fire

MDPI

Data Descriptor

Extensible Database of Validated Biomass Smoke Events for Health Research

Ivan C. Hanigan [1,2,3,*], Geoffrey G. Morgan [1,2], Grant J. Williamson [4], Farhad Salimi [1,2], Sarah B. Henderson [5], Murray R. Turner [6], David M. J. S. Bowman [4] and Fay H. Johnston [2,7]

1 The University of Sydney, University Centre for Rural Health, School of Public Health, Sydney 2006, Australia; Geoffrey.Morgan@sydney.edu.au (G.G.M.); Farhad.Salimi@sydney.edu.au (F.S.)
2 Centre for Air Pollution, Energy and Health Research (CAR), Sydney 2006, Australia; Fay.Johnston@utas.edu.au
3 Centre for Research and Action in Public Health, University of Canberra, Canberra 2617, Australia
4 School of Natural Sciences, University of Tasmania, Hobart 7005, Australia; Grant.Williamson@utas.edu.au (G.J.W.); David.Bowman@utas.edu.au (D.M.J.S.B.)
5 Environmental Health Services, British Columbia Centre for Disease Control, Vancouver, BC V5Z 4R4, Canada; Sarah.Henderson@bccdc.ca
6 Research and Information Services, University of Canberra, Canberra 2617, Australia; Murray.Turner@canberra.edu.au
7 Menzies Institute for Medical Research, University of Tasmania, Hobart 7005, Australia
* Correspondence: Ivan.Hanigan@sydney.edu.au; Tel.: +61-428-265-976

Received: 2 November 2018; Accepted: 3 December 2018; Published: 6 December 2018

Abstract: The extensible Biomass Smoke Validated Events Database is an ongoing, community driven, collection of air pollution events which are known to be caused by vegetation fires such as bushfires (also known as wildfire and wildland fires), or prescribed fuel reduction burns, and wood heaters. This is useful for researchers of health impacts who need to distinguish smoke from vegetation versus other sources. The overarching aim is to study statistical associations between biomass smoke pollution and health. Extreme pollution events may also be caused by dust storms or fossil fuel smog events and so validation is necessary to ensure the events being studied are from biomass. This database can be extended by contribution from other researchers outside the original team. There are several available protocols for adding validated smoke events to the database, to ensure standardization across datasets. Air pollution data can be included, and free software was created for identification of extreme values. Protocols are described for reference material needed as supporting evidence for event days. The utility of this database has previously been demonstrated in analyses of hospitalization and mortality. The database was created using open source software that works across operating systems. The prospect for future extensions to the database is enhanced by the description in this paper, and the availability of these data on the open access Github repository enables easy addition to the database with new data by the research community.

Keywords: health impacts; bushfire; smoke; dust; extreme air pollution

1. Introduction

Epidemiological Studies of Outdoor Air Pollution

For decades, researchers have studied the public health impacts of ambient outdoor air pollution, especially from the effects of particulate matter (PM) and gaseous pollutants associated with the combustion of coal, petroleum and biomass used for cooking [1]. Far fewer studies have examined the

effect of outdoor air pollution affected by intermittent extreme smoke events from biomass burning, such as that which occurs in bushfires, or from woodsmoke trapped by inversion layers during winter months as wood is burned for heating [2].

The epidemiological literature of health effects from ambient outdoor air pollution relating to smoke from biomass burning is much more limited than the vast literature on ambient PM from sources such as industrial and transport emissions [3]. Most literature available that focuses on biomass smoke health impacts looks at indoor pollution from cooking [4]. Particles in outdoor pollution from biomass smoke can directly influence the respiratory system through their inhalation and lodgment in the lungs where they promote inflammation and oxidative stress. Impacts on the cardiovascular system can take place through a range of processes including the promotion of systemic inflammation, blood coagulation, and impairing blood vessel function and autonomic reactivity [5]. Associations with diabetic, neurological, perinatal and other outcomes are increasingly being characterized [6].

