Why China Needs Data Sharing to Address Its Air Quality Challenge


Data sharing within China and bringing scientists together to come up with policy recommendations based on that data is at least as important as data collection in the fight against air pollution. The scale of data publicly available can have dramatic implications for the insights possible: a series of research studies in limited data contexts can lead to the wrong conclusions if the contexts do not encompass the broader picture or if they are not connected. Despite the clear benefits of data sharing, there are systemic hurdles in China: In the past, collection of air pollution data was fraught with distortion. Further, promotion and tenure requirements disincentivize data sharing essential to informative results. Finally, overcoming the data sharing and collaboration problem is not enough. As demonstrated by the California Air Resources Board and in air quality management in Houston, scientists must then come together around that shared data to inform the policy-making process.

As data sharing and collaboration there around can take on multiple forms, to elucidate our argument, we first differentiate between forms of data sharing. *Data Sharing* is the exchange of data between members of a group for an express purpose. Data sharing can take many shapes with varying numbers of researchers at the same or different institutions. A data sharing program may or may not involve making data public. *Making data public* is the most extreme form of data sharing, in as much as it makes the data available to everyone. In between public data sharing and data shared amongst only two entities is a data sharing club, where data is shared based on a set of common principles. Traditionally, an ambitious international collaboration addressing an issue with the complexity of air pollution would be the product of a small group with the ability to expand as the project develops. The types and scale of questions answerable with shared data can be significantly greater than any single group could collect and analyze on their own. Data sharing can conflict with traditional cultural standards of data collection and ownership in academia, without incentives otherwise. Houston, Texas’ history of addressing air pollution, discussed later in this paper, is a prime case-study in data sharing, without making the data public.
Data Pooling is the construction of a database to hold data that members, or in some extreme cases the general public, can access at will, rather than for a particular research or policy goal. It may be thought of as a more sophisticated extension of data sharing. Data pooling, when executed well, can have the advantage of ensuring validity and uniformity of data. Data pooling also has the potential to leverage shared resources to answer more complex, far-reaching questions, as well as allowing for new questions to be addressed as scientific understanding develops, encouraging the emergence of new knowledge. Pooling data can be difficult to implement as it requires buy-in from participants who may be accustomed to holding data independently to make the most individual use of it. It also requires significant maintenance and management to build, curate, and continually validate such a database. Databases such as those created by the Convention on Long-range Transboundary Air Pollution (CLRTAP) or the Aerosols, Clouds, and Trace gases Research Infrastructure (ACTRIS) -- both of which are discussed later in the paper -- are examples of data pooling.¹²

The Problem

In the midst of heightened tensions in United States (U.S.)-China relations on trade and the South China Sea, on June 5, 2016, John Holdren, Assistant to the U.S. President for Science and Technology and Director of the U.S. Office of Science and Technology Policy, and Wan Gang, Chinese Minister of Science and Technology, convened the U.S.-China Innovation Dialogue to discuss opportunities where collaboration could lead to improved outcomes for both nations. One focus of the meeting of the Dialogue’s expert group was the role that data sharing and collaborative analysis could play in tackling climate change and conventional air pollution.

Particulate matter in the atmosphere sized <2.5μm (PM_{2.5}) and ozone caused ~4.5 million premature deaths in 2005.³ A component of PM_{2.5}, black carbon, is a top contributor to global warming, second only to CO_{2}.⁴ Reduction in these pollutants leads to beneficial results for both the quality of the air we breathe and the pace and extent of climate change.⁵ Data sharing can be critical in effectively tackling this widespread issue. The scale of data that is currently publicly available can have dramatic implications
for the insights possible: a series of research studies in limited data contexts can lead to the wrong conclusions if the contexts do not encompass the broader picture or if they are not connected with each other. For example, in the early understanding of ozone depletion, many scientists and policy makers questioned the chlorofluorocarbon hypothesis, but with data collection across scales and platforms to inform a new policy approach, the hypothesis was confirmed.\textsuperscript{6}

Despite the clear benefits of data sharing and pooling, there are systemic hurdles in China. In the past, collection of air pollution data in China was fragmented and fraught with distortions.\textsuperscript{7,8,9} Most of the data was not publicly available. Moreover, China’s academic promotion practices have unintentionally hindered data sharing. For example, first or corresponding authorship is necessary for promotion and tenure. Because one, and at most two or three authors can be listed first or as corresponding authors, and because data sharing is typically conditional on being named as first author, publishing of data is discouraged, highlighting that co-authorship is not always a sufficient solution. Researchers and institutions also guard data in order to maximize the publishable material from its use. With high-quality sensors owned by different groups across China, and air pollution analysis requiring high-quality data from multiple sites within each region, these data sharing and collaboration barriers must be overcome to better reduce and manage air pollution in China.

