Understanding Climate Change with Statistical Downscaling and Machine Learning*

Julie Jebeile[1,2], Vincent Lam[1,2,3], and Tim Räz[†4]

[1]Institute of Philosophy, University of Bern, Länggassstrasse 49a, 3012 Bern, Switzerland
[2]Oeschger Centre for Climate Change Research, University of Bern, Hochschulstrasse 4, 3012 Bern, Switzerland
[3]School of Historical and Philosophical Inquiry, The University of Queensland, St Lucia QLD 4072, Australia.
[4]Institute of Biomedical Ethics and History of Medicine, University of Zürich, Winterthurerstrasse 30, 8006 Zürich, Switzerland

September 4, 2020

Forthcoming in *Synthese*.

Abstract

Machine learning methods have recently created high expectations in the climate modelling context in view of addressing climate change, but they are often considered as non-physics-based 'black boxes' that may not provide any understanding. However, in many ways, understanding seems indispensable to appropriately evaluate climate models and to build confidence in climate projections. Relying on two case studies, we compare how machine learning and standard statistical techniques affect our ability to understand the climate system. For that purpose, we put five evaluative criteria of understanding to work: intelligibility, representational accuracy, empirical accuracy, coherence with background knowledge, and assessment of the domain of validity. We argue that the two families of methods are part of the same continuum where these various criteria of understanding come in degrees, and that therefore machine learning methods do not necessarily constitute a radical departure from standard statistical tools, as far as understanding is concerned.

*We thank the participants of the philosophy of science research colloquium in the Spring semester 2020 at the University of Bern for valuable feedback on an earlier draft of the paper. We also wish to thank the participants of the seminar 'Philosophy of science perspectives on the climate challenge' and the workshop 'Big data, machine learning, climate modelling & understanding' in the Fall semester 2019 at the University of Bern and supported by the Oeschger Centre for Climate Change Research. JJ and VL are grateful to the Swiss National Science Foundation for financial support (grant PP00P1_170460). TR was funded by the cogito foundation.

†Corresponding author, tim.raez@posteo.de

Keywords: climate models, understanding, dynamical and statistical downscaling, deep neural networks, machine learning, climate change

Author contributions: All authors contributed equally.

1 Introduction

The topic of this paper is understanding with climate models. More specifically, we are interested in the question of how the use of statistical techniques and machine learning in climate models affects our ability to understand the climate system and its response to external forcings. It is tempting to deny that understanding is important to climate modelling in the first place. The central task of climate modelling, it could be argued, is to provide climate projections in order to inform decision-makers about future climate change. Accordingly, it seems that careful evaluation of the trustworthiness of climate model projections is a more important matter to address than understanding.

But in many ways, understanding is not at all secondary for climate modelling, quite to the contrary (e.g. Held 2005, Parker 2014). It can be convincingly argued that understanding is an essential aspect of climate model evaluation: it is crucial to the identification of the reasons for the successes or for the failures of climate models to fit empirical data[1] and furthermore, understanding is essential to building confidence in (e.g. long-term, high-forcing) climate projections (Baumberger et al. 2017, Knutti 2018).

In recent years, the application of machine learning methods in climate modelling have created high expectations in terms of more reliable climate (change) projections. The use of machine learning in science has also been met with suspicion, in particular by those who emphasise the importance of understanding. The predictive power we gain using these methods may come at the cost of understanding, because machine learning models are non-physics-based, 'black box' models (see e.g. López-Rubio and Ratti 2019 and Alain and Bengio 2016). This potentially prevents understanding and the identification of proper explanatory mechanisms whenever these methods are used in climate modelling.

However, the claim that machine learning prevents understanding with climate models altogether is an extreme position that should be avoided. For one, difficulties with understanding also occur in climate modelling without machine learning, for instance at the level of regional climate models, which are central to tackling climate change and also use various standard statistical methods (and hence constitute an interesting case study). We should expect a certain continuity with respect to understanding between the use of machine learning and more traditional statistical methods. Overstating the contrast between these methods may lead us to neglect a critical perspective on understanding in climate modelling, independently of the use of machine learning methods.

[1] Such identification may actually be constrained by a form of holism of confirmation and refutation that generally characterizes (complex) climate models (Lenhard and Winsberg 2010).

What is more, it is generally acknowledged in the philosophical literature, if only implicitly, that understanding comes in degrees (although this may not have been taken seriously enough, see Baumberger 2019). One of the main points of the present paper is that there is a continuum in climate modelling with respect to various criteria of understanding. Taking this point seriously will allow us to avoid the two extreme positions: neither is understanding completely absent from or unimportant to climate modelling, nor is it completely eradicated by the use of machine learning models.

Climate models are best considered with respect to their adequacy for a certain purpose (Parker 2020), and many climate models (in particular regional climate models relying on statistical techniques and climate models incorporating machine learning methods) may not have understanding as their primary purpose. From this point of view, it can be asked to what extent and in what sense these climate models can nevertheless provide some understanding; addressing these questions will allow for a more detailed characterization of the understanding gap in climate modelling (highlighted in Held 2005).

The strategy of this paper is not to elaborate a general philosophical theory of scientific understanding, valid across domains (as discussed in, e.g., Wilkenfeld 2017 and de Regt 2017). Rather, we focus on five evaluative criteria of understanding in the context of climate modelling.[2] They include intelligibility in the sense of the ability to anticipate qualitative behaviour, representational accuracy, empirical accuracy, coherence with background knowledge and assessment of the domain of validity. We endeavour to articulate those criteria in section 2. We then discuss how standard statistical techniques in regional climate modelling affect these five criteria of understanding in section 3, before turning to climate modelling involving machine learning methods in section 4. We finally discuss in section 5 how the five criteria come in degrees in a similiar way in both cases, before concluding in section 6.

