ✉email: zunannan@hainanu.edu.cn

Introduction

Green development has become a global consensus. As direct contributors to ecological degradation, firms face increasing social attention and legitimacy pressure regarding their environmental protection actions (Delmas and Toffel, 2004). Environmental information disclosure is crucial for firms to effectively communicate these actions to stakeholders (Kuo and Chen, 2013). However, greenwashing has become a strategic practice for some firms, who utilize their information advantages to selectively disclose favorable environmental information or use misleading language to conceal insubstantial environmental protection efforts (Delmas and Burbano, 2011; Seele and Gatti, 2017; Zhang, 2023). As the global digital economy grows, big data drives a new wave of technological revolution. At its core, big data involves the generation, integration, analysis, and application of massive data, and its importance in fighting environmental challenges is becoming increasingly prominent (Wu et al., 2016; Gao et al., 2023). Does big data affect firm greenwashing? If so, what is the underlying mechanism? Investigating these questions is crucial for explaining how the digital factor drives green development at the firm level.

The literature relevant to this paper is divided into two main streams: one examines the determinants affecting firm greenwashing. Existing literature indicates that firms embellish environmental performance in response to environmental regulations (Hu et al., 2023; Zhou et al., 2024), seek financial resources (Zhang, 2023), and cater to the green preferences of investors and consumers (Delmas and Burbano, 2011). While on the other hand, existing literature discovers factors such as social media (Fernando et al., 2014), independent directors and institutional investors (Yu et al., 2020), and retail investor activism (Zhang, 2024) that inhibit corporate greenwashing, the practice remains covert and complex, posing challenges for effective regulation and governance. With the rise of the digital economy, existing literature finds that fintech (Xie et al., 2023), digital finance (Yin and Yang, 2024), government digitalization (Xu et al., 2024), corporate digital transformation (Wang et al., 2024), and artificial intelligence (Zhang, 2024) are effective tools for governing corporate greenwashing. The second stream probes how big data affects business operations and development. Existing literature explores the positive effects of big data on innovation from various perspectives, including business model innovation (Ciampi et al., 2021), green technology innovation (Gao et al., 2023), process innovation, innovation by diverse recombination (Wu et al., 2020), and organizational service innovation (Troilo et al., 2017). Furthermore, some studies demonstrate that big data reshapes traditional decision-making paradigms (Yaqoob et al., 2016), enables product sales forecasting to be more accurate, and enhances firm supply chain agility and adaptability (Wamba et al., 2020). Additionally, big data increases the productivity of the firms and corporate performance (Müller et al., 2018; Vitari and Raguseo, 2020). The literature also examines the feasibility of big data in green development through qualitative analysis (Wu et al., 2016). Overall, existing literature has begun to examine how the digital technologies of various organizations, such as governments (Xu et al., 2024) and enterprises (Yin and Yang, 2024; Zhang, 2024), affect corporate greenwashing. Although big data provides abundant data resources for digital technologies, there are notable differences in the attributes between big data and digital technologies. Big data is the deep integration of data sets and technological systems, which allows data mining and efficient application through high-speed access, integration, and analysis of large and diverse data resources (Chen et al., 2012). The core characteristics of big data include high volume, high velocity, wide variety, significant veracity, and low value density (McAfee and Brynjolfsson, 2012; Wamba et al., 2015). In contrast, digital

technologies rely on exact processing and efficient transmission of information via binary encoding, discretization, and algorithmic implementation (Fitzgerald et al., 2014). Consequently, the effect of big data on corporate greenwashing and its underlying mechanism needs to be further clarified.

To achieve this, we use the launch of China's national big data comprehensive pilot zone in 2016 as a policy shock and employ the difference-in-differences specification to examine its effect on corporate greenwashing. Difference-in-differences is a causal inference method used to estimate the net effect of policy interventions by comparing changes before and after the policy between the treatment group (the group affected by the policy) and the control group (the group not affected by the policy). We define enterprises located within the pilot zones as the treatment group and those outside the pilot zones as the control group. Then, we estimate the net effect of big data policy on firm greenwashing by comparing the changes in firm greenwashing of the two groups before and after the establishment of pilot zones.

This study provides three marginal contributions as follows. First, this article contributes to the existing literature by examining how policy-driven data technology development at the regional level affects corporate greenwashing. Although existing literature has begun to explore the effect of digital factors, such as fintech (Xie et al., 2023), digital finance (Yin and Yang, 2024), government digitalization (Xu et al., 2024), corporate digital transformation (Wang et al., 2024), and artificial intelligence (Zhang, 2024), on firm greenwashing, it is primarily limited to the digitalization or digital technology applications within different organizations, including government agencies and enterprises. China's national big data comprehensive pilot zones are a regional pilot policy aimed at promoting the regional development of big data techniques, providing a more objective indicator for measuring regional big data development, so we essentially examine the effect of the regional development of big data techniques at the macro-level policy on corporate greenwashing. The empirical finding extends our understanding of how the digital factor governs corporate greenwashing. Second, it contributes to the studies on the effect of big data on green development. The existing literature indicates that big data enhances green technology innovation (Gao et al., 2023) and qualitatively discusses its role in driving green changes (Wu et al., 2016). According to our research, big data significantly inhibits firm greenwashing, adding to the knowledge of how big data affects green development from the perspective of green speculation. Finally, our study clarifies the underlying mechanism through which big data affects firm greenwashing. We find that big data inhibits corporate greenwashing by enhancing environmental performance and alleviating information asymmetry, offering deeper insights into the effect of big data on greenwashing.

The remainder of this paper is organized as follows. **Institutional background and hypothesis development** cover the institutional background of China's national big data comprehensive pilot zones and hypothesis development. **Data, variable construction, and descriptive statistics** describe the data, variable construction, and descriptive statistics. **Empirical specification and results** describe the empirical specification and results. **Further analysis** provides more analysis, while **Conclusion** concludes.

