HHS Public Access

Author manuscript

Environ Model Softw. Author manuscript; available in PMC 2016 December 01.

Published in final edited form as:

Environ Model Softw. 2015 December 1; 74: 238–246. doi:10.1016/j.envsoft.2015.06.003.

Integrating modelling and smart sensors for environmental and human health

Stefan Reis[a,j], **Edmund Seto**[b], **Amanda Northcross**[c], **Nigel W.T. Quinn**[d], **Matteo Convertino**[e], **Rod L. Jones**[f], **Holger R. Maier**[g], **Uwe Schlink**[h], **Susanne Steinle**[k], **Massimo Vieno**[a], and **Michael C. Wimberly**[i]

[a]Natural Environment Research Council, Centre for Ecology & Hydrology, Bush Estate, Penicuik, EH26 0QB, United Kingdom

[b]Department of Occupational & Environmental Health Sciences, School of Public Health, University of Washington, 1959 Pacific Street, Seattle, WA 98195, USA

[c]Department of Environmental and Occupational Health, George Washington University, 950 New Hampshire Ave. NW, Washington, DC 20052, USA

[d]HydroEcological Engineering Advanced Decision Support, Berkeley National Laboratory, 1 Cyclotron Road Bld 64-209, Berkeley, CA 94720, USA

[e]HumNat Lab, School of Public Health - Division of Environmental Health Sciences & Health Informatics Program, Institute on the Environment and Institute for Engineering in Medicine University of Minnesota, Twin-Cities USA

[f]Department of Chemistry, Cambridge University, United Kingdom

[g]School of Civil, Environmental and Mining Engineering, The University of Adelaide, Adelaide SA 5005, Australia

[h]UFZ Helmholtz Centre for Environmental Research, Permoserstraße 15, 04318 Leipzig, Germany

[i]Geospatial Sciences Center of Excellence, South Dakota State University, Brookings, SD 57007-3510, USA

[j]University of Exeter Medical School, Knowledge Spa, Truro, TR1 3HD, United Kingdom

[k]Institute of Occupational Medicine, Research Avenue North, Riccarton, Edinburgh, EH14 4AP, United Kingdom

Abstract

Sensors are becoming ubiquitous in everyday life, generating data at an unprecedented rate and scale. However, models that assess impacts of human activities on environmental and human

Corresponding author: Stefan Reis, srei@ceh.ac.uk - phone: +44 131 445 8507.

health, have typically been developed in contexts where data scarcity is the norm. Models are essential tools to understand processes, identify relationships, associations and causality, formalize stakeholder mental models, and to quantify the effects of prevention and interventions. They can help to explain data, as well as inform the deployment and location of sensors by identifying hotspots and areas of interest where data collection may achieve the best results. We identify a paradigm shift in how the integration of models and sensors can contribute to harnessing 'Big Data' and, more importantly, make the vital step from 'Big Data' to 'Big Information'. In this paper, we illustrate current developments and identify key research needs using human and environmental health challenges as an example.

Keywords

integrated modelling; environmental sensors; population health; environmental health; big data

1. Introduction

1.1. Background

Models have become widely used and indispensable tools to assess effects of environmental factors on human and ecosystem health. Applications include, but are not limited to, the modelling of environmental processes, such as the emission, dispersion and environmental fate of pollutants in atmospheric (e.g., Vieno *et al.*, 2010, 2014), terrestrial and aquatic environments (e.g., Wu *et al.*, 2014a, b; Perelman and Ostfeld, 2013), the quantification of human exposures to these pollutants (e.g., McKone, 1993; MacIntosh *et al.*, 1995), the risks and public health burdens from exposures to environmental pollutants (e.g., Lim *et al.*, 2012; Schlink *et al.*, 2010, the dynamics of biomarkers in relation to drugs and pathogens, and the efficacy of efforts to control the consequences of these processes on human health (e.g., May *et al.*, 2008; Wu *et al.*, 2014b), and the quantification of stakeholder mental models for optimal decision making (Wood *et al.*, 2012; Voinov *et al.*, 2014; Boschetti, 2015). Models have important uses in examining the accidental or natural release of chemicals, radionuclides, volcanic ash, or pathogens in the environment. Generally, both physical process-based and statistical models are calibrated and validated against observed environmental data, which have traditionally been obtained from few, typically sparsely distributed routine monitoring stations, or from costly short-term field measurement studies. In both cases, the spatial and temporal performance of models is evaluated against relatively few directly measured data points.

Conversely, the capabilities and availability of cheaper, more sensitive and sophisticated sensors for gases, particulates, water quality, noise and other environmental measurements have improved and are enabling researchers to collect data in unprecedented spatial, temporal and contextual detail (Stocker *et al*., 2014). These sensors range from bespoke devices designed for specific applications, to those found on more mainstream personal devices, such as smartphones. In some cases, people may act as environmental sensors by reporting what they see, hear and feel by participating in the crowdsourcing of environmental conditions (Salathe' *et al.*, 2012). By leveraging widely available computing, networking and sensor technologies, many new sensor systems are relatively low-cost

compared with technologies used in established monitoring networks. Low-cost sensing has the potential to broaden the scale of environmental measurements, both through improving the feasibility of larger scale monitoring networks and by empowering non-traditional researchers, such as community groups, environmental justice organizations and citizen scientists to participate in collecting environmental, biological and clinical data. Hence, new sensors may potentially solve the limitations of traditional environmental monitoring by improving data collection in currently under-monitored areas, including urban areas with large spatio-temporal variations in pollutant concentrations and exposures, as well as rural areas and developing countries where few conventional monitoring sites may be available. One challenge of ubiquitous sensing is a potential explosion of data collected by multiple groups for different purposes, with differing accuracy, precision and hence data quality. Advances in data science and data fusion are vital to enable researchers to make best use of the vast amounts of additional, heterogeneous measurement data. Environmental models will potentially play an important role in integrating these data as inputs to refine and quantify important environmental relationships and processes (Banzhaf *et al.*, 2014; Galelli *et al.*, 2014). Models may also benefit from having new data to use as calibration, validation, and assimilation points to improve the outputs of increasingly complex and downscaled models. Documenting, understanding and implementing quality assurance and quality control processes that are responsive to heterogeneous sensor data will be critical if they are to be used for modelling. Modellers are not only users of sensor data, but can also help to inform the sensor community by identifying existing modelling uncertainties, sensitivities, and constraints that could benefit from improved empirical data, so as to guide what, when and where sensors should measure. Ultimately, data from both sensors and models provides evidence to policy decision-makers, hence the role of stakeholders and their interaction with the scientific community is a vital area for discussion in this context.

1.2. Approach

This paper presents the potential benefits and opportunities available to the modelling community through improved adoption and integration of sensor technologies. For the purpose of this paper, we use the term ‘data’ to specifically identify raw and unprocessed observations specifically, and ‘information’ to illustrate data that has undergone validation, quality assurance/quality control (QA/QC) and (objective-based) interpretation to be used for decision making. Finally, as ‘Big Data’ does not have a concise and generally accepted, scientific definition to date (the moving target presented by defining a volume of data that is pushing the boundaries of current processing capabilities), we adopt the widely used definition by Doug Laney and applied by industry (e.g. SAS, 2015), which stipulates ‘Big Data’ as being determined by the three Vs, *volume*, *velocity* and *variety*. These three aspects are important when monitoring a wide variety of data and are therefore highly relevant to the purposes of this paper.

We discuss cases in which models may benefit from large datasets emerging from new sensor networks, particularly in terms of increased model accuracy through better calibration/validation and global uncertainty/sensitivity analyses (Saltelli *et al.*, 2010), while also benefiting groups designing, deploying, and analysing data from sensor networks. Figure 1 presents a conceptual framework in which both the sensing and modelling

communities play integral roles in information science, with this science ultimately operating within and informing policy. Critically, missing from this conceptual diagram are the details of data management, processing and flows.

