Supporting Information for

ES&T in the 21st century: A data-driven analysis of research topics, interconnections, and trends in the past 20 years

Jun-Jie Zhu[†], Willow Dressel[‡], Kelee Pacion[‡], and Zhiyong Jason Ren[†,*]

[†]Department of Civil and Environmental Engineering and Andlinger Center for Energy and the Environment, Princeton University, Princeton, NJ 08544, USA.

[‡]Princeton University Library, Princeton University, Princeton, NJ 08544, USA

[*]Corresponding author; email: zjren@princeton.edu

Summary

Number of pages	28
Number of figures	7
Number of tables	12
Number of texts	4

Table S1. Eleven major challenges identified in raw keyword data and their corresponding six-step preprocessing approaches developed in this study.

Challenges		Corresponding preprocessing approaches (with inspection applied to all)
Description	Examples	
Same stem but in different forms	*system* vs. *systems (system)*; *contamination* vs. *contaminants (contamin)*	**Standard word stemming**. All keywords were lowercased and keywords with more than four letters were stemmed before other steps. Python NLP package *nltk*[1] and the "SnowballStemmer" algorithm was used. For example, *contamination* and *contaminants* were both normalized to their root *contamin*. A few words with irregular plural forms were manually corrected, such as bacterium (bacteria), consortium (consortia), and medium (media).
Prefix or isomer	*3,3'-dichlorobiphenyl* vs. *dichlorobiphenyl*; *alpha alumina* vs. *alumina*	**Excess component removal**. *ChemListem*,[2] a deep neural networks-based Python NLP package for chemical named entity recognition (NER), was adopted to pre-select organic chemicals to avoid affecting terms like *16s ribosomal-rna* and *25-degrees-c*. A rule was applied to typical isomers (such as *1,1,1-trichloroethan* or *2,2',4,4'-tetrabromodiphenyl ether*) that contain number, hyphen, and more than three letters while the first element must be number, and number and letters are not successive. Prefix like *alpha*, *beta*, and *gamma* were removed, initial words such as *contaminated*, *environmental*, and *polluted* and ending words such as *atom*, *concentration(s)*, *emission(s)*, *formation*, *ion(s)*, *level(s)*, *production*, *reduction*, and *removal*, were eliminated for all non-single-word keywords.
Excess ending word	*lead concentration* vs. *lead*; *copper ion* vs. *copper*	
Acronyms and abbreviations	*PAH* vs. *polycyclic aromatic hydrocarbon*; *DBP* vs. *disinfection byproduct*	**Acronym identification and replacement**. A text-based method was used to detect initial letters-based acronyms. The primary step to identify *X*-letter acronyms (*X* = 2, 3, 4, or 5) was to screen all *X*-letter candidates with defined stop words, such as *air* and *gas* (*X* = 3). Candidates were further selected while each of them should have corresponding, first-letters-matched *X*-word term(s). Corresponding articles were identified and reviewed to determine the final acronyms based on domain knowledge. An acronym is a combination of initial letters (e.g., PAH) or partial-initial letters (e.g., TCE) of a terminology, typically from three to five letters. The same method without the first-letters-matched step was used to detect the partial initial letters-based acronyms. **Table S2** lists 45 acronyms identified in this study with special case explained.

<table>
<tr><td>Different chemical expressions</td><td>carbon dioxide vs. CO_2; Hg vs. mercury</td><td rowspan="2">Chemical recognition and unification. Inorganic chemicals had different expressions (name or formula) in the raw data. Identifying chemical formula using Chemical NER (ChemListem) was determined not effective in this case, instead detecting chemical name was used to screen chemicals (e.g., use carbon dioxide rather than co2) with relatively high frequency. In 377 frequent chemicals, only 22 of them were required to be unified (Table S3). Words that contain any typical formats of roman numerals (e.g., i, …, vii, (i), …, (vii)) or charges (+, 2+, …, 7+) were identified and replaced correspondingly (Table S4). Specifically, different formats (e.g., chromium(iii), cr(iii), chromium(vi), cr(vi), and cr) of a metal were unified to a single, base name (chromium) only except when the metal (e.g., iron) has different names in different valences (ferrous oxide or ferric oxide) and is not the single element in the chemical.</td></tr>
<tr><td>Chemicals with charges or roman numerals</td><td>mercury(ii) versus Hg(ii) versus Hg^{2+}</td></tr>
<tr><td>Similar terms that may be combined</td><td>organic compound and organic chemical were combined, whereas organic contaminant and organic compound were not</td><td rowspan="2">Principal term detection and combination. The method involves detecting the same first (several) word(s), which are denoted them as “principal terms”. Any frequent keywords had the same principal terms (only the last word varies) were identified if they have the same number of words. For example, acid rain and acid deposition or dissolved humic substance and dissolved humic material were combined, respectively. The method was applied to keywords from one- to four-word, leading to 62 groups of synonyms based on domain knowledge (Table S5). A similar method was applied to the keywords with the same last (several) word(s), such as carbon nanotube and walled carbon nanotube, leading to another 56 groups of synonyms (Table S6).</td></tr>
<tr><td>Subset terms that may be combined</td><td>in situ bioremediation and in situ remediation were combined</td></tr>
<tr><td>Terms that have the same meaning</td><td>sewage water vs. wastewater; physical chemical vs. physicochemical</td><td rowspan="2">Inspection and post-hoc correction. In each of the above five steps, an initial inspection was used to prepare an effective preprocessing method, and a final inspection was taken to determine and combine variants and synonyms. In addition, a post-hoc inspection and correction was conducted to refine the final treated keywords database and improve the reliability of results. This post-hoc step helped to address many issues, such as same-meaning terms, subset terms, and other miscellaneous issues.</td></tr>
<tr><td colspan="2">Other miscellaneous challenges include excess parenthesis (e.g. poly(dimethylsiloxane) vs. polydimethylsiloxane), excess hyphen (waste-water vs. wastewater), irregular space (zero valent iron vs. zerovalent iron), and repeat-word acronym (trinitrotoluene tnt)</td></tr>
<tr><td colspan="2">All keywords were capitalized in the raw data which makes the above issues more challenging</td><td>Solved with other issues by the above approaches. For example, use chemical name rather than chemical formula in the chemical NER; acronym was not identified based on capital letters.</td></tr>
</table>

Table S2. Acronyms that were identified (frequency ≥ 5) and their full descriptions (punctuations were removed and remain as singular form).

Acronym	Full description	Acronym	Full description
AFM	Atomic force microscopy	PCB	Polychlorinated biphenyl
AHTN	Acetyl hexamethyl tetralin	PCDD*	Polychlorinated dibenzodioxin
BHT	Butylated hydroxytoluene	PCDF*	Polychlorinated dibenzofuran
BPA	Bivalve potamocorbula amurensis	PCE	Perchloroethylene
DDT	Dichlorodiphenyltrichloroethane	PCN	Polychlorinated naphthalene
DOC	Dissolved organic carbon	PCR	Polymerase chain reaction
DOM	Dissolved organic matter	PFAS**	Perfluorinated alkylated substance
EPS	Extracellular polymeric substance	PFC	Per/poly fluorinated compound
GAC	Granular activated carbon	PFOA	Perfluorooctanoic acid
GC	Gas chromatography	PFOS	Perfluorooctane sulfonate
GIS	Geographic information system	PM	Particulate matter
HBCD	Hexabromocyclododecane	RDX	Hexahydro trinitro triazine
HCH	Hexachlorocyclohexane	RO	Reverse osmosis
LCA	Life cycle assessment	SCR	Selective catalytic reduction
MBR	Membrane bioreactor	SMP	Soluble microbial product
MEA	Monoethanolamine	SOA	Secondary organic aerosol
NAPL	Nonaqueous phase liquids	TCDD	Tetrachlorodibenzo p dioxin
NDMA	N nitrosodimethylamine	TCE	Trichloroethylene
NF	Nanofiltration	THM	Trihalomethanes
NMR	Nuclear magnetic resonance	TNT	Trinitrotoluene
NOM	Natural organic matter	UV	Ultraviolet
PAH	Polycyclic aromatic hydrocarbon	VOC	Volatile organic compound
PBDE	Polybrominated diphenyl ether		

* PCDD and PCDF were equal frequently studied together, so all the relevant keywords were replaced and combined as *pcdd/pcdfs*.

**PFAS: Perfluorinated alkylated substance; polyfluorinated alkylated substance; perfluoroalkyl substance; or polyfluoroalkyl substance

Table S3. Chemical names that were identified using *ChemListem* (combined frequency ≥ 10) and unified with their formulas.

Chemical name	Chemical formula	Chemical name	Chemical formula
Ammonia	NH_3	Nitrate radical	NO_3
Bromide	Br	Nitric oxide	NO
Carbon dioxide	CO_2	Nitrogen dioxide	NO_2
Carbon monoxide	CO	Nitrogen oxide	NO_x
Cerium oxide	CeO_2	Nitrous oxide	N_2O
Cesium	Cs	Palladium	Pd
Chloride	Cl	Rhodium	Rh
Hydrogen peroxide	H_2O_2	Selenium	Se
Hydrogen sulfide	H_2S	Sulfur dioxide	SO_2
Hydroxyl radical	OH radical	Titanium dioxide	TiO_2
Methane	CH_4	Zinc oxide	ZnO

Table S4. Identified metals (combined frequency ≥ 10) that had different forms (in raw, low-cased texts and separated by semicolons) and their unified forms.