Biomass smoke generally forms only a small fraction of the mixture of pollutants in the air. However, when a bushfire or inversion layer event occurs there is often a coincident spike in the pollution to extreme levels, primarily composed of biomass smoke. Such events provide the opportunity to study statistical associations between these pollution spikes and the health outcomes around those days. Anomalous levels of pollution can be arbitrarily defined using a threshold such as the 95th percentile, and days meeting this criterion might be assumed to be biomass smoke days, although other events might cause such spikes, such as dust storms, factory fires or even sea salt being driven by certain wind events. Therefore, a need exists to validate the dates on which events are ascribed in any correlational study of biomass smoke pollution spikes with health.

This paper aims to describe how a database was created to enable the collection of evidence linking historical spikes in air pollution with vegetation fire smoke. Our group has previously used the database in published research papers that quantify the health impacts of these extreme air pollution events from biomass fires [7,8]. This current paper describes how the database has been extended to be able to be distributed in an open, extensible format that allows the research community to add to the history of these events.

2. Materials and Methods

2.1. The Available Validation Protocols

The contributors to this database are requested to follow a protocol that defines the criteria they used to judge if an event date is considered validated. Further details about accessing the protocols, along with installation of the database, tools and documentation, are found in the Supplementary Materials below. The following protocols are described so that future studies that enter events into the database can follow a standardized protocol as part of their review of the evidence supporting each event. This information about the protocol used for each event's validation will then inform future users who can assess different events on the basis of the methods used and the amount (and types) of information used to validate them. Therefore, every event in the database has a link to the protocol used to identify it and an extraction from the database will show both the event and protocol(s) used.

2.2. The Johnston 2011 Protocol

The Johnston 2011 Protocol was the first method our team developed for this project and was published as a peer-reviewed journal article [9]. Detailed description of the Johnston 2011 protocol is included with the database in a document available from the internal web2py URL ./biomass_smoke_events_db/static/protocols.html. This protocol is considered the most conceptually appealing method to-date. In this protocol, the first step is to collect the longest available time-series of daily PM air pollution for each location. In our original study there were up to 13 years (between 1994 and 2007) of daily air quality data measured as Particulate Matter (PM) less than 10 μm (PM_{10}) or less than 2.5 μm ($PM_{2.5}$) in aerodynamic diameter. Air pollution data were provided by government

agencies in each Australian state. Daily averages for each site were calculated, excluding days with less than 75% of hourly measurements available.

In Sydney and Perth, where data were collected from several monitoring stations, the daily site-specific PM concentrations had missing data gaps which were filled using imputation using available data from other proximate monitoring sites in the network. The daily city-wide PM concentrations were then estimated following the protocol of the Air Pollution and Health: a European Approach (APHEA) studies [10,11]. A range of sources were then examined to validate these by identifying the cause of air pollution. Sources included online news archives, Internet searches for reports by government and research agencies, satellite imagery and the 'Dust Event Database' (DEDB) held at Griffith University, Brisbane, Australia [12]. Remote sensing products were obtained from NASA, primarily from the MODIS instruments [13,14], including both visible and aerosol optical depth (AOD) images, and these were examined to provide further information about days for which the other methods did not find any valid references. All Internet and news archive searches used the event type search word terms of 'bushfire' or 'fuel-reduction/prescribed burn' or 'smog' or 'dust' or 'haze'.

2.3. The Morgan 2010 Protocol

This protocol was developed by one of our colleagues and a co-author (Dr. Morgan) for a study in Sydney, Australia [15]. The procedure is very similar to the Johnston 2011 Protocol in that the 'potential event dates' are identified as days with city-wide 24 h average PM concentrations greater than the 99th percentile for the study period. These dates are then validated as either bushfires or fuel-reduction burns on or immediately prior to these days by checking newspaper archives and any other sources. The main differences are that the 99th percentile is only used (instead of 95th and 99th), event type search word terms are restricted to 'bushfire' or 'fuel-reduction/prescribed burn' (not extended to include 'smog', 'dust' or 'haze'), and there was no systematic review of satellite images.

