China’s environmental policy intensity for 1978–2019

Guoxing Zhang1,2✉, Yang Gao1, Jiexun Li3, Bin Su4,5✉, Zhanglei Chen1 & Weichun Lin1

Improving the measurement of environmental policy intensity would affect not only the selection of variables in environmental policy research but also the research conclusions when evaluating policy effects. Because direct evaluation is lacking, the existing research usually applies data such as pollutant emission data, or the number of policies to construct proxy variables. However, these proxy variables are affected by many assumptions and different selection criteria, and they are inevitably accompanied by endogeneity problems. In this study, China’s environmental policy is comprehensively collected for the first time, and a machine learning algorithm is applied to evaluate the policy intensity. We provide all the policies issued by the Chinese government from 1978 to 2019 and the quantified intensity for each policy. We also distinguish all policies into three types according to their attributes. This dataset can help researchers to further understand China’s environmental policy system. In addition, it provides a valuable dataset for related research on evaluating environmental policy and recommending actions for further improvement.

Background & Summary
Environmental issues have become a global challenge that threatens the health and livelihood of humans and all living creatures on earth. In particular, during decades of rapid industrialization, China has been suffering from an environmental crisis, including air pollution, water scarcity, etc. These issues, in return, have limited the sustainability of economic and social development in China1. Over recent decades, China’s fight against environmental problems has advanced through the implementation of a series of environmental policies. These policies have achieved significant successes as indicated by the reduction in energy intensity and pollutant emissions. However, which policies or which parts of the policies have led to success? It would be of tremendous value to study the relationships between environmental policies and their outcomes: this would require a quantitative analysis of policies.

Quantitative analyses of environmental policies directly affect the assessment of environmental policy outcomes, providing the ability to continuously monitor policy effectiveness and improve the environmental policy system. China has proposed the goals of “carbon peaking” and “carbon neutral”, which will inevitably set higher requirements for future environmental policies and lead to increasingly complex policies being promulgated2,3. With the emergence of big data and large samples, it is essential that suitable ideas and methods be found to improve quantitative research on policies.

Policy instruments are described as “building blocks” to transfer the rather abstract principles and rules into concrete and substantive action4. In the research on policy change5,6, policy outcomes7,8, policy outputs9, policy mix10, many researchers have focused on the importance of policy instruments. For example, Knill et al.8 proposed measuring policy intensity by six indicators: objectives, scope, integration, budget, implementation, and monitoring. These indicators form a content-based coding procedure that allows the systematic assessment of policy intensity over time and space as well as across policy fields10.

Policy intensity has also emerged as a powerful tool for quantitative policy research. Policy intensity is an index that weights policy instruments according to measures such as whether the instrument has measurable targets, designated budgets, clear objectives and timelines; its integration with larger policy initiatives; and the enactment of policy monitoring. In the literature, there are several related concepts, such as “importance”, “significance”, or “stringency”11–13. For example, the Organization for Economic Co-operation and Development

1School of Management, Lanzhou University, Lanzhou, 730000, China. 2Institute of Green Finance, Lanzhou University, Lanzhou, 730000, China. 3Department of Decision Sciences, Western Washington University, Bellingham, WA, 98225, USA. 4Energy Studies Institute, National University of Singapore, Singapore, 119620, Singapore. 5Department of Industrial Systems Engineering and Management, National University of Singapore, Singapore, Singapore. ✉e-mail: guoxingzh@lzu.edu.cn; subin@nus.edu.sg
OECD) has developed an index of environmental policy stringency. Based on these definitions, policy intensity in this paper is the degree of stringency in the delivery and output of a policy document. The higher the policy intensity of a document is, the more stringency there is confronting the policy stakeholders. Quantitative analysis of policy often involves a significant amount of manual reviewing, measuring and annotating comprehensive indices from policy text. For environmental policy, empirical studies rely on a variety of regulatory impact data. In recent years, a growing number of studies have suggested the use of policy output data to assess the influence in a more direct fashion. Terms such as policy strength, policy objectives, policy measures have gradually attracted more attention. However, the processes of collecting and coding policy instruments are not sufficiently systematic and are somewhat fragmented. The manual process is often tedious and time consuming. It is challenging to ensure the efficiency and accuracy of quantification methods when facing a large collection of policy documents. More importantly, policy assessment is heavily dependent on policy resources. To the best of our knowledge, there has been no study that systematically organizes and discloses a comprehensive collection of environmental policy data in China.

To fill those gaps, we build a comprehensive collection of environmental policies in China over 40 years and develop a novel and systematic method to assess the policy intensity quantitatively. This indicator provides an excellent way to understand and appreciate China's environmental policy system. The quantitative results would help researchers to systematically observe the change and development of China's environmental policies over time. In addition, our dataset provides a detailed inventory of environmental policies, allowing researchers to reduce the labour and time costs of collecting environmental policies in China. Our dataset will facilitate the advancement of environmental policy-related research, and researchers can use it further to develop a systematic assessment of China's environmental policies.

