Studies in History and Philosophy of Science 84 (2020) 46–56

Contents lists available at ScienceDirect

Studies in History and Philosophy of Science

journal homepage: http://www.elsevier.com/locate/shpsa

Understanding climate phenomena with data-driven models

Benedikt Knüsel [a,b,*], Christoph Baumberger [a]

[a] Institute for Environmental Decisions, ETH Zürich, Universitätstrasse 16, 8092, Zürich, Switzerland
[b] Institute for Atmospheric and Climate Science, ETH Zürich, Universitätstrasse 16, 8092, Zürich, Switzerland

ARTICLE INFO

Keywords:
Understanding
Climate models
Machine learning
Data-driven models
Representation
Grasping

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.

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, Meehl, Bony, & Easterling, 2011). This allows climate scientists to use machine learning; specifically it allows them to construct data-driven models 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 Knüsel et al., 2019; Northcott, 2019; Pietsch, 2015). 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 to what extent 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, leads to a decrease in model intelligibility, which, in turn impairs 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 the distinction between data-driven models and process-based models. We then outline 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. We then apply these criteria to data-driven models and compare their fitness as vehicles for understanding to that of process-based models. Based on this discussion, we argue that at first glance, data-driven models appear to pose a dilemma: They can be fit for providing understanding of simple phenomena. However, in these cases, researchers typically have sufficient background knowledge to construct

* Corresponding author. ETH Zürich, CHN J70.1 Universitätstrasse 16, 8092, Zürich, Switzerland.
E-mail address: benedikt.knuesel@alumni.ethz.ch (B. Knüsel).

https://doi.org/10.1016/j.shpsa.2020.08.003
Received 20 December 2019; Received in revised form 11 July 2020; Accepted 2 August 2020
Available online 23 August 2020

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 take it that it is this alleged dilemma of data-driven models that causes many researchers and philosophers to view machine learning as being mostly useful for prediction but not for understanding. The difficulties in using data-driven models for understanding in such more complex cases are partly based on the problems with model intelligibility discussed by López-Rubio and Ratti (2019). The lack of background knowledge to argue for the representational accuracy of a model can be seen as an instance of what Sullivan (2019) labels "link uncertainty". 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. Namely, for data-driven models to be useful for understanding phenomena, researchers should be in a position to argue from the coherence of the model with background knowledge to its representational accuracy. This can for example be achieved if important bivariate relationships are known. This sort of reasoning provides exactly the kind of evidence that reduces the link uncertainty discussed by Sullivan (2019). Furthermore, researchers can take steps to limit model complexity, e.g. by reducing the number of independent variables of their models. By doing so, researchers can obtain more intelligible models.

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 phenomena, 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 and briefly discuss how the insights from this example can be generalized. 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.

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 (Weisberg, 2013, chapter 1, argues that 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).

Like process-based models, data-driven models are models of phenomena. However, in contrast to process-based models, data-driven models are not constructed by explicitly representing processes in the form of equations. Instead they are built with machine learning. Machine learning is a set of methods at the interface of statistics and computer science that aim to algorithmically extract useful information from datasets. Data-driven modeling of phenomena, in the context of the present paper, starts with the training of a supervised machine learning algorithm with data that describes the phenomenon.[1] Training, here, refers to the step of algorithmically learning how to predict the values of a dependent variable 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 & Hartmann, 2012).

We note here 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 of machine learning include exploring a dataset and finding patterns and associations between variables. There is a wide variety of machine learning algorithms. On one end of the spectrum, there are simple tools such as linear regression and LASSO regression, which is a linear regression technique that performs a regularization selecting the most important variables automatically. At the other end of the spectrum, there are complex non-linear artificial neural networks, including deep learning, i.e., neural networks with many hidden layers (James, Witten, Hastie, & Tibshirani, 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 3.2.

Data-driven models have two main advantages over 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 when scientists do not have sufficient process understanding to construct a process-based model. The reason for this is that 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 (for a discussion of predictions with machine learning, see Northcott, 2019; Pietsch, 2015), there is skepticism about their usefulness for understanding phenomena as outlined in the introduction (see López-Rubio & 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

[1] In this paper, the focus is on data-driven modeling that relies on supervised machine learning, 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 it to predict the dependent variable based on independent variables for new samples with an unknown value of the dependent variable. 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.

sometimes be unclear. 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 can be used e.g. to represent cloud formation or vegetation in the form of plant functional types. In these parameterizations, certain parameter values are empirically set such that overall model behavior matches observational data, at least if the parameter values cannot be theoretically constrained. Parameterizations are "data-driven" in that observational data can be crucial in their construction and they 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, Pritchard, Rasp, Reinaudi, & Yacalis, 2018). These approaches further blur the line between process-based and data-driven models. Although the distinction may not always be clear, it is a useful one because clear-cut cases of either type of model exist, especially of data-driven models. Such examples will be discussed in more detail in sections 4 and 5.1.