Institutional background and hypothesis development

Institutional background. China's central government recognized big data technologies as a disruptive force for the future. The Action Outline for Promoting Big Data Development was issued in 2015, which makes it evident that the government in

China attempted to build the foundations for big data technologies and their applications. Despite the initial top-to-bottom push from the government, the development of big data technologies was still subject to some constraints, such as lagging law and regulation, integrated planning, openness, and data sharing. As a result, the push for big data development has been almost ineffective. To ease these constraints, the central government switched to selecting some regions in a balanced manner across the country to pilot big data technologies and their applications in 2016. The national comprehensive big data pilot zones were officially launched. The local governments in these regions were given special administrative authority to ease the constraints on developing big data technologies and their applications. The eight regions included the Pearl River Delta, Beijing, Tianjin, Hebei, Shanghai, Chongqing, Henan, Inner Mongolia, Guizhou, and Shenyang. These eight regions launched a pilot in several major aspects, such as big data systematic innovation, the opening and sharing of public data, the application and advancement of big data, the circulation of data factors, the integration and utilization of data centers, and international big data collaboration, aiming to promote regional development of big data technologies. Compared to the others, these regions establish comprehensive development systems of big data technologies through pilot policies, which enable them to have significant advantages in data acquisition, integration, analysis, application, circulation, and collaboration. As a result, they make data mining more effective and promote the deeper and broader application of data.

Hypothesis development. The two most basic questions about greenwashing are: Why do companies greenwash? And what is the nature of greenwashing? When the accounting benefits of corporate environmental governance are insufficient to compensate for the accounting costs, corporate management is under pressure for short-term financial performance, and they are strongly motivated to greenwash. Thus, environmental performance declines (Clatworthy and Jones, 2003; Hu et al., 2023). If corporate environmental governance becomes more accessible and affordable, firms are more likely to improve their environmental governance performance, and the incentive to greenwash is reduced. In essence, greenwashing is the distortion, exaggeration, or concealment of a company's environmental information (Seele and Gatti, 2017). When it becomes difficult to falsify, exaggerate, or conceal information, corporate greenwashing becomes more difficult to achieve. Therefore, the impact of the regional development of big data technologies on corporate greenwashing depends on whether big data enhances corporate environmental performance and whether it mitigates environmental information asymmetry between firms and stakeholders. We develop a null hypothesis against an alternative hypothesis as follows.

H1₀: Big data significantly inhibits firm greenwashing vs. H1₁: Big data significantly drives firm greenwashing.

In traditional production models, corporate environmental governance activities are usually characterized by high uncertainty, high expenditures, and long cycles. These characteristics make it difficult for the returns on environmental governance activities to cover investment costs, thereby weakening the effort for companies to fulfill their environmental responsibilities (Gray, 1987; Palmer et al., 1995). In sustainable production models, however, businesses with poor environmental performance face legitimacy pressures such as negative public perception and social ostracism (Lin et al., 2016), substantial fines from regulatory authorities (Shimshack and Ward, 2005), and more stringent credit requirements from financial institutions (Chava, 2014). As a result, these businesses are highly motivated to adopt low-cost

greenwashing strategies, attempting to embellish their poor environmental performance and portray a false image of environmental friendliness (Meng et al., 2014).

The development of big data techniques makes it easier and cheaper for companies to improve environmental performance, thus weakening the incentive to greenwash. On the one hand, big data makes it easier and lower-cost for companies to adopt a data-driven model to reduce pollutant emissions in their production processes. Big data offers real-time dynamic monitoring of resource inputs and pollutant emissions. Companies use environmental data analysis and frequent feedback to adjust their production strategies on time, optimizing the balance between resource availability, demand, and pollutant emissions. This approach reduces resource waste and pollutant emissions (Wu et al., 2016). Data sharing empowers enterprises and their supply chain partners to establish close alliances dedicated to environmental governance, which facilitates the accurate measurement and dynamic control of energy consumption and carbon emissions across the supply chain from procurement to warehousing to logistics. This data-driven green collaborative control model enhances the enterprises' efficiency of environmental governance and also promotes the optimization of environmental performance across the entire supply chain network (Raut et al., 2019). Moreover, big data integrates environmental pollution data from various sources, which clarifies the origins and dispersion channels of pollution. It provides a timely warning of pollution incidents and motivates companies to develop cooperative prevention and control of environmental pollution (Wu et al., 2016). On the other hand, big data enhances corporate green innovation. Big data makes it easier and less expensive to access and analyze a large amount of consumer data, which enables companies to more accurately predict consumers' green preferences so that the direction of R&D can be more in line with consumer preferences and reduce the trial-and-error costs of green innovation (Troilo et al., 2017; Yaqoob et al., 2016). Big data also transcends traditional spatial boundaries and enables enterprises to develop closer connections with consumers, suppliers, universities, and academic institutions. Big data offers an efficient platform for knowledge and information exchange, resource sharing, and R&D collaboration. It accelerates the green R&D process and, at the same time, promotes the dissemination and use of green technologies (Chen et al., 2024). Based on the above analysis, we develop the second hypothesis.

H2: Big data significantly improves firm environmental performance, thereby inhibiting corporate greenwashing.

In a traditional business environment, corporate environmental governance is a secondary function, and the environmental information is often hidden in a large volume of primary business information. This information is often discontinuous, confusing, and diverse. Meanwhile, the lack of standardized environmental information disclosure norms and regulations grants management considerable discretion in determining what and how much information to disclose (Yu et al., 2020). Stakeholders face multiple challenges when collecting environmental information, such as narrow search scopes, ambiguous results, and high costs, all of these challenges make it difficult to accurately assess whether companies are fulfilling or not fulfilling their environmental obligations. Consequently, this information asymmetry creates a fertile ground for corporate greenwashing (Li et al., 2023; Wu et al., 2020).

The development of big data techniques mitigates information asymmetry between the firm and stakeholders, thus enhancing the completeness and reliability of environmental information disclosure. First, big data improves firms' ability to process environmental information. Big data captures massive information throughout the production-to-sales process. It then

standardizes and processes huge amounts of information using technologies such as data aggregation, hierarchical classification, and multi-dimensional analysis, enabling enterprises to integrate and use diverse forms of data (Oussous et al., 2018). This shifts the long-standing practice of firms relying on human experience and subjective judgment to process environmental information, transforming it from fragmented and ambiguous to comprehensive and verifiable. Second, big data increases the flow of information and communication between the firm's departments, decreasing the possibility of management manipulating environmental information. The data network enables the board of directors, employees, and other internal stakeholders to access diverse environmental data. Therefore, it is more likely to stop the distortion and concealment of environmental information within the company and also make it more effective in reducing management discretion in environmental decision-making. Finally, in the big data environment, the volume of environmental information released by enterprises to stakeholders outside the company has surged. Environmental information disclosure has become more diversified, both in terms of content and format, as well as through communication channels and tools (Blazquez and Domenech, 2018). These stakeholders access firm environmental information more easily through lower-cost and diversified channels. To assess and verify information through big data, stakeholders can make precise and effective decisions about whether firms are achieving their environmental obligations. For example, big data enables verification of consistency between environmental indicators voluntarily disclosed by enterprises and those reported by government regulatory agencies. It also facilitates a comparative analysis of a company's environmental performance against industry benchmarks, thereby identifying significant deviations that may indicate anomalies or discrepancies in environmental practices (Chen et al., 2012). Moreover, big data facilitates the collection and integration of social media reports and consumer feedback, providing real-time insights into diverse public perceptions of corporate environmental behavior (Chen et al., 2012). Therefore, big data enables the company's outside stakeholders to identify potential corporate greenwashing. Based on the above analysis, we develop the third hypothesis.