Methods
An overview of our methods is shown in Fig. 1. The research framework consists of four modules: Manual quantification, Text data preparation, Modeling, and Validation.

**Data collection.** We first collect environmental policies from the Global Legal and Regulatory Network (http://policy.mofcom.gov.cn/), China Legal Resources Database (http://www.lawyee.org), Wanfang Database (http://c.g.wanfangdata.com.cn), China National Knowledge Infrastructure (CNKI, https://www.cnki.net/),

![Fig. 1 Research framework. (1) Manual Quantification: we first manually read and score each policy. Then, we categorize policies into different types according to their attributes. (2) Text Data Preparation: We analyze the text content by a series of text-mining operators, e.g., cleaning, tokenization, stop-word removal, word frequency analysis, and feature extraction from the text. The specific lexicon of environmental policy is screened and constructed. (3) Modeling: We deploy a series of prediction models to quantify the intensity of environmental policies. (4) Validation: To validate the credibility of the results, we evaluate and compare models based on their performance and key features.](http://www.nature.com/scientificdata)
**Manual quantification.** The manual quantification of environmental policy intensity mainly involved combinations with pre-set dimensions, such as policy objectives and policy measures. Through an interpretation of the policy text considering the enforceability and content details of different policy measures and objectives, each policy is rated on a scale of 1 to 5 in terms of its intensity. Policy measures include personnel measures, administrative measures, fiscal and tax measures, financial measures, guiding measures, and other economic measures. Policy objectives include preventing and controlling pollution, improving the effectiveness of energy conservation and emission reduction, promoting industrial upgrading, improving energy use efficiency, optimizing the energy consumption structure, and promoting the technological transformation of energy conservation and emission reduction.

We trained a group of personnel to manually read and rate each policy for the intensity of measures and objectives. Each policy was rated by multiple raters independently and validated for interrater reliability. The ratings not only reflect the degree to which the policy emphasizes certain measures or objectives but also to a certain extent solve the problem of weight selection in the process of constructing indicators. The formula for calculating the intensity of environmental policy is as follows:

\[
ERI = \frac{\sum_{k=1}^{n} M_k O_{kn}}{n}
\]

where \(t\) represents the year, and \(i\) represents a policy. \(M_{ik}\) is the sum of the intensity of \(k\) policy measures in a certain year, and \(O_{kn}\) is the sum of the intensity of \(n\) policy objectives in a certain year.

These environmental regulation intensity (ERI) scores rated by our personnel can serve as a valuable and reliable resource for future research related to China’s environmental policy. Furthermore, these ERI scores, along with the policy text, can be used to train a machine learning model that can estimate the intensity of future environmental policies.

**Environmental policy types.** The policy instrument is the “carrier” of policy, a channel through which policy science researchers can study the main content of policy, the policy formulation process and the policy tools, and it serves as an objective, accessible, traceable written record of the policy system and policy process. Most research on environmental policy focuses on policy tools, policy types, and game behaviour among the central bodies of governments at all levels in implementation. For example, the World Bank divides environmental policy tools into four types: environmental regulations, market application, market creation, and public participation. Based on the driving mechanisms, environmental policy can be divided into command-control policy and market-based policy.

In this paper, we followed related research and divided environmental policy into three types: command-control environmental policy (CCEP), market-based environmental policy (MBEP) and public participation environmental policy (PPEP). CCEP and MBEP have been extensively investigated in the literature. The main characteristic of CCEP is that it is mandatory, and thus it relies on administrative instruments such as certain types of governance and standards. By contrast, MBEP mainly uses market-based instruments for environmental governance, such as fiscal policies related to environmental governance, emissions fees, emissions trading, product promotion catalogues, etc.

PPEP refers to the public and private sectors that have the right to participate in environmental protection. The Environmental Protection Act implemented in 2015 stipulated the principle of “public participation” in the “General Regulations” section: “All organizations and individuals have the obligation to protect the environment and have the right to report and accuse the organizations and individuals who pollute or destroy the

| Departments that promulgate environmental policies (list only a portion of all departments) |
|-----------------------------------|-------------------------------------|
| The National People’s Congress     | The Ministry of Supervision         |
| The State Council                 | The Ministry of Agriculture         |
| The Ministry of Ecological Environ| The Forestry Bureau                 |
| The National Development and Refo| The State Administration of Taxation|
| The Ministry of Finance           | The China Banking Regulatory Commi|
| The Ministry of Transport         | The Electricity Regulatory Commission|
| The Ministry of Industry and Infor| The State Administration for Industry and Commerce |
| The Ministry of Housing and Urban-Rural Development | The State Council Administration of Organ Affairs |
| The Ministry of Science and Techn| The Ministry of Education           |

**Table 1.** List of departments that have promulgated environmental policies.
Therefore, we consider an environmental policy in the category of public-participation if it contains terms such as “public participation,” “citizens,” “opinions,” “seek advice,” “supervision,” “hearing,” “argument,” and “subject declaration.”