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). There are different accounts of scientific understanding, which differ from each other because they require different characteristics from theories or models and from cognitive agents for understanding. While some accounts require that the theory be factive (Strevens, 2013), others require that the agent have certain abilities in handling the theory (de Regt, 2017), or both (Wilkenfeld, 2017). In the following, we draw on this literature and develop a framework that allows us to assess to what extent a model is fit for providing a user with understanding. Hence, the framework allows us to perform a fitness-for-purpose (or an adequacy-for-purpose) assessment where the purpose is understanding.[2]

The focus of the present paper is specifically on the understanding of phenomena, such as when an agent understands global warming, cloud formation, or the ice-albedo feedback. Understanding a phenomenon is typically related to having an explanation of the phenomenon (Baumberger, Beisbart, & Brun, 2017; de Regt, 2017). This kind of understanding would be attributed to people who can explain why global warming occurs, how cloud formation works, etc.[3] Hence, a model is considered fit for providing this kind of understanding if it allows the model user to construct an explanation of the phenomenon by providing her with explanatory information.[4] 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. Specifically, 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 our framework does not propose individually necessary and jointly sufficient conditions for (outright) understanding based on these dimensions and the respective criteria. Rather, they are evaluative criteria to assess how good the understanding is that can be obtained with a given model. 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 depends on the context. 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. Thus, when using phrases like "reasonably fit to serve as a vehicle for understanding", we mean that the model can give a degree of understanding that would typically be considered satisfactory in the context of scientific research.[5]

In subsections 3.1, 3.2, and 3.3, we introduce these dimensions and the respective evaluative criteria to assess how well a model fares with respect to the dimension.

3.1. Representational accuracy

The representational accuracy of a model is the degree to which the model is similar to its target in relevant respects (Giere, 2004; Wilkenfeld, 2017). As representational accuracy is defined with respect to a specific phenomenon, the relevant respects have to be determined with respect to that phenomenon. Specifically, 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 one to assess the representational accuracy of a model. They are based on work by 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

[2] Here, we use the term "fitness" instead of "adequacy" because fitness-for-purpose is a matter of degrees (see Parker, 2020). Understanding, too, is generally recognized to come in degrees. By making the assessment of the usefulness of a model for understanding explicitly gradual, the account of understanding allows for degrees, too.

[3] 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 the term.

[4] 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 an explanation of a phenomenon.

[5] We note, here, that in a specific context, it may be possible to define what the minimum requirements of a model are in terms of representational accuracy, representational depth, and graspability and the respective criteria, such that the model can serve as a vehicle for outright understanding. Then, these criteria would in fact be individually necessary and jointly sufficient conditions for outright understanding. This, however is not the focus of our paper. We thank an anonymous reviewer of this journal for raising this issue.

guarantee that a model is representationally accurate. Rather, they each come in degrees 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. This 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 a stronger reason 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 & 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 common core assumptions and they agree on some hypothesis, then this agreement can provide evidence for these core causal assumptions in the model. This is because, at least in the presence of other lines of evidence, it is likely that the shared core assumptions (rather than some of the auxiliary assumptions that the models do not share) are responsible for the agreement between the models (see Baumberger, Knutti, & 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 it is unlikely that the agreement on a hypothesis would occur if the models did not accurately represent the relevant aspects of the target (which could be the case 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.

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 is usually spelled out in terms of certain abilities, such as the ability to make use of the theory or model. 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 that theory. De Regt and Dieks suggest that a theory is intelligible for a scientist if she can estimate qualitatively the consequences of the theory without performing any calculations. As different scientists may weigh the aforementioned qualities differently, a theory can be intelligible for some scientist but not for others (see also de Regt, 2017). We propose two evaluative criteria for graspability. We briefly introduce them below and explain why they are relevant for 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? This is a model-specific adaptation of de Regt and Dieks' (2005) criterion for intelligibility. 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. A similar notion has also been suggested by Lenhard (2006) for understanding with simulation models.