2 Understanding with scientific models

Historically, the distinction between explanation and understanding, viz. "erklären" and "verstehen", was introduced to emphasise the methodological and disciplinary oppositions between natural science and humanities; there was a lively philosophical debate about this distinction at the beginning of the 20th century, in particular in the German-speaking world. Later, the logical positivists (e.g. Hempel and Oppenheim 1948) considered understanding to be a mere psychological by-product of explanation, and a mental process that is internal and hardly communicable. For this reason, understanding has been dismissed in order to avoid human and subjective influence from the hypothetico-deductive

[2]The approach here is similar to the framework proposed in Knüsel and Baumberger (2020), although our aims are different (and complementary): whereas they want to show that climate models involving machine learning can provide some understanding in certain cases (they discuss a case study, on which part of their argument crucially relies), we aim to emphasise that this is all a matter of degree, already within 'standard' climate modelling *without* machine learning.

reconstruction of science. The psychological effect of cognitively grasping an explanation (of 'haha!' or 'Eureka!' type) has later been called 'sense of understanding'. It has been shown to artificially increase one's confidence in explanations, and therefore to be a fallible criterion for good explanations (e.g. Trout 2002; and Kuorikoski 2011 in the context of computer simulations). Recent philosophical accounts of understanding solve this problem by elaborating a multidimensional concept, in particular in the context of the use of models (cf. references in the next paragraph). By putting evaluative criteria forward, they offer a concept of understanding that cannot be confounded anymore with a mere subjective feeling, since one's understanding can be measured by people with respect to the very criteria. The criteria are supposed to account for both (i) the adequacy of the model to produce explanations about the target phenomenon or system, and (ii) the ability of the agents in drawing these explanations from the model.

In this section, in order to compare understanding with machine learning and understanding with common statistical tools, we use evaluative criteria for understanding with a model in general. We provide five criteria that we mostly adopt (with the notable exception of the fifth one) from the recent philosophical literature on understanding and more particularly understanding with climate models (e.g. Wilkenfeld 2017; Baumberger et al. 2017; Knüsel and Baumberger 2020). As we will argue, these criteria are not categorical, but come in degrees depending on how well they are individually met.

2.1 Intelligibility

Understanding with a model presumes adequacy of the model to provide explanations. But understanding with a model also requires understanding *of* the model, and this bears on the intelligibility of the model (see, e.g., de Regt and Dieks 2005; de Regt 2017; Wilkenfeld 2017). While adequacy concerns the representational function of the model with respect to the target phenomenon, intelligibility is based on the ability and skill of the agent to use the model and to obtain explanations from it, and on the features of the model that enable its manipulability (in agreement with the literature on the topic, we will focus on these latter, and not on the more subjective features of the agent).[3] The first criterion we want to highlight is thus related to the intelligibility of the model.

A prominent criterion of intelligibility is provided by de Regt and Dieks (2005). De Regt and Dieks (2005) suggest that a "scientific theory T is intelligi-

[3] Adequacy and intelligibility are commonly considered as the two central 'pillars' of understanding. Thus, de Regt distinguishes understanding a phenomenon–that is, having an adequate explanation of the phenomenon–and understanding a theory–that is, being able to use the theory (2017, 23). When Wilkenfeld (2017) introduces his Multiple Understanding Dimensions (MUD) theory as "a natural synthesis of existing views" of understanding, he argues that "representational-accuracy (of which we assume truth is one kind) and intelligibility (which we will define so as to entail abilities) are good-making features of a state of understanding" (Wilkenfeld 2017, 1274). Following the example, Knüsel and Baumberger (2020, §3) offer three "dimensions" of understanding encompassing representational accuracy, representational depth, and graspability.

ble for scientists (in context C) if they can recognise qualitatively characteristic consequences of T without performing exact calculations" (2005, 151). On this view, understanding results from the reasoning of a computationally unaided agent with the help of an external representation, i.e., a theory. However, in the case of complex modelling, it is worth questioning whether it is possible to anticipate qualitatively consequences without computer assistance. In many contemporary scientific domains, e.g. in astrophysics and in climate science, some predictions are impossible without running a computer simulation. Scientists can familiarise themselves with the model, and probably learn to anticipate, to some extent, the qualitative behaviour of the model, but only by using the model in the first place, i.e. by making local changes in the model inputs, and running computer simulations.

In such circumstances, intelligibility can rather be obtained through relevant manipulation of the model, which then may allow one to anticipate its qualitative behaviour. Following Kuorikoski and Ylikoski (2015), we assume that understanding with a model should be viewed as an extended cognitive activity that relies on the "inferential aid" of an implemented model used to produce explanations and on the agent's ability to manipulate the model. In Kuorikoski and Ylikoski (2015)'s inferentialist account of model-based understanding, the understanding of a model is obtained by the ability of the agent to answer what-if-things-would-have-been-different questions about the target phenomenon by manipulating the model. Manipulating a model means 'playing around' with the model, i.e. varying the model parameters, the parameterisations, the discretisation-based numerical schemes, or the initial conditions, and then running computer simulations in order to explore the resulting changes in the simulated phenomenon and to get a sense of the qualitative behaviour of the model. Then, a model is more or less intelligible and can thus provide more or less understanding (along this criterion) depending on the number of what-if-things-would-have-been-different questions it can enable the agent to answer.

That said, as Kuorikoski and Ylikoski (2015) make it clear, "it is not enough to be able to make just any inferences one wishes; one must get those inferences right" (3819). Therefore, evaluative criteria, used to assess to which extent the model is an accurate representation of the target phenomenon, are also required. In what follows, our aim is to articulate additional evaluative criteria of understanding that assess the connection of the model with the target phenomenon.

2.2 Representational accuracy

Models can provide scientists with understanding if they are adequate representations for providing explanations. In this regard, representational accuracy is a second important criterion for understanding with a model. It is evaluated with regard to how well a model captures the relevant physical processes at work in the target system under investigation.