Pre-processing of the text data. The encoded information in the text is enriched and supplemented in traditional research with structured data. In recent years, there have been many studies based on text, such as financial news, social media, speeches by political and corporate documents, and so on. The most widely used method in text analysis is the bag-of-words model based on the document-word matrix. Researchers have discussed the practical application effects of various word lists and specific words in the field. However, as used method in text analysis is the bag-of-words model based on the document-word matrix. Researchers have pointed out, many studies rely on classification dictionaries derived from other disciplines, and such applications may produce false results. The selection or construction of a lexicon suitable for a specific field presents a problem that needs to be fully considered before research can commence.

To avoid this problem, a specific lexicon applicable to environmental policies was constructed during the processing of textual data. The text preprocessing process in this paper includes several steps, such as text cleaning, tokenization, and word frequency statistics. Since the term-document matrix contains all the words that appear in the policy documents and many words have little importance to policy intensity, it is necessary to screen the words before forming a specific lexicon:

- Words that appear in more than 99% and no more than 1% of the documents are screened and deleted;
- The correlation coefficients between word frequency and policy intensity (manual quantification intensity) are calculated;
- Based on the results of the coefficients, words with high correlation (>0.1) with the intensity are selected;
- The selected word is divided into two aspects, policy objectives and policy measures, and the words representing objectives and measures were not repeated;
- Words that are not substantive, such as the names of people, places, departments, etc., are removed.

Machine learning methods. In the literature, the use of text language to calculate indicators is gradually increasing. The combination of this method and machine learning tools also enables the testing and verification of corresponding prediction scores based on specific structures. For example, Harrison et al. introduced a new language-based method to measure the five personality traits of CEOs. We first use the text analysis method to construct a specific lexicon suitable for environmental policy. Second, we construct a model to explore the relationship between the lexicon and policy intensity. Because of the nature of the task, traditional measurement methods and machine learning algorithms can both be applied. In this paper, the traditional linear regression model is selected, and eight groups of algorithm models are selected, including Ridge (Ridge), Lasso (Lasso), robust linear model (RLM), partial least squares (PLS), generalized linear model (GLM), support vector machine (SVM), eXtreme gradient boosting (XgbLinear), and random forest (RF).

Ridge and Lasso are shrinkage methods for prediction models and are particularly useful for datasets with a large number of explanatory variables. RLM and GLM relax the conditions of least squares regression and are better able to handle outliers or external variances in the data. PLS projects prediction variables and observation variables to a new space to find a linear regression; this method is suitable for the linear regression task with high-dimensional data. SVM is also suitable for dealing with high-dimensional data and finding fitted curves. XgbLinear and RF are mainly tree-based methods. When out-of-sample prediction ability is essential, these are both prevalent and effective methods for flexibly estimating regression functions. Several of the representative algorithms selected have also achieved good performance in other fields, and can thus well support this research.

Data Records
Our dataset contains a total of 1,912 environmental policies from 1978 to 2019, along with their intensity scores quantified by experts and machine-learning models. In particular, the dataset includes 1,391 CCEP policies, 292 MBEP policies, and 229 PPEP policies. All data records have been uploaded to the public data repository Figshare, specifically including the following:

- Dataset of environmental policy intensity measurement results for China (1978–2019) ["China’s environmental policies intensity, 1978–2019"];
- Dataset of the Chinese environmental policy lexicon ["Featured words in China's environmental policies"];
- Dataset of the importance of key variables ["Key variables importance"].

Environmental policy intensity. This dataset provides the results over 40 years of China’s environmental policy intensity from Reform and Opening Up (1978) to 2019. Specifically, this dataset includes all environmental policies issued by the Chinese government every year and their corresponding policy intensity. The policy number (e.g., 2019–086) in the dataset reflects the number of environmental policies issued each year. Figure 2a–d show a comparison of the three policy types in terms of their policy count, mean and trend in the distribution of policy intensity by year, respectively. As shown in Fig. 2a, China has implemented an increasing number of environmental policies across all three types, especially during the last ten years, and the trend appears to be slowing down. CCEP (red) is still the predominant type of environmental policy in China.
As shown in Fig. 2b–d, the yearly variation in policy intensity is significant. Environmental policy intensity in earlier years shows greater volatility, and the degree of volatility decreases over time. In terms of the average trend in policy intensity, all policies have shown an upward trend in recent years. This trend reflects that China has been reinforcing its regulations to fight environmental issues via various types of policy. The difference between different types of policy is also becoming more visible, which reflects that China is adopting a differentiated strategy in the process of environmental governance.

**Fig. 2** The distributions of policy types and policy intensity. The four subplots show: (a) the distribution of the three policy types, (b) box plots of CCEP policy intensity, (c) box plots of MBEP policy intensity, and (d) box plots of PPEP policy intensity.
Environmental policy lexicon. We create a lexicon for quantifying the intensity of environmental policy in China that is included in the dataset. Specifically, we divide the words in this lexicon into two categories: policy objectives and policy measures. We also construct detailed subcategories under the objectives and measures categories. Table 2 provides the number of words in these categories.