H3: Big data significantly alleviates information asymmetry between enterprises and stakeholders, thereby inhibiting corporate greenwashing.

Data, variable construction, and descriptive statistics

Data. We chose A-share firms listed on China's Shanghai and Shenzhen Stock Exchanges from 2012 to 2022 as the initial research sample. We remove the outliers from the original sample using the approach taken in existing literature (for example, He et al., 2025). Financial institutions are highly regulated, and therefore, we exclude them. Considering that ST firms exhibit financial anomalies, such as consecutive years of losses or insolvency, we also exclude them from the original sample. We exclude enterprises with significant missing data to ensure that all variables can be measured. We finally obtained 1293 A-share listed enterprises, totaling 10509 sample observations. The data on measuring the greenwashing indicator are sourced from the Bloomberg and WIND databases, respectively. Other data are sourced from the China Stock Market and Accounting Research (CSMAR) database. To avoid extreme observation interference, all continuous variables are winsorized at the 1st and 99th percentiles.

Variable construction. The dependent variable (GW) in this paper is firm greenwashing. Existing literature (Hu et al., 2023; Zhang, 2023; Yu et al., 2020) measures ESG (environmental,

social, and governance dimensions) greenwashing by calculating the gap between a company's standardized ESG disclosure score and its standardized ESG real performance, both of which are standardized relative to peer enterprises.¹ However, this paper explores corporate greenwashing from an environmental aspect. Since the environmental dimension of ESG directly reflects a company's environmental practices, following the approach of Hu et al. (2023), we measure greenwashing by calculating a firm's standardized environmental dimension disclosure score and subtracting its standardized environmental dimension actual performance. The design of greenwashing is as follows.

$$GW_{i,t} = (E_{dis\ i,t} - \bar{E}_{dis})/\sigma_{dis} - (E_{per\ i,t} - \bar{E}_{per})/\sigma_{per} \quad (1)$$

Where i denotes firm, t refers to year, GW is firm greenwashing. E_{dis} is a company's environmental disclosure score, which is indicated by the Bloomberg environmental dimension score.² \bar{E}_{dis} represents the average environmental disclosure score of peer companies. σ_{dis} denotes the standard deviation of environmental disclosure scores among peer companies. E_{per} represents a firm's real environmental performance, as indicated by the Huazheng environmental dimension rating.³ \bar{E}_{per} is the average of the real environmental performance of peer enterprises. σ_{per} denotes the standard deviation of real environmental performance among peer enterprises.

We construct the primary explanatory variable (Post×Treat) using the launch of China's national big data pilot zone in 2016, which is the interaction between the dummy variable (Post) representing the launch time of the pilot zone and the dummy variable (Treat) indicating the firm situated in the pilot zone. Post is equal to 1 if the year is 2016 or later and 0 otherwise. Treat is equal to 1 if the firm is located in the pilot zone, which includes Pearl River Delta, Beijing, Tianjin, Hebei, Shanghai, Chongqing, Henan, Inner Mongolia, Guizhou, and Shenyang, indicating the treatment group; otherwise, Treat is 0, indicating the control group.

According to the existing literature (Hu et al., 2023; Zhang, 2023), we control for several variables including firm size (Size), asset-liability ratio (Leverage), the growth rate of operating income (Growth), return on common stockholders' equity (ROE), operating cashflow (Cashflow), current ratio (Liquidity), fixed asset ratio (PPE), firm age (AGE), the proportion of independent directors (Director), proportion of executives' holdings (Holdings), and proportion of the largest shareholder's holdings (First). Table 1 presents the variable definitions.

Descriptive statistics. Table 2 reports the descriptive statistics of the main variables. For GW , the mean is -0.0177 , which indicates that sample firms engage in low-level greenwashing on average. The values range from a minimum of -3.6927 to a maximum of 3.4330 , with a standard deviation of 1.2000 , which suggests that greenwashing varies widely among the sample firms. The mean of Post×Treat is 0.3058 , which shows that 30.58% of the sample firms are affected by the establishment of the pilot zones. Additionally, the mean, standard deviation, minimum, median, and maximum of the control variables are generally consistent with those reported in the existing literature, which supports the reliability of our sample.

Empirical specification and results

Baseline specification. Utilizing the 2016 establishment of China's national big data pilot zone as a policy shock, we use the difference-in-differences method to examine how big data affects firm greenwashing. Compared to the classic difference-in-differences model, the difference-in-differences model with firm and year fixed effects is more precise in estimation and inference, so we set firm and year fixed effects. The model is specified as

follows.

$$GW_{i,t} = \beta_0 + \beta_1 Post_t \times Treat_i + \sum \beta_i Controls_{i,t} + \mu_i + \tau_t + \varepsilon_{i,t} \quad (2)$$

Where i denotes firm, t refers to year. GW is the dependent variable that represents firm greenwashing. $Post$ is the dummy variable for the establishment time of the pilot zone. $Treat$ is the dummy variable for a firm located in the pilot zone. We are most interested in the variable $Post \times Treat$, as the sign, magnitude, and statistical significance of the coefficient β_1 capture the effect of big data on firm greenwashing. $Controls$ represent a vector of control variables including $Size$, $Leverage$, $Growth$, ROE , $Cashflow$, $Liquidity$, PPE , AGE , $Director$, $Holdings$, and $First$. μ_i and τ_t represent firm-fixed effects and year-fixed effects, respectively. To control for the time-varying characteristics of different industries and regions, we also include the interaction fixed effects between industry and year ($Industry \times Year$), and the interaction fixed effects between province and year ($Province \times Year$).

