DISS. ETH NO. 26626

Epistemological Issues in Data-Driven Modeling in Climate Research

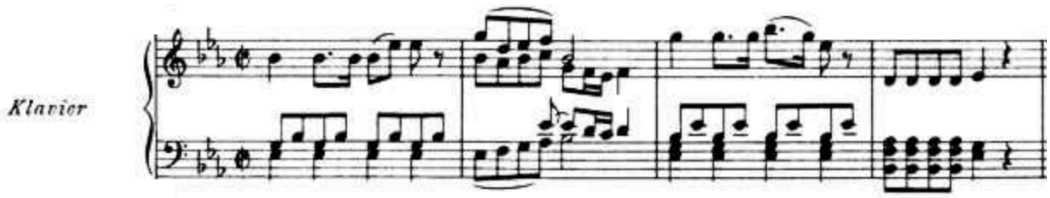
A thesis submitted to attain the degree of
DOCTOR OF SCIENCES of ETH ZURICH (Dr. sc. ETH Zurich)

presented by
BENEDIKT KNÜSEL
MSc ETH Environ. Sc, ETH Zurich

Born on 31.05.1990
citizen of
Gisikon (LU)

accepted on the recommendation of
Prof. Dr. David N. Bresch, examiner
Prof. Dr. Reto Knutti, co-examiner
Dr. Christoph Baumberger, co-examiner
Prof. Dr. Mathias Frisch, co-examiner

2020



Principal theme of the second movement (Larghetto)
Piano Concerto No. 24 in C minor (K. 491)
Wolfgang Amadeus Mozart

(copyright information on next page)

The sheet music excerpt on the previous page is adapted from the *Neue Mozart-Ausgabe* (Series V/15, Volume 7, published in 1959, edited by Hermann Beck, page 124), which is published here:

Digital Interactive Mozart Edition, published by the Internationale Stiftung Mozarteum, Salzburg (https://dme.mozarteum.at/nmaonline/, accessed on January 6, 2020).

Abstract

Recent years have seen a dramatic increase in the volumes of data that are produced, stored, and analyzed. This advent of *big data* has led to commercial success stories, for example in recommender systems in online shops. However, scientific research in various disciplines including environmental and climate science will likely also benefit from increasing volumes of data, new sources for data, and the increasing use of algorithmic approaches to analyze these large datasets. This thesis uses tools from philosophy of science to conceptually address epistemological questions that arise in the analysis of these increasing volumes of data in environmental science with a special focus on data-driven modeling in climate research. Data-driven models, here, are defined as models of phenomena that are built with machine learning. While epistemological analyses of machine learning exist, these have mostly been conducted for fields characterized by a lack of hierarchies of theoretical background knowledge. Such knowledge is often available in environmental science and especially in physical climate science, and it is relevant for the construction, evaluation, and use of data-driven models. This thesis investigates predictions, uncertainty, and understanding from data-driven models in environmental and climate research and engages in in-depth discussions of case studies. These three topics are discussed in three topical chapters.

The first chapter addresses the term "big data", and rationales and conditions for the use of big-data elements for predictions. Namely, it uses a framework for classifying case studies from climate research and shows that "big data" can refer to a range of different activities. Based on this classification, it shows that most case studies lie in between classical domain science and pure big data. The chapter specifies necessary conditions for the use of big data and shows that in most scientific applications, background knowledge is essential to argue for the constancy of the identified relationships. This constancy assumption is relevant both for new forms of measurements and for data-driven models. Two rationales for the use of big-data elements are identified. Namely, big-data elements can help to overcome limitations in financial, computational, or time resources, which is referred to as the rationale of efficiency. Big-data elements can also help to build models when system understanding does not allow for a more theory-guided modeling approach, which is referred to as the epistemic rationale.

The second chapter addresses the question of predictive uncertainties of data-driven models. It highlights that existing frameworks for understanding and characterizing uncertainty focus on specific locations of uncertainty, which are not informative for the predictive uncertainty of data-driven models. Hence, new approaches are needed for this task. A framework is developed and presented that focuses on the justification of the fitness-for-purpose of the models for the specific kind of prediction at hand. This framework uses argument-based tools and distinguishes between first-order and second-order epistemic uncertainty. First-order uncertainty emerges when it cannot be conclusively justified that the model is maximally fit-for-purpose. Second-order uncertainty emerges

when it is unclear to what extent the fitness-for-purpose assumption and the underlying assumptions are justified. The application of the framework is illustrated by discussing a case study of data-driven projections of the impact of climate change on global soil selenium concentrations. The chapter also touches upon how the information emerging from the framework can be used in decision-making.

The third chapter addresses the question of scientific understanding. A framework is developed for assessing the fitness of a model for providing understanding of a phenomenon. For this, the framework draws from the philosophical literature on scientific understanding and focuses on the representational accuracy, the representational depth, and the graspability of a model. Then, based on the framework, the fitness of data-driven and process-based climate models for providing understanding of phenomena is compared. It is concluded that data-driven models can, under some conditions, be fit to serve as vehicles for understanding to a satisfactory extent. This is specifically the case when sufficient background knowledge is available such that the coherence of the model with background knowledge provides good reasons for the representational accuracy of the data-driven model, which can be assessed e.g. through sensitivity analyses. This point is illustrated by discussing a case study from atmospheric physics in which data-driven models are used to better understand the drivers of a specific type of clouds.

The work of this thesis highlights that while big data is no panacea for scientific research, data-driven modeling offers new tools to scientists that can be very useful for a variety of questions. All three studies emphasize the importance of background knowledge for the construction and evaluation of data-driven models as this helps to obtain models that are representationally accurate. The importance of domain-specific background knowledge and the technical challenges of implementing data-driven models for complex phenomena highlight the importance of interdisciplinary work. Previous philosophical work on machine learning has stressed that the problem framing makes models theory-laden. This thesis shows that in a field like climate research, the model evaluation is strongly guided by theoretical background knowledge, which is also important for the theory-ladenness of data-driven modeling. The results of the thesis are relevant for a range of methodological questions regarding data-driven modeling and for philosophical discussions of models that go beyond data-driven models.

Zusammenfassung

In den vergangenen Jahren ist die Menge an neu generierten, gespeicherten und analysierten Daten stark angestiegen. «Big Data», wie diese Datenströme genannt werden, hat im Privatsektor zu kommerziellen Erfolgen geführt, beispielsweise in der Form von personalisierten Empfehlungen in Online-Shops. Big Data ist aber auch für die Forschung von Bedeutung, beispielsweise in den Umwelt- und Klimawissenschaften, da grössere Datenvolumen, neue Datenquellen und die zunehmende Verwendung von algorithmischen Instrumenten zur Analyse dieser Datenströme neue wissenschaftliche Möglichkeiten bieten. Diese Doktorarbeit befasst sich konzeptionell mit epistemologischen Fragen, die bei der Analyse von immer grösseren Datenvolumen in den Umweltwissenschaften entstehen. Ein spezieller Fokus liegt auf der datengetriebenen Modellierung in der Klimaforschung. Der Begriff «datengetriebene Modelle» steht dabei für Modelle von Phänomenen, die mittels maschinellen Lernens konstruiert werden. Abhandlungen zu epistemologischen Fragen des maschinellen Lernens fokussierten bisher meist auf Disziplinen ohne Hierarchien von theoretischem Hintergrundwissen. Solches Wissen ist in den Umwelt- und vor allem in den physikalischen Klimawissenschaften aber oft vorhanden und für das Konstruieren, Bewerten und Anwenden von datengetriebenen Modellen wichtig. In der vorliegenden Arbeit werden spezifisch für die Umwelt- und Klimawissenschaften Vorhersagen und Unsicherheiten von datengetriebenen Modellen sowie das wissenschaftliche Verständnis, das aus solchen Modellen gewonnen werden kann, untersucht und anhand von Fallstudien vertieft diskutiert. Diese drei Themen werden in je einem eigenen Kapitel abgehandelt.

Das erste Kapitel beschäftigt sich einerseits mit dem Begriff «Big Data» und andererseits mit den Voraussetzungen für erfolgreiche Vorhersagen mit Big-Data-Elementen und den Gründen für deren Verwendung. Anhand einer Klassifikation von Fallstudien aus der Klimaforschung wird gezeigt, dass eine Reihe verschiedener Aktivitäten im weitesten Sinne unter den Begriff «Big Data» fällt, wobei die meisten Studien im Bereich zwischen der klassischen Wissenschaft und reinem Big Data liegen. Wir schlagen notwendige Bedingungen für das Anwenden von Big-Data-Elementen vor und zeigen, dass Hintergrundwissen in vielen wissenschaftlichen Anwendungen zentral ist, um die Konstanz der identifizierten Beziehungen zu begründen. Diese Konstanzannahme ist sowohl für neue Formen von Messungen wie auch für die datengetriebene Modellierung wichtig. In unserer Analyse kristallisieren sich zwei Gründe für das Verwenden von Big-Data-Elementen heraus. Einerseits können diese Elemente nützlich sein, wenn beschränkte Ressourcen – seien es finanzielle Ressourcen, die Zeit oder Rechenleistung – andere Ansätze erschweren. Dies ist der Effizienzgrund für das Verwenden von Big Data. Andererseits können Big-Data-Elemente die Modellierung von Phänomenen ermöglichen, wenn aufgrund des Systemverständnisses keine stärker auf Theorien basierende Modellierung möglich ist. Dies ist der epistemische Grund für das Verwenden von Big Data.

Das zweite Kapitel widmet sich der Unsicherheit der Vorhersagen aus datengetriebenen Modellen. Etablierte Ansätze zur Charakterisierung von Unsicherheiten setzen bei den Stellen im Modellierungsprozess an, an denen die Unsicherheit auftritt. Für die Unsicherheit datengetriebener Modelle sind diese Stellen aber nicht informativ, weshalb neue Ansätze nötig sind. Wir schlagen einen Ansatz vor, der auf die Rechtfertigung der Eignung des Modelles für einen bestimmten prädiktiven Zweck fokussiert und diese mittels einer Argumentanalyse untersucht. Dabei wird zwischen epistemischen Unsicherheiten erster und zweiter Ordnung unterschieden. Unsicherheiten erster Ordnung treten auf, wenn nicht zwingend begründet werden kann, dass das Modell maximal für den anvisierten Zweck geeignet ist. Unsicherheiten zweiter Ordnung treten auf, wenn unklar ist, in welchem Grad die Annahme der Eignung des Modells und weitere dieser Annahme zugrundeliegende Annahmen gerechtfertigt sind. Der Ansatz wird mit einer Fallstudie zu datengetriebenen Vorhersagen der Klimaauswirkungen auf die globale Konzentration von Selen in Böden illustriert. Zudem wird diskutiert, wie die Information aus diesem Ansatz für Entscheidungssituationen verwendet werden kann.

Das dritte Kapitel behandelt das wissenschaftliche Verständnis von Phänomenen. Konkret wird ein Ansatz entwickelt, um die Eignung von Modellen für das Verstehen von Phänomenen zu beurteilen. Dieser Ansatz baut auf der philosophischen Literatur zum Thema wissenschaftliches Verstehen auf und beurteilt Modelle anhand ihrer repräsentationalen Genauigkeit, ihrer repräsentationalen Tiefe und ihrer Intelligibilität. Anhand dieses Ansatzes vergleichen wir die Eignung von prozessbasierten und datengetriebenen Modellen als Vehikel für das Verstehen von Phänomenen. Es zeigt sich, dass datengetriebene Modelle unter bestimmten Voraussetzungen hinreichend geeignet sein können, um Verständnis zu liefern. Dies verlangt allerdings, dass das Phänomen bereits soweit verstanden ist, dass die Kohärenz des Modells mit dem Hintergrundwissen gute Gründe für die repräsentationale Genauigkeit des Modells liefert. Diese Beurteilung kann zum Beispiel mithilfe von Sensitivitätsanalysen durchgeführt werden. Diese Punkte illustrieren wir mit einer Fallstudie aus der Atmosphärenphysik, in der datengetriebene Modelle von Wolken zu einem besseren Verständnis der Einflussfaktoren dieser Wolken führen.

Die vorliegende Doktorarbeit unterstreicht, dass Big Data zwar kein Allheilmittel für wissenschaftliche Probleme ist, die datengetriebene Modellierung aber für viele wissenschaftliche Fragen nützlich sein kann. In allen drei Kapiteln wird die Bedeutung des theoretischen Hintergrundverständnisses für das Konstruieren und Evaluieren von datengetriebenen Modellen betont, da dies nötig ist, um Modelle zu entwickeln, die ihr Zielsystem hinreichend genau abbilden. Aus diesem Grund ist die interdisziplinäre Zusammenarbeit wichtig. Bestehende philosophische Arbeiten betonen, dass mittels maschinellen Lernens konstruierte Modelle aufgrund des *Problemframings* theoriegeladen sind. Die Resultate dieser Arbeit zeigen, dass in Disziplinen wie der Klimaforschung auch die stark vom Hintergrundverständnis getriebene Modellevaluation wichtig ist. Die hier erarbeiteten Resultate sind für eine Reihe methodologischer Fragen bezüglich der

datengetriebenen Modellierung sowie für die philosophischen Diskussionen von Modellen, die über datengetriebene Modelle hinausgehen, von Bedeutung.

Acknowledgments

I am deeply grateful for the support I received from my supervisors David Bresch, Reto Knutti, and Christoph Baumberger. Having three supervisors from diverse backgrounds who all feel responsible for the success of a thesis was a gift that helped me in reaching the goals I initially set for this thesis. I am especially grateful to Reto Knutti and David Bresch for their generosity and for the guidance they provided. I thank Christoph Baumberger for the supervision and friendship over the last three years. The support that he provided went beyond what I ever imagined from a supervisor!

I am grateful to the Swiss National Science Foundation for its financial support through the NRP75. I specifically acknowledge the support by program manager Christian Mottas.

Many researchers that I look up to have helped me in clarifying my thoughts and to write such an interdisciplinary thesis. I am deeply grateful to Gertrude Hirsch Hadorn. She was not only supportive during the time of this thesis but also crucially involved in making this project happen in the first place and advised me in many situations. I thank Marius Zumwald for his friendship and for the good collaboration we had on this project. I thank Roman Frigg and the Centre for Philosophy of Natural and Social Science at the London School of Economics and Political Science for providing such a welcoming environment during my research stay in London. I thank Wendy Parker who provided help for the overall framing of this thesis and for each of the chapters of this thesis.

I thank all of my colleagues in the Weather and Climate Risks Group and the Climate Physics Group at ETH for their friendship and support. I am especially grateful to my PhD colleagues from the WCR group, Samuel Eberenz, Luise Fischer, Thomas Röösli, Maurice Skelton, and Marius Zumwald. I would also like to thank all members of the D-USYS TdLab for tolerating me in their break room and for the many lively discussions over coffee or lunch. The last three years would not have been the same without them! I thank Sandro Bösch, Sarah Spitzauer, and Barbara Wittneben for all of the administrative and IT support.

Finally, I am deeply grateful to my family and friends. I thank Philippe Knüsel and Annina Michel for reviewing chapters of this thesis and helping with the layout. Furthermore, I would like to specifically acknowledge the invaluable support I received particularly during the more difficult times of my PhD work, especially from my parents Hans and Giselle Knüsel, as well as Annina Michel, Philippe Knüsel, Evelyne Knüsel, Daniela Seiler, Mischa Kaspar, and Xenia Wietlisbach. Without them, the obstacles of the last three years would not look so insignificant in hindsight.

Contents

Figures

Tables

1. Introduction

The volumes and complexity of data that society produces and stores have been increasing dramatically over recent decades. Technological innovations have created opportunities to efficiently handle these massive datasets and to draw useful inferences from them, but as the pace of data generation keeps increasing, so does the need for technological innovation (National Research Council 2013). There are numerous examples of successful predictions made based on big-data analysis, many of which have taken place in the private sector. These include e-commerce platforms which use data on their customers to provide individual recommendations based on past purchases (Mayer-Schönberger and Cukier 2013), social media platforms where content is prioritized based on a user's interests, and image and speech recognition (National Research Council 2013).

Big data becomes more and more important for scientific research, too, in the form of increasing volumes of data and new methods to analyze them. As Pietsch and Wernecke (2017) have argued, scientific research is currently experiencing a move away from more complex models that are rooted in scientific theory and use comparatively little data, towards simpler models that are not explicitly rooted in theory but use lots of data. Hence, comparatively theory-free, data-driven models constructed with machine learning could become increasingly more important in scientific research (see Lyon 2015). However, even if such models are not constructed by prescribing the relationships between the variables based on theory, some aspects of model construction are still guided by theory, for example the choice of variables (Pietsch 2015). The same might hold true also for the evaluation and use of models for specific purposes. This changing role of scientific theory in data-driven modeling implies important epistemological shifts regarding the construction, evaluation, and use of models and the justification of model results, which requires in-depth analysis.

The fact that data-driven modeling techniques become more and more relevant in scientific research seems particularly relevant for the broad field of environmental science in general and for climate science in particular. There are several reasons for this. First, more and more data is available to Earth scientists (Reichstein et al. 2019) and climate scientists (Overpeck et al. 2011). Second, research in environmental science is

inherently interdisciplinary. This interdisciplinary character often requires that heterogeneous datasets be linked, which is one of the features typically associated with big data (Sun and Scanlon 2019). This is also relevant for environmental social sciences, where human behavior can be analyzed using new forms and sources of data, e.g., in the wake of natural disasters. Social media data, for example, has already been analyzed to understand the response to a number of different disasters (see, e.g., Preis et al. 2013; Shelton et al. 2014; Kryvasheyeu et al. 2016). Third, there are many examples of data-driven modeling techniques and new forms of data used in environmental and climate research (for overviews, see Huntingford et al. 2019; Reichstein et al. 2019; Sun and Scanlon 2019). Thus, environmental and climate science offer many big-data case studies, and these are likely to be particularly interesting to study epistemological shifts due to big data. The reason for this is that classical models in environmental science are constructed based on equations that describe system behavior. Accordingly, model evaluation in environmental science also considers the degree to which models are coherent with background knowledge (for climate models, in-depth discussions of this point can be found in Baumberger, Knutti, and Hirsch Hadorn 2017; Knutti 2018; Carrier and Lenhard 2019). The evaluation of data-driven models, in contrast, cannot consider the coherence of models with background knowledge in the same way since the modeled relationships are not derived from theory (Pietsch 2015). Hence, comparing data-driven models with classical process-based environmental models can offer interesting insights about model construction, evaluation, and use.

Big-data case studies in environmental and climate science have tackled a number of scientific problems. They include the downscaling of results from numerical climate models with machine learning (Mearns et al. 2018) or machine learning predictions of extreme weather events (Gagne II et al. 2012; 2017; McGovern et al. 2014; 2017). Such machine learning applications are often developed in interdisciplinary teams in order to construct models that make ideal use of the available data and are at the same time physically plausible (see Faghmous and Kumar 2014; Gibert, Horsburgh, et al. 2018; McGovern et al. 2019; Reichstein et al. 2019; see also Karpatne et al. 2017). Big-data case studies in environmental and climate science have also used new forms of data, e.g., from social media or cell phone GPS signals. For example, different studies have used social-media data to assess the damages from natural disasters (e.g., Shelton et al. 2014; Kryvasheyeu et al. 2016) or GPS signals to detect climate adaptation (Lu et al. 2016). Consequently, there have been calls for the use of this kind of big data for researching and measuring climate adaptation (Ford et al. 2016) and natural disaster management (Yu, Yang, and Li 2018). Furthermore, crowdsourcing data has been used to measure environmental conditions, for example urban temperatures (Overeem et al. 2013; Elmore et al. 2014; Muller et al. 2015).

This thesis aims to explore epistemological aspects of data-driven modeling techniques in environmental and climate science in more detail. For this, it uses tools from philosophy of science to analyze the construction, evaluation, and use of data-driven models but also the use of big data more generally in environmental science. The focus of the

thesis is on climate research, ranging from physical climate science to environmental and social impacts of climate change. Such epistemological issues have, thus far, not been addressed specifically for these fields. This thesis aims to fill this gap in a way that is relevant for an interdisciplinary audience, particularly for environmental scientists, data scientists, and philosophers of science.

The remainder of this introductory chapter begins by outlining the objectives of this thesis in section 1.1. Then, section 1.2 introduces the methodological, conceptual, and philosophical background to this thesis. Three individual studies are presented in the form of self-contained papers in chapters 2, 3, and 4. Finally, in chapter 5, I first summarize the central findings of the thesis and then conclude by discussing the implications of these findings and by providing an outlook to open research questions.

1.1. Thesis' Context and Objectives

The advent of big data has led to many open scientific, technological, legal, ethical, and conceptual questions. This led the Swiss National Science Foundation to launch a National Research Program on big data (NRP75). The goal of this program is to obtain a better understanding of big data and the challenges and opportunities it brings to society from the point of view of various disciplines.[1] This thesis is part of a project funded under the NRP75 that investigates big data from a perspective that lies at the interface of philosophy of science and climate science. The ongoing trend of increasing generation and storage of data in environmental and climate science motivates three main objectives related to big data and data-driven modeling, which this thesis tackles in three individual studies presented in chapters 2, 3, and 4, respectively. The first of these three studies (chapter 2) looks at predictions in the context of big data. It aims to clarify what big data is and where in climate research it can most fruitfully be applied and for what reasons. A special focus of this study is on the conditions for successful predictions and the role of domain-specific background knowledge. The second and third study take a narrower perspective. They investigate data-driven modeling and try to assess important epistemological questions that have received attention for process-based models. The second study (chapter 3) looks more closely at the conclusions of chapter 2 and introduces a framework to assess the uncertainty of the predictions from data-driven models. The third study (chapter 4), finally, addresses the question of what makes a model fit for providing understanding of climate phenomena and then evaluates to what extent data-driven models are fit for this purpose.

[1] See the website of the NRP 75: http://www.nfp75.ch/en/the-nrp, accessed on January 16, 2020.

1.1.1. Discussing Big-Data Elements and Predictions in Climate Research

Original publication:
Benedikt Knüsel, Marius Zumwald, Christoph Baumberger, Gertrude Hirsch Hadorn, Erich M. Fischer, David N. Bresch, Reto Knutti. Applying Big Data Beyond Small Problems in Climate Research. *Nature Climate Change* 9, no. 3 (2019): 196-202.

The term "big data" lacks a clear definition (see section 1.2.2). While some clear-cut cases of big data exist, for example commercial applications like recommender systems (see Mayer-Schönberger and Cukier 2013), it remains unclear what big-data elements can be used in scientific research and what problems they can help to address. Approaches that are related to big data and data-intensive science have already been applied in various contexts in climate research. However, their relation to "classical" domain science and big data has not been discussed. Chapter 2 has several objectives. It aims to highlight which big-data elements are used for which kind of questions in climate research and why. It further aims to clarify the conditions under which big-data elements can be used for successful predictions. For this, we develop a framework to categorize case studies based on three dimensions, namely the measurements, the datasets, and the models. We use this framework to show what different elements of big data are used in climate research. Based on the application of this framework, we draw conclusions regarding the conditions for predictions based on big-data elements and the problems with the biggest potential for the application of big-data elements.

1.1.2. Developing a Framework for Characterizing the Predictive Uncertainty of Data-Driven Models

Original publication:
Benedikt Knüsel, Christoph Baumberger, Marius Zumwald, David N. Bresch, Reto Knutti. Assessment of Predictive Uncertainty of Data-Driven Environmental Models. Submitted for publication to *Environmental Modelling & Software*.

The topic of uncertainty has been widely discussed for model-based decision support (see section 1.2.5). All of these established tools for uncertainty classification focus, among other dimensions, on specific locations of uncertainty like the model structure and model parameters. These dimensions cannot be readily applied to understand predictive uncertainties of data-driven models (see chapter 3 for a detailed discussion on this point). Chapter 3 aims to develop a framework to characterize the predictive uncertainty of data-driven models. For this, we first highlight why existing frameworks are not readily applicable to data-driven models. We then develop a framework for assessing and characterizing the uncertainty of predictions from data-driven models. We illustrate the application of the framework using a case study from environmental

science. Furthermore, we outline some implications of the uncertainty assessment with this framework for scientific decision-support.

1.1.3. Assessing the Fitness of Data-Driven Models for Providing Understanding of Climate Phenomena

Original publication:
Benedikt Knüsel and Christoph Baumberger. Understanding Climate Phenomena with Data-Driven Models. Submitted for publication to *Studies in History and Philosophy of Science*.

Big-data proponents have argued that in the age of big data, the *what* may be more important than the *why*, and that hence, understanding of phenomena is only of minor interest (see, e.g., Mayer-Schönberger and Cukier 2013). However, this seems unlikely to be the case in science since understanding is seen as an important epistemic aim of science (Dellsén 2016; de Regt 2017, see section 1.2.6). However, whether data-driven models can be useful tools for understanding or whether they are merely tools for predictions is unclear. Philosophers and scientists have questioned the achievability of understanding with machine learning, specifically because of the lack of interpretability of machine learning algorithms (Ratti and López-Rubio 2018; López-Rubio and Ratti 2019), but also because it can be difficult to link models built with machine learning to a phenomenon of interest (Sullivan 2019). Chapter 4 aims to clarify the fitness of data-driven models to serve as vehicles for understanding phenomena in the climate system. We address this issue by developing a framework to evaluate models in terms of their fitness for providing understanding. We then apply the framework to both process-based and data-driven climate models. Based on this, chapter 4 discusses the extent to which data-driven models can provide scientists with understanding of phenomena in the climate system, and how this compares to process-based models.

1.2. Methodological, Conceptual, and Philosophical Background

In this section, I introduce and discuss methodological, conceptual, and philosophical topics relevant for this thesis. In section 1.2.1, I provide a short introduction to argument analysis as this method will be important in chapter 3 and, to some extent in chapters 2 and 4. In the more general sections 1.2.2 and 1.2.3, I introduce conceptual and philosophical issues concerning big data and data-intensive science and relevant topics from the philosophy of climate science, respectively. Sections 1.2.4, 1.2.5, and 1.2.6 are more specific. They each provide the background to the topic of one of the three studies presented in this thesis, namely predictions, uncertainty, and scientific understanding, respectively.

1.2.1. Argument Analysis

In chapter 3, we present a framework for assessing the uncertainty of predictions made with data-driven models. This framework is based on an analysis of arguments that can be made to support the assumption that a given model is fit-for-purpose. Argument analysis is also important in chapters 2 and 4 although there, we engage with the method less explicitly. That an analysis of arguments is important for assessing the adequacy of models for certain purposes has, for example, been argued by Baumberger, Knutti, and Hirsch Hadorn (2017). In this section, I provide a short introduction to the method of argument analysis and focus on those aspects that will be relevant in later chapters. This section is based on Brun and Betz (2016) and Brun and Hirsch Hadorn (2014).

The first step in argument analysis consists in reconstructing the arguments. This means that individual arguments in favor or against a proposition need to be identified and stated as inferences. By doing so, it becomes clear which propositions are to be justified and which propositions perform this justification. The former ones are called "conclusions" and the latter ones are called "premises". If and only if the relation between the premises and the conclusion is a correct inference, the premises justify the conclusion. Then, the relationship between the different individual arguments of the complex argumentation is reconstructed. This highlights whether the conclusion of one argument functions as a premise in a different argument, or whether the conclusion of one argument attacks or supports the premises of a different argument. It is also possible that several arguments that are based on different premises all support the same conclusion. This complex argumentation can for example be presented in an argument map that gives an overview of how the individual arguments are related to each other.

The second step of argument analysis consists in evaluating the arguments. In the evaluation step of argument analysis, the goal is to assess the quality of an argumentation. For this, the truth and acceptability of individual premises has to be evaluated. Furthermore, the validity or strength of individual arguments needs to be assessed, as well as the contribution of individual arguments to the overall argumentation. For determining the validity or strength of individual arguments, one needs to assess the relationship between the premises and the conclusion of an argument and identify whether the argument is deductively valid or non-deductively correct. An argument is deductively valid if the following condition holds: if all the premises are true, the conclusion must be true, too. A climate-related example of a deductively valid argument is the following (this example is taken from Baumberger, Knutti, and Hirsch Hadorn 2017, 7):

P1 *If a model is adequate for projecting X for the far future, then the model reliably indicates X for past and present.*

P2 *Model M does not reliably indicate X for past and present.*

C *Hence, M is not adequate for projecting X for the far future.*

In contrast to deductively valid arguments, non-deductively correct arguments are risky in the sense that even if all the premises are true, the conclusion need not necessarily be true, too. Instead, the evaluation focuses on whether the premises provide sufficiently good (but not conclusive) reasons for the truth of the conclusion. Hence, the focus of the evaluation of non-deductive arguments is not on the validity of the argument, but on the strength of the argument. In this thesis, inductive arguments are the most important form of non-deductive arguments. Other forms of non-deductive arguments include analogies or inferences to the best explanation. An example of an inductive argument is the following (this example is also taken from Baumberger, Knutti, and Hirsch Hadorn 2017, 7):

P1	*Model M reliably indicates X and climate quantities upon which X depends for past and present.*
C	*So probably, M is adequate for projecting X for the near future.*

If an argument is deductively valid and has true premises, that argument is referred to as "sound". A non-deductively correct argument whose premises are true is referred to as "cogent" (Baumberger, Knutti, and Hirsch Hadorn 2017).

1.2.2. Big Data and Data-Driven Modeling

The volumes and complexity of data that society produces and stores have been increasing dramatically over recent decades. According to an IBM blog post (Jacobson 2013), 2.5 exabytes of data ($2.5 \cdot 10^{18}$ bytes) were created every day in 2013. Technological progress has created opportunities to handle, store, and analyze these massive datasets efficiently and to draw useful inferences from them (National Research Council 2013). In this section, I introduce conceptual and epistemological issues regarding big data and then regarding machine learning and data-driven modeling.

Big Data and Data-Intensive Science

The term "big data" is often used to describe massive datasets with fine-grained information on individuals and new technological tools to analyze these datasets (Durán 2018, chap. 6.2). However, there is no generally accepted definition of "big data". Most attempts at defining the term have focused on characteristics of datasets. Often, big data is characterized by a number of V-words, most notably the volume, velocity, and variety of data (see Kitchin and McArdle 2016). The high volume refers to the size of the datasets, the velocity refers to the pace with which new data is created, and the variety refers to the diversity of data types and structures (Laney 2001). Some authors have added further characteristics besides these three Vs such as the veracity of the data (Lukoianova and Rubin 2014) or value, exhaustivity, relationality, and others (see Kitchin and McArdle 2016). However, while all of these characteristics of big data have intuitive appeal, focusing on such characteristics is unlikely to yield a satisfactory definition. Namely, as Floridi (2012) has argued, these characteristics are relative concepts.

If a definition is based on such relative concepts, what counts as big data today need not be considered big data in the future. This concern certainly holds for each of the three V-words listed above (Pietsch 2016). More importantly, focusing on the data only might yield an incomplete picture. The change brought about by big data lies not only in greater data availability, but also in how the datasets are made use of with analytic methods like machine learning (see Veltri 2017).

Although a clear-cut definition of the term "big data" is lacking, most authors seem to agree on certain characteristics of big data. One of these characteristics is the size and complexity of the datasets (see, Floridi 2012; Mayer-Schönberger and Cukier 2013; Kitchin 2014; De Mauro, Greco, and Grimaldi 2016; Pietsch 2016; Holmes 2017; Northcott 2019). Another characteristic that has often been referred to is the use of machine learning and data mining as techniques to analyze these large volumes of data (see, Boyd and Crawford 2012; Kitchin 2014; Pietsch 2015; 2016; Holmes 2017; Veltri 2017; Northcott 2019). In a recent book on big data in the series *Very Short Introductions* by Oxford University Press, Holmes (2017) illustrates these two aspects nicely. She explains, first, how the size and the complexity of datasets inhibit the storage in traditional, relational databases, and what technical challenges one faces when handling such massive datasets. Then she goes on to illustrate how big data analytics based on machine learning works to extract value from these datasets. More work is needed to clarify the term "big data" (and potentially, related terms like "data-intensive science" and "data science"). While this thesis does not explicitly contribute to clarifying the term, chapter 2 provides an overview of scientific studies that use certain big-data elements and provides a framework to classify the studies. By doing so, the study highlights the range of activities that fall under the term "big data" in some sense.

The ubiquity of increasing volumes of data has led to a veritable hype around big data, including a widespread belief that increasing volumes of data can automatically contribute to objective insights about phenomena that were previously not accessible (Boyd and Crawford 2012). This concerns not only commercial big-data applications, but also scientific research due to the increasing volumes of data available to scientists.[2] In fact, various disciplines experience rapidly increasing volumes of data, ranging from biology (see Callebaut 2012; Canali 2016) to particle physics (see Radovic et al. 2018), to Earth science (see Reichstein et al. 2019), to the social sciences and humanities (Kitchin

[2] This belief was perhaps most strongly expressed in a now-infamous editorial by then-editor-in-chief of Wired magazine, Chris Anderson, which appeared in 2008. Anderson (2008) argued that the ever increasing volumes of available data allow us to predict essentially every aspect of interest to us. As Anderson claims, "[w]ith enough data, the numbers speak for themselves". This would, as the provocative title of Andersons's editorial claimed, result in "the end of theory" and "make the scientific method obsolete". Scholars from a range of disciplines were quick to highlight the flaws in Anderson's argument, and, as Google Director of Research Peter Norvig wrote in his blog (see https://norvig.com/fact-check.html, accessed on November 18, 2019) at the time, Anderson never really endorsed the position he described in his editorial. Rather, he was aiming to provoke a conversation.