Ability to explain model behavior: To what degree can the behavior of the model be explained by the user? This second aspect of graspability is relevant because the ability to qualitatively anticipate model outputs could also be obtained for a black-box model, at least if the input-output relationships are not too complex.[6] It seems obvious that a model whose behavior cannot only be anticipated qualitatively but also be explained is more graspable than one whose behavior cannot be explained. As a consequence of being more graspable, the model would be useful to obtain a higher degree of understanding. Namely, 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. This aspect of graspability is particularly relevant when a computer is involved in the modeling. For such models, it has been argued that the use of complex computer simulations can be harmful to the ability to explain model behavior, specifically because computer simulations are epistemically opaque (Humphreys, 2004, 2009) and because it can be difficult to attribute the reasons for successes and failures of climate model simulations to specific submodels (Lenhard & Winsberg, 2010).

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 well-versed model user. While 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 than the evaluation of its representational accuracy.

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. Generally, a model is considered representationally deeper, the more it describes a phenomenon not only on a phenomenological level but also in terms of the lower-level mechanisms producing the phenomenon. The representationally deeper a model is in this sense, the more comprehensively it represents the processes producing the phenomenon. Therefore, a representationally deeper model generally

[6] This means that a black-box model can, if it is also representationally accurate to some degree, have at least some degree of fitness as a vehicle for understanding. Whether or not this degree is sufficiently high to say that the black-box model can provide outright understanding depends on the context. All that our framework says is that, ceteris paribus, a model for which it is possible to anticipate model outputs has a higher degree of fitness-for-purpose than one for which this is not possible. We thank an anonymous reviewer of this journal for drawing our attention to this important point.

allows for more mechanistic understanding, which we take to be better than mere phenomenological understanding about how changing inputs relate to changing outputs. A phenomenon can be described at very different levels. For example, climate models range from models with no explicit representation of the spatial dimensions to models with a very high spatial resolution. Thus, representational depth, like the previous two dimensions, is also a matter of degree.[7]

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 a process, 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.

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. This discussion will also be helpful for the assessment of data-driven models. For this, we discuss the example of a zero-dimensional energy-balance model of the Earth's climate. Such simple models of the Earth's climate can help to better understand global warming (see, for example, Winsberg, 2018; chap. 3). We show how by discussing 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:

$$C \cdot \frac{dT}{dt} = F - \lambda \cdot T \tag{1}$$

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., Forster et al., 2013; Knutti & Rugenstein, 2015; for an overview, see; Knutti, Rugenstein, & Hegerl, 2017). The model equation prescribes that in equilibrium, any increase in the radiative forcing (the term "radiative forcing" refers to "the net change in the energy balance of the Earth system due to some imposed perturbation", e.g., due to increased CO_2 levels in the atmosphere, Myhre et al., 2013, p. 8) must be equal to the 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 1931–1980. The factor F, here, captures solar irradiance and volcanic activity (labeled "natural forcing factors" because they are independent of human activity, see Myhre et al., 2013) as well as three important greenhouse gases, namely CO_2, CH_4, and N_2O (called "anthropogenic forcing factors" because they are emitted primarily from human activities, see Myhre et al., 2013). Further details about model parameters and data are provided in the supplementary material.

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 a kind of causal hypothesis test, which can reveal which causes can and in fact have produced a certain phenomenon (see Parker, 2014). In the first simulation, both natural and anthropogenic radiative forcing factors follow their actual values over time; in the second, anthropogenic forcing factors are kept constant at their pre-industrial averages, and only natural forcing factors evolve according to historical records. The results of these two scenarios are displayed in Fig. 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 of global warming depends on the fitness of the model to serve as a vehicle for understanding, which can be assessed using the framework introduced in the previous section.

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. Equation (1) emerges when approximating a conservation of energy equation with a Taylor expansion under the assumption that the feedback parameter is constant (and thus, that higher-order expressions can be ignored, see Knutti & Rugenstein, 2015). This approximated equation states that, in equilibrium, incoming shortwave radiation and outgoing longwave radiation balance each other out. The factors considered in F were aggregated based on their radiative forcing. Thus, the model equation is derived from a fundamental physical law. It is coherent with background knowledge as long as the idealizations made do not impair the representational accuracy of the model. In equation (1), the idealizations concern the assumption of a constant feedback parameter and heat capacity. As Knutti and Rugenstein (2015, p. 7) argue, the assumption of a constant feedback parameter may be problematic when the climate system is taken far outside its current state (e.g., for high-emission scenario or paleoclimate studies). However, here, the model was only used to reproduce historical temperature records and 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. 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 Fig. 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 in which 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 & 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.

