Applying big data beyond small problems in climate research

Article *in* Nature Climate Change · February 2019

DOI: 10.1038/s41558-019-0404-1

CITATIONS
36

READS
500

7 authors, including:

Benedikt Knüsel
ETH Zurich
7 PUBLICATIONS **54** CITATIONS
SEE PROFILE

Christoph Baumberger
ETH Zurich
50 PUBLICATIONS **305** CITATIONS
SEE PROFILE

Gertrude Hirsch Hadorn
ETH Zurich
73 PUBLICATIONS **3,585** CITATIONS
SEE PROFILE

Erich M Fischer
ETH Zurich
135 PUBLICATIONS **12,884** CITATIONS
SEE PROFILE

Applying Big Data Beyond Small Problems in Climate Research

Benedikt Knüsel[1,2], Marius Zumwald[1,2], Christoph Baumberger[1], Gertrude Hirsch Hadorn[1], Erich M. Fischer[2], David N. Bresch[1,3], Reto Knutti[2]

[1] Institute for Environmental Decisions, ETH Zurich, Switzerland

[2] Institute for Atmospheric and Climate Science, ETH Zurich, Switzerland

[3] Federal Office of Meteorology and Climatology MeteoSwiss, Zurich, Switzerland

Accepted for publication in Nature Climate Change.

Author's accepted manuscript.

Preface

Commercial success of big data has led to speculation that big-data-like reasoning could partly replace theory-based approaches in science. Big data typically has been applied to "small problems", well-structured cases characterized by repeated evaluation of predictions. Here, we show that in climate research, intermediate categories exist between classical domain science and big data, and that big-data elements have also been applied without the possibility of repeated evaluation. Big-data elements can be useful for climate research beyond small problems if combined with more traditional approaches based on domain-specific knowledge. The biggest potential for big-data elements, we argue, lies in socioeconomic climate research.

Big data affects increasingly many aspects of our lives. The large volumes of data gathered and stored are the basis of the recommendations we receive when shopping online and the way in which we connect to people all over the world via social media [1]. Naturally, this has led to debates about how increasing volumes of data and new analytic tools might impact scientific research. An emerging view is that largely theory-free data-driven models will supplant models

that explicitly start from theory [2,3]. Big data could have a big potential in various scientific disciplines [4] including climate research [5,6], but it remains unclear what questions big data can potentially help to answer. The usefulness of big data and the associated epistemological shifts are of particular importance for climate research for three reasons. First, the already large volumes of current climate data are expected to increase further in both volume and complexity over the coming years and decades [7]. Second, approaches typically associated with big data have already entered climate research[5] (for examples see refs. 8–10). And third, climate models are rooted in scientific theory, which is one of the key reasons for confidence in their projections [11]. This makes climate research an interesting test case for the suggested shift from process-based to largely theory-free modeling.

A prevailing problem concerning big data is the fuzziness around the terminology. To date, there is no consensus definition of big data or related concepts such as data-intensive science, data-driven science, and big-data science. Based on suggested definitions of these terms [12,13], we adopt a conception of big data that focuses on the characteristics of data and the tools used to analyze them. The data are often voluminous streams of partly unstructured and heterogeneous data (characterized by the so-called three Vs, volume, velocity, and variety) and can be noisy and uncertain compared to more standardized datasets (indicated by a fourth V, veracity) [14,15]. The tools used to analyze them are machine learning and data mining, ranging from simple linear regression tools to complex non-linear models in deep learning [16,17].

Small problems

Many commercial problems are solved using pure big data approaches**;** a typical example is the problem of predicting online consumer preferences such as online book recommendations with pure big data, which use data on how customers react to different books. An algorithm analyzes these data and automatically identifies similar books. Both successful and unsuccessful recommendations inform future recommendations [1]. The problem of recommending the right

books to the right customers constitutes a well-posed problem with a clear measure of success and fast evaluation of the predictions: the customer looks at the book or buys it. As wrong predictions are hard to avoid and contribute to improving the predictions, pure big data is usually applied when the impact or the probability of wrong predictions is small. Due to their narrow scope, their clear measure of success, the small impact of wrong predictions, and the repeated evaluation of the predictions, we refer to such problems as “small problems”, even if the statistical techniques may be complex and the computational and storage cost may be very large. The following set of conditions is necessary for reliably solving small problems with big data:

1. The system is predictable for the questions of interest.
2. Sufficient data is available for the initial training of the model.
3. Sufficient new data is available to periodically evaluate the predictions against observations and make adjustments to the relationships if necessary.

Condition 1 is necessary for any kind of reliable prediction. If book choices were fundamentally unpredictable, an algorithmic prediction could not outperform a random book recommendation. Condition 2 is necessary to identify and train an initial model for predicting a given variable of interest. In the case of online book recommendations, data engineers can employ a so-called “item-to-item” approach which uses individual books as the unit of comparison rather than other traditional recommender systems [18].

While conditions 1 and 2 are not unique to our notion of “small problems”, condition 3 is. In small problems, the repeated evaluation of the predictions and the consequent adaptation has an important epistemic function as it allows to detect and correct relationships between variables that are not represented adequately. In the example of book recommendations, two books with a large shared readership today will not necessarily also be read by the same people in the future. Furthermore, new books are released for which no data is available. Thus, the predictions need continuous evaluation and adaptation.