2014). Pietsch (2016) has argued that in scientific contexts, the term "big data" merely describes that science has more data available. Hence, he has suggested to focus on the term "data-intensive science" instead. According to Leonelli (2012, 1), the two characteristic features of data-intensive science (or "data-driven science", as she calls it) are the central role of inductive inference from data and the importance of automation, as machines are used to automatically extract useful information from data.[3] The methodological techniques used in data-intensive science can be divided into methods for data acquisition, data storage, and data analysis (Pietsch 2016).

The availability of massive volumes of data brings about interesting epistemological issues with respect to all three methodological aspects, i.e., data acquisition, data storage, and data analysis. Epistemological issues regarding data acquisition arise due to the possibility for scientists to increasingly collect data in novel ways. For example, data gathered in citizen science and crowdsourcing projects (see, for example, Overeem et al. 2013; Elmore et al. 2014; Muller et al. 2015) could be met with concerns regarding data quality. Furthermore, the participants in a citizen science project could potentially influence the data collection based on non-epistemic motives (Elliott and Rosenberg 2019). Further issues with respect to data acquisition concern the modeling activities necessary to obtain an adequate dataset (Bokulich 2018; Leonelli 2019a) and related to that, whether data obtained from computer simulations should be considered on a par with observation-based data (Lusk 2016; Parker 2016; 2017). Epistemological issues regarding data storage could arise for example because increasing volumes of data stored on distributed servers could impact the portability of the data, which is essential for the data to serve as prospective evidence for or against scientific claims (Leonelli 2015). This might be exacerbated by data produced and owned by private companies and institutions (see Leonelli 2019b). Furthermore, scientists may have to be increasingly selective about which parts of the massive streams of data they should retain as the volumes of data can become too large to be stored. Finally, epistemological issues regarding data analysis concern, for example, the sense in which models built with machine learning can be considered theory-free (Callebaut 2012; Lyon 2015; Pietsch 2015) and the importance of interpretability and transparency of algorithms for different purposes (Krishnan 2019; Creel, forthcoming).

Machine Learning and Data-Driven Modeling

The focus of this thesis lies on issues that relate to data analysis and data-driven modeling, i.e., the third class of methodological techniques mentioned above. Throughout this thesis, a model of a phenomenon built with machine learning is called a "data-driven

[3] Pietsch (2015) agrees with these two aspects but specifies that data-intensive science is particularly concerned with a specific kind of inductive reasoning, namely eliminative induction. This is related to the method of difference proposed by John Stuart Mill. For an explanation of eliminative induction, see Pietsch (2015, 909–11).

model".[4] Machine learning is a set of methods at the interface of computer science and statistics, with which useful information can be extracted from a dataset. In this thesis, the focus lies on the application of supervised regression methods for the modeling of environmental phenomena. Supervised methods are ones in which a dependent variable is provided, which can be either a categorical variable in classification tasks or a continuous variable in regression tasks. Supervised machine learning algorithms extract information from a dataset to predict the dependent variable based on the independent variables. In contrast to supervised machine learning, unsupervised methods aim to extract useful patterns in a dataset without being provided a dependent variable (Hastie, Tibshirani, and Friedman 2008, chap. 1; James et al. 2013, chap. 2). Unsupervised machine learning will not be further discussed in this thesis.

A range of different supervised machine learning algorithms exist, ranging from simple linear regression and regularized techniques that perform a variable selection before fitting a linear model, to non-linear methods including neural networks and ensemble approaches like random forest (James et al. 2013, chap. 2). In recent years, deep learning, i.e., neural networks with multiple hidden layers of neurons (see Buckner 2019 for a philosophical introduction to deep learning), have become increasingly popular, including in environmental science (Reichstein et al. 2019). While more complex methods are very flexible and can learn highly non-linear behavior from the data, this increase in flexibility generally comes at the expense of model interpretability. This is because in contrast to simpler linear approaches, complex methods do not provide an equation that could be studied to learn about the behavior of the model (James et al. 2013, chap. 2).

When constructing a data-driven model, a modeler usually splits the available dataset into a training dataset, to which a model is fit, and a test or validation dataset, which is used to select a best-fitting model or to estimate the error rate of the model (James et al. 2013, chaps. 2, 5). Because the machine learning algorithm learns the relationships between the variables directly from the training dataset, the modeler need not know these relationships in advance. Hence, data-driven models can be considered theory-free in the sense that the relationships between the variables are not derived from some background theory (Lyon 2015; Pietsch and Wernecke 2017). However, this does not make data-driven modeling a theory-free activity. Rather, as Pietsch (2015) has argued, these models are theory-laden in a different way than rule-based or process-based models. Because the relationships between the variables are not prescribed from theory, data-driven models are not internally theory-laden. However, data-driven models are theory-laden in an external sense because of the theory-guided problem framing, including the selection of input variables (Pietsch 2015; see also Hosni and Vulpiani 2018). Data-driven modeling is further theory-laden because of the theory-ladenness of observational

[4] A more detailed discussion of the term "data-driven model" is provided in chapter 4.2 of this thesis.

data and because of the interpretation of results in terms of the target system (Callebaut 2012).

Machine learning models can provide reliable predictions of phenomena. According to Pietsch (2016), this predictive success is rooted in the causal nature of data-driven modeling. As he argues, machine learning is able to identify the causal relevance of factors for a given phenomenon and hence, to extract the underlying causal structure of a phenomenon when some conditions are fulfilled. Namely, first, the terms on which the variables are based need to be well-defined; second, all relevant variables need to be included; third, data on all relevant configurations of the target system need to be available; and fourth, the background conditions need to be sufficiently constant. Northcott (2019) has suggested that additional conditions might be required, for example, he suggests that appropriate methods to handle the data need to be available.

The conditions developed by Pietsch and Northcott indicate under what conditions models built with machine learning can be adequate for predictive purposes. However, these conditions were specifically developed for disciplines without generalized theoretical laws (see Pietsch 2016, 160). In environmental and climate science, however, this kind of background knowledge is available for many processes and phenomena. It is an open question what the availability of this background knowledge means for successful predictions based on big-data elements. We address this question in chapter 2. Background knowledge can also be helpful in assessing the predictive uncertainty of data-driven models, which, to the best of my knowledge, has not been discussed in the literature, thus far (see section 1.2.5). This issue is addressed in chapter 3. Pietsch (2016, 168) has argued that due to the absence of general theoretical laws, models in data-intensive science cannot typically be used to derive phenomena from theoretical laws. Hence, data-driven models cannot provide explanations that achieve unification (Pietsch 2016, 168). Again, this is likely different in a field in which data-driven models can be constructed against a wealth of theoretical background knowledge. Hence, the role that data-driven models can play as vehicles for understanding in a field like climate science deserves attention. This issue is addressed in chapter 4.

1.2.3. Philosophy of Climate Science

The Summary for Policy-Makers of the Fifth Assessment Report by Working Group 1 of the Intergovernmental Panel on Climate Change (IPCC 2013, 4) states that "[w]arming of the climate system is unequivocal, and since the 1950s, many of the observed changes are unprecedented over decades to millennia." The dominant factor for this observed warming trend in global temperatures since the mid-20th century has been human influence (IPCC 2013). These changes will continue as long as humans keep emitting greenhouse gases to the atmosphere. Hence, "[l]imiting climate change will require substantial and sustained reductions of greenhouse gas emissions" (IPCC 2013, 19). In 2015, most countries of the world agreed to limit global warming to well below 2°C above preindustrial levels in the Paris Climate Agreement (United Nations 2015).

However, fast, substantial, and sustained reductions of greenhouse gas emissions are required if these goals are to be met. And even if the goals are met, societies will have to respond to various impacts of climate change with adaptation measures as the world has already warmed considerably (IPCC 2018).

Quotes like the ones presented in the previous paragraph show that climate science is a highly policy-relevant discipline. However, confidently basing policy decisions on scientific claims like these requires a good understanding of their epistemic foundations and reliability. At the same time, climate science relies on large computer models to simulate the behavior of the Earth's climate system. These computer-intensive methods raise many interesting epistemological questions. Both because of the policy-relevance and the complex methodological approaches of the discipline, philosophers have taken increasing interest in climate science (Winsberg 2018b, chap. 1). In recent years, the philosophy of climate science has become an active subdiscipline of philosophy of science. For example, over the year 2018, this subdiscipline has produced an anthology (Lloyd and Winsberg 2018), a monograph (Winsberg 2018b), and an entry in the Stanford Encyclopedia of Philosophy (Parker 2018).

The philosophy of climate science has tackled a multitude of issues. By far the most philosophical attention has been directed at climate models and issues relating to them (for an overview, see Frigg, Thompson, and Werndl 2015b). However, philosophers have addressed many other questions, too, regarding for example the theoretical foundations of climate science, such as how to define key concepts like "climate" and how to draw the system boundaries of the climate system (Werndl 2016; Katzav and Parker 2018). Climate data have attracted philosophical interest, too. Many forms of climate data are obtained through extensive modeling and processing. As these activities can introduce errors and uncertainty, disagreements between climate models and climate data have to be inspected carefully since the disagreements need not necessarily reflect a problem with the models (Winsberg 2018b, chap. 2; for a case study investigating such discrepancies, see Lloyd 2012). Finally, philosophers have also addressed questions related to social epistemology, for example, whether dissent in climate science, especially in the context of climate denialism, is epistemically detrimental (Biddle and Leuschner 2015), and concerning the value of the scientific consensus regarding the anthropogenic origin of global warming (Oreskes 2018).

A full review of the topics of philosophy of climate science is beyond the scope of this introduction. Interested readers are referred to the good, comprehensive introductions to these topics provided elsewhere (Frigg, Thompson, and Werndl 2015a; 2015b; Parker 2018; Winsberg 2018b). The remainder of this section will introduce topics from the philosophical literature on climate models that are relevant for the purpose of this thesis.

Climate Models and Computer Simulation

Different things can serve as scientific models, for example concrete objects or sets of differential equations (see Frigg and Hartmann 2012; Weisberg 2013). Many scientific models represent some real-world target in an idealized way (Contessa 2011; Frigg and

Hartmann 2012). Idealizations are characterized as either Aristotelian or Galilean. An idealization is Aristotelian if the act of idealizing consists in deliberately abstracting from real-world properties and focusing on those aspects of the real world that are deemed to be most important. An example from classical physics is when a real-world object is represented as having only a mass and a shape and all other properties are ignored. An idealization is Galilean if an aspect of the real world is presented in a deliberately distorted manner in the model for the purpose of tractability. An example from classical mechanics is to represent the movement of a real-world object as the movement of a point mass on a frictionless plane. These two types of idealizations are not mutually exclusive, and many models contain both (see Frigg and Hartmann 2012, where these and further examples of the two kinds of idealizations are presented). An important question with respect to representation is what standards exist to assess the representational accuracy of a model (Frigg and Nguyen 2016). In chapters 3 and 4, this question will be addressed specifically for data-driven models of phenomena in the climate system.

Due to their complexity, some models cannot be solved by an unaided human agent in reasonable time. This is for example the case for models in meteorology and climate science, where the models rely on coupled differential equations that describe processes such as the dynamics of the atmosphere. Due to the complexity and spatiotemporal nature of these equations, approximating the solution of the equations by hand would be prohibitively expensive. Instead, the equations are implemented on a computer that numerically approximates a solution (see Parker 2014). The program that performs this approximation to the solution of a mathematical model is called a "computer simulation" (at least in a narrow definition of the term "computer simulation", see Winsberg 2019). While some view the use of computers to approximate the solution of model equations as a practical nuisance (Frigg and Reiss 2009; Knutti 2018), others argue that it has important epistemological consequences. For example, Humphreys (2004; 2009) argues that computer simulations suffer from epistemic opacity because it is not possible to have knowledge of all epistemically relevant details of computer simulations. Issues concerning the opacity of simulations and the transparency of computational systems more generally will be addressed in section 1.2.6 on scientific understanding.

State-of-the art climate models are among the largest simulation models in use today (Parker 2018). At the core of these models is a set of coupled differential equations. These equations represent, for example, the flow of air masses through the atmosphere and heat-transfer processes. The equations are discretized on a three-dimensional grid. Certain important processes take place on scales smaller than the grid size, for example cloud formation. To account for the effect of these processes, so-called "parameterizations" are implemented in climate models (Winsberg 2018b, chap. 4). Parameterizations also allow to include processes like vegetation for which no equations are available (Knutti 2018). State-of-the-art climate models consist of different sub-models, representing for example the atmosphere, the ocean, or land surfaces (Winsberg 2018b, chap. 4). Depending on which sub-models they contain, climate models are referred to as

general circulation models (GCMs) or Earth system models (ESMs). While GCMs resolve the general circulation of the Earth's atmosphere and ocean, ESMs resolve additional processes, for example chemical and biogeochemical ones. These models are run on powerful supercomputers (Parker 2018). Due to their complexity, today's climate models have been developed over a long time by adding new submodules. This path-dependency of climate model development makes it difficult to clearly identify which parts or sub-models of a climate model are responsible for the success or failure of a given simulation (Lenhard and Winsberg 2010).

Climate Model Evaluation

In order to establish confidence in the results of climate models, the models need to be evaluated. An important question regarding climate model evaluation is what the target of confirmation should be. Lloyd (2009) has argued that many instances of fit between climate model results and empirical data as well as robustness considerations confirm the models. This is in line with an influential account of model evaluation by Oreskes et al. (1994). Oreskes et al. have argued that scientific models cannot be verified in the sense of establishing their truth or the truth of their results since verification is only possible in a closed system. Nor can models be conclusively validated (Oreskes et al. 1994). Rather, instances of fit between model results and observational data should be understood as confirming, but not verifying or conclusively validating, the model. However, Parker (2009) has argued that the hypotheses about the climate system's inner workings embedded in climate models are not plausible candidates for truth. The reason for this is that models contain hypotheses that are known to be false from the outset, and hence, they should not be seen as the target of confirmation. Rather, Parker argues, what should be confirmed is the hypothesis that a given model is adequate for some specific purpose.

This change of perspective from confirming the models to confirming hypotheses about the model's adequacy for a purpose has important implications because it can be difficult to assess what exactly makes a climate model adequate for a given purpose (Parker 2009). Baumberger, Knutti, and Hirsch Hadorn (2017) have developed a framework to assess the adequacy of climate models for long-term projections (i.e., predictions of future climate depending on socioeconomic boundary conditions, see section 1.2.4). According to their framework, models should be evaluated not only based on their empirical accuracy, but also based on the robustness of model results, i.e., the agreement between different models or model versions. Furthermore, their framework stresses the importance of the coherence of the models with background knowledge, which should also be considered when evaluating their adequacy for long-term projections. A similar point has been made by Knutti (2018), who has emphasized the importance of qualitative process understanding to establish the adequacy of climate models for long-term projections. Examples of how to evaluate models in terms of their coherence with background knowledge are provided by Carrier and Lenhard (2019). The three aspects (i.e., empirical accuracy, robustness, and coherence with background knowledge) emphasized by Baumberger, Knutti, and Hirsch Hadorn (2017) are specifically relevant

because they are all related to the representational accuracy of a model, i.e., to how accurately a model represents processes that are important for some purpose (e.g., long-term projections of temperature values). However, aspects other than representational accuracy may be relevant for the adequacy of a model for a given purpose. As Parker (forthcoming) argues in a general account of model evaluation, practical considerations such as sufficient computational power can be relevant, too. The topic of adequacy-for-purpose will return in chapters 3 and 4, where the issues of predictive uncertainty and scientific understanding are addressed, respectively. Note that in these chapters, the term "fitness-for-purpose" will be used rather than "adequacy-for-purpose" since fitness-for-purpose admits of degrees (Parker, forthcoming).

As has been argued above, the empirical accuracy and the robustness of climate model results are important aspects for evaluating the representational accuracy of climate models. In other words, climate model evaluation considers how well model results agree with observational data (empirical accuracy) and how well the results from different climate models or model versions agree with each other (robustness) (see Baumberger, Knutti, and Hirsch Hadorn 2017). However, neither empirical accuracy nor robustness provide straightforward support of the representational accuracy of climate models in every instance. As both of these considerations are relevant for the evaluation of fitness-for-purpose in chapters 3 and 4, I provide a brief overview of the concerns surrounding the two criteria in the next two paragraphs.

Empirical accuracy can provide support of the representational accuracy of a model only to the extent that the agreement between models and data occurs for the right reasons. This is why the topic of model tuning has received attention in the philosophy of climate science. Parameter values related to parameterizations mentioned above are often not well constrained. Hence, their values are determined such that the model as a whole behaves in accordance with available observational data (Frigg, Thompson, and Werndl 2015b; Winsberg 2018b, chaps. 4, 10). Due to this process of model tuning (or calibration), the models are forced, to some extent, to reproduce observations. An intuitive position is that data that has been used to tune climate models should not also be used to evaluate the models as this would essentially be double-counting. Although the distinction between model tuning and model evaluation may not be so clear (Steele and Werndl 2013), most philosophers and scientists agree that independent data provides a better benchmark for model evaluation (Frisch 2015; Schmidt and Sherwood 2015; see also Lloyd 2010 who emphasizes the importance of results for which climate models cannot be tuned). Traditional model selection criteria, for example cross-validation, give importance to use-novel data for model evaluation, too, but they permit some double-counting (Steele and Werndl 2016). Thus, the intuitive view about double-counting should be expressed with more nuance (Steele and Werndl 2018). This debate about use-novel data for model evaluation is specifically important in chapter 4 of this thesis, where empirical accuracy is a criterion used to evaluate the representational accuracy of a model.

Like empirical accuracy, the robustness of climate model results is a criterion to assess the representational accuracy of the models (see Baumberger, Knutti, and Hirsch Hadorn 2017). As Lloyd (2009; 2010; 2015) has argued, if a number of models share a common causal core and their outputs show a similar behavior, this confirms the causal core of the models under some conditions. Namely, this confirmation depends on whether there is independent observational and experimental evidence for the shared causal core as well as for other model assumptions and for the behavior of the models. This is the kind of robustness consideration that is relevant for evaluating the representational accuracy of models in chapters 3 and 4. However, there is a worry that climate models might agree for the wrong reasons, specifically because of the interdependence of climate models (see Parker 2011). Baumberger, Knutti, and Hirsch Hadorn (2017) acknowledge this and consequently suggest that the independency and diversity of climate models should be increased to strengthen these robustness considerations.[5]

1.2.4. Predictions

Predictions are an important topic in chapter 2 of this thesis. In the philosophical literature, predictions have mostly been discussed in the context of the confirmation of theories and hypotheses (see M. Forster 2010). However, both in the context of climate science and big data, predictions are important goals of model construction that are not tied to the confirmation of theories or models, at least not exclusively. Climate models are often used to make predictions of the future climate conditional on some boundary conditions (i.e., a kind of what-if inference is performed). These conditional predictions are made, among other things, to convey to decision-makers what the future climatic conditions on Earth will be in response to a certain level of greenhouse gas emissions. The term "climate projection" is used to denote these conditional predictions (Knutti 2018).

In the context of big data, prediction is often seen as the main goal of modeling (Northcott 2019). As has been discussed above in section 1.2.2, Pietsch (2015) and Northcott (2019) have suggested that models in data-intensive science can extract the causal structure of a phenomenon and make reliable predictions about it if certain conditions are met. These conditions imply that more data and algorithms to analyze the data do not necessarily lead to better predictions. For example, the openness of the system, chaotic behavior, and non-stationarity can limit predictability (Northcott 2019). Pietsch (2016) has argued that the suggested conditions are specifically important for scientific applications for which no well-established theoretical laws are available. For many climate phenomena, theoretical background knowledge of this kind is available.

---

[5] Note that robustness considerations to confirm hypotheses indicated by model outputs are not discussed in detail in this thesis. The focus, here, solely lies on robustness considerations to confirm hypothesis about the shared causal core of the models. For overviews regarding robustness considerations in climate science, see Parker (2011), Lloyd (2015), and Winsberg (2018b, chaps. 11, 12).

Hence, the role of background knowledge for big-data predictions deserves investigation. This is one of the topics that we discuss in chapter 2. Specifically, we contrast "classical" big-data predictions with climate projections and identify conditions for successful predictions based on big-data elements. A focus of this analysis lies on the role of background knowledge for the justification of confidence in the predictions.

1.2.5. Uncertainty in Model-Based Decision Support

Uncertainty of model-based scientific inferences is the topic of chapter 3, where we develop and present a framework for assessing predictive uncertainties of data-driven models. Although uncertainty is a central concept in many scientific areas, it remains a poorly understood one as Frigg, Thompson, and Werndl (2015b) note with respect to climate models. In a general decision situation, there are various reasons for uncertainty, for example concerning the available actions from which to choose, the consequences of each course of action, the values that are attached to each potential outcome, and the rule with which to choose between the different courses of action (Hirsch Hadorn et al. 2015; Bradley and Drechsler 2014; Hansson and Hirsch Hadorn 2018). In such decision situations, it is common to differentiate between situations of risk and situations of uncertainty. According to this distinction, a decision under risk is one in which precise probabilities of different outcomes are known, whereas a decision under uncertainty is one in which information on outcomes is not available in the form of precise probabilities (Hansson 2007). As Hansson (2009) argues, hardly any real-world decision situation has probabilities that are known with precision, underlining why uncertainty is a central topic for many practical applications.

Conceptual work for uncertainty assessments has aimed to provide a full account of what kinds of uncertainty emerge in a generic decision situation and for what reasons. Recently, philosophers have suggested to take an "argumentative turn" in policy analysis and approach the topic of uncertainty and decision support with a rigorous analysis of arguments (Hansson and Hirsch Hadorn 2016; 2018). The framework for uncertainty assessment introduced in chapter 3 follows this approach. This argument-based approach to uncertainty analysis consists of strategies to handle several sorts of uncertainty (Hansson 2016; Hirsch Hadorn 2016), for example strategies to evaluate the framing of a problem (Grüne-Yanoff 2016), and strategies to characterize different kinds of uncertainty such as value uncertainty, e.g., uncertainty about which values are at stake (Möller 2016), and uncertainty regarding what the actual or potential states of the world are (Betz 2016a).

In this thesis, the focus will be on the last of these kinds of uncertainties, namely uncertainties concerning the actual and potential states of the world. This is specifically relevant in policy analysis because, as Betz (2016a, 138) has put it, an analysis of policy options requires a balancing act of basing decisions on "no more and no less than what one actually knows". In scientific policy advice, information of this kind is often derived from model-based inferences. Uncertainty is inherent to such model-based inferences

because models are idealized representations of their target systems, as has been discussed in section 1.2.3 above.

Philosophers, policy analysists, and scientists have suggested ways to characterize and categorize uncertainties of model-based inferences. Uncertainty is routinely characterized as either aleatory or epistemic. Aleatory uncertainty is due to the inherent variability of the target system, and epistemic uncertainty arises because of our imperfect knowledge about the target system (Walker et al. 2003). However, more refined characterizations are possible. For example, Walker et al. (2003) have suggested that uncertainty be characterized along three dimensions that can be arranged as an uncertainty matrix, namely the location of uncertainty (where in the research process does the uncertainty arise?), the nature of uncertainty (is the uncertainty aleatory or epistemic?), and the severity of uncertainty (how uncertain is the information on the spectrum from determinism to total ignorance?). This framework has been very influential and has been adapted and applied by many researchers (see Kwakkel, Walker, and Marchau 2010). This uncertainty matrix aims to make explicit what kinds of uncertainties there are and why they emerge before any attempt is undertaken to quantify them. By carefully considering all sources and kinds of uncertainties, it can be avoided that uncertainty quantification results in overconfident estimates that disregard important sources of uncertainty that are difficult to quantify (Walker et al. 2003; Bradley and Drechsler 2014).

Predictive uncertainties have also been a topic in the philosophy of climate science (for an overview of characterizations of uncertainties in climate science, see Frigg, Thompson, and Werndl 2015b). Different locations of uncertainty (in the parlance of Walker et al. 2003) of climate model projections have been identified, namely, the model structure, the numerical approximation of the model equations, the choice of parameter values, the internal variability of the climate system, the available observations, and natural and socioeconomic boundary conditions (Knutti 2018; Winsberg 2018b, chap. 7). Uncertainty from the first three sources is often referred to by climate scientist as "model uncertainty" (see, e.g., Hawkins and Sutton 2009). However, it is important to recognize that climate models themselves are strictly speaking not uncertain, but the relationship between the models and the real climate system is. In other words, what is commonly referred to as "model uncertainty" should be understood as "representational uncertainty" (Parker 2010a; Knutti 2018).

As the process of model construction is a non-unique one (Frigg and Hartmann 2012), this representational uncertainty can be estimated by considering different models of the same target system (see Parker 2010a for a discussion of this in the context of weather predictions and climate projections). Such ensemble methods for uncertainty quantification are routinely used in climate science. Specifically, climate scientists can run models with different initial conditions or employ an ensemble of models that have a different model structure or different parameter values. The spread of the model results can then be used as a basis to quantify uncertainty (Parker 2010b). However, the spread

of a climate model ensemble is recognized to be smaller than the actual uncertainty because the available models do not sample from the entire space of possible model structures or parameter values. This is partly due to model interdependence, which was briefly mentioned in section 1.2.3 when discussing robustness considerations in climate science. Hence, the spread of model ensembles cannot be readily turned into a probability density function (Parker 2010b). Remedies to this issue can be to weight climate models according to their performance and independence (Knutti, Baumberger, and Hirsch Hadorn 2019) or to use structured expert elicitation (Thompson, Frigg, and Helgeson 2016).

In the IPCC reports, uncertainty about specific scientific statements is expressed using a two-dimensional terminology. Namely, the IPCC expresses uncertainty using both a quantitative likelihood and a qualitative confidence metric (Mastrandrea et al. 2010). This practice has attracted criticism due to the inconsistent use by IPCC authors and because the relationship between the two metrics is not clear (Adler and Hirsch Hadorn 2014; Wüthrich 2017). Nevertheless, Winsberg (2018a) has argued that it is a good practice to express uncertainty using two metrics. These two metrics should be understood, Winsberg argues, as first-order uncertainty statements that provide imprecise probabilities and second-order qualifications of the first-order uncertainty estimate depending on the quality and consistency of evidence. Winsberg argues that the two metrics allow scientists to deliver information that is useful for policy-makers by communicating either narrower uncertainty bands with larger second-order uncertainty or vice versa. Which of these approaches the scientists choose is partly based on value judgments (see also Winsberg 2018b, chap. 9).

Due to the increasing use of data-driven modeling techniques, uncertainties of inferences from data-driven models are an important topic. However, to the best of my knowledge, no accounts exist that offer suggestions on how to understand and characterize the predictive uncertainty of data-driven models. Chapter 3 of this thesis offers a discussion of this topic and highlights why existing frameworks that characterize uncertainty from process-based models in terms of different locations, as outlined above, cannot be readily applied to data-driven models. It then discusses how representational uncertainty should be assessed for data-driven models using argument-based tools. The framework considers first-order and second-order uncertainty, similarly to the uncertainty estimates provided by the IPCC (as discussed in the previous paragraph).

1.2.6. Scientific Understanding

In chapter 4 of this thesis, we address the topic of scientific understanding and specifically the question whether data-driven models can contribute to scientific understanding of phenomena in the climate system. Understanding is considered an important epistemic aim of science (Dellsén 2016; de Regt 2017). Different types of understanding are distinguished. Often, one distinguishes between interrogative understanding (understanding why something is the case or understanding how something came about),

objectual understanding (understanding a system or a subject matter), and propositional understanding (understanding that something is the case) (see Baumberger 2014). Understanding-why, one type of interrogative understanding, is also described as "understanding a phenomenon" and is tied to having an explanation of that phenomenon (de Regt 2009; 2017), although some have suggested that it is possible to understand phenomena without having an explanation of them (see Lipton 2009; Gijsbers 2013).

Understanding a real-world phenomenon requires some theory or model that serves as vehicle for understanding. There are different positions on what qualities make a theory or model adequate for serving as a vehicle for understanding. De Regt (2009; 2015; 2017) has argued that an epistemic agent needs to be able to use a theory in order for that theory to provide understanding of a phenomenon. In his words, the theory needs to be intelligible. Intelligibility of a theory is relevant for understanding because epistemic agents need to be able to apply the theory to a concrete phenomenon in order explain that phenomenon, and it is generally not obvious how a general theory can be applied to a phenomenon.[6] De Regt (2009) suggests that a theory is intelligible if a scientist can qualitatively anticipate the consequences of that theory without performing any calculations.

Other authors took issue with the focus on the relationship between the scientist and the theory or model that is to serve as vehicle for understanding. Instead, they argued that the important relationship to be considered is that between the theory and the target. E.g., according to Strevens (2013), understanding consists in grasping a true explanation of a phenomenon. Hence, understanding requires an explanation that is true of the world. Similarly, Khalifa (2012) holds that even the most promising accounts of scientific understanding can be reduced to insights gained in the literature on scientific explanation.

Wilkenfeld (2017) has taken a middle position in this debate and argued that both the relationship between a model or theory and the target system on the one hand, and the relationship between the model or theory and the epistemic agent on the other hand are important for an account of understanding. Thus, understanding according to Wilkenfeld should be taken to be a multidimensional concept. According to this account, intelligibility and representational accuracy are good-making features of understanding, meaning that both factors determine what degree of understanding can be obtained from a model or theory. A similar approach was taken by Baumberger (2019), who has provided an explication of objectual understanding which considers three dimensions. Namely, understanding, according to his account, depends on the extent to which an agent grasps a theory or model (which is related to intelligibility), the extent to which

---

[6] De Regt (2009) discusses two ways of applying a theory to a phenomenon. One is to construct an explanation according to the deductive-nomological model. A different approach is to construct a model for a model-based explanation. Both of these approaches, according to de Regt, require that the epistemic agent can use the theory.

the theory or model answers to the facts (which is related to representational accuracy), and the degree to which the agent's commitment to the theory or model is justified.[7]

It has not only been discussed under what conditions a model can give scientists understanding of phenomena, but also what kind of understanding they can obtain with different models. What kind of understanding a model can provide an agent with depends on the relationship between the model and its target. Even highly idealized models can be used to obtain how-possibly understanding of a phenomenon, meaning that they can show possible causal mechanisms that should be considered as hypotheses to explain actual instances of the phenomenon (Ylikoski and Aydinonat 2014). This is possible even if the models are largely autonomous from a scientific theory (Reutlinger, Hangleiter, and Hartmann 2018). Reutlinger, Hangleiter, and Hartmann (2018) have argued that highly idealized models can also provide scientists with how-actually understanding of phenomena, i.e., with an actual explanation for a specific instance of a phenomenon. For this, the models need to be embedded in a theoretical framework. Parker (2014) has discussed the distinction between how-possibly and how-actually understanding for climate models. Considering hypothesis tests, she has argued that climate models can provide how-possibly understanding if the relationships between the candidate causal factors considered in the model are represented accurately. Climate models can also give how-actually understanding of a phenomenon if all candidate causal factors are considered in the model.

As discussed above, according to many accounts of scientific understanding (e.g., de Regt 2009; Ylikoski 2014; Baumberger 2019), the degree to which a model or theory is intelligible to an agent is relevant to determine to what degree the agent can understand a phenomenon through that model or theory. As I have mentioned in section 1.2.3, many scientific models including climate models have to be implemented and run on computers as simulation models due to their complexity and the spatiotemporal resolution. However, computer simulations have inherent features that are relevant for a discussion of model intelligibility. Namely, as Humphreys (2004; 2009) has argued, computer simulations are epistemically opaque. The epistemic opacity of computer simulations arises because it is not generally possible for a human agent to know all the epistemically relevant aspects of a simulation. This is due to the many steps involved in running a simulation, but also for example due to emerging patterns that could not be anticipated from microlevel behavior in agent-based simulations. Furthermore, not only characteristics of the simulation run itself seem relevant for the epistemic opacity of simulation models, but also characteristics of the underlying model such as its complexity and its modularity (Beisbart, in preparation; Lenhard and Winsberg 2010). Model opacity may

---

[7] Note that whereas de Regt (2009) holds that theories need to be intelligible in order to construct a model from the theory, according to Baumberger's (2019) account, both theories and models can serve as vehicles for understanding and hence, need to be graspable. Note further that Baumberger also argues that the agent's commitment to the model or theory is a necessary condition for understanding. This point is not important for the present purposes.

reduce the extent to which a model is intelligible, even if it does not make the model entirely unintelligible.