Physics-based equations aim at explicitly describing the physical processes of interest in mathematical terms, and are supposed to contribute to a better

representational accuracy in virtue of their high degree of confirmation. In contrast, statistical techniques deliver model outputs based on functional relations between model inputs and outputs that rely on statistical rather than physical considerations.

Representational accuracy depends on the way a model captures the physical processes, but also the extent to which the system's aspects being omitted are relevant for the purpose under consideration, and the degree of idealisation. Thus, a physics-based equation can be more accurate than another one if, for instance, it describes additional variables of interest in a comprehensive manner, or if it corrects for previous misrepresentations such as parameterisations or model biases. Representational accuracy can be constrained by the available computational means: stronger computational power allows higher resolution, which, in turn, is supposed to generate smaller discretisation errors, thus increasing overall representational accuracy.

Representational accuracy is one evaluative criterion of understanding with a model. But understanding clearly involves further criteria. For instance, a physics-based equation may be better than a mere statistical correlation at describing some of the physical processes at stake in the target system, but can still fail to account for observational data in a satisfactory way (e.g. because of various biases). More criteria are therefore needed to explicate how we can legitimately gain understanding from models: they include empirical accuracy, physical consistency and delimiting the domain of validity.

2.3 Empirical accuracy

Empirical accuracy is evaluated with regard to how well the model outputs match the available observations. This third criterion is evaluated on the basis of the model outputs (while representational accuracy can be assessed on the basis of the model, before even running the computer programme).

However, it is recognized that meeting empirical accuracy is not sufficient to provide understanding. Indeed, a model matching the available data may remain unable to yield explanations of phenomena that occur beyond the domain covered by the available data. Thus, for the model to be adequate beyond this domain (e.g. in view of making future climate projections), one should appeal to additional criteria, in particular coherence with background knowledge, as recently emphasised by Baumberger et al. (2017).

2.4 Physical consistency

Physical consistency of the model outputs constitutes an additional criterion for understanding with a model. In particular, when data is missing, it is worth questioning whether the model outputs are physically plausible, i.e. coherent with background knowledge. This latter typically includes fundamental physical laws like conservation laws, available empirical relationships, and observed behaviour of relevant physical processes.

From our perspective, physical consistency partly contributes to building confidence in climate projections (as rightly defended by Baumberger et al. 2017), in virtue of providing (some degree of) understanding of the relevant phenomena with the considered model. It should be noted that if physical consistency is closely related to (and clearly not independent from) representational accuracy, the two should be clearly distinguished and considered separately (as we will also see in the case studies below in sections 3 and 4)—in particular, representational accuracy involves certain more pragmatic aspects, such as idealisation, that are absent from (and possibly in tension with) the physical consistency criterion.[4] Now, we believe that an additional important criterion, which has not yet been much discussed in relation to understanding, is especially relevant in the climate context.

2.5 Delimiting the domain of validity

Delimiting the domain of validity, we believe, is a genuine token of understanding. Models are not supposed to be reliable in any circumstances but, since they are only partial and idealised representations of the target phenomena, they should be adequate for the specific purpose (or set of purposes) at stake (Parker 2020). If scientists don't know the domain of validity of the model they are using, they encounter the risk of misusing it, e.g., of running the simulation in a physical domain in which the underlying model fails to apply. Insofar as simulations are 'doomed to succeed', they would deliver misleading outputs when used in domains on which they are not supposed to apply.

In particular, users who did not participate in the model development process may not be aware of the precise model idealisations being involved, and of the extent to which they are valid in the target domain. For example, a fluid dynamics model assuming that a fluid is a macroscopic continuum does not apply in transitional flow regime or collisionless flow regime (Meiburg 1986). But there are subtler cases for which the scope of the underlying model can be difficult to assess, as we will discuss in the next sections.

Importantly, we should specify that a narrow scope of validity is no reason to believe that the model does not yield understanding; it yields understanding to some (limited) extent. As soon as one is able to delimit the scope of validity, be it narrow or large, then one certainly gains some understanding with the model. This fifth criterion also bears on the agent's ability and skill of grasping the extent to which the model correctly applies, which may involve the first criterion of understanding we have discussed above, namely intelligibility—here, the ability to manipulate the model and identify qualitatively the model's possible problematic behaviour, in view of evaluating its domain of validity.

[4] In Knüsel and Baumberger (2020), empirical accuracy and representational accuracy are also closely related notions: they see the first as an evaluative criterion for the second. What we mean with representational accuracy here also partly includes what they call—but leave on the side—representational depth. Note that we do not make the distinction between dimensions of understanding and evaluative criteria for understanding, since we reckon that the former, as we define them here, are also directly evaluative.

As we see, the five criteria of understanding we have articulated—intelligibility, representational accuracy, empirical accuracy, physical consistency, delimiting the domain of validity—are not independent of each other; for example, delimiting the domain of validity also relies on the assessment of representational accuracy and empirical accuracy.

However, it could still be thought that intelligibility in particular is independent from the other criteria of adequacy; indeed, some philosophers, e.g. Sullivan (2019), even claim that, in the case of machine learning models, understanding with a model can be achieved without understanding these models better than we currently do. However, we think that this claim is inaccurate. For example, increasing our understanding through manipulability could lead to an understanding of how robust a model behaves under small changes in the input; this, in turn, is relevant for empirical adequacy, say, if the model is not robust and this is an artifact of the model itself, not a feature of the target system.

We will now apply these five evaluative criteria to two important case studies in the climate modelling context, involving statistical downscaling and machine learning methods. We will argue on this basis that understanding (according to the criteria defined here) actually comes in degrees and that, therefore, machine learning methods do not necessarily constitute a radical departure from standard statistical tools, as far as understanding is concerned.

3 Understanding with statistical downscaling

In the face of the climate challenge, information about climate change at the regional scale is crucially needed, in particular in view of adaptation (but also, to some extent, for mitigation). Indeed, adaptation measures naturally take place at the regional scale: for instance, it is the change in precipitation patterns at the local rather than global scale that is relevant in order to devise appropriate measures (e.g. against possible future flooding). In this context, downscaling constitutes a family of methods that aim to provide regional climate change information in view of impact assessments; in this sense, downscaling techniques can furnish decision makers with relevant tools to address (the adaptation side of) the climate challenge.