Different forms	Unified form
al; al(iii)	aluminum
sb; sb(iii)	antimony
as; as(iii); as(v); arsenic(iii); arsenic(v)	arsenic
cd; cd(ii); cd 2+; cadmium(ii)	cadmium
cr; cr(iii); cr(vi); chromium(iii); chromium(vi); hexavalent chromium	chromium
eu; eu(iii); europium(iii)	europium
au; au iii; gold(iii)	gold
pb; pb(ii); lead(ii)	lead
mn; mn(ii); mn(iii); mn(iv); manganese(ii); manganese(iii); manganese(iv)	manganese
hg; hg(ii); hg ii; hg2+; mercury(ii); inorganic mercury; elemental mercury	mercury
np(v); neptunium(v)	neptunium
ni; ni(ii)	nickel
pu; pu(iv); pu(v); plutonium(iv)	plutonium
ag; ag i	silver
tc; tc(vii)	technetium
u(iv); u(vi); u vi; uranium(iv); uranium(vi)	uranium
zn; zn(ii); zinc(ii)	zinc
fe; fe(ii); fe(iii); iron(iii); fe ii; iron(ii); ferrous iron; ferric iron	iron
fe(ii); fe ii; iron(ii); ferrous iron	ferrous*
fe(iii); iron(iii); ferric iron	ferric*
cu; cu(ii); cu ii; cu2+; copper(ii)	copper
cu(ii); cu ii; cu2+; copper(ii)	cupric*

*Only converted to this form if it is part of a binary chemical form, such as *fe(ii) oxide*

Table S5. Keywords (frequency ≥ 10) the same first (several) word(s) identified based on the principal term method and their final replaced term (bold). Keywords may be listed as their singular forms while the actual text replacement also included their plural forms

No.	Keywords
1	acid deposition; acid rain
2	advanced oxidation; advanced oxidation process
3	aerobic biodegradation; aerobic biotransformation
4	anaerobic biodegradation; anaerobic degradation; anaerobic digestion
5	aquatic ecosystem; aquatic environment; aquatic system
6	aromatic compound; aromatic hydrocarbon
7	atmospheric oxidation; atmospheric photooxidation
8	chemical analysis; chemical characteristics; chemical characterization
9	chlorophyll; chlorophyll a; chlorophyll alpha
10	climate; climate change
11	competitive adsorption; competitive sorption
12	contaminated aquifer; contaminated groundwater
13	cryptosporidium; cryptosporidium parvum; cryptosporidium parvum oocysts
14	dissolution kinetic; dissolution rate
15	dissolved humic material; dissolved humic substance; humic substance
16	dissolved organic compound; dissolved organic carbon
17	endocrine disrupting compound; endocrine disrupting chemical; endocrine disruption; endocrine disruptor
18	energy; energy consumption; energy use
19	environmental contaminant; environmental pollutant
20	fecal contamination; fecal pollution
21	fluidized bed; fluidized bed reactor
22	food chain; food web
23	green alga; green algae
24	greenhouse gas; greenhouse gas emission
25	geographic information system; geographical information system
26	human serum; human serum albumin
27	in situ bioremediation; in situ degradation; in situ hybridization; in situ remediation
28	ion exchange; ion exchange membrane
29	land application; land use; land use change
30	life cycle; life cycle analysis; life cycle assessment
31	marine; marine environment; marine ecosystem; marine water

32	messenger RNA; messenger RNA expression
33	microbial degradation; microbial oxidation; microbial transformation
34	mytilus edulis; mytilus edulis l.
35	nano; nanoscale; nanosized
36	nanofiltration; nanofiltration membrane; NF membrane
37	organic acid; organic carbon; organic chemical;
38	organic compound; organic material; organic matter
39	organic contaminant; organic micropollutant; organic pollution
40	organochlorine; organochlorine compound; organochlorine contaminant
41	PCB; PCB congener
42	perfluoroalkyl; perfluoroalkyl compound; perfluoroalkyl contaminant; perfluoroalkyl substance; polyfluoroalkyl chemical; polyfluorinated alkyl substance; polyfluoroalkyl compound; polyfluoroalkyl substance; PFAS
43	petroleum; petroleum hydrocarbon
44	photocatalytic activity; photocatalytic degradation; photocatalytic oxidation
45	photochemical oxidation; photochemical transformation
46	photo fenton; photo fenton reaction
47	quantitative analysis; quantitative determination
48	reduced sulfur; reduced sulfur groups
49	rate coefficient; rate constant
50	RO; RO membrane; reverse osmosis; reverse osmosis membrane
51	seasonal trend; seasonal variation
52	solid phase microextraction; solid phase extraction
53	spatial distribution; spatial pattern; spatial trend; spatial variability; spatial variation
54	spectroscopic characterization; spectroscopic evidence; spectroscopic properties
55	steroid estrogens; steroid hormones
56	surface chemistry; surface properties
57	temporal trend; temporal variability
58	thermal decomposition; thermal degradation
59	treatment process; treatment system; treatment work
60	ultrafiltration; ultrafiltration membrane; UF membrane
61	UV; UV light
62	volatile organic compound; volatile organic contaminant

Table S6. Keywords (frequency ≥ 10) with the same last (several) word(s) identified based on the principal term method and their final replaced term (bold). Keywords may be listed as their singular forms while the actual text replacement also included their plural forms

No.	Keywords	No.	Keywords
1	**activated carbon**; granular activated carbon	26	**children**; preschool children; young children
2	**aerosol**; ambient aerosol; atmospheric aerosol	27	**China**; north China; south China
3	**algae**; blue green algae; green algae	28	coated silver nanoparticle; **silver nanoparticle**
4	**alkane**; n-alkane	29	**desalination**; seawater desalination; water desalination
5	ambient air; **atmospheric air**; outdoor air	30	dolphins tursiops truncatus; **tursiops truncatus**
6	**anaerobic bacteria**; strictly anaerobic bacteria	31	**estuary**; river estuary
7	**Asia**; east Asia	32	**exposure**; human exposure
8	**Atlantic**; north Atlantic	33	**ferrihydrite**; line ferrihydrite
9	Atlantic salmon; **salmon**	34	**fish**; marine fish
10	bears ursus maritimus; **ursus maritimus**	35	**groundwater**; shallow groundwater
11	biodiversity; **diversity**; microbial diversity	36	gulls larus argentatus; **larus argentatus**
12	**biofilm reactor**; membrane biofilm reactor	37	**health**; human health
13	**biofilm**; microbial biofilm	38	**in vitro**; vitro
14	biofuel cell; fuel cell; **microbial fuel cell**	39	**in vivo**; vivo
15	**biomass**; microbial biomass	40	***Lake Michigan**; southern Lake Michigan
16	**black carbon**; environmental black carbon	41	m-xylene; p-xylene; **xylene**
17	blue mussel; **mussel**	42	minnow pimephales promelas; **pimephales promelas**
18	**California**; southern California	43	municipal wastewater; **wastewater**
19	carbon sequestration; **CO_2 sequestration**	44	**nanomaterial**; engineered nanomaterial
20	carp cyprinus carpio; **cyprinus carpio**	45	**nanoparticle**; engineered nanoparticle
21	capture; carbon capture; dioxide capture; **CO_2 capture**	46	**nitrosamine**; n-nitrosamine
22	chain PFAA; **PFAA**	47	**nonylphenol**; p-nonylphenol
23	chain PFCA; **PFCA**	48	northern Sweden; **Sweden**
24	chemical ion; **ion**	49	**Ontario;** southern Ontario
25	**chemistry**; environmental chemistry	50	**temporal trend**; time trend
51	airborne particulate matter; ambient particulate matter; atmospheric particulate matter; **particulate matter**		
52	**carbon nanotube**; multiwalled carbon nanotube; walled carbon nanotube		
53	**liquid chromatography**; performance liquid chromatography		
54	**magnetic resonance spectroscopy**; nuclear magnetic resonance spectroscopy		
55	midwestern USA; northeastern USA; southeastern USA; **USA**; western USA		
56	rainbow trout; **trout**; trout oncorhynchus mykiss; oncorhynchus mykiss; salvelinus namaycush; trout salvelinus namaycush		

*Lake Erie, Lake Michigan, Lake Ontario, and Lake Superior were combined with “great lakes”

Text S1. Stemming

Stemming is a crude process to cut off the last several characters.[3] Stemming is a better way in our case, and all keywords were lowercased and keywords that were more than four letters were stemmed before other preprocessing steps. Python NLP package *nltk*[1] was used to perform the stemming and the "SnowballStemmer" algorithm was used. Specific rules used in stemming can be complex, here we briefly introduce several basic rules.[4,5] Porter's algorithm is the most popular algorithm to perform the stemming of English. Some typical rules:

- *sses* → *ss*; *ies* → *i*; *ational* → *ate*; *tional* → *tion*
- Weight of word sensitive rules
- ($m > 1$) EMENT: *replacement* → *replac*; *cement* → *cement*

Given that a word is in a form of $[C](VC)^m[V]$, where C and V are consonant and vowel, respectively; *m* is the *measures* of a word or part of a word. The rules for removing a suffix, (condition) S1 → S2, are usually based on *m*. This means that S1 will be replaced by S2 if the word ends with S1 and the stem before S1 meets the condition. In the above example, S1 is 'EMENT' and S2 is null, which maps *replacement* to *replac*, but not *cement* to *c*, because *replac* is a word part with m = 2. There are many other specific rules and information associated with the Porter's algorithm.[5] Snowball was a revised and improved version of the Porter's algorithm when the inventor, Martin Porter, realized that the original algorithm could give incorrect results in many researchers' published works.[6]

Text S2. Selection of trending up topics

Majority of trending up keywords were determined based on moderate values of the trend factor (> 0.4) and $F_{current}$ (> 4). The two criteria helped to ensure a general growing popularity in selected keywords when comparing their normalized frequencies during the current period (2010-2019) with the past period (2000-2009). To guarantee a steady popularity, an additional criterion ($F_{2015-2019}/F_{2010-2014} > 90\%$) was applied to exclude keywords with a much lower frequency in the most recent years. The proposed trending analyzing method simplified the selection processes, but the break point may cause an "edge effect". In other words, it is possible to miss a potential trending up keywords if its frequency rapidly increases over the years just before 2009 but slowly increases subsequently. Although most of this type of keywords can be still detected using the above approach, some of them have a trend factor of between 0.2 and 0.4, below the defined threshold. To address this issue, we considered two additional criteria to screen the candidates that did not meet the original trend factor (> 0.4):

a. The normalized frequency in the current period (2010-2019) should be slightly higher ($0.1 < \text{trend factor}_{2007-2009}^{2010-2019} < 0.25$) than the normalized frequency during 2007-2009 (years just before 2010);
b. The normalized frequency in the current period (2010-2019) should be significantly higher ($\text{trend factor}_{2000-2006}^{2010-2019} > 0.4$) than the normalized frequency during 2000-2006.