The notion of small problems introduced here is closely related to the kind of problems solved by narrow (or weak) artificial intelligence [19]. Note that characterizing a problem as a "small problem" does neither imply that it is unimportant, nor that it is easy to solve. In fact, building a well-functioning machine learning model, the so-called training step, can be technically challenging in terms of data collection, preparation and storage, modeling, and computation. While we acknowledge the challenges associated with these issues [20,21], we do not elaborate on them here because our focus lies on making predictions for new cases using an already developed model, the so-called inference step.

Also, some problems falling under our category of "small problems" can be very complex (such as speech recognition). Other big-data applications, such as personalized medicine, are not small problems because the impact of wrong predictions can be large. We will return to these cases in a later section. What is common to "small problems" is that their solutions have a clearly defined purpose, and that success can periodically be measured against new observations in order to evaluate and improve predictions. In many scientific applications, this is not possible because it is not clear what constitutes a successful prediction and because the time horizon is too long to wait for observational data to test the prediction.

Contrasting domain science and big data

In this section, we introduce a conceptual framework to better understand to what extent big-data elements have already been applied in climate research and to classify case studies. The framework components are introduced by contrasting, on the one hand, how scientists construct and use general circulation models (GCMs) to project future states of the climate system as an example of classical domain science, and, on the other hand, the case of online book recommendations, introduced in the previous section, as an example of pure big data. This comparison is also intended to resolve some confusion about the difference between big data and "lots of data" common among domain scientists who are experienced in handling large

volumes of data.

Measurements: In classical domain science, the measurements assign numerical values to phenomena described by theory-based concepts. For example, cloud albedo values indicate the fraction of reflected radiation by clouds based on calibrated satellite readings. In most climate datasets, this operationalization is complex and involves modeling, hence domain-specific knowledge is required for domain-science measurements. This differs from online book recommendations. In this case, internet traces are analyzed that assess whether a customer has clicked on a given book recommendation and whether she has proceeded to actually buying the book. These features are engineered based on everyday reasoning, which is the foundation of measurements in pure big data.

Datasets: Climate scientists use datasets to determine the initial conditions of variables of interest [22] and to determine the values of certain parameters whose values are insufficiently constrained by theoretical considerations [23], a process usually referred to as tuning or calibration. These datasets can be quite large in volume but they are fixed sets of data fitting into a pre-defined structure, for example a relational table. In the case of online book recommendations, the datasets are used for identifying a suitable model structure as well as for training the model. Furthermore, since periodic evaluation of the predictions is needed to correct the relationships between variables if necessary, a flow of new data is required. Hence, in this case, a data stream is analyzed rather than a fixed set. The constant inflow of new data and its ongoing analysis is often referred to as its velocity, a typical characteristic associated with big data. Furthermore, in pure big data applications, the data are often partly unstructured.

Models: In the case of climate model construction, the phenomena are described in terms of theory-based concepts, such as temperature, air pressure, and condensation. The relationships between these variables are whenever possible established from theory, for example from physical equations [24], although empirical parameterizations are necessary for certain processes.

For online book recommendations, the phenomena that are put into relation to each other are based on everyday language concepts, such as which of the recommended books a customer clicks on. The relationships between different books are automatically detected, typically by a machine learning algorithm, rather than imposed from theory.

The three components of this framework also highlight differences between classical statistical approaches and pure big data. Classical statistical approaches usually handle fixed sets of theory-based measurements. Also, classical statistics makes strict assumptions regarding the distribution of the data or the residuals and hence the model. This is not the case in pure big data, where the data are more important. We do note, however, that there is some overlap between regression analysis and machine learning tools, and even more so when considering non-parametric statistical modeling [3].

Big-data elements in climate research

We applied our conceptual framework to categorize scientific studies from atmospheric science, climate science, and climate impact research. A total of 45 studies were reviewed that we obtained through the search terms “big data weather”, “big data climate”, “data mining weather”, “data mining climate”, “machine learning weather”, and “machine learning climate” in ISI web of science and Google Scholar, published between January 2006 and April 2017. However, the goal was to provide an overview of big-data elements in the climate science literature rather than a full review. Hence, we excluded weather-related technical applications such as data-driven forecasting of renewable energy production from wind or solar power, and weather and climate impacts on biodiversity and agriculture to contain the set of studies to a manageable size.

Table 1 provides an overview of the categories and indicates which studies fall into the respective categories. In between the two extreme cases of classical domain science and pure big data, we identify four intermediate categories, each of which we present below using an illustrative case

study.