This section investigates to what extent regional climate modelling—and more specifically: downscaling techniques—can provide some understanding of the target regional climate system, where understanding is articulated using the five criteria we have introduced in section 2. In particular, regional climate modelling and downscaling allow us to focus on the impact of standard statistical techniques on understanding in the climate context. After introducing the idea of downscaling in the main lines—taking the recent climate scenarios for Switzerland CH2018 as an illustration—we discuss to what extent our five criteria for understanding are affected by downscaling techniques.

3.1 Dynamical and statistical downscaling

In the context of producing climate change information, downscaling aims to bridge the modelling gap between, on the one hand, the large scales where global circulation models (GCM) operate and, on the other hand, the regional and local scales relevant for impact assessments. There are two main families of downscaling techniques, namely dynamical downscaling and statistical downscaling. In very schematic terms, dynamical downscaling involves high-resolution regional climate models (RCMs) whose boundary conditions are prescribed ('driven') by GCMs (the RCM is said to be 'nested' into the GCM); in contrast, statistical downscaling aims at identifying empirical-statistical relationships between relevant climate variables at the large and local scales, in view of applying these relationships to future climate projections. Dynamical and statistical downscaling are not exclusive; the two can be combined in a two-step process (dynamical downscaling first, then statistical downscaling, as in the example discussed below). Indeed, from a post-processing perspective, statistical downscaling naturally includes the correction or adjustment of regional climate model output biases with the help of an empirical-statistical link with observations (bias correction, also sometimes denoted 'model output statistics').[5] Within the framework of climate change projections, it is crucial to emphasise that such bias correction makes sense only under the assumption that the empirical-statistical link that is used remains stationary.

3.2 Example: CH2018

These are very general considerations and it can be helpful to briefly discuss a concrete example. We consider the recent climate scenarios CH2018 for Switzerland. These scenarios constitute an example of state-of-the-art regional climate change information in view of climate change impact assessment and decision-making. Moreover, the geographical location, small size and complex Alpine topography make regional climate modelling for Switzerland particularly interesting and relevant. Among other outputs, CH2018 provides localised projections at meteorological stations and on a high-resolution (2 km) grid for various climate variables, for three future 30-year periods, 2020-2049, 2045-2074 and 2070-2099 (1981-2010 being the reference period), and for three standard "Representative Concentration Pathways" (RCPs), which encode the anthropogenic forcing corresponding to different emission scenarios up to 2100, from 2°C compliant mitigation to unabated emissions (RCP2.6, RCP4.5 and RCP8.5, see

[5] Biases in model outputs can have different origins, one of the most obvious being the finite resolution of climate models, leading to various types of model errors at the global, regional and local scales. It is important to emphasise that bias correction "cannot overcome errors from a substantial misrepresentation of relevant processes" (Maraun and Widman 2018, 117), in particular such as global scale circulation biases or missing (or misrepresented) local scale processes (e.g. linked to complex orography, as in the example discussed below); to evaluate precisely when relevant processes are being substantially misrepresented can of course be a tricky issue (especially in the climate change context) and actually lies at the heart of the discussion below.

IPCC 2013, ch. 12.3).[6] These localised projections are based on regional climate models (RCMs) from the standardised ensemble EURO-CORDEX, with boundary conditions prescribed by GCMs from the CMIP5 ensemble.[7] In the post-processing phase, RCM outputs are bias-corrected, in particular using quantile mapping techniques, and further statistically downscaled (to the stations or the high-resolution grid).

The central elements of quantile mapping are (quantile-based) transfer or correction functions matching the model simulation (quantiles) with the reference observations (quantiles) in the historical calibration period; climate change projections can then be bias-corrected using these transfer functions, crucially assuming their time-invariance. In the context of CH2018, quantile mapping also includes a downscaling step for localised projections: it relates the model outputs at a certain scale (e.g. on a certain grid) to observations at a smaller scale, e.g. on a higher-resolution grid or directly to individual meteorological stations. For instance, in the latter case, the quantile mapping reference for daily mean temperature consists of observations of this variable in the period 1981-2010 from 85 weather stations distributed across Switzerland; these reference observations allow for the calibration of the bias correction, which can then be applied to 'raw' regional climate model (simulation) outputs in order to provide bias corrected (and downscaled) regional climate change signals for the considered variable (e.g. mean temperature) at individual meteorological stations (see CH2018, ch. 5).

3.3 Understanding with statistical downscaling

How does downscaling impact our understanding of the regional climate system under consideration and of the related regional climate change signal? We can evaluate this impact using the five criteria for understanding we have defined in section 2. This topic can be seen as part of the discussion on the added value of regional climate modelling and downscaling (e.g. see recently Rummukainen 2016 and Maraun and Widman 2018, in particular ch. 15 & 17), with a focus on the explanatory and understanding-related dimensions. It should also be mentioned that there is actually a plethora of methods that are referred to as downscaling in the climate science literature, and we do not aim to discuss them all in detail (see Table 1 in Hewitson et al. 2014, 546-547 for a good overview). We rather discuss generic features of downscaling—and of statistical downscaling in particular[8]—related to the issue of understanding (partly relying

[6] We closely follow CH2018 here, to which we refer for more details.

[7] CMIP5 is the fifth phase of the Coupled Model Intercomparison Project, which provides a standardised framework for comparing GCM simulations; EURO-CORDEX is the European branch of the Coordinate Regional Climate Downscaling Experiment, which is the regional counterpart of CMIP5 for RCM simulations.