It is also worthwhile to mention that the above approaches helped to determine the most trending up topics, while there are many other less popular, trending up topics.

Text S3. Rule-based classification method

The title, abstract, and keywords of a paper were treated and combined to develop the corpus; keywords were preprocessed as described previously; the abstract was also tokenized by *n*-grams (n = 1, 2, 3, and 4), lowercased, stop-worded, and stemmed. To accurately classify the papers, specific terms, denoted as *domain surrogates*, were carefully and rigorously selected to label every individual domain. The selected surrogates should be representative. For example, compared to *disinfection*, *disinfection byproduct* is a better surrogate to label a water-specific study. Selection of surrogates followed an iterative procedure comprised of the following steps:

1) Initial, typical surrogates were brainstormed and prepared;

2) Because the keywords *water* and *air* are less representative, more specific, frequent terms that included “water” or “air”, such as *drinking water* or *air quality*, were identified;

3) New surrogates were identified from frequent terms of pre-classified papers based on pre-identified surrogates;

4) A manual inspection was used to serve as an additional expansion on the list of surrogates based on unlabeled papers;

5) Steps 3 and 4 were iteratively conducted until a minimum document retrieval rate (80%) was achieved and no more than five new surrogates were identified.

6) A post-hoc validation was taken to improve the classification accuracy. Fifty sample papers were randomly selected for review at each iteration, and inappropriate surrogates were removed or corrected afterward. A sample classification accuracy (correct number/sample size) was calculated and the validation was iteratively conducted until 90% accuracy was achieved.

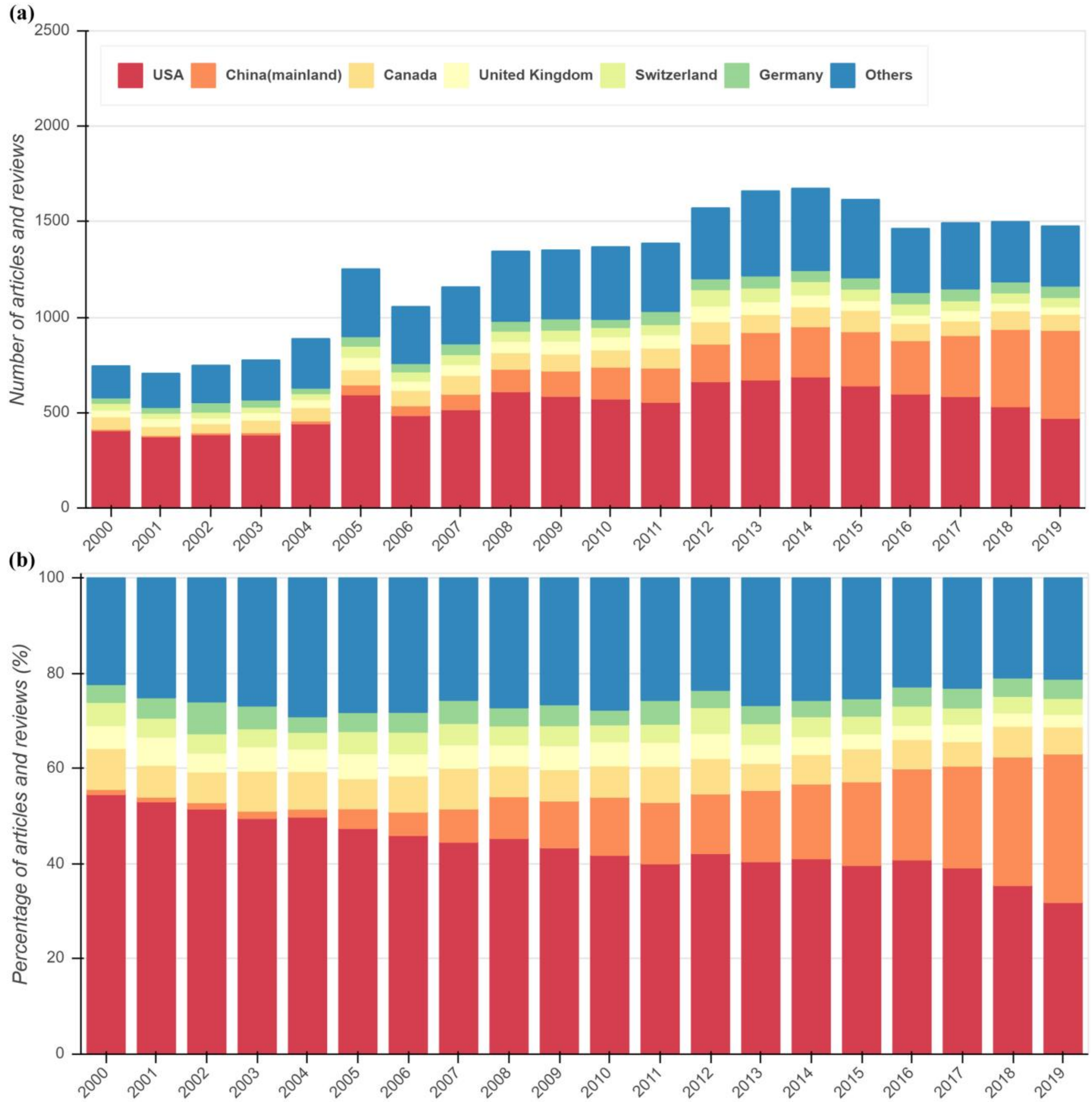


Figure S1. Temporal and geospatial variations of articles and reviews published in ES&T from 2000 to 2019. (a) Actual number of papers; (b) percentage of valid papers.

Table S7. Top 100 frequent keywords (lowercased, stemmed form) and their frequencies

No.	Keyword	Freq.	No.	Keyword	Freq.	No.	Keyword	Freq.
1	*water*	3106	35	*speciat*	960	69	*natural organic matt*	603
2	*sorption*	2581	36	*chemic*	925	70	*bioaccumul*	598
3	*soil*	2383	37	*surfac*	916	71	*energi*	592
4	*emiss*	2187	38	*air*	912	72	*ozon*	590
5	*oxid*	1969	39	*fate*	901	73	*mass spectroscopi*	569
6	*surface wat*	1852	40	*bacteria*	888	74	*urban*	555
7	*sediment*	1704	41	*identif*	871	75	*organic pollut*	553
8	*exposur*	1703	42	*china*	871	76	*implic*	553
9	*organic compound*	1639	43	*drinking wat*	848	77	*copper*	551
10	*remov*	1587	44	*transform*	809	78	*analysi*	550
11	*pah*	1560	45	*matter*	790	79	*growth*	543
12	*model*	1560	46	*atmospher*	787	80	*catalyst*	534
13	*degrad*	1503	47	*metal*	784	81	*nitrat*	529
14	*mechan*	1449	48	*pbde*	784	82	*nitrogen*	529
15	*kinet*	1447	49	*product*	752	83	*air pollut*	528
16	*wastewat*	1438	50	*accumul*	745	84	*e coli*	527
17	*toxic*	1417	51	*biodegrad*	732	85	*life cycle assess*	511
18	*impact*	1414	52	*aerosol*	711	86	*spectroscopi*	510
19	*reduct*	1384	53	*hydrocarbon*	698	87	*arsenic*	509
20	*pcb*	1374	54	*perform*	690	88	*persistent organic pollut*	508
21	*contamin*	1368	55	*plant*	682	89	*co2*	490
22	*transport*	1351	56	*chemistri*	663	90	*sampl*	486
23	*particulate matt*	1335	57	*deposit*	648	91	*aromatic compound*	480
24	*carbon*	1288	58	*mercuri*	646	92	*h2o2*	475
25	*system*	1253	59	*fish*	646	93	*distribut*	474
26	*particl*	1180	60	*concentr*	643	94	*fraction*	465
27	*humic subst*	1127	61	*behavior*	636	95	*black carbon*	465
28	*iron*	1094	62	*nanoparticl*	635	96	*climat*	464
29	*usa*	1052	63	*temperatur*	632	97	*pharmaceut*	461
30	*acid*	1039	64	*bioavail*	628	98	*great lak*	456
31	*groundwat*	1019	65	*complex*	626	99	*ion*	455
32	*pollut*	1012	66	*dissolved organic carbon*	616	100	*miner*	449
33	*aqueous solut*	980	67	*wastewater treatment process*	607			
34	*environ*	968	68	*heavy met*	604			