Technical Validation

Data retrieval and collection. In this research, we build a comprehensive collection of environmental policies issued by multiple levels of governments and ministries in China from 1978 to 2019 by retrieving policy documents from different official databases and websites. This process was carefully designed and executed to ensure comprehensive coverage and integrity in the data collection. Next, by carefully reading the policy text, we sort out those that are relevant to environmental regulation, which ensures the accuracy of data collection.

Model performance. In this research, we empirically train and test learning-based models that allow automatic estimation of the intensity of environmental policy in China (from 1978 to 2016). To allow comparison, our experiments include several popular statistical learning algorithms, from simple linear regression models (LM) and regularization techniques (Ridge regression and LASSO) to machine learning models (e.g., SVM and random forest). For evaluation, we randomly split the entire data collection into a training set (75%) and a test set (25%). We apply each learning algorithm to the training set and estimate the training error using 10-fold cross validation. Then, we apply the trained prediction model to the test set and estimate the testing error. Table 3 summarizes the training and testing errors for the different prediction models. Among all models, RF achieved the best performance in terms of root mean squared error (RMSE) for both the training and test sets. Therefore, we choose RF to measure policy intensity.

Identifying the key features. After showing that learning-based models can estimate the intensity of environmental policy, we assess the importance of features and identify the key indicators of policy intensity. The inquiry results based on the importance of variables can help to validate and interpret the prediction models. For evaluating variable importance, there are two types of methods: model-specific and model-agnostic. Model-specific methods use certain elements of the model structure to evaluate the importance of variables. For example, in a linear regression model, the value of normalized coefficients and their corresponding p-values can be interpreted as reflecting the importance of variables, whereas in tree-based models, top decision nodes are considered more important variables for prediction. In contrast, without making any assumptions about the model, model-agnostic methods measure the impact on the model fitness when a variable (or a subset of

| Category          | Subcategory                                                                 | Number of words |
|-------------------|-----------------------------------------------------------------------------|-----------------|
| Policy Objectives | 1. preventing and controlling pollution                                      | 23              |
|                   | 2. improving the effectiveness of energy conservation and emission reduction | 30              |
|                   | 3. establishing awareness of energy conservation and emission reduction       | 19              |
|                   | 4. promoting industrial upgrading                                             | 39              |
|                   | 5. improving energy use efficiency                                            | 32              |
|                   | 6. optimizing the energy consumption structure                                 | 20              |
|                   | 7. promoting the technological transformation of energy conservation and emission reduction | 28              |
|                   | 8. comprehensive goals                                                        | 55              |
| Policy Measures   | 1. personnel measures                                                         | 22              |
|                   | 2. administrative measures                                                    | 67              |
|                   | 3. fiscal and tax measures                                                    | 16              |
|                   | 4. financial measures                                                         | 4               |
|                   | 5. guiding measures                                                           | 42              |
|                   | 6. other economic measures                                                    | 21              |
|                   | 7. measure behavior                                                          | 37              |

Table 2. Number of words in the lexicon for measuring the intensity of environmental policy.

| Training Set | LM  | Ridge | Lasso | RLM  | PLS  | GLM  | SVM  | XGBoost | RF    |
|--------------|-----|-------|-------|------|------|------|------|---------|-------|
| RMSE         | 72.98 | 51.78 | 46.03 | 71.17 | 47.05 | 45.82 | 47.23 | 42.03   | 38.71 |
| Testing Set  | 99.84 | 61.28 | 54.06 | 97.28 | 41.58 | 52.91 | 40.93 | 35.63   | 34.23 |

Table 3. Performance of each model in the training and test sets.
variables) is removed from the model. Suppose \( L = f(\tilde{y}, y) \) is the loss function of prediction model \( f(x) \). Then, the importance of variable \( x_i \) can be defined as:

\[
\text{vip}_{\text{Diff}}(x_i) = L^{x_i} - L
\]

(2)

or

\[
\text{vip}_{\text{Ratio}}(x_i) = \frac{L^{x_i}}{L}
\]

(3)

where \( L \) is the value of the loss function calculated by all variables \( x \), and \( L^{x_i} \) is that calculated by variables except \( x_i \).

In this research, for the different models we build for prediction, we use the model-agnostic method for evaluating the importance of variables. For each variable, we calculate the \( \text{vip} \) score by dropping it out of the model. “Data Record 3” summarizes the 20 most important variables in terms of contribution to the model.

In particular, for policy objectives, we can see that the variables relevant to energy and technology objectives play the most important roles, such as the optimization of the energy consumption structure (2 words belong to this category) and promoting the technological transformation of energy conservation and emission reduction (3 words belong to this category). For policy measures, administrative measures (3 words belong to this category), fiscal and tax measures (3 words belong to this category), and financial measures (2 words belong to this category) appear to play the most important roles.