2.4. The Salimi 2016 and 2017 Protocols

In 2016 one of our colleagues and a co-author (Dr. Salimi) extended the biomass smoke database for Sydney. That project developed a refinement of the Johnston 2011 Protocol in which only satellite images and news archives were used. In the Salimi 2016 Protocol the air pollution data was processed in the same way as the Johnston Protocol. In 2017 Dr. Salimi applied these techniques to the city of Melbourne, Australia and in addition to satellite and news data on the same day and days prior, evidence was sought in searches of the government Environmental Protection Agency (EPA) reports.

2.5. The Bare Minimum Protocol

The Bare Minimum Protocol was developed for this paper. In this protocol all that is required for an event to be validated is any reference that the contributor deems relevant. It is desirable that they add as much detail as possible to the database (e.g., author, title, publisher, year, URL, and date accessed). To date only two references have been inserted using this protocol [16,17] but considering the greater ease with which contributors may validate events in this way it is envisaged that this protocol will prove popular. It is envisaged that this method will allow the database to capture more events in an opportunistic way as many sources of information will become available in an ad hoc fashion. However, this method is the least conceptually appealing because it results in a collection of events from times and places that have had unequal amounts of research effort expended on finding evidence (e.g., differential sampling intensity), and may not be 'missing at random' and therefore contain systematic biases.

2.6. Selecting Which of the Protocols to Follow

Any of the protocols defined above can be used, or the contributor can create their own. If they create their own protocol then this must be shared along with the events data contributed for inclusion

back into the master copy of the database. This flexibility to allow multiple approaches is a strength of this database, but also presents the users with some challenges. On the one hand it is beneficial that decisions on the protocol to be used can be made based on the resources available that can be allocated to the effort. On the other hand, this presents difficulty to select which protocol to follow.

3. Results

The Biomass Smoke Events Database System Design

The system is described in Figure 1. The procedure starts with the master copy of the database that is maintained by the Data Manager (DM) in our group. The DM extracts a snapshot of the database (with a specific version identifier from the Git version control system) and makes a 'standalone' version available on Github. This standalone version uses web2py and SQLite, open source applications, so that it is capable of being downloaded and run on any operating system used by other computers. Contributors may download that version and use it as a local database.