The results of the baseline regression. Table 3 reports the results of the baseline. Column (1) presents the result controlling only for fixed effects for Firm, Year, $Industry \times Year$, and

Table 1 Variable definitions.	
Variable	Variable definitions
GW	The calculation obtained by Eq. (1)
Post	If the year is 2016 or later, the value is 1 and 0 otherwise
Treat	If the firm is located in the pilot zone, the value is 1 and 0 otherwise
Post \times Treat	Post multiplied by Treat
Size	The natural logarithm of the total assets
Leverage	Total liabilities divided by total assets
Growth	Operating income divided by the previous year's operating income minus 1
ROE	Net profit divided by common stockholders' equity
Cashflow	Net operating cash flow divided by total assets
Liquidity	Current assets divided by current liabilities
PPE	Fixed asset divided by total assets
AGE	The natural logarithm of the number of years since the enterprise's establishment
Director	Number of independent directors divided by the number of board members
Holdings	Number of shares held by executives divided by the number of common shares
First	Number of shares held by the largest shareholder divided by the number of common shares

$Province \times Year$. The coefficient on $Post \times Treat$ is -0.8214 and is significant at the 1% level. Column (2) shows that after adding controls for firm-level characteristics, the coefficient on $Post \times Treat$ is -0.8130 and significant at the 1% level. Column (3) suggests that further adding controls for corporate governance characteristics, the coefficient on $Post \times Treat$ is -0.8187 and significant at the 1% level. These empirical results indicate that big data significantly inhibits firm greenwashing, and $H1_0$ is statistically tested at conventional levels.

Table 3 The results of the baseline regression.

Variables	(1) GW	(2) GW	(3) GW
Post \times Treat	-0.8214^{***} (0.2893)	-0.8130^{***} (0.2912)	-0.8187^{***} (0.2923)
Size		0.0512^* (0.0281)	0.0593^{**} (0.0283)
Leverage		-0.0766 (0.1351)	-0.0848 (0.1351)
Growth		0.0105 (0.0143)	0.0100 (0.0143)
ROE		0.2061^{**} (0.0875)	0.1899^{**} (0.0883)
Cashflow		0.0766 (0.1665)	0.0856 (0.1667)
Liquidity		0.0196^* (0.0104)	0.0184^* (0.0103)
PPE		0.0993 (0.1531)	0.0974 (0.1532)
AGE		-0.1683 (0.1828)	-0.0857 (0.1851)
Director			-0.0187 (0.1526)
Holdings			0.6318^{***} (0.2251)
First			0.0757 (0.1749)
Firm/Year/ Industry \times Year /Province \times Year	Yes	Yes	Yes
Constant	0.2336^{***} (0.0890)	-0.5211 (0.8185)	-0.9902 (0.8438)
N	10,509	10,509	10509
Adj. R ²	0.5134	0.5139	0.5142

Numbers in parentheses are robust standard errors. ***, ** and * are significant 1%, 5%, and 10% levels, respectively.

Table 2 Descriptive statistics.

Variables	Observations	Mean	Standard deviation	Minimum	Median	Maximum
GW	10,509	-0.0177	1.2000	-3.6927	-0.0499	3.4330
Post \times Treat	10,509	0.3058	0.4608	0.0000	0.0000	1.0000
Size	10,509	23.2464	1.3254	19.5411	23.1326	27.0484
Leverage	10,509	0.4743	0.1980	0.0156	0.4867	0.8825
Growth	10,509	0.3362	0.8654	-2.0377	0.1218	5.7912
ROE	10,509	0.0772	0.1230	-0.6169	0.0821	0.3630
Cashflow	10,509	0.0613	0.0708	-0.4208	0.0580	0.2640
Liquidity	10,509	1.9455	1.8291	0.1305	1.4289	12.5305
PPE	10,509	0.2283	0.1774	0.0019	0.1846	0.7251
AGE	10,509	2.8981	0.3487	0.6931	2.9444	3.9890
Director	10,509	0.3792	0.0746	0.2000	0.3636	0.6000
Holdings	10,509	0.0471	0.1139	0.0000	0.0000	0.5509
First	10,509	0.3709	0.1613	0.0408	0.3566	0.7732

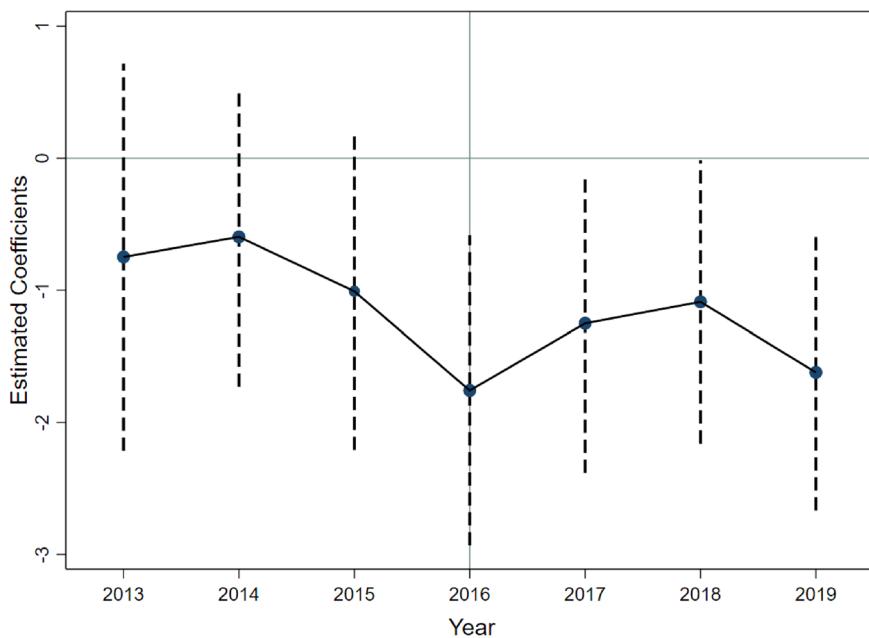


Fig. 1 Parallel trend test.