[8] From the point of view of understanding, regional climate modelling involving 'only' dynamical downscaling raises similar issues as global climate modelling (e.g. about model complexity, parameterisation and opacity); in contrast, and to a certain extent, statistical downscaling involves some different—typically statistical—issues for understanding, akin to those encountered in machine learning approaches (see section 4).

on CH2018 as a concrete example).

- *Intelligibility.* Several features of both dynamical and statistical downscaling may hinder the ability to get a sense of the qualitative behaviour of the model through manipulability in counterfactual situations. For instance, regional climate models may inherit biases from their driving GCMs through the boundary conditions and thus have their qualitative behaviour affected—as already mentioned in footnote 8, dynamical downscaling raises similar issues for intelligibility (and understanding in general) as global climate modelling. Statistical downscaling further limits the manipulability of the downscaled model in view of meaningfully addressing what-if-things-would-have-been-different questions, and hence limits the intelligibility of the model: in generic terms, the main worry is that answers to what-if-things-would-have-been-different questions can turn out to be mere statistical artifacts (especially if the stationarity assumption of the empirical-statistical link or bias correction is violated, such as possibly in a climate change context). An example of a statistical artifact, mentioned in CH2018, is that quantile mapping "tends to amplify temperature changes at high elevations along the Alpine ridge and to dampen change signals in valleys", while noting that these "modifications cannot be explained by physical reasoning in a straightforward sense" (85)—this issue is also clearly related to representational accuracy, to which we turn next. However, more generally, this does not mean that statistical downscaling prevents any intelligibility of the downscaled models, but the issue requires a very careful and case-by-case evaluation.[9]

- *Representational accuracy.* Within the dynamical downscaling component, increasing resolution and complexity *can* involve an increase in representational accuracy to some extent, depending on the context (variables, regions, etc.), for instance through the better resolution of the topography of the region under consideration. The important point is that dynamical downscaling can unveil physical mechanisms that are left unseen by lower resolution models; as a consequence, the climate change signal from RCMs and GCMs can be relevantly different (see e.g. the examples discussed in Rummukainen 2016, 151-153). This is in stark contrast with the statistical downscaling component of regional climate modelling, which is mainly statistical and data-driven. In this sense, statistically downscaled models can perform poorly in terms of representational accuracy—and, in general, statistical downscaling does not improve representational accuracy—since the sub-grid physical processes (beyond the scale resolved by the underlying regional climate models) that can be at the origin of the biases to be corrected are simply not considered. As already mentioned above, statistical downscaling can actually introduce certain statistical artifacts and

[9] Various limitations to statistical downscaling (and to the intelligibility of statistically downscaled models) arise in particular in "topographically structured terrain" such as the Alpine region of Switzerland.

additional biases; for example, quantile mapping in CH2018 can "misrepresent spatial climate variability on short timescales" (99). Indeed, it seems that adequate application of statistical downscaling and bias correction requires some prior knowledge of the representational accuracy of the relevant underlying processes as well as of the biases themselves (Maraun et al. 2017).

- *Empirical accuracy.* Generally speaking, the discussion about empirical accuracy is similar in many ways in the case of regional climate modelling as in the case of global climate modelling.[10] An important aspect of the discussion concerns future climate projections, for which there is obviously no direct way to evaluate their empirical accuracy; moreover, empirical accuracy with past and current observations is clearly not a sufficient condition for warranting confidence in future climate (change) projections, see the discussion in e.g. Baumberger et al. (2017).[11] There is also a straightforward sense in which, in principle, statistical downscaling involving bias correction (e.g. such as quantile mapping in the case of CH2018), can directly improve empirical accuracy, since, very schematically, models outputs are corrected toward observations; however, the evaluation of bias correction techniques can be difficult—e.g. cross-validation may not be relevant, because "model and observations are not in synchrony" (Maraun et al. 2017, 765)—especially under climate change conditions (see Maraun and Widman 2018, ch. 15 & 16).

- *Physical consistency.* To the extent that "statistical downscaling methods [...] do not represent the fundamental laws of thermodynamics and fluid dynamics" (Maraun and Widman 2018, 273), physical consistency is not guaranteed a priori in this context, and so needs careful evaluation. Specific aspects of this issue include inter-variable consistency and inheritance of large-scale circulations biases (affecting the driving GCMs) or consistency with the large-scale flow (as described by the driving GCMs) (for instance, see the discussion in CH2018, §5.7 in the case of the climate scenarios for Switzerland; see also Maraun et al. 2017).

- *Domain of validity.* Within the statistical downscaling framework, one of the crucial assumptions in this respect is the stationarity of the climate model biases and hence of the correction function. From a climate change perspective, this assumption is meaningful only up to some point—in very

[10] For instance, empirical accuracy is not the same for all variables; e.g., it is in general better for temperature than for precipitation. It should be noted that, overall, empirical accuracy is better at the global scale than at the regional and local scales.

[11] One reason has to do with the role of calibration of parameter values in achieving empirical accuracy with past and current observations. Another important reason relates to the criteria concerning representational accuracy and the domain of validity (see below): in the context of radically different boundary conditions, such as high forcing scenarios, certain empirical parameterisation procedures may not be valid anymore and important feedbacks may be missing.

rough and intuitive terms: when training or historical data are not representative anymore. But it is extremely difficult to pin down precisely this point in a given concrete situation; consequently, it is extremely difficult to define precisely the domain of validity of (statistical) downscaling techniques for future climate projections.

Evaluating these criteria for understanding suggests that understanding with climate models comes in degrees—and, as we have seen in this section, this is very much so when statistical techniques like statistical downscaling are involved. We will see in the next section that the situation is—perhaps surprisingly—similar when machine learning methods are involved, despite their aura of novelty and opacity.

4 Understanding with machine learning

In this section, we examine how the use of machine learning (ML) in climate models affects our ability to understand with climate models. We will do this with the help of a case study. First, we will briefly review relevant aspects of ML.