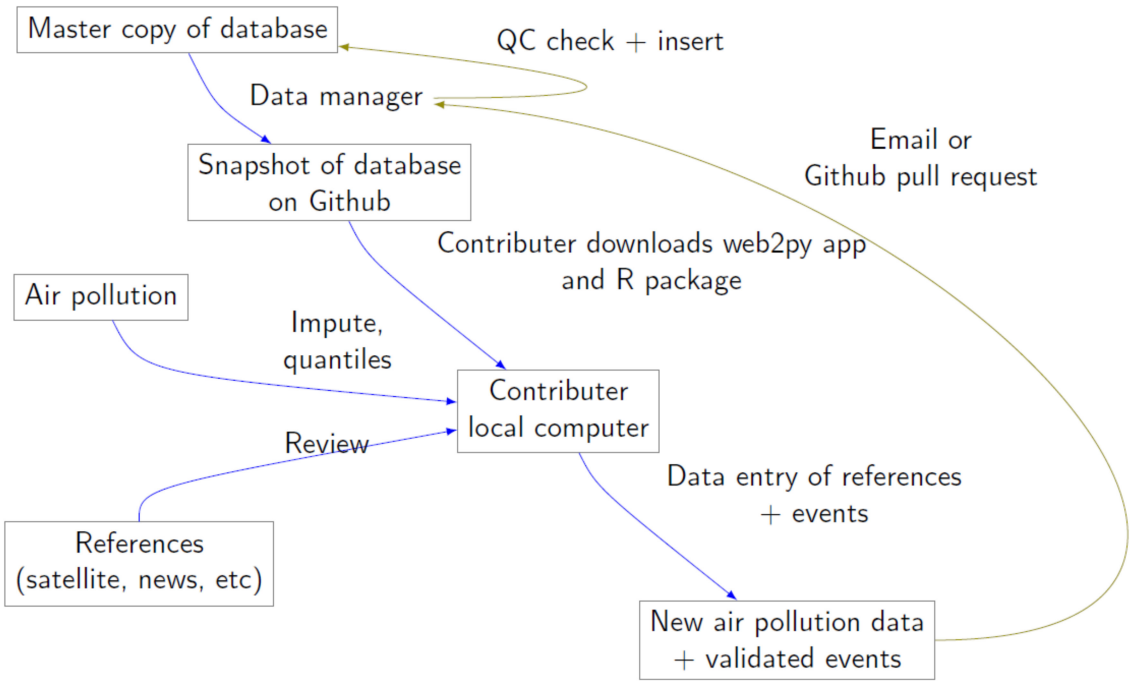

Figure 1. Schematic diagram of the processes for extending the database.

If following the Johnston 2011 or Morgan 2010 Protocols, the contributor needs to have daily air pollution data available, and access to the required reference materials for validation (e.g., satellite images, newspaper archives, the Dust Storms database). If the user follows the Salimi 2016 or 2017 Protocols they only require daily air pollution, news, reports and satellite images. If they are following the Bare Minimum Protocol then they only require the validation reference document.

The R package we developed (also available as open source code at the Github repository: https://github.com/swish-climate-impact-assessment/BiosmokeValidatedEvents) contains functions that may be used to impute any missing data gaps using the procedure as per the APHEA2 study protocol [11]. The R package is used by the Johnston and Salimi Protocols to compute the quantiles of the new extended time-series of imputed pollution data, to identify events above the 95th percentile threshold that has been set to define 'extreme events'. The contributor then must review these 'potential event dates' to find validation references, and then uses the web2py data entry forms to add the information. Once they complete their review of all events they can notify the DM either with email or by using the Github 'pull request' feature. The DM performs some basic Quality Control (QC)

checks to ensure the protocol is described and that the new events data is easy to merge with the old database (for example place names are spelt the same). and then uploads the new data to the online database. The procedure then starts again when a new version is loaded into the Github repository with descriptions of the additional changes that have been incorporated.

4. Discussion

4.1. Examples of Studies That Will Benefit from This Database

The availability of this database makes it easier to do new research that builds on top of the old database to incorporate new events and add to knowledge of the health impacts of biomass fires. This is shown by a recent paper on the health burden associated with fire smoke in Sydney, New South Wales (NSW), Australia's largest city [18]. In that study our events database was extended by the new authors from the original end-point in 2007 to include six more years of data and the time series now finishes in 1 January 2014 for that city. These new data have been added to the database. Using the extended dataset the authors estimated that around 200 premature deaths were attributable to fire smoke over the 13 years studied, compared to just 77 deaths in NSW previously estimated to have been directly attributable to bushfires during the entire 110 years between 1901 and 2011 [19]. This indicates a previously underappreciated burden of disease from biomass fire smoke.

In a related example the recently published Australian 'Countdown on health and climate change' [20] was limited for the indicator of lethal weather-related disasters because it is known that estimates of the deaths attributable to bushfires and biomass smoke in Australia based on historical data is underestimated. This was noted as a problem in the international 'EM-DAT' database of disasters whereby the estimate of deaths from bushfires was found to have large discrepancies with estimates from other data sources. For example, Blanchi et al. [19] identified 825 deaths from direct exposure to bushfires in Australia between 1901 and 2011, compared to just 501 between 1900 and 2017 from EM-DAT, and so the results in the Countdown report are likely underestimated. As the Countdown report aims to release annual updates which will track these indicators of the health impacts of weather-related disasters such as bushfires due to climate change, the future work will be able to combine multiple data sources, including our biomass smoke events database, to mitigate the major limitations of the EM-DAT database in terms of identifying bushfire events in Australia. Such a data integration task would have been much more difficult without the development of this extensible validated events database.