[7] It might be that describing a phenomenon at a lower level will not uniformly increase the fitness of a model for understanding, as at some level of description, the target phenomenon may be "lost". For example, describing individual atoms and molecules (rather than, say their concentration) in a climate model is unlikely to make a model a better vehicle for understanding. Similar points have been raised for example by Hilary Putnam's (1975) famous peg and hole example. If this is the case, the criterion of representational depth needs to be weakened to allow for this possibility.

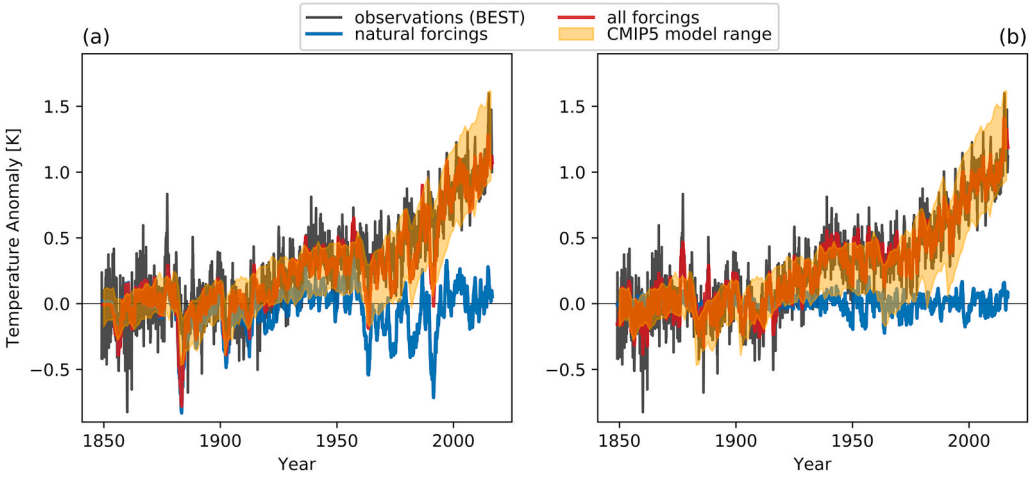

Fig. 1. Simulation runs of the energy-balance model (a) and of the data-driven model (b) for the scenario with all forcing factors corresponding to historical observations, and for the scenario where anthropogenic forcing factors are held constant at their preindustrial average values. Temperature anomalies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

When constructing this model, the values of the feedback parameter and the heat capacity were determined by calibrating the model to observations from the period from 1931 to 1980, displayed in Fig. 1(a). Hence, the model's good performance, especially outside of this range, gives us some confidence that it represents the climate system sufficiently accurately for our purposes (see Frisch, 2015).

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 Fig. 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. Nevertheless, the two modeling approaches rely partly on similar core causal assumptions, regarding e.g. energy conservation and radiative forcing of greenhouse gases. Thus, the robustness, here, also gives us some confidence in the representational accuracy of the energy-balance model.

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. Thus, it seems comparatively easy to learn to anticipate the behavior of the energy-balance model.

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 $\lambda \cdot T$. Hence, the ability to explain the behavior of the energy-balance model can also be obtained in a comparatively easy way.

Thus, there are indications that the simple energy-balance model represents the climate system accurately for an account of 20th century global mean surface temperature. 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 understanding of the 20th century global mean temperature evolution. This understanding is of a more phenomenological character due to the low degree of representational depth of the model, i.e., due to the fact that the model describes global warming at a high level of description.

In practice, climate scientists often employ models that are more complex than the simple energy-balance model just discussed. These state-of-the-art climate models (either general circulation models or Earth system models, see Parker 2018) are not only representationally accurate to a high degree, they also provide a lower-level description of phenomena than the energy-balance model discussed in this section. This means that they are generally representationally deeper than the energy-balance model, allowing them to provide a more mechanistic understanding of 20th century global warming than the energy-balance model.[8,9] At the same time, the complexity arising from the number of processes included in these models, as well as the layered development histories of climate models, can make it difficult for scientists to understand why a given model is successful or not, as both scientists (see Held, 2005) and philosophers (see Lenhard & Winsberg, 2010) have argued. This difficulty directly impairs the ability to explain model behavior (our second evaluative criterion of graspability), which implies that the graspability of state-of-the-art climate models is generally lower compared to the simple energy-balance model discussed above. This reveals that increases in representational accuracy and representational depth generally seem to be associated with a decrease in graspability, indicating an important trade-off regarding the three dimensions of our framework. Note that although there is a practical connection between

[8] 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, they allow for a deeper understanding of the phenomenon of global warming. However, in practice it can be difficult to attribute simulated phenomena to specific parts of the model. Thus, whether this understanding is better than understanding obtained with the energy-balance model depends on how the dimensions of representational depth and graspability are weighted, which, as outlined earlier, depends on the context.