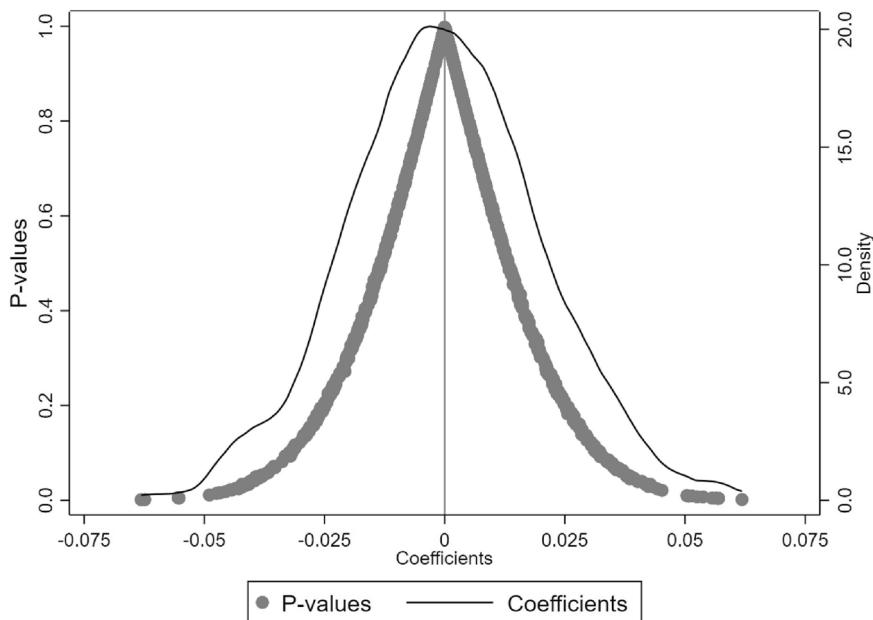


Fig. 2 Placebo test.

Robustness checks. We perform a series of checks to validate the robustness of the baseline regression results.

Parallel trend test. A fundamental assumption of the difference-in-differences design is that the treatment and the control group should exhibit the same trend before policy implementation. Following the approach of Liu and Qiu (2016), we specify model (3) to test the parallel trend.

$$GW_{i,t} = \beta_0 + \sum_{t=2013}^{2019} \beta_t Year_t \times Treat_i + \beta_1 Controls + \mu_j + \tau_t + \varepsilon_{i,t} \quad (3)$$

Where, we introduce interactions between each year and Treat, using 2012 as the base period. The β_t denotes estimated

coefficients for Year \times Treat from 2013 to 2019. The definitions of all other variables remain consistent with those in model (2). As shown in Fig. 1, the β_t from 2013 to 2015 is not significant at conventional levels, which indicates that corporate greenwashing between the treatment and control groups has no statistically significant differences. This finding supports the parallel trend assumption before the establishment of pilot zones. In addition, the β_t from 2016 to 2019 is significant at least at the 10% level, which demonstrates that the establishment of pilot zones has a significant inhibitory effect on firm greenwashing.

Placebo test. To examine whether the baseline finding is affected by unobserved factors, we randomly construct a false treatment group and repeat the procedure 1000 times. Figure 2 presents a distribution plot of false β_1 and their corresponding p -values. The

Table 4 The results of additional robustness checks.

Variables	PSM	EBM	Re-measurement of greenwashing	Exclusion of computer-related firms
	(1) GW	(2) GW	(3) GW	(4) GW
Post×Treat	−0.8715*** (0.3192)	−0.8212*** (0.3047)	−0.7182** (0.3044)	−0.8239*** (0.2972)
Controls	Yes	Yes	Yes	Yes
Firm/Year/Industry×Year /Province×Year	Yes	Yes	Yes	Yes
N	7923	10,509	10,509	9600
Adj. R ²	0.5150	0.5211	0.4795	0.5027

Numbers in parentheses are robust standard errors. *** and ** indicate significance at the 1% and 5% levels, respectively.

false β_1 displays a normal distribution centered around 0, and the majority of p -values exceed 0.1, which indicates statistical insignificance at the 10% level. Therefore, other unobserved factors have no significant effect on the finding that big data inhibits firm greenwashing.

Additional robustness checks. To rectify sample selection bias arising from observable variables between the treatment and control groups, we apply the propensity score matching (PSM) method. Specifically, we use the control variables from the previous regression as covariates and perform the 1:2 nearest-neighbor procedure to match the two groups. We also perform the robustness test using the entropy balancing matching (EBM) method. As shown in columns (1) and (2) of Table 4, the regression results show that the coefficients of Post×Treat are −0.8715 and −0.8212, both significant at the 1% level. Moreover, we remeasure firm greenwashing based on Eq. (1) using disclosure scores of ESG and actual performance of ESG. Column (3) of Table 4 presents the empirical finding, which indicates that the coefficient of Post×Treat is −0.7182 and significant at the 5% level. Finally, given that computer information technology firms are directly related to big data industries, they might influence the baseline regression result. To accomplish this, we conduct the robustness test by excluding these sample companies.⁴ The result in column (4) of Table 4 suggests that the coefficient of Post×Treat remains significantly negative at the 1% level. Overall, the results of the robustness check consistently show that big data has a statistically significant inhibitory effect on company greenwashing.

Further analysis

Mechanism analysis. As mentioned above, the inhibitory effect of big data on firm greenwashing depends on whether or not big data effectively enhances corporate environmental performance and reduces information asymmetry between the firm and stakeholders. We empirically investigate the effect of big data on corporate environmental performance and information asymmetry.

Environmental performance. To examine the causal relationship between big data and corporate environmental performance, we use environmental tax (ET) and green innovation (GI) as proxies for firm environmental performance.⁵ Columns (1) and (2) of Table 5 present the empirical results, the coefficients of Post×Treat are −0.1090 and 0.3432, which are significant at least at the 10% level. These results suggest that big data significantly reduces environmental taxes and improves green innovation. As is evident, big data inhibits enterprise greenwashing by improving their environmental performance. Thus, H2 is statistically tested.

Table 5 The results of the mechanism analysis.

Variables	(1) ET	(2) GI	(3) DA	(4) ES
Post×Treat	−0.1090** (0.0519)	0.3432* (0.1997)	−0.0255* (0.0151)	−11.9693*** (3.6221)
Controls	Yes	Yes	Yes	Yes
Firm/Year/ Industry×Year /Province×Year	Yes	Yes	Yes	Yes
N	8204	10509	10026	8782
Adj. R ²	0.5665	0.7322	0.2003	0.4340

Numbers in parentheses are robust standard errors. **, *, and * are significant at 1%, 5%, and 10% levels, respectively.

Information asymmetry. We follow the literature (Dechow et al., 1995) and measure firm information asymmetry between the firm and stakeholders using the discretionary accruals (DA) estimated by the modified Jones model.⁶ Column (3) of Table 5 reports the result of the effect of big data on discretionary accruals, the coefficient of Post×Treat is −0.0255 and significant at the 10% level. Referring to existing literature (Bhattacharya et al., 2003; Francis et al., 2004), we also utilize earnings smoothing (ES) as a proxy for information asymmetry.⁷ The coefficient of Post×Treat in column (4) is −11.9693 and significant at the 1% level. These empirical results show that big data significantly reduces corporate discretionary accruals and earnings smoothing. The mitigation of information asymmetry as the channel through which big data inhibits corporate greenwashing is statistically tested at conventional levels, thus supporting H3.