Table S8. Summary of annual top ten keywords from 2000 to 2019

Top#	2000	2001	2002	2003	2004
1	*soil*	*water*	*sorption*	*water*	*soil*
2	*water*	*soil*	*water*	*soil*	*sorption*
3	*sorption*	*sorption*	*soil*	*sorption*	*water*
4	*sediment*	*sediment*	*organic compound*	*pah*	*pcb*
5	*organic compound*	*organic compound*	*pah*	*sediment*	*surface wat*
6	*pah*	*emiss*	*oxid*	*pcb*	*sediment*
7	*pcb*	*pah*	*surface wat*	*surface wat*	*pah*
8	*surface wat*	*surface wat*	*sediment*	*remov*	*organic compound*
9	*kinet*	*oxid*	*humic subst*	*kinet*	*degrad*
10	*oxid*	*pcb*	*emiss*	*emiss*	*kinet*
Top#	2005	2006	2007	2008	2009
1	*water*	*water*	*water*	*water*	*water*
2	*sorption*	*sorption*	*sorption*	*sorption*	*sorption*
3	*soil*	*soil*	*soil*	*soil*	*emiss*
4	*pah*	*surface wat*	*oxid*	*oxid*	*soil*
5	*organic compound*	*sediment*	*surface wat*	*emiss*	*oxid*
6	*surface wat*	*contamin*	*pah*	*organic compound*	*sediment*
7	*sediment*	*pcb*	*sediment*	*sediment*	*exposur*
8	*pcb*	*pah*	*emiss*	*contamin*	*organic compound*
9	*oxid*	*oxid*	*degrad*	*reduct*	*model*
10	*model*	*remov*	*contamin*	*degrad*	*remov*
Top#	2010	2011	2012	2013	2014
1	*water*	*water*	*water*	*water*	*water*
2	*sorption*	*sorption*	*sorption*	*sorption*	*emiss*
3	*soil*	*soil*	*emiss*	*emiss*	*impact*
4	*emiss*	*emiss*	*soil*	*soil*	*exposur*
5	*oxid*	*exposur*	*exposur*	*exposur*	*sorption*
6	*surface wat*	*oxid*	*toxic*	*impact*	*soil*
7	*sediment*	*sediment*	*oxid*	*oxid*	*oxid*
8	*degrad*	*toxic*	*surface wat*	*toxic*	*toxic*
9	*transport*	*surface wat*	*mechan*	*surface wat*	*surface wat*
10	*pcb*	*contamin*	*impact*	*kinet*	*carbon*
Top#	2015	2016	2017	2018	2019
1	*emiss*	*water*	*water*	*water*	*water*
2	*water*	*sorption*	*emiss*	*emiss*	*exposur*
3	*impact*	*exposur*	*exposur*	*exposur*	*oxid*
4	*oxid*	*emiss*	*wastewat*	*impact*	*emiss*
5	*exposur*	*impact*	*surface wat*	*sorption*	*remov*
6	*wastewat*	*oxid*	*soil*	*wastewat*	*impact*
7	*sorption*	*soil*	*particulate matt*	*soil*	*degrad*
8	*model*	*toxic*	*toxic*	*oxid*	*mechan*
9	*surface wat*	*wastewat*	*remov*	*surface wat*	*sorption*
10	*soil*	*model*	*impact*	*remov*	*particulate matt*

Table S9. Summary of the 79 couples of high frequent (≥ 200) co-occurring keywords

Keyword 1	Keyword 2	Freq.	Keyword 1	Keyword 2	Freq.
sorption	*soil*	538	*degrad*	*biodegrad*	259
water	*sorption*	467	*water*	*degrad*	254
emiss	*particulate matt*	434	*pah*	*aromatic compound*	250
sorption	*remov*	426	*sorption*	*mechan*	249
pcb	*pah*	409	*mercuri*	*methylmercuri*	247
soil	*sediment*	369	*water*	*organic compound*	238
pcb	*pbde*	345	*usa*	*emiss*	237
particl	*particulate matt*	342	*pah*	*organic compound*	236
pbde	*brominated flame retard*	334	*sorption*	*iron*	232
soil	*organic compound*	325	*particl*	*emiss*	224
pah	*hydrocarbon*	321	*sorption*	*reduct*	223
pcb	*persistent organic pollut*	317	*sorption*	*pah*	222
sorption	*oxid*	315	*wastewat*	*surface wat*	222
oxid	*mechan*	312	*sorption*	*aqueous solut*	221
reduct	*iron*	312	*water*	*aqueous solut*	220
water	*soil*	309	*sorption*	*kinet*	219
impact	*emiss*	308	*speciat*	*soil*	219
water	*remov*	307	*sediment*	*pcb*	218
sorption	*organic compound*	306	*water*	*kinet*	218
oxid	*kinet*	305	*water*	*acid*	217
sorption	*sediment*	303	*water*	*groundwat*	217
water	*sediment*	303	*pcb*	*biphenyl*	217
air pollut	*particulate matt*	300	*emiss*	*china*	214
surface wat	*sediment*	298	*kinet*	*degrad*	214
sediment	*pah*	298	*humic subst*	*dissolved organic carbon*	213
water	*oxid*	296	*soil*	*humic subst*	213
sorption	*humic subst*	292	*soil*	*degrad*	212
oxid	*degrad*	288	*sorption*	*activated carbon*	211
mechan	*kinet*	287	*humic subst*	*acid*	209
aerosol	*particulate matt*	286	*pcb*	*contamin*	208
remov	*oxid*	284	*water*	*contamin*	207
reduct	*oxid*	282	*speciat*	*sorption*	205
soil	*pah*	281	*soil*	*bioavail*	205
oxid	*iron*	280	*particl*	*aerosol*	203
humic subst	*natural organic matt*	278	*sediment*	*organic compound*	203
wastewat	*remov*	276	*transport*	*soil*	203
hydrocarbon	*aromatic compound*	274	*reduct*	*kinet*	202
surfac	*sorption*	270	*water*	*mechan*	202
matter	*humic subst*	266	*remov*	*reduct*	201
toxic	*exposur*	266			

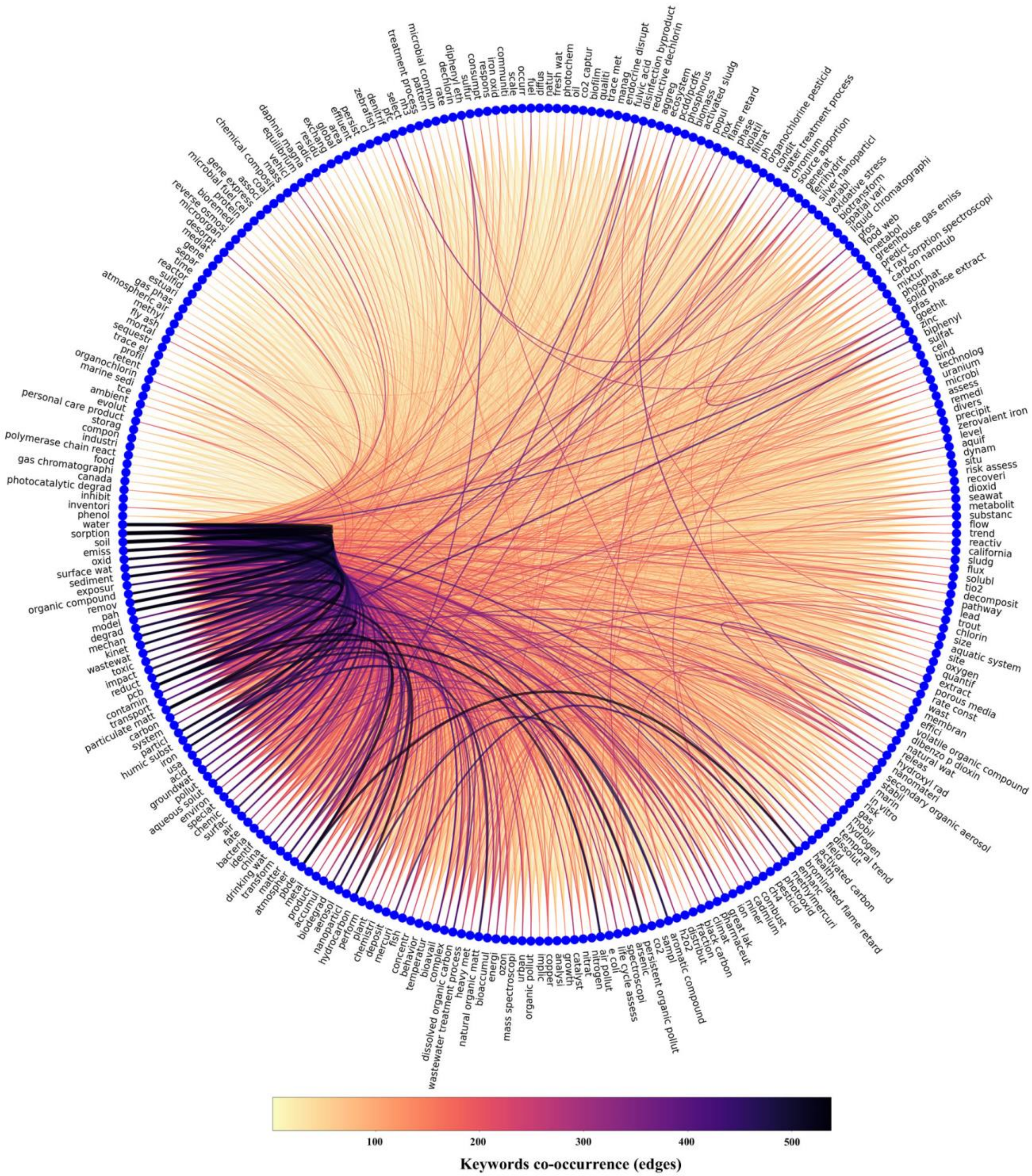


Figure S2. Co-occurrence of the top 300 frequent keywords (stemmed form) based on the circos plot. The keywords (nodes) are ordered by their overall frequency. Edge width and color are used to indicate the co-occurrence between keywords. High quality figures and other related materials can also be found or downloaded via the author's webpage[7].