Both policy objectives and policy measures have large impacts on policy intensity. The setting of objectives and corresponding measures are critical to environmental policymaking. Our results show that policies with more specific objectives and measures tend to have higher intensity. In particular, when the policy content involves details related to energy, technology, industry, fiscal and tax policy, finances, etc., the policy intensity is often higher.

To test the robustness of our findings, we used two additional metrics, i.e., mean square error (IncMSE) and node purity (IncNodePurity), to measure the importance of variables. Among the top 20 variables identified by IncMSE, nine belong to policy objectives: promoting industrial upgrades, optimizing the energy consumption structure, improving energy efficiency, establishing awareness of energy conservation and emission reduction, and comprehensive goals; and 11 variables belong to the policy measures, including measure behaviour, administrative measures, fiscal and tax measures, financial measures, and guiding measures. Among the top 20 variables identified by IncNodePurity, 14 belong to policy objectives, and six belong to policy measures. The results of IncMSE and IncNodePurity indicate that policy objectives have a more significant impact on policy intensity than measures. However, the specific objectives and measures, involving technological transformation, industrial upgrading, fiscal and tax measures, and others, continue to play an important role in policy intensity.

Comparison with existing databases. To our knowledge, our dataset is the first large collection of environmental policies with their intensity scores estimated by both human experts and learning-based models. To evaluate the degree of difference between the two results (manual quantification and machine quantification), we objectively compare the results of the policy intensity from 1978–2016. From the perspective of sequence similarity measured by the shortest distance, the basic standard is that the closer the distance is, the higher the similarity. The classical measurement methods generally include two types: lockstep measures and elasticity measures. Euclidean distance is a typical lockstep measure used to compare the point-to-point distance between time series, and standardized Euclidean distance is an improvement of Euclidean distance, as the latter is affected by the inconsistency of the multidimensional data scale. Elasticity measures show an improvement compared with lockstep measures, as they can compare two time series “one to one” or “one to many”. Dynamic time warping (DTW) is a typical representative elastic measure that can calculate the similarity between time series and is especially suitable for time series of different lengths and rhythms. Compared with the traditional Euclidean distance, this method can describe the similarity of time series more accurately.

We select two indicators, standardized Euclidean distance and DTW, to compare the two time series of policy intensity scores manually rated by experts and those estimated by the learning-based models. The results show that the standard Euclidean distance is 0.05, and the DTW distance is 5.59, which indicates that the policy intensity estimated by the model is highly consistent with the results of expert scoring.

By depicting the neat path that minimizes the DTW distance between sequences, the difference between the manually and machine-quantified intensity can be further judged (see Fig. 3). The blue line is the path with the lowest warping cost. When it coincides with the main diagonal or the gap is small, there can be considered to be high consistency between the two values. As seen from the results, the neat path of the model is relatively smooth, the cost of the neat path is small, and the neat path is parallel to the main diagonal. This shows that the intensity measured by the machine learning algorithms has a high degree of consistency with the results of manual quantification, reflecting the real changing trend in policy intensity.

Limitations and future work. We quantify the policy intensity from policy text, and the analysis process and results have several implications for the application of text analysis and machine learning methods in economics and management. Simultaneously, the environmental policy lexicon constructed by the text analysis method can provide a reference for the quantitative study of policy intensity. We also provide the necessary dataset for research on environmental policy evaluation, which can help for researchers to systematically understand and evaluate China’s environmental policy system and will promote related researches on environmental policy. This dataset can greatly reduce the difficulty for researchers to conduct research on China’s environmental policy.
and provide researchers with a booklet to find relevant policies conveniently and efficiently. Although our environmental policy dataset mainly includes the policies issued by the central government, it can be re-used and inspire further research in several aspects:

- First, by quantifying the policy texts, we introduce the concept of “policy intensity” that, rather than simply indicating the existence of a policy, allows measuring environmental policies with a scale. Several studies have demonstrated the power of analyzing policy texts for analyzing environmental, social, and economic impacts. Using our dataset and proposed methodology, policy analysts can study the evolution of policy intensity over time and gain insights into governments’ attitudes towards environmental issues at different time points. The comparison of the intensity of different policy types over years can also provide a basis for a more insightful analysis of environmental policies.

- Second, environmental policies are the documentation of the policy makers’ ideas and the basis of administration and execution. For their prior constraints and guidance nature, policies impose certain restrictions on stakeholders’ behaviors. While the policy intensity measured by the policy texts can indicate the strength of regulation, the quantification of the policy itself is only the starting point of evaluating environmental policy. At the end of the day, it is the execution of environmental policies that can directly determine the effects on the environment. Future research could look into regulatory actions in the policy execution process and measure the strength of these actions as well as their environmental impacts. By comparing the intensity of policy documents and their execution, we can gain more insights into the key factors to the success of environmental regulation.