4.1 Machine learning: basic ideas and challenges

ML is a technique to automatically extract rules for classification and prediction from data.[12] Here we will only be concerned with supervised learning for prediction. The goal is to automatically find a rule $\hat{f}$ that takes a variable x as an input, and outputs a variable $\hat{y} = \hat{f}(x)$, such that the distance between $\hat{f}(x) = \hat{y}$ and the actual value y is small. In the case we will consider here (Gentine et al. 2018), x is a vector of physical quantities like temperature and humidity at a given time, while the output $\hat{y}$ is a vector of physical quantities like temperature and humidity tendencies.

Gentine et al. (2018) use deep neural networks (DNNs), a kind of ML model (LeCun et al. 2015; Goodfellow et al. 2016). In DNNs, the input is processed through many connected layers of a network. Each layer consists of many nodes. In a simple, fully connected DNN, the value of each node is computed as follows: First, a weighted sum (plus bias) of the values of the previous layer is computed; the resulting value is then further processed through a non-linear activation function. The parameters of a DNN are the weights and biases. The basic idea of supervised learning is to train the DNN on a data set $X = \{(x_i, y_i)\}_{i \in I}$, where x_i is an input instance, and y_i is the correct output for that instance. The DNN is fed with a subset of X, a so-called batch, and computes the output $\hat{y_i}$ for the inputs x_i of this batch. Then the distance between the output $\hat{y_i}$ and the correct answer y_i is calculated using a loss function; this provides us with a

[12] Note that this section provides a standard overview of some basic, well-known facts about machine learning; similar, more detailed accounts can be found in any good introduction to machine learning or deep learning (Hastie et al. 2009; Goodfellow et al. 2016).

measure of the error made by the DNN. In the next, learning step, the weights and biases of the network are adapted to make the error a little smaller. The error correction is propagated backwards through the network. This procedure is repeated until the set X is exhausted. Trained DNNs are evaluated using a test set X', which is drawn from the same distribution as X, but disjoint from it. This procedure is the foundation of the recent surge of applications of deep learning.

Despite their success, many questions with respect to ML models and their use in science remain open. Leading figures in ML research have called DNNs black boxes (Alain and Bengio 2016). One of the main issues is that while DNNs have been applied successfully, a theoretical understanding of this success and many of their properties is still missing (Goodfellow et al. 2016; Vidal et al. 2017). There are also challenges that arise from the use of ML models in the context of climate modelling in particular (Reichstein et al. 2019). These challenges include guaranteeing physical consistency, the heterogeneity and high-dimensional nature of climate data, and obtaining labeled training data for the application of supervised learning techniques.

4.2 Case study: machine learning approach to convective parameterisation

To better understand the advantages and challenges of understanding with ML in climate science, we now turn to a case where DNNs are used in climate modelling (Gentine et al. 2018). In this case, an existing, physics-based climate model is taken as a starting point. Physical sub-models of this climate model are then replaced by DNNs in order to make parameterisation computationally tractable. We first give a short account of this case; then we turn to the question how the use of DNNs affects our ability to understand climate phenomena.[13]

Current climate models with parameterised convection cannot capture some aspects of convection that are relevant to climate predictions. For example, so-called mesoscale convective systems (MCS), systems of thunderstorms of the order of 100 km in diameter, are not accurately represented. Problems with parameterised convection could in principle be overcome by using climate models with higher resolution, specifically, horizontal grids of 2 km or less. Cloud resolving models (CRMs) are able to accurately represent relevant aspects of MCSs. The use of global CRMs would therefore be highly desirable. However, such models are computationally intractable for timescales relevant to climate modelling (several decades and more).

One approach to overcome this problem is superparameterisation (SP). The idea behind SP is to add a layer of fine-grained CRMs to a large-scale global

[13]It could be asked whether this case is representative of the use of machine learning in climate science. It is difficult to answer this question, because the use of DNNs in climate science is still relatively novel, see Reichstein et al. (2019). The results by Gentine et al. (2018) can be interpreted as proof of concept: they show that DNNs have the potential to address some computational problems. Consequently, climate projections under different forcings are not considered.

circulation model (GCM); see Khairoutdinov et al. (2005) for details. The GCM is spatially coarse-grained with horizontal grids of the order of 100 km; the CRMs are more fine-grained, physics-based models, which are embedded into each grid cell. The CRMs run independent of each other, with time-steps of about 20s; the GCM runs on time-steps of about 1h. The two layers are coupled: At each time step of the GCM, the CRMs are forced by the large-scale tendencies of the GCM. The CRMs, in turn, return an average of the locally calculated physical variables to the GCM. In this way, the evolution of the entire model is a mixture of local and global tendencies.

Of course, SP comes with its own problems. SP is computationally expensive, such that several idealisations have to be introduced. For one, there is no direct interaction between the CRMs except via the global level. Some versions of SP are 2D, which means that the physical evolution is only calculated for one horizontal direction, e.g., east-west; in calculating the spatial average, the CRMs output is a statistical sample. Finally, aspects of convection are still not accurately represented by GCMs augmented by SP due to physical idealisations.

This is where machine learning comes in. Gentine et al. (2018) replaced the 8129 CRM modules of an SP model with DNNs in order to obtain a better convection parameterisation. Specifically, they used the so-called SuperParameterised Community Atmosphere Model (SPCAM-3), a well-known GCM, to obtain training and test data for the DNNs. The SPCAM was set up in a configuration which simulates an aqua planet with a resolution of the order of 100 km. 8129 CRMs are embedded in this version of SPCAM; they interact as described above. Gentine et al. (2018) ran this model for a period of two years and made a record of the CRM's input and output. Half of these data, the first year, was set aside for training the DNNs, and the other half as test data (validation). They then prepared 8129 copies of DNNs, and trained them on the training set obtained from the original SPCAM. Finally, they added the trained DNNs to a global model, which they called CBRAIN. This model was validated by comparing the output of CBRAIN with the output of SPCAM-3 of the second year.