[9] We note, here, that complex climate models explicitly represent a range of different phenomena and can hence be used as vehicles for understanding many different phenomena. While this makes them broadly applicable, it does not influence their fitness as vehicles for understanding one specific phenomenon.

these dimensions, the dimensions are nevertheless independent in a preferential sense (for the difference between preferential and statistical independence, see Eisenführ, Weber, & Langer, 2010, chap. 3): 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. This trade-off between representational accuracy and depth on the one hand and graspability on the other hand is the reason why idealizations do not always reduce the fitness of a model for providing explanations: although idealizations can reduce the model's representational accuracy or depth, they can increase the graspability of the model (see Jebeile & Kennedy, 2015).

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 5.1 and compare it to the energy-balance model from section 4. In subsection 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 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.

5.1. An illustrative example

For the example discussed here, we again take up the example of hypothesis testing of the causes of 20th global warming discussed in section 4. As we have mentioned in subsection 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 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 & Seneviratne, 2015). It is a machine learning algorithm that uses subsets of the data to create many individual regression (or decision) trees. 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 section 4 as well as with modes of internal variability. As the dependent variable, global mean surface temperature was used. Details are provided in the supplementary material. Then two simulations were run in analogy to the example in section 4. The model results are displayed in Fig. 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 3.

In Table 1, 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 ass assessment of the empirical accuracy, the 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. Here, 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.

While the assessment of the two models in terms of their empirical accuracy, their robustness, and the ability to qualitatively anticipate model outputs 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. Thus, these two criteria deserve an

Table 1
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	•model is based on conservation of energy •idealizations seem justified for the case at hand	•model behavior is consistent with background knowledge •model outputs are consistent with background knowledge •most relevant variables are considered •sufficiently flexible functional form is used •sufficiently many configurations of the target system are considered
Graspability	ability to qualitatively anticipate model outputs	•users can familiarize themselves with the model through manipulation •users can analyze the equation	•users can familiarize themselves with the model through manipulation
	ability to explain model behavior	•users can analyze the equation •users can learn to explain model behavior through manipulation	•users can study the variable importance plot •users can learn to explain model behavior through manipulation •users can make inferences from the working of the optimization algorithm to model behavior

in-depth discussion.

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.[10] 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 point 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 with other input factors. Since the representational accuracy, 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 quite flexible. We add two notes of caution: using a very flexible algorithm comes with the drawback of lower 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, p. 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 degree that, in combination with the other criteria, gives some confidence in the representational accuracy of the model.

Ability to explain model behavior: Whereas for 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 behaves. For example, it is possible to learn in this way that the model does treat anthropogenic forcing factors as the most important factor of global mean surface temperature in the 20th century. Similar insights can also be gained by e.g. considering the variable importance plot; a tool that graphically displays which independent variables cause the largest variation in the dependent variable (see supplementary material). 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 model user can also learn to explain such model artefacts and distinguish between features of a model which reflect actual behavior of the target and features which are due to the specific modeling approach at hand. Thus, although we acknowledge the difficulty in explaining model behavior, a well-versed model user can explain model behavior at least to some extent. Yet, 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 degree of 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. 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, 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.

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 subsection, data-driven models can be fit for providing understanding of phenomena in the climate system at least to some extent. They might therefore seem like a good choice of tools for achieving a better understanding of ill-understood phenomena. But to what extent can the example from the previous subsection be generalized? Generally, the evaluation of the empirical accuracy and the robustness of model results 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 1. However, in cases where model users have less background knowledge, arguing from the coherence with background knowledge to representational accuracy makes for a considerably weaker argument. Specifically, in more complex cases, the available background knowledge will often be insufficient to assess whether the most 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 1) 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 with model graspability. It will be more difficult to explain 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

[10] In the case of random forest, the variable importance plot can be assessed to learn about model behavior. This tool shows which variables contribute most strongly to variation in the dependent variable. While it is possible to learn about model behavior in this way (see below) and thus to increase the graspability of the model, we do not discuss it here, because it does not help with the assessment of the coherence of the model with background knowledge in this example. The reason for this is that to evaluate whether the variable importance plot shows a model behavior that is consistent with background knowledge, one would have to know which variables have the strongest influence on the dependent variable. However, this is precisely the understanding we are after in the example at issue.

model users than the one presented above. Furthermore, in more complex cases, it can also be more difficult to qualitatively anticipate model outputs. The reason for this is that these more complex cases may be characterized by a large number of variables and complex interactions between them. Thus, learning to anticipate how inputs and outputs are related might be very difficult. The concerns about the lack of graspability are therefore especially relevant for models with a large number of independent variables.