4.2. Differential Sampling Intensity and Potential Exposure Misclassification Bias

Providing a number of Event Validation Protocols options for contributors to follow is important. In addition to the database's open format the flexibility of following a protocol that meets the resources available to a contributor increases the likelihood of data being contributed. This will add to the utility of the database for those researching the health effects from ambient outdoor air pollution relating to smoke from biomass burning.

The Event Validation Protocols described in this paper are all conceptually appealing because they allow a collection of events from times and places if evidence is available from the sources. Unfortunately, the end result of combining these data into a single database is that the derived dataset is made up of components which have had unequal amounts of research effort expended on finding evidence (e.g., differential sampling intensity), as well as different search criteria used for finding the references to support events, and may not be 'missing at random' and therefore contain systematic biases, which is a problem for statistical analysis.

This raises the potential for bias by exposure misclassification, which would occur by classifying actual fire smoke/dust days as non-fire smoke/dust days, or classifying non-fire/dust days as actual fire/dust days. The impact of exposure misclassification will of course be related to the particular study design implemented with the fire smoke database. For time series studies the issue is discussed

briefly in Morgan et al. [15]. They explain that missing some bushfire days would reduce the power of the analysis to find an effect (if one is present), but it would be unlikely to bias the result. Because fire smoke/dust incidents are rare and PM is usual relatively low in Sydney (and in most other Australian cities) it is possible to categorize any day as having either "Biomass Smoke Event" PM or "background" PM.

Morgan et al. [15] included this background PM explicitly in their model to capture differences with the Biomass Smoke Event days. It is possible such an approach will include a small number of extra bushfire days with days categorized as background days.

Morgan et al. argue that any such inclusions would be unlikely to influence the background PM results due to the large number of non-bushfire days in a multi-year study period. The sensitivity analysis they conducted did not categorize daily PM into bushfire PM and background PM. They found results similar to those reported for background PM. This suggests that including additional bushfire days with non-bushfire days in the background PM analysis would not bias their PM results.

As part of our database design we aimed to minimize this risk because the database clearly identifies the amount and type of information sources used for each event and the validation protocol used in each review. This serves as a flag to communicate to the user the amount of certainty there is about each event being from biomass smoke (and whether from wildfires, prescribed burns, woodheaters or dust, or some combination). Future users can then assess if the set of events they extract from the database meet the level of certainty required for their study, which will be based on their research questions and the inferences they aim to make from their results. A more detailed discussion of these issues is outside the scope of this brief database description paper.

5. Conclusions

This open and extensible database was developed by the authors to identify historical spikes in particulate matter concentrations and to evaluate whether they were caused by vegetation fire smoke or by other means. A detailed explanation of the development of the original protocol used to create the database and a summary of the data we originally collected is published already [9]. This current paper describes how the database has been extended to be able to be distributed in an open, extensible format that allows the research community to add to the history of these events. Having this database constantly updated will allow new researchers to study the health impacts of biomass smoke based on a strong historical foundation of validated events as well as take advantage of (and add to) new knowledge about recent trends and changes of biomass burning. Changes in biomass burning are already occurring associated with climate change and there is the beginning of an 'energy transition' that is hoped will mitigate the emissions of greenhouse gases as well as produce co-benefits by reducing exposure of humans to the harmful fire smoke pollution. This database will support public health research on this important topic.

Supplementary Materials: The database generated during the current study is available in the Github repository (https://github.com/swish-climate-impact-assessment/biomass_smoke_events_db). An R package was written to support the air pollution data processing and is available at Github: https://github.com/swish-climate-impact-assessment/BiosmokeValidatedEvents. Different operating systems are catered for by both the R package and Web2py Data Entry forms (Unix, Mac and Windows operating systems). The main programming languages (R and SQL) required to use the data are widely used in the environmental science community. The PostgreSQL database (with PostGIS spatial extension) is recommended, but the more accessible SQLite database can also be used. The license is open access Creative Commons Attribution version 4.0 (CCBY4.0). There are no major restrictions to use, however any amendments of errors are invited but will be vetted before insertion into the master database by the original project team.