- Third, from the perspective of local governments and industry, most environmental policies promulgated by provincial/local governments are largely aligned with those by the central government in China. Local policies often involve further refinement and enforcement of the policies from the above, without significant adjustments in essence. Hence, researchers can further investigate and design new comprehensive indicators of policy intensity, by combining our dataset with the strength of execution by local governments (e.g., the number of environmental enforcement officers, pollution emission monitoring level, or the degree of pollution control). Such an indicator could become an essential variable for evaluating the impact of environmental policies on company behaviors. In addition, researchers can apply our proposed analytical approach to collect and analyze more environmental policies from local governments or other countries and further expand the dataset.

Code availability
The relevant code data used for calculation and analysis in this paper can be obtained from the supplementary information.

Received: 8 October 2021; Accepted: 27 January 2022; Published: 11 March 2022

References
1. Zhou, Q., Zhang, X., Shao, Q. & Wang, X. The non-linear effect of environmental regulation on haze pollution: Empirical evidence for 277 Chinese cities during 2002–2010. *Journal of Environmental Management* **248**, 109274, https://doi.org/10.1016/j.jenvman.2019.109274 (2019).
2. Duan, H. et al. Assessing China’s efforts to pursue the 1.5 °C warming limit. *Science* **372**, 378–385, https://doi.org/10.1126/science.aba8767 (2021).
3. Mo, J. et al. The role of national carbon pricing in phasing out China’s coal power. *iScience* **24**, 102655, https://doi.org/10.1016/j.isci.2021.102655 (2021).
4. May, P. J. Policy design and implementation. *In Handbook of public administration* **223**, 233 (2003).
5. Howlett, M. & Cashore, B. The Dependent Variable Problem in the Study of Policy Change: Understanding Policy Change as a Methodological Problem. *Journal of Comparative Policy Analysis: Research and Practice* **11**, 33–46, https://doi.org/10.1080/13876980802648144 (2009).
6. Jones, B. D. & Baumgartner, E. R. From There to Here: Punctuated Equilibrium to the General Punctuation Thesis to a Theory of Government Information Processing. *Policy Studies Journal* **40**, 1–20, https://doi.org/10.1111/j.1541-0072.2011.00431.x (2012).
7. Jahn, D. & Kuitto, K. Taking stock of policy performance in Central and Eastern Europe: Policy outcomes between policy reform, transitional pressure and international influence. European Journal of Political Research 50, 719–748, https://doi.org/10.1111/1475-6765.2010.01981.x (2011).
8. Knill, C., Schulze, K. & Tosun, J. Regulatory policy outputs and impacts: Exploring a complex relationship. Regulation & Governance 6, 427–444, https://doi.org/10.1111/1748-9911.2012.01150.x (2012).
9. Schaffrin, A., Sewerin, S. & Seubert, S. Toward a Comparative Measure of Climate Policy Output. Policy Studies Journal 43, 257–282, https://doi.org/10.1111/psj.12095 (2015).
10. Schmidt, T. S. & Sewerin, S. Measuring the temporal dynamics of policy mix – An empirical analysis of renewable energy policy mixes’ balance and design features in nine countries. Research Policy 48, 103557, https://doi.org/10.1016/j.respol.2018.03.012 (2019).
11. Carley, S. & Miller, C. J. Regulatory Stringency and Policy Drivers: A Reassessment of Renewable Portfolio Standards. Policy Studies Journal 40, 730–756, https://doi.org/10.1111/1541-0072.2012.00471.x (2012).
12. Clinton, J. D. & Lapinski, J. S. Measuring Legislative Accomplishment, 1877–1994. American Journal of Political Science 50, 232–249, https://doi.org/10.1111/1540-5907.2006.00181.x (2006).
13. Grant, J. T. & Kelly, N. J. Legislative Productivity of the U.S. Congress, 1789–2004. Political Analysis 16, 303–323, https://doi.org/10.1093/pan/mpn035 (2008).
14. OECD. Environmental Policy Stringency index (Edition 2017). https://doi.org/10.1787/b4f0fdcc-en (2018).
15. OECD. How stringent are environmental policies?, https://www.oecd.org/economy/greeneconomy/how-stringent-are-environmental-policies.htm (2016).
16. Debrun, X., Moulin, L., Turriani, A., Ayuso-i-Casals, J. & Kumar, M. S. Tied to the mast? National fiscal rules in the European Union. Economic Policy 23, 298–362, https://doi.org/10.1111/eopi.12017 (2008).
17. Neumayer, E. Are left-wing party strength and corporatism good for the environment? Evidence from panel analysis of air pollution in OECD countries. The Economic Journal 128, 2755–2784, https://doi.org/10.1111/ejoc.12538 (2018).
18. Du, W. & Li. M. Assessing the impact of environmental regulation on pollution abatement and collaborative emissions reduction: Micro-evidence from Chinese industrial enterprises. Environmental Impact Assessment Review 82, 106382, https://doi.org/10.1016/j.eiar.2020.106382 (2020).
19. Zhao, J., Jiang, Q., Dong, X. & Dong, K. Would environmental regulation improve the greenhouse gas benefits of natural gas use? A Chinese case study. Energy Economics 87, 104712, https://doi.org/10.1016/j.eneco.2020.104712 (2020).