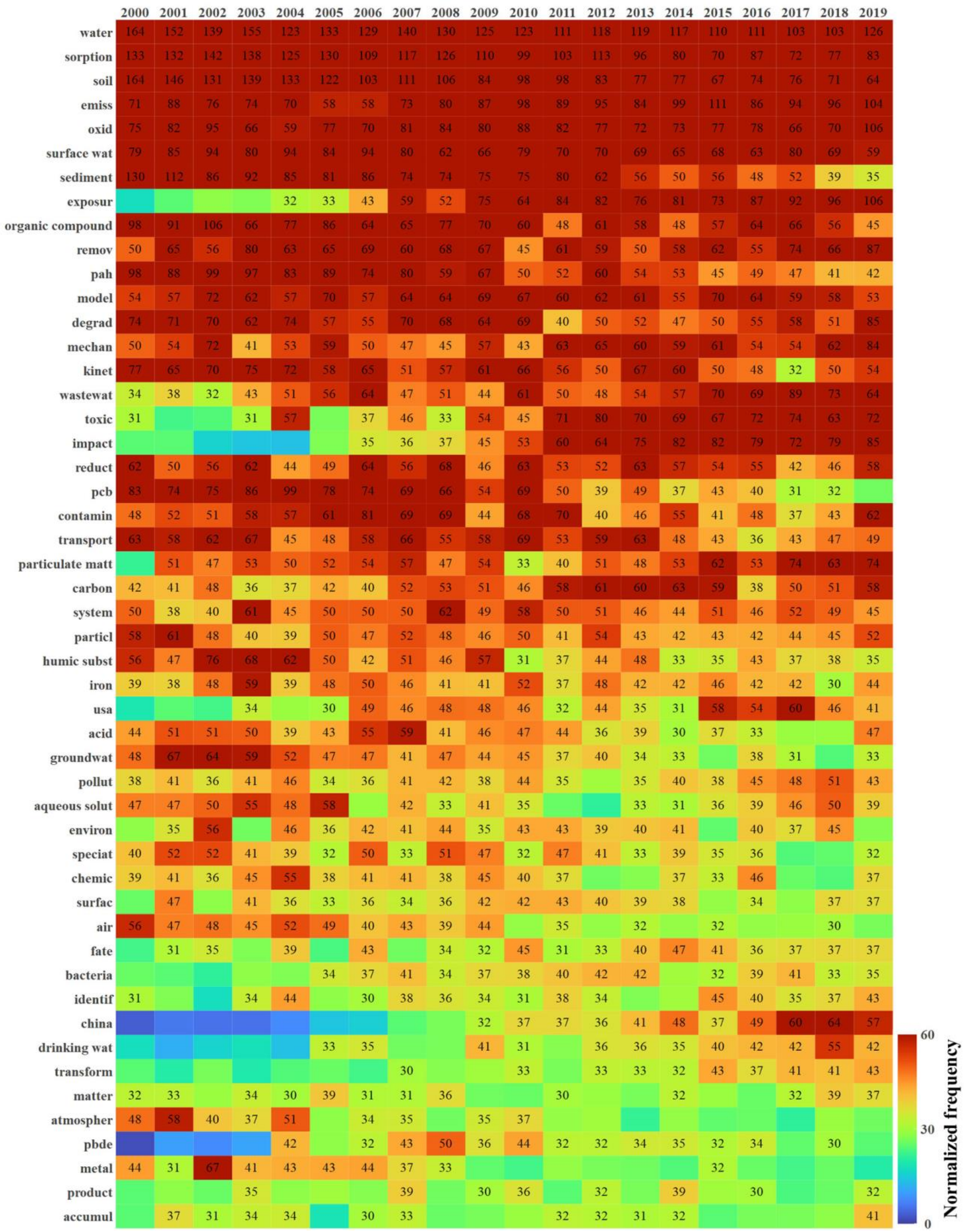


Figure S3. Temporal trend of the top 50 frequent keywords based on normalized annual frequency. Higher frequencies (≥ 30) are labeled.

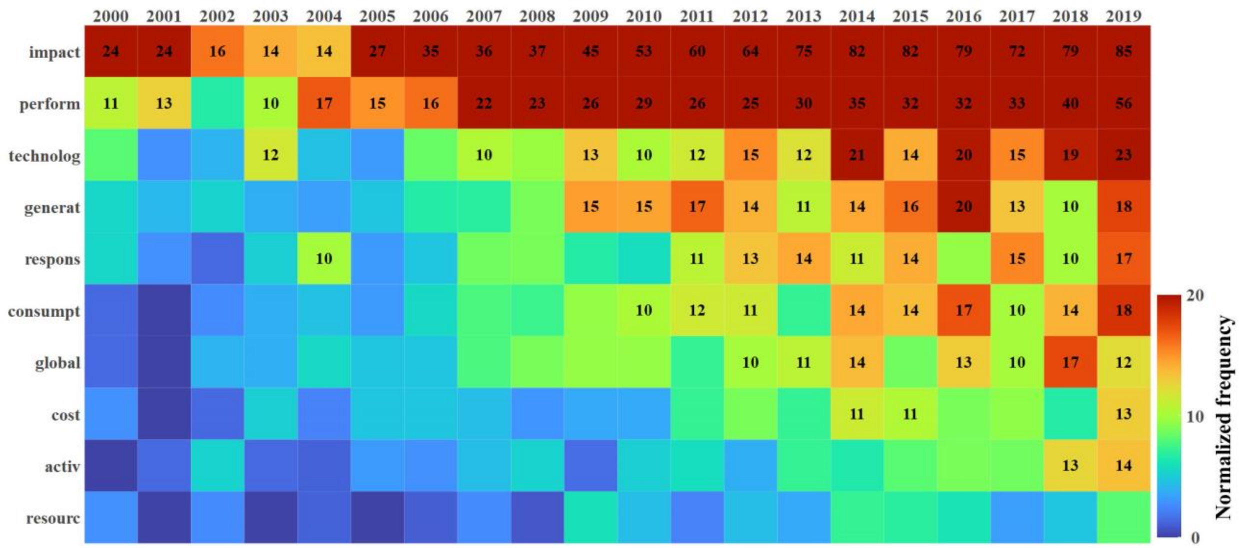


Figure S4. Temporal trend of ten other "general" keywords that have been trending up over the time based on annual normalized frequency. Higher frequencies (≥ 10) are labeled; keywords are ordered by the cumulative frequency.

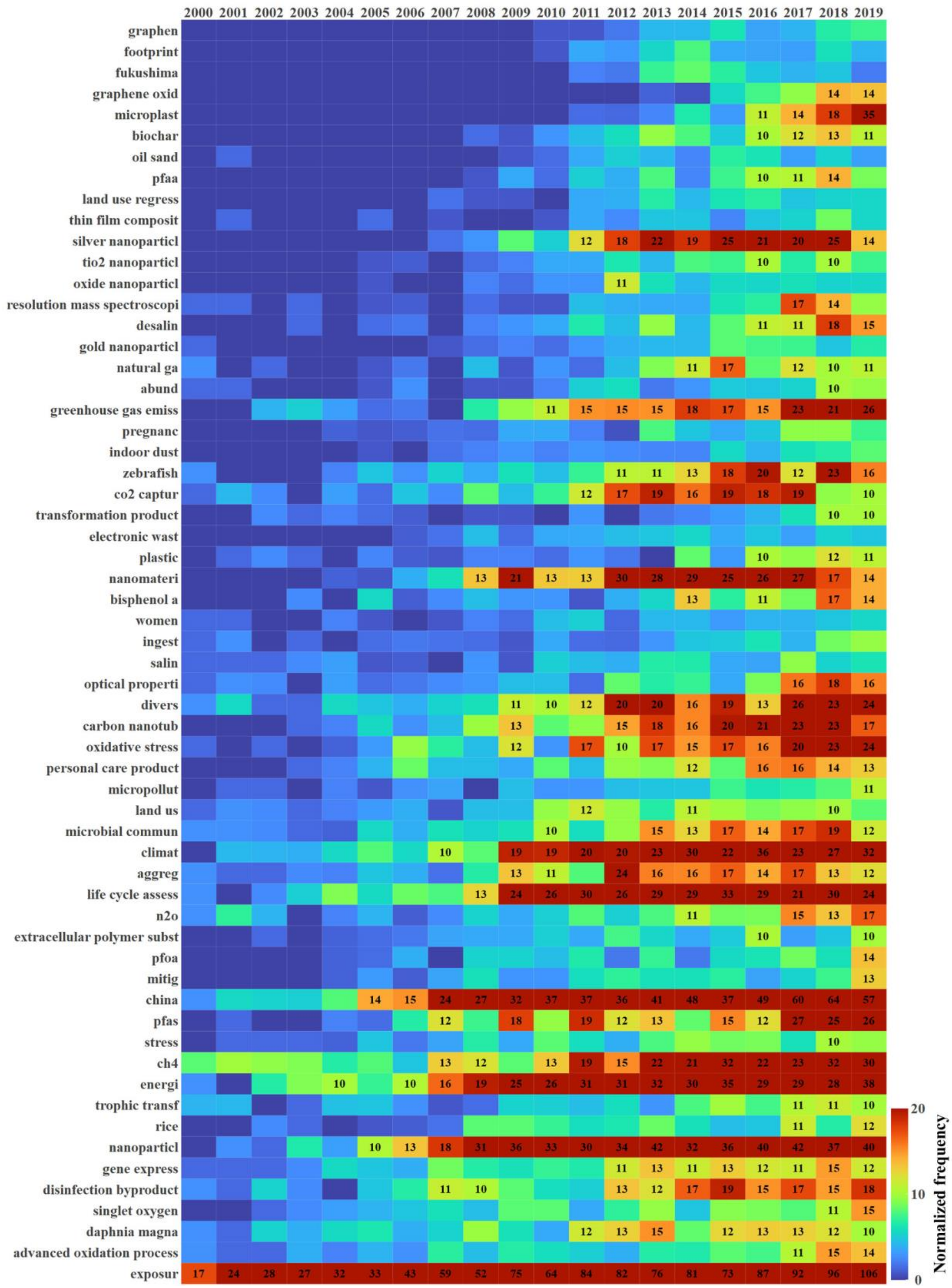


Figure S5. Temporal trend of keywords that have been trending up over the time based on annual normalized frequency. Higher frequencies (≥ 10) are labeled; keywords are ordered by the trend factor.