Thus, when serving as vehicles for understanding, data-driven models can face two problems. First, the difficulty explaining model behavior and anticipating model results qualitatively 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 because on its own, it is a rather weak argument for the representational accuracy of the model. . These two problems give rise to an (alleged) dilemma for data-driven models 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.

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 didactic 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, Cermak, Fuchs, Knutti, and Lohmann (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 depend on the region, 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. 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. Specifically, the complexity of aerosol-cloud interactions, and the spatial and temporal scales at which cloud processes take place limit the overall understanding of aerosol-cloud interactions as Andersen et al. (2017, p. 9535) explain. A poor understanding of these factors puts limits on how they could be represented in process-based models.

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 a better understanding.[11] Can the authors convincingly argue that their models provided 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. 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 3, i.e., 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 good 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 by making it easier to learn to anticipate model outputs qualitatively. 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 it in terms of representational accuracy.[12] What is more, data-driven models can also

[11] A second attack against the conclusion that data-driven models can provide understanding in some cases would claim that the models employed by Andersen et al. (2017) were not truly data-driven models. For example, one might argue that because a model user has to believe that the models accurately represent certain processes for the models to serve as vehicles for understanding, the models become process-based ones. However, we do not believe that such an argument would be convincing. The models employed by Andersen et al. (2017) are clearly data-driven according to the distinction discussed in section 2, because no equations were explicitly prescribed. Rather, the relationships between the variables were algorithmically extracted from a dataset by artificial neural networks. Yet, it is true that process knowledge is crucial for constructing data-driven models, e.g. in the choice of variables, and for evaluating them, e.g. for the sensitivity analyses. Hence, while such data-driven models should not be considered process-based, we note that they are clearly heavily theory-laden (cf. Pietsch, 2015). We thank an anonymous reviewer of this journal for raising this issue.

[12] 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, Haupt, Nychka, & Thompson, 2019). While graspability becomes a larger issue for these more complex models, the justification of representational accuracy can be possible.

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, this may indicate that processes not represented in their model must be relevant, and that additional variables may be 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 & 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). Advances in explainable machine learning can generally be expected to further increase the fitness of data-driven models for the purpose of understanding as they will increase the graspability of data-driven models (see Creel, forthcoming).

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 using the framework to compare a process-based and a data-driven model, we considered an alleged dilemma for data-driven models. According to this dilemma, 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, data-driven models can play an important role for researchers aiming to better understand climate phenomena.

We largely ignored the role that machine learning methods can play in applications other than as representational models of phenomena. For example, clustering algorithms, a class of unsupervised machine learning methods, can be used to identify homogeneous groups in climate datasets, e.g. regions of similar climatic conditions (Zscheischler, Mahecha, & Harmeling, 2012). 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 their use for the 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 state-of-the-art 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.

Funding

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

CRediT authorship contribution statement

Benedikt Knüsel: Conceptualization, Writing - original draft, Software, Visualization, Investigation. **Christoph Baumberger:** Conceptualization, Writing - review & editing, Supervision.

Declaration of competing interest

None.

Acknowledgments

We thank Hendrik Andersen, David N. Bresch, Roman Frigg, Mathias Frisch, Gertrude Hirsch Hadorn, Reto Knutti, Wendy Parker, Joe Roussos, 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 of the conference 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 Modeling & Understanding at the University of Bern for feedback on related talks.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.shpsa.2020.08.003.

References

Andersen, H., Cermak, J., Fuchs, J., Knutti, R., & Lohmann, U. (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.

Baumberger, C. (2019). Explicating objectual understanding taking degrees seriously. *Journal for General Philosophy of Science, 50*, 367–388. https://doi.org/10.1007/s10838-019-09474-6.

Baumberger, C., Beisbart, C., & Brun, G. (2017). What is understanding? An overview of recent debates in epistemology and philosophy of science. In S. R. Grimm, C. Baumberger, & S. Ammon (Eds.), *Explaining understanding: New perspectives from epistemology and philosophy of science*. New York ; London: Routledge,Taylor & Francis Group.