20. Zhou, Q., Song, Y., Wan, N. & Zhang, X. Non-linear effects of environmental regulation and innovation – Spatial interaction evidence from the Yangtze River Delta in China. Environmental Science & Policy 114, 263–274, https://doi.org/10.1016/j.envsci.2020.08.006 (2020).
21. Neumayer, E. Are left-wing party strength and corporatism good for the environment? Evidence from panel analysis of air pollution in OECD countries. Ecological Economics 45, 203–220, https://doi.org/10.1016/S0921-8009(03)00012-0 (2003).
22. Bernauer, T. & Koubi, V. Effects of political institutions on air quality. Ecological Economics 68, 1355–1365, https://doi.org/10.1016/j.ecolecon.2008.09.003 (2009).
23. Liefersink, D., Arts, B., Kamastra, J. & Ouwel, J. Leaders and laggards in environmental policy: a quantitative analysis of domestic policy outputs. Journal of European Public Policy 16, 677–700, https://doi.org/10.1080/13501760902983283 (2009).
24. Zhang, G., Zhang, P., Zhang, Z. G. & Li, J. Impact of environmental regulations on industrial structure upgrading: An empirical study on Beijing-Tianjin-Hebei region in China. Journal of Cleaner Production 238, 117,848, https://doi.org/10.1016/j.jclepro.2019.117848 (2019).
25. Zhang, G., Deng, N., Mou, H., Zhang, Z. G. & Chen, X. The impact of the policy and behavior of public participation on environmental governance performance: Empirical analysis based on provincial panel data in China. Energy Policy 129, 1347–1354, https://doi.org/10.1016/j.enpol.2019.03.030 (2019).
26. Tang, M., Li, X., Zhang, Y., Wu, Y. & Wu, B. From command-and-control to market-based environmental policies: Optimal transition timing and China’s heterogeneous environmental effectiveness. Economic Modelling 90, 1–10, https://doi.org/10.1016/j.econmod.2020.04.021 (2020).
27. Guo, R. & Yuan, Y. Different types of environmental regulations and heterogeneous influence on energy efficiency in the industrial sector: Evidence from Chinese provincial data. Energy Policy 145, 111,747, https://doi.org/10.1016/j.enpol.2020.111747 (2020).
28. Mou, H., Atkinson, M. M. & Tapp, S. Do Balanced Budget Laws Matter in Recession? Public Budgeting & Finance 38, 28–46, https://doi.org/10.1111/pbaf.12163 (2018).
29. Huang, C. Quantitative Research on Policy Literature. (Science Press, 2016).
30. Huang, C. et al. A bibliometric study of China’s science and technology policies: 1949–2010. Scientometrics 102, 1521–1539, https://doi.org/10.1007/s11192-014-1406-4 (2015).
31. Sheng, J., Zhou, W. & Zhu, B. The coordination of stakeholder interests in environmental regulation: Lessons from China’s environmental regulation policies from the perspective of the evolutionary game theory. Journal of Cleaner Production 249, 119385 (2020).
32. Sheng, J. & Webber, M. Incentive-compatible payments for watershed services along the Eastern Route of China’s South-North Water Transfer Project. Ecosystem Services 25, 213–226, https://doi.org/10.1016/j.ecoser.2017.04.006 (2017).
33. World Bank. Five Years after Rio Innovations in Environmental Policy. (1997).
34. Xie, R.-h., Yuan, Y.-j & Huang, J.-J. Different Types of Environmental Regulations and Heterogeneous Influence on “Green” Productivity: Evidence from China. Ecological Economics 132, 104–112, https://doi.org/10.1016/j.ecolecon.2016.10.019 (2017).
35. Li, H.-I., Zhu, X.-h., Chen, J.-y. & Jiang, F.-t Environmental regulations, environmental governance efficiency and the green transformation of China’s iron and steel enterprises. Ecological Economics 165, 106397, https://doi.org/10.1016/j.ecolecon.2019.106397 (2019).
36. Berman, E. & Bui, L. T. M. Environmental Regulation and Productivity: Evidence from Oil Refineries. The Review of Economics and Statistics 83, 498–510, https://doi.org/10.1162/0034653012480144 (2001).
37. Ryan, S. P. The Costs of Environmental Regulation in a Concentrated Industry. Econometrica 80, 1019–1061, https://doi.org/10.3982/ECTA6750 (2012).
38. Acemoglu, D., Aghion, P., Bursztyn, L. & Hemous, D. The Environment and Directed Technical Change. American Economic Review 102, 131–166, https://doi.org/10.1257/aer.102.1.131 (2012).
39. Acemoglu, D., Akcigit, U., Hanley, D. & Kerr, W. Transition to Clean Technology. Journal of Political Economy 124, 52–104, https://doi.org/10.1086/684511 (2016).
40. Aghion, P., Dechelepletr, A., Héamous, D., Martin, R. & Van Reenen, J. Carbon Taxes, Path Dependency, and Directed Technical Change: Evidence from the Auto Industry. Journal of Political Economy 124, 1–51, https://doi.org/10.1086/684581 (2016).
41. Sun, L., Zhu, D. & Chan, E. H. Public participation impact on environment Nimby conflict and environmental conflict management: Comparative analysis in Shanghai and Hong Kong. Land use policy 58, 208–217, https://doi.org/10.1016/j.landusepol.2016.07.025 (2016).
42. Jegadeesh, N. & Wu, D. Word power: A new approach for content analysis. Journal of Financial Economics 110, 712–729, https://doi.org/10.1016/j.jfineco.2013.08.018 (2013).