Heterogeneity analysis. The above sections explore the effect of big data on corporate greenwashing and its underlying mechanism. We proceed to examine whether this causal effect exhibits heterogeneity across different ownership structures and its competitive intensity.

Ownership structure. Compared to non-state-owned enterprises, state-owned enterprises reflect government willingness and are subject to stricter regulatory oversight. Thus, state-owned enterprises pay more attention to their environmental legitimacy. When environmental performance is poor, state-owned enterprises are more likely to engage in greenwashing to present an environmentally friendly image (Cheng et al., 2017). Big data enables the government to monitor and assess the environmental performance of state-owned enterprises more accurately, thereby driving these enterprises to genuinely fulfill environmental governance activities. Therefore, we predict that big data significantly

Table 6 The results of heterogeneity analysis.

Variables	State-owned enterprises	Non-state-owned enterprises	High market competition	Low market competition
	(1) GW	(2) GW	(3) GW	(4) GW
Post×Treat	−1.1536*** (0.3341)	−0.9314 (0.5807)	−1.2087*** (0.3809)	−0.4360 (0.3668)
Controls	Yes	Yes	Yes	Yes
Firm/Year/Industry×Year /Province×Year	Yes	Yes	Yes	Yes
N	5391	4823	4948	5074
Adj. R ²	0.5101	0.5427	0.5443	0.5312

Numbers in parentheses are robust standard errors. *** is significant at the 1% level.

inhibits state-owned enterprise greenwashing. To test this, we follow the literature (Cheng et al., 2025) and divide the sample into state-owned enterprise and non-state-owned enterprise groups according to the ownership structure. Column (1) of Table 6 shows that the coefficient of Post×Treat for the state-owned enterprise group is −1.1536 and significant at the 1% level. The coefficient of Post×Treat in column (2) for the non-state-owned enterprise group is −0.9314, but it is not statistically significant at the conventional level. These results suggest that the inhibitory effect of big data on greenwashing is more prominent in the state-owned enterprise group.

Market competition. When faced with consumer demand for green products and competitors' pressure, firms are more motivated to claim environmental sustainability (Delmas and Burbano, 2011). Big data increases the risk of greenwashing being exposed, which reduces firms' competitive advantage, thereby driving them to improve their environmental performance. Compared to firms with low market competition, we anticipate that big data significantly inhibits the greenwashing of firms with high market competition. Following the approach of Aghion et al. (2005), we use the Lerner index to measure market competition and categorize firms into high and low market competition groups according to whether their Lerner index exceeds the industry median.⁸ The regression results in columns (3) and (4) of Table 6 indicate that the coefficients of Post×Treat are −1.2087 and −0.4360, but only the high market competition group is significant at the 1% level. It implies that the inhibitory effect is more pronounced in firms with high market competition.

Conclusions

In the era of the digital economy, it is crucial to examine the effect of big data on firm greenwashing and its underlying mechanism. We use the launch of China's national big data comprehensive pilot zone in 2016 as a policy shock to investigate its effect on corporate greenwashing. The empirical finding indicates that big data has a significant inhibitory effect on firm greenwashing, which still holds after conducting parallel trend tests, placebo tests, and additional robustness checks. The mechanism analysis demonstrates that improving environmental performance and reducing information asymmetry between the firm and stakeholders are underlying channels through which big data inhibits corporate greenwashing. Heterogeneity analysis across various types of enterprises shows that state-owned enterprises and those with high market competition are particularly predisposed to this inhibitory effect.

Our study is based on the Chinese scenario, where the design and implementation of the policy in the pilot zones are geographically constrained due to different market environments and regulatory policies, but the firms' responses to this policy are

genuine and representative of the firms' behavior. These findings suggest the following implications. On the one hand, the pursuit of green development goals should consider the inhibitory effect of big data on corporate greenwashing. The government should further update policy designs to support the extensive application of big data in green governance. This includes using big data to improve corporate environmental performance and break down environmental information barriers, thus effectively solving the green development challenges posed by greenwashing. On the other hand, big data development has created a favorable technological environment for firms' green governance. Firms should actively integrate big data into environmental protection efforts, environmental information processing, and disclosure, thereby promoting the coordinated development of digitalization and green transformation.

Data availability

All supporting data and codes are available in the supplementary material.

Received: 14 December 2024; Accepted: 22 May 2025;

Published online: 05 June 2025

Notes

1 Peer-relative greenwashing mitigates the adverse effects of methodological differences among rating agencies and industry heterogeneity on measurement accuracy (Yu et al., 2020).

2 The Bloomberg score is calculated based on the environmental information voluntarily disclosed by firms. Consequently, firms that disclose more environmental information tend to receive higher scores (Yu et al., 2020).

3 Huazheng rating employs AI, web scraping, semantic analysis, and integrated analysis to collect comprehensive environmental information on firms, which includes self-disclosed environmental information, media coverage of both positive and negative environmental events, and records of penalties from local regulatory authorities for environmental violations. With distinct advantages in localization, coverage, and timeliness, this rating objectively reflects firms' environmental performance.

4 We exclude the sample firms from the computer, communications, other electronic equipment manufacturing industries, the information transmission, software, and information technology services industries.

5 Environmental tax is measured as the natural logarithm of the sum of the environmental tax plus one, with pollution fees before 2018. The higher the value of this indicator, the poorer the enterprise's environmental performance. Green innovation is expressed as the natural logarithm of the sum of the green patent application numbers plus one.

6 The model is $\frac{TA_{i,t}}{Asset_{i,t-1}} = \alpha_0 \frac{1}{Asset_{i,t-1}} + \alpha_1 \frac{\Delta REV_{i,t} - \Delta REC_{i,t}}{Asset_{i,t-1}} + \alpha_2 \frac{PPE_{i,t}}{Asset_{i,t-1}} + \epsilon_{i,t}$, where TA is the company's total accrual profit, which is equal to the enterprise's operating profit minus net cash flow from operating. ΔREV is the firm's growth in operating income, ΔREC is the firm's increase in accounts receivable, PPE is the company's fixed assets, Asset is the company's total assets. The regression is conducted by year and industry according to the model, with the absolute value of the residuals representing discretionary accruals (DA). A higher value of DA indicates firms are more likely to manipulate earnings, thereby increasing information asymmetry (Dechow et al., 1995).