Baumberger, C., Knutti, R., & Hirsch Hadorn, G. (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.

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

Creel, K. A. (forthcoming). Transparency in complex computational systems. Philosophy of Science, 37.

Dellsén, F. (2016). Scientific progress: Knowledge versus understanding. *Studies In History and Philosophy of Science Part A, 56*, 72–83. https://doi.org/10.1016/j.shpsa.2016.01.003.

Eisenführ, F., Weber, M., & Langer, T. (2010). *Rational decision making*. Berlin ; London: Springer.

Forster, P. M., Andrews, T., Good, P., Gregory, J. M., Jackson, L. S., & Zelinka, M. (2013). Evaluating adjusted forcing and model spread for historical and future scenarios in the CMIP5 generation of climate models. *Journal of Geophysical Research: Atmosphere, 118*(3), 1139–1150. https://doi.org/10.1002/jgrd.50174.

Frigg, R., & Hartmann, S. (2012). Models in science. In E. N. Zalta (Ed.), *The stanford Encyclopedia of philosophy (summer 2018)*. Retrieved from https://plato.stanford.edu/archives/sum2018/entries/climate-science/.

Frisch, M. (2015). Predictivism and old evidence: A critical look at climate model tuning. *European Journal for Philosophy of Science, 5*(2), 171–190. https://doi.org/10.1007/s13194-015-0110-4.

Gagne, D. J., II, Haupt, S. E., Nychka, D. W., & Thompson, G. (2019). Interpretable deep learning for spatial analysis of severe hailstorms. *Monthly Weather Review, 147*(8), 2827–2845. https://doi.org/10.1175/MWR-D-18-0316.1.

Gentine, P., Pritchard, M., Rasp, S., Reinaudi, G., & Yacalis, G. (2018). Could machine learning break the convection parameterization deadlock? *Geophysical Research Letters*. https://doi.org/10.1029/2018GL078202.

Giere, R. N. (2004). How models are used to represent reality. *Philosophy of Science, 71*(5), 742–752. https://doi.org/10.1086/425063.

Gijsbers, V. (2013). Understanding, explanation, and unification. *Studies in History and Philosophy of Science, 44*(3), 516–522. https://doi.org/10.1016/j.shpsa.2012.12.003.

Gudmundsson, L., & Seneviratne, S. I. (2015). Towards observation-based gridded runoff estimates for Europe. *Hydrology and Earth System Sciences, 19*(6), 2859–2879. https://doi.org/10.5194/hess-19-2859-2015.

Held, I. M. (2005). The gap between simulation and understanding in climate modeling. *Bulletin of the American Meteorological Society, 86*(11), 1609–1614. https://doi.org/10.1175/BAMS-86-11-1609.

Humphreys, P. (2004). *Extending Ourselves*. Oxford University Press.

Humphreys, P. (2009). The philosophical novelty of computer simulation methods. *Synthese, 169*(3), 615–626. https://doi.org/10.1007/s11229-008-9435-2.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning*. https://doi.org/10.1007/978-1-4614-7138-7.

Jebeile, J., & Kennedy, A. G. (2015). Explaining with models: The role of idealizations. *International Studies in the Philosophy of Science, 29*(4), 383–392. https://doi.org/10.1080/02698595.2015.1195143.

Katzav, J., & Parker, W. S. (2015). The future of climate modeling. *Climatic Change, 132*(4), 475–487. https://doi.org/10.1007/s10584-015-1435-x.

Katzav, J., & Parker, W. S. (2018). Issues in the theoretical foundations of climate science. *Studies in History and Philosophy of Modern Physics, 63*, 141–149. https://doi.org/10.1016/j.shpsb.2018.02.001.

Knüsel, B., Zumwald, M., Baumberger, C., Hirsch Hadorn, G., Fischer, E. M., Bresch, D. N., et al. (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, R., & Rugenstein, M. A. A. (2015). Feedbacks, climate sensitivity and the limits of linear models, 2054 *Philosophical Transactions of the Royal Society A: Mathematical, Physical & Engineering Sciences, 373*. https://doi.org/10.1098/rsta.2015.0146, 20150146.

Knutti, R., Rugenstein, M. A. A., & Hegerl, G. C. (2017). Beyond equilibrium climate sensitivity. *Nature Geoscience, 10*(10), 727–736. https://doi.org/10.1038/ngeo3017.