Table 1: Categories of climate-related research employing big-data elements. Notable examples are listed in the last column. The top row corresponds to classical domain science, the bottom row to pure big data.

	models	datasets	measurements	examples
constructing and using theory-based models	theory-based concepts and relations[a]	structured and fixed set[a]	measurements of theory-based concepts[a]	
identifying some model relations with machine learning	theory-based concepts, some automatically detected correlations[b]	structured and fixed set[a]	measurements of theory-based concepts[a]	[8,25]
identifying all model relations with machine learning	theory-based concepts, automatically detected correlations[c]	structured and fixed set[a]	measurements of theory-based concepts[a]	[10,26–57]
finding proxies for missing data	theory-based concepts and relations[a]	structured and fixed set[a]	measurements of theory-based concepts, some measurements based on everyday reasoning[b]	[58]
theory-structured big-data analysis	partly everyday language concepts, partly automatically detected correlations[b]	partly unstructured data stream[c]	measurements based on everyday reasoning[c]	[59,60]
big-data analysis	everyday language concepts, automatically detected correlations[c]	partly unstructured data stream[c]	measurements based on everyday reasoning[c]	[9,61–63]

[a] use theory-based background knowledge	[b] use only partially theory-based background knowledge	[c] do not use theory-based background knowledge

Identifying some model relations with machine learning

A study [25] exemplifying this category created a “hybrid general circulation model”. The parameterizations for longwave and shortwave radiation were replaced by a machine learning

emulator, namely artificial neural networks. This made the simulation process substantially more efficient without adversely affecting the model's accuracy. While the *datasets* and the *measurements* correspond to classical domain science, the modeling partly depends on automatically detected correlations. The variables are, however, still theory-based concepts. Hence, the *models* lie in between classical domain science and pure big data. Another type of studies falling into this category uses machine learning for hypothesis creation as suggested by Caldwell et al [8].

Identifying all model relations with machine learning

This is the category to which we attributed most of the considered case studies. For example, one study [10] created real-time warm wind ("Foehn") forecasting in the Swiss alps using a machine learning algorithm. Two types of forecasts were compared, one of them using 133 predictors from reanalysis datasets, the other one using the air pressure gradients between all surrounding stations, leading to approximately 2,500 predictors. Both approaches worked with a reasonable accuracy. In this category, while the *measurements* and the *datasets* correspond to classical domain science, the model is built entirely upon automatically detected correlations between the variables. Hence, the *models* lie between classical domain science and pure big data. Other examples for this category include the use of machine learning for downscaling of GCM results to a finer spatial or temporal scale [26,31,39,38]; and for predicting climatic variables such as rainfall [33,40] and drought [43].

Finding proxies for missing data

An example for this category is a study [58] which created an indicator to measure the vulnerability of European cities to different climate risks. Background knowledge suggested citizens' awareness of climate change and climate-induced risks should be included, but no data existed. Thus, the authors used standardized frequency with which a city name in combination with the specific climate risks was searched for on Google as a proxy for this variable. The *models* and

datasets correspond to classical domain science because the model relies on domain-specific knowledge for the relations used to construct the indicator, and the datasets are fixed sets with a pre-defined structure. However, the *measurements* were partly based on everyday reasoning due to the inclusion of data from the Google search.

Theory-structured big-data analysis

An example for this category is a study [59] that sought to estimate impacts from Hurricane Sandy in 2012 in New York City using Twitter data but structured the data analysis according to a theoretical framework from human geography, focusing on territory, place, scale, and network. This analysis revealed that while there is a good correlation between hurricane impacts and changes in Twitter activity, this correlation is scale-dependent. The framework allowed the authors to take a critical look at big data for such analyses and also to embed their research into the body of existing literature from human geography. Studies in this category analyze streams of unstructured data and the measurements are based on everyday reasoning. The model relies on automatically detected correlations, but model construction is partly informed by domain-specific scientific knowledge. Other examples belonging to this category use new forms of data, e.g., from video cameras, in order to detect meteorological phenomena such as fog [60].

Big-data analysis

Few reviewed studies fall into the category of pure big data. An example is a further study linking Twitter data to impacts from Hurricane Sandy [9]. Unlike the study in the previous section, it did not structure the analysis according to a theory-based framework but relied fully on automatically detected correlations between everyday-language concepts for the modeling. The study concludes that social media data might provide a useful tool for rapid post-disaster assessment of impacts due to the good correlation of changes in Twitter activity and hurricane impacts. Hence, in this study, the modeling was guided by everyday reasoning without appeal to scientific theory. Of course, the authors hypothesized in advance that social media activity and natural disaster

impacts might be correlated, but this is based on an everyday rather than a theory-based understanding of the system.

General findings

Some reviewed studies [64,65] relied on so-called crowdsourced weather information. Crowdsourcing refers to the process of collecting data from a large number of people [66]. This is potentially relevant in the context of big data because crowdsourced data typically constitute streams of data with measurements based on everyday reasoning. However, these studies can still fall into different categories because the datasets could still be analyzed with different types of models.

The reviewed studies reveal that big data enters scientific research with individual elements such as machine learning methods and new forms of data. While machine learning is already a well-established tool in climate research, new forms of data such as crowdsourced weather data and social media data have rarely been used so far. Based on the studies evaluated, we identify two rationales for inclusion of big-data elements. First, they are included when a more theory-based modeling or data collection would have been too time-consuming, or computationally or financially expensive. Examples include studies that used machine learning to speed up the simulation of GCMs, or when missing data was proxied using big data, even if in principle it could also have been collected in a classical way. We refer to this as the rationale of efficiency. Second, big-data elements were used when the understanding of the target system prohibited a more theory-based modeling approach or measurement process. Examples include the application of machine learning to weather nowcasting, or the analysis of social media data for climate impact assessment studies, as it is unclear how social media activity relates to natural disaster damages. We refer to this as the epistemic rationale. Hence, big data can support an analysis when facing limitations in resources and/or limitations in scientific understanding.