The environmental monitoring community produces data that are subject to QA/QC, which then could be used on their own as empirical data related to environmental processes. However, data could also flow to the modelling community as inputs and calibration and validation points for modelling. The combination of quality data and a validated conceptual model that incorporates state of the science understanding of environmental and disease processes can be explored via simulation, scenario, and global sensitivity and uncertainty analyses to produce information relevant for policy and planning. In this framework, we acknowledge that all measurement data are subject to error, and can benefit from QA/QC to filter the data for errors and anomalies leading to the use of models for data synthesis. Models can also vary in complexity and accuracy, however, even spatial and temporal smoothing of data can serve as a useful, yet relatively simple, form of model to aid in visualizing temporal trends and spatial gradients relevant to many environmental processes. The combined use of data and models at different spatio-temporal scales can serve to identify scale-dependent and universal relationships between potential causal factors and outcomes of interest.

Conceptual models can be extended to be more sophisticated, coupling separate sub- or component models of pollutant emissions, fate and transport, multiple routes of exposure and dose-response relationships to assess health impacts, and might utilize a variety of sensor data to inform the processes and relationships coded into each sub model. At different points in our model framework, there are uncertainties in the sensing, data collection, and modelling processes that ultimately affect the confidence with which we are able to apply information to the planning and policy process. Thus, the information required should ultimately be the driver of any step, rather than what raw data can be generated, and global sensitivity and uncertainty analyses can guide information creation and model and surveillance network design. We are guided by the following key questions, which are addressed in the remainder of the paper:

1. How can modellers best make use of the ‘Big Data’ emerging from current and next-generation smart sensor networks?(Section 2)
2. What are the key challenges for model-sensor integration across temporal and spatial scales? (Section 3)
3. Can model-sensor integration improve the quantification of uncertainty by addressing issues of precision and accuracy in current exposure assessment techniques for health impact assessment? (Section 4)
4. Can integrated model-sensor approaches improve understanding of the associations and causality of environmental determinants for human health effects? (Section 5)
5. What are the critical research questions and knowledge gaps that can improve progress with model-sensor integration? (Section 6)

6. How can sensor and model outputs be best communicated to a wide variety of potential decision makers and stakeholders? (Section 7)

While answering these questions in detail is outside of the scope of a single paper, we will provide examples of applications and approaches which can contribute to a better, more comprehensive integration of sensors and models in the subsequent sections.

2. The emergence of 'Big Data' and what it means for modelling

One of the most important questions posed by the advent of 'Big Data' that integrate modelling and smart sensors is: *How can modellers make best use of the additional data?* In order to answer this question, we discuss the potential paradigm shift precipitated by the availability of 'big data', we examine the potential benefits 'Big Data' offers to the field of modelling and we highlight the importance of considering the implications of 'Big Data' on quality assurance and control processes.

2.1 How can smart sensor networks change the game

Today's earth systems science is an archetype for how sensing and modelling systems might be integrated in near real-time. Global climate models are available and interconnected with sensing data systems, and are currently used for science and policy purposes. Satellite and surface-based measurement campaigns have adopted standards and recognized practices for data collection, metadata documentation, production of data products and access to data and products via the Internet (e.g., U.S. Geological Survey Global Earth Observation System of Systems (GEOSS) atmospheric, land cover, land- and sea surface temperature, albedo, and other remotely sensed products, National Centre for Atmospheric Research (NCAR) and World Meteorological Organisation (WMO) databases for global surface monitoring data). The aforementioned standards and practices are critical, due to the volume of data that is produced by these sensing systems and the speed with which they become available. For example, the National Aeronautics and Space Administration's (NASA) *Earth Observing System and Data Information System* (EOSDIS), as of September 2013, contained 9.8 petabytes of data and served 1.7 million users (EOSDIS, 2014).

Consequently, the modelling community has incorporated these data. The *MM5 Community Model* (http://www2.mmm.ucar.edu/mm5/) and *Community Multi-scale Air Quality Model* (CMAQ, http://www.epa.gov/AMD/Research/RIA/cmaq.html) meteorology and air pollution modelling groups routinely make use of existing monitoring data, both for model input, calibration, and validation. Moreover, these groups routinely develop solutions to the challenges of sensor fusion and the merging of data collected from different instruments with different spatial and temporal resolutions, and as a result they are able to up/downscale their models to best fit available data and answer policy questions. The hydrological sciences provide another example in which there are already large datasets and initiatives underway to develop data-driven modelling methods, such as the *Panta Rhei - Change in Hydrology and Society* initiative (http://distart119.ing.unibo.it/pantarhei/).

While established *Earth System Modelling* (ESM) demonstrates that even complex data-model coupling is possible, new, rapidly developing movements like Smart Cities and

Citizen Science are potentially game-changing in terms of the amount, variety, and improved spatial temporal resolution of sensor data that can potentially be integrated into models. Urban processes, such as roadway traffic, are already monitored regularly using sensors, producing data useful for traffic demand management. Unlike the aforementioned ESMs, there is great potential to couple these real-time data with real-time models to create "closed-loop" control systems, allowing for sensed data to result in immediate actions (Hilty *et al.*, 2014). Although this has yet to be fully realized, some applications, such as congestion-driven road and parking pricing schemes, are candidates to employ models based on real-time traffic sensor data. Next-generation semi and fully autonomous vehicles may also rely heavily on sensor systems for efficient routing and collision avoidance.

Additionally, citizen crowdsourcing of information is now commonplace. Some examples that many may be familiar with include internet-based services, such as the *Great Internet Mersenne Prime Search* (GIMPS, http://www.mersenne.org/) or the internet-based protein folding activity *FoldIt* (http://fold.it/portal/). Similar examples are used widely by the general public, for instance *Yelp* (http://www.yelp.com), where users both provide and make use of reviews of restaurants and other establishments, or the online retailer *Amazon*, which provides users with peer-reviews of items for sale. A range of crowdsourcing efforts for environmental variables, including weather, traffic, noise, radiation, and air quality has recently emerged. The *Safecast* group (http://blog.safecast.org/) is an example of a citizen sensing project aimed at collecting environmental data that emerged in response to the concerns about radiation exposures after the earthquake of March 11, 2011, in Japan and catastrophic system failure at the *Fukushima Daichi* nuclear power plant. The UK *Biological Records Centre* (BRC) utilises citizen science based on mobile applications to conduct country-wide surveys on ladybird occurrence (http://www.ladybird-survey.org/recording.aspx). Finally, the *International Cooperative Programme on Effects of Air Pollution on Natural Vegetation and Crops* (ICP Vegetation) has recently launched a mobile app to record location and time of ozone related plant damage (http://icpvegetation.ceh.ac.uk/record/mobile-app-ozone-injury).

The large numbers of mobile device users provides additional opportunities to quantify the time-location patterns of many individuals, which in the past has been a challenge for exposure assessment. While there are ethical concerns related to privacy, increasingly, mobile users are allowing third parties access to their location data when there is perceived value in doing so. The increasing use of location-based services (e.g., map and navigation applications, such as Google Maps and Apple Maps), nearest friend (e.g., Foursquare) and service-finding applications (Uber, Next Bus, bank ATMs, etc.) are examples of such cases.

While the above systems illustrate the pervasiveness of modern sensors and computing that are producing huge amounts of data, many are proprietary, have not been designed with model integration in mind, and lack the standards and protocols (Laniak *et al.*, 2013) to enable them to be coupled with other sensor data or models. Enabling the successful use of such data in order to achieve scientific breakthroughs will depend on approaches taken towards the accessibility, integration and analysis of large datasets (Horsburgh *et al.*, 2009).