State-of-the-art climate models are certainly affected by a lack of intelligibility because of their complexity and modularity. As has been mentioned in section 1.2.3, climate models are not routinely built from scratch. Rather, existing models are used for a variety of purposes and are extended with new submodules over time. This makes it difficult to assess which parts of a model are responsible for the success or failure of a climate model (Lenhard and Winsberg 2010). Consequently, in an essay entitled "The Gap Between Simulation and Understanding in Climate Modeling", climate scientist Isaac Held (2005) has claimed that the tendency to make climate models increasingly more complex by adding processes and increasing their resolution has made it more difficult to understand the models, and accordingly, phenomena in the climate system through the models. Nevertheless, Parker (2014) has argued that climate simulations can contribute to scientific understanding. They can do so by allowing climate scientists to actually use available theories that would be too complex to an unaided human, by providing surrogate observational data, by allowing to run hypothesis tests with counterfactuals, and by allowing to explore model hierarchies (a point that has also been suggested by Held 2005). All of these activities, Parker (2014) argues, provide scientists with explanatory information which contributes to scientific understanding. In chapter 4, we explicitly address how the difficulty in grasping climate models affects their fitness to serve as vehicles for understanding.

The worry about model intelligibility seems even more acute for data-driven models. Machine learning algorithms, especially complex ones like deep learning, are notoriously difficult to interpret (Krishnan 2019; Creel, forthcoming). This has led to some skepticism about the role that data-driven models can play for obtaining explanations, at least if these explanations are to be mechanistic. Namely, Ratti and López-Rubio (2018) and López-Rubio and Ratti (2019) have argued that modeling complex phenomena with machine learning requires that the data-driven models take a complex form, too. However, this increase in complexity leads to a decrease in model intelligibility. Hence, obtaining explanations from machine learning models might not be possible, they argue. In contrast, Sullivan (2019) has argued that what can prevent deep learning from providing a model user with understanding of a phenomenon is not model interpretability, but rather "link uncertainty". With this term, Sullivan refers to the lack of evidence available that model users have to link the model to the phenomenon of interest. Hence, for Sullivan, the main problem lies not in the relationship between the model and the model user, but in the relationship between the model and its target phenomenon (or more specifically, in the justification of this relationship).

More work is needed that assesses data-driven models in terms of their ability to provide understanding of phenomena and what factors make this understanding better or worse. In chapter 4, we contribute to this debate. We do this by developing a framework that explicitly assesses different types of climate models in terms of their fitness for serving

as vehicles for understanding phenomena. We then apply this framework to data-driven models of climate phenomena and compare them to process-based models.

2. 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 Fischer[2], David N. Bresch[1,3], Reto Knutti[2]

[1] Institute for Environmental Decisions, ETH Zurich

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

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

(Published in Nature Climate Change, 2019, 9(3), 196-202.
doi: https://doi.org/10.1038/s41558-019-0404-1)

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 (Mayer-Schönberger and Cukier 2013). 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 (Lyon 2015; Pietsch and Wernecke 2017). Big data could have a big potential in various scientific disciplines (Karpatne et al. 2017) including climate research (Faghmous and Kumar 2014; Ford et al. 2016), 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 (Overpeck et al. 2011). Second, approaches typically associated with big data have already entered climate research (Faghmous and Kumar 2014; for examples see Caldwell et al. 2014; Kryvasheyeu et al. 2016; Sprenger et al. 2017). And third, climate models are rooted in scientific theory, which is one of the key reasons for confidence in their projections (Baumberger, Knutti, and Hirsch Hadorn 2017). 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 (Boyd and Crawford 2012; De Mauro, Greco, and Grimaldi 2016), 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) (Kitchin and McArdle 2016; Lukoianova and Rubin 2014). 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 (Hastie, Tibshirani, and Friedman 2008; LeCun, Bengio, and Hinton 2015).

2.1. 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 (Mayer-Schönberger and Cukier 2013). 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 (Linden, Smith, and York 2003).

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 (Goertzel and Pennachin 2007). 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 (Manogaran and Lopez 2018; Manogaran, Lopez, and Chilamkurti 2018), 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.

2.2. 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.

2.2.1. 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.

2.2.2. Datasets

Climate scientists use datasets to determine the initial conditions of variables of interest (McGuffie and Henderson-Sellers 2005) and to determine the values of certain parameters whose values are insufficiently constrained by theoretical considerations (Müller 2010), 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.

2.2.3. 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 (Knutti 2008), 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 (Pietsch and Wernecke 2017).

2.3. 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]	(Caldwell et al. 2014; Krasnopolsky and Fox-Rabinovitz 2006)
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]	[1]
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]	(Tapia et al. 2017)
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]	(Shelton et al. 2014; Castelli et al. 2016)
big-data analysis	everyday language concepts, automatically detected correlations[c]	partly unstructured data stream[c]	measurements based on everyday reasoning[c]	(Kryvasheyeu et al. 2016; Lu et al. 2016; Preis et al. 2013; Tkachenko, Jarvis, and Procter 2017)

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

[1] examples of studies in which all model relations are identified with machine learning: (Sprenger et al. 2017; Tripathi, Srinivas, and Nanjundiah 2006; Ghosh and Mujumdar 2008; Mendes and Marengo 2010; Chen, Yu, and Tang 2010; Wenzel and Schröter 2010; Chadwick, Coppola, and Giorgi 2011; Raje and Mujumdar 2011; Abbot and Marohasy 2012; Gagne II et al. 2012; Rasouli, Hsieh, and Cannon 2012; Mekanik et al. 2013; Merz, Kreibich, and Lall 2013; Nasseri, Tavakol-Davani, and Zahraie 2013; Tavakol-Davani, Nasseri, and Zahraie 2013; Abbot and Marohasy 2014; McGovern et al. 2014; Abbot and Marohasy 2015; Deo and Şahin 2015; Mohammadi et al. 2015; Patil and Deka 2016; Salcedo-Sanz et al. 2016; Andersen et al. 2017; Das, Chakraborty, and Maitra 2017; Dayal, Deo, and Apan 2017; Eghdamirad, Johnson, and Sharma 2017; Majdzadeh Moghadam 2017; Kashiwao et al. 2017; Park et al. 2017; Rahmati and Pourghasemi 2017; Roodposhti, Safarrad, and Shahabi 2017; Wu et al. 2013; Zhou et al. 2017)

2.3.1. Identifying Some Model Relations with Machine Learning

A study (Krasnopolsky and Fox-Rabinovitz 2006) 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. (2014).

2.3.2. 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 (Sprenger et al. 2017) 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 (Tripathi, Srinivas, and Nanjundiah 2006; Chadwick, Coppola, and Giorgi 2011; Tavakol-Davani, Nasseri, and Zahraie 2013; Nasseri, Tavakol-Davani, and Zahraie 2013); and for predicting climatic variables such as rainfall (Abbot and Marohasy 2012; 2014) and drought (Deo and Şahin 2015).

2.3.3. Finding Proxies for Missing Data

An example for this category is a study (Tapia et al. 2017) 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.

2.3.4. Theory-Structured Big-Data Analysis

An example for this category is a study (Shelton et al. 2014) 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 (Castelli et al. 2016).

2.3.5. 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 (Kryvasheyeu et al. 2016). 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.

2.3.6. General Findings

Some reviewed studies (Overeem et al. 2013; Elmore et al. 2014) relied on so-called crowdsourced weather information. Crowdsourcing refers to the process of collecting data from a large number of people (Muller et al. 2015). 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 (Bunn et al. 2015) 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 (Bunn et al. 2015).

Further studies used machine learning to assess climate change impacts on the global distribution of selenium in soils (Jones et al. 2017) and for the prediction of power output from wind (Foley et al. 2012) and solar power (Inman, Pedro, and Coimbra 2013) 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.

2.4. 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 (Tripathi, Srinivas, and Nanjundiah 2006; Ghosh and Mujumdar 2008; Mendes and Marengo 2010; Chen, Yu, and Tang 2010; Chadwick, Coppola, and Giorgi 2011; Raje and Mujumdar 2011; Nasseri, Tavakol-Davani, and Zahraie 2013; Tavakol-Davani, Nasseri, and Zahraie 2013), or the wide scope of the analyzed problems with unclear measures of success, as Shelton et al. (2014) 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 (Pietsch 2016), 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 (Karpatne et al. 2017).

2.5. 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.

2.5.1. Analyzing Increasing Volumes of Climate Data

The volume and complexity of data produced and stored are large and expected to further increase (Overpeck et al. 2011). 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 (Caldwell et al. 2014). However, this requires appropriate background knowledge to distinguish between potentially meaningful and meaningless patterns (Masson and Knutti 2013). This is especially true for datasets with a very large number of variables.

2.5.2. 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 (Lu et al. 2016). 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 (Welker et al. 2016). 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 (Arbuthnott et al. 2016), the constancy assumption could still be fulfilled for the estimation of a ceteris paribus baseline scenario.

2.5.3. Climate Services

Increasing volumes of climate data make it possible to provide more tailored information to users, often referred to as "climate services" (Vaughan and Dessai 2014). 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 (Benestad et al. 2017). 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 (Wahabzada et al. 2016). Similarly, machine learning and a dense network of climate and weather data might render farming practices more efficient (Walter et al. 2017), and hence contribute to more climate-resilient agriculture (often labeled "climate-smart agriculture", Lipper et al. 2014). In such cases, the understanding of the target system might justify

the constancy assumption especially when the forecasting horizon is comparatively short.

2.5.4. 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 (Katzav and Parker 2015). This approach could be assisted by machine learning (Schneider et al. 2017). This would essentially solve a small problem within the framework of a very complex problem.

2.6. 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.

Acknowledgments

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.

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.

3. Assessment of Predictive Uncertainty of Data-Driven Environmental Models

Benedikt Knüsel[1,2], Christoph Baumberger[1], Marius Zumwald[1,2], David N. Bresch[1,3], Reto Knutti[2]

[1] Institute for Environmental Decisions, ETH Zurich

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

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

(Submitted for publication to Environmental Modelling & Software)

Abstract

Increasing volumes of data allow environmental scientists to use machine learning to construct data-driven models of phenomena. These models can provide decision-relevant predictions, but confident decision-making requires that the involved uncertainties are understood. We argue that existing frameworks for characterizing uncertainties are not appropriate for data-driven models because of their focus on distinct locations of uncertainty. We propose a framework for uncertainty assessment that uses argument analysis to assess the justification of the assumption that the model is fit for the predictive purpose at hand. Its flexibility makes the framework applicable to data-driven models. The framework is illustrated using a case study from environmental science. We show that data-driven models can be subject to substantial second-order uncertainty, i.e., uncertainty in the assessment of the predictive uncertainty, because they are often applied to ill-understood problems. We close by discussing the implications of the predictive uncertainties of data-driven models for decision-making.

3.1. Introduction

Real-world policy decisions are taken in light of great uncertainties. Several aspects of a decision situation can be uncertain, including the problem framing, the available options from which to choose, the actual or potential states of the world, and the values that decision-makers attach to these states of the world (see Bradley and Drechsler 2014; Hirsch Hadorn et al. 2015; Hansson and Hirsch Hadorn 2016). In this paper, we are concerned with uncertainty about the actual or potential states of the world. Often, this uncertainty is related to uncertainty of scientific information on which a decision is to be based. This information relates to what the world will be like, either in the form of a prediction of environmental conditions, such as a severe weather forecast, or in the form of the conditional prediction (so-called projection) of environmental conditions in response to a policy measure, such as climatic conditions in response to a certain socioeconomic pathway and associated greenhouse gas emissions. In order to take decisions under uncertainty, the uncertainty should be analyzed and, if possible, quantified. An important part of the uncertainty analysis is the characterization of what kind of uncertainty emerges because of which features of the research process. Only such a thorough treatment of uncertainties can help to ensure that societal decisions are based on "no more and no less than what one actually knows" (Betz 2016a, 138).

Recent years have seen large increases in data produced and stored. This trend is also apparent in the sciences in general and in the environmental sciences in particular (e.g., in climate sciences, see Overpeck et al. 2011). The increase in the availability of data about environmental systems enables the use of machine learning to analyze the data and to construct data-driven models of phenomena using machine learning (see Gibert, Horsburgh, et al. 2018; Gibert, Izquierdo, et al. 2018). In this paper, we will discuss applications of machine learning for the data-driven modeling of a phenomenon. In contrast to a process-based model, in which the relationships between variables are prescribed in the form of equations specified by an expert, a data-driven model is constructed by algorithmically inferring the relationships or parameters from a dataset using machine learning (for a more detailed distinction on process-based and data-driven models, see Knüsel and Baumberger, under review). There have been numerous applications of machine learning in the environmental sciences (see, e.g., Reichstein et al. 2019). If predictions from data-driven models are reliable, which seems to be the case at least under certain conditions (see Pietsch 2015; Northcott 2019), these predictions are potentially useful for decision-making related to the modeled phenomena. One of the advantages of data-driven models is that they can be constructed when the processes producing a phenomenon are not fully understood. Hence, they might provide decision-relevant information specifically about phenomena which are quantitatively not understood well enough to construct process-based models. However, no tools to appropriately evaluate data-driven models in terms of their uncertainties are available to date, which reduces their usefulness for decision-making.

In this paper, we address the characterization of uncertainties of predictions from data-driven models. The focus of the present paper lies on uncertainties of predictions, and we will not distinguish between "pure" predictions of environmental conditions, such as a forecast of severe weather conditions, and predictions of the environmental response conditional on a human intervention, such as climatic conditions in response to a certain path of greenhouse gas emissions. The approach presented here is fairly general and can be applied in both situations, equally. In section 3.2, we argue that existing frameworks for the characterization of the uncertainties of model-based predictions are not appropriate for predictions from data-driven models. We hence introduce a new, more general approach in section 3.3. It focuses on the justification of the assumptions underlying a prediction that is possible based on the available data and background knowledge. The assumptions that need to be justified include the fitness-for-purpose of the used model. In sections 3.4 and 3.5, we demonstrate the application of this framework to a toy example and to a case study from environmental science. In section 3.6, we discuss the implications of this framework for the quantification of uncertainties and for decision-making more generally. We conclude in section 3.7.

3.2. Existing Frameworks

Existing frameworks for model-based decision-support and specifically uncertainty analysis have been developed for process-based models. An influential framework to analyze uncertainties of model-based predictions is due to Walker et al. (2003). It distinguishes three different dimensions of uncertainty that are arranged as an uncertainty matrix, namely the location (where in the modeling complex does the uncertainty manifest itself?), the nature (is the uncertainty due to the inherent variability of the phenomenon or due to imperfect knowledge of it?), and the level of uncertainty (how severe is the uncertainty, ranging from complete certainty to complete ignorance?). Several variations of this matrix have been developed (see e.g. Refsgaard et al. 2007), and versions of it have been applied in different contexts (see Kwakkel, Walker, and Marchau 2010). The locations of uncertainty introduced by Walker et al. (2003) are the context, the model structure and the technical model, the driving forces and the system data, the parameters, and finally the model outcomes, which obtain their uncertainty from the preceding locations. Kwakkel, Walker, and Marchau (2010) have synthesized adaptations and criticism of the uncertainty matrix introduced by Walker et al. and have, amongst other things, suggested that some of the dimensions, including the locations of uncertainty, be rearranged.

Similar locations of uncertainty have been discussed in specific disciplines. For climate models, for example, Knutti (2018) suggests that the relevant locations of uncertainty are model structural uncertainty, numerical approximations, parameterizations, natural variability due to initial conditions, emission scenario, boundary conditions, and observational data uncertainty, all of which have an analog in the locations discussed above.

Similar locations of uncertainty have been discussed for climate models by other authors (Winsberg 2018b, chap. 7).

What the approaches presented above have in common is that they suggest that different model-related aspects, the locations of uncertainty, are investigated in order to characterize the uncertainty of the model results. Namely, at each location, aspects should be identified that are either intrinsically or epistemically uncertain, i.e., uncertain due to system properties or due to our imperfect understanding of the system, and the level of this uncertainty should be determined. Then, this uncertainty is propagated in order to characterize the uncertainty of the processed model outputs.

Thinking about uncertainty in terms of specific locations is not informative for data-driven models for three reasons. First, for many data-driven models the model structure cannot readily be tied to processes in the target system or be interpreted in terms of the target system. Yet, model structure is one of the locations of uncertainty common to the established frameworks. This is especially obvious for models based on neural networks or bagging approaches like random forest. While neural networks have a model structure consisting of a number of neurons and layers, this structure cannot be interpreted in terms of the target system in a straightforward way. For bagging approaches like random forest, it is unclear what the relevant model structure would be because they make predictions based on the average of multiple individual estimators. For process-based models, structural model uncertainty arises because different model specifications might seem equally plausible and it is unclear how to best represent the target system for a specific purpose. This representational uncertainty is related to predictive uncertainty (see Parker 2010a). The best model setup can be more radically underdetermined in the case of data-driven models since a very large number of model setups can be consistent with the available training data. Hence, the best model setup for a specific purpose has to be chosen based on background knowledge. As data-driven models are often employed when processes are ill-understood (see Knüsel et al. 2019), the model setup of data-driven models and the resulting representational uncertainty of a target seem relevant for an analysis of predictive uncertainty even though the model structure, if there is an accessible model structure, cannot directly be tied to the target system. However, it is unclear how exactly representational uncertainty and predictive uncertainty are related for data-driven models. Second, similar to the model structure, model parameters are not an obvious location of uncertainty either for many data-driven models. Some machine learning approaches like random forest are non-parametric and hence, this location of uncertainty is simply not defined for them. Such non-parametric approaches do have meta-parameters, e.g., the number of trees or the number of considered variables at every split in a random forest model, but these, again, cannot readily be interpreted in terms of the target system. Other machine learning approaches like deep neural networks can have an overwhelming number of parameters. As for the model structure, many of the model parameters have no obvious interpretation in terms of the target system, and the uncertainty in each parameter does not directly translate to predictive uncertainty in the same sense. Third, machine learning is, at least sometimes, preoccupied mainly with

good predictions and less with the reasons for these predictions. But as we will see below, for some predictions, the reasons for the predictions can matter, too. A framework for assessing predictive uncertainties of data-driven models should thus reflect this purpose-dependence of the representational character of the model.

Thus, while existing typologies of uncertainty provide a good starting point to think about the uncertainty of predictions from data-driven models, they do not seem adequate as analytical tools. From the above discussion, two requirements become clear for an analytic framework for the predictive uncertainties of data-driven models. First, such a framework needs to do justice to the fact that under some, but not all, circumstances, a data-driven model has an instrumental character in the sense that modelers are preoccupied only with the predictive success of data-driven models and not with the reasons for this success. Second, in cases in which a modeler is not only preoccupied with the predictions but also with the reasons for predictive success, the representational accuracy of data-driven models needs to be assessed even though, as we have argued above, neither the model structure nor model parameters can readily be interpreted in terms of the target system, at least not for models with many degrees of freedom. In these cases, model users will be forced to adopt a more realistic interpretation of the model in order to justify confidence in model predictions. Hence, a framework to characterize the uncertainty from these predictions should allow to assess representational uncertainty without directly interpreting the model structure or the model parameters in terms of the target system. However, as outlined above, the representational uncertainty of a data-driven model should not be inferred from the underdetermination of its structure or its parameters. Specifically, this means that the framework needs to be able to assess to what extent the *behavior* (as opposed to the structure or the parameters) of the data-driven model is coherent with background knowledge and to what extent it is possible to argue from the coherence with background knowledge to representational accuracy.

Uncertainty has also been a topic in the computer science literature on machine learning. To the best of our knowledge, the literature on uncertainties in machine learning has mainly discussed the role of Bayesian approaches in order to quantify uncertainty (see Blundell et al. 2015; Ghahramani 2015; Gal and Ghahramani 2016). Kendall and Gal (2017) have explicitly distinguished between epistemic and aleatory uncertainty in the context of Bayesian deep learning for computer vision – a distinction we will take up in later sections. These Bayesian approaches are certainly useful for some contexts, such as computer vision and for cases in which uncertainty can be characterized easily because it only comes from certain sources, such as noisy data. However, for the present purposes, they seem insufficient as a basis of the uncertainty assessment of data-driven environmental models as it is unclear how they can be used to assess representational uncertainty in cases in which a relatively realistic interpretation of the model is adopted. However, we will return to how these methods could complement the ideas presented in our framework in section 3.6.

3.3. An Argument-Based Framework for Uncertainty Analysis

Above, we have argued that a framework for analyzing the predictive uncertainty of data-driven models needs to do justice to the fact that some, but not all, data-driven models have a purely instrumental character. Hence, the framework necessarily has to be more general than existing frameworks in order to allow for different representational characters in different contexts. Here, we put forward such a framework. The framework suggests to analyze uncertainties in three steps. The first step consists in reconstructing what assumptions have to be made when using a model for a specific purpose and how these assumptions can be justified. The second step consists in evaluating how well justified the assumptions are. The third step consists in assessing the uncertainty based on the previous two steps. The basic assumption that has to be justified in any modeling application is an assumption regarding the fitness-for-purpose of the model used. As Parker (2009) has argued for climate models, the goal of model evaluation should *not* be to confirm the model itself, generally, but rather to confirm that a model is adequate for a specific purpose. In this paper, we use the term "fitness-for-purpose", which, in contrast to adequacy-for-purpose, admits of various degrees of fitness (meaning that models are not just fit-for-purpose or not, but fit-for-purpose to a larger or smaller degree, see Parker, forthcoming).

For the present purposes, a model is considered maximally fit for a specific predictive purpose if it can reliably predict the variable of interest with errors lying in some small range. If a model is maximally fit for purpose, this range, here, should be understood to depend solely on the inherent variability of the target system, meaning that it arises due to aleatory uncertainty. In other words, the model skill converges toward the theoretical limit of predictability of the system as the fitness-for-purpose increases. An example of such a fitness-for-purpose assumption would be: Model M is fit for predicting the total precipitation amount of the next 24 hours at location L with errors in some small range. If it can be conclusively justified that the model is maximally fit-for-purpose in this sense, the model predictions exhibit no epistemic uncertainty. If the degree of fitness-for-purpose is lower, the range will increase, and this will be due to epistemic uncertainty. Hence, according to the framework presented here, epistemic uncertainty arises because of factors that reduce the fitness-for-purpose to a lower degree than maximal fitness-for-purpose or factors that make it unclear what degree of fitness-for-purpose a model has.

In the following, we will introduce the three steps of the framework in more detail. The first step is the reconstruction of the assumptions and their possible justifications. In order to do this, the following two questions have to be addressed: First, what modeling assumptions does a prediction rely on? Second, how can these assumptions be justified? As noted above, the basic assumption is a fitness-for-purpose assumption concerning the model that was used to produce a prediction. Depending on the circumstances, the fitness of the model for the predictive task at hand will be justified differently. If the justification is itself (partly) based on assumptions that require further justification, the

two questions above need to be addressed several times in order to assess how well the basic assumption is justified. Once this is done, an argument map can be drawn in which it becomes apparent which arguments and assumptions justify the fitness-for-purpose assumption.

In the second step, based on the argument map from the first step (or in simpler cases, based on a direct assessment of the arguments), it is evaluated how well the different assumptions and specifically the fitness-for-purpose assumption are justified. For this, it is relevant to check whether the premises used to justify a conclusion are true and whether they provide sufficiently good reason to accept the truth of the conclusion. Whenever possible, arguments should be reconstructed as deductively valid arguments, i.e., arguments for which the following condition holds: if all the premises are true, then the conclusion must be true. This way, the assumptions become more explicit and the actual sources of uncertainty can be identified more easily. An example of a deductively valid argument is the following (this example is taken from Baumberger, Knutti, and Hirsch Hadorn 2017, 7): If a model is adequate for projecting X for the far future, then the model reliably indicates *X* for past and present. Model *M* does not reliably indicate *X* for past and present. Hence, *M* is not adequate for projecting *X* for the far future. Some arguments cannot easily be reconstructed as deductively valid, e.g., because one would have to add premises that are suspected or even known to be false. In these cases, the arguments can be reconstructed as non-deductively correct, i.e. as arguments that provide sufficiently strong (but not conclusive) reasons for the truth of the conclusion. Non-deductively correct arguments are risky in the sense that even if all of their premises are true, the truth of the conclusion is not guaranteed. An example of a non-deductive argument is the following (this example, too, is taken from Baumberger, Knutti, and Hirsch Hadorn 2017, 7): Model *M* reliably indicates *X* and climate quantities upon which *X* depends for past and present. So probably, *M* is adequate for projecting *X* for the near future.

Finally, the third step consists in an assessment of the uncertainty. By analyzing the possible epistemic justification of assumptions, the focus of the framework presented here is on epistemic uncertainties. The framework distinguishes between two types of epistemic uncertainty that we refer to as "first-order uncertainty" and "second-order uncertainty".[8] These two types of uncertainty are distinguished by their objects. First-order uncertainty is the epistemic uncertainty of the prediction. Consequently, first-order uncertainty is defined, here, as the extent to which the assumptions underlying a prediction are not or insufficiently justified. Thus, epistemic first-order uncertainty arises if it cannot be conclusively justified that the model is maximally fit-for-purpose. Second-order uncertainty is defined, here, as the extent to which the assessment of first-order uncertainty is impaired by context-specific factors. Specifically, second-order uncertainty

[8] Note that the justifications always have to rely on arguments that can be constructed from what is known. Hence, the framework introduced here is not helpful to identify uncertainty arising from unknown unknowns.

depends on difficulties in assessing to which extent the assumptions are justified, e.g. because of a lack of evidence, contradictory evidence, or expert disagreement. Thus, epistemic second-order uncertainty arises if the degree of fitness-for-purpose cannot be conclusively determined. The assessment of these two types of uncertainty is based on the evaluation of the uncertainty from the second step, as will be illustrated below. Based on the framework, the expressions of first-order and second-order uncertainty will be purely qualitative. However, in section 6 below, we address the questions of uncertainty quantification, which is desired for many applications.[9]

Before we proceed to illustrating the application of the framework, two clarifications are in order. First, in practical applications, the fitness-for-purpose of a model will not only depend on its predictive accuracy. Further considerations such as practical concerns can play a role, for example ease-of-use or computational cost. However, in the present paper, we are only concerned with the uncertainty of the predictions of a model and we will, hence, discuss fitness-for-purpose only as it relates to predictive accuracy. Second, note that if a model is found to be less-than-maximally fit-for-purpose, this does not mean that the model is outright unfit-for-purpose. Whether a given degree of fitness-for-purpose is sufficient to consider a model outright fit (or adequate) for a specific purpose depends on the context.

The identification and evaluation of how the assumptions are justified requires expertise from domain scientists concerned with the phenomenon at hand and from modelers and data scientists, but also expertise in argument analysis (an introduction to argument analysis can be found in Brun and Betz 2016). In the literature, it has been recognized that an analysis of assumptions can be important for a better understanding of uncertainties (see Kloprogge, van der Sluijs, and Petersen 2011). However, as will be shown below, since the framework presented here is concerned with justifying the fitness-for-purpose, it is not only concerned with assumptions that a modeler makes in the process of model construction. Instead, it is concerned with assumptions that a model user is relying on, at least implicitly, when using a given model for a specific predictive purpose. For example, a model user might have to assume that certain processes are accurately represented in a data-driven model. However, this assumption is not made explicitly in the process of building a data-driven model as the relationships between variables are not explicitly prescribed (see Pietsch 2015). Hence, there are usually no explicit representations of processes in data-driven models.

---

[9] We note here that uncertainties of higher orders could be defined in analogy to second-order uncertainty. For example, third-order uncertainty could be defined as the uncertainty of the assessment of second-order uncertainty. As the discussion below will highlight, second-order uncertainty is an important concept to better understand the epistemic uncertainty of model predictions. Uncertainties of higher order are likely irrelevant in practical applications.

3.4. Toy Example for Illustration

In this section, we discuss a simple toy example of a data-driven model to illustrate how the framework can be applied. Our toy model is a random forest algorithm that is trained to predict the maximum daily air temperature at a given location with a lead-time of one day. For this, the current air temperature and pressure at the location, the season, and an index on the general weather conditions are used as predictors. Imagine that the model predictions are successful and actually measured maximum daily air temperatures are always close to the predicted value (within a small error range), and we are able to repeatedly use the model and evaluate its performance.

We might generally wish to better understand the uncertainty of this kind of prediction in order to base decisions on them.[10] For example, the prediction of an unusually cold or hot maximum temperature might be relevant for public health. In order to obtain a better understanding of the uncertainty of the predictions, we can apply the framework introduced in the previous section. First, the assumptions and their possible justifications need to be identified and graphically arranged. As explained above, the most fundamental assumption is a fitness-for-purpose assumption. In this case, the fitness-for-purpose assumption states that the model is fit for making predictions of the daily maximum temperature with a lead time of a day for a specific location up to some range. How can this assumption be justified? Since we have used the model and evaluated the accuracy of its predictions repeatedly in the past, this past performance can be used to justify the fitness of the model for the predictions at hand. Namely, the model has predicted many past instances of maximum daily temperature accurately.[11] This justification can now be illustrated, for example in a table as shown in Table 2.

Table 2: Application of the conceptual framework to the toy example with maximum daily temperature predictions.

assumption	**justification of the assumption**
The model is fit for predicting maximum daily temperature with a lead time of one day.	The model has accurately predicted many past instances of maximum daily temperature with a lead time of one day.

[10] The same considerations could also be applied in an analysis for this specific instance of a prediction from the model. Then, the reconstruction of the assumptions would have to include that the conditions of the specific prediction are sufficiently similar to past conditions. Furthermore, the fitness-for-purpose assumption would have to be reformulated to refer to the specific prediction instance.

[11] Implicitly, this can be read as saying that the model does not make an extrapolation far outside the range of values for which it has been trained.

This justification can be reconstructed as a deductively valid argument of the following form:

P1	*If a model has predicted many past instances of a variable accurately and the conditions for the predictions remain sufficiently similar to the past instances, the model is fit for predicting that variable.*
P2	*Model M has predicted many past instances of T_{max} accurately.*
P3	*The conditions for the predictions of T_{max} remain sufficiently similar to past instances.*
C	*M is fit for predicting T_{max}.*

Now that the argumentation is reconstructed, we can proceed to the second step, namely evaluating to what extent the fitness-for-purpose assumption is justified. For the evaluation of the justification, we need to assess whether the premises are true and to what extent they provide good reasons for the fitness-for-purpose assumption. As the justification can be reconstructed as a deductively valid argument, the conclusion must be true if the premises are true. Thus, the evaluation consists in determining whether all premises are true. P1 makes a conditional claim about the fitness-for-purpose of the model. This premise, we take it, is uncontroversial. The premise P2 is related to the evaluation of past predictions of the model. As mentioned above, the model has been extensively used in past cases and has been predictively successful. Thus, P2 is true, too. The truth of P3, finally, has to be justified based on domain-specific background knowledge, namely that the past cases are sufficiently representative to be confident about the models' performance more generally for cases that are similar to the ones considered thus far. The short lead-time of the predictions makes it likely that P3 is true, too, for two reasons. First, on a practical level, the short lead-time allows for repeated evaluations of the predictions. Second and more fundamentally, the short lead-time increases the chance that the evaluated predictions of the past are representative of the current predictions because the system is less likely to have experienced large changes in boundary conditions over a short period of time.[12] Hence, we have a deductively valid argument that justifies the fitness-for-purpose of the model, and there are good reasons to assume that all of its premises are true. Hence, the argument seems to be sound (a sound argument is a deductively valid argument with true premises), which means that the conclusion of the argument is true.

Finally, in the third step, the epistemic uncertainty of the prediction can be assessed. The first-order uncertainty depends on the extent to which the fitness-for-purpose assumption is justified. As we have seen, the justification of the fitness-for-purpose, here,

[12] Note that the argument could as well be reconstructed without premise P3. This would turn the argument into an inductive argument whose strength would have to be assessed based on the representativeness of the past cases. We choose and recommend the deductive reconstruction as it makes the uncertainty more explicit.

seems to a sound argument, meaning that the model can safely be considered close to maximally fit-for-purpose. This means that epistemic first-order uncertainty is very small or even absent in the toy example. Epistemic second-order uncertainty is also small or even absent as it is straightforward to reconstruct and evaluate the argument for the fitness-for-purpose assumption. There may be some second-order uncertainty related to premise P3 if it is unclear to what extent the past performance is representative of future performance. The justification of the truth of P3 fundamentally depends on the system understanding.

Note that in this simple toy example, the framework can also help to identify the aleatory uncertainty. The reason for this is that the total uncertainty of the predictions can be estimated in a straightforward way based on the evaluated predictions. Since the epistemic uncertainty of the predictions can be reliably estimated as a result of the low level of second-order uncertainty, the aleatory uncertainty is simply the difference between total uncertainty and epistemic first-order uncertainty. In the present case, aleatory uncertainty corresponds to the range that can be estimated from the small random errors of model predictions. As Kendall and Gal (2017) have argued, in machine learning aleatory uncertainty is best understood as uncertainty that cannot be reduced by collecting additional samples of data, which is exactly the kind of uncertainty that remains here. In this toy example, the aleatory uncertainty can readily be quantified based on records of past model performance.

In sum, all things considered, both epistemic first-order and second-order uncertainty turn out to be very small in this example. The reason for the small epistemic uncertainty is that the model predictions have been evaluated repeatedly for similar cases. This allows model users to adopt a purely instrumental view of the model, meaning that the focus of the model evaluation lies purely on its predictive success and not on the reasons for why the model makes certain predictions, nor on its structure. Hence, in such a case, an evaluation of uncertainty in terms of model structure or model parameters would make little sense. This highlights that the framework introduced here can deal with cases where model users have a purely instrumental view of their models.