As noted above, we have not categorized data-driven studies dealing with weather-related

technical applications or analyzing climate impacts upon agriculture. We believe that including these studies would not change the insights gained from the overview provided above. For example, a study [67] assessing coffee production in a warmer climate has relied on data from Google Earth and used machine learning methods to identify suitable production locations. Hence, the study combines the categories “finding proxies for missing data” and “identifying all model relations with machine learning”. While our review contains no such a category, the study corroborates our findings about how big-data elements are used in research. Furthermore, the rationales are the same, proxy data are used for efficiency reasons, machine learning is used for efficiency and epistemic reasons [67].

Further studies used machine learning to assess climate change impacts on the global distribution of selenium in soils [68] and for the prediction of power output from wind [69] and solar power [70] based on weather parameters. In these three studies, fixed sets of classical variables that were hypothesized to be relevant were related to the target variable using automatically detected correlations, meaning that they fall into the category “identifying all model relations with machine learning”.

In conclusion, we believe that the sample of categorized studies is sufficiently broad to give an overview of how and why big-data elements are used in climate research. While some studies might fall into categories lying in-between those in Table 1, they are unlikely to yield major new insights.

Conditions for adequacy

There are numerous issues in climate research where researchers are confronted with limitations in either resources or scientific understanding of the target system, indicating potential for big-data elements. However, most of the problems faced by climate researchers do not fall into the realm of “small problems” because repeated evaluation of predictions is not possible. The

reasons for this include the long lead times of climate predictions, for instance when using machine learning for downscaling climate model outputs [26–29,31,32,38,39], or the wide scope of the analyzed problems with unclear measures of success, as Shelton et al [59] demonstrate when using a theoretical framework in the analysis of social media activity and hurricane impacts. Yet, as our review of case studies has highlighted, big-data elements have been employed in climate research also when repeated evaluation was not possible. In these cases, confidence in the predictions is established not through constant evaluation of predictions against new data but by assuming that the identified relationships remain constant over the forecasting horizon, an assumption often only made implicitly. The adapted conditions for successfully applying big-data elements are as follows:

1. The system is predictable for the questions of interest.
2. Sufficient data is available to train the algorithm.
3. The identified relationships between the variables remain sufficiently constant over the relevant configurations of the target system (a), or
   sufficient new data is available to periodically evaluate the predictions against observations and make adjustments to the relationships if necessary (b).

These conditions are necessary for successfully applying big-data elements for predictions, and we assume that they are also jointly sufficient for this purpose. Big data can thus reliably be used beyond “small problems” if scientists have arguments in favor of condition 3a. This condition is quite straightforward and corresponds to an intuition many scientists have concerning statistical tools. Since there is no repeated evaluation, and hence no adaptation of the predictions, the identified relationships need to remain constant over the temporal and spatial horizon of interest. For machine learning algorithms, this is fairly obvious. The constancy assumption is, however, also crucial for other big-data elements, namely for new forms of data. Also, the constancy of the relationships identified do not affect the first and the second condition, as the target system still needs to be predictable, and sufficient data for fitting the algorithm is still needed.

The necessary condition for going beyond small problems has important epistemological implications. Contrary to the repeated evaluation (3b), the constancy assumption (3a) cannot be made based on the data. Rather, the constancy assumption relies on the relevant background knowledge about the target system. Scientists can appeal to notions of a system's linearity or argue that the training dataset at hand covered sufficiently many states of the target system to assume that the relationship identified are of causal nature and hence remain constant [71], at least over configurations of the target system sufficiently similar to the ones covered by the training dataset. When applying big-data elements in such cases, background knowledge is crucial for ensuring robust measurements and reliable model results. Hence, in order to profit from the advantages of big-data elements, namely that they can help to handle limitations in resources and scientific understanding, an optimal path consists in combining theory and more classical scientific approaches with new data-science tools [4].

Going beyond small problems

Classical domain science can be applied beyond problems that require continuous evaluation of the predictions because the theory embedded into its measurements and models justifies extrapolations beyond the observed range. In pure big data however, each component is largely detached from domain-specific scientific theory. This makes it very difficult, and in many cases even impossible, to argue for the constancy assumption. Hence, pure big data is mainly applicable to what we have defined as "small problems". However, our review of case studies shows that in climate research, big-data elements have been applied beyond "small" problems. Based on our considerations, this is justified when these elements are combined with theory-based approaches, which helps to argue for the constancy of the identified relationships. But for which specific areas of climate research could big-data elements be useful? The two rationales identified above suggest that they can be useful whenever scientists face limitations in their resources or their understanding of the target system. The review of case studies shows that the

most common big-data element in climate research is machine learning used with standard climate data, but we believe that other interesting but yet unexploited applications for big-data elements exist. In the following, we speculate on where specifically we see the biggest potential.