The general conclusion drawn by Gentine et al. (2018) is positive. They found that SPCAM-3 and CBRAIN agree surprisingly well. They claim to have demonstrated that DNNs can "skillfully represent many of the effects of unresolved clouds and convection" (Gentine et al. 2018, 5748). In their discussion, Gentine et al. (2018) highlight the computational efficiency of CBRAIN in comparison to SPCAM as the main benefit; tests show that CBRAIN is ten times faster than SPCAM. They identify two main challenges: First, DNNs do not accurately represent some physical properties; in particular, they do not capture energy and moisture conservation, which is required for climate prediction. Second, it is not clear whether CBRAIN will generalise well in situations not represented in the training data; for example, it is not clear whether CBRAIN would accurately represent convection over continents, because it has been trained on data from an aqua planet.

4.3 Understanding with machine learning

What does this case tell us about the prospect of gaining understanding from climate models that use DNNs? We will now evaluate the proposal by Gentine et al. using the five criteria we have outlined in section 2.

- *Intelligibility.* First, it is a general, well-known fact that DNNs are not robust with respect to manipulations of the input. A prominent example for this deficiency is the existence of adversarial examples (Szegedy et al. 2014). These are inputs designed to fool the models into making classification errors; usually, the manipulations are such that, in the case of input images, the perturbed inputs cannot be distinguished from the original by humans. Importantly, while computer scientists know how to construct adversarial examples, they do not yet fully understand why they occur and how models can be made robust against them. Thus, arguably, intelligibility with respect to manipulability in DNNs is low in general. Then, the CBRAIN modellers did not try to improve their understanding of the ML components of the model through manipulating either inputs or parameters. Presumably, one reason for this is that the input and parameter space are just too big to be investigated systematically, and it is not clear what would be achieved by carrying out just a few manipulations. Finally, what researchers working with DNNs do routinely is tweak these models to improve performance, by adapting optimisation procedures and other features of the learning algorithm. However, this does not improve intelligibility per se, because it is mostly geared towards a better performance, not towards understanding how the model behaves in different kinds of circumstances. All these points suggest that intelligibility decreases, or at least does not increase, by using DNNs in CBRAIN.

- *Representational accuracy.* In the case we just considered, representational accuracy is the degree to which the SPCAM, and CBRAIN, respectively, faithfully represent the physical processes producing the output variables, convective heating and moistening, and longwave and shortwave heating rates. The representational accuracy of CBRAIN is lower in comparison to SPCAM. The latter in itself is already highly idealised in several respects; and the GCM part is the same in both models. However, the CRM submodels of SPCAM are based on physical equations. This is not the case for the DNNs that replace the CRMs in CBRAIN. The only objective of the DNNs is to minimise the empirical error with respect to the output variables. There is no requirement for representational accuracy.

- *Empirical accuracy.* By decreasing computational costs, CBRAIN makes it possible to obtain predictions for longer time periods, while the empirical accuracy is comparable to the physical simulation by SPCAM in the periods for which we have empirical data. Thus, for this criterion, there is an increase of understanding because CBRAIN provides comparable empirical accuracy, but, potentially, for a longer period of time, due to

computational gains. However, the increase in understanding has to be qualified, as Gentine et al. (2018) acknowledge: there is no guarantee that CBRAIN is empirically adequate if boundary conditions change, e.g., if the planet had land masses. Thus, there is uncertainty with respect to this criterion. We will return to this point below.[14]

- *Physical consistency.* Gentine et al. (2018) point out that that the DNNs do not intrinsically satisfy energy and moisture conservation, which is a relevant kind of physical consistency. They write: "This can be fine for implementation in a weather forecast model but energy and moisture conservation are required for climate prediction" (Ibid, 5748). Thus, we have a loss of understanding in CBRAIN in comparison to SPCAM, where conservation laws are easier to check. Also, physical consistency becomes more relevant at the timescales of climate predictions.

- *Domain of validity.* Along this criterion, CBRAIN will presumably perform worse than SPCAM. The reason for this is that the DNNs have only been trained on a dataset that represents a very specific type of scenario, viz., an aqua planet with a very specific range of physical variables. As soon as boundary conditions change, or if global temperatures change, there is no guarantee that CBRAIN will still agree with SPCAM. In fact, it is well known that DNNs are very unreliable when applied to out-of-distribution data (see also Kawamleh (2021) on this very limit of DNNs). Thus, the use of DNNs may be particularly problematic in the case of climate change.

At this point, an important feature of this case should be stressed. Gentine et al. (2018) replace physics-based CRMs with DNNs. In order to do this, they first generate training and test data using SPCAM. Thus, this case highlights advantages and drawbacks of understanding with DNNs if we have access to clean, synthetically generated data. However, this is not always the case, as highlighted in Reichstein et al. (2019). If we were to use real data, there are additional problems with obtaining labeled training data, and also with messy, high-dimensional data. Thus, the present case only provides a partial picture of the challenges of understanding with DNNs in the climate context.

5 Understanding in degrees

In this section, we systematically compare how the respective (statistical) methods used in the two case studies affect our ability to understand climate phenomena.

[14] Gentine et al. (2018) do not examine or discuss how CBRAIN would perform with respect to different forcings, or how well it is suited to address the issue of climate change in general. In principle, it is possible to evaluate how well CBRAIN performs in comparison to SPCAM for different boundary conditions, because SPCAM can generate test data for a variety of boundary conditions, and one could evaluate CBRAIN on these test sets.

Before we begin with the comparison, we should clarify what exactly it is that we are comparing. Our question is how the use of the two (sets of) methods—statistical downscaling in RCMs and the use of DNNs as an alternative to superparameterisation—affects our ability to understand with the respective climate models. We should note a dissimilarity between the two case studies: in the machine learning example, we examined the difference between a 'base case', which does not use DNNs (that is, SPCAM) and the case in which DNNs are used (that is, CBRAIN). In contrast, downscaling techniques provide regional climate (change) information at resolutions that are in general not available without them (e.g. by GCMs); in this latter case, we can however investigate to what extent downscaling techniques provide additional understanding compared to GCMs.