2.2 How can models benefit from ubiquitous sensing

Models are often viewed as 'black boxes' that obscure complexity and lack transparency, instead of 'tools to think with' (McIntosh *et al.*, 2007). However, models should be seen as technologies that can help in any step of the scientific process of understanding and solving complex systems problems. Considering the current advancements in theory and computational power, models should be seen as virtual reality technologies for a global system science in which sensors serve the fundamental sensing function for (i) understanding system dynamics, (ii) early detection and response to system malfunctions, and (iii) building resilient systems via enhanced adaptive management approaches that learn from past events and associated decision alternatives. We particularly emphasize the primary roles of physically-based systems models – for instance based on reaction-diffusion-dispersal processes and information theory (Shannon and Weaver, 1949; Vespignani, 2012; Quax *et al.*, 2013; Hill *et al.*, 2011) - versus statistical/data-driven models (Wu *et al.*, 2014b) for robust investigation of causal relationships between the environment and populations and system design to minimize systemic health burdens.

The additional data provided by ubiquitous sensing will enable both physically-based and data-driven models to be calibrated and validated over a wider range of inputs and outputs, thereby increasing model performance. In addition, these data will enable new information and relationships to be discovered using data-driven modelling and analysis approaches. In many instances, there will be more data than can be utilised by physically-based modelling approaches, and data-driven approaches can be used to extract the information that is locked up in these large datasets (Galelli *et al.*, 2014). This information can be used to enhance understanding, develop predictive and forecasting capability and improve physically-based models (Maier *et al.*, 2010; Wu *et al.*, 2014a).

2.3 The importance of data quality assurance and quality control with increasingly dense and complex sensor networks

QA/QC processes are critical to the dissemination and utilization of sensor data, but are also highly dependent on the application and the perceived or real-harm that might arise from public use of the data. First, recognizing that all measurement is prone to some degree of error, the quality of sensors used and the data produced should be documented. Second, data that fall outside reasonable ranges should be identified through the QA/QC process. However, this can be challenging with new emerging sensor deployments, for which it may be difficult to distinguish between data that are erroneous and data that are correct but have never observed before, because previous measurements were too sparse to identify extreme values and stochastic phenomena. Third, modellers will need to consider how measurement error may affect the error of their model predictions.

The public has become accustomed to forecasts of phenomena, such as skiing conditions, precipitation and typhoon storm paths and generally has an understanding, developed over time, of the reliability of this information without requiring a public disclaimer each time this information is disseminated. Hence, the public is disinclined to seek remedy when forecasts prove to be flawed and the public experiences inconvenience or loss because of actions undertaken in response to these forecasts. QA/QC are typically undertaken by those

providing the technical analysis and simulation modelling behind these forecasts in the form of a comparison with alternate models and analytical approaches. While expert analysts may have their own internal metrics of what constitutes acceptable deviation from parallel forecasts, ideally, more quality data will help modellers identify models that produce valid forecasts.

The dissemination of erroneous data can cause real harm or could be used against the agency reporting the data in future litigation. In places where litigation is a common means of resolving resource management or protection conflicts, this real or perceived concern may limit the willingness of potential data providers to share. One example of addressing this issue is with hydrologic time series management software, which is capable of performing continuous error processing of real-time sensor data and taking data censoring actions based on an established set of rules and procedures. Although this does not entirely eliminate the human factor in data processing, it does make this more efficient and significantly reduces the processing time leading to dissemination of real-time sensor data. There are a number of commercial software vendors such as Kisters Inc. (http://www.kisters.net/) based in Germany (WISKI and HYDSTRA) and Aquatic Informatics in Canada (AQUARIUS), as well as public agencies such as the US Army Corps of Engineers (DATAVUE), that offer real-time quality assurance processing capability.

3. Key challenges associated with issues of scale

As more data become available from sensors, the scales at which measurements are made potentially change. This will require rethinking about the scales, the objects and processes we model. It will also require creative thinking about fusing data available at different scales, which may potentially change the scale of inference we make on environmental processes.

3.1 Spatio-temporal resolution

Applied challenges, such as the prediction of the associations, causes or consequences of environmental pollution for global health, require interfacing of phenomena that occur on very different scales of space, time, and managerial organization (Levin, 1995). However, patterns that are unique to any range of scales will have potential unique causes and biological consequences. The key to prediction and understanding lies in the elucidation of mechanisms underlying observed patterns. Typically, these mechanisms operate at different scales than those on which patterns are observed; in many cases, patterns are better understood as emerging from the collective behaviours (interactions) of smaller scale units coupled with scale-dependent constraints. Examination of such phenomena requires the study of how pattern and variability change with the scale of description, and the development of laws for simplification, aggregation, universality and scaling. With such laws, it is easier to move from one scale to another and from one region to another without the necessity to run the model multiple times. In general, *a priori* there is no single spatio-temporal scale or resolution at which population health issues should be studied. Whereas disease usually affects the individual, public health challenges require different approaches. The scale and resolution of analysis should be related to the objective of the problem of interest, considering the scale of validity of control strategies. However, scale should always

consider the population as a whole in solving complex system issues and detecting 'true' causality and connectivity among system components (Rose, 1985; Helbing, 2013).

3.2 Coupling and integration

The integration of satellite and ground-based sensing provides an example of how data and models can be coupled at different spatio-temporal scales. Because of the different types of measurements obtained by different sensor networks implemented at distinctive spatial and temporal scales, multi-sensor integration is often desirable for environmental health applications, including air pollution monitoring. One of the main limitations of ground-based air pollution sensors is their sparse spatial distribution. In contrast, geospatial data products derived from space borne sensors provide spatially explicit 'Big Data' on a variety of pollutants, including NO_x, SO_2, CO, methane, ammonia, volatile organic compounds, and particulate matter (Duncan *et al*,. 2014). However, current air-pollution measurements from satellite-borne sensors also have significant limitations, including data gaps resulting from clouds, limited temporal resolution, and lack of information about the vertical distribution of many pollutants (Duncan *et al.* 2014). Models are frequently used to link these observations across spatial and temporal scales, with the aim of leveraging the strengths and minimizing the limitations of each type of sensor. For example, statistical modelling can be used to develop interpolated maps that blend spatially continuous *aerosol optical depth* (AOD) data with more precise local measurements of $PM_{2.5}$ from ground-based sensors (Puttaswamy *et al.*, 2014). Process-based chemical transport models can also be used to derive conversion factors that translate satellite-based AOD observations into ground-level $PM_{2.5}$ values, allowing estimates to be derived in areas where ground-based monitors are sparse or non-existent (van Donkelaar *et al.*, 2010). As novel ground-based and satellite sensors are developed, they will provide new opportunities for multi-scale sensor integration, but will also present conceptual and computational challenges requiring the extension of current models and the development of new modelling approaches.

3.3 Nesting and dynamically moving between systems at different scales

The use of hierarchical systems models and compartmental model nesting using a single model has become more ubiquitous with the advent of improved numerical techniques that allow efficient transfer of flux and pressure boundaries within regional models to form new boundary conditions for small scale, more highly disaggregated models. This is also true considering the advancement of a socially and computationally driven global system science (Convertino *et al.*, 2014, 2015; Helbing *et al.*, 2015). This has paved the way for some of the coupling and integration activities described in section 3.2 (for details on conceptual approaches for model integration see Argent, 2004a, b; Kelly *et al.*, 2013). The advantages of nesting models and different scales can be both computational and political. Stakeholder involvement in decision making and in the modelling process has long been sought not only as a means of improving models over time, but also to ensure that these models get used effectively for decision support (McIntosh *et al.*, 2011) at all levels of analysis. Changing the scale and improving the resolution of a model to the point that stakeholders start to recognize the characteristics of their own system in an airshed, catchment or landscape can yield significant long-term benefits to the decision-making process by increasing the potential for early stakeholder buy-in to the decision making process. Stakeholders rarely

have the technical skills to understand the nuances of the complex modelling tools being applied, but will often respond when presented with data or information that they are familiar with at a scale that allows them to evaluate its accuracy. Engagement even at this level can positively influence trust later in the resource management process. Such engagement can also be leveraged with the quantitative incorporation of stakeholder preferences via stakeholder belief assessment models in a direct and/or indirect way.