3.5. Case Study: Long-Term Global Selenium Predictions

In this section, we present a case study from environmental science to demonstrate the application of the framework to a long-term prediction to illustrate how the framework works in more complex situations than the toy example from the previous section. Namely, we discuss the case of predictions of changes in global soil selenium content by Jones et al. (2017). The study used data-driven models for three goals, namely "(i) to test hypothesized drivers of soil Se [selenium] concentrations, (ii) to predict global soil Se [selenium] concentrations quantitatively, and (iii) to quantify potential changes in soil Se [selenium] concentrations resulting from climate change" (Jones et al. 2017, 2848). For illustration, we will discuss the last of these three goals, the impact of climate change on soil selenium.

Jones et al. (2017) used data from over 30.000 samples worldwide to train three different models, namely, two artificial neural networks and a random forest model. All three data-driven models relate environmental variables to soil selenium concentrations. The authors performed a variable selection procedure, after which the seven most important predictors were retained and trained the three models using historical data. The chosen predictors were the aridity index, the clay content, evapotranspiration, lithology, pH, precipitation, and soil organic carbon. These trained models were then used to project changes in soil selenium concentrations due to climate change, hereby using changes in precipitation and evapotranspiration from climate models (for RCP6.0) and an accompanying scenario for the development of soil organic carbon. Doing this, they estimated that average soil selenium concentrations will decline by 4.3% under the chosen boundary conditions. Selenium is an essential micronutrient, which makes information on future selenium loss of this magnitude potentially decision-relevant. For example, changes to farming practices might be required to counter the climate impacts to ensure nutritionally adequate crops. Such measures could include fertilization and relying on crops that can take up selenium from the soil even if soil selenium concentrations are lower.[13] However, taking decisions about such measures requires confidence in the predictions, which, in turn, requires an analysis of the uncertainties of the predictions. For this, the framework introduced here can be applied.

Again, in a first step, the assumptions underlying the predictions and their possible justification need to be reconstructed and graphically represented. An overview of all of the assumptions and their justification is provided in Table 3. The first assumption is the fitness-for-purpose assumption. The fitness-for-purpose assumption states that the models constructed with the given set of drivers and historical data allows to project (i.e., make a conditional prediction of) future selenium concentrations.[14] However, in this case, the fitness-for-purpose assumption can no longer be conclusively justified based on the repeated evaluation of the model predictions (as it was in the toy example) because it is unclear whether the future cases are sufficiently similar to past cases. Hence, further ways of justifying the fitness-for-purpose are required besides past model performance. A plausible justification is that the modeled relationships can be assumed to be sufficiently constant over time and hence, can be extrapolated into the far future (see

---

[13] Which of these methods would be best suited to address selenium losses is, of course, fraught with uncertainties, including uncertainty about what is most important for local populations directly affected by potential selenium losses. For reasons of simplicity we will not engage in discussions of specific policy measures here but only discuss the uncertainties of the prediction.

[14] The reconstruction presented here assumes that the boundary conditions require no further justification and can just be regarded as given. This is done for simplicity. A different reconstruction would be possible in which the adequacy of the boundary conditions, which depends on the internal consistency of the scenario and on how informative the scenario is for the purpose at hand, could be included. This assumption would then have to be justified based on an independent evaluation of the respective models that were used to create these scenarios. As this point is not essential for the present purposes, we will not engage with this discussion in more detail.

Knüsel et al. 2019). However, this assumption itself requires further justification. A possible justification of this constancy assumption is that the model accurately represents the most relevant causal processes that drive selenium concentrations and that these mechanisms will not change in response to changing environmental conditions.[15]

Table 3: Application of the conceptual framework to the example with projections of long-term selenium concentrations.

iteration	assumption	justification of the assumption
1	The model is fit for projecting soil selenium concentrations in the far future for the given boundary conditions.	The model predicts past instances of selenium concentrations well.
		The model relationships are sufficiently constant over time.
2	The model relationships are sufficiently constant over time.	The model represents most relevant processes driving selenium concentrations accurately and these remain sufficiently constant under changing environmental conditions.
3	The model represents most relevant processes driving selenium concentrations accurately and these remain sufficiently constant under changing environmental conditions.	The most relevant predictors were included in the model.
		Data from many regions was used for training.
		Sufficiently flexible machine learning algorithms were used.
		The model is empirically accurate when tested with data from the past.
		Model behavior is consistent with background knowledge.
		The model results are robust to the modeling assumptions.
4	The most relevant predictors were included in the model.	The variables were identified based on domain-specific background knowledge and a variable selection procedure.
	Data from many regions was used for training.	These regions are sufficiently representative of possible configurations.
	Sufficiently flexible machine learning algorithms were used.	Neural networks and random forest are very flexible methods.

[15] We note here that a different justification for the constancy of the described relationship could also be that two factors whose relationship is modeled have a common cause instead of being directly causally related. However, the direct path is the more plausible one here.

Yet, that the model represents most of the relevant causal processes accurately is itself an assumption since no mechanisms were explicitly included in the model, and hence, it needs to be justified. Furthermore, even if the model represents important causal processes, it may be unclear to what extent these mechanisms remain constant when extrapolated to changing environmental conditions. This is relevant, here, because the model is used to make projections for values of the variables that are somewhat different from today's values because of climate change. The justification of the assumption that most of the relevant causal processes are accurately represented leads to a third argumentation iteration, and hence a third row in Table 3.

Due to the lack of explicit representations of processes, the assumption needs to be justified indirectly. There are several ways to justify this assumption, namely (1) that the most relevant variables were included, (2) that data from many different regions was analyzed, (3) that sufficiently flexible machine learning algorithms were used, (4) that the models are empirically accurate (i.e., the cross-validation error is low), (5) that model behavior, as assessed through sensitivity analysis, is consistent with background knowledge about the system, and finally (6) that three different machine learning algorithms were used and largely agreed (i.e., a robustness argument). These reasons have been suggested as conditions for the adequacy of machine learning approaches in the philosophical literature. For example, Pietsch (2015) has stressed the importance of (1) and (2), while Knüsel and Baumberger (under review) have stressed the importance of all six points for evaluating to what extent a data-driven model is coherent with background knowledge. Now, (1), (2), and (3) again require further justification, which leads to a fourth iteration. In order to justify them, scientists have to rely on both background knowledge on the behavior of trace elements in the environment and understanding of and experience with machine learning.

With this, the possible justification of the fitness-for-purpose has been identified. For space reasons, we do not provide an explicit reconstruction of the arguments here. A reconstruction of the complex argumentation is provided in the appendix. We note here that the arguments in the first two iterations can be reconstructed as deductively valid arguments. This is not easily possible for the third iteration. In the appendix, we also provide an argument map that shows how the different arguments relate to each other. Note that the empirical accuracy of the model predictions now appears twice as a justification: It is a necessary condition for considering the model fit-for-purpose in the first iteration, but it also gives some indication that the relevant processes are represented accurately in the model in the third iteration.

In the second step, the arguments for justifying the fitness-for-purpose have to be evaluated. The truth of the individual premises (see the reconstruction in the appendix) has to be evaluated based on domain-specific background knowledge, the evaluation of the models with available data, and the comparison of the behavior of the three individual models. Based on the model evaluation as discussed by Jones et al. (2017), all the premises of the arguments seem to be at least approximately true. There is, however, one

exception to this, concerning the choice of variables: As Jones et al. (2017) note, their model lacks selenium sources, which leads to an underprediction of global average selenium values. The reason for this is that data on selenium sources like atmospheric deposition or biomass deposition was missing. That there is an underprediction of average soil selenium concentrations attacks the inference from the empirical accuracy of the models to the conclusion that most relevant processes are accurately represented in the models. That data on selenium sources was lacking attacks the premise that most relevant predictors were considered because the global underprediction shows that selenium sources are important for soil selenium concentrations. This means that the strength of the argument from empirical accuracy is reduced somewhat, and the argument about the most important variables being considered has a premise that is strictly speaking false. Furthermore, while several arguments can be provided to argue for the assumption that most relevant mechanisms are accurately represented in the model, they neither individually nor jointly guarantee that the mechanisms are represented accurately. This is because in the third iteration, the arguments are not deductively valid but provide only more or less strong reasons for the truth of their conclusion. Hence, also the preceding assumptions about the constancy of the identified relationships and the fitness-for-purpose cannot be justified conclusively.

Now, in a third step, we can proceed to assessing the uncertainty based on the argumentation reconstructed above. As noted, some of the provided premises are known to be false, strictly speaking, namely the assumption that all relevant variables were included. This also somewhat affects the empirical accuracy of the models. Furthermore, the justifications provided can neither individually nor jointly guarantee that the models are really fit for purpose. The lack of good justification of some of the assumptions leads to a lower-than-maximal fitness-for-purpose, which means that there is more epistemic first-order uncertainty in this case study compared to the toy example from above. The framework introduced here does not only highlight that the epistemic uncertainty of the prediction is comparatively large, it also highlights which specific aspects of the justification are responsible for this. Second-order uncertainty is also substantially larger than in the toy example above. The reason for this is that it is not clear to what extent the modeling assumptions are justified by the provided evidence. This has to do on the one hand with opacity of the models, as it is not entirely clear what relations they actually represent and on what grounds.[16] More importantly, it has to do with the lack of background knowledge to judge to what degree the assumptions are justified by the provided arguments. This lack of background knowledge makes it difficult to assess the strength of the non-deductive arguments in the third and fourth iterations shown in the Table 3.

All uncertainties considered, we see that in this example, both first-order and second-order epistemic uncertainty are present to a larger degree than in the toy example. First-

[16] The reason for this is that the model does not provide an explicit equation or a set of rules that could be analyzed. This is especially true of the models used by Jones et al. (2017), random forest and neural networks.

order uncertainty relates to a lower degree of fitness-for-purpose because the justification of fitness-for-purpose is less strong. There are two reasons for this. First, some of the arguments provided are non-deductive, meaning that the conclusion need not be true even if all the premises are true. This means that the provided justification cannot guarantee that the model is generally fit for the kind of prediction of interest. Second, the lacking data on selenium sources and, relatedly, the global underprediction of average selenium concentrations attacks two arguments for the overall fitness-for-purpose, which reduces the fitness-for-purpose to a less-than-maximal level. Note again, here, that even though the degree of fitness-for-purpose is less-than-maximal, the models are not outright unfit-for-purpose. Second-order uncertainty relates to certain arguments whose strength is difficult to evaluate. For example, it is unclear whether the models really do capture important causal processes that are sufficiently constant under changing environmental conditions. This is due to a lack of domain-specific background knowledge. Also, the arguments provided for the statement that the model represents important causal processes accurately are all non-deductive. These non-deductive arguments introduce first-order uncertainty because the justification of fitness-for-purpose becomes less conclusive. They also introduce second-order uncertainty because the strength of the justification is difficult to evaluate due to a lack of system understanding.

The arguments discussed above can also be found in the paper by Jones et al. (2017) who discussed them in order to understand the uncertainties of the inferences. For example, they provide a discussion of missing variables (concerning (1) above) and of the representativeness of the available samples (concerning (2) above). Furthermore, they discussed the empirical accuracy of the model in a cross-validation setting (concerning (4) above) and conducted sensitivity analyses to assess whether model behavior is consistent with background knowledge (concerning (5) above). Finally, they only considered predictions for pixels where the three data-driven models agreed in the sign of change. Hence, they considered the robustness of the predictions (concerning (6) above). This shows that the arguments provided here were actively engaged with by the authors of the original study. Note, however, that the assumptions discussed are not explicitly made during the process of model construction by the modelers. Rather, they are assumptions that modelers need to make once they apply the models for certain kinds of long-term predictions.

As noted, both first-order and second-order uncertainty were considerably larger in this case study than in the toy example above. The reason for this is that the data-driven models were constructed for the selenium prediction but due to the long lead-time of the prediction and the lack of evaluation of the model predictions for the desired purpose, model users had to adopt a more realist interpretation of model behavior (compared to the more instrumental view of the model in the toy example). Hence, the reason for the increase in both types of epistemic uncertainty is not simply that data-driven models were used, but rather that data-driven models were used in a context where the uncertainty cannot be estimated from the past performance of the model alone. The conclusion of this discussion is likely to hold more generally in cases where background knowledge

is insufficient to provide conclusive justification of the fitness-for-purpose assumption. As Knüsel et al. (2019) argue, data-driven models are often constructed when background knowledge is insufficient for constructing process-based models. Hence, the points about increasing second-order uncertainty are likely to hold more generally, not just for this case study. The flexibility of the framework presented here might lead to different arguments being relevant in different contexts. However, the arguments highlighted in this case study are likely to show up in different contexts again, specifically the six reasons provided for assuming that the model presents most of the relevant processes accurately. In some examples, it is well possible that further iterations are required to justify some of the six points raised above.

3.6. Implications for Decision-Making

One of the key reasons for better understanding the uncertainties of scientific inferences is that this understanding is required for epistemically confident decision-making. Hence, more needs to be said about how the kind of information provided by the framework presented here can be handled in decision-making. This is specifically important because the framework delivers two types of epistemic uncertainty, first-order and second-order uncertainty and characterizes them in a purely qualitative form. In this section we address how the information on uncertainty provided by our framework can be used effectively for decision-making and point to areas where further research is necessary.

Many decision-principles require that information on first-order uncertainties be quantified. However, as Walker et al. (2003, 8) state, quantified "statistical uncertainty should not be accorded as much attention as other levels of uncertainty in the uncertainty analysis" if there are more severe levels of uncertainty present. This means that uncertainties should only be quantified when researchers are in a position to do so confidently. Doing this first requires a good understanding of what the relevant sources of uncertainty are, i.e., it requires an understanding of which assumptions lead to uncertainty. In this sense, the framework presented here can be used to build the groundwork for uncertainty quantification because it highlights where uncertainties come from and what the relevant uncertainties are.

Approaches exist to quantify uncertainties from machine learning predictions. Some of these, such as quantile regression forests, directly provide probabilistic information by predicting not only the best estimate but also the quantiles of the probability distribution function, which is learned from the data (Meinshausen 2006). There are also approaches for estimating uncertainty that are based on Bayesian reasoning that account for the uncertainty of individual parameter values in deep learning (see e.g. Blundell et al. 2015; Gal and Ghahramani 2016). These approaches are useful to quantify uncertainty that can directly be inferred from the available data. We recognize that they yield valuable information and can provide a full account of uncertainties in some settings, e.g., in image classification tasks. However, in cases such as the case study considered in this paper, these approaches would not be able to quantify the full uncertainty. While these

methodological approaches help to assess the robustness of the results, they do not address all of the sources of uncertainty identified above. Hence, in cases such as the case study discussed above, additional approaches for uncertainty quantification are needed.

One promising approach might be to rely on structured expert elicitation in order to estimate quantitative information on uncertainties from the qualitative information that the framework presented here provides (see Morgan 2014; Thompson, Frigg, and Helgeson 2016; Oppenheimer, Little, and Cooke 2016). As it will generally be difficult to create exact uncertainty estimates based on the framework, experts will likely be inclined to provide imprecise probability estimates. This would require experts to consider a graphical representation such as the argument map provided in the appendix and assess the strength of the arguments provided for the fitness-for-purpose assumption at hand. The aforementioned methods for uncertainty quantification based on the robustness of the results can provide a good starting point here. Based on the expert assessment, the intervals obtained would have to be widened or narrowed accordingly. The strength of the arguments discussed above should be assessed by domain experts. For some of the factors leading to uncertainty, it can suffice to specify plausible scenarios without quantitative information on their probability (this would be scenario uncertainty in the matrix of Walker et al. 2003). This is for example the case for the information on boundary conditions regarding the changing climatic conditions in the case study introduced above.

The framework does not only provide information on first-order but also on second-order uncertainty. When quantifying first-order uncertainty, second-order uncertainty should be considered, too. A large second-order uncertainty means that it is difficult to judge the degree of fitness-for-purpose of the model. This means that first-order uncertainty will be only weakly constrained. If first-order uncertainty is less well constrained, a trade-off emerges. Experts can either provide narrower estimates of first-order uncertainty and be less confident about it (i.e., they face more second-order uncertainty) or provide a wider estimate of first-order uncertainty with more confidence (see Winsberg 2018b, chap. 7). Balancing this trade-off has to be based on what is perceived to be the most useful for decision-makers (Winsberg 2018a).

Research into the development of decision principles that can be used with the two-tiered information on uncertainty discussed here is still needed (Winsberg 2018b, chap. 8). A candidate approach is the confidence approach that considers different models depending on decision makers' risk attitude (Roussos, Bradley, and Frigg, under review). A different approach is decision-making with possibilistic information, i.e., with information on what is and what is not consistent with our understanding of a system (Betz 2016a). If model-based information is handled with such a possibilistic mindset, a greater second-order uncertainty implies that it is more difficult to distinguish between outcomes that are consistent with our background knowledge, outcomes that are inconsistent with our background knowledge, and outcomes that cannot be put in either of these categories. Hence, a possible outcome (an event with some epistemic

first-order uncertainty) with a large second-order uncertainty might have to be considered by a risk-averse decision-maker even if its largest possible likelihood, as estimated based on the first-order uncertainty, seems small. The reason for this is that its (first-order) uncertainty assessment is uncertain and might need to be revised in light of new information.

Predictions are not the only way in which models can provide decision-relevant information. Namely, knowledge of causal connections and exploratory modeling can guide policy decisions (Weaver et al. 2013). In such cases, models are needed that represent the processes responsible for producing a phenomenon with sufficient accuracy. Data-driven models can be fit for providing this kind of information, too. In these cases, the evaluation of the models' fitness is similar to the uncertainty analysis seen in section 3.5 (see Knüsel and Baumberger, under review).

3.7. Conclusions

In this paper, we have presented an argument-based framework for assessing the uncertainties of model-based predictions. We hereby focused on features of data-driven models and showed that the framework is able to analyze the uncertainty of predictions from data-driven models. Based on a toy example and the extensive discussion of a case study from environmental science, we highlighted how the application of the framework works in practice. Constructing data-driven models is possible also when a phenomenon is comparatively ill-understood. However, this lack of background knowledge and the opacity of data-driven models can lead to substantial second-order uncertainty, as we have shown here. We then discussed what the framework implies for the quantification of uncertainties and for decision-making based on information from data-driven models. Open questions remain specifically with respect to the quantification of uncertainties. We encourage attempts at using structured expert elicitation as suggested here and further research into decision principles.

Environmental scientists working with data-driven models are often aware of the limitations and uncertainties of their models. However, the lack of conceptual tools for uncertainty assessments may inhibit a clear understanding of how large these uncertainties are. Thus, there can potentially be overconfidence about results obtained with data-driven models. The lack of conceptual tools can also impair a better understanding of the factors that lead to the uncertainty. Understanding these factors can be useful for researchers, e.g., to identify what steps they could take to reduce the impact of a specific factor that leads to uncertainty. The framework presented here provides tools to perform such uncertainty assessments and communicate the uncertainty of predictions from data-driven models more transparently. Hence, we encourage researchers developing and working with data-driven models to employ the framework provided here to assess the predictive uncertainties of their models. Being more explicit about uncertainties increases the usefulness of data-driven models both for scientific and policy purposes. At the same time, explicitly discussing the representational function of models may reveal

that data-driven models are more skillful in some applications than one might have expected initially. Hence, the argument-based framework provided here can help to make good use of data-driven models in environmental science.

Kwakkel, Walker, and Marchau (2010) have emphasized the importance of using a common language in uncertainty assessments in order to provide information to decision-makers that is easier for them to compare to other cases and contexts. We agree with this view. However, in the case of data-driven models, it seems unlikely that locations of uncertainty similar to the ones from other frameworks can be defined that can be applied to data-driven models generally and are informative of their predictive uncertainty. For example, it might be intuitive, here, to speak of "model uncertainty" and "extrapolation uncertainty". However, as the discussion of the case study has shown, how much uncertainty the extrapolation introduces directly depends on properties of the model. Hence, these two terms would not refer to distinct locations of uncertainty. However, we encourage future work that aims to find a terminology for the information from our framework that can consistently be related to uncertainties from other frameworks. Future research should also address decision principles that can handle the kind of uncertainty that the framework presented here provides.

The considerations made in this paper are likely relevant beyond data-driven models. The framework discussed here is quite general. It can hence be applied to other types of models, too and could hence complement existing discussions of uncertainty of environmental models. Furthermore, as the framework focuses on assumptions and how they are justified, it can potentially reveal that some of the analyzed assumptions concern value judgments and hence highlight cases of value uncertainty.

Acknowledgments

We thank Richard Bradley, Roman Frigg, Gertrude Hirsch Hadorn, David Stainforth, and Lenny Winkel for discussions and/or feedback on earlier versions of this manuscript. We further thank the participants of the workshop *Uncertainty in Data-Driven Environmental Modeling* at ETH Zurich in August 2019.

Funding

This research was funded by the Swiss National Science Foundation under the National Research Programme "Big Data" (NRP75), project no. 167215.

4. Understanding Climate Phenomena with Data-Driven Models

Benedikt Knüsel[1,2] and Christoph Baumberger[1].

[1] Institute for Environmental Decisions, ETH Zurich

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

(Submitted for publication to Studies in History and Philosophy of Science)

Abstract

In climate science, climate models are one of the main tools for understanding phenomena. Here, we develop a framework to assess the fitness of a climate model for providing understanding. The framework is based on three dimensions: representational accuracy, representational depth, and graspability. We show that this framework does justice to the intuition that classical process-based climate models give understanding of phenomena. While simple climate models are characterized by a larger graspability, state-of-the-art models have a higher representational accuracy and representational depth. We then compare the fitness-for-providing understanding of process-based to data-driven models that are built with machine learning. We show that at first glance, data-driven models seem either unnecessary or inadequate for understanding. However, a case study from atmospheric research demonstrates that this is a false dilemma. Data-driven models can be useful tools for understanding specifically for phenomena for which scientists can argue from the coherence of the models with background knowledge to their representational accuracy and for which the model complexity can be reduced such that they are graspable to a satisfactory extent.

4.1. Introduction

Recent years have seen increasing volumes of climate information produced and stored, driven by satellite data and results from numerical climate models (Overpeck et al. 2011). This makes data analysis based on machine learning possible, including data-driven modeling of phenomena in the climate system (Knüsel et al. 2019). Machine learning is often said to be useful for predictions of complex ill-understood phenomena (at least under some conditions, see Pietsch 2015; Knüsel et al. 2019; Northcott 2019). However, climate scientists aim not only at predicting phenomena but also at understanding them. Whereas process-based climate models can be useful for understanding phenomena (Parker 2014), it is unclear whether and under what conditions data-driven models can provide understanding. In fact, skepticism is often expressed about the fitness of data-driven models for understanding and explaining. For example, López-Rubio and Ratti (2019) argue that the complexity of machine learning models, which generally increases with the models' predictive skill for complex phenomena, impairs their intelligibility and hence, their usefulness for yielding mechanistic explanations. In contrast, Sullivan (2019) argues that the usefulness of machine learning models for understanding is primarily impaired by what she calls "link uncertainty", i.e., a lack of evidence linking the model to the target system. Still, Sullivan (2019) argues that it can be possible to reduce this link uncertainty and successfully use machine learning models for understanding.

In this paper, we address the fitness of data-driven models for providing understanding. We do so by first clarifying in a general way what criteria determine the fitness of a model for providing understanding. As process-based climate models are routinely used to obtain understanding of phenomena (see Parker 2014), we illustrate the application of the framework to process-based climate models of different complexities. We then apply these criteria to data-driven models and compare their fitness as vehicles for understanding to that of process-based models. We argue that at first glance, there seems to be a dilemma to data-driven models: While they are fit for providing understanding of some simple phenomena, in these cases, researchers typically have sufficient background knowledge to construct process-based models and hence do not need data-driven models. In other, more complex cases, the lack of background knowledge and the lack of model intelligibility impair the fitness of data-driven models for providing understanding. Thus, stated boldly, data-driven models seem either unnecessary or inadequate for understanding. We go on to show that, while intuitively plausible, this is a false dilemma, which we illustrate using a case study from climate science where a data-driven model was successfully used to obtain understanding. Generalizing the insights from this example to other cases, we conclude that data-driven models can be useful for understanding phenomena under certain conditions.

The remainder of this paper is structured as follows. In section 2, we clarify the distinction between process-based and data-driven models. In section 3, we introduce a

framework to assess to what degree a model can provide understanding of aspects of a target system, which builds upon three dimensions: the representational accuracy and the representational depth of a model, and model graspability. We illustrate the application of this framework in section 4 for a simple energy-balance model of the global climate system but also discuss state-of-the-art global climate models. In section 5, we assess the potential of data-driven models for providing understanding, which we compare to the fitness of process-based models for this purpose. We conclude in section 6.

4.2. Process-Based and Data-Driven Models

The focus of the present paper is on data-driven models as opposed to process-based models. Process-based models are mathematical models that explicitly represent with equations processes taking place in the target system. Examples include state-of-the-art climate models, general equilibrium models in economics, and the Lotka-Volterra model in ecology. While the equations of process-based models are often derived from theory, they need not necessarily be so (as argued by Weisberg 2013, chap. 1, this was the case for the original formulation of the Lotka-Volterra model in ecology). While in principle, pen and paper suffice to formulate and use a process-based model, in many applications the models are implemented on a computer as a simulation model. Using a computer is necessary when analytical solutions for the model equations are out of reach or when the problem at hand is too complex to analyze the model equations directly, e.g. because of its temporal and spatial resolution (Parker 2014).

Data-driven models, in contrast, are built with machine learning. Note that machine learning can be used for a variety of purposes. Modeling phenomena, e.g. in order to make reliable predictions of new cases, is only one of them. Other purposes are e.g. to explore a dataset, and to find patterns and associations between variables. A variety of machine learning algorithms exists, ranging from simple tools such as linear regression and LASSO regression, a linear regression technique that performs a regularization that selects the most important variables automatically, to complex non-linear artificial neural networks, including deep learning (James et al. 2013; Reichstein et al. 2019). Generally, there is a trade-off between the flexibility of machine learning algorithms and model interpretability (James et al. 2013). Flexibility refers to the ability of a machine learning algorithm to extract complex, non-linear relations between variables. Interpretability refers to how much insight a model allows a user into its inner workings. In this paper, we focus on algorithms that lie on the more flexible and less interpretable end of this spectrum, which are often non-parametric methods (for a philosophical discussion of non-parametric machine learning models, see Pietsch 2015). More on interpretability and related terms will be said below in section 4.3.2.

Hence, data-driven models of phenomena are not constructed with equations or other explicit representations of processes but by training a machine learning algorithm, which we refer to as "data-driven modeling of phenomena". Training refers to the step of algorithmically learning how to predict the values of the dependent variables from the set

of independent variables. We use the term "data-driven model" only for this trained model, not for the machine learning algorithm prior to the training step. Note that data-driven models should not be confused with models of data (for a discussion of models of phenomena and models of data, see Frigg and Hartmann 2012). In this paper, the focus is on data-driven modeling that relies on supervised machine learning,[17] which is a set of methods for datasets that consist of labeled samples of independent and dependent variables. An algorithm is used to learn generalizable rules that allow to predict the dependent variable based on independent variables for new samples with an unknown value of the dependent variable.

The use of data-driven models has two main advantages compared to process-based models. First, running an already trained data-driven model is usually inexpensive from a computational perspective. Second, training a data-driven model is possible also when scientists do not have sufficient process understanding to construct a process-based model. This is because in principle it suffices to be able to specify which variables are potentially important for producing a phenomenon without knowledge of their relative contributions and the processes responsible for the connections (Knüsel et al. 2019). While it is undisputed that data-driven models can be useful for predictions, at least under certain conditions, there is skepticism about their usefulness for understanding phenomena as outlined in the introduction (see López-Rubio and Ratti 2019; Sullivan 2019).

In this paper, we draw a sharp distinction between process-based and data-driven models, but we note that in practice the distinction may not always be clear. For example, state-of-the-art Earth system models in climate science are process-based models as far as e.g. the large-scale flow dynamics are concerned. However, empirical parameterizations are used e.g. to represent cloud formation or vegetation in the form of plant functional types. These parameterizations have a phenomenological character that is similar to that of data-driven models. Furthermore, climate models exist in which some parameterizations have been replaced by machine learning (e.g., Gentine et al. 2018). These approaches further blur the line between process-based and data-driven models.

4.3. Models and Understanding

In recent years, understanding has received increasing attention from philosophers of science and has been recognized as an important epistemic aim of science (de Regt 2017; Dellsén 2016). Different accounts of scientific understanding exist. These accounts require different characteristics from theories or models and from cognitive agents for understanding, such as that the theory is factive (Strevens 2013), that the agent has certain abilities in handling the theory (de Regt 2017), or both (Wilkenfeld 2017). In the

[17] Besides supervised learning, there are unsupervised machine learning algorithms, which do not require labeled output data. Instead, patterns in the datasets are detected. Examples of unsupervised learning are clustering algorithms and principal component analysis.

following, we draw on this literature and develop a framework that allows to assess to what extent a model is fit for providing a user with understanding. Hence, the framework allows to perform an adequacy-for-purpose assessment where the purpose is understanding (however, we use the term "fitness" instead of "adequacy" because fitness-for-purpose is a matter of degrees, see Parker, forthcoming).

The focus of the present paper is on understanding of phenomena. Examples for this kind of understanding are when an agent understands global warming, cloud formation, or the ice-albedo feedback. Understanding of a phenomenon is typically related to having an explanation of the phenomenon (de Regt 2017; Baumberger, Beisbart, and Brun 2017). This kind of understanding would be attributed to people who can explain why global warming occurs, how cloud formation works, etc.[18] Hence, a model that is fit for providing this kind of understanding provides the model user with explanatory information that enables her to construct an explanation of the phenomenon.[19] This is possible, for example, by highlighting which causal factors are most relevant for producing a phenomenon or how different causal factors interact in producing a phenomenon.

The framework for assessing the fitness of a model for providing understanding acknowledges that understanding comes in degrees and takes understanding to be a multidimensional concept (see Baumberger 2019; Wilkenfeld 2017). We take the fitness of a model for understanding a phenomenon to depend upon the three dimensions of representational accuracy, representational depth, and graspability. Representational accuracy and representational depth concern the relationship between the model and its target, and graspability concerns the relationship between the model and its user. Thus, we suggest that the extent to which a model M is capable of providing a user S with understanding of a phenomenon P in target T depends on (a) how accurately M represents T for an account of P, (b) how graspable M is for S, and (c) how comprehensively M represents the processes producing P. We want to emphasize here that these dimensions and the respective criteria should not be taken to be necessary conditions for understanding but evaluative criteria to assess how good the understanding is that can be obtained with a given model. It depends on the context how well a model needs to perform with respect to the three dimensions in order to be capable of providing a user with outright understanding. Typically, in a research context, representational accuracy and depth might outweigh graspability, whereas in an educational context graspability might well need to be higher than in a research context. In the following, we discuss the fitness-for-purpose of models for a general context of scientific research.

---

[18] We leave open whether agents can understand a phenomenon without having an explanation of it (see Lipton 2009) because we will show that data-driven models can enable the construction of explanations, and, hence, can be fit for providing understanding even in this stronger sense of understanding.

[19] Note that we use the term "explanatory information" in a more restrictive sense than Parker (2014). We refer to information as "explanatory" if it allows a scientist to construct to an explanation of a phenomenon.

In subsections 4.3.1, 4.3.2, and 4.3.3, we introduce these dimensions and the respective evaluative criteria to assess how well a model fares with respect to the dimension.

4.3.1. Representational Accuracy

Representational accuracy of a model is the degree to which the model is similar to its target in relevant respects (Giere 2004; Wilkenfeld 2017). When the goal is understanding, the relevant respects will often be related to causal processes that are potentially relevant for producing the phenomenon under investigation. If they are accurately represented, the researcher can obtain information via the model that aids the construction of how-possibly or how-actually explanations of the phenomenon (Parker 2014). For instance, if the model generates a phenomenon very similar to that observed in the target, this suggests that the causal factors represented in the model are sufficient for producing the phenomenon (irrespective of whether they in fact are responsible for the actual instances of the phenomenon thus far observed). Likewise, running the model with a process "turned off" can reveal that the process is necessary for the production of the phenomenon – because it no longer appears in the simulation – at least in the presence of the other causal factors represented in the model.

The accuracy of a model's representation of particular causal processes is not directly accessible and needs to be justified indirectly. Our framework offers three evaluative criteria that allow to assess the representational accuracy of a model, which are based on Baumberger, Knutti, and Hirsch Hadorn (2017) and Baumberger (2019). These three criteria are the coherence of a model with background knowledge, the empirical accuracy of relevant model results, and the robustness of model results. Below, we introduce these three criteria and explain why they can be used to evaluate the representational accuracy of a model. The three criteria can neither individually nor jointly guarantee that a model is representationally accurate. Rather, they are gradual and provide more or less strong non-deductive reasons for thinking that a model has a certain degree of representational accuracy. They should thus be seen as indicators of representational accuracy.