Overcoming the data sharing problem is not enough. As demonstrated by the California Air Resources Board in Los Angeles and in air quality management in Houston, scientists must then together analyze the accumulated data to inform the policy-making process. Without this knowledge transfer, policy can easily be misinformed, misguided, and lead to undesired outcomes. Finally, sharing and using data permits improvement in the process of data collection itself.

**Lessons from History: Data Sharing Frameworks**

China is not alone in facing these challenges. Mandated data sharing protocols are still relatively new internationally in air quality research. In 1950, the World Meteorological Organization began to emphasize improved international collaboration through data pooling in the areas of weather, climate, and
water. This early effort faced challenges because the data was published without checks and balances to ensure equal quality and reliability, making it difficult to use for research. The WMO did instate its Global Atmospheric Watch program as a high-quality data pooling project that is still publicly available. The Convention on Long-range Transboundary Air Pollution was signed in 1979 and has been implemented by the European Monitoring and Evaluation Program (EMEP), creating a shared data network to address air quality issues across nations. In 1983, NASA’s Global Tropospheric Experiment (GTE) was created to study tropospheric chemistry and its effects on the U.S. GTE mandated a data management protocol as a condition of funding to ensure data is shared across parties with funded projects, and funding was also explicitly allocated for data management. Field data collection ended in 2001, but the data is still publicly available, highlighting the potential for a pooled and public resource. Eionet in the EU was created in 1997 as a data consortium among EU member and cooperating countries to collect, share, and study environmental data. In 2011, ACTRIS (Aerosols, Clouds, and Trace gases Research Infrastructure) was created to merge a group of programs (not including Eionet) dating back to 2000 into a combined database and network. ACTRIS is funded by the European Commission and jointly run by member-state representatives. Standards for data and meta-data reporting (including location, instrumentation, uncertainties/percentiles, etc.) are implemented in the management of the network in order to maintain quality across contributing sites. Top journals have also brought data sharing practices into their requirements for publishing. Nature’s policy reads “A condition of publication in a Nature journal is that authors are required to make materials, data, code, and associated protocols promptly available to readers without undue qualifications.” There is still work, however, to be done internationally. The United States does not have the equivalent of ACTRIS in place. Further the U.S. NSF’s funding for air quality monitoring projects has not historically allowed for dedicated data management roles, though sharing and eventually publicizing data is a current requirement.
Lessons from History: Scientists informing the policy process

There are important international examples of how scientists can inform the policy process, and how shared data practices between scientists and government regulators improve air quality policy. In the United States, the California Air Resources Board (CARB) has a long history of promoting and funding air quality research in California. Their work led to significant improvements in air quality throughout the state, often leading the nation. In the mid-20th century, the air quality in Los Angeles was degraded to an extent comparable to the worst found in Beijing today; annual-average PM$_{10}$ reached $\sim$150 µg/m$^3$ and peak ozone exceeded 600 ppb. The CARB set the first automobile NOx standards in the nation as a result of research into the effects of NOx on ozone formation and PM$_{2.5}$. As a consequence of these and subsequent effective emissions controls, LA air pollution levels are now less than one fourth of those in the past, even though population has doubled and vehicle miles traveled quadrupled. In California as a whole, the collective cancer risk from exposure to major toxic air contaminants declined 76% in the past 23 years alone. With the close ties that exist in China between government institutions and academia, the CARB’s experience could be a data sharing model from which to draw insights in the international context toward building cross-boundary projects and institutions.