SCIENTIFIC DATA | (2022) 9:75 | https://doi.org/10.1038/s41597-022-01183-y
44. Loughran, T. I. M. & McDonald, B. When Is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks. *The Journal of Finance* **66**, 35–65, https://doi.org/10.1111/j.1540-6261.2010.01625.x (2011).

45. Baker, S. R., Bloom, N. & Davis, S. J. Measuring Economic Policy Uncertainty*. The Quarterly Journal of Economics** **131**, 1593–1636, https://doi.org/10.1093/qje/qjw024 (2016).

46. Gentzkow, M., Kelly, B. & Taddy, M. Text as Data. *Journal of Economic Literature*** **57**, 535–574, https://doi.org/10.1257/jel.20181020 (2019).

47. Loughran, T., McDonald, B. & Yun, H. A Wolf in Sheep’s Clothing: The Use of Ethics-Related Terms in 10-K Reports. *Journal of Business Ethics*** **89**, 39–49, https://doi.org/10.1007/s10551-008-9910-1 (2009).

48. Malhotra, S., Reus, T. H., Zhu, P. & Roelofsen, E. M. The Acquisitive Nature of Extraverted CEOs. *Administrative Science Quarterly*** **63**, 370–408, https://doi.org/10.1177/0001839217712240 (2017).

49. Gamache, D. L., McNamara, G., Mannor, M. J. & Johnson, R. E. Motivated to Acquire? The Impact of CEO Regulatory Focus on Firm Acquisitions. *Academy of Management Journal*** **58**, 1261–1282, https://doi.org/10.5465/amj.2013.0377 (2014).

50. Park, G. et al. Automatic personality assessment through social media language. *Journal of Personality and Social Psychology*** **108**, 934, https://doi.org/10.1037/pspp0000020 (2015).

51. Harrison, J. S., Thurgood, G. R., Boivie, S. & Pfarrr, M. D. Measuring CEO personality: Developing, validating, and testing a linguistic tool. *Strategic Management Journal*** **40**, 1316–1330, https://doi.org/10.1002/smj.3023 (2019).

52. Storm, H., Baylis, G., Mannor, K. & Heckelei, T. Machine learning in agricultural and applied economics. *European Review of Agricultural Economics*** **47**, 849–892, https://doi.org/10.1093/erae/jbz033 (2020).

53. Athey, S. & Imbens, G. W. Machine Learning Methods That Economists Should Know About. Annual *Review of Economics*** **11**, 683–723, https://doi.org/10.1146/annurev-economics-080217-053433 (2019).

54. Zhang, G. China’s environmental policy intensity for 1978–2019, figshare, https://doi.org/10.6084/m9.figshare.16740376.v1 (2022).

55. Buczek, P. & Burzykowski, T. *Explanatory Model Analysis*, 2020.

56. Bai, S., Qi, H.-D. & Xiu, N. Constrained Best Euclidean Distance Embedding on a Sphere: A Matrix Optimization Approach. *SIAM Journal on Optimization*** **25**, 439–467, https://doi.org/10.1137/13094918X (2015).

57. Berthold, M. & Höppner, F. On Clustering Time Series Using Euclidean Distance and Pearson Correlation. *arXiv preprint arXiv:1601.02213* (2016).

Acknowledgements

This work was funded by grants from the National Natural Science Foundation of China (72034003, 7184096, 71834003).

Author contributions

G.Z. came up with the idea of this study and responsible for overall scientific direction. G.Z., Y.G., Z.C. and W.L. designed the study and compiled the dataset. J.L. performed the technical guidance. B.S. guided and reviewed every draft of this study. All the authors made significant contributions to manuscript editing and approving the final version of the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41597-022-01183-y.

Correspondence and requests for materials should be addressed to G.Z. or B.S.

Reprints and permissions information is available at 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 4.0 International License, which permits use, sharing, adaptation, 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 license, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons license 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 license, visit http://creativecommons.org/licenses/by/4.0/.

The Creative Commons Public Domain Dedication waiver http://creativecommons.org/publicdomain/zero/1.0/ applies to the metadata files associated with this article.

© The Author(s) 2022, corrected publication 2022