4. Addressing precision, accuracy, uncertainty and relevance in integrated model-sensor systems

As more data become available from sensor systems, questions arise as to variations in the precision and accuracy of different sensor instruments between different sensor platforms or networks, and whether data and the models that utilize these data are suitable for a specific intended purpose. Personal sensors are able to provide immediate and relevant data for instance related to physical activity and mobility, and enable the generation, as well as derivation of location-based information on environmental factors.

4.1 Precision vs. information content – rethinking instrument and data quality

Air quality measurement provides an example for considering sensor data quality and information content. A frequently stated view is that all air quality instruments, whether used as part of static networks or for personal exposure monitoring, should perform at a level equivalent to the reference instruments used for compliance, a principle enshrined in many legislative and regulatory contexts (e.g., in the EU Air Quality Directive 2008/50/EC). To establish and maintain a network of instruments based on this premise can be extremely expensive, a consequence of which is that fixed site networks are currently extremely sparse, with e.g. only 109 sites currently active in the UK (http://uk-air.defra.gov.uk). This raises several issues. The first, obvious one, is whether, given that major pollutant emission sources are related to traffic emissions, such fixed sites are (or can ever be) a true indicator of the spatial and temporal variability likely to be present in air quality. The answer, in most cases, is probably no. For example the UK, in its assessment of air quality compliance, places increasing reliance on physical and statistical models. The second, increasingly pressing and arguably more profound issue, given the advent of low cost air quality sensors and air quality sensor networks and the advent of highly sophisticated but efficient air quality models, is whether adherence to the ‘equivalence’ principle for air quality measurements, dominated by consideration of instrument accuracy and precision, is any longer the optimum approach.

Data assimilation, where various observations are combined, often with models, to produce optimal solutions, have been widely used (e.g. for determination of regional greenhouse gas (GHG) emissions or vertical profile retrieval of GHGs) and recently applied for human exposure assessment in urban regions (Schlink and Fischer, 2014). While instruments must have well quantified error characteristics, the discussion is now centred on *information content*, and how this could be optimised by suitable deployments of instruments and networks (Convertino *et al.*, 2014, 2015). The issue is no longer just *‘how good is an instrument’*, but *‘have we placed it where it provides maximum information’*? Translating

this to the context of low cost sensors, the question is whether a relatively poor measurement in the correct place can provide more useful information than a high precision measurement sited incorrectly. Recognising that sensors can be readily deployed as networks, the issue is whether information can be obtained by exploiting the higher density of measurements. There is an increasing body of literature suggesting that while this question is not fully resolved, low cost sensors and sensor networks have an increasing role to play (Schimak *et al.*, 2010; Ostfeld *et al.*, 2013; Diaz *et al.*, 2013; Austen, 2015). It has to be noted that information content is a function dependent on the relevance of data to one or multiple objectives of stakeholders involved in the decision making process. The trilemma of model relevance, accuracy and uncertainty can typically be solved numerically using global sensitivity and uncertainty analysis methods (Muller *et al.*, 2011), but stakeholder engagement should always be present.

There has also been discussion about whether low cost sensors and sensor networks should displace high precision instruments. There are strong arguments against this, both for historical reasons for ensuring continuity of data records, and because there are strong synergistic advantages in running low cost and high quality measurements (integrated using numerical models) side by side. However, there is a strong argument that the discussion should move away from purely considering the accuracy and precision of individual instruments towards assessing the information content of integrated measurement networks (including low cost as well as high precision instruments) and modelling systems.

4.2 Sensing individual activity and its relevance for health impact assessments

The increasing pervasiveness of personal computing devices has created new opportunities for sensing individual activity, which is relevant for estimating human exposures to environmental conditions (Schlink *et al.*, 2014) and characterizing health-related responses that may be associated with exposures. The mobile phone is the most common of these devices. Network service providers collect data on the time-location patterns of their mobile phone subscribers. In terms of precision, service providers typically know the location of their subscribers to the nearest cell tower. For Global Positioning System (GPS) and Wireless Network (WiFi)-enabled smartphones, handset manufacturers are collecting more detailed time-location data on the users of these devices. GPS systems are fairly accurate in outdoor environments, and are able to locate individuals to within city-block distances <30 m. Furthermore, assisted-GPS (aGPS) uses both cellular and WiFi-networks to improve location estimation in outdoor, as well as indoor, environments. In addition, as mentioned before, various smartphone apps that offer location-based services have the capability to record the time-location patterns of their users. Aside from the mobile phone, many wearable personal monitoring devices are now available that can measure a variety of physiological and health-related parameters, including those related to motion, muscle activity, cardiovascular health, respiration, perspiration, temperature, glucose, brain activity, emotion and affect, diet, sleep quality, and vision. Some of these parameters, e.g., motion, heart rate, diet, and emotion, can be determined via mobile phones. The quality and usability of these devices are improving rapidly, and efforts are already underway to promote better standards for data collection and the use of metadata within the academic community.

The analysis of human time-activity patterns and disease has benefitted from previous ecological modelling of animal populations, including classic predator-prey, biogeographic, and metapopulation models that are spatially and temporally explicit (Loos *et al.*, 2010). Already, mobile phone data have been used to parameterize population movement networks (Barabasi, 2005), relevant to the spread of malaria (Wesolowski *et al.*, 2012). The coupling of monitored personal time-activity patterns with modelled air pollution concentrations has improved the characterizations of air pollution exposures (Dons *et al.*, 2011; Engel-Cox *et al.*, 2013; Steinle *et al.*, 2013; 2014; Schlink and Ragas, 2011), and in some cases, has incorporated physiological sensing, such as energy expenditure, to improve exposure estimates (de Nazelle *et al.*, 2013). In addition, recently Citizen Scientists have begun leveraging population mobility and mobile phone data to conduct air pollution monitoring (http://aircasting.org/).

5. Integrated modelling of human and environmental health

In this section, we discuss how integrated model-sensor approaches can improve the understanding of associations and outcome-driven causality of environmental determinants for human health effects. Various models have been used for this purpose, in particular for associating toxicological and systems' information to epidemiological patterns. At one end of the spectrum statistical epidemiological models identify correlations (i.e., "associations") between environmental exposures and disease outcomes. These models are particularly important because the biological mechanisms for many diseases are not completely understood. Yet, strong statistical associations between environmental exposures and health outcomes can motivate health protective policies. The regulation of particulate matter air pollution is an example where our understanding of disease mechanisms is not yet clear, but the epidemiological evidence of the association between $PM_{2.5}$ and cardiovascular mortality has led to ambient air quality standards. A common pitfall of association studies is the occurrence of exposure misspecification, as most measurements are not individual-specific and this can make the results insignificant and/or biased (Begg and Lagakos, 1990).

At the other end of the spectrum are more mechanistic and dynamic multi-compartment models that discretize populations into susceptible, exposed, diseased (i.e., infected for infectious diseases) and recovered subpopulations (Rothman *et al.*, 2008). Such models rely on data for estimating non-physical transition rates among compartments, lack a spatial component, and very rarely are coupled to environmental and agent-causing-disease compartments whose information can be derived from integrated models (Convertino *et al.*, 2014, 2015; Rinaldo *et al.*, 2012; Helbing *et al.*, 2014).