Analyzing increasing volumes of climate data

The volume and complexity of data produced and stored are large and expected to further increase [7]. Increasingly, scientists will face difficulties in analyzing these data following more traditional methodological pathways. Machine learning can help scientists to find patterns in large volumes of data from climate models or satellites and potentially to formulate hypotheses [8]. However, this requires appropriate background knowledge to distinguish between potentially meaningful and meaningless patterns [72]. This is especially true for datasets with a very large number of variables.

Climate impact research

As the drivers and physical consequences of climate change are better understood, researchers increasingly turn to socio-economic impacts of climate change. Big-data elements could prove useful in this area of research because for such target systems, there are no well-confirmed universal theories. Hence the ability to construct theory-based impact models is limited, but researchers still have some understanding of how the target system works. Pertinent background knowledge might be sufficient for making the constancy assumption regarding the identified relationships for certain timescales and spatial scales.

There are different ways in which big-data elements could improve climate impact modeling. New forms of data are useful for calibrating impact models. Data from crowdsourcing and crowdsensing specifically collected for a given purpose might be useful as the constancy assumption can be justified by appealing to the user basis. An example for such a study would be the use of GPS data from phones to track where and how people move [61]. Furthermore, machine learning might be a promising choice of method for assessing the impacts of extreme weather

events on technical and other complex systems. For instance, machine learning could be used to assess asset damages from severe weather events and extrapolate these results into future climatic regimes given that scientists have some understanding of the relationships between these variables and might hence be able to justify the constancy assumption in impact processes. Studies on asset damages from severe weather events typically use damage curves to link the weather parameters and the damages to exposed assets such as insured financial losses [73]. Using machine learning instead of simpler damage curves could lead to a more fine-grained and more accurate analysis. While in such climate impact studies, adaptation measures can run contrary to the constancy assumption [74], the constancy assumption could still be fulfilled for the estimation of a ceteris paribus baseline scenario.

Climate services

Increasing volumes of climate data make it possible to provide more tailored information to users, often referred to as "climate services" [75]. In order for climate scientists to deliver information that fits users' needs, big-data elements could become increasingly important. There are several case studies employing machine learning for downscaling of GCM results to a more local scale. It has already been suggested that large volumes of climate data could improve climate services in this way [76]. However, one could go one step further by combining these localized variables with user-specific data and thus providing tailor-made climate services to users as is being developed in personalized medicine. For example, farmers' decisions on specific farming practices depend on climatological variables. A useful climate service would be to partly automate this decision by considering a few key variables that can be predicted at the time of planting seeds. Such variables could be identified by combining climatological data with observed data at the farm level with machine learning. Decision trees can help to identify crop diseases in plants [77]. Similarly, machine learning and a dense network of climate and weather data might render farming practices more efficient [78], and hence contribute to more climate-resilient agriculture (often labeled "climate-smart agriculture", see ref 79). In such cases, the understanding of the target

system might justify the constancy assumption especially when the forecasting horizon is comparatively short.

Small problems in climate research

Finally, there is also room for solving relevant "small" problems in climate research, which neither implies that they are unimportant, nor that they are easy to solve. For instance, it has been suggested to compare forecasts from high-resolution models to observations when they become available and make corrections either to model output or to parameterizations in these models if necessary [80]. This approach could be assisted by machine learning [81]. This would essentially solve a small problem within the framework of a very complex problem.

Conclusion

In this article, we have reviewed case studies from climate research and shown that many categories exist between classical domain science and pure big data. While pure big data requires constant evaluation of the predictions, combining big-data elements with more classical theory-driven approaches can help to justify the constancy assumption that allows going beyond "small problems." Hence, big-data elements can potentially be beneficial to overcome limitations in resources and scientific understanding in climate research but most likely not replace approaches based on theory and understanding. Many of the points raised in this article can be extended beyond climate research and transferred to research domains investigating complex phenomena with increasing volumes of stored data. Certain aspects of climate research make the use of big data particularly challenging, in particular the long forecasting lead times relative to the short periods for which data is available. However, we expect that the framework used here, as well as the rationales and conditions for using big data could be fruitfully used by other fields.

Corresponding Author

Correspondence should be addressed to B.K..

Author Contributions

B.K. reviewed and classified the studies and led the writing with contributions from all authors. All authors contributed to the framing and the development of the ideas of the paper.

Competing interests

The authors declare no competing interests.

Acknowledgements

We thank Claus Beisbart, Anna Merrifield, Sebastian Sippel, Rosemarie McMahon, and Johan Lilliestam for discussions and comments which have improved the quality of this manuscript. The research was supported by the Swiss National Science Foundation, National Research Programme 75 Big Data, project No 167215.

References

1. Mayer-Schönberger, V. & Cukier, K. *Big Data: A Revolution that Will Transform how We Live, Work and Think*. (John Murray, 2013).
2. Lyon, A. Data. in *The Oxford Handbook of the Philosophy of Science* (ed. Humphreys, P.) (Oxford University Press, 2015).
3. Pietsch, W. & Wernecke, J. Introduction: Ten Theses on Big Data and Computability. in *Berechenbarkeit der Welt? Philosophie und Wissenschaft im Zeitalter von Big Data* (eds. Pietsch, W., Wernecke, J. & Ott, M.) 37–57 (Springer VS, 2017).
4. Karpatne, A. *et al.* Theory-Guided Data Science: A New Paradigm for Scientific Discovery

from Data. *IEEE Trans. Knowl. Data Eng.* **29**, 2318–2331 (2017).