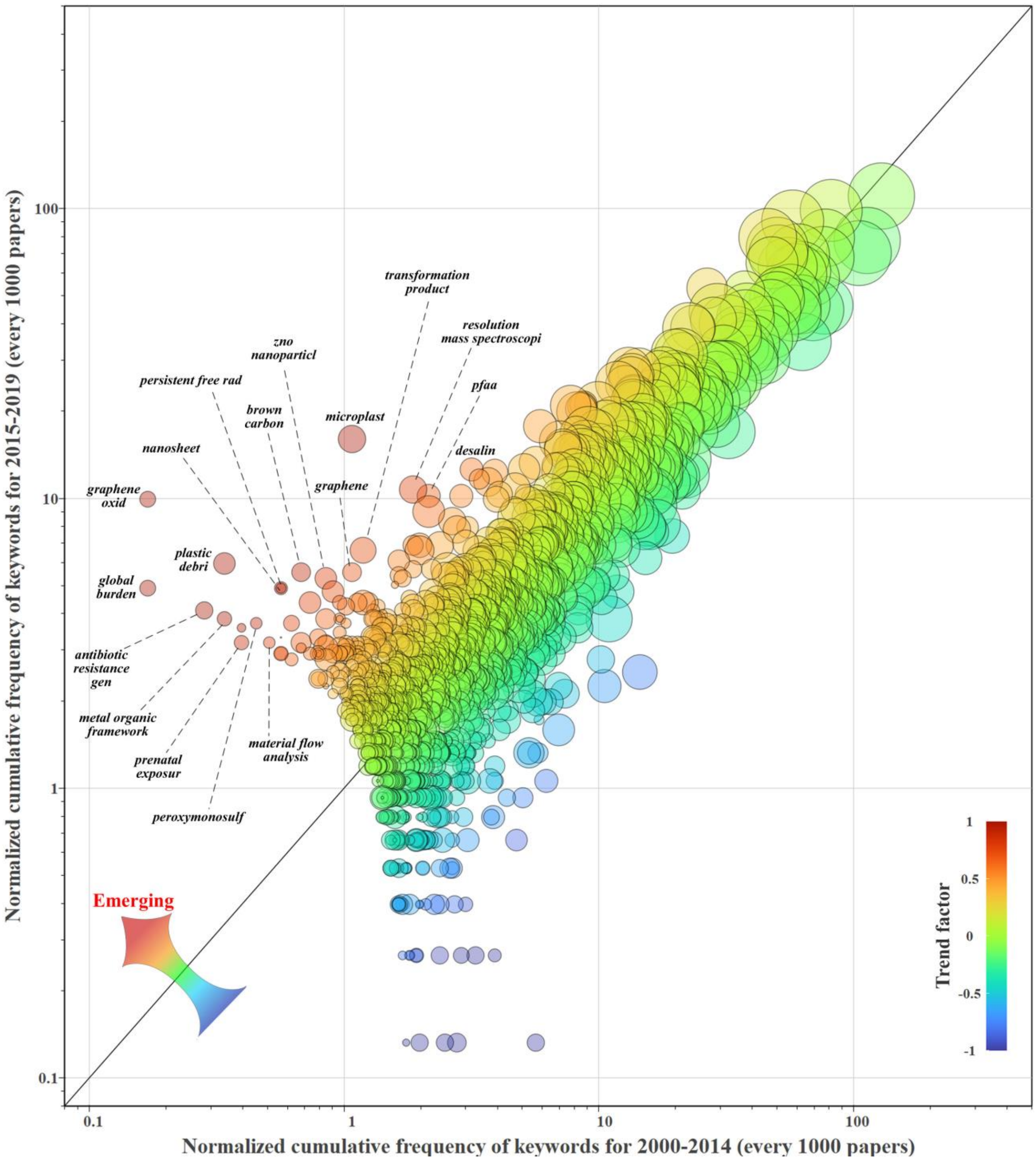


Figure S6. Normalized cumulative frequencies of the top 1500 frequent keywords (bubbles) in the earlier (2000-2014) and most recent (2015-2019) periods. Trend factor value is shown by color; keywords rendered by the red color are more likely to be emerging research topics. The size of bubble reflects the geospatial popularity of the keyword.

Table S10. Major domain surrogates (#influenced documents ≥ 5) identified during the rule-based classification method based on ES&T data. Different forms or abbreviations of surrogates might be used.

Domain	Domain surrogates
Air	acid deposition; acid rain; aerosol; air emission; air mass; air pollution; air quality; air sample; airborne; ambient air; atmospheric; co2 capture; co2 emission; clean air; coal fired power plant; downwind; dry deposition; dust sample; emission control; emission factor; emission inventory; emission rate; emission reduction; emissions inventory; emissions reduction; exhaust; flue gas; fly ash; fossil fuel combustion; indoor; light duty vehicle; long range transport; marine boundary layer; meteorological; multimedia model; nitrogen dioxide emission; nitrogen oxide emission; nitrous oxide emission; particulate matter; plume model; reactive gaseous; semivolatile organic compound; smog; source apportionment; sulfur dioxide; ultrafine particle; vehicle emission; volatile organic compound; water vapor
Soil	acid volatile sulfide; clay; contaminated land; contaminated sediment; contaminated site; contaminated soil; enrichment factor; glacier; multimedia model; peat; plant root; plant uptake; porewater; porous heterogeneous medium; remobilization; rhizosphere; root cell; sediment; sedimentary; snowpack; soil; subsurface; superfund
Solid waste	agricultural waste; animal waste; bottom ash; composting; electronic waste; food waste; hazardous waste; landfill; livestock waste; mine waste; mining waste; municipal solid waste; nuclear waste; organic waste; plastic waste; solid waste; waste incinerator; waste management; waste material; waste pcb; waste repository; wastes disposal
Water	acid mine drainage; aquaculture; aquatic ecosystem; aquatic environment; aquatic life; aquatic organism; aquatic system; aquatic toxicity; aqueous stream; brackish water; coastal water; contaminated water; creek; cryptosporidium; deepwater; deionized water; desalination; disinfection byproduct; drinking water; estuary; eutrophication; flood; freshwater; groundwater; gulf of mexico; hydrology; injection well; irrigation water; lagoon; lake; marine environment; marine food web; marine mammal; marine water; multimedia model; mussel; natural water; phytoplankton; polluted water; potable water; rainwater; receiving water; river; riverine; sea; seawater; softening; source water; stormwater; surface water; tap water; trout; water act; water consumption; water disinfection; water dispersion; water distribution; water environment; water footprint; water management; water pollution; water purification; water resource; water sample; water source; water supply; water suspension; water treatment; water use; water velocity; watershed; waterway; wetland
Wastewater	activated sludge; anammox; biosolid; granular sludge; membrane bioreactor; mine water; sequencing batch reactor; sewage; sewer; waste stream; wastewater; wastewater treatment process

Additional notes:

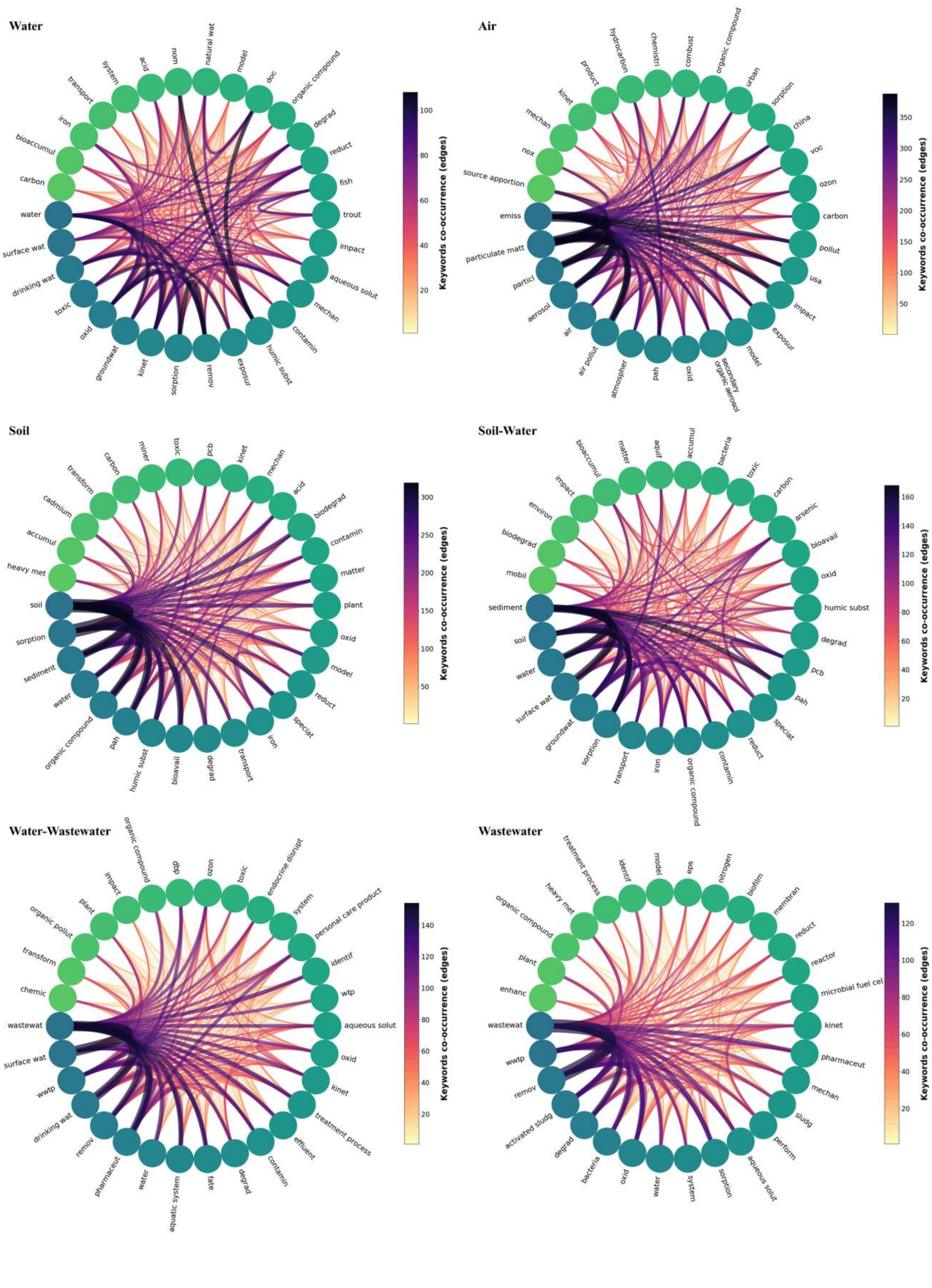
- Many initial surrogates were not included because there are more influential surrogates can be used to label the same papers. For example, “phosphorus recovery” was not used because “wastewater” covered all of the relevant papers.
- Glacier and snowpack are grouped to the soil domain in this study.
- “sediment” belonged to the soil domain when it appeared together with water-related surrogates.
- Hazardous wastes (e.g., electronic waste, nuclear waste) were also included in the solid waste domain.