Coherence with background knowledge

To what degree is the model as an account of the phenomenon under investigation coherent with background knowledge and assumptions? A direct comparison of the inner workings of the model with the inner workings of its target is not possible because the target's inner workings are generally not directly accessible. Hence, the inner workings of the model are instead compared to available background knowledge and assumptions. The considered background knowledge can include anything from approximately true fundamental physical laws such as conservation of energy to well-established empirical relationships such as the near-linear relationship between total carbon emissions and temperature change.

Empirical accuracy

How well do model results relevant for the phenomenon under investigation resemble the states of the target system as depicted in observational and observation-based data of sufficient quality? Empirical accuracy indicates whether the model behaves approximately the way it is expected to. However, not all instances of empirical accuracy provide equally strong reasons for thinking that the model is representationally accurate. A good fit of model results to observations provides stronger reasons for the representational accuracy of the model if the observational data was not previously used for model tuning or training (Frisch 2015). Hence, use-novel data has a special status in model evaluation, for example in cross-validation, even though this does not mean that double-counting of data for model construction and evaluation is impermissible (Steele and Werndl 2016). Note that the argument from empirical accuracy to representational accuracy alone can be weak due to the problem of underdetermination. Even in combination with the other criteria, the argument from empirical accuracy to representational accuracy is not conclusive.

Robustness

To what degree are model results relevant for the phenomenon under investigation dependent on the specific model implementation? This can often be assessed by checking whether model results agree with the outputs from other models (Weisberg 2006). If models share a common core and agree on some hypothesis, then this can provide evidence for these core causal assumptions in the model (see Baumberger, Knutti, and Hirsch Hadorn 2017; Lloyd 2010; Weisberg 2006). In order for this agreement to increase our confidence that the model is representationally accurate in the relevant respects, we need to believe that the other models can serve as benchmarks because they are themselves sufficiently representationally accurate or that it is unlikely that the agreement would occur even though the models were not accurately representing the relevant aspects of the target (e.g. due to shared biases). This is an important caveat in climate modeling due to recognized model interdependence (see Parker 2011). Robustness considerations can be especially important when little data is available to assess empirical accuracy.

4.3.2. Graspability

A common view holds that in order to understand a phenomenon with the help of a theory or model, an agent needs to grasp the theory or model to some degree. What it means to grasp a theory or model to a relevant degree is usually spelled out in terms of certain abilities, such as the ability to make use of the theory or model, and hence, different authors associate or even equate understanding with these abilities. The most prominent suggestion along these lines is due to de Regt and Dieks (2005). It states that a scientist needs a theory that is intelligible in order to use it as a vehicle to understand a phenomenon, where intelligibility is the value that scientists attribute to the cluster of qualities of a theory (e.g. simplicity, scope, familiarity, causation, mechanism, and

visualizability) that facilitate the use of the theory. De Regt and Dieks suggest that a sufficient condition for the intelligibility of a theory for a scientist is that the scientist can estimate qualitatively the consequences of the theory without performing any calculations (see also de Regt 2017). We propose an adapted version of this suggestion, namely the ability to qualitatively anticipate model outputs. However, as Lenhard (2006) has pointed out, it is possible to gain this ability without being able to explain the behavior of the model by simply familiarizing oneself sufficiently with the model. Since the ability to further explain model behavior would increase the graspability of the model, we suggest that this ability is a second criterion to assess the graspability of a model.

Ability to qualitatively anticipate model outputs

To what degree can model outputs be anticipated by the user without performing calculations or running a simulation of the model? That this ability is important for understanding has been argued by de Regt and Dieks (2005). It holds that if a model user accumulates experience with a model, she can learn to anticipate how the model behaves in response to changing inputs. This notion has also been suggested by Lenhard (2006) for simulation models.

Ability to explain model behavior

To what degree can the behavior of the model be explained by the user? This aspect of graspability has been the focus of philosophical research on computer simulations. It has been argued that computer simulations are epistemically opaque (Humphreys 2004, 2009) and that it is difficult to attribute the reasons for successes and failures of climate model simulations to specific submodels (Lenhard and Winsberg 2010). All else being equal, if a model allows the user to explain its behavior, the model is fitter as a vehicle for understanding. If the model is also representationally accurate to a sufficient degree, explaining model behavior allows a user to explain the behavior of the target system to some extent as well.

Obviously, model graspability does not only depend on characteristics of the model but also on a specific model user. Here, we focus on characteristics of the model and lay out general considerations that are relevant for a versed model user. While for representational accuracy, the three criteria from above only provide more or less strong reasons to assume representational accuracy, performing well in terms of the two criteria considered here constitutes grasping. Hence, the evaluation of the extent to which a user actually grasps a model is more direct and certain compared to the evaluation of its representational accuracy.

4.3.3. Representational depth

Representational depth is defined in terms of the level at which a model describes the processes producing the phenomenon that is to be understood. A representationally deeper model describes a phenomenon not only on a phenomenological level but also

describes the lower-level mechanisms producing the phenomenon and hence provides in this sense a more comprehensive representation of the processes. Therefore, a representationally deeper model generally allows for more mechanistic understanding, which we take to be better than mere phenomenological understanding about how changing inputs relate to changing outputs.

Representational depth becomes relevant only when discriminating between two models that describe the same target but do so at different levels of description. This is for example the case when comparing two climate models with different spatial resolutions that are both used to study the phenomenon of global warming. The model with the higher resolution offers a more complete representation of the processes that produce global warming. This is because more processes need to be represented at a lower level of description as a result of the increased resolution. An example of such processes relevant for understanding global warming is the formation of clouds that can be more comprehensively represented in a model with higher spatial resolution. In the remainder of this article, we will compare a process-based and a data-driven model that describe the same phenomenon at the same level of description and, hence, have the same representational depth. This ensures that the comparison of the two models is a fair one. Thus, we will discuss models mainly in terms of the other two dimensions, representational accuracy and graspability, but the dimension of representational depth is required to ensure a fair comparison. We will briefly return to representational depth in section 4.4.2, where we discuss state-of-the-art climate models.

4.4. Understanding with Process-Based Climate Models

There is a general intuition that process-based models are useful tools for understanding phenomena. In this section, we illustrate that the framework introduced in the previous section does justice to this intuition. We do this by first presenting the example of a simple, zero-dimensional energy-balance model in subsection 4.4.1 and then extend the discussion to state-of-the-art climate models in subsection 4.4.2.

4.4.1. Zero-Dimensional Energy-Balance Model

In this subsection, we use the example of a simple, zero-dimensional linear energy-balance model and discuss how well it performs with respect to the first two dimensions of the framework introduced above. We discuss the example of the same type of hypothesis test of the causes of 20th century global warming as discussed by Parker (2014). The model is based on just one linear differential equation:

$$C \cdot \frac{dT}{dt} = F - \lambda \cdot T \qquad (1)$$

In equation (1), C denotes the heat capacity of the Earth's climate system, T denotes the global mean surface temperature perturbation (relative to some baseline), t denotes time, F is a term capturing a linear combination of all radiative forcing factors, and λ is a

constant feedback parameter. Variations of this model have been used and discussed extensively in climate physics (e.g., P. M. Forster et al. 2013; Knutti and Rugenstein 2015; for an overview, see Knutti, Rugenstein, and Hegerl 2017). The model equation prescribes that any change in the total heat content of the Earth's climate system must equal some positive radiative forcing (e.g., increased CO_2 in the atmosphere) minus energy that leaves the climate system due to a response in the radiative budget, parameterized here as linear feedback. The values of C and λ were identified by calibrating the model to data for the years 1931 to 1980. Details about model parameters and data are provided in the appendix.

Using this model, it is possible to run a hypothesis test for the causes of 20th century global warming. The radiative forcing factors are usually classified as either natural or anthropogenic depending on their origin. Natural forcing factors considered here are changes in solar irradiance and aerosols from volcanic activity. Anthropogenic forcing factors include greenhouse gases, aerosols, ozone, black carbon, and land use (Myhre et al. 2013). Here, three important greenhouse gases are considered, namely CO_2, CH_4, and N_2O. Now, suppose we want to determine whether anthropogenic factors caused the measured increase in global mean surface temperature over the 20th century. This question can be addressed by comparing two model simulations. In the first, both natural and anthropogenic radiative forcing factors follow their actual values over time; in the second, anthropogenic forcing factors are kept constant at their preindustrial averages, and only natural forcing factors evolve according to historical records.

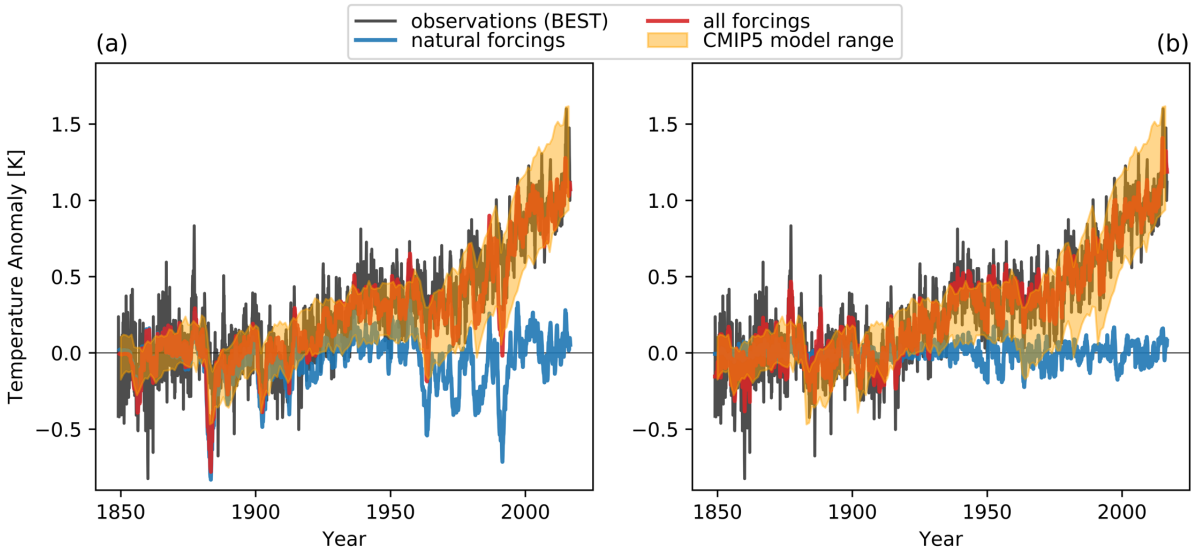


Figure 1: Simulation runs of the energy-balance model (a) and of the data-driven model (b) for the scenario with all forcing factors corresponding to historical observations, and for the scenario where anthropogenic forcing factors are held constant at their preindustrial average values. Temperature anomalies are relative to 1851 – 1880.

The results of the simulations for these two scenarios are displayed in Figure 1(a). The blue curve displays the simulation in which only natural forcing factors correspond to historical observations. It is unable to reproduce the evolution in 20th century global

mean surface temperature. In contrast, the simulation run with all forcing factors following historical records generally tracks the observations closely, if appropriate values for the feedback and heat capacity are chosen. The extent to which these results provide understanding depends on the fitness of the model to serve as a vehicle of understanding, which can be assessed using the framework introduced in the previous two subsections.

Coherence with background knowledge

The model only consists of one equation. For the model to be representationally accurate for this type of hypothesis test, the equation needs to consider all relevant causal factors and needs to adequately reflect the relationships between them. In order to argue for this, according to the framework, the equation needs to be assessed in terms of its coherence with background knowledge. The model states that in equilibrium, incoming shortwave radiation and outgoing longwave radiation balance each other out, and is hence based on conservation of energy. Equation (1) emerges when approximating a conservation of energy equation with a Taylor expansion under the assumption that the feedback parameter is constant and that higher-order expressions can be ignored (see Knutti and Rugenstein 2015). Thus, the equation is based on assumptions that idealize certain properties of the real-world climate system for an account of global mean temperature change. As Knutti and Rugenstein (2015) argue, the assumption of a constant feedback parameter only holds within limits since many feedbacks are state-dependent. However, here, the model was only used to reproduce historical temperature records. Furthermore, there are some uncertainties in estimating the radiative forcing F. Hence, the assumption of a constant feedback parameter and a constant heat capacity is probably not a problematic idealization. The factors considered in F were aggregated based on their radiative forcing. Thus, the model is coherent with well-established background knowledge, at least to a certain extent.

Empirical accuracy

In order to assess the empirical accuracy of the results of the energy-balance model, model results (red) need to be compared to observations or observation-based data (black curve in Figure 1(a)). The two curves are in good agreement, meaning that the model is empirically accurate. Some deviations are apparent, especially for the time before 1950 where the model exhibits less variation. This can partly be explained by modes of internal variability, i.e., factors that are related to the chaotic nature of the climate system and cause natural internal fluctuations (see Katzav and Parker 2018) for which no data was available for the time before 1950. Furthermore, uncertainties in forcing and observed warming for this early period may contribute to the deviations. Hence, the deviations should not be a major reason for concern.

In model creation, the values of some parameters cannot be fixed based on background knowledge. Instead, they have to be inferred from observations through so-called model calibration or model tuning (Frisch 2015). Most authors hold that for model evaluation, data should be considered that has not been used for model calibration (see Parker 2018 for an overview). Specifically, Frisch argues that if complex climate models have

predictive success (i.e., success in reproducing independent observations), this provides better reasons to think that they represent fundamental processes accurately than does achieving the same success through tuning/calibration. Only the observations from 1931 to 1980 displayed in Figure 1(a) were used for model calibration. Hence, especially the model's good performance outside of this range gives us some confidence that it represents the climate system sufficiently accurately for our purposes.

Robustness

In order to judge the robustness of model results, one needs to assess the extent to which model results correspond to the results of other models. For this, the results from the energy-balance model can for example be compared to the results from the Coupled Model Intercomparison Project (CMIP5 members) (plotted as a yellow area in Figure 1(a)), which is an ensemble of state-of-the art climate models. This reveals that the energy-balance model tracks the spread of the CMIP5 members closely, hence the results of the energy-balance model are robust with respect to CMIP5 models. Due to the different modeling approach in the construction of the energy-balance model and of the CMIP5 models, shared biases seem rather unlikely.

Ability to qualitatively anticipate model outputs

A model user can familiarize herself with a simulation model by way of systematically varying the inputs into the model and observing the outputs. It is certainly possible to learn about the behavior of the energy-balance model in this way. Furthermore, as the model is comparatively simple, a model user versed in mathematics will be able to anticipate model outputs qualitatively if she knows details about the development of the factors contained in the expression F in equation (1), i.e., the radiative forcing factors.

Ability to explain model behavior

The final criterion to consider is a model user's ability to explain model behavior, which, again, is dependent on the specific model user. In a simple case like the energy-balance model discussed here, it is possible to explain much of the behavior based on equation (1). For example, as concentrations of greenhouse gases in the atmosphere rise, so will the radiative forcing imposed by them, which is entailed in the values of F. As the climate system approaches a new equilibrium, it will balance the excess energy input through a rising temperature, which leads to a larger feedback term $\lambda \cdot T$.

Thus, there are indications that the simple energy-balance model provides a representation of the climate system for an account 20th century global mean surface temperature that is accurate to a satisfactory degree. Also, the model can be grasped to a satisfactory degree by a versed user. Hence, based on the framework we conclude that the model can be considered reasonably fit to serve as a vehicle for (phenomenological) understanding of 20th century global mean temperature evolution.

4.4.2. State-of-the-Art Climate Models

In this section, we briefly discuss to what extent and why our framework also allows for more complex state-of-the-art climate models to provide understanding. As Parker (2014) discusses, general circulation models are routinely used to provide understanding of phenomena in climate research, and their use for this purpose rests mainly on the assumption of their representational accuracy. As outlined above, representational accuracy here means that all candidate causal factors (i.e. candidate causes of the phenomenon to be understood) are included and the relationships between them are sufficiently adequately modeled. For example, complex climate models can be used for the same types of hypothesis tests that we have discussed in the previous section (Parker 2014). State-of-the-art climate models routinely consider more causal factors relevant for 20th century global warming than the energy-balance model discussed above, such as additional greenhouse gases from industrial processes and black carbon. Thus, they can be considered more representationally accurate for an account of 20th century global warming than the simple energy-balance model. However, if all of these factors were included in the energy-balance model, too, it would be similarly representationally accurate for an account of 20th century global warming.

State-of-the-art climate models differ from the simple energy-balance model in terms of their graspability. As Lenhard and Winsberg (2010) have argued, the complexity and layered development histories of climate models can makes it difficult to attribute reasons for model success or model failure to individual model components or submodels. This is a drawback of complex climate models in terms of graspability because it limits the ability of the model user to explain model behavior. This difficulty in explaining model behavior makes clear why climate scientist Isaac Held (2005) states that there is a gap between the ability of (process-based) climate models to accurately simulate the climate and scientists' understanding of the models. Hence, the tendency to explicitly resolve more processes comes at the expense of model graspability.

State-of-the-art climate models also differ from the energy-balance model in that they have a larger degree of representational depth. This means that they represent phenomena on lower levels of description, which is why they can, in principle, provide more mechanistic understanding by revealing how, in a mechanistic sense, a phenomenon is produced. For example, complex climate models allow a model user to see how increasing levels of greenhouse gases have changed the temperatures over land and over the ocean, and how this in turn has led to an increase in global mean surface temperature. Hence, in principle, they allow for a better understanding of the phenomenon of global warming. However, in practice, as argued above, it can be difficult to attribute simulated phenomena to specific parts of the model.

Representational depth can be understood as a kind of "vertical completeness" because it is defined through the comprehensive representation of the processes that produce the phenomenon of interest. Besides this "vertical completeness", which specifically concerns the description of one phenomenon, state-of-the-art climate models are also more

complete than the energy-balance model in a second, "horizontal" sense because they describe many additional processes in the target system. This "horizontal completeness" is not relevant for understanding a given phenomenon and is hence not part of our framework. It does, however, generally make one model applicable to a range of different phenomena and can also make it adequate for predictive purposes. An example for this form of completeness is that state-of-the-art climate models may represent the melting of the Greenland ice sheets, which is not directly relevant for producing 20th century global warming. Hence, while this "horizontal completeness" is not directly relevant for accounts of specific phenomena, it makes a specific model more broadly applicable.

Hence, even though graspability is a problem for state-of-the-art climate models, they are still reasonably fit for providing a versed user with understanding of a wide range of phenomena, assuming something like the causal-hypothesis-testing approach to advancing understanding. This fitness stems from the accuracy with which many different processes are represented and from the representational depth of the models for accounts of a range of phenomena, which allows for mechanistic understanding. Hence, state-of-the-art climate models generally have a higher degree of representational depth than simpler models like the energy-balance model, and they are often also representationally more accurate for accounts of specific phenomena. However, the energy-balance model fares better than state-of-the-art climate models in terms of graspability. This tendency reveals a dependence between the three dimensions of understanding. They are independent in a preferential sense because, in principle, a model that is representationally accurate, representationally deep, and graspable to a high degree would be preferable to one that only performs well with respect to one dimension. However, the dimensions do not seem independent in a statistical sense because, in practice, representational accuracy and representational depth often run counter to model graspability (for the difference between preferential and statistical independence, see Eisenführ, Weber, and Langer 2010). This trade-off is the reason why idealizations in models do not always reduce their fitness for providing explanations: although idealizations reduce the model's representational accuracy or depth, they can increase the graspability of the model (see Jebeile and Kennedy 2015).

4.5. Data-Driven Models and Understanding

In this section, we discuss the fitness of data-driven models as vehicles for understanding phenomena in the climate system and compare it to that of process-based models. We start with an example in subsection 4.5.1 and compare it to the energy-balance model from subsection 4.4.1. In subsection 4.5.2, we generalize the insights from the example to other cases of data-driven models and discuss the alleged dilemma of data-driven models. Finally, in subsection 4.5.3, we use an example from climate research to show that the alleged dilemma of data-driven models is a false dilemma and make some comments on how data-driven models can best be used as vehicles for understanding climate phenomena.

4.5.1. An Illustrative Example

We again take up the example of hypothesis testing of the causes of 20th global warming discussed in subsection 4.4.1. As we have mentioned in section 4.3.3, we compare two models that describe the same phenomenon at the same level of description – i.e., two models with the same representational depth. This ensures a fair comparison. The example will inform the general discussion about the role of data-driven models for obtaining understanding. As discussed in section 4.2, there is a general trade-off between the flexibility and the graspability (interpretability) of a machine learning algorithm. At the same time, the skepticism regarding the role that data-driven models can play in understanding phenomena stems partly from the lack of graspability. Hence, for the example here, we use a data-driven model constructed using the random forest algorithm (Breiman 2001) because it lies on the flexible and non-interpretable end of the spectrum (see James et al. 2013). Furthermore, random forest is an algorithm that is used in environmental science (e.g., Gudmundsson and Seneviratne 2015). It is a machine learning algorithm that uses subsets of the data to create many individual regression (or decision) trees and averages their vote. It starts by creating random subsets of the data (so-called bootstrapping). Then, a tree is trained on each subset. Each of these trees aims to predict the dependent variable based on the independent variables. However, at each split in the decision tree, only a random subset of all the variables is considered, which helps to train trees that are decorrelated. Once many trees are trained, the prediction of the dependent variable is made by averaging the predictions from all the individual trees (so-called aggregation). This combination of bootstrapping and aggregation is referred to as "bagging".

We trained a random forest model with data on the anthropogenic and natural forcing factors discussed in the example in subsection 4.4.1 as well as with modes of internal variability. As the dependent variable, global mean surface temperature was used. Details are provided in the appendix. Then two simulations were run in analogy to the example in subsection 4.4.1. The model results are displayed in Figure 1(b). At first glance, they look similar to the ones obtained from the energy-balance model, as only the red curve considering all forcing factors tracks observations closely. Hence, the question emerges whether the same degree of understanding can be obtained from the random forest model as was obtained from the energy-balance model. In order to know to what extent the model is fit to serve as a vehicle for understanding the causes 20th century global warming, the random forest model has to be assessed with the framework introduced in section 4.3.

In Table 4, we compare the fitness of the energy-balance model and the random forest model for providing understanding of the observed warming. As can be seen, the required considerations for empirical accuracy, robustness, and the ability to qualitatively anticipate model outputs are identical or very similar between the two models: not only do they fare similarly well with respect to these criteria, the considerations necessary to perform the evaluation are also similar. One difference is that the ability to qualitatively

anticipate model outputs can mainly be gained by manipulating the model in the case of the random forest model. We note here that it is even more important in the case of the data-driven model to evaluate the empirical accuracy of the model with novel data. Because the model structure crucially depends on the data, using the same data for model training and model evaluation would hide cases of overfitting. This is why in machine learning, researchers routinely split the datasets into training, test, and validation sets. A random split of the data was performed in order to have different training and test sets. This procedure employs use-novel data for model selection but double-counting occurs when the full model is evaluated as depicted in Figure 1(b).

Table 4: Comparison of the fitness of the energy-balance model and the random forest model of global mean surface temperature to serve as a vehicle for understanding.

dimension of understanding	**evaluative criterion**	**energy-balance model (process-based model)**	**random forest model (data-driven model)**
representational accuracy	empirical accuracy	• model reproduces observations well	• model reproduces observations well
	robustness	• model outputs are similar to CMIP5 models	• model outputs are similar to CMIP5 models
	coherence with background knowledge	• based on conservation of energy • idealizations seem justified for the case at hand	• model behavior consistent with background knowledge • model outputs consistent with background knowledge • all relevant variables considered • sufficiently flexible functional form used • sufficiently many configurations of the target system considered
graspability	ability to qualitatively anticipate model outputs	• familiarizing oneself with the model through manipulation • analysis of equation	• familiarizing oneself with the model through manipulation
	ability to explain model behavior	• analysis of differential equation • manipulation of model	• variable importance plot • manipulation of model • working of optimization algorithm

While the assessment of the two models in terms of three criteria is similar, the two model types differ when assessing the remaining two criteria, i.e., coherence with background knowledge and the ability to explain model behavior.

Coherence with background knowledge

As data-driven models do not explicitly incorporate background knowledge in the form of equations, the coherence of the models with background knowledge needs to be assessed indirectly. Different aspects can be addressed for this. First, model behavior, to the extent that it is accessible, can be checked for its consistency with background knowledge. This can, admittedly, be difficult in the case of many machine learning applications.[20] Second, model outputs can be checked for their consistency with background knowledge. Here, model outputs show no obvious violations of background knowledge. Third, the relevant variables, as judged from the point of view of background knowledge, should be considered in model development. This question is of importance for data-driven models because excluding causally relevant variables from the model can lead to a biased estimation of the contribution from other factors if the excluded factor is correlated to other input factors. Since understanding here depends on estimating the contributions of different factors to global mean surface temperature, the relevant causal factors should be included. In the example illustrated above, important natural and anthropogenic forcing agents were included (see Myhre et al. 2013). Fourth, the machine learning method used should be sufficiently flexible to model the relationships between the variables. Here, a bagging method was used that is generally very flexible. We add two notes of caution: using a very flexible algorithm to make it likely to capture the true dependencies comes with the drawback of lost model graspability, and it can lead to overfitting. Fifth, a sufficient number of configurations of the target system should be considered in the training dataset. The importance of this point has been stressed by Pietsch (2016, 138), who has claimed that data-intensive science "requires data covering all configurations of a phenomenon that are relevant with respect to a specific research question". Variation within one variable without covariation with the other variables is especially important. Here, to avoid the problem of correlated variables, the considered anthropogenic forcings were aggregated into one time series based on the respective radiative forcings. This step to decorrelate the variables makes it likely that sufficiently many configurations were considered. Hence, based on these considerations, the model seems coherent with background knowledge to a satisfactory degree.

Ability to explain model behavior

Whereas for the process-based models, some direct assessments of model behavior is possible, this is not straightforward for the data-driven model considered here. The reason for this is that random forest does not provide a set of rules or a model equation that could be analyzed. However, by manipulating the input and assessing the resulting outputs, one can conduct sensitivity analyses and learn about model behavior beyond the ability to simply anticipate model outputs and actually learn to explain how the model

[20] In the case of random forest, the variable importance plot could be assessed to learn about model behavior (see appendix). This shows which variables contribute most strongly to variation in the dependent variable. As this is precisely the understanding we are after in the example, we do not show and further discuss this here.

behaves.[21] Also, by knowing how the underlying machine learning algorithm works, it can be possible to know at least to some extent what drives model behavior. For example, in random forests, one can generally expect that sufficiently small variations in model inputs will not impact the values of the dependent variable due to the stepwise predictions. Hence, a versed model user can explain model behavior to some extent. But the data-driven model certainly performs worse compared to the process-based energy-balance model in terms of this criterion. We note here that advances in explainable artificial intelligence might further contribute to the graspability of data-driven models (for a more detailed discussion of different types of transparency of computational systems, including machine learning, see Creel, forthcoming).

Hence, despite the difficulties in explaining model behavior, we conclude that in the above example the data-driven model has a reasonable fitness to serve as a vehicle for understanding 20th century global warming. This is because the model performs similarly to a process-based model with the same representational depth in terms of empirical accuracy, robustness, and the ability to qualitatively anticipate model outputs. The obstacles for the fitness-for-understanding are the difficulty in explaining model behavior and in assessing the coherence of the model with background knowledge. However, at least the coherence with background knowledge can be assessed to a reasonable degree. Hence, while we acknowledge that the difficulties with respect to the two criteria can have an impact on the fitness of the model for providing understanding of phenomena, they do not seem sufficient to make the model in this example entirely unfit-for-purpose. However, it is yet unclear what the considerations from this example tell us about data-driven models more generally. We address this question in the following section.

4.5.2. Generalization: The Alleged Dilemma of Data-Driven Models

Constructing data-driven models does not require that all processes are quantitatively understood to the same extent that is necessary for constructing process-based models. Hence, it is possible to construct data-driven models of comparatively ill-understood phenomena. As seen in the previous section, data-driven models can be fit for providing understanding of phenomena in the climate system. They might therefore seem like a good choice of tools for achieving a better understanding of ill-understood phenomena. Thus, the question needs to be addressed to what extent the example from the previous section can be generalized. Generally, the evaluation of the empirical accuracy and the robustness of model results, and the ability to qualitatively anticipate model outputs would be very similar in other cases. Furthermore, the evaluation of the coherence of data-driven models with background knowledge will have to consider points similar to the ones presented in Table 4. However, in cases where model users have less

[21] Tools like the variable importance plot could further help here. We do not discuss this further as it tells about which the most important variables are, which is precisely the understanding we are after, here (see footnotes 20 and 22).

background knowledge, arguing from the coherence with background knowledge to representational accuracy is considerably weaker. Specifically, in more complex cases, the available background knowledge will often be insufficient to assess whether all relevant variables have been included and whether sufficiently many configurations of the target system have been considered. This problem also affects the first two points (in Table 4) about judging the consistency of model behavior and model results with background knowledge. Hence, in ill-understood cases, the coherence of a model with background knowledge can be a very weak argument as a justification of representational accuracy.

In more complex cases, there will often also be more difficulties in explaining model behavior compared to the example above. In such cases, model users might employ more flexible methods, e.g., models constructed with deep learning. These more flexible methods can be even less graspable for model users than the one presented above. The concerns about the lack of access to model behavior are especially relevant for models with a large number of independent variables.

Thus, when serving as vehicles for understanding, data-driven models can face several problems. First, the difficulty to explain model behavior can limit the graspability of the model. Second, when background knowledge is quite limited, the argument from coherence with background knowledge to representational accuracy is weaker. This also impacts the argument from empirical accuracy to representational accuracy due to the problem of underdetermination. This gives rise to an (alleged) dilemma of data-driven models in terms of their usefulness as vehicles for understanding. Namely, data-driven models can be fit for providing understanding of climate phenomena in simple, well understood cases. However, in these cases, scientists can typically construct and work with process-based models, whose evaluation in terms of coherence with background knowledge is more direct and hence, more certain, and which are more graspable. In more complex, ill-understood cases, it is not possible to construct process-based models. However, in these cases, the difficulty in grasping data-driven models and in justifying their representational accuracy seriously impairs their fitness for providing understanding. Thus, it seems that in simple cases there is no need for data-driven models since we can construct process-based models to provide understanding, and in more complex cases, where process-based models are out of reach, data-driven models are not fit for providing understanding. Stated boldly, data-driven models seem either unnecessary or inadequate for understanding. Hence, this indicates limited scope for data-driven models to provide understanding in practice.

4.5.3. Overcoming the Dilemma

If this dilemma holds, it restricts the role of data-driven models as vehicles for understanding to cases where computational cost is limiting or to didactical applications. However, while it might seem intuitively plausible, it is a false dilemma. This is because there are applications in practice where sufficiently restrictive background knowledge is available for scientists to distinguish between representationally accurate and

inaccurate models. At the same time, this background knowledge is insufficient for the construction of satisfactory process-based models. Such a case is presented by Andersen et al. (2017) in a paper labeled "Understanding the Drivers of Marine Liquid-Water Cloud Occurrence and Properties with Global Observations Using Neural Networks". The authors use satellite and reanalysis data and train multilayer perceptrons, a type of artificial neural networks, to reproduce cloud fraction and different cloud properties such as the optical thickness. The neural networks in the study have one hidden layer with five neurons. In terms of the criteria of our framework, this machine learning method has a similar graspability to the random forest model introduced above. The independent variables were chosen based on a review of other studies. They included the aerosol index, relative humidity and vertical vorticity at different pressure levels, boundary level height, and the lower-tropospheric stability. The authors chose not to train a single artificial neural network but instead to construct regionally specific models because some relationships were known to be regionally specific, e.g. pertaining to seasonal effects.

The resulting models achieved comparatively good empirical accuracy. Furthermore, the authors performed sensitivity analyses in which they systematically varied the values of one input variable while holding the others constant. In this way, they were able to learn to anticipate model outputs and, to some extent, to explain model behavior to some degree. Thus, the model was graspable to a satisfactory extent along both of the evaluative criteria for graspability introduced above. Finally, and most importantly, several bivariate relationships between individual predictor variables and the dependent variables were well constrained in the literature. Thus, the authors had sufficient background knowledge such that the evaluation of the coherence of the models with background knowledge gave strong arguments for the representational accuracy of the models. Nevertheless, while process-based models of such clouds exist, their usefulness for this kind of analysis is impaired by imperfect knowledge of the processes and computational costs.

Hence, this study shows that the dilemma introduced above is a false one, and that data-driven models can successfully be used as vehicles for understanding phenomena in the climate system, unless one claims that the study did not lead to new understanding. Can the authors convincingly argue that their models yielded new understanding of aspects of the climate system? Andersen et al. (2017) showed how different processes interact, and hence provided a better understanding of the formation of and processes within marine liquid-water clouds. E.g., they were able to show which predictors are the most important drivers of cloud fraction and estimate the individual contributions of different factors. Hence, this study provides explanatory information as introduced in section 4.3, namely information that can be used to construct explanations of cloud formation.