5. Faghmous, J. H. & Kumar, V. A Big Data Guide to Understanding Climate Change: The Case for Theory-Guided Data Science. *Big Data* **2**, 155–163 (2014).
6. Ford, J. D. *et al.* Big data has big potential for applications to climate change adaptation. *Proc. Natl. Acad. Sci.* **113**, 10729–10732 (2016).
7. Overpeck, J. T., Meehl, G. A., Bony, S. & Easterling, D. R. Climate Data Challenges in the 21st Century. *Science* **331**, 700–702 (2011).
8. Caldwell, P. M. *et al.* Statistical significance of climate sensitivity predictors obtained by data mining. *Geophys. Res. Lett.* **41**, 1803–1808 (2014).
9. Kryvasheyeu, Y. *et al.* Rapid assessment of disaster damage using social media activity. *Sci. Adv.* **2**, 1–11 (2016).
10. Sprenger, M., Schemm, S., Oechslin, R. & Jenkner, J. Nowcasting Foehn Wind Events Using the AdaBoost Machine Learning Algorithm. *Weather Forecast.* **32**, 1079–1099 (2017).
11. Baumberger, C., Knutti, R. & Hirsch Hadorn, G. Building confidence in climate model projections: an analysis of inferences from fit. *Wiley Interdiscip. Rev. Clim. Change* **8**, e454 (2017).
12. Boyd, D. & Crawford, K. Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Inf. Commun. Soc.* **15**, 662–679 (2012).
13. De Mauro, A., Greco, M. & Grimaldi, M. A formal definition of Big Data based on its essential features. *Libr. Rev.* **65**, 122–135 (2016).
14. Kitchin, R. & McArdle, G. What makes Big Data, Big Data? Exploring the ontological characteristics of 26 datasets. *Big Data Soc.* **3**, 1–10 (2016).
15. Lukoianova, T. & Rubin, V. L. Veracity Roadmap: Is Big Data Objective, Truthful and Credible? *Adv. Classif. Res. Online* **24**, 4 (2014).
16. Hastie, T., Tibshirani, R. & Friedman, J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* (Springer, 2008).
17. LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521**, 436–444 (2015).

18. Linden, G., Smith, B. & York, J. Amazon.com recommendations: item-to-item collaborative filtering. *IEEE Internet Comput.* **7**, 76–80 (2003).
19. Goertzel, B. & Pennachin, C. *Artificial general intelligence*. (Springer, 2007).
20. Manogaran, G. & Lopez, D. Spatial cumulative sum algorithm with big data analytics for climate change detection. *Comput. Electr. Eng.* **65**, 207–221 (2018).
21. Manogaran, G., Lopez, D. & Chilamkurti, N. In-Mapper combiner based MapReduce algorithm for processing of big climate data. *Future Gener. Comput. Syst.* **86**, 433–445 (2018).
22. McGuffie, K. & Henderson-Sellers, A. *A Climate Modelling Primer*. (John Wiley & Sons, 2005).
23. Müller, P. Constructing climate knowledge with computer models. *Wiley Interdiscip. Rev. Clim. Change* **1**, 565–580 (2010).
24. Knutti, R. Should we believe model predictions of future climate change? *Philos. Trans. R. Soc. Math. Phys. Eng. Sci.* **366**, 4647–4664 (2008).
25. Krasnopolsky, V. M. & Fox-Rabinovitz, M. S. Complex hybrid models combining deterministic and machine learning components for numerical climate modeling and weather prediction. *Neural Netw.* **19**, 122–134 (2006).
26. Tripathi, S., Srinivas, V. V. & Nanjundiah, R. S. Downscaling of precipitation for climate change scenarios: A support vector machine approach. *J. Hydrol.* **330**, 621–640 (2006).
27. Ghosh, S. & Mujumdar, P. P. Statistical downscaling of GCM simulations to streamflow using relevance vector machine. *Adv. Water Resour.* **31**, 132–146 (2008).
28. Mendes, D. & Marengo, J. A. Temporal downscaling: a comparison between artificial neural network and autocorrelation techniques over the Amazon Basin in present and future climate change scenarios. *Theor. Appl. Climatol.* **100**, 413–421 (2010).
29. Chen, S.-T., Yu, P.-S. & Tang, Y.-H. Statistical downscaling of daily precipitation using support vector machines and multivariate analysis. *J. Hydrol.* **385**, 13–22 (2010).
30. Wenzel, M. & Schröter, J. Reconstruction of regional mean sea level anomalies from tide

gauges using neural networks. *J. Geophys. Res.* **115**, (2010).