Table S11. List of the 31 classified domain groups (A: air; S: soil; SW: solid waste; W: water; WW: wastewater) and their numbers of papers, highest, average, and standard deviation (SD) of the normalized citation (NC) counts (#/year). Groups that have more than 200 papers are shown in grey shaded cells.

Domain type	Specific domain(s)	#Papers	Highest NC	Average NC	SD NC
Mono-	A	4454	79	5.7	6.0
	S	2481	120	5.5	6.3
	SW	298	37	5.3	5.0
	W	4458	126	6.2	7.7
	WW	1006	244	8.7	12.3
Bi-	A-S	632	57	5.4	5.8
	A-SW	229	45	4.4	4.3
	A-W	796	41	5.4	4.7
	A-WW	109	65	5.7	7.3
	S-SW	149	40	4.9	5.8
	S-W	2445	103	5.6	6.1
	SW-W	76	35	5.4	5.4
	S-WW	167	33	7.1	6.2
	SW-WW	58	34	7.4	7.1
	W-WW	1213	309	9.0	12.4
Tri-	A-S-SW	65	42	6.1	6.4
	A-SW-W	25	18	6.4	4.6
	A-S-WW	15	31	6.7	7.3
	A-SW-WW	17	12	5.2	3.1
	A-S-W	653	96	5.3	6.5
	A-W-WW	98	62	8.6	9.3
	S-SW-W	91	27	5.8	5.5
	S-SW-WW	22	54	9.2	11.0
	S-W-WW	371	156	9.3	15.1
	SW-W-WW	35	116	10.0	19.4
Quad-	A-S-SW-W	25	39	7.5	8.5
	A-S-SW-WW	5	9	4.2	3.2
	A-S-W-WW	53	197	10.1	26.6
	A-SW-W-WW	3	9	5.0	2.8
	S-SW-W-WW	17	49	13.0	13.7
All domains		5	13	4.4	4.5

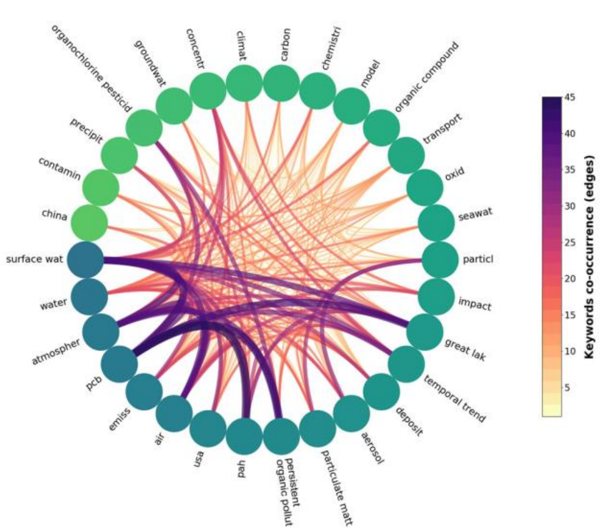
Table S12. Summary of the top ten keywords and their frequencies for the 12 major groups (#papers ≥ 200, groups are ordered by number of papers).

Top#	water	air	soil	soil-water
1	*water, 765*	*emiss, 1325*	*soil, 1144*	*sediment, 690*
2	*surface wat, 579*	*particulate matt, 1106*	*sorption, 574*	*soil, 530*
3	*drinking wat, 494*	*particl, 600*	*sediment, 435*	*water, 430*
4	*toxic, 390*	*aerosol, 561*	*water, 327*	*surface wat, 425*
5	*oxid, 364*	*air, 464*	*organic compound, 313*	*groundwat, 399*
6	*groundwat, 363*	*air pollut, 457*	*pah, 281*	*sorption, 354*
7	*kinet, 361*	*atmospher, 430*	*humic subst, 255*	*transport, 254*
8	*sorption, 338*	*pah, 425*	*bioavail, 225*	*iron, 226*
9	*remov, 333*	*oxid, 392*	*degrad, 206*	*organic compound, 224*
10	*exposur, 308*	*secondary organic aerosol, 388*	*transport, 200*	*contamin, 220*
Top#	water-wastewater	wastewater	air-water	air-soil-water
1	*wastewat, 613*	*wastewat, 448*	*surface wat, 174*	*surface wat, 215*
2	*surface wat, 264*	*wwtp, 238*	*water, 126*	*sediment, 160*
3	*wwtp, 218*	*remov, 236*	*atmospher, 107*	*soil, 128*
4	*drinking wat, 205*	*activated sludg, 142*	*pcb, 102*	*water, 91*
5	*remov, 197*	*degrad, 131*	*emiss, 98*	*pcb, 90*
6	*pharmaceut, 194*	*bacteria, 116*	*air, 98*	*pah, 81*
7	*water, 155*	*oxid, 114*	*usa, 77*	*deposit, 79*
8	*aquatic system, 125*	*water, 110*	*pah, 74*	*transport, 77*
9	*fate, 117*	*system, 92*	*persistent organic pollut, 71*	*contamin, 74*
10	*degrad, 109*	*sorption, 88*	*particulate matt, 65*	*organic compound, 69*
Top#	air-soil	soil-water-wastewater	solid waste	air-solid waste
1	*soil, 255*	*wastewat, 150*	*wast, 61*	*emiss, 74*
2	*emiss, 113*	*sediment, 97*	*msw, 40*	*fly ash, 67*
3	*pah, 100*	*surface wat, 92*	*china, 35*	*pcdd/pcdfs, 63*
4	*air, 90*	*soil, 73*	*electronic wast, 30*	*combust, 61*
5	*pcb, 89*	*fate, 72*	*pbde, 29*	*dibenzo p dioxin, 51*
6	*atmospher, 82*	*sorption, 54*	*system, 25*	*china, 38*
7	*particulate matt, 72*	*wwtp, 48*	*sorption, 24*	*msw, 37*
8	*deposit, 69*	*remov, 45*	*manag, 24*	*inciner, 34*
9	*sediment, 59*	*pharmaceut, 44*	*energi, 23*	*pcb, 29*
10	*model, 59*	*degrad, 41*	*product, 23*	*waste inciner, 28*

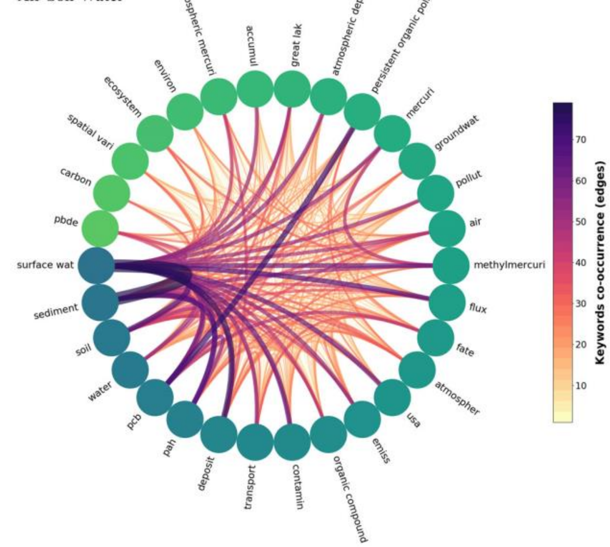


Water
nom
natural wat
model
doc
organic compound
degrad
reduct
fish
trout
impact
aqueous solut
mechan
contamin
humic subst
exposur
remov
sorption
kinet
groundwat
oxid
toxic
drinking wat
surface wat
water
carbon
bioaccumul
iron
transport
system
acid
Keywords co-occurrence (edges)
Air
hydrocarbon
chemistri
combust
organic compound
urban
sorption
china
voc
ozon
carbon
pollut
usa
impact
exposur
model
secondary organic aerosol
oxid
pah
atmospher
air pollut
air
aerosol
particl
particulate matt
emiss
source apportion
nox
mechan
kinet
product
Keywords co-occurrence (edges)
Soil
miner
toxic
pcb
kinet
mechan
acid
biodegrad
contamin
matter
plant
oxid
model
reduct
speciat
iron
transport
degrad
bioavail
humic subst
pah
organic compound
water
sediment
sorption
soil
heavy met
accumul
cadmium
transform
carbon
Keywords co-occurrence (edges)
Soil-Water
bioaccumul
matter
aquif
accumul
bacteria
toxic
carbon
arsenic
bioavail
oxid
humic subst
degrad
pcb
pah
speciat
reduct
contamin
organic compound
iron
transport
sorption
groundwat
surface wat
water
soil
sediment
mobil
biodegrad
environ
impact
Keywords co-occurrence (edges)
Water-Wastewater
organic compound
dbp
ozon
toxic
endocrine disrupt
system
personal care product
identif
wtp
aqueous solut
oxid
kinet
treatment process
effluent
contamin
degrad
fate
aquatic system
water
pharmaceut
remov
drinking wat
wwtp
surface wat
wastewat
chemic
transform
organic pollut
plant
impact
Keywords co-occurrence (edges)
Wastewater
treatment process
identif
model
eps
nitrogen
biofilm
membran
reduct
reactor
microbial fuel cel
kinet
pharmaceut
mechan
sludg
perform
aqueous solut
sorption
system
water
oxid
bacteria
degrad
activated sludg
remov
wwtp
wastewat
enhanc
plant
organic compound
heavy met
Keywords co-occurrence (edges)