What more general points can be learned from this example? Data-driven models can be good tools to understand phenomena in the climate system (and potentially in other scientific fields) if researchers can take steps to increase the graspability of the model

and if their background knowledge is sufficiently large so that coherence with it provides a good argument for the representational accuracy of the model. One step that researchers can take is to restrict the set of independent variables based on their background knowledge of the phenomenon of interest. This increases the graspability of the models. If model users further have knowledge of some bivariate relationships (as was the case for Andersen et al. (2017)), sensitivity analyses or other techniques to explore the behavior of the data-driven model can help to both increase the graspability of the model and to evaluate to what extent it is coherent with background knowledge.[22] What is more, data-driven models can also potentially serve as a good starting point for understanding phenomena. For example, if a large dataset of some independent variables and a dependent variable is available, but a sufficiently flexible machine learning algorithm fails to accurately represent the phenomenon of interest, researchers know that processes not represented in their model must be relevant, and that additional variables are potentially relevant. If two data-driven models are compared that differ only because one of them also considers a specific variable that is not considered in the other model and the first of the two models is much more empirically accurate, this can give scientists some understanding of the respective processes. A similar point about hierarchies of process-based climate models of different complexity was made by Katzav and Parker (2015).

Thus, we agree that both model interpretability (cf. López-Rubio and Ratti 2019) and the lack of evidence linking the model to the target (cf. Sullivan 2019) pose difficulties for the fitness of data-driven models for understanding. However, we argued here that neither problem necessarily precludes data-driven models from serving as vehicles for understanding in specific instances. Creating data-driven models in situations where sufficient background knowledge is available to argue from the coherence of the model with background knowledge to its representational accuracy can provide exactly the kind of evidence that reduces the link uncertainty discussed by Sullivan (2019). Furthermore, Krishnan (2019) argued that in many cases, the ability to explain model behavior is often not necessary for a machine learning model to be fit for a specific purpose. Based on our framework, we agree with this position for the purpose of understanding phenomena because the models can have a sufficient degree of graspability even if their behavior cannot be explained. Nevertheless, advances in explainable machine learning would generally increase the fitness of data-driven models for the purpose of understanding (see Creel, forthcoming).

---

[22] Note that evaluating the coherence of a model with background knowledge can also be possible for complex methods like deep neural networks, e.g., based on variable importance (see Gagne II et al. 2019). While graspability becomes a larger issue for these more complex models, the justification of representational accuracy can be possible.

4.6. Conclusion

In this paper, we have proposed a framework for assessing the fitness of climate models to serve as vehicles for understanding. This framework is built upon three dimensions of understanding, namely the model's representational accuracy, its graspability, and its representational depth. We introduced several evaluative criteria to assess how well a particular model performs along each of these dimensions. After applying the framework to process-based models, we considered an alleged dilemma for data-driven models, according to which they are either irrelevant or inadequate because they can only provide understanding for cases in which process-based models could more confidently be applied. Using a case study, we showed that this is a false dilemma. Hence, there is a role for data-driven models to play for researchers aiming to better understand climate phenomena.

We largely ignored the role that machine learning methods can play in applications other than representational models of phenomena. For example, unsupervised machine learning can be used to identify clusters in climate datasets (Zscheischler, Mahecha, and Harmeling 2012), i.e., as models of data. If one holds that identifying such groups constitutes understanding (see Gijsbers 2013), then machine learning can play a role for obtaining understanding that goes beyond the use for data-driven modeling of phenomena.

The concerns raised in this paper have consequences that go beyond data-driven models. They might also be relevant for more classical statistical modeling practices in the sciences. Furthermore, as discussed, the framework for assessing the fitness for the purpose of understanding applies equally to process-based climate models. The extent to which the representational accuracy of global climate models is impaired by empirical parameterizations and to which the complexity of the models impairs model users' ability to grasp the model are open questions.

Acknowledgments

We thank Hendrik Andersen, David N. Bresch, Roman Frigg, Gertrude Hirsch Hadorn, Reto Knutti, Wendy Parker, and Marius Zumwald for feedback on earlier versions of this manuscript. Furthermore, we are grateful to the participants of the workshop *Science and Art of Simulation 2019* in Stuttgart and *EPSA19* in Geneva, the participants of the Weekly Research Meeting of CHESS at the University of Durham, the participants of the PhD Seminar of the Department of Philosophy, Logic and Scientific Method at the London School of Economics and Political Science, and the participants of the workshop *Big Data, Machine Learning, Climate Modelling & Understanding* at the University of Bern for feedback on related talks.

Funding

This research was funded by the Swiss National Science Foundation National Research Programme Big Data (NRP 75) grant number 167215.

5. Conclusions and Outlook

Data generated worldwide is increasing in both volume and complexity. This does not only lead to technical, legal, and ethical challenges, it also opens up new opportunities, including for scientific research. This thesis provides some insights into what these opportunities consist in for environmental science and specifically for climate research. It addresses questions of predictions, uncertainty, and scientific understanding, and discusses what challenges researchers face when applying data-driven modeling techniques for these purposes.

In this chapter, I conclude by highlighting central findings and implications of the thesis and by providing an outlook to future research. The chapter is organized as follows. I first present the central findings emerging from chapters 2, 3, and 4 in section 5.1. Section 5.2 presents implications of the findings of the thesis for broader scientific and societal debates. I then present suggestions for further research in section 5.3 before ending with some closing remarks in section 5.4.

5.1. Central Findings

The central finding of this thesis is that neither extreme optimism nor extreme pessimism is warranted regarding the possibilities emerging from big data in general and data-driven modeling techniques in particular. It does not seem likely, based on the findings presented in this thesis, that all or most modeling in climate science will or should become data-driven for all purposes. Nor is it true that data-driven modeling is only useful for narrow short-term predictions. Rather, thanks to increasing volumes of data, scientists can add techniques for data-driven modeling to their toolbox and use them for a range of scientific tasks, including for long-range projections and for obtaining understanding of phenomena. Although new forms of data, e.g. from crowdsourcing or from social media, were not a central topic of this thesis, similar conclusions regarding these data sources were reached in chapter 2. Making good use of big-data elements in environmental and climate science requires that the research process be guided by domain-specific background knowledge. This helps to evaluate whether models and data are fit for the purpose to which they are put.

In chapter 2, we argued that big data does not affect scientific research in climate science in an all-or-nothing way. Big data is present in research in the form of individual elements and hence affects research to larger or smaller degrees. Big-data elements can help scientists to overcome two kinds of limitations they may face in their research. First, big-data elements can help them to model or measure a phenomenon when the financial, computational, or time resources are limiting a more classical approach. We referred to this as the rationale of efficiency. Second, big-data elements can allow scientists to model a phenomenon when their understanding is insufficient for a more theory-based modeling approach. We referred to this as the epistemic rationale. In chapters 3 and 4, we have discussed cases in which the epistemic rationale was the reason for using data-driven modeling techniques. We have shown that also in such cases, researchers can confidently make arguments from the coherence with background knowledge to the representational accuracy of the model. Thanks to this evaluation, data-driven environmental models can be useful for purposes such as making long-term projections and obtaining scientific understanding.

In chapter 2, we further showed that in most scientific applications, predictions made with big-data elements, i.e. machine learning and new forms of data, cannot be constantly evaluated against new data. Hence, in these situations, researchers need to assume the constancy of the identified relationship and need to justify this assumption. This justification can only be based on background knowledge. In chapter 3, we specifically addressed the question of how this constancy assumption can be justified by extensively discussing the case study of a data-driven environmental model. We have shown that the constancy assumption can be justified based on the accuracy with which the data-driven model represents important causal processes in the target system. This representational accuracy, in turn, can be justified by at least six points. These are (1) that most relevant variables have been included, (2) that data covering many configurations of the phenomenon of interest were included in the training dataset, (3) that sufficiently flexible machine learning methods were used, (4) that the results of the model are consistent with background knowledge, (5) that the model is empirically accurate, and (6) that the results are robust across different models. In chapter 4, the same six points were provided to justify that a data-driven model provides an accurate representation of a target system for an account of a specific phenomenon. While these aspects can be helpful to justify the representational accuracy of a data-driven model, we have seen in chapter 3 that they do not necessarily guarantee representational accuracy, nor, consequently, the constancy of the identified relationships. This is because the reconstructed arguments in chapter 3 were non-deductive.

These points about the representational accuracy of data-driven models are related to the external theory-ladenness of data-driven models introduced by Pietsch (2015). External theory-ladenness in Pietsch's account is explicated mainly in terms of the framing of the problem. Namely, Pietsch argues that when constructing machine learning models of phenomena, scientists need to include the right variables and have training data covering all states of the target system that are of interest. Furthermore, stable background

conditions and stable causal categories are required. If these four conditions hold, Pietsch (2015) argues, no further internal theoretical assumptions are needed for constructing successful machine learning models. Pietsch (2015; 2016) has specifically addressed disciplines where no well-established hierarchies of scientific theories are available. In disciplines like climate science where such theoretical background knowledge is available to a larger extent, we have shown that additional considerations are important, specifically the consistency of model results with background knowledge. As the modeled relationships are still not prescribed from theory, the evaluation of the consistency of model results with background knowledge leads to an external theory-ladenness of data-driven modeling that goes beyond what Pietsch has described.

The aspects to evaluate the representational accuracy of data-driven models in chapters 3 and 4 are important specifically in environmental and climate science because there are reasons to believe that often, not all of Pietsch's (2015) four conditions hold for environmental systems. This specifically concerns the requirement to have data covering all configurations of interest, and potentially also the requirement to have stable background conditions. Due to global environmental change, at least for long-term projections such as the selenium projections discussed in chapter 3, it seems unlikely that environmental scientists can train a data-driven model using data that covers all configurations or that they encounter stable background conditions. However, the aspects that we discussed in chapters 3 and 4 to evaluate models in terms of their representational accuracy can be a potential remedy that helps to confidently use data-driven models despite this extrapolation. This is also why, rather than emphasizing constant background conditions or data covering all configurations of interest, in chapter 2 we have emphasized the importance of the constancy of the identified relationships in models and measurements.

One of the recurring topics in the chapters of this thesis has been the role of background knowledge in model construction, model evaluation, and in the interpretation of model results. There is an apparent paradox emerging regarding of the role of background knowledge. On the one hand, we have argued in chapter 2 that there is an epistemic rationale for the use of data-driven models because they allow the modeling of phenomena that researchers are unable to model otherwise because of their lack of system understanding. On the other hand, we have stressed the importance of background knowledge for model evaluation in chapters 3 and 4. One might worry that this undercuts the epistemic rationale and reduces its importance in practical applications. It is this worry that gave rise to the alleged dilemma of data-driven models that we encountered in chapter 4. However, as the discussions in chapters 3 and 4 showed, the situation is less dramatic than it seems at first glance. There are cases, such as the case studies discussed in these two chapters, where data-driven models are constructed due to the epistemic rationale, yet the available background knowledge allows to make sufficiently strong arguments from coherence with background knowledge to representational accuracy.

In many debates, machine learning and data-science tools more generally are referred to as black boxes. Since many machine learning algorithms do not explicitly represent processes and because they are often applied to large datasets with data on ill-understood phenomena, this is understandable. However, Krishnan (2019) has cogently argued that model interpretability is only an instrumental goal, and the ultimate goals are often achievable without model interpretability. While Krishnan has made this point mainly for goals related to non-epistemic values, for example for the goal of non-discrimination, the findings of this thesis support her point of view for certain epistemic purposes. Namely, in chapters 3 and 4, we have shown that making long-term projections with and obtaining understanding from data-driven models is possible with data-driven models despite difficulties with model interpretability. The reason for this is that interpretability in the sense of explaining every predicted instance is not a requirement for an assessment of the representational accuracy of a data-driven model. A high-level transparency regarding how a computer system turns inputs into outputs is often sufficient for this kind of model evaluation (see Creel, forthcoming). In more complex data-science contexts, the lack of transparency and interpretability of the models and the model construction may be a serious issue (see Hutson 2018). However, in research contexts such as in environmental and climate science, the kind of interpretability of data-driven models that is required for epistemic goals is often achievable with existing tools, as the case studies in chapters 3 and 4 have shown.

5.2. Implications

One of the central findings of this thesis, as discussed above, is the importance of background knowledge in model construction, evaluation, and use. Especially for complex modeling tasks, this finding can be read as a call for interdisciplinary collaborations between data scientists and domain scientists. While in some simple cases, domain scientists are able to construct data-driven models on their own, this could be increasingly difficult with unstructured data or very large datasets in general. Modeling in these cases thus requires interdisciplinary collaboration, as has also been suggested elsewhere (Faghmous and Kumar 2014; Faghmous et al. 2014; Karpatne et al. 2017). These interdisciplinary collaborations can ensure that the models are physically plausible and to some extent interpretable by domain scientists (see Gagne II et al. 2019; McGovern et al. 2019). Hence, based on the findings provided here, attempts should be welcomed and encouraged that bring domain scientists and data scientists together to collaborate on projects. Such attempts are made, for example, by the Swiss Data Science Center.[23]

For the most part, in this thesis, I have addressed issues of data-driven models, i.e., applications of machine learning to model phenomena in scientific contexts. However, the findings of the thesis have implications for other applications of machine learning, too. This is particularly true of the framework for assessing predictive uncertainties

[23] See https://datascience.ch/academic-projects/, accessed on January 16, 2020.

introduced in chapter 3. In machine learning applications in which the algorithm is not interpreted as the representation of a phenomenon, the uncertainty can still be assessed using the argument-based framework introduced in this thesis. Applying this framework might be helpful, e.g., in image classification tasks by highlighting that there is second-order uncertainty because it is unclear to what extent the model can generalize or fall prey to adversarial examples. This insight can guide further research to reduce the influence of the factors that lead to this uncertainty. In other applications, for example when machine learning is used to assess the credit-worthiness of people or for predictive policing, the argument-based framework may yield useful insights into uncertainties that are relevant from an ethical perspective. Hence, the work presented in this thesis could provide good tools for ethical analyses of big data and machine learning. However, assessing how useful the argument-based framework is for these tasks requires further work.

In chapters 3 and 4, we have discussed the predictive uncertainty of data-driven models and the fitness of data-driven models to serve as vehicles for understanding phenomena. In both chapters, the frameworks used were quite general and could be applied to process-based models, too. In the case of the framework for understanding, an explicit discussion of process-based models was even provided. The differences between the two types of models consisted in how exactly the assessments were performed. This comparison reveals that the epistemological questions that arise with data-driven models concern issues that philosophers and scientists are familiar with because they arise in other, more traditional modeling activities, too. These are, for example, questions related to representational accuracy or model opacity. Thus, it seems that in many scientific contexts, the epistemological issues arising in the construction, evaluation, and use of data-driven models and machine learning more generally are best understood in the context of the existing literature on models and computer simulations.[24]

The evaluation of models in terms of background knowledge proposed in this thesis has implications for the evidence that data-driven models can provide for or against specific hypotheses. For example, Mazzocchi and Pasini (2017) have argued that climate model ensembles should be taken beyond dynamical, i.e., process-based, models and also include data-driven models. Pasini et al. (2017) have discussed such an application. Namely, the authors trained artificial neural networks to model global mean surface temperature dependent on different forcing factors similar to the application demonstrated with a random-forest model in chapter 4. As Pasini et al. (2017, 1) explain, "… to achieve robustness we need to obtain a common result from independent means of investigation (models) and GCMs [i.e., general circulation models] do not seem so independent from each other […]. Thus, attribution results from different approaches could be compared with GCMs' ones for understanding if we have robust results." In other words, investigating climate phenomena using data-driven models could,

[24] This is of course not true for developments that go beyond narrow applications of machine learning, i.e., efforts to develop general artificial intelligence (Goertzel and Pennachin 2007).

according to Pasini et al., provide independent evidence of the findings from process-based models. However, it is not prima facie clear to what extent data-driven models can provide independent evidence in such cases. For data-driven models to provide independent evidence, they would have to help exclude rival hypotheses for global warming (i.e., hypotheses that claim that anthropogenic greenhouse gas emissions are not the main factor driving global warming) that are consistent with results from process-based models (see Schupbach 2016). At the same time, the kind of reasoning employed here demands that the data-driven models be sufficiently fit for serving as vehicles for understanding. As the discussion in chapter 4 has shown, the fitness of data-driven models for this purpose requires that they provide an accurate representation of the climate system for an account of global warming. This representational accuracy is evaluated, among other things, in terms of the models' coherence with background knowledge, and this background knowledge is largely embedded in the available process-based models. Background knowledge that is used for the evaluation of data-driven models may also be obtained from the process-based climate models. If the data-driven models are coherent with this background knowledge to a sufficient degree, they are unlikely to be useful in robustness analysis because they probably cannot help to exclude rival hypotheses. If, however, the data-driven models are not coherent with background knowledge to a sufficient degree, then their fitness for providing understanding of the desired phenomenon will be rather low. Hence, while data-driven models are independent from process-based models in how they are constructed, it is doubtful that they can provide independent evidence for or against a hypothesis in such a robustness analysis. Yet, the independence of the two modeling types is an issue that deserves attention in future work.[25]

In the age of big data, researchers will increasingly be confronted with situations in which they have too many variables at their disposal that they can potentially include into data-driven models. Including all of them would likely lead to models that are partly based on spurious correlations (see Calude and Longo 2017). This is related to the problem of overfitting in machine learning. In such cases, a sensible variable selection procedure has to be implemented. While the problem of having too many variables available has not been explicitly discussed in this thesis, the uncertainty framework provided in chapter 3 has implications for how the uncertainty related to variable selection can be addressed. The selection of variables is one of the aspects that may have to be justified as it is relevant for the fitness-for-purpose of the model and, consequently, for the uncertainty assessment. Two approaches are possible for this. The first approach is to base the selection of variables entirely on researchers' background knowledge of the target system. This requires that their system understanding is good enough to decide

[25] Note that the robustness reasoning employed in chapter 4 of this thesis is different from the robustness reasoning employed by Pasini et al. (2017). In chapter 4, we recommend that robustness reasoning is used as a means to judge the causal core of the models that leads to robust properties. Hence, the robustness reasoning that we employed is not about accepting or rejecting hypotheses indicated by model results, it is about the models themselves (see Lloyd 2015).

which variables to include and which ones to exclude. The second approach is to base this decision on automated variable selection procedures such as LASSO or RIDGE regression (see James et al. 2013). In climate science, the latter approach has, for example, been employed by Sippel et al. (2020). Regardless of which of these two paths is chosen, it would appear in the framework presented in chapter 3 as a justification of the representational accuracy of the model, and it would itself require further justification. If the first path is deemed appropriate, the justification has to be provided by domain scientists based on domain-specific background knowledge. If the second path is chosen, data scientists and domain scientists may have to jointly judge the appropriateness of a given method for variable selection. Hence, deciding on which path to take may be a task that is best solved in an interdisciplinary manner. Second-order uncertainty emerges due to this decision if it is unclear how well justified the chosen approach is.

A final implication of the results here concerns scientific objectivity. In a recent account of scientific objectivity, Koskinen (2018) has suggested that something, be it a result, a method, or a scientific community, should be considered objective if it successfully averts the influence of our imperfections as cognitive agents, i.e., epistemic risks. Inductive risks are one type of epistemic risks that have to be managed, but other individual and collective biases are relevant, too. In chapter 3 of this thesis, we have argued that results from data-driven models can be fraught with second-order uncertainties because of the epistemic rationale that motivates the use of data-driven models, and because of model opacity. In cases with large second-order uncertainties, it might be difficult for researchers to assess the epistemic risks they face. Hence, second-order uncertainty could have implications for scientific objectivity in data-intensive science. However, this is a point that merits further investigation in the future.

5.3. Future Research

In chapter 2, we have concluded that machine learning provides a good set of tools for modeling phenomena when scientific understanding of the involved processes is insufficient for the construction of process-based models. We have then stressed the importance of background knowledge in justifying the assumption that the identified relationships remain sufficiently constant. I have noted in section 5.1 that this may seem like a paradox at first glance, but that there are cases where the representational accuracy of a model can be evaluated based on background knowledge even though no satisfactory process-based models can be constructed. Yet, more work is needed to understand the relationship between background knowledge needed to construct satisfactory process-based models and background knowledge needed to justify the representational accuracy of data-driven models. This issue could be addressed in future research by domain scientists, data scientists, and philosophers. In this assessment, non-epistemic values are likely to play a role. This is because a less conclusive justification of representational accuracy may suffice in cases where the consequences of wrong inferences are low. The larger the consequences of wrong inferences become, the more conclusive

the justification may have to be, requiring a better initial understanding of the target system.

In chapter 3, we have provided some thoughts on the quantification of uncertainties that builds upon the proposed framework, which analyzes uncertainty in a purely qualitative way. Chapter 3 specifically recommends using structured expert elicitation to obtain quantified information on uncertainties. Future research could assess the usefulness of expert elicitation for such quantifications. Related to this, we have provided some initial thoughts on decision-making with the kind of information on uncertainties that is provided by our framework. We have noted that more work is needed to develop decision frameworks that can handle the kind of information that results from this framework. Hence, research into decision frameworks that consider both first-order and second-order uncertainty will be highly useful.

In chapters 3 and 4, we have touched upon the transparency and interpretability of machine learning models. We have argued that the lack of transparency of data-driven models can be a source of second-order uncertainty and somewhat reduces the fitness of data-driven models as vehicles for scientific understanding. Recently, Creel (forthcoming) has suggested that three different types of transparency of complex computational systems should be distinguished, which are not all equally important in all contexts and for all purposes. It has also been argued, recently, that different stakeholders might have different requirements of what constitutes a satisfactory explanation in explainable machine learning (Zednik 2019). Future work could more explicitly link questions about the transparency of machine learning with the scientific issues discussed in this thesis, namely uncertainty assessments and scientific understanding. This work could engage in a discussion of the type of transparency that is needed for these tasks.

As noted in section 1.2.2, the term "big data" is not well defined. In chapter 2, we have touched on the issue of what exactly the term "big data" refers to. While we have identified different big-data elements that are characteristic of what the term refers to, we have not proposed a definition of the term. However, given the range of categories we have presented in which big-data elements are used, the descriptive approach chosen in chapter 2 yields insights that can be relevant for future work aiming to clarify the term "big data". Such work should be welcomed in order to have more clarity about what big data is and is not.

In chapters 3 and 4, we have distinguished between process-based and data-driven models. In the previous section, I have noted that it seems unlikely that data-driven models of phenomena can provide evidence for or against a hypothesis that is independent of the evidence derived from process-based models. The independence of the two modeling approaches is thus an issue that deserves attention in future work. What is more, future research could also address the distinction between process-based and data-driven models more explicitly. In chapter 2, we have illustrated that machine learning has been embedded into existing climate models to replace or improve parameterization schemes (see Krasnopolsky and Fox-Rabinovitz 2006; Schneider et al. 2017; Gentine et al. 2018).

This shows that the distinction between the two types of models need not always be clear in practice. Interesting research in the future could address epistemological issues concerning models that combine process-based and data-driven aspects, specifically hybrid models in which machine learning is embedded into process-based models. It could be interesting to investigate, for example, how the evaluation of such models would work in terms of their representational accuracy.

Finally, in chapters 3 and 4, we have introduced new frameworks to address epistemological issues in data-driven modeling. Both of these frameworks are fairly general and could be applied in various other contexts in future work, including in other disciplines in which data-driven models are employed and for process-based models in climate science. Future research could also scrutinize the two frameworks and assess, e.g., whether epistemic uncertainty can generally be thought of as a lack of conclusive justification as suggested in chapter 3 or further elaborate on the dimension of representational depth that we have introduced in the framework for understanding in chapter 4.

5.4. Closing Remarks

In November 2019, Adrian Daub, a professor of literature at Stanford University, wrote an op-ed for the Zurich-based newspaper *Neue Zürcher Zeitung* about the tendency of businesses to collect more and more data (Daub 2019). He argued that these massive datasets are affecting us, today, mostly in hypothetical terms. We are affected by the promise that these datasets might one day be used for important discoveries, but the promises of big data are mostly just that: promises. Regardless of whether this is true for the private sector, it does not seem true of scientific applications of big data, based on the findings of this thesis. While big data is no panacea for scientific research, it provides scientists with tools that are already used for a range of interesting research questions.

This thesis has investigated epistemological issues in data-driven modeling in climate research. The topics discussed show that the epistemological issues arising in relation to data-driven models are not fundamentally different from traditional topics discussed for other models even though aspects of model evaluation may be different. The age of big data brings about many opportunities for scientific research, and interdisciplinary research teams will likely obtain interesting insights thanks to data-science tools. However, data-driven models will remain one tool among many available to scientists that can be useful for some but certainly not all research questions.

References

Abbot, John, and Jennifer Marohasy. 2012. "Application of Artificial Neural Networks to Rainfall Forecasting in Queensland, Australia." *Advances in Atmospheric Sciences* 29 (4): 717–30. https://doi.org/10.1007/s00376-012-1259-9.

———. 2014. "Input Selection and Optimisation for Monthly Rainfall Forecasting in Queensland, Australia, Using Artificial Neural Networks." *Atmospheric Research* 138 (March): 166–78. https://doi.org/10.1016/j.atmosres.2013.11.002.

———. 2015. "Using Artificial Intelligence to Forecast Monthly Rainfall under Present and Future Climates for the Bowen Basin, Queensland, Australia." *International Journal of Sustainable Development and Planning* 10 (1): 66–75. https://doi.org/10.2495/SDP-V10-N1-66-75.

Adler, Carolina E., and Gertrude Hirsch Hadorn. 2014. "The IPCC and Treatment of Uncertainties: Topics and Sources of Dissensus: IPCC: Topics and Sources of Dissensus." *Wiley Interdisciplinary Reviews: Climate Change* 5 (5): 663–76. https://doi.org/10.1002/wcc.297.

Andersen, Hendrik, Jan Cermak, Julia Fuchs, Reto Knutti, and Ulrike Lohmann. 2017. "Understanding the Drivers of Marine Liquid-Water Cloud Occurrence and Properties with Global Observations Using Neural Networks." *Atmospheric Chemistry and Physics*, 9535–9546. https://doi.org/10.5194/acp-2017-282.

Anderson, Chris. 2008. "The End of Theory: The Data Deluge Makes the Scientific Method Obsolete." *WIRED Magazine*, June, 16.07. http://archive.wired.com/science/discoveries/magazine/16-07/pb_theory.

Arbuthnott, Katherine, Shakoor Hajat, Clare Heaviside, and Sotiris Vardoulakis. 2016. "Changes in Population Susceptibility to Heat and Cold over Time: Assessing Adaptation to Climate Change." *Environmental Health* 15 (S1). https://doi.org/10.1186/s12940-016-0102-7.

Baumberger, Christoph. 2014. "Types of Understanding: Their Nature and Their Relation to Knowledge." *Conceptus* 40 (98). https://doi.org/10.1515/cpt-2014-0002.

———. 2019. "Explicating Objectual Understanding Taking Degrees Seriously." *Journal for General Philosophy of Science*. https://doi.org/10.1007/s10838-019-09474-6.

Baumberger, Christoph, Claus Beisbart, and Georg Brun. 2017. "What Is Understanding? An Overview of Recent Debates in Epistemology and Philosophy of Science." In *Explaining Understanding: New Perspectives from Epistemology and Philosophy of Science*, edited by Stephen R. Grimm, Christoph Baumberger, and Sabine Ammon, 1–34. New York; London: Routledge,Taylor & Francis Group.

Baumberger, Christoph, Reto Knutti, and Gertrude Hirsch Hadorn. 2017. "Building Confidence in Climate Model Projections: An Analysis of Inferences from Fit." *Wiley Interdisciplinary Reviews: Climate Change* 8 (3): e454. https://doi.org/10.1002/wcc.454.

Beisbart, Claus. in preparation. "Opacity Thought Through – Two Ways of Understanding the Intransparency of Computer Simulations."

Benestad, Rasmus, Kajsa Parding, Andreas Dobler, and Abdelkader Mezghani. 2017. "A Strategy to Effectively Make Use of Large Volumes of Climate Data for Climate Change Adaptation." *Climate Services* 6 (April): 48–54. https://doi.org/10.1016/j.cliser.2017.06.013.

Betz, Gregor. 2016a. "Accounting for Possibilities in Decision Making." In *The Argumentative Turn in Policy Analysis*, edited by Sven Ove Hansson and Gertrude Hirsch Hadorn, 10:135–69. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-30549-3_6.

———. 2016b. "Logik und Argumentationstheorie." In *Neues Handbuch des Philosophie-Unterrichts*, edited by Jonas Pfister and Peter Zimmermann, 1. Auflage. UTB Philosophie, Ethik, Didaktik 4514. Bern: Haupt Verlag.

Biddle, Justin B., and Anna Leuschner. 2015. "Climate Skepticism and the Manufacture of Doubt: Can Dissent in Science Be Epistemically Detrimental?" *European Journal for Philosophy of Science* 5 (3): 261–78. https://doi.org/10.1007/s13194-014-0101-x.

Blundell, Charles, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. 2015. "Weight Uncertainty in Neural Networks." In *Proceedings of the 32nd International Conference on Machine Learning*. Vol. 32. Lille. http://arxiv.org/abs/1505.05424.

Bokulich, Alisa. 2018. "Using Models to Correct Data: Paleodiversity and the Fossil Record." *Synthese*, May. https://doi.org/10.1007/s11229-018-1820-x.

Boyd, Danah, and Kate Crawford. 2012. "Critical Questions for Big Data: Provocations for a Cultural, Technological, and Scholarly Phenomenon." *Information, Communication & Society* 15 (5): 662–79. https://doi.org/10.1080/1369118X.2012.678878.

Bradley, Richard, and Mareile Drechsler. 2014. "Types of Uncertainty." *Erkenntnis* 79 (6): 1225–48. https://doi.org/10.1007/s10670-013-9518-4.

Breiman, Leo. 2001. "Random Forests." *Machine Learning* 45 (1): 5–32. https://doi.org/10.1023/A:1010933404324.

Brun, Georg, and Gregor Betz. 2016. "Analysing Practical Argumentation." In *The Argumentative Turn in Policy Analysis*, edited by Sven Ove Hansson and Gertrude Hirsch Hadorn, 10:39–77. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-30549-3_3.

Brun, Georg, and Gertrude Hirsch Hadorn. 2014. *Textanalyse in den Wissenschaften: Inhalte und Argumente analysieren und verstehen*. UTB Schlüsselkompetenzen, Arbeitshilfen 3139. Zürich: vdf Hochschulverl.

Buckner, Cameron. 2019. "Deep Learning: A Philosophical Introduction." *Philosophy Compass* 14 (10). https://doi.org/10.1111/phc3.12625.

Bunn, Christian, Peter Läderach, Oriana Ovalle Rivera, and Dieter Kirschke. 2015. "A Bitter Cup: Climate Change Profile of Global Production of Arabica and Robusta Coffee." *Climatic Change* 129 (1–2): 89–101. https://doi.org/10.1007/s10584-014-1306-x.

Caldwell, Peter M., Christopher S. Bretherton, Mark D. Zelinka, Stephen A. Klein, Benjamin D. Santer, and Benjamin M. Sanderson. 2014. "Statistical Significance of Climate Sensitivity Predictors Obtained by Data Mining." *Geophysical Research Letters* 41 (5): 1803–8. https://doi.org/10.1002/2014GL059205.

Callebaut, Werner. 2012. "Scientific Perspectivism: A Philosopher of Science's Response to the Challenge of Big Data Biology." *Studies in History and Philosophy of Biological and Biomedical Sciences* 43 (1): 69–80. https://doi.org/10.1016/j.shpsc.2011.10.007.

Calude, Cristian S., and Giuseppe Longo. 2017. "The Deluge of Spurious Correlations in Big Data." *Foundations of Science* 22 (3): 595–612. https://doi.org/10.1007/s10699-016-9489-4.

Canali, Stefano. 2016. "Big Data, Epistemology and Causality: Knowledge in and Knowledge out in EXPOsOMICS." *Big Data & Society* 3 (2): 1–11. https://doi.org/10.1177/2053951716669530.

Carrier, Martin, and Johannes Lenhard. 2019. "Climate Models: How to Assess Their Reliability." *International Studies in the Philosophy of Science* 32 (2): 81–100. https://doi.org/10.1080/02698595.2019.1644722.

Castelli, Roberto, Peter Frolkovič, Christian Reinhardt, Christiaan C Stolk, Jakub Tomczyk, and Arthur Vromans. 2016. "Fog Detection from Camera Images." In *SWI 2016*, 19. http://www.swi-wiskunde.nl/swi2016/download-our-proceedings/.

Chadwick, R., E. Coppola, and F. Giorgi. 2011. "An Artificial Neural Network Technique for Downscaling GCM Outputs to RCM Spatial Scale." *Nonlinear Processes in Geophysics* 18 (6): 1013–28. https://doi.org/10.5194/npg-18-1013-2011.

Chen, Shien-Tsung, Pao-Shan Yu, and Yi-Hsuan Tang. 2010. "Statistical Downscaling of Daily Precipitation Using Support Vector Machines and Multivariate Analysis." *Journal of Hydrology* 385 (1–4): 13–22. https://doi.org/10.1016/j.jhydrol.2010.01.021.

Contessa, Gabriele. 2011. "Scientific Models and Representation." In *The Continuum Companion to the Philosophy of Science*, edited by Steven French and Juha Saatsi. New York: Continuum.

Creel, Kathleen A. forthcoming. "Transparency in Complex Computational Systems." *Philosophy of Science*. http://philsci-archive.pitt.edu/16669/.