31. Chadwick, R., Coppola, E. & Giorgi, F. An artificial neural network technique for downscaling GCM outputs to RCM spatial scale. *Nonlinear Process. Geophys.* **18**, 1013–1028 (2011).
32. Raje, D. & Mujumdar, P. P. A comparison of three methods for downscaling daily precipitation in the Punjab region. *Hydrol. Process.* **25**, 3575–3589 (2011).
33. Abbot, J. & Marohasy, J. Application of artificial neural networks to rainfall forecasting in Queensland, Australia. *Adv. Atmospheric Sci.* **29**, 717–730 (2012).
34. Gagne II, D. J., McGovern, A., Basara, J. B. & Brown, R. A. Tornadic Supercell Environments Analyzed Using Surface and Reanalysis Data: A Spatiotemporal Relational Data-Mining Approach. *J. Appl. Meteorol. Climatol.* **51**, 2203–2217 (2012).
35. Rasouli, K., Hsieh, W. W. & Cannon, A. J. Daily streamflow forecasting by machine learning methods with weather and climate inputs. *J. Hydrol.* **414–415**, 284–293 (2012).
36. Mekanik, F., Imteaz, M. A., Gato-Trinidad, S. & Elmahdi, A. Multiple regression and Artificial Neural Network for long-term rainfall forecasting using large scale climate modes. *J. Hydrol.* **503**, 11–21 (2013).
37. Merz, B., Kreibich, H. & Lall, U. Multi-variate flood damage assessment: a tree-based data-mining approach. *Nat. Hazards Earth Syst. Sci.* **13**, 53–64 (2013).
38. Nasseri, M., Tavakol-Davani, H. & Zahraie, B. Performance assessment of different data mining methods in statistical downscaling of daily precipitation. *J. Hydrol.* **492**, 1–14 (2013).
39. Tavakol-Davani, H., Nasseri, M. & Zahraie, B. Improved statistical downscaling of daily precipitation using SDSM platform and data-mining methods. *Int. J. Climatol.* **33**, 2561–2578 (2013).
40. Abbot, J. & Marohasy, J. Input selection and optimisation for monthly rainfall forecasting in Queensland, Australia, using artificial neural networks. *Atmospheric Res.* **138**, 166–178 (2014).
41. McGovern, A., Gagne, D. J., Williams, J. K., Brown, R. A. & Basara, J. B. Enhancing understanding and improving prediction of severe weather through spatiotemporal relational

learning. *Mach. Learn.* **95**, 27–50 (2014).

42. Abbot, J. & Marohasy, J. Using artificial intelligence to forecast monthly rainfall under present and future climates for the bowen basin, Queensland, Australia. *Int. J. Sustain. Dev. Plan.* **10**, 66–75 (2015).
43. Deo, R. C. & Şahin, M. Application of the extreme learning machine algorithm for the prediction of monthly Effective Drought Index in eastern Australia. *Atmospheric Res.* **153**, 512–525 (2015).
44. Mohammadi, K. *et al.* Extreme learning machine based prediction of daily dew point temperature. *Comput. Electron. Agric.* **117**, 214–225 (2015).
45. Patil, A. P. & Deka, P. C. An extreme learning machine approach for modeling evapotranspiration using extrinsic inputs. *Comput. Electron. Agric.* **121**, 385–392 (2016).
46. Salcedo-Sanz, S., Deo, R. C., Carro-Calvo, L. & Saavedra-Moreno, B. Monthly prediction of air temperature in Australia and New Zealand with machine learning algorithms. *Theor. Appl. Climatol.* **125**, 13–25 (2016).
47. Andersen, H., Cermak, J., Fuchs, J., Knutti, R. & Lohmann, U. Understanding the drivers of marine liquid-water cloud occurrence and properties with global observations using neural networks. *Atmospheric Chem. Phys.* 9535–9546 (2017) doi:10.5194/acp-2017-282.
48. Das, S., Chakraborty, R. & Maitra, A. A random forest algorithm for nowcasting of intense precipitation events. *Adv. Space Res.* **60**, 1271–1282 (2017).
49. Dayal, K., Deo, R. & Apan, A. A. Drought Modelling Based on Artificial Intelligence and Neural Network Algorithms: A Case Study in Queensland, Australia. in *Climate Change Adaptation in Pacific Countries* (ed. Leal Filho, W.) 177–198 (Springer International Publishing, 2017). doi:10.1007/978-3-319-50094-2_11.
50. Eghdamirad, S., Johnson, F. & Sharma, A. Using second-order approximation to incorporate GCM uncertainty in climate change impact assessments. *Clim. Change* **142**, 37–52 (2017).
51. Majdzadeh Moghadam, F. Neural Network-Based Approach for Identification of Meteorological Factors Affecting Regional Sea-Level Anomalies. *J. Hydrol. Eng.* **22**,

04016058-1-04016058–15 (2017).