Regardless of whether the goal is to utilize modelling for human health risk assessment or for health impact assessment, there are great opportunities to leverage the emerging sensor system data to improve estimates of micro-environmental levels of hazards and estimates of human mobility and time-location patterns – collectively these can improve exposure assessment science. Improved exposure estimates for individuals can be linked to 'Big Data' (e.g., electronic medical records), as well as emerging fields of biomedical science, such as exposomics, as discussed below. In addition, physiological sensing data from wearable sensors may improve our understanding of pre-clinical effects of exposure, enhanced by

model-data fusion, for instance taking into account high resolution meteorological information (Johansson *et al.*, 2015).

5.1 Missing link for association and causality

A criticism of statistical epidemiologic models is their focus on identifying association, while causality remains difficult to assess, despite the fact that many information theoretical and physical based models have been developed recently for dissecting spatio-temporal correlation time series more deeply than with traditional statistical models. Models, such as, for instance, transfer entropy models (Villaverde *et al.*, 2014), maximal information-based nonparametric exploration models, and global uncertainty and sensitivity analysis models (Saltelli *et al.*, 2010) are able to explore lags in space and time of data considering variable uncertainty and any combination of variable dependency, creating non-linear variations in the monitored output, that is, for instance disease incidence. These models embrace the idea that interactions of factors matter much more than single factor effects in shaping population health trajectories, thus traditional factor ranking based on one-time sensitivity analyses has limited validity and applicability. Note that such models can also provide predictions and can screen variable importance and interaction before any physical-based model is built. They can inform the design of modelling systems beyond the analysis of causality in data. Perhaps the most relevant aspect of sensors to these spatio-temporal predictive models is that there is the potential for sensor data to improve our understanding of the timing and context for exposures to a particular hazard or mixtures of hazards, confirming that exposures precede disease, and are not confounded by other competing risk factors to improve causal inference. Finally, it is important to recognise that the strength of causality is always a function of the stated objective, rather than a universal value valid across any domain and temporal scale of analysis if scaling analysis is not performed.

5.2 ‘Big data’ and exposomics

Future environmental health models may obtain relevant information for decision making through several linkages to ‘Big Data’, e.g. using web technologies (Vitolo *et al.*, 2015). One very likely linkage involves the increasing movement of clinical data to electronic medical records (EMR). EMRs have the potential to greatly improve our ability to access and query populations to compare the health outcomes of individuals living in different environments with different environmental exposures.

Another possible direction for health modelling involves ‘Big Data’, not at the population level, but rather at the individual-level. There is a small but emerging subculture, the *Quantified Self* movement (http://quantifiedself.com/), who are individuals interested in collecting large amounts of behavioural health data about themselves. Empowered by personal sensor devices, these Quantified Self persons may collect gigabytes of data over several years about their physical activity levels, time-location patterns, etc., for the purposes of understanding behavioural patterns and optimizing efficiency in their life. Individual-based dynamic models may be helpful in understanding these patterns.

Advances in the biomedical sciences have enabled the new field of exposomics (Wild, 2005; Rappaport, 2011), which aims to understand, through biology, the mixture of exposures to

different environmental hazards throughout an individual's stages of life. As the methods within exposomics are refined to the point where an individual's environmental exposures can be characterized (from analysis of biological samples), there will likely be an increasing need to model the relationship between these biological exposure factors to exposure factors outside of the body, such as behaviours and environmental processes that can be more appropriately dealt with through environmental policies and planning.

In each of these 'Big Data' examples, and with sensing in general, the prospects for exciting new data-integrative modelling must still be balanced with the practicalities and need for ethical use of sensitive data. For instance, in the U.S., there are federal laws that govern the disclosure of protected health information in electronic medical records. In addition, for data from individual-level sensing and exposure biology, the ethical concerns regarding what can and cannot be inferred from disclosure remains largely unexplored.

6. Models and data at the science-policy interface

Key for successful science-policy interaction is to establish the science-policy interface to include all aspects of the policy decision development cycle: starting from issue framing, in the adequate institutional setting, building of trust is an essential step. *Salience* and *timing* of the scientific evidence agreed upon within the science community and presented to policy stakeholders are equally important to ensure uptake.

In the case of transboundary air pollution, long-term monitoring activities and developing modelling capability have supported the framing of the issue and delivered robust data to derive salient policy information. A crucial role in communicating risk quantification concepts and providing input data for integrated assessment modeling (IAM) has been fulfilled by the application of both sensor networks and models documenting the environmental fate and effects of air pollution. IAMs integrated this information and – by providing high-level summary evaluations of different policy options - highlight the cause for action and the costs of inaction. The institutional setting provided by, for instance, the *United Nations Economic Commission for Euro*pe's (UNECE) *Convention on Long-range Transboundary Air Pollution* (CLRTAP) has been essential both in building trust between different scientific fields, and between science and policy stakeholders. The flow of information is not unidirectional from science to policy: the explicit and implicit values expressed by national and international political processes find their way into the priority setting process for modelling and research, and the valuation of different, at times conflicting, policy targets (Voinov *et al.*, 2014).

Much of the success of CLRTAP in integrating science and policy can be attributed to scientific results, assessments, and technological solutions, forming an integral part of the agendas of negotiating meetings. Scientists are present in negotiation meetings, and policy-makers participate in scientific meetings and thus can make sure that the science remains focused on the needs of the policy process. Such meetings typically start with an update of the available science and end with further requests to scientists (Reis *et al.*, 2012).

7. Conclusions and outlook

In this paper, we have outlined a number of exciting developments within environmental sensing that offer new opportunities for data-intensive modelling, particularly involving the incorporation of 'Big Data' from sensors and health-related datasets. The variety of both sensor and model systems is too large to provide a comprehensive review, however, we have attempted to provide useful examples in which sensor systems have been integrated with models, as well as sensor systems producing data that have not been modelled, and models that may benefit from sensor data. There are real challenges related to data QA/QC, metadata, standards, and spatio-temporal scaling (Schimak *et al.*, 2010) that will require continual development in the upcoming years as models and sensor systems are increasingly integrated. With such integration across different dimensions (Hamilton *et al.*, 2015), there is the possibility to better understand uncertainty, and to improve model predictions, particularly in the estimate of human exposures to environmental hazards, which is a fundamental step in human health risk assessment and health impact assessment. Particularly exciting will be the development of systems that so tightly couple real-time sensor data with models, that they produce information that actively engages with the public and informs stakeholders (Voinov *et al*., 2010; Boschetti, 2015) towards improving public health in a seamless and transparent manner – true ubiquitous sensing and computing. Here, recent developments for instance in the development of Geospatial Information Infrastructures (Diaz *et al.*, 2013) provide useful examples and can inform progress towards integrated environmental modelling (Laniak *et al.*, 2013). At the same time, the motivation for integration needs to be clear and demand driven, to avoid the emergence of 'integronsters' (Voinov *et al.*, 2013), i.e. integrated models which have become too complex and convoluted to be transparent or useful.

'Big Data' and sensors are without doubt hot topics in the scientific community, as recently illustrated by the discussion of 'Big Data' in relation to public health (Khoury and Ioannidis, 2014; Fung *et al.*, 2014) and the spotlight on the use of low-cost sensors for crowdsourcing air pollution data in developing countries (Austen, 2015). In order to move forward and realise the substantial (potential) benefits offered by embracing these concepts, we identify these key research areas:

1. Developing metadata and access standards.
2. Understanding and developing QA/QC frameworks for sensor data that are adaptable to different purposes, and informative to modelling applications.
3. Continuing to develop improvements in modelling architectures for working with data of different spatio-temporal resolutions.
4. Continuing to develop improvements in the coupling of model systems with sensor systems for real-time control.
5. Improving 'Big Data' science, including data management, access, fusion, and analytics.
6. Addressing the ethical challenges of balancing privacy with data accessibility to improve public health.