Das, Saurabh, Rohit Chakraborty, and Animesh Maitra. 2017. "A Random Forest Algorithm for Nowcasting of Intense Precipitation Events." *Advances in Space Research* 60 (6): 1271–82. https://doi.org/10.1016/j.asr.2017.03.026.

Daub, Adrian. 2019. "Im Rausch des Sammelns von Nullen und Einsen." *Neue Zürcher Zeitung*, November 11, 2019. https://www.nzz.ch/meinung/im-rausch-des-sammelns-von-einsen-und-nullen-ld.1519978.

Dayal, Kavina, Ravinesh Deo, and Armando A. Apan. 2017. "Drought Modelling Based on Artificial Intelligence and Neural Network Algorithms: A Case Study in Queensland, Australia." In *Climate Change Adaptation in Pacific Countries*, edited by Walter Leal Filho, 177–98. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-50094-2_11.

De Mauro, Andrea, Marco Greco, and Michele Grimaldi. 2016. "A Formal Definition of Big Data Based on Its Essential Features." *Library Review* 65 (3): 122–35. https://doi.org/10.1108/LR-06-2015-0061.

Dellsén, Finnur. 2016. "Scientific Progress: Knowledge versus Understanding." *Studies in History and Philosophy of Science* 56 (April): 72–83. https://doi.org/10.1016/j.shpsa.2016.01.003.

Deo, Ravinesh C., and Mehmet Şahin. 2015. "Application of the Extreme Learning Machine Algorithm for the Prediction of Monthly Effective Drought Index in Eastern Australia." *Atmospheric Research* 153 (February): 512–25. https://doi.org/10.1016/j.atmosres.2014.10.016.

Durán, Juan Manuel. 2018. *Computer Simulations in Science and Engineering: Concepts - Practices - Perspectives*. The Frontiers Collection. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-90882-3.

Eghdamirad, Sajjad, Fiona Johnson, and Ashish Sharma. 2017. "Using Second-Order Approximation to Incorporate GCM Uncertainty in Climate Change Impact Assessments." *Climatic Change* 142 (1–2): 37–52. https://doi.org/10.1007/s10584-017-1944-x.

Eisenführ, Franz, Martin Weber, and Thomas Langer. 2010. *Rational Decision Making*. Berlin ; London: Springer.

Elliott, Kevin C., and Jon Rosenberg. 2019. "Philosophical Foundations for Citizen Science." *Citizen Science: Theory and Practice* 4 (1): 1–9. https://doi.org/10.5334/cstp.155.

Elmore, Kimberly L., Z. L. Flamig, V. Lakshmanan, B. T. Kaney, V. Farmer, Heather D. Reeves, and Lans P. Rothfusz. 2014. "MPING: Crowd-Sourcing Weather Reports for Research." *Bulletin of the American Meteorological Society* 95 (9): 1335–42. https://doi.org/10.1175/BAMS-D-13-00014.1.

Faghmous, James H., Arindam Banerjee, Shashi Shekhar, Michael Steinbach, Vipin Kumar, Auroop R. Ganguly, and Nagiza Samatova. 2014. "Theory-Guided Data Science for Climate Change." *Computer* 47 (11): 74–78. https://doi.org/10.1109/MC.2014.335.

Faghmous, James H., and Vipin Kumar. 2014. "A Big Data Guide to Understanding Climate Change: The Case for Theory-Guided Data Science." *Big Data* 2 (3): 155–63. https://doi.org/10.1089/big.2014.0026.

Floridi, Luciano. 2012. "Big Data and Their Epistemological Challenge." *Philosophy & Technology* 25 (4): 435–37. https://doi.org/10.1007/s13347-012-0093-4.

Foley, Aoife M., Paul G. Leahy, Antonino Marvuglia, and Eamon J. McKeogh. 2012. "Current Methods and Advances in Forecasting of Wind Power Generation." *Renewable Energy* 37 (1): 1–8. https://doi.org/10.1016/j.renene.2011.05.033.

Ford, James D., Simon E. Tilleard, Lea Berrang-Ford, Malcolm Araos, Robbert Biesbroek, Alexandra C. Lesnikowski, Graham K. MacDonald, Angel Hsu, Chen Chen, and Livia Bizikova. 2016. "Big Data Has Big Potential for Applications to Climate Change Adaptation." *Proceedings of the National Academy of Sciences* 113 (39): 10729–32. https://doi.org/10.1073/pnas.1614023113.

Forster, Malcolm. 2010. "Prediction." In *The Routledge Companion to Philosophy of Science*, edited by Martin Curd and Stathis Psillos. Routledge. https://doi.org/10.4324/9780203744857.ch42.

Forster, Piers M., Timothy Andrews, Peter Good, Jonathan M. Gregory, Lawrence S. Jackson, and Mark Zelinka. 2013. "Evaluating Adjusted Forcing and Model Spread for Historical and Future Scenarios in the CMIP5 Generation of Climate Models." *Journal of Geophysical Research: Atmospheres* 118 (3): 1139–50. https://doi.org/10.1002/jgrd.50174.

Frigg, Roman, and Stephan Hartmann. 2012. "Models in Science." In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta. Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/sum2018/entries/climate-science/.

Frigg, Roman, and James Nguyen. 2016. "Scientific Representation." In *Stanford Encyclopedia of Philosophy*, edited by Edward N Zalta. Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/entries/scientific-representation/.

Frigg, Roman, and Julian Reiss. 2009. "The Philosophy of Simulation: Hot New Issues or Same Old Stew?" *Synthese* 169 (3): 593–613. https://doi.org/10.1007/s11229-008-9438-z.

Frigg, Roman, Erica Thompson, and Charlotte Werndl. 2015a. "Philosophy of Climate Science Part I: Observing Climate Change: Observing Climate Change." *Philosophy Compass* 10 (12): 953–64. https://doi.org/10.1111/phc3.12294.

———. 2015b. "Philosophy of Climate Science Part II: Modelling Climate Change: Modelling Climate Change." *Philosophy Compass* 10 (12): 965–77. https://doi.org/10.1111/phc3.12297.

Frisch, Mathias. 2015. "Predictivism and Old Evidence: A Critical Look at Climate Model Tuning." *European Journal for Philosophy of Science* 5 (2): 171–90. https://doi.org/10.1007/s13194-015-0110-4.

Gagne II, David John, Sue Ellen Haupt, Douglas W. Nychka, and Gregory Thompson. 2019. "Interpretable Deep Learning for Spatial Analysis of Severe Hailstorms." *Monthly Weather Review* 147 (8): 2827–45. https://doi.org/10.1175/MWR-D-18-0316.1.

Gagne II, David John, Amy McGovern, Jeffrey B. Basara, and Rodger A. Brown. 2012. "Tornadic Supercell Environments Analyzed Using Surface and Reanalysis Data: A Spatiotemporal Relational Data-Mining Approach." *Journal of Applied Meteorology and Climatology* 51 (12): 2203–17. https://doi.org/10.1175/JAMC-D-11-060.1.

Gagne II, David John, Amy McGovern, Sue Ellen Haupt, Ryan A. Sobash, John K. Williams, and Ming Xue. 2017. "Storm-Based Probabilistic Hail Forecasting with Machine Learning Applied to Convection-Allowing Ensembles." *Weather and Forecasting* 32 (5): 1819–40. https://doi.org/10.1175/WAF-D-17-0010.1.

Gal, Yarin, and Zoubin Ghahramani. 2016. "Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning." In *Proceedings of the 33rd International Conference on Machine Learning*, 33:10. New York.

Gentine, P., M. Pritchard, S. Rasp, G. Reinaudi, and G. Yacalis. 2018. "Could Machine Learning Break the Convection Parameterization Deadlock?" *Geophysical Research Letters* 45 (11): 5742–51. https://doi.org/10.1029/2018GL078202.

Ghahramani, Zoubin. 2015. "Probabilistic Machine Learning and Artificial Intelligence." *Nature* 521 (May): 452–59. https://doi.org/10.1038/nature14541.

Ghosh, Subimal, and P.P. Mujumdar. 2008. "Statistical Downscaling of GCM Simulations to Streamflow Using Relevance Vector Machine." *Advances in Water Resources* 31 (1): 132–46. https://doi.org/10.1016/j.advwatres.2007.07.005.

Gibert, Karina, Jeffery S. Horsburgh, Ioannis N. Athanasiadis, and Geoff Holmes. 2018. "Environmental Data Science." *Environmental Modelling & Software* 106 (August): 4–12. https://doi.org/10.1016/j.envsoft.2018.04.005.

Gibert, Karina, Joaquín Izquierdo, Miquel Sànchez-Marrè, Serena H. Hamilton, Ignasi Rodríguez-Roda, and Geoff Holmes. 2018. "Which Method to Use? An Assessment of Data Mining Methods in Environmental Data Science." *Environmental Modelling & Software* 110 (December): 3–27. https://doi.org/10.1016/j.envsoft.2018.09.021.

Giere, Ronald N. 2004. "How Models Are Used to Represent Reality." *Philosophy of Science* 71 (5): 742–52. https://doi.org/10.1086/425063.

Gijsbers, Victor. 2013. "Understanding, Explanation, and Unification." *Studies in History and Philosophy of Science* 44 (3): 516–22. https://doi.org/10.1016/j.shpsa.2012.12.003.

Goertzel, Ben, and Cassio Pennachin. 2007. *Artificial General Intelligence*. Edited by Ben Goertzel and Cassio Pennachin. Cognitive Technologies. Berlin ; New York: Springer.

Grüne-Yanoff, Till. 2016. "Framing." In *The Argumentative Turn in Policy Analysis*, edited by Sven Ove Hansson and Gertrude Hirsch Hadorn, 10:189–215. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-30549-3_8.

Gudmundsson, L., and S. I. Seneviratne. 2015. "Towards Observation-Based Gridded Runoff Estimates for Europe." *Hydrology and Earth System Sciences* 19 (6): 2859–79. https://doi.org/10.5194/hess-19-2859-2015.

Hansson, Sven Ove. 2007. "Risk." In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta. Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/sum2018/entries/climate-science/.

———. 2009. "From the Casino to the Jungle: Dealing with Uncertainty in Technological Risk Management." *Synthese* 168 (3): 423–32. https://doi.org/10.1007/s11229-008-9444-1.

———. 2016. "Evaluating the Uncertainties." In *The Argumentative Turn in Policy Analysis*, edited by Sven Ove Hansson and Gertrude Hirsch Hadorn, 10:79–104. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-30549-3_4.

Hansson, Sven Ove, and Gertrude Hirsch Hadorn, eds. 2016. *The Argumentative Turn in Policy Analysis*. 1st Edition. Vol. 10. Logic, Argumentation & Reasoning. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-30549-3.

———. 2018. "Argument-Based Decision Support for Risk Analysis." *Journal of Risk Research* 21 (12): 1449–64. https://doi.org/10.1080/13669877.2017.1313767.

Hastie, Trevor, R Tibshirani, and J Friedman. 2008. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* 2nd Edition. Springer. http://www.myilibrary.com?id=18743.

Hawkins, Ed, and Rowan Sutton. 2009. "The Potential to Narrow Uncertainty in Regional Climate Predictions." *Bulletin of the American Meteorological Society* 90 (8): 1095–1108. https://doi.org/10.1175/2009BAMS2607.1.

Held, Isaac M. 2005. "The Gap between Simulation and Understanding in Climate Modeling." *Bulletin of the American Meteorological Society* 86 (11): 1609–14. https://doi.org/10.1175/BAMS-86-11-1609.

Hirsch Hadorn, Gertrude. 2016. "Temporal Strategies for Decision-Making." In *The Argumentative Turn in Policy Analysis*, edited by Sven Ove Hansson and Gertrude Hirsch Hadorn, 10:217–42. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-30549-3_9.

Hirsch Hadorn, Gertrude, Georg Brun, Carla Riccarda Soliva, Andrea Stenke, and Thomas Peter. 2015. "Decision Strategies for Policy Decisions under Uncertainties: The Case of Mitigation Measures Addressing Methane Emissions from Ruminants." *Environmental Science & Policy* 52 (October): 110–19. https://doi.org/10.1016/j.envsci.2015.05.011.

Holmes, Dawn E. 2017. *Big Data: A Very Short Introduction*. First edition. Very Short Introductions 539. Oxford, United Kingdom: Oxford University Press.

Hosni, Hykel, and Angelo Vulpiani. 2018. "Data Science and the Art of Modelling." *Lettera Matematica* 6 (May): 121–129. https://doi.org/10.1007/s40329-018-0225-5.

Humphreys, Paul. 2004. Extending Ourselves: Computational Science, Empiricism, and Scientific Method. New York: Oxford University Press.

———. 2009. "The Philosophical Novelty of Computer Simulation Methods." *Synthese* 169 (3): 615–26. https://doi.org/10.1007/s11229-008-9435-2.

Huntingford, Chris, Elizabeth S Jeffers, Michael B Bonsall, Hannah M Christensen, Thomas Lees, and Hui Yang. 2019. "Machine Learning and Artificial Intelligence to Aid Climate Change Research and Preparedness." *Environmental Research Letters* 14: 124007. https://doi.org/10.1088/1748-9326/ab4e55.

Hutson, Matthew. 2018. "AI Researchers Allege That Machine Learning Is Alchemy." Science | AAAS. May 2, 2018. http://www.sciencemag.org/news/2018/05/ai-researchers-allege-machine-learning-alchemy. Accessed on May 7, 2018.

Inman, Rich H., Hugo T.C. Pedro, and Carlos F.M. Coimbra. 2013. "Solar Forecasting Methods for Renewable Energy Integration." *Progress in Energy and Combustion Science* 39 (6): 535–76. https://doi.org/10.1016/j.pecs.2013.06.002.

IPCC. 2013. "Summary for Policy-Makers." In Climate Change 2013: The Physical Science Basis: Working Group I Contribution to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change, edited by Thomas Stocker, D. Qin, G.-K. Plattner, M. Tignor, S.K. Allen, J. Boschung, A. Nauels, Y. Xia, V. Bex, and P.M. Midgley. Stocker, T.F., D. Qin, G.-K. Plattner, M. Tignor, S.K. Allen, J. Boschung, A. Nauels, Y. Xia, V. Bex and P.M. Midgley (eds.). New York: Cambridge University Press.

———. 2018. "Summary for Policymakers." In Global Warming of 1.5°C. An IPCC Special Report on the Impacts of Global Warming of 1.5°C above Pre-Industrial Levels and Related Global Greenhouse Gas Emission Pathways, in the Context of Strengthening the Global Response to the Threat of Climate Change, Sustainable Development, and Efforts to Eradicate Poverty, edited by V. Masson-Delmotte, P. Zhai, H.-O. Pörtner, D. Roberts, J. Skea, P.R. Shukla, A. Pirani, et al. http://www.ipcc.ch/report/sr15/.

Jacobson, Ralph. 2013. "2.5 Quintillion Bytes of Data Created Every Day. How Does CPG & Retail Manage It?" *IBM Consumer Products Industry Blog* (blog). April 24, 2013. https://www.ibm.com/blogs/insights-on-business/consumer-products/2-5-quintillion-bytes-of-data-created-every-day-how-does-cpg-retail-manage-it/. Accessed on November 18, 2019.

James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani. 2013. *An Introduction to Statistical Learning*. Springer Texts in Statistics (STS Volume 103). New York, NY: Springer New York. https://doi.org/10.1007/978-1-4614-7138-7.

Jebeile, Julie, and Ashley Graham Kennedy. 2015. "Explaining with Models: The Role of Idealizations." *International Studies in the Philosophy of Science* 29 (4): 383–92. https://doi.org/10.1080/02698595.2015.1195143.

Jones, Gerrad D., Boris Droz, Peter Greve, Pia Gottschalk, Deyan Poffet, Steve P. McGrath, Sonia I. Seneviratne, Pete Smith, and Lenny H. E. Winkel. 2017. "Selenium Deficiency Risk Predicted to Increase under Future Climate Change." *Proceedings of the National Academy of Sciences* 114 (11): 2848–53. https://doi.org/10.1073/pnas.1611576114.

Karpatne, Anuj, Gowtham Atluri, James H. Faghmous, Michael Steinbach, Arindam Banerjee, Auroop Ganguly, Shashi Shekhar, Nagiza Samatova, and Vipin Kumar. 2017. "Theory-Guided Data Science: A New Paradigm for Scientific Discovery from Data." *IEEE Transactions on Knowledge and Data Engineering* 29 (10): 2318–31. https://doi.org/10.1109/TKDE.2017.2720168.

Kashiwao, Tomoaki, Koichi Nakayama, Shin Ando, Kenji Ikeda, Moonyong Lee, and Alireza Bahadori. 2017. "A Neural Network-Based Local Rainfall Prediction System Using Meteorological Data on the Internet: A Case Study Using Data from the Japan Meteorological Agency." *Applied Soft Computing* 56 (July): 317–30. https://doi.org/10.1016/j.asoc.2017.03.015.

Katzav, Joel, and Wendy S. Parker. 2015. "The Future of Climate Modeling." *Climatic Change* 132 (4): 475–87. https://doi.org/10.1007/s10584-015-1435-x.

———. 2018. "Issues in the Theoretical Foundations of Climate Science." *Studies in History and Philosophy of Modern Physics* 63: 141–49. https://doi.org/10.1016/j.shpsb.2018.02.001.

Kendall, Alex, and Yarin Gal. 2017. "What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision?" In *Proceedings of the 31st Conference on Neural Information Processing Systems*, 11. Long Beach.

Khalifa, Kareem. 2012. "Inaugurating Understanding or Repackaging Explanation?" *Philosophy of Science* 79 (1): 15–37. https://doi.org/10.1086/663235.

Kitchin, Rob. 2014. “Big Data, New Epistemologies and Paradigm Shifts.” *Big Data & Society* 1 (1): 1–12. https://doi.org/10.1177/2053951714528481.

Kitchin, Rob, and Gavin McArdle. 2016. “What Makes Big Data, Big Data? Exploring the Ontological Characteristics of 26 Datasets.” *Big Data & Society* 3 (1): 1–10. https://doi.org/10.1177/2053951716631130.

Kloprogge, Penny, Jeroen P. van der Sluijs, and Arthur C. Petersen. 2011. “A Method for the Analysis of Assumptions in Model-Based Environmental Assessments.” *Environmental Modelling & Software* 26 (3): 289–301. https://doi.org/10.1016/j.envsoft.2009.06.009.

Knüsel, Benedikt, and Christoph Baumberger. under review. “Understanding Climate Phenomena with Data-Driven Models.”

Knüsel, Benedikt, Marius Zumwald, Christoph Baumberger, Gertrude Hirsch Hadorn, Erich M. Fischer, David N. Bresch, and Reto Knutti. 2019. “Applying Big Data beyond Small Problems in Climate Research.” *Nature Climate Change* 9: 196–202. https://doi.org/10.1038/s41558-019-0404-1.

Knutti, Reto. 2008. “Should We Believe Model Predictions of Future Climate Change?” *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 366 (1885): 4647–64. https://doi.org/10.1098/rsta.2008.0169.

———. 2018. “Climate Model Confirmation: From Philosophy to Predicting Climate in the Real World.” In *Climate Modelling*, edited by Elisabeth A. Lloyd and Eric Winsberg, 325–59. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-65058-6_11.

Knutti, Reto, Christoph Baumberger, and Gertrude Hirsch Hadorn. 2019. “Uncertainty Quantification Using Multiple Models—Prospects and Challenges.” In *Computer Simulation Validation: Fundamental Concepts, Methodological Frameworks, and Philosophical Perspectives*, edited by Claus Beisbart and Nicole J. Saam. Simulation Foundations, Methods and Applications. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-70766-2.

Knutti, Reto, and Maria A. A. Rugenstein. 2015. “Feedbacks, Climate Sensitivity and the Limits of Linear Models.” *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 373 (2054): 1–20. https://doi.org/10.1098/rsta.2015.0146.

Knutti, Reto, Maria A. A. Rugenstein, and Gabriele C. Hegerl. 2017. “Beyond Equilibrium Climate Sensitivity.” *Nature Geoscience* 10 (10): 727–36. https://doi.org/10.1038/ngeo3017.

Koskinen, Inkeri. 2018. "Defending a Risk Account of Scientific Objectivity." *The British Journal for the Philosophy of Science* axy053 (August). https://doi.org/10.1093/bjps/axy053.

Krasnopolsky, Vladimir M., and Michael S. Fox-Rabinovitz. 2006. "Complex Hybrid Models Combining Deterministic and Machine Learning Components for Numerical Climate Modeling and Weather Prediction." *Neural Networks* 19 (2): 122–34. https://doi.org/10.1016/j.neunet.2006.01.002.

Krishnan, Maya. 2019. "Against Interpretability: A Critical Examination of the Interpretability Problem in Machine Learning." *Philosophy & Technology*, August. https://doi.org/10.1007/s13347-019-00372-9.

Kryvasheyeu, Y., H. Chen, N. Obradovich, E. Moro, P. Van Hentenryck, J. Fowler, and M. Cebrian. 2016. "Rapid Assessment of Disaster Damage Using Social Media Activity." *Science Advances* 2 (3): 1–11. https://doi.org/10.1126/sciadv.1500779.

Kwakkel, Jan H., Warren E. Walker, and Vincent A.W.J. Marchau. 2010. "Classifying and Communicating Uncertainties in Model-Based Policy Analysis." *International Journal of Technology, Policy and Management* 10 (4): 299. https://doi.org/10.1504/IJTPM.2010.036918.

Laney, Doug. 2001. "3D Data Management: Controlling Data Volume, Velocity, and Variety." *Application Delivery Strategies. Meta Group.* (blog). February 6, 2001. https://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf. Accessed on November 1, 2019.

LeCun, Yann, Yoshua Bengio, and Geoffrey Hinton. 2015. "Deep Learning." *Nature* 521 (May): 436–44. https://doi.org/10.1038/nature14539.

Lenhard, Johannes. 2006. "Surprised by a Nanowire: Simulation, Control, and Understanding." *Philosophy of Science* 73 (5): 605–16. https://doi.org/10.1086/518330.

Lenhard, Johannes, and Eric Winsberg. 2010. "Holism, Entrenchment, and the Future of Climate Model Pluralism." *Studies in History and Philosophy of Science Part B: Studies in History and Philosophy of Modern Physics* 41 (3): 253–62. https://doi.org/10.1016/j.shpsb.2010.07.001.

Leonelli, Sabina. 2012. "Introduction: Making Sense of Data-Driven Research in the Biological and Biomedical Sciences." *Studies in History and Philosophy of Biological and Biomedical Sciences* 43 (1): 1–3. https://doi.org/10.1016/j.shpsc.2011.10.001.

———. 2015. "What Counts as Scientific Data? A Relational Framework." *Philosophy of Science* 82 (5): 810–21. https://doi.org/10.1086/684083.

———. 2019a. “What Distinguishes Data from Models?” *European Journal for Philosophy of Science* 9 (22). https://doi.org/10.1007/s13194-018-0246-0.

———. 2019b. “Data — from Objects to Assets.” *Nature* 574 (October): 317–20. https://doi.org/10.1038/d41586-019-03062-w.

Linden, G., B. Smith, and J. York. 2003. “Amazon.Com Recommendations: Item-to-Item Collaborative Filtering.” *IEEE Internet Computing* 7 (1): 76–80. https://doi.org/10.1109/MIC.2003.1167344.

Lipper, Leslie, Philip Thornton, Bruce M. Campbell, Tobias Baedeker, Ademola Braimoh, Martin Bwalya, Patrick Caron, et al. 2014. “Climate-Smart Agriculture for Food Security.” *Nature Climate Change* 4 (12): 1068–72. https://doi.org/10.1038/nclimate2437.

Lipton, Peter. 2009. “Understanding Without Explanation.” In *Scientific Understanding. Philsophical Perspectives*, edited by Henk W. de Regt, Sabina Leonelli, and Kai Eigner. Pittsburgh University Press.

Lloyd, Elisabeth A. 2009. “I—Elisabeth A. Lloyd: Varieties of Support and Confirmation of Climate Models.” *Aristotelian Society Supplementary Volume* 83 (1): 213–32. https://doi.org/10.1111/j.1467-8349.2009.00179.x.

———. 2010. “Confirmation and Robustness of Climate Models.” *Philosophy of Science* 77 (5): 971–84. https://doi.org/10.1086/657427.

———. 2012. “The Role of ‘Complex’ Empiricism in the Debates about Satellite Data and Climate Models.” *Studies in History and Philosophy of Science* 43 (2): 390–401. https://doi.org/10.1016/j.shpsa.2012.02.001.

———. 2015. “Model Robustness as a Confirmatory Virtue: The Case of Climate Science.” *Studies in History and Philosophy of Science* 49 (February): 58–68. https://doi.org/10.1016/j.shpsa.2014.12.002.

Lloyd, Elisabeth A., and Eric Winsberg, eds. 2018. *Climate Modelling: Philosophical and Conceptual Issues*. 1st edition 2018. Cham: Springer International Publishing.

López-Rubio, Ezequiel, and Emanuele Ratti. 2019. “Data Science and Molecular Biology: Prediction and Mechanistic Explanation.” *Synthese*, May. https://doi.org/10.1007/s11229-019-02271-0.

Lu, Xin, David J. Wrathall, Pål Roe Sundsøy, Md. Nadiruzzaman, Erik Wetter, Asif Iqbal, Taimur Qureshi, et al. 2016. “Detecting Climate Adaptation with Mobile Network Data in Bangladesh: Anomalies in Communication, Mobility and Consumption Patterns during Cyclone Mahasen.” *Climatic Change* 138 (3–4): 505–19. https://doi.org/10.1007/s10584-016-1753-7.

Lukoianova, Tatiana, and Victoria L. Rubin. 2014. "Veracity Roadmap: Is Big Data Objective, Truthful and Credible?" *Advances in Classification Research Online* 24 (1): 4. https://doi.org/10.7152/acro.v24i1.14671.

Lusk, Greg. 2016. "Computer Simulation and the Features of Novel Empirical Data." *Studies in History and Philosophy of Science* 56 (April): 145–52. https://doi.org/10.1016/j.shpsa.2015.10.005.

Lyon, Aidan. 2015. "Data." In *The Oxford Handbook of the Philosophy of Science*, edited by Paul Humphreys. New York: Oxford University Press.

Majdzadeh Moghadam, Farhad. 2017. "Neural Network-Based Approach for Identification of Meteorological Factors Affecting Regional Sea-Level Anomalies." *Journal of Hydrologic Engineering* 22 (3): 04016058-1-04016058–15. https://doi.org/10.1061/(ASCE)HE.1943-5584.0001472.

Manogaran, Gunasekaran, and Daphne Lopez. 2018. "Spatial Cumulative Sum Algorithm with Big Data Analytics for Climate Change Detection." *Computers & Electrical Engineering* 65 (January): 207–21. https://doi.org/10.1016/j.compeleceng.2017.04.006.

Manogaran, Gunasekaran, Daphne Lopez, and Naveen Chilamkurti. 2018. "In-Mapper Combiner Based MapReduce Algorithm for Processing of Big Climate Data." *Future Generation Computer Systems* 86 (September): 433–45. https://doi.org/10.1016/j.future.2018.02.048.

Masson, David, and Reto Knutti. 2013. "Predictor Screening, Calibration, and Observational Constraints in Climate Model Ensembles: An Illustration Using Climate Sensitivity." *Journal of Climate* 26 (3): 887–98. https://doi.org/10.1175/JCLI-D-11-00540.1.

Mastrandrea, Michael D., Christopher B. Field, T. F. Stocker, Othmar Edenhofer, Kristie L. Ebi, D J Frame, Hermann Held, et al. 2010. "Guidance Note for Lead Authors of the IPCC Fifth Assessment Report on Consistent Treatment of Uncertainties." https://www.ipcc.ch/site/assets/uploads/2017/08/AR5_Uncertainty_Guidance_Note.pdf.

Mayer-Schönberger, V., and K. Cukier. 2013. *Big Data: A Revolution That Will Transform How We Live, Work and Think*. John Murray. https://books.google.ch/books?id=DReelwEACAAJ.

Mazzocchi, Fulvio, and Antonello Pasini. 2017. "Climate Model Pluralism beyond Dynamical Ensembles: Climate Model Pluralism beyond Dynamical Ensembles." *Wiley Interdisciplinary Reviews: Climate Change* 8 (6): e477. https://doi.org/10.1002/wcc.477.

McGovern, Amy, Kimberly L. Elmore, David John Gagne, Sue Ellen Haupt, Christopher D. Karstens, Ryan Lagerquist, Travis Smith, and John K. Williams. 2017. "Using Artificial Intelligence to Improve Real-Time Decision-Making for High-Impact Weather." *Bulletin of the American Meteorological Society* 98 (10): 2073–90. https://doi.org/10.1175/BAMS-D-16-0123.1.

McGovern, Amy, David J. Gagne, John K. Williams, Rodger A. Brown, and Jeffrey B. Basara. 2014. "Enhancing Understanding and Improving Prediction of Severe Weather through Spatiotemporal Relational Learning." *Machine Learning* 95 (1): 27–50. https://doi.org/10.1007/s10994-013-5343-x.

McGovern, Amy, Ryan Lagerquist, David John Gagne, G. Eli Jergensen, Kimberly L. Elmore, Cameron R. Homeyer, and Travis Smith. 2019. "Making the Black Box More Transparent: Understanding the Physical Implications of Machine Learning." *Bulletin of the American Meteorological Society*, November, 2175–99. https://doi.org/10.1175/BAMS-D-18-0195.1.

McGuffie, K., and A. Henderson-Sellers. 2005. *A Climate Modelling Primer*. Research and Developments in Climate. John Wiley & Sons.

Mearns, L. O., M. Bukovsky, S. C. Pryor, and V. Magaña. 2018. "Downscaling of Climate Information." In *Climate Modelling*, edited by Elisabeth A. Lloyd and Eric Winsberg, 199–269. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-65058-6_8.

Meinshausen, Nicolai. 2006. "Quantile Regression Forests." *Journal of Machine Learning Research* 7: 983–99.

Mekanik, F., M.A. Imteaz, S. Gato-Trinidad, and A. Elmahdi. 2013. "Multiple Regression and Artificial Neural Network for Long-Term Rainfall Forecasting Using Large Scale Climate Modes." *Journal of Hydrology* 503 (October): 11–21. https://doi.org/10.1016/j.jhydrol.2013.08.035.

Mendes, David, and José A. Marengo. 2010. "Temporal Downscaling: A Comparison between Artificial Neural Network and Autocorrelation Techniques over the Amazon Basin in Present and Future Climate Change Scenarios." *Theoretical and Applied Climatology* 100 (3–4): 413–21. https://doi.org/10.1007/s00704-009-0193-y.

Merz, B., H. Kreibich, and U. Lall. 2013. "Multi-Variate Flood Damage Assessment: A Tree-Based Data-Mining Approach." *Natural Hazards and Earth System Science* 13 (1): 53–64. https://doi.org/10.5194/nhess-13-53-2013.

Mohammadi, Kasra, Shahaboddin Shamshirband, Shervin Motamedi, Dalibor Petković, Roslan Hashim, and Milan Gocic. 2015. "Extreme Learning Machine Based Prediction of Daily Dew Point Temperature." *Computers and Electronics in Agriculture* 117 (September): 214–25. https://doi.org/10.1016/j.compag.2015.08.008.

Möller, Niklas. 2016. "Value Uncertainty." In *The Argumentative Turn in Policy Analysis*, edited by Sven Ove Hansson and Gertrude Hirsch Hadorn, 10:105–33. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-30549-3_5.

Morgan, M. G. 2014. "Use (and Abuse) of Expert Elicitation in Support of Decision Making for Public Policy." *Proceedings of the National Academy of Sciences* 111 (20): 7176–84. https://doi.org/10.1073/pnas.1319946111.

Muller, C.L., L. Chapman, S. Johnston, C. Kidd, S. Illingworth, G. Foody, A. Overeem, and R.R. Leigh. 2015. "Crowdsourcing for Climate and Atmospheric Sciences: Current Status and Future Potential." *International Journal of Climatology* 35 (11): 3185–3203. https://doi.org/10.1002/joc.4210.

Müller, Peter. 2010. "Constructing Climate Knowledge with Computer Models." *Wiley Interdisciplinary Reviews: Climate Change* 1 (4): 565–80. https://doi.org/10.1002/wcc.60.

Myhre, Gunnar, Drew Shindell, François-Marie Bréon, William Collins, Jan Fuglestvedt, Jianping Huang, Dorothy Koch, et al. 2013. "Anthropogenic and Natural Radiative Forcing." In *Climate Change 2013 - The Physical Science Basis*, edited by T. F. Stocker, D. Qin, G.-K. Plattner, M. Tignor, S.K. Allen, J. Boschung, A. Nauels, Y. Xia, V. Bex, and P.M. Midgley, 659–740. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9781107415324.018.

Nasseri, M., H. Tavakol-Davani, and B. Zahraie. 2013. "Performance Assessment of Different Data Mining Methods in Statistical Downscaling of Daily Precipitation." *Journal of Hydrology* 492 (June): 1–14. https://doi.org/10.1016/j.jhydrol.2013.04.017.