Lenhard, J. (2006). Surprised by a nanowire: Simulation, control, and understanding. *Philosophy of Science, 73*(5), 605–616. https://doi.org/10.1086/518330.

Lenhard, J., & Winsberg, E. (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), 253–262. https://doi.org/10.1016/j.shpsb.2010.07.001.

Lipton, P. (2009). Understanding without explanation. In H. W. de Regt, S. Leonelli, & K. Eigner (Eds.), *Scientific understanding. Philsophical perspectives*. Pittsburgh University Press.

Lloyd, E. A. (2010). Confirmation and robustness of climate models. *Philosophy of Science, 77*(5), 971–984. https://doi.org/10.1086/657427.

López-Rubio, E., & Ratti, E. (2019). Data science and molecular biology: Prediction and mechanistic explanation. *Synthese*. https://doi.org/10.1007/s11229-019-02271-0.

Myhre, G., Shindell, D., Bréon, F.-M., Collins, W., Fuglestvedt, J., Huang, J., et al. (2013). Anthropogenic and natural radiative forcing. In *Intergovernmental panel on climate change (ed.), climate change 2013—the physical science basis*. https://doi.org/10.1017/CBO9781107415324.018.

Northcott, R. (2019). *Big data and prediction: Four case studies. Studies in history and philosophy of science Part A, S0039368119300652*. https://doi.org/10.1016/j.shpsa.2019.09.002.

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

Parker, W. S. (forthcoming). Model evaluation: An adequacy-for-purpose view. Philosophy of Science.

Parker, W. S. (2011). When climate models agree: The significance of robust model predictions*. *Philosophy of Science, 78*(4), 579–600. https://doi.org/10.1086/661566.

Parker, W. S. (2014). Simulation and understanding in the study of weather and climate. *Perspectives on Science, 22*(3). https://doi.org/10.1162/POSC_a_00137.

Parker, W. S. (2018). Climate science. In E. N. Zalta (Ed.), *The stanford Encyclopedia of philosophy (summer 2018)*. Retrieved from https://plato.stanford.edu/archives/sum2018/entries/climate-science/.

Pietsch, W. (2015). Aspects of theory-ladenness in data-intensive science. *Philosophy of Science, 82*(5), 905–916. https://doi.org/10.1086/683328.

Pietsch, W. (2016). The causal nature of modeling with big data. *Philosophy & Technology, 29*(2), 137–171. https://doi.org/10.1007/s13347-015-0202-2.

Putnam, H. (1975). 14. Philosophy and our mental life. In *Mind, language and reality* (Vol. 2, pp. 291–303). London: Cambridge University Press. https://doi.org/10.1017/CBO9780511625251.016.

de Regt, H. W. (2017). *Understanding scientific understanding*. New York: Oxford University Press.

de Regt, H. W., & Dieks, D. (2005). A contextual approach to scientific understanding. *Synthese, 144*(1), 137–170. https://doi.org/10.1007/s11229-005-5000-4.

Reichstein, M., Camps-Valls, G., Stevens, B., Jung, M., Denzler, J., Carvalhais, N., et al. (2019). Deep learning and process understanding for data-driven Earth system science. *Nature, 566*(7743), 195–204. https://doi.org/10.1038/s41586-019-0912-1.

Steele, K., & Werndl, C. (2016). The diversity of model tuning practices in climate science. *Philosophy of Science, 83*(5), 1133–1144. https://doi.org/10.1086/687944.

Strevens, M. (2013). No understanding without explanation. *Studies In History and Philosophy of Science Part A, 44*(3), 510–515. https://doi.org/10.1016/j.shpsa.2012.12.005.

Sullivan, E. (2019). Understanding from machine learning models. *The British Journal for the Philosophy of Science*. https://doi.org/10.1093/bjps/axz035. axz035.

Weisberg, M. (2006). Robustness analysis. *Philosophy of Science, 73*, 730–742.

Weisberg, M. (2013). *Simulation and similarity: Using models to understand the world*. New York: Oxford University Press.

Wilkenfeld, D. A. (2017). MUDdy understanding. *Synthese, 194*(4), 1273–1293. https://doi.org/10.1007/s11229-015-0992-x.

Winsberg, E. (2018). *Philosophy and Climate Science*. Cambridge University Press.

Zscheischler, J., Mahecha, M. D., & Harmeling, S. (2012). Climate classifications: The value of unsupervised clustering. *Procedia Computer Science, 9*, 897–906. https://doi.org/10.1016/j.procs.2012.04.096.