52. Kashiwao, T. *et al.* A neural network-based local rainfall prediction system using meteorological data on the Internet: A case study using data from the Japan Meteorological Agency. *Appl. Soft Comput.* **56**, 317–330 (2017).
53. Park, S., Im, J., Park, S. & Rhee, J. Drought monitoring using high resolution soil moisture through multi-sensor satellite data fusion over the Korean peninsula. *Agric. For. Meteorol.* **237–238**, 257–269 (2017).
54. Rahmati, O. & Pourghasemi, H. R. Identification of Critical Flood Prone Areas in Data-Scarce and Ungauged Regions: A Comparison of Three Data Mining Models. *Water Resour. Manag.* **31**, 1473–1487 (2017).
55. Roodposhti, M. S., Safarrad, T. & Shahabi, H. Drought sensitivity mapping using two one-class support vector machine algorithms. *Atmospheric Res.* **193**, 73–82 (2017).
56. Wu, J. *et al.* Establishing and assessing the Integrated Surface Drought Index (ISDI) for agricultural drought monitoring in mid-eastern China. *Int. J. Appl. Earth Obs. Geoinformation* **23**, 397–410 (2013).
57. Zhou, L. *et al.* Quantitative and detailed spatiotemporal patterns of drought in China during 2001–2013. *Sci. Total Environ.* **589**, 136–145 (2017).
58. Tapia, C. *et al.* Profiling urban vulnerabilities to climate change: An indicator-based vulnerability assessment for European cities. *Ecol. Indic.* **78**, 142–155 (2017).
59. Shelton, T., Poorthuis, A., Graham, M. & Zook, M. Mapping the data shadows of Hurricane Sandy: Uncovering the sociospatial dimensions of 'big data'. *Geoforum* **52**, 167–179 (2014).
60. Castelli, R. *et al.* Fog detection from camera images. in *SWI 2016* 19 (2016).
61. Lu, X. *et al.* Detecting climate adaptation with mobile network data in Bangladesh: anomalies in communication, mobility and consumption patterns during cyclone Mahasen. *Clim. Change* **138**, 505–519 (2016).
62. Preis, T., Moat, H. S., Bishop, S. R., Treleaven, P. & Stanley, H. E. Quantifying the Digital Traces of Hurricane Sandy on Flickr. *Sci. Rep.* **3**, 1–3 (2013).

63. Tkachenko, N., Jarvis, S. & Procter, R. Predicting floods with Flickr tags. *PLOS ONE* **12**, e0172870 (2017).

64. Overeem, A. *et al.* Crowdsourcing urban air temperatures from smartphone battery temperatures. *Geophys. Res. Lett.* **40**, 4081–4085 (2013).

65. Elmore, K. L. *et al.* MPING: Crowd-Sourcing Weather Reports for Research. *Bull. Am. Meteorol. Soc.* **95**, 1335–1342 (2014).

66. Muller, C. L. *et al.* Crowdsourcing for climate and atmospheric sciences: current status and future potential. *Int. J. Climatol.* **35**, 3185–3203 (2015).

67. Bunn, C., Läderach, P., Ovalle Rivera, O. & Kirschke, D. A bitter cup: climate change profile of global production of Arabica and Robusta coffee. *Clim. Change* **129**, 89–101 (2015).

68. Jones, G. D. *et al.* Selenium deficiency risk predicted to increase under future climate change. *Proc. Natl. Acad. Sci.* **114**, 2848–2853 (2017).

69. Foley, A. M., Leahy, P. G., Marvuglia, A. & McKeogh, E. J. Current methods and advances in forecasting of wind power generation. *Renew. Energy* **37**, 1–8 (2012).

70. Inman, R. H., Pedro, H. T. C. & Coimbra, C. F. M. Solar forecasting methods for renewable energy integration. *Prog. Energy Combust. Sci.* **39**, 535–576 (2013).

71. Pietsch, W. The Causal Nature of Modeling with Big Data. *Philos. Technol.* **29**, 137–171 (2016).

72. Masson, D. & Knutti, R. Predictor Screening, Calibration, and Observational Constraints in Climate Model Ensembles: An Illustration Using Climate Sensitivity. *J. Clim.* **26**, 887–898 (2013).

73. Welker, C. *et al.* Modelling economic losses of historic and present-day high-impact winter windstorms in Switzerland. *Tellus Dyn. Meteorol. Oceanogr.* **68**, 29546 (2016).

74. Arbuthnott, K., Hajat, S., Heaviside, C. & Vardoulakis, S. Changes in population susceptibility to heat and cold over time: assessing adaptation to climate change. *Environ. Health* **15**, (2016).

75. Vaughan, C. & Dessai, S. Climate services for society: origins, institutional arrangements,

and design elements for an evaluation framework: Climate services for society. *Wiley Interdiscip. Rev. Clim. Change* **5**, 587–603 (2014).

76. Benestad, R., Parding, K., Dobler, A. & Mezghani, A. A strategy to effectively make use of large volumes of climate data for climate change adaptation. *Clim. Serv.* **6**, 48–54 (2017).
77. Wahabzada, M. *et al.* Plant Phenotyping using Probabilistic Topic Models: Uncovering the Hyperspectral Language of Plants. *Sci. Rep.* **6**, (2016).
78. Walter, A., Finger, R., Huber, R. & Buchmann, N. Opinion: Smart farming is key to developing sustainable agriculture. *Proc. Natl. Acad. Sci.* **114**, 6148–6150 (2017).
79. Lipper, L. *et al.* Climate-smart agriculture for food security. *Nat. Clim. Change* **4**, 1068–1072 (2014).
80. Katzav, J. & Parker, W. S. The future of climate modeling. *Clim. Change* **132**, 475–487 (2015).
81. Schneider, T., Lan, S., Stuart, A. & Teixeira, J. Earth System Modeling 2.0: A Blueprint for Models That Learn From Observations and Targeted High-Resolution Simulations. *Geophys. Res. Lett.* **44**, 12,396-12,417 (2017).