Author Contributions: Conceptualization, I.C.H., G.G.M., G.J.W., S.B.H., D.M.J.S.B. and F.H.J.; Methodology, I.C.H., G.G.M., G.J.W., S.B.H., D.M.J.S.B. and F.H.J.; Validation, I.C.H., G.G.M., G.J.W., S.B.H. and M.R.T.; Writing-Original Draft Preparation, I.C.H., G.G.M., G.J.W., M.R.T. and F.H.J.; Writing-Review & Editing, I.C.H., G.G.M., G.J.W., F.S., S.B.H., M.R.T., D.M.J.S.B. and F.H.J.; Funding Acquisition, G.G.M., M.R.T., D.M.J.S.B. and F.H.J.

Funding: This research was funded by grants from the Australian Research Council (LP0882048) and the National Health and Medical Research Council (490057). We would like to gratefully acknowledge the Australian National Data Service (ANDS) for funding support for this project through the Collection Enhancement Projects program and also the ANDS funding received for the SWISH Climate Impacts project (project code AP07). The funding bodies had no role in the design of the study, collection, analysis, and interpretation of data or writing the manuscript.

Acknowledgments: We thank Talia Portner for valuable assistance in literature review for the original database.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, and in the decision to publish the results.

References

1. Pope, C.A., III; Dockery, D.W. Health effects of fine particulate air pollution: Lines that connect. *J. Air Waste Manag. Assoc.* **2006**, *56*, 709–742.
2. Naeher, L.P.; Brauer, M.; Lipsett, M.; Zelikoff, J.T.; Simpson, C.D.; Koenig, J.Q.; Smith, K.R. Woodsmoke health effects: A review. *Inhal. Toxicol.* **2007**, *19*, 67–106. [CrossRef] [PubMed]
3. Reid, C.E.; Brauer, M.; Johnston, F.H.; Jerrett, M.; Balmes, J.R.; Elliott, C.T. Critical review of health impacts of wildfire smoke exposure. *Environ. Health Perspect.* **2016**, *124*, 1334–1343. [CrossRef] [PubMed]
4. Smith, K.R. Fuel combustion, air pollution exposure, and health: The situation in developing countries. *Annu. Rev. Energy Environ.* **1993**, *18*, 529–566. [CrossRef]
5. Franklin, B.A.; Brook, R.; Pope, C.A. Air pollution and cardiovascular disease. *Curr. Probl. Cardiol.* **2015**, *40*, 207–238. [CrossRef] [PubMed]
6. Rückerl, R.; Schneider, A.; Breitner, S.; Cyrys, J.; Peters, A. Health effects of particulate air pollution: A review of epidemiological evidence. *Inhal. Toxicol.* **2011**, *23*, 555–592. [CrossRef] [PubMed]
7. Johnston, F.; Hanigan, I.; Henderson, S.; Morgan, G.; Bowman, D. Extreme air pollution events from bushfires and dust storms and their association with mortality in Sydney, Australia 1994–2007. *Environ. Res.* **2011**, *111*, 811–816. [CrossRef] [PubMed]
8. Martin, K.L.; Hanigan, I.C.; Morgan, G.G.; Henderson, S.B.; Johnston, F.H. Air pollution from bushfires and their association with hospital admissions in Sydney, Newcastle and Wollongong, Australia 1994–2007. *Aust. N. Z. J. Public Health* **2013**, *37*, 238–243. [CrossRef] [PubMed]
9. Johnston, F.H.; Hanigan, I.C.; Henderson, S.B.; Morgan, G.G.; Portner, T.; Williamson, G.J.; Bowman, D.M. Creating an integrated historical record of extreme particulate air pollution events in Australian cities from 1994 to 2007. *J. Air Waste Manag. Assoc.* **2011**, *61*, 390–398. [CrossRef] [PubMed]