Air quality management in Houston, Texas offers another story of the importance of collaboration between academia and government to drastically improve air quality. In 1999, while experiencing the worst ozone pollution in the United States, Houston embarked on an ambitious research collaboration among a variety of state, federal, and academic groups. Rather than NOx being the dominant cause of ozone formation, as in California, highly reactive volatile organic compounds (VOC) were found to be more critical ozone-forming agents. By rapidly summarizing the results of over 300 scientists’ work prior to any publications, the correct policy solution was effectively communicated to and acted upon by lawmakers. By regulating the highly reactive VOCs rather than NOx, industry saved an estimated 1 billion dollars. The LA and Houston cases underscore the importance of engaging scientists in the policy process and sharing data along the way. In these cases, opposite interventions were necessary to reduce ozone, due to regional geographic and industry differences.
The United States and China have a history of working together to bring science into the regulatory process. In 2003, the U.S. Environmental Protection Agency and China State Environmental Protection Agency (the predecessor to the Ministry of Environmental Protection) made history by signing a memorandum of understanding focused on collaboration on fuel and vehicle technology and standards. Although not shared with the greater public in China, American and Chinese scientists working on the projects shared data. Reports of some of the work were also published in the United States. In 2008, the Department of Energy’s Atmospheric Radiation Measurement Climate Research Mobile Facility was moved to southeastern China. While this was a positive step, data use and ownership issues forced the project to close down before it could gather the long-term data needed to gain scientific or policy insight. Most recently, U.S.-China cooperation on climate and energy under the Climate Change Working Group (CCWG) contributed to the successful negotiation of the Paris Agreement. The CCWG has already begun to promote data sharing between research institutions: The China Energy Modeling Forum was established in 2015 at Tsinghua to put modeling teams and policy makers on the same platform in order to better inform policy development. China has also begun the process of creating a pooled database of independently collected environmental data as well as consolidated data from a variety of sources made available through a government data center. Such efforts will greatly benefit from a broad approach addressing the systemic barriers to data sharing and pooling, as well as scientific informing of policy. While the challenges are not unique to China, as one of the top pollution emitters, China uniquely holds the opportunity to solve some of the most important air pollution and climate challenges of our time -- leading the world in these innovations and this dialogue. China has an impressive history of leading the world in the engineering and implementation of large-scale infrastructure innovations, such as it’s impressive achievements with high-speed rail. By leading the world in data pooling infrastructure and giving the very best scientists from China (and possibly also select friends from around the world) access to that pooled data with which to advance science and inform policy-makers, China has the potential to lead the world in solving arguably the most important problem facing the globe today.
**Proposed Actions**

1) Create incentives, such as national awards, for Chinese scientists undertaking joint research papers and the sharing of important data (e.g. to demonstrate that not only first or corresponding authorship count for promotion in China) in areas such as air pollution science and policy analysis.

2) Establish goals for all government-funded air pollution science data to be made public and shared two years after collection. Such goals might include that all data from publically funded research must be made public after two years.

3) Create the infrastructure and network for a commonly shared database aggregating data from both research supersites and government sites across China into a common, high-quality format. This infrastructure should be developed with the air quality - climate change - health nexus in mind. A model for the infrastructure network would be the European Aerosol, Clouds, and Trace gases Research Infrastructure (ACTRIS), CLRTAP, or the Global Atmospheric Watch.\(^{1,2,11}\)

4) Build a core platform for data collaboration among the countries with the most intensive commitment to scientific funding for combating climate change. Studies of international cooperation emphasize that smaller groups with higher initial investments in problem solving can speed up the process of establishing effective collaboration and create the conditions for successful subsequent expansion to a more inclusive global model.\(^{27,28}\)

i) Have the appropriate National Academies (including from the U.S., China, and Europe) invite their best scientists working on the air quality – climate change – health nexus to join a Blue Ribbon Panel to collaborate on development of the shared data infrastructure and accelerated policy analysis. A Blue Ribbon Panel is an expert advisory group convened to impartially inform a policy decision. Past U.S. Blue Ribbon Panel examples include the Kennedy Assassination committee and the 9/11 commission. Related examples, not necessarily called out as Blue Ribbon Panels include the National Academies committee which wrote “Rising Above the Gathering Storm,” and among
other things helped instigate the founding of ARPA-E, and the standing EPA Scientific Advisory Board, which advises the EPA on scientific methods and research programs.29,30

ii) Include a commitment by the Blue Ribbon Panel of communicating within one year consensus scientific findings (prior to publication) to policy makers.

iii) Include a commitment to have the data from the network fully shared and public.

iv) Set appropriate safeguards to ensure scientific rigor and transparency, including that the data would be made public in a timely manner.

v) Include a commitment to a timeline for full global inclusion in the Blue Ribbon Panel.

Finally, while international participation could bring the best minds from around the world to the urgent air quality issue in China, alone mandating data sharing within China and bringing Chinese scientists together to inform air quality policy would be an enormous step forward for air quality improvements in China and thus world-wide.

Acknowledgements:

Thanks to David Allen, Michael Brauer, Greg Carmichael, Neil Donahue, Baruch Fischhoff, Denise Mauzerall, Tom Mitchell, Chris Nielsen, Allen Robinson, Robert Sawyer, Jay Turner, and Alfred Wiedensohler for their time and input into the content of the document.

References


3. Lelieveld, J., Barlas, C., Giannadaki, D., and Pozzer, A. Atmos. Chem. Phys. 13, 7023–7037,