7 The calculation equation of earnings smoothing (ES) is

$$ES_{i,t} = SD(\frac{CFO_{i,t-3}}{Asset_{i,t-3}}, \frac{CFO_{i,t-2}}{Asset_{i,t-3}}, \frac{CFO_{i,t-1}}{Asset_{i,t-3}}) / SD(\frac{NI_{i,t-3}}{Asset_{i,t-4}}, \frac{NI_{i,t-2}}{Asset_{i,t-3}}, \frac{NI_{i,t-1}}{Asset_{i,t-2}}, \frac{NI_{i,t}}{Asset_{i,t-1}})$$
,
 where SD(.) denotes the standard deviation of the value within the parentheses, NI is the company's net profit, CFO is the company's net cash flow from operating, and Asset is the firm's total assets. The larger the ES, the more likely management is to intentionally mask earnings volatility to make false impressions of stable operations, thereby increasing information asymmetry (Bhattacharya et al., 2003; Francis et al., 2004).

8 The calculation equation of the Lerner index is

$$LI_{i,t} = (OR_{i,t} - OC_{i,t} - SE_{i,t} - AE_{i,t}) / OR_{i,t}$$
, where LI is the Lerner index, OR is the firm's operating revenue, SE is the company's selling expenses, and AE is the enterprise's administrative expense. The larger the Lerner index, the stronger the firm's pricing capacity and the higher its market competitive position.

References

Aghion P, Bloom N, Blundell R, Griffith R, Howitt P (2005) Competition and innovation: an inverted-U relationship. *Q J Econ* 120(2):701–728. <https://doi.org/10.1093/qje/120.2.701>

Bhattacharya U, Daouk H, Welker M (2003) The world price of earnings opacity. *Acc Rev* 78(3):641–678. <https://doi.org/10.2308/accr.2003.78.3.641>

Blazquez D, Domenech J (2018) Big data sources and methods for social and economic analyses. *Technol Forecast Soc* 130:99–113. <https://doi.org/10.1016/j.techfore.2017.07.027>

Chava S (2014) Environmental externalities and cost of capital. *Manag Sci* 60(9):2223–2247. <https://doi.org/10.1287/mnsc.2013.1863>

Chen H, Chiang RHL, Storey VC (2012) Business intelligence and analytics: from big data to big impact. *MIS Quart* 1165–1188. <https://doi.org/10.2307/41703503>

Chen L, Shen Q, Yu X, Chen X (2024) Knowledge spillovers along the sustainable supply chain of China's listed companies: the role of long-term orientation. *J Innov Knowl* 9(2):100478. <https://doi.org/10.1016/j.jik.2024.100478>

Cheng H, Cao A, Hong C, Liu D, Wang M (2025) Can green investors improve the quality of corporate environmental information disclosure? *Int Rev Econ Financ* 103901. <https://doi.org/10.1016/j.iref.2025.103901>

Cheng Z, Wang F, Keung C, Bai Y (2017) Will corporate political connection influence the environmental information disclosure level? Based on the panel data of A-shares from listed companies in Shanghai stock market. *J Bus Ethics* 143:209–221. <https://doi.org/10.1007/s10551-015-2776-0>

Ciampi F, Demi S, Magrini A, Marzi G, Papa A (2021) Exploring the impact of big data analytics capabilities on business model innovation: the mediating role of entrepreneurial orientation. *J Bus Res* 123:1–13. <https://doi.org/10.1016/j.jbusres.2020.09.023>

Clatworthy M, Jones MJ (2003) Financial reporting of good news and bad news: evidence from accounting narratives. *Acc Bus Res* 33(3):171–185. <https://doi.org/10.1080/00014788.2003.9729645>

Dechow PM, Sloan RG, Sweeney AP (1995) Detecting earnings management. *Acc Rev* 70(2):193–225. <https://www.jstor.org/stable/248303>

Delmas MA, Burbano VC (2011) The drivers of greenwashing. *Calif Manag Rev* 54(1):64–87. <https://doi.org/10.1525/cmr.2011.54.1.64>

Delmas M, Toffel MW (2004) Stakeholders and environmental management practices: an institutional framework. *Bus Strateg Environ* 13(4):209–222. <https://doi.org/10.1002/bse.409>

Fernando AG, Suganthi L, Sivakumaran B (2014) If you blog, will they follow? Using online media to set the agenda for consumer concerns on “green-washed” environmental claims. *J Advert* 43(2):167–180. <https://doi.org/10.1080/00913367.2013.841088>

Fitzgerald M, Kruschwitz N, Bonnet D, Welch M (2014) Embracing digital technology: a new strategic imperative. *MIT Sloan Manag Rev* 55(2):1

Francis J, LaFond R, Olsson PM, Schipper K (2004) Costs of equity and earnings attributes. *Acc Rev* 79(4):967–1010. <https://doi.org/10.2308/accr.2004.79.4.967>

Gao Q, Cheng C, Sun G (2023) Big data application, factor allocation, and green innovation in Chinese manufacturing enterprises. *Technol Forecast Soc* 192:122567. <https://doi.org/10.1016/j.techfore.2023.122567>

Gray WB (1987) The cost of regulation: OSHA, EPA and the productivity slowdown. *Am Econ Rev* 77(5):998–1006. <https://www.jstor.org/stable/1810223>

He G, Li Z, Yu L, Zhou Z (2025) Does commercial reform embracing digital technologies mitigate stock price crash risk? *J. Corp Financ* 102741. <https://doi.org/10.1016/j.jcorfin.2025.102741>

Hu S, Wang A, Du K (2023) Environmental tax reform and greenwashing: evidence from Chinese listed companies. *Energ Econ* 124:106873. <https://doi.org/10.1016/j.eneco.2023.106873>

Hu X, Hua R, Liu Q, Wang C (2023) The green fog: environmental rating disagreement and corporate greenwashing. *Pac -Basin Financ J* 78:101952. <https://doi.org/10.1016/j.pacfin.2023.101952>

Kuo L, Chen YV (2013) Is environmental disclosure an effective strategy on establishment of environmental legitimacy for organization? *Manag Decis* 51(7):1462–1487. <https://doi.org/10.1108/MD-06-2012-0395>

Li W, Li W, Seppänen V, Koivumäki T (2023) Effects of greenwashing on financial performance: moderation through local environmental regulation and media coverage. *Bus Strateg Environ* 32(1):820–841. <https://doi.org/10.1002/bse.3177>

Lin H, Zeng S, Wang L, Zou H, Ma H (2016) How does environmental irresponsibility impair corporate reputation? A multi-method investigation. *Corp Socl Resp Env Ma* 23(6):413–423. <https://doi.org/10.1002/csr.1387>