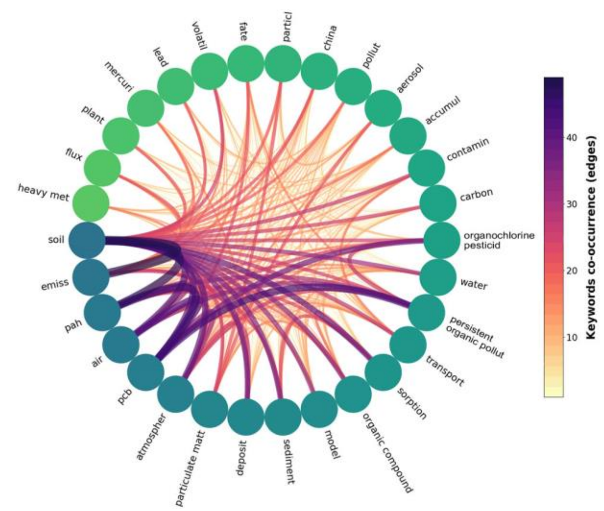
Air-Water



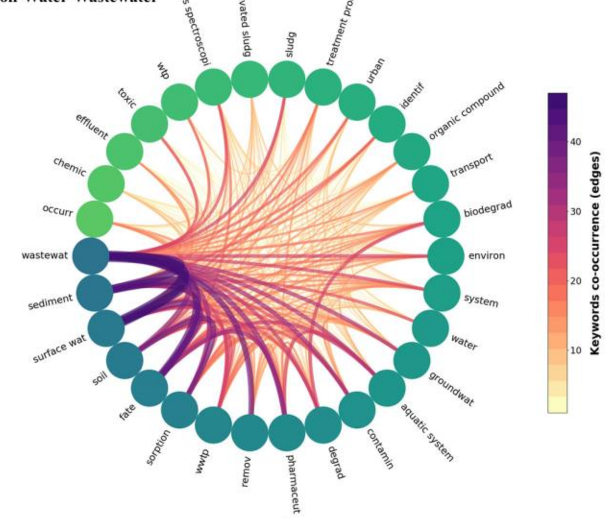
Air-Soil-Water



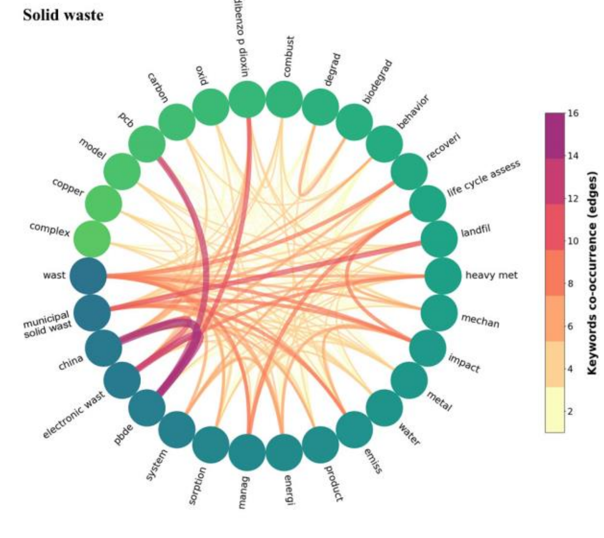
Air-Soil



Soil-Water-Wastewater



Solid waste



Air-Solid waste

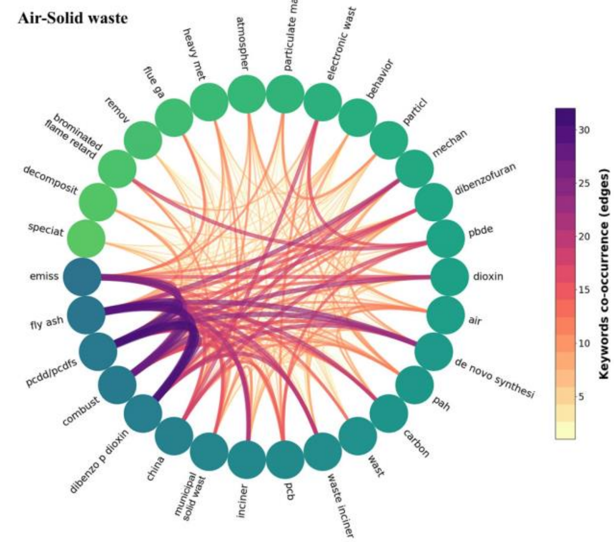


Figure S7. Co-occurrence of the top 30 frequent keywords (stemmed form) for each of the major 12 groups based on the circos plot. The keywords (nodes) are ordered by their overall frequency. Edge width and color are used to indicate the co-occurrence between keywords.

Text S4. Library science analyses

In library science, traditional methods for analyzing literature include bibliometric analysis such as those cited in the introduction, systematic reviews which synthesize the results of several similar studies, meta-analysis which uses statistical methods to analyze results of similar studies, and analysis tools provided by databases such as Web of Science. A search in Web of Science for the journal *Environmental Science & Technology* from 2000-2019 provides analysis of fields such as categories, publication years, document types, authors, organizations, countries of origin, and more.[8] Web of Science's automated analysis has limitations on selecting specific document types, so the analysis includes more documents than were used in this study. Web of Science Categories are included in the analysis instead of keywords. For the journal *Environmental Science & Technology* only two categories, "Engineering Environment" and "Environmental Studies", are applied across all articles published between 2000-2019. This analysis was not able to reveal emerging topics or research gaps. Similarly, the Web of Science automated analysis of the publication over time only provides data on the number of articles published as opposed to the analysis of keywords over time performed in this study. Web of Science limits the number of countries analyzed to 25. The numbers are slightly different because of the inability to select specific document types, but the rankings provided by Web of Science match those in this study. Scopus indexing of *Environmental Science & Technology* for the years 2000-2019 seems to be incomplete. Analysis provided by Scopus for a similar dataset provides the same level of granularity as compared to Web of Science.[9] In Scopus it is possible to view and limit based on keywords but no advanced analysis of keywords is available. In fact the top keyword available in Scopus is "Article" with 16,076 results. It is clear that the text mining approach presented in this study has provided a more in depth understanding of emerging topics and research gaps than searching directly in the database would provide.

Environmental Science & Technology is one journal among a whole ecosystem of interdisciplinary research. In addition to other peer reviewed journals related to the environment, research results are also disseminated through technical reports, government documents such as U.S. Geological Survey sources, and state government agencies.[10] Like the literature cited in the introduction, the analysis on Environmental Science & Technology in this study provides insight into a slice of environmental research. Other text mining studies vary widely in scope and breadth, but few are related to environmental studies. Rabiei et al. used text mining on search queries performed on a database in Iran to analyze search behavior.[11] Other studies examine text mining as a research tool, but using research from another discipline. In a text mining study on 15 million articles comparing the results of using full text versus abstracts, Westgaard et al. found that "text-mining of full text articles consistently outperforms using abstracts only".[12]

References

(1) NLTK is a natural language toolkit based on Python programs to work with human language data. https://www.nltk.org.

(2) Corbett, P.; Boyle, J. Chemlistem: Chemical Named Entity Recognition Using Recurrent Neural Networks. *J Cheminform* **2018**, *10* (1), 59. https://doi.org/10.1186/s13321-018-0313-8.

(3) Manning, C.; Raghavan, P.; Schütze, H. The term vocabulary and postings lists. In Introduction to Information Retrieval (pp. 18-44). Cambridge: Cambridge University Press. **2008**. https://doi.org/10.1017/CBO9780511809071.003.

(4) CS 276 / LING 286: Information Retrieval and Web Search. Course materials of Information retrieval at the Stanford University can be free retrieved from http://web.stanford.edu/class/cs276/.

(5) Porter, M. The Porter stemming algorithm. **2006**. http://snowball.tartarus.org/algorithms/porter/stemmer.html.

(6) Porter, M. Snowball: A language for stemming algorithms. **2001**. https://perun.pmf.uns.ac.rs/radovanovic/dmsem/cd/install/PorterStemmer/snowball/doc/introduction.html.

(7) High quality figures and other related materials may also be found or downloaded via the author's webpage. https://junjiezhublog.wordpress.com/tm/.

(8) Clarivate. *Web of Science Core Collection*. Results analysis of SO=(ENVIRONMENTAL SCIENCE TECHNOLOGY) AND PY=2000-2019. https://wcs.webofknowledge.com/RA/analyze.do?product=WOS&SID=5AbFDjnXgCO2d6ToezW&field=TASCA_JCRCategories_JCRCategories_en&yearSort=false.

(9) Elsevier. Scopus. Results analysis of ISSN(0013936x) AND ISSN(15205851) AND LIMIT-TO(PUBYEAR, 2000-2019) https://www.scopus.com/.

(10) Wild, E.; Havener, W. M. Online Bibliographic Sources in Hydrology. *Science & Technology Libraries*. **2001**, *21*, 63-86. https://doi.org/10.1300/J122v21n03_05.

(11) Rabiei, M.; Hosseini-Motlagh, S. M.; Haeri, A. Using Text Mining Techniques for Identifying Research Gaps and Priorities: a Case Study of the Environmental Science in Iran. *Scientometrics*. **2017**, *110*, 815-842. https://link.springer.com/article/10.1007/s11192-016-2195-8.

(12) Westergaard, D.; Stærfeldt, H.H.; Tønsberg, C.; Jensen, L. J.; Brunak, S. A. Comprehensive and Quantitative Comparison of Text-mining in 15 Million Full-text Articules Versus their Corresponding Abstracts. *PLOS Computational Biology*. **2018**. https://doi.org/10.137/journal.pcbi.1005962.