10. Atkinson, R.W.; Anderson, H.R.; Sunyer, J.; Ayres, J.; Baccini, M.; Vonk, J.M.; Boumghar, A.; Forastiere, F.; Forsberg, B.; Touloumi, G. Acute effects of particulate air pollution on respiratory admissions. *Am. J. Respir. Crit. Care Med.* **2012**. [CrossRef] [PubMed]
11. Katsouyanni, K.; Schwartz, J.; Spix, C.; Touloumi, G.; Zmirou, D.; Zanobetti, A.; Wojtyniak, B.; Vonk, J.; Tobias, A.; Pönkä, A. Short term effects of air pollution on health: A European approach using epidemiologic time series data: The APHEA protocol. *J. Epidemiol. Community Health* **1996**, *50*, S12–S18. [CrossRef] [PubMed]
12. McTainsh, G.H. Dust storm index. In *Sustainable Agriculture: Assessing Australia's Recent Performance: A Report to Standing Committee on Agriculture and Resource Management (SCARM) of the National Collaborative Project on Indicators for Sustainable Agriculture*; National Collaborative Project on Indicators for Sustainable Agriculture: Collingwood, Australia, 1998; pp. 65–72.
13. Justice, C.; Giglio, L.; Korontzi, S.; Owens, J.; Morisette, J.; Roy, D.; Descloitres, J.; Alleaume, S.; Petitcolin, F.; Kaufman, Y. The MODIS fire products. *Remote. Sens. Environ.* **2002**, *83*, 244–262. [CrossRef]
14. Justice, C.; Townshend, J.; Vermote, E.; Masuoka, E.; Wolfe, R.; Saleous, N.; Roy, D.; Morisette, J. An overview of MODIS land data processing and product status. *Remote. Sens. Environ.* **2002**, *83*, 3–15. [CrossRef]
15. Morgan, G.; Sheppeard, V.; Khalaj, B.; Ayyar, A.; Lincoln, D.; Jalaludin, B.; Beard, J.; Corbett, S.; Lumley, T. Effects of bushfire smoke on daily mortality and hospital admissions in Sydney, Australia. *Epidemiology* **2010**, *21*, 47–55. [CrossRef] [PubMed]
16. Kjellstrom, T.; Kingsland, S.; Hanigan, I. *Health Impacts on Canberra Residents of Smoke from the January 2003 Bushfires: Report to ACT Health*; ACT Government: Canberra, Australia, 2004.

17. Williamson, G.; Bowman, D.M.S.; Price, O.F.; Henderson, S.; Johnston, F. A transdisciplinary approach to understanding the health effects of wildfire and prescribed fire smoke regimes. *Environ. Res. Lett.* **2016**, *11*, 125009. [CrossRef]
18. Horsley, J.A.; Broome, R.A.; Johnston, F.H.; Cope, M.; Morgan, G.G. Health burden associated with fire smoke in Sydney, 2001–2013. *Med. J. Aust.* **2018**, *208*, 309–310. [CrossRef] [PubMed]
19. Blanchi, R.; Leonard, J.; Haynes, K.; Opie, K.; James, M.; Oliveira, F.D. Environmental circumstances surrounding bushfire fatalities in Australia 1901–2011. *Environ. Sci. Policy* **2014**, *37*, 192–203. [CrossRef]
20. Zhang, Y.; Beggs, P.J.; Bambrick, H.; Berry, H.L.; Linnenluecke, M.K.; Trueck, S.; Alders, R.; Bi, P.; Boylan, S.M.; Green, D.; et al. The MJA-Lancet Countdown on health and climate change: Australian policy inaction threatens lives. *Med. J. Aust.* **2018**, *209*, 1.e1–1.e21. [CrossRef]

© 2018 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (http://creativecommons.org/licenses/by/4.0/).