Liu Q, Qiu LD (2016) Intermediate input imports and innovations: evidence from Chinese firms' patent filings. *J Int Econ* 103:166–183. <https://doi.org/10.1016/j.intecon.2016.09.009>

McAfee A, Brynjolfsson E (2012) Big data: the management revolution. *Harv Bus Rev* 90(10):60–68

Meng XH, Zeng SX, Shi JJ, Qi GY, Zhang ZB (2014) The relationship between corporate environmental performance and environmental disclosure: an empirical study in China. *J Environ Manag* 145:357–367. <https://doi.org/10.1016/j.jenvman.2014.07.009>

Müller O, Fay M, Brocke JV (2018) The effect of big data and analytics on firm performance: an econometric analysis considering industry characteristics. *J Manag Inf Syst* 35(2):488–509. <https://doi.org/10.1080/07421222.2018.1451955>

Oussous A, Benjelloun FZ, Lahcen AA, Belfkih S (2018) Big data technologies: a survey. *J King Saud Univ -Com* 30(4):431–448. <https://doi.org/10.1016/j.jksuci.2017.06.001>

Palmer K, Oates WE, Portney PR (1995) Tightening environmental standards: the benefit-cost or the no-cost paradigm? *J Econ Perspect* 9(4):119–132. <https://doi.org/10.1257/jep.9.4.119>

Raut RD, Mangla SK, Narwane VS, Gardas BB, Priyadarshinee P, Narkhede BE (2019) Linking big data analytics and operational sustainability practices for sustainable business management. *J Clean Prod* 224:10–24. <https://doi.org/10.1016/j.jclepro.2019.03.181>

Seele P, Gatti L (2017) Greenwashing revisited: in search of a typology and accusation-based definition incorporating legitimacy strategies. *Bus Strateg Environ* 26(2):239–252. <https://doi.org/10.1002/bse.1912>

Shimshack JP, Ward MB (2005) Regulator reputation, enforcement, and environmental compliance. *J Environ Econ Manag* 50(3):519–540. <https://doi.org/10.1016/j.jeem.2005.02.002>

Troilo G, De Luca LM, Guenzi P (2017) Linking data-rich environments with service innovation in incumbent firms: a conceptual framework and research propositions. *J Prod Innov Manag* 34(5):617–639. <https://doi.org/10.1111/jpim.12395>

Vitari C, Raguseo E (2020) Big data analytics business value and firm performance: linking with environmental context. *Int J Prod Res* 58(18):5456–5476. <https://doi.org/10.1080/00207543.2019.1660822>

Wamba SF, Akter S, Edwards A, Chopin G, Gnanzou D (2015) How 'big data' can make big impact: findings from a systematic review and a longitudinal case study. *Int J Prod Econ* 165:234–246. <https://doi.org/10.1016/j.ijpe.2014.12.031>

Wamba SF, Dubey R, Gunasekaran A, Akter S (2020) The performance effects of big data analytics and supply chain ambidexterity: the moderating effect of environmental dynamism. *Int J Prod Econ* 222:107498. <https://doi.org/10.1016/j.ijpe.2019.09.019>

Wang S, Lai Y, Zhang S (2024) Greening by digitization? Exploring the effect of enterprise digital transformation on greenwashing. *Bus Strateg Environ* 33:6616–6639. <https://doi.org/10.1002/bse.3825>

Wu J, Guo S, Li J, Zeng D (2016) Big data meet green challenges: big data toward green applications. *IEEE Syst J* 10(3):888–900. <https://doi.org/10.1109/JSYST.2016.2550530>

Wu L, Hitt L, Lou B (2020) Data analytics, innovation, and firm productivity. *Manag Sci* 66(5):2017–2039. <https://doi.org/10.1287/mnsc.2018.3281>

Wu Y, Zhang K, Xie J (2020) Bad greenwashing, good greenwashing: corporate social responsibility and information transparency. *Manag Sci* 66(7):3095–3112. <https://doi.org/10.1287/mnsc.2019.3340>

Xie J, Chen L, Liu Y, Wang S (2023) Does fintech inhibit corporate greenwashing behavior? Evidence from China. *Financ Res Lett* 55:104002. <https://doi.org/10.1016/j.frl.2023.104002>

Xu T, Sun Y, He W (2024) Government digitalization and corporate greenwashing. *J Clean Prod* 452:142015. <https://doi.org/10.1016/j.jclepro.2024.142015>

Yaqoob I, Hashem IAT, Gani A, Mokhtar S, Ahmed E, Anuar NB, Vasilakos AV (2016) Big data: from beginning to future. *Int J Inf Manag* 36(6):1231–1247. <https://doi.org/10.1016/j.ijinfomgt.2016.07.009>

Yin L, Yang Y (2024) How does digital finance influence corporate greenwashing behavior? *Int Rev Econ Financ* 93:359–373. <https://doi.org/10.1016/j.ijref.2024.05.014>

Yu EP, Luu BV, Chen CH (2020) Greenwashing in environmental, social and governance disclosures. *Res Int Bus Financ* 52:101192. <https://doi.org/10.1016/j.ribaf.2020.101192>

Zhang D (2023) Subsidy expiration and greenwashing decision: is there a role of bankruptcy risk? *Energ Econ* 118:106530. <https://doi.org/10.1016/j.eneco.2023.106530>

Zhang D (2024) The pathway to curb greenwashing in sustainable growth: the role of artificial intelligence. *Energ Econ* 133:107562. <https://doi.org/10.1016/j.eneco.2024.107562>

Zhang X (2024) Can retail investor activism inhibit corporate greenwashing behavior: evidence from investor interactive platforms in China. *J Clean Prod* 461:142617. <https://doi.org/10.1016/j.jclepro.2024.142617>

Zhou K, Qu Z, Liang J, Tao Y, Zhu M (2024) Threat or shield: environmental administrative penalties and corporate greenwashing. *Financ Res Lett* 61:105031. <https://doi.org/10.1016/j.frl.2024.105031>

Author contributions

JS provided overall guidance and revised the manuscript, NZ wrote the main manuscript text, and CZ contributed to data collection. All authors reviewed the manuscript.

Competing interests

The authors declare no competing interests.

Ethical approval

This article does not contain any studies with human participants performed by any of the authors.

Informed consent

This article does not contain any studies with human participants performed by any of the authors.

Additional information

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1057/s41599-025-05107-0>.

Correspondence and requests for materials should be addressed to Nannan Zu.

Reprints and permission information is available at <http://www.nature.com/reprints>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2025