7. Improving partnerships between Citizen Science, Community Crowdsourcing, and other public data collection campaigns to improve the quality of sensing data and their usability for open source modelling.
8. Examining the differences and potential disparities between developed versus developing country adoption of sensor technologies, and their impacts on modelling of environmental health processes.
9. Evaluating the performance of integrated dynamical model-sensor systems and their use in policy making via the coupling of decision-analytical with biophysical models.
10. Demonstrating robustness and establishing trust in information generated from integrated modelling and ubiquitous sensing data.

This list is not exhaustive, but highlights the key areas we identify as critical for a better integration of models, sensors and stakeholders with the ultimate objective to provide better information for evidence based decision making. In closing, it is important to highlight the potential ethical challenges for integrated sensor-model systems, for instance in relation to personal data, privacy and individual autonomy (Vayena *et al.*, 2015). While some of these challenges are not novel and well known in the context of public health and data use, others are new and emerging due to the recent advances in the capabilities of sensors and models

Acknowledgements

E.S. is funded by NIH R21ES024715. M.C. gratefully acknowledges the Minnesota Discovery, Research and InnoVation Economy (MnDRIVE) ''Global Food Venture'' funding and the Institute on the Environment ''Discovery Grant'' funding at the University of Minnesota Twin-Cities. S.R. and S.S. acknowledge the support for the conceptual development and testing of personal exposure monitoring methods by the UK Natural Environment Research Council through *National Capability* funding.

References

Argent RM. Concepts, methods and applications in environmental model integration. Environ Model Softw. 2004a; 19(3):217.

Argent RM. An overview of model integration for environmental applications—components, frameworks and semantics. Environ Model Softw. 2004b; 19(3):219–234.

Austen K. Pollution Patrol. Nature. 2015; 136(517) http://www.nature.com/news/environmental-science-pollution-patrol-1.16654.

Banzhaf E, de la Barrera FJ, Kindler A, Reyes-Paecke S, Schlink U, Welz J, Kabisch S. A conceptual framework for integrated analysis of environmental quality and quality of life. Ecol Indic. 2014; 45:664–668.

Barabási AL. The origin of bursts and heavy tails in human dynamics. Nature. 2005; 435:207–211. [PubMed: 15889093]

Begg MD, Lagakos S. On the consequences of model misspecification in logistic regression. Environmental Health Perspectives. 1990; 87:69–75. [PubMed: 2269243]

Boschetti F. Models and people: An alternative view of the emergent properties of computational models. Complexity. 2015

Convertino M, Liu Y, Hwang H. Optimal Surveillance System Design for Outbreak Source Detection Maximization: a Value of Information Model. Complex Adaptive Systems Modeling 2014. 2014; 2:6.

Convertino M, Muñoz-Carpena R, Kiker GA, Perz SG. Design of optimal ecosystem monitoring networks: hotspot detection and biodiversity patterns. Stochastic Environmental Research and Risk Assessment. 2015; 29(4):1085–1101.

de Nazelle A, Seto E, Donaire-Gonzalez D, Mendez M, Matamala J, Nieuwenhuijsen MJ, Jerrett M. Improving estimates of air pollution exposure through ubiquitous sensing technologies. Environ Pollut. 2013; 176:92–99. [PubMed: 23416743]

Díaz L, Bröring A, McInerney D, Libertá G, Foerster T. Publishing sensor observations into Geospatial Information Infrastructures: A use case in fire danger assessment. Environ Model Softw. 2013; 48(0):65–80.

Dons E, Int Panis L, Van Poppel M, Theunis J, Willems H, Torfs R, Wets G. Impact of time–activity patterns on personal exposure to black carbon. Atmos Environ. 2011; 45(21):3594–36021.

Duncan BH, Prados AI, Lamsal LN, Liu Y, Streets DG, Gupta P, Hilsenrath E, Kahn RA, Nielsen JE, Beyersdorf AJ, Burton SP, Fiore AM, Fishman J, Henze DK, Hostetler CA, Krotkov NA, Lee P, Lin M, Pawson S, Pfister G, Pickering KE, Pierce RB, Yoshida Y, Ziemba LD. Satellite data of atmospheric pollution for U.S. air quality applications: Examples of applications, summary of data end-user resources, answer to FAQs, and common mistakes to avoid. Atmos Environ. 2014; 94:647–662.

Engel-Cox J, Nguyen TKO, vanDonkelaar A, Martin RV, Zell E. Toward the next generation of air quality monitoring: Particulate matter. Atmos Environ. 2013; 80:584–590.

EOSDIS. Earth Observation System Data and Information System. 2014 https://earthdata.nasa.gov/about-eosdis/performance.

Fung IC-H, Zion Tsz Ho Tse ZTH, Fu K-W. Converting Big Data into public health. Science. 2015; 347(6222):620. [PubMed: 25657237]

Galelli S, Humphrey GB, Maier HR, Castelletti A, Dandy GC, Gibbs MS. An evaluation framework for input variable selection algorithms for environmental data-driven models. Environ Model Softw. 2014; 62:33–51.

Hamilton SH, ElSawah S, Guillaume JHA, Jakeman AJ, Pierce SA. Integrated assessment and modelling: Overview and synthesis of salient dimensions. Environ Model Softw. 2015; 64(0):215–229.

Helbing D. Globally networked risks and how to respond. Nature. 2013; 497:51–59. [PubMed: 23636396]

Helbing D, Brockmann D, Chadefaux T, Donnay K, Blanke U, Woolley-Meza O, Moussaid M, Johansson A, Krause J, Schutte S, Perc M. Saving Human Lives: What Complexity Science and Information Systems can Contribute. Journal of Statistical Physics. 2015; 158:735–781. [PubMed: 26074625]

Hill DJ, Liu Y, Marini L, Kooper R, Rodriguez A, Futrelle J, Minsker BS, Myers J, McLaren T. A virtual sensor system for user-generated, real-time environmental data products. Environ Model Softw. 2011; 26(12):1710–1724.

Hilty LM, Aebischer B, Rizzoli AE. Modeling and evaluating the sustainability of smart solutions. Environ Model Softw. 2014; 56(0):1–5.

Horsburgh JS, Tarboton DG, Piasecki M, Maidment DR, Zaslavsky I, Valentine D, Whitenack T. An integrated system for publishing environmental observations data. Environ Model Softw. 2009; 24(8):879–888.

Johansson L, Epitropou V, Karatzas K, Karppinen A, Wanner L, Vrochidis S, Bassoukos A, Kukkonen J, Kompatsiaris I. Fusion of meteorological and air quality data extracted from the web for personalized environmental information services. Environ Model Softw. 2015; 64(0):143–155.

Kelly (Letcher) RA, Jakeman AJ, Barreteau O, Borsuk ME, ElSawah S, Hamilton SH, Henriksen HJ, Kuikka S, Maier HR, Rizzoli AE, vanDelden H, Voinov AA. Selecting among five common modelling approaches for integrated environmental assessment and management. Environ Model Softw. 2013; 47:159–181.

Khoury MJ, Ioannidis JPA. Big data meets public health. Science. 2014; 346(6213):1054–1055. [PubMed: 25430753]

Laniak GF, Olchin G, Goodall J, Voinov A, Hill M, Glynn P, Whelan G, Geller G, Quinn N, Blind M, Peckham S, Reaney S, Gaber N, Kennedy R, Hughes A. Integrated environmental modeling: A vision and roadmap for the future. Environ Model Softw. 2013; 39(0):3–23.

Lazer D, Kennedy R, King G, Vespignani A. The Parable of Google Flu: Traps in Big Data Analysis. Science. 2014 Mar; 343(14):1203–1205. Copy at http://j.mp/1ii4ETo. [PubMed: 24626916]

Levin S. The problem of pattern and scale in ecology. Ecological Time Series. 1995:277–326.

Lim SS, Vos T, Flaxman AD, Danaei G, Shibuya K, et al. A comparative risk assessment of burden of disease and injury attributable to 67 risk factors and risk factor clusters in 21 regions, 1990–2010: a systematic analysis for the Global Burden of Disease Study 2010. The Lancet. 2012; 380(9859): 2224–2260.