National Research Council, ed. 2013. *Frontiers in Massive Data Analysis*. Washington, D.C: The National Academies Press. https://doi.org/10.17226/18374.

Northcott, Robert. 2019. "Big Data and Prediction: Four Case Studies." *Studies in History and Philosophy of Science* in press (September). https://doi.org/10.1016/j.shpsa.2019.09.002.

Oppenheimer, Michael, Christopher M. Little, and Roger M. Cooke. 2016. "Expert Judgement and Uncertainty Quantification for Climate Change." *Nature Climate Change* 6 (5): 445–51. https://doi.org/10.1038/nclimate2959.

Oreskes, Naomi. 2018. "The Scientific Consensus on Climate Change: How Do We Know We're Not Wrong?" In *Climate Modelling*, edited by Elisabeth A. Lloyd and Eric Winsberg, 31–64. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-65058-6_2.

Oreskes, Naomi, K. Shrader-Frechette, and K. Belitz. 1994. "Verification, Validation, and Confirmation of Numerical Models in the Earth Sciences." *Science* 263 (5147): 641–46. https://doi.org/10.1126/science.263.5147.641.

Overeem, A., J. C. R. Robinson, H. Leijnse, G. J. Steeneveld, B. K. P. Horn, and R. Uijlenhoet. 2013. "Crowdsourcing Urban Air Temperatures from Smartphone Battery Temperatures." *Geophysical Research Letters* 40 (15): 4081–85. https://doi.org/10.1002/grl.50786.

Overpeck, J. T., G. A. Meehl, S. Bony, and D. R. Easterling. 2011. "Climate Data Challenges in the 21st Century." *Science* 331 (6018): 700–702. https://doi.org/10.1126/science.1197869.

Park, Seonyoung, Jungho Im, Sumin Park, and Jinyoung Rhee. 2017. "Drought Monitoring Using High Resolution Soil Moisture through Multi-Sensor Satellite Data Fusion over the Korean Peninsula." *Agricultural and Forest Meteorology* 237–238 (May): 257–69. https://doi.org/10.1016/j.agrformet.2017.02.022.

Parker, Wendy S. forthcoming. "Model Evaluation: An Adequacy-for-Purpose View." *Philosophy of Science*.

———. 2009. "Confirmation and Adequacy-for-Purpose in Climate Modelling." *Aristotelian Society Supplementary Volume* 83 (1): 233–49. https://doi.org/10.1111/j.1467-8349.2009.00180.x.

———. 2010a. "Predicting Weather and Climate: Uncertainty, Ensembles and Probability." *Studies in History and Philosophy of Science Part B: Studies in History and Philosophy of Modern Physics* 41 (3): 263–72. https://doi.org/10.1016/j.shpsb.2010.07.006.

———. 2010b. "Whose Probabilities? Predicting Climate Change with Ensembles of Models." *Philosophy of Science* 77 (5): 985–97. https://doi.org/10.1086/656815.

———. 2011. "When Climate Models Agree: The Significance of Robust Model Predictions." *Philosophy of Science* 78 (4): 579–600. https://doi.org/10.1086/661566.

———. 2014. "Simulation and Understanding in the Study of Weather and Climate." *Perspectives on Science* 22 (3): 336–56. https://doi.org/10.1162/POSC_a_00137.

———. 2016. “Reanalyses and Observations: What’s the Difference?” *Bulletin of the American Meteorological Society* 97 (9): 1565–72. https://doi.org/10.1175/BAMS-D-14-00226.1.

———. 2017. “Computer Simulation, Measurement, and Data Assimilation.” *The British Journal for the Philosophy of Science* 68 (1): 273–304. https://doi.org/10.1093/bjps/axv037.

———. 2018. “Climate Science.” In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta. Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/sum2018/entries/climate-science/.

Pasini, Antonello, Paolo Racca, Stefano Amendola, Giorgio Cartocci, and Claudio Cassardo. 2017. “Attribution of Recent Temperature Behaviour Reassessed by a Neural-Network Method.” *Scientific Reports* 7 (1). https://doi.org/10.1038/s41598-017-18011-8.

Patil, Amit Prakash, and Paresh Chandra Deka. 2016. “An Extreme Learning Machine Approach for Modeling Evapotranspiration Using Extrinsic Inputs.” *Computers and Electronics in Agriculture* 121 (February): 385–92. https://doi.org/10.1016/j.compag.2016.01.016.

Pietsch, Wolfgang. 2015. “Aspects of Theory-Ladenness in Data-Intensive Science.” *Philosophy of Science* 82 (5): 905–16. https://doi.org/10.1086/683328.

———. 2016. “The Causal Nature of Modeling with Big Data.” *Philosophy & Technology* 29 (2): 137–71. https://doi.org/10.1007/s13347-015-0202-2.

Pietsch, Wolfgang, and Jörg Wernecke. 2017. “Introduction: Ten Theses on Big Data and Computability.” In *Berechenbarkeit Der Welt? Philosophie Und Wissenschaft Im Zeitalter von Big Data*, edited by Wolfgang Pietsch, Jörg Wernecke, and Maximilian Ott, 37–57. Springer VS.

Preis, Tobias, Helen Susannah Moat, Steven R. Bishop, Philip Treleaven, and H. Eugene Stanley. 2013. “Quantifying the Digital Traces of Hurricane Sandy on Flickr.” *Scientific Reports* 3 (3141): 1–3. https://doi.org/10.1038/srep03141.

Radovic, Alexander, Mike Williams, David Rousseau, Michael Kagan, Daniele Bonacorsi, Alexander Himmel, Adam Aurisano, Kazuhiro Terao, and Taritree Wongjirad. 2018. “Machine Learning at the Energy and Intensity Frontiers of Particle Physics.” *Nature* 560 (August): 41–48. https://doi.org/10.1038/s41586-018-0361-2.

Rahmati, Omid, and Hamid Reza Pourghasemi. 2017. “Identification of Critical Flood Prone Areas in Data-Scarce and Ungauged Regions: A Comparison of Three Data Mining Models.” *Water Resources Management* 31 (5): 1473–87. https://doi.org/10.1007/s11269-017-1589-6.

Raje, Deepashree, and P. P. Mujumdar. 2011. "A Comparison of Three Methods for Downscaling Daily Precipitation in the Punjab Region." *Hydrological Processes* 25 (23): 3575–89. https://doi.org/10.1002/hyp.8083.

Rasouli, Kabir, William W. Hsieh, and Alex J. Cannon. 2012. "Daily Streamflow Forecasting by Machine Learning Methods with Weather and Climate Inputs." *Journal of Hydrology* 414–415 (January): 284–93. https://doi.org/10.1016/j.jhydrol.2011.10.039.

Ratti, Emanuele, and Ezequiel López-Rubio. 2018. "Mechanistic Models and the Explanatory Limits of Machine Learning." Presented at PSA2018: The 25th Biennial Meeting of the Philosophy of Science Association in Seattle, WA. http://philsci-archive.pitt.edu/14452/. Accessed on April 25, 2019.

Refsgaard, Jens Christian, Jeroen P. van der Sluijs, Anker Lajer Højberg, and Peter A. Vanrolleghem. 2007. "Uncertainty in the Environmental Modelling Process – A Framework and Guidance." *Environmental Modelling & Software* 22 (11): 1543–56. https://doi.org/10.1016/j.envsoft.2007.02.004.

Regt, Henk W. de. 2015. "Scientific Understanding: Truth or Dare?" *Synthese* 192 (12): 3781–97. https://doi.org/10.1007/s11229-014-0538-7.

———. 2017. *Understanding Scientific Understanding*. New York: Oxford University Press.

Regt, Henk W. de, and Dennis Dieks. 2005. "A Contextual Approach to Scientific Understanding." *Synthese* 144 (1): 137–70. https://doi.org/10.1007/s11229-005-5000-4.

Regt, Henk W. de. 2009. "The Epistemic Value of Understanding." *Philosophy of Science* 76 (5): 585–97. https://doi.org/10.1086/605795.

Reichstein, Markus, Gustau Camps-Valls, Bjorn Stevens, Martin Jung, Joachim Denzler, Nuno Carvalhais, and Prabhat. 2019. "Deep Learning and Process Understanding for Data-Driven Earth System Science." *Nature* 566 (February): 195–204. https://doi.org/10.1038/s41586-019-0912-1.

Reutlinger, Alexander, Dominik Hangleiter, and Stephan Hartmann. 2018. "Understanding (with) Toy Models." *The British Journal for the Philosophy of Science* 69 (4): 1069–99. https://doi.org/10.1093/bjps/axx005.

Roodposhti, Majid Shadman, Taher Safarrad, and Himan Shahabi. 2017. "Drought Sensitivity Mapping Using Two One-Class Support Vector Machine Algorithms." *Atmospheric Research* 193 (September): 73–82. https://doi.org/10.1016/j.atmosres.2017.04.017.

Roussos, Joe, Richard Bradley, and Roman Frigg. under review. “Making Confident Decisions with Model Ensembles.”

Salcedo-Sanz, S., R. C. Deo, L. Carro-Calvo, and B. Saavedra-Moreno. 2016. “Monthly Prediction of Air Temperature in Australia and New Zealand with Machine Learning Algorithms.” *Theoretical and Applied Climatology* 125 (1–2): 13–25. https://doi.org/10.1007/s00704-015-1480-4.

Schmidt, Gavin A., and Steven Sherwood. 2015. “A Practical Philosophy of Complex Climate Modelling.” *European Journal for Philosophy of Science* 5 (2): 149–69. https://doi.org/10.1007/s13194-014-0102-9.

Schneider, Tapio, Shiwei Lan, Andrew Stuart, and João Teixeira. 2017. “Earth System Modeling 2.0: A Blueprint for Models That Learn From Observations and Targeted High-Resolution Simulations.” *Geophysical Research Letters* 44 (24): 12,396-12,417. https://doi.org/10.1002/2017GL076101.

Schupbach, Jonah N. 2016. “Robustness Analysis as Explanatory Reasoning.” *The British Journal for the Philosophy of Science* 69 (1): 275–300. https://doi.org/10.1093/bjps/axw008.

Shelton, Taylor, Ate Poorthuis, Mark Graham, and Matthew Zook. 2014. “Mapping the Data Shadows of Hurricane Sandy: Uncovering the Sociospatial Dimensions of ‘Big Data.’” *Geoforum* 52 (March): 167–79. https://doi.org/10.1016/j.geoforum.2014.01.006.

Sippel, Sebastian, Nicolai Meinshausen, Erich M. Fischer, Enikő Székely, and Reto Knutti. 2020. “Climate Change Now Detectable from Any Single Day of Weather at Global Scale.” *Nature Climate Change* 10 (1): 35–41. https://doi.org/10.1038/s41558-019-0666-7.

Sprenger, Michael, Sebastian Schemm, Roger Oechslin, and Johannes Jenkner. 2017. “Nowcasting Foehn Wind Events Using the AdaBoost Machine Learning Algorithm.” *Weather and Forecasting* 32 (3): 1079–99. https://doi.org/10.1175/WAF-D-16-0208.1.

Steele, Katie, and Charlotte Werndl. 2013. “Climate Models, Calibration, and Confirmation.” *The British Journal for the Philosophy of Science* 64 (3): 609–35. https://doi.org/10.1093/bjps/axs036.

———. 2016. “The Diversity of Model Tuning Practices in Climate Science.” *Philosophy of Science* 83 (5): 1133–44. https://doi.org/10.1086/687944.

———. 2018. “Model-Selection Theory: The Need for a More Nuanced Picture of Use-Novelty and Double-Counting.” *The British Journal for the Philosophy of Science* 69 (2): 351–75. https://doi.org/10.1093/bjps/axw024.

Strevens, Michael. 2013. "No Understanding without Explanation." *Studies in History and Philosophy of Science* 44 (3): 510–15. https://doi.org/10.1016/j.shpsa.2012.12.005.

Sullivan, Emily. 2019. "Understanding from Machine Learning Models." *The British Journal for the Philosophy of Science* axz035 (August). https://doi.org/10.1093/bjps/axz035.

Sun, Alexander Y, and Bridget R Scanlon. 2019. "How Can Big Data and Machine Learning Benefit Environment and Water Management: A Survey of Methods, Applications, and Future Directions." *Environmental Research Letters* 14 (7): 073001. https://doi.org/10.1088/1748-9326/ab1b7d.

Tapia, Carlos, Beñat Abajo, Efren Feliu, Maddalen Mendizabal, José Antonio Martinez, J. German Fernández, Txomin Laburu, and Adelaida Lejarazu. 2017. "Profiling Urban Vulnerabilities to Climate Change: An Indicator-Based Vulnerability Assessment for European Cities." *Ecological Indicators* 78 (July): 142–55. https://doi.org/10.1016/j.ecolind.2017.02.040.

Tavakol-Davani, H., M. Nasseri, and B. Zahraie. 2013. "Improved Statistical Downscaling of Daily Precipitation Using SDSM Platform and Data-Mining Methods." *International Journal of Climatology* 33 (11): 2561–78. https://doi.org/10.1002/joc.3611.

Thompson, Erica, Roman Frigg, and Casey Helgeson. 2016. "Expert Judgment for Climate Change Adaptation." *Philosophy of Science* 83 (5): 1110–21. https://doi.org/10.1086/687942.

Tkachenko, Nataliya, Stephen Jarvis, and Rob Procter. 2017. "Predicting Floods with Flickr Tags." Edited by Guy J-P. Schumann. *PLOS ONE* 12 (2): e0172870. https://doi.org/10.1371/journal.pone.0172870.

Tripathi, Shivam, V.V. Srinivas, and Ravi S. Nanjundiah. 2006. "Downscaling of Precipitation for Climate Change Scenarios: A Support Vector Machine Approach." *Journal of Hydrology* 330 (3–4): 621–40. https://doi.org/10.1016/j.jhydrol.2006.04.030.

United Nations. 2015. *The Paris Agreement*. https://unfccc.int/process-and-meetings/the-paris-agreement/the-paris-agreement. Accessed on January 16, 2020.

Vaughan, Catherine, and Suraje Dessai. 2014. "Climate Services for Society: Origins, Institutional Arrangements, and Design Elements for an Evaluation Framework: Climate Services for Society." *Wiley Interdisciplinary Reviews: Climate Change* 5 (5): 587–603. https://doi.org/10.1002/wcc.290.

Veltri, Giuseppe Alessandro. 2017. "Big Data Is Not Only about Data: The Two Cultures of Modelling." *Big Data & Society* 4 (1): 1–16. https://doi.org/10.1177/2053951717703997.

Wahabzada, Mirwaes, Anne-Katrin Mahlein, Christian Bauckhage, Ulrike Steiner, Erich-Christian Oerke, and Kristian Kersting. 2016. "Plant Phenotyping Using Probabilistic Topic Models: Uncovering the Hyperspectral Language of Plants." *Scientific Reports* 6 (1). https://doi.org/10.1038/srep22482.

Walker, W.E., P. Harremoës, J. Rotmans, J.P. van der Sluijs, M.B.A. van Asselt, P. Janssen, and M.P. Krayer von Krauss. 2003. "Defining Uncertainty: A Conceptual Basis for Uncertainty Management in Model-Based Decision Support." *Integrated Assessment* 4 (1): 5–17. https://doi.org/10.1076/iaij.4.1.5.16466.

Walter, Achim, Robert Finger, Robert Huber, and Nina Buchmann. 2017. "Opinion: Smart Farming Is Key to Developing Sustainable Agriculture." *Proceedings of the National Academy of Sciences* 114 (24): 6148–50. https://doi.org/10.1073/pnas.1707462114.

Weaver, Christopher P., Robert J. Lempert, Casey Brown, John A. Hall, David Revell, and Daniel Sarewitz. 2013. "Improving the Contribution of Climate Model Information to Decision Making: The Value and Demands of Robust Decision Frameworks: The Value and Demands of Robust Decision Frameworks." *Wiley Interdisciplinary Reviews: Climate Change* 4 (1): 39–60. https://doi.org/10.1002/wcc.202.

Weisberg, Michael. 2006. "Robustness Analysis." *Philosophy of Science* 73: 730–742. https://doi.org/10.1086/518628.

———. 2013. *Simulation and Similarity: Using Models to Understand the World*. Oxford Studies in Philosophy of Science. New York: Oxford University Press.

Welker, Christoph, Olivia Martius, Peter Stucki, David Bresch, Silke Dierer, and Stefan Brönnimann. 2016. "Modelling Economic Losses of Historic and Present-Day High-Impact Winter Windstorms in Switzerland." *Tellus A: Dynamic Meteorology and Oceanography* 68 (1): 29546. https://doi.org/10.3402/tellusa.v68.29546.

Wenzel, Manfred, and Jens Schröter. 2010. "Reconstruction of Regional Mean Sea Level Anomalies from Tide Gauges Using Neural Networks." *Journal of Geophysical Research* 115 (C8). https://doi.org/10.1029/2009JC005630.

Werndl, Charlotte. 2016. "On Defining Climate and Climate Change." *The British Journal for the Philosophy of Science* 67 (2): 337–64. https://doi.org/10.1093/bjps/axu048.

Wilkenfeld, Daniel A. 2017. "MUDdy Understanding." *Synthese* 194 (4): 1273–93. https://doi.org/10.1007/s11229-015-0992-x.

Winsberg, Eric. 2018a. "Communicating Uncertainty to Policymakers: The Ineliminable Role of Values." In *Climate Modelling*, edited by Elisabeth A. Lloyd and Eric Winsberg, 381–412. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-65058-6_13.

———. 2018b. *Philosophy and Climate Science*. Cambridge, United Kingdom ; New York, NY: Cambridge University Press.

———. 2019. "Computer Simulations in Science." In *Stanford Encyclopedia of Philosophy*, edited by Edward N Zalta. https://plato.stanford.edu/entries/simulations-science/.

Wu, Jianjun, Lei Zhou, Ming Liu, Jie Zhang, Song Leng, and Chunyuan Diao. 2013. "Establishing and Assessing the Integrated Surface Drought Index (ISDI) for Agricultural Drought Monitoring in Mid-Eastern China." *International Journal of Applied Earth Observation and Geoinformation* 23 (August): 397–410. https://doi.org/10.1016/j.jag.2012.11.003.

Wüthrich, Nicolas. 2017. "Conceptualizing Uncertainty: An Assessment of the Uncertainty Framework of the Intergovernmental Panel on Climate Change." In *EPSA15 Selected Papers*, edited by Michela Massimi, Jan-Willem Romeijn, and Gerhard Schurz, 5:95–107. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-53730-6_9.

Ylikoski, Petri. 2014. "Agent-Based Simulation and Sociological Understanding." *Perspectives on Science* 22 (3): 318–35. https://doi.org/10.1162/POSC_a_00136.

Ylikoski, Petri, and N. Emrah Aydinonat. 2014. "Understanding with Theoretical Models." *Journal of Economic Methodology* 21 (1): 19–36. https://doi.org/10.1080/1350178X.2014.886470.

Yu, Manzhu, Chaowei Yang, and Yun Li. 2018. "Big Data in Natural Disaster Management: A Review." *Geosciences* 8 (165). https://doi.org/10.3390/geosciences8050165.

Zednik, Carlos. 2019. "Solving the Black Box Problem: A Normative Framework for Explainable Artificial Intelligence." *Philosophy and Technology*. https://doi.org/10.1007/s13347-019-00382-7.

Zhou, Lei, Jianjun Wu, Xinyu Mo, Hongkui Zhou, Chunyuan Diao, Qianfeng Wang, Yuanhang Chen, and Fengying Zhang. 2017. "Quantitative and Detailed Spatiotemporal Patterns of Drought in China during 2001–2013." *Science of The Total Environment* 589 (July): 136–45. https://doi.org/10.1016/j.scitotenv.2017.02.202.

Zscheischler, Jakob, Miguel D. Mahecha, and Stefan Harmeling. 2012. "Climate Classifications: The Value of Unsupervised Clustering." *Procedia Computer Science* 9: 897–906. https://doi.org/10.1016/j.procs.2012.04.096.

A. Supplementary Material to Chapter 3

Here, we present the individual arguments that can be made to justify the fitness-for-purpose in the case study presented in section 3.5. It is essentially an explicit reconstruction of the justifications shown in Table 3. The method used to reconstruct the arguments and relate them to each other in the argument map is largely based on Betz (2016b) with the exception that we also consider non-deductive arguments here. For a general introduction into the analysis of practical arguments, the reader is referred to Brun and Betz (2016).

In this reconstruction, the variable *M* denotes the model ensemble used by Jones et al. (2017), consisting of an ensemble of data-driven models, one of which was built using random forest, and two of which were built using artificial neural networks. The variable *S* refers to soil selenium concentrations. The first argument corresponds to the first row of Table 3 and directly concerns the fitness-for-purpose of the model and is similar to the one presented in the toy example:

	Argument 1
P1.1	*If a model has predicted many past instances of a phenomenon accurately and the modeled relationships remain sufficiently constant over time, that model is fit for predicting the phenomenon in the far future.*
P1.2	*M has predicted many past instances of S accurately.*
P1.3	*The modeled relationships in M remain sufficiently constant.*
C1	*M is fit for predicting S in the far future.*

In argument 1, premise P1.3 requires further justification. A possible justification is based on the fact that the relevant causal processes are represented in the model in a sufficiently accurate manner. This argument can again be reconstructed as a deductively valid argument:

	Argument 2
P2.1	*If a model represents the most important causal processes producing a phenomenon accurately and these processes are unaffected by changing environmental conditions, the modeled relationships remain sufficiently constant.*
P2.2	*The causal processes represented in M are unaffected by changing environmental conditions.*
P2.3	*M accurately represents the important causal processes that produce S.*
C2	*The modeled relationships in M remain sufficiently constant. (= P1.3)*

While this argument is deductively valid, it is not clear whether its premises are true. Premise P2.1 seems uncontroversial. Premise P2.2 requires some further justification. This can for example be justified based on background knowledge, e.g., if the processes represented are consistent with current scientific understanding and there is reason to believe that they are not dependent on current environmental conditions. Premise P2.3 in argument 2 also requires further justifications. There are four arguments that can be made in favor of P2.3, all of which are non-deductive. Hence, in these arguments, even if all the premises are true, they neither individually, nor jointly guarantee the truth of the conclusion. The first of these arguments refers the reasons (1), (2), and (3) presented in the main text and concerns how the machine learning algorithms were trained to construct the ensemble of data-driven models:

	Argument 3
P3.1	*M was constructed using data that represents sufficiently many configurations of S.*
P3.2	*M was constructed using the most important variables.*
P3.3	*M was constructed using sufficiently flexible methods.*
C3	*M represents most important mechanisms that produce S. (=P.23)*

In argument 3, the individual premises require further justification, too. This justification has to be made by referring to background knowledge. The expertise of both domain scientists and data scientists is necessary who need to judge whether the considered samples are sufficiently diverse (P3.1), whether relevant variables were omitted (P3.2), and whether the used methods were sufficiently flexible (P3.3).

A second argument that can be made in favor of P2.2 refers to the empirical accuracy of the model:

	Argument 4
P4.1	*M is empirically accurate with respect to the data from the past.*
C4	*M represents most important mechanisms that produce S. (= P2.3)*

Note, here, that there is a thesis that weakens argument 4. Namely, *M* has a low bias and underpredicts global average soil selenium concentration. This underprediction attacks P4.1 to some extent.

A third argument considers the consistency of the model with background knowledge. The truth of P5 can be established by conducting sensitivity analyses of the models. Furthermore, Jones et al. (2017) also use existing samples to show that the rate of change predicted by their models has historical precedents, which also serves as evidence for the truth of P5.

	Argument 5
P5	*M behaves in consistency with background knowledge about S.*
C5	*M represents most important mechanisms that produce S. (= P2.3)*

Finally, a fourth argument can be made that refers to the robustness of the models because predictions were only considered for the regions in which all three machine learning algorithms agreed.

	Argument 6
P6	*The predictions are only considered if the ensemble members of M agree on the sign of change of S.*
C6	*M represents most important mechanisms that produce S. (= P2.3)*

As mentioned above, the premises of Argument 3 all require further justification. For each of these, that justification has to come from background knowledge.

	Argument 7
P7	*M was constructed using over 30.000 samples from different continents.*
C7	*M was constructed using data that represents sufficiently many configurations of S. (=P3.1)*

	Argument 8
P8.1	*M was constructed using seven variables chosen based on a variable selection procedure.*
P8.2	*Most potentially relevant variables were included in the variable selection procedure.*
C8	*M was constructed using the most important variables. (= P3.2)*

	Argument 9
P9	*M was constructed using artificial neural networks and random forest.*
C9	*M was constructed using sufficiently flexible methods. (= P3.3)*

A problem emerges with respect to argument 8. Namely, as has been noted, data on selenium sources was lacking. This is what leads the models to underpredict global average selenium concentration (see argument 4 above). The low bias of the models shows that these sources of selenium are important for soil selenium concentrations. Hence,

that data on these sources was lacking directly attacks premise P8.2, which states that all potentially relevant variables were included in the variable selection procedure.

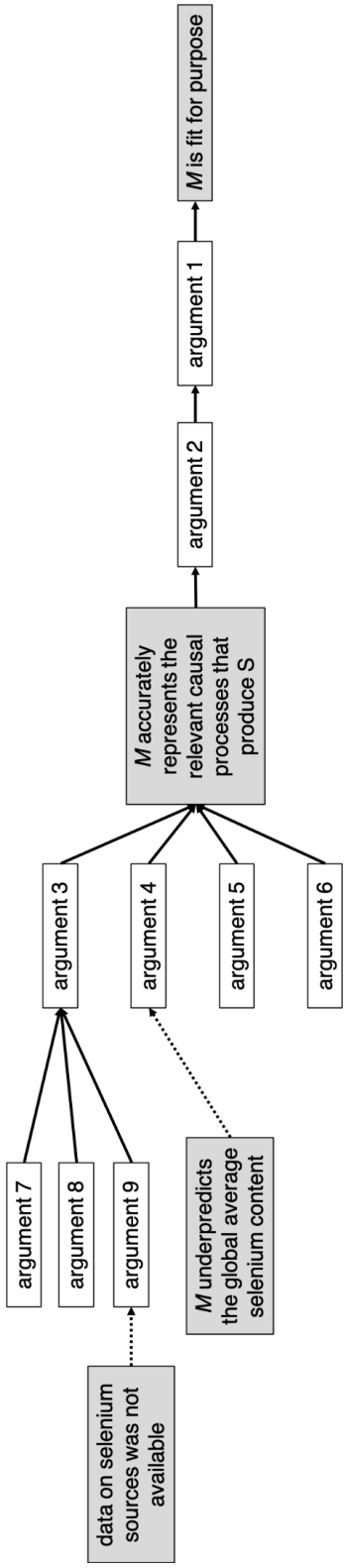


Figure 2: Argument map of the justification of the fitness-for-purpose of the models in the case study from section 3.5.

All of these arguments can then be arranged in an argument map as shown in Figure 2. This map is created as introduced by Betz (2016b). White boxes refer to arguments, and grey boxes to theses. Solid arrows denote that the content of one box, be it an argument or a thesis, supports the content of the other box. Dashed arrows denote that the content of one box attacks the content of the other box. Note that the solid arrows here do not differentiate between deductive and non-deductive arguments.

If an arrow goes from a thesis to an argument, this means that the thesis is a premise of the argument (support) or that the thesis contradicts a premise of the argument (attack). If the arrow goes from an argument to a thesis, this means that the thesis is the conclusion of the argument (support) or that the thesis is contradicted by the conclusion of the argument (attack).

B. Supplementary Material to Chapter 4

B.1. Energy-Balance Model

The energy balance model represents the Earth's energy balance with the following equation:

$$C \cdot \frac{dT}{dt} = F - \lambda \cdot T$$

The expression F consists of anthropogenic and natural forcing factors, F_{ant} and F_{nat}:

$$F = F_{ant} + F_{nat}$$

Data $F_{ant}$was obtained from the IPCC. For F_{ant}, a time series of mid-year radiative forcing of CO_2, CH_4, and N_2O combined was used. For F_{nat}, time series of mid-year radiative forcing from stratospheric aerosol optical depth and solar radiation were used. As for both these expressions, these time series contained annual values, they were interpolated linearly in order to obtain monthly values.

Then, using the BEST monthly temperature data and the forcing values from the IPCC, the values of the parameters C and λ were determined based on a least-squares optimization for the data from January 1931 to December 1980. As the value for C was not well constrained, the bounds for the optimization were based on the literature (see here: O. Geoffroy et al. "Transient climate response in a two-layer energy-balance model. Part I: Analytical solution and parameter calibration using CMIP5 AOGCM experiments." Journal of Climate 26, no. 6 (2013): 1841-1857.).

The IPCC data was retrieved from here:

http://www.pik-potsdam.de/~mmalte/rcps/data/, accessed on September 27, 2019.

The identified values were:

$$C = 6.0 \ \frac{J}{K \cdot m^2 \cdot year}$$

$$\lambda = \ 2.4 \frac{W}{m^2 \cdot K}$$

Note that the values of the heat capacity were divided by twelve for the simulation model as monthly time series was used.

The simulation was conducted by discretizing the equation as follows:

$$C \cdot \Delta T_t = F_t - \lambda \cdot (T_{t-1} + \Delta T_t) + (\beta_1 \cdot ENSO_t + \beta_2 \cdot AMO_t + \beta_3 \cdot PDO_t) \cdot \frac{W}{m^2}$$

$$\Delta T_t = T_{t+1} - T_t$$

The linear terms added with the parameters β_1, β_2, and β_3 account for internal variability. ENSO denotes the El Niño Southern Oscillation, AMO denotes the Atlantic Meridional Oscillation, and PDO denotes the Pacific Decadal Oscillation. The values were identified in the same way as C and λ, but only in a later step. The identified values are:

$$\beta_1 = 0.200, \beta_2 = 1.001, \beta_3 = 0.021$$

The time series for ENSO, AMO, and PDO were obtained from NOAA from the following websites (all accessed on September 27, 2019):

https://origin.cpc.ncep.noaa.gov/products/analysis_monitoring/ensostuff/ONI_change.shtml

https://www.esrl.noaa.gov/psd/data/timeseries/AMO/

https://www.ncdc.noaa.gov/teleconnections/pdo/

They contain unitless indices of the respective modes of internal variability.

B.2. Data-Driven Model

The same monthly data was used in a random forest regression model obtained from the open source python library SciKitLearn. The following parameter values were specified:

min_samples_split = 12 (meaning that the 12 observations had to be in a node at least to further split it into two nodes when training a tree; this parameter was chosen to reduce the noisiness of the predictions that resulted from predictions with the default value of 2).

n_estimators = 150 (meaning that 150 individual trees are trained).

The data was randomly split into a test and training dataset, with 70% of the data being used for training and 30% for testing. The testing procedure tested different model setups in which the parameter max_features was varied from 2 to 6 variables. This parameter denotes how many variables are maximally considered at each split. The model with a value of max_features = 4 achieved the lowest root mean squared error on the test set and was hence selected.

The other parameters are set according to the default values. Details can be found here: https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html, accessed on September 27, 2019.

The model was trained using BEST temperature data.

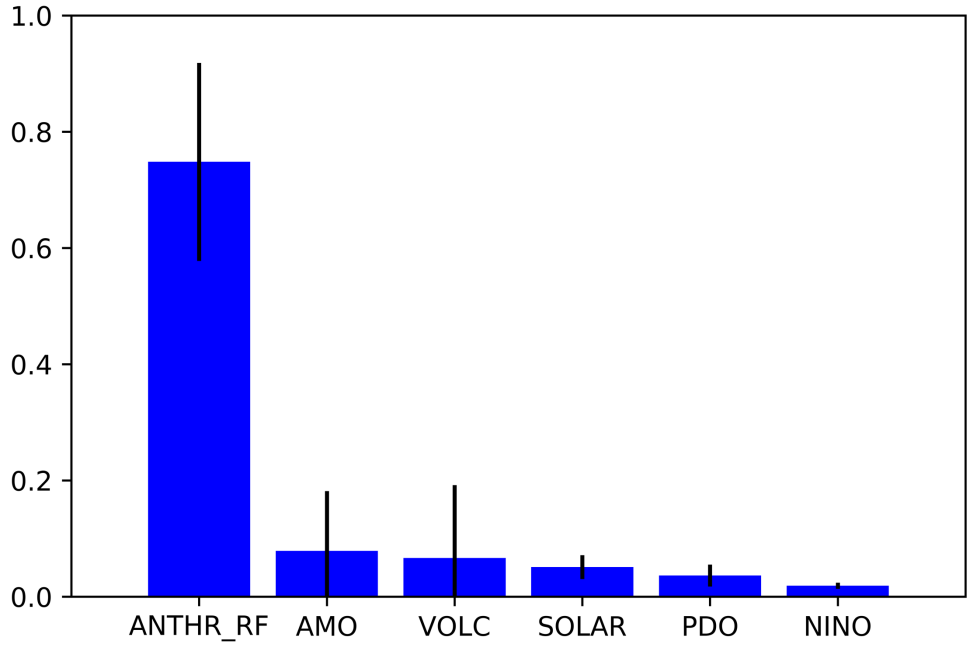


Figure 3: Variable importance plot of the model in the example in section 4.5.1.