Loos M, Schipper AM, Schlink U, Strebel K, Ragas AMJ. Receptor-oriented approaches in wildlife and human exposure modelling: A comparative study. Environ Model Softw. 2010; 25:369–382.

MacIntosh DL, Xue J, Ozkaynak H, Spengler JD, Ryan PB. A population-based exposure model for benzene. J Expo Anal Environ Epidemiol. 1995; 5(3):375–403. [PubMed: 8814777]

Maier HR, Jain A, Dandy GC, Sudheer KP. Methods used for the development of neural networks for the prediction of water resource variables in river systems: Current status and future directions. Environ Modell Softw. 2010; 25(8):891–909.

May RJ, Dandy GC, Maier HR, Nixon JB. Application of partial mutual information variable selection to ANN forecasting of water quality in water distribution systems. Environ Modell Softw. 2008; 23(10–11):1289–1299.

McIntosh BS, Seaton RAF, Jeffrey P. Tools to think with? Towards understanding the use of computer-based support tools in policy relevant research. Environ Modell Softw. 2007; 22(5):640–648.

McIntosh BS, Ascough JC II, Twery M, Chew J, Elmahdi A, Haase D, Harou JJ, Hepting D, Cuddy S, Jakeman AJ, Chen S, Kassahun A, Lautenbach S, Matthews K, Merritt W, Quinn NWT, Rodriguez-Roda I, Sieber S, Stavenga M, Sulis A, Ticehurst J, Volk M, Wrobel M, van Delden H, El-Sawah S, Rizzoli A, Voinov A. Environmental decision support systems (EDSS) development – Challenges and best practices. Environ Model Softw. 2011; 26(12):1389–1402.

McKone, TE. Livermore, CA: Lawrence Livermore National Laboratory; 1993. CalTOX, A Multimedia Total-Exposure Model for Hazardous-Wastes Sites Part III: The Multiple-Pathway Exposure Model. UCRL-CR-111456Pt III. https://www.dtsc.ca.gov/AssessingRisk/upload/techman3.pdf. [Accessed 7 June 2015]

Muller, SJ.; Muñoz-Carpena, R.; Kiker, GA. Model relevance: Frameworks for exploring the complexity-sensitivity-uncertainty trilemma. In: Linkov, I., editor. Climate Change: Global Change and Local Adaptation. Netherlands: Springer; 2011. Adaptive Management for Climate Change (NATO series).

Perelman L, Ostfeld A. Operation of remote mobile sensors for security of drinking water distribution systems. Water Research. 2013; 47(13):4217–4226. [PubMed: 23764572]

Puttaswamy SJ, Nguyen HM, Braverman A, Hu X, Liu Y. Statistical data fusion of multi-sensor AOD over the continental United States. Geocarto International. 2014; 29:48–64.

Quax R, Apolloni A, Sloot PMA. Towards understanding the behavior of physical systems using information theory. The European physical journal. Special topics. 2013; 222(6):1389–1401.

Rappaport SM. Implications of the exposome for exposure science. J Expo Sci Environ Epidemiol. 2011; 21(1):5–9. [PubMed: 21081972]

Reis S, Grennfelt P, Klimont Z, Amann M, ApSimon H, Hettelingh J-P, Holland M, LeGall A-C, Maas R, Posch M, Spranger T, Sutton MA, Williams M. From Acid Rain to Climate Change. Science. 2012; 338(6111):1153–1154. [PubMed: 23197517]

Rinaldo A, Bertuzzo E, Mari L, Righetto L, Blokesch M, Gatto M, Casagrandi R, Murray M, Vesenbeckh SM, Rodriguez-Iturbe I. Reassessment of the 2010–2011 Haiti cholera outbreak and multi-season projections via inclusion of rainfall and waning immunity. PNAS. 2012; 109(17): 6602–6607. [PubMed: 22505737]

Rose G. Sick individuals and sick populations. International Journal of Epidemiology. 1985; 14:32–38. [PubMed: 3872850]

Rothman, K.; Greenland, S.; Lash, T. Modern Epidemiology. Lippincott W and Wilkins; 2008.

Salathé M, Bengtsson L, Bodnar TJ, Brewer DD, Brownstein JS, Buckee C, Campbell EM, Cattuto C, Khandelwal S, Mabry PL, Vespignani A. Digital epidemiology. PLoS Comput Biol. 2012; 8:e1002616. pmid:22844241. [PubMed: 22844241]

Saltelli A, Paola Annoni, Ivano Azzini, Francesca Campolongo, Marco Ratto, Stefano Tarantola. Variance based sensitivity analysis of model output. Design and estimator for the total sensitivity index. Computer Physics Communications. 2010; 181(2):259–270.

SAS. Big Data - What it is & why it matters. http://www.sas.com/en_us/insights/big-data/what-is-bigdata.html.

Schimak G, Rizzoli AE, Watson K. Sensors and the environment – Modelling & ICT challenges. Environ Model Softw. 2010; 25(9):975–976.

Stocker M, Baranizadeh E, Portin H, Komppula M, Rönkkö M, Hamed A, Virtanen A, Lehtinen K, Laaksonen A, Kolehmainen M. Representing situational knowledge acquired from sensor data for atmospheric phenomena. Environ Model Softw. 2014; 58(0):27–47.

Schlink, U.; Fischer, G. A Bayesian Maximum Entropy scheme for the assimilation of mobile recordings with simulations of urban micrometeorological data. In: Ames, DP.; Quinn, NWT.; Rizzoli, AE., editors. Proceedings of the 7th International Congress on Environ Modell Softw (iEMSs). San Diego, California, USA: International Environ Modell Softw Society (iEMSs), Manno; 2014 Jun 15–19. p. 1-6.

Schlink U, Kindler A, Großmann K, Schwarz N, Franck U. The temperature recorded by simulated mobile receptors is an indicator for the thermal exposure of the urban inhabitants. Ecol Indic. 2014; 36:607–616.

Schlink U, Ragas AMJ. Truncated Levy flights and agenda-based mobility are useful for the assessment of personal human exposure. Environ Pollut. 2011; 159(8–9):2061–2070. [PubMed: 21429644]

Schlink U, Strebel K, Loos M, Tuchscherer R, Richter M, Lange T, Wernicke J, Ragas AMJ. Evaluation of human mobility models, for exposure to air pollutants. Sci. Total Environ. 2010; 408:3918–3930. [PubMed: 20417545]

Shannon, CE.; Weaver, W. The Mathematical Theory of Communication. University of Ilinois Press; 1949.

Steinle S, Reis S, Sabel C. Quantifying human exposure to air pollution - moving from static monitoring to spatio-temporally resolved personal exposure assessment. Sci Total Environ. 2013; 443:184–193. [PubMed: 23183229]

Steinle S, Reis S, Sabel C, Semple S, Twigg MM, Braban CF, Leeson AE, Heal MR, Harrison D, Lin C, Wu H. Application of a low-cost method to quantify human exposure to ambient particulate matter concentrations across a wide range of microenvironments. Sci Total Environ. 2015; 508:383–394. [PubMed: 25497678]

Van Donkelaar A, Martin RV, Brauer M, Kahn R, Levy R, Verduzco C, Villenevue PJ. Global estimates of ambient fine particulate matter concentrations from satellite-based aerosol optical depth: Development and application. Environ Health Persp. 2010; 118:847–855.

Vayena E, Salathé M, Madoff LC, Brownstein JS. Ethical Challenges of Big Data in Public Health. PLoS Comput Biol. 2015; 11(2):e1003904. [PubMed: 25664461]

Vespignani A. Modeling dynamical processes in complex socio-technical systems. Nature Physics. 2012; 8:32–39.

Vieno M, Heal MR, Hallsworth S, Famulari D, Doherty RM, Dore AJ, Tang YS, Braban CF, Leaver D, Sutton MA, Reis S. The role of long-range transport and domestic emissions in determining atmospheric secondary inorganic particle concentrations across the UK. Atmos Chem Phys. 2014; 14:8435–8447.

Vieno M, Dore AJ, Stevenson DS, Doherty R, Heal MR, Reis S, Hallsworth S, Tarrason L, Wind P, Fowler D, Simpson D, Sutton MA. Modelling surface ozone during the 2003 heat-wave in the UK. Atmos Chem Phys. 2010; 10:7963–7978.

Villaverde AF, Ross J, Morán F, Banga JR. MIDER: Network Inference with Mutual Information Distance and Entropy Reduction. PLoS ONE. 2014; 9(5):e96732. [PubMed: 24806471]

Voinov A, Bousquet F. Modelling with stakeholders. Environ Model Softw. 2010; 25(11):1268–1281.

Voinov A, Shugart H. 'Integronsters', integral and integrated modeling. Environ Model Softw. 2013; 39:149–158.

Voinov A, Seppelt R, Reis S, Nabel JEMS, Shokravi S. Values in socio-environmental modelling: Persuasion for action or excuse for inaction. Environ Model Softw. 2014; 53:207–212.

Vitolo C, Elkhatib Y, Reusser D, Macleod CJA, Buytaert W. Web technologies for environmental Big Data. Environ Model Softw. 2015; 63(0):185–198.

Wesolowski A, Eagle N, Tatem AJ, Smith DL, Noor AM, Snow RW, Buckee CO. Quantifying the impact of human mobility on malaria. Science. 2012; 338(6104):267–270. [PubMed: 23066082]

Wild CP. Complementing the genome with an "exposome": the outstanding challenge of environmental exposure measurement in molecular epidemiology. Cancer epidemiology, biomarkers & prevention: a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology. 2005; 14(8):1847–1850.

Wood M, Kovacs D, Bostrom A, Convertino M, Linkov I. A Moment of Mental Model Clarity, Response to Jones *et al*. 2011 “Mental Models: An Interdisciplinary Synthesis of Theory and Methods, 2011, 16-1, Ecology and Society) in the special issue “Mental models in human - environment interactions: Theory, policy implications, and methodological explorations”; Editor: L. Gunderson), http://dx.doi.org/10.5751/ES-05122-170407. Ecology and Society. 2012

Wu W, Dandy GC, Maier HR. Protocol for developing ANN models and its application to the assessment of the quality of the ANN model development process in drinking water quality modeling. Environ Model Softw. 2014a; 54:108–127.

Wu W, Dandy GC, Maier HR. Optimal control of total chlorine and free ammonia levels in a water transmission pipeline using artificial neural networks and genetic algorithms. J. Water Resour. Plann. Manage. 2014b 10.1061/(ASCE)WR.1943-5452.0000486, 04014085.

Highlights

1. Sensors and models play vital roles in harnessing ‘Big Data’ to extract information
2. Data analytics can help to diminish monitoring burden and support locating sensors
3. Exploring ‘Big Data’ is essential to detect universal associations across space and time
4. Ethical challenges and issues of standards and harmonisation need to be addressed
5. Citizen science needs robust sensors and models to crowd-source and interpret data

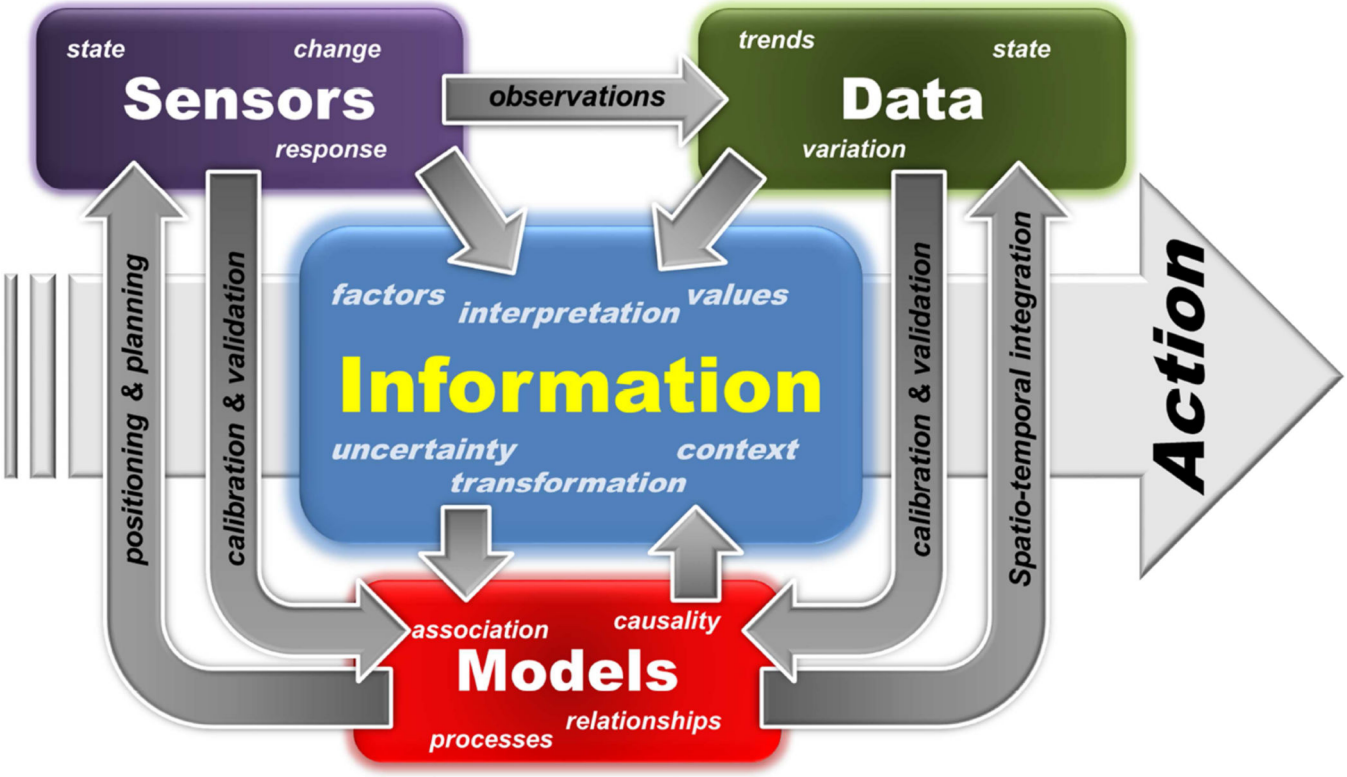

Fig. 1.

A conceptual model for sensor-model integration illustrating the complex system required for the development of evidence and data based action (e.g. policy development and implementation). The central role of information (factors, interpretation, values, uncertainty, transformation and context) is highlighted. Here, *information* is also depicted as input to the modelling stage, e.g., to reduce the size of 'Big Data' by extracting only data with high information value for the question being asked (Shannon and Weaver 1949, Lazer *et al*., 2014, Galelli *et al.*, 2014; Convertino *et al.*, 2014, 2015). Information in general and the policy questions to be assessed in particular include value judgements (Voinov *et al*., 2014). This can affect the interpretation of data, for instance by identifying priorities and setting the context for analyses. A robust science-policy interface (Reis *et al.*, 2012) can establish trust in data and information generated by sensors and models. This is essential, as transparency and traceability of data flows and processing methods are key requirements to assess the quality of data. Such science-policy interfaces need to reflect stakeholders' conceptual and mental models (alternatives, preferences, utility, and drivers) embedded in decision science frameworks, integrating those (mainly) qualitative models with (quantitative) biophysical models and decisions (see Wood *et al.*, 2012; Boschetti, 2015 and section 7).