How do the two methods affect understanding? Beginning with the criterion of *intelligibility*, both statistical downscaling and DNNs in GCMs do not fare well with respect to manipulability and anticipating the qualitative behavior of the model; the use of both methods arguably makes things worse. The specific reasons—biases in the case of statistical downscaling, errors for small perturbations in the case of DNNs—are different for the two methods, but the common root of the problem is that both run the danger of exhibiting statistical artifacts. Intelligibility may decrease more in the case of DNNs in comparison to statistical downscaling,[15] particularly due to robustness (see section 4).

Along the criterion of *representational accuracy*, we can observe that both statistical downscaling and DNNs in GCMs create challenges because they are not designed to capture the processes producing the output variables. In a nutshell, both are statistical techniques that try to reproduce input-output patterns. However, in both cases, there is not a total loss of representational accuracy. In the case of statistical downscaling, the GCM, as well as the dynamical downscaling step, are based on physical equations and thus representationally accurate to a certain extent. The same is true for the GCM in the second case, which is combined with DNNs.

Turning to *empirical accuracy*, we see that there is a (qualified) increase in both cases. In the case of statistical downscaling, variables are bias-corrected to match observations, which increases empirical accuracy. The use of DNNs makes it possible to obtain predictions that are comparable to predictions without DNNs, but for longer periods (thanks to their smaller computational costs compared to superparameterisation). Thus, arguably, empirical accuracy is increased. In both cases, a qualification related to the domain of validity has to be added: it is difficult to gauge the empirical accuracy of both methods for climate (change) projections.

At this point, it should be stressed that the purpose of using the methods

[15] The precise extent to which statistical downscaling methods allow for some manipulability (and hence for some intelligibility) is an open (and a case-by-case) issue; there is actually a call in the climate modelling community for designing "ensembles of statistical downscaling methods or even ensembles combining GCMs, RCMs and a range of statistical methods" (Maraun and Widman 2018, 284)—such ensembles would help to get a clearer picture on the manipulability issue.

in the two cases is different. In the case of statistical downscaling in RCMs, the goal is to adapt a GCM to a regional context such that climate projections can be obtained at a higher spatial resolution. Accordingly, empirical accuracy increases because model output is changed to match observations. In the case of DNNs, model output is not corrected. Rather, the goal is to overcome a computational deadlock, and DNNs make it possible to obtain predictions which would not be available otherwise. Still, both methods are geared towards improving predictive capabilities—higher precision in one case, more predictions in the other.

Physical consistency is not guaranteed in both cases. This is certainly a drawback in comparison to physics-based modelling. Lack of physical consistency can lead to problems in the context of climate predictions specifically. However, it should be noted that while physical consistency is not guaranteed due to the statistical nature of the methods—this is related to the lack of representational accuracy—we can expect that physical consistency is satisfied to a certain degree, because we have a certain degree of empirical accuracy.

Finally, the *domain of validity* criterion is a reason for concern in both cases. For both statistical downscaling and DNNs in GCMs, it is necessary to make stationarity assumptions, viz., that training or observational data are valid for future scenarios (this is critical in the context of climate change projections). In both cases, it is extremely hard to pin down when we should expect this assumption to be violated. One difference between the two methods is that DNNs are known to be very sensitive to changes in the underlying distribution, i.e., to non-stationarity. We should expect that applying DNNs outside a known domain will lead to inaccurate predictions along all criteria discussed above.

Taking stock, we can see that both statistical downscaling and the use of DNNs affect the five criteria of understanding in a qualitatively similar manner. Both are statistical methods that are not based on physical principles, and thus may involve a decrease of understanding along the criteria of intelligibility, representational accuracy and physical consistency. The strength of both methods is to increase empirical accuracy. This supports the thesis that there is not a categorial difference between machine learning methods and more traditional statistical techniques, as far as understanding is concerned. This does not mean, however, that the two methods are on a par.

The two cases also support the thesis that understanding as a whole comes in degrees. All five criteria we considered should not be interpreted as yielding categorial results, but relative increases or decreases. In both cases, the methods are applied in the context of complex models, and in combination with physics-based components, idealisations, further statistical techniques, and thus do not affect our ability to understand with these models in a categorical manner, but gradually. In particular, it is wrong to say that the use of statistical downscaling, or of machine learning, leads to an overall loss of understanding.

We can also observe that in both cases, there are similar tradeoffs between the five criteria of understanding. In a nutshell, the tradeoff is between an increase of empirical accuracy and a decrease along the other four criteria. This also implies that we cannot interpret the use of any one of these methods as

yielding a loss or a gain in understanding *tout court*. The five criteria we have proposed in this article precisely aim to articulate and shed some light on these tradeoffs as well as the various aspects of understanding. And the latter, in turn, show that, with respect to understanding, regional climate models obtained from 'common' statistical downscaling[16] and climate models using 'fancy' machine learning methods such as deep neural networks are actually part of the same continuum, where the various criteria of understanding come in degrees.

Finally, a word of caution is in order. We have not argued that the use of statistical techniques or of machine learning in the context of climate models is unproblematic because these methods do not necessarily lead to a loss of understanding. Rather, we have argued that these methods do affect certain criteria of understanding, and in particular what could be called the physical (and explanatory) criteria of understanding. This, of course, may well be problematic, because, for instance, a loss of process understanding affects our confidence in climate predictions (see Baumberger et al. 2017).

6 Conclusion

A central goal of climate modelling is to provide projections in view of decision-making with respect to climate change. But understanding is not secondary. Understanding is indispensable to appropriately evaluate climate models and to build confidence in climate projections.

Contemporary techniques in climate modelling, including statistical methods and machine learning techniques, are, to some extent, like black boxes and, despite the fact that they can considerably enhance our predictive abilities, they affect our ability to understand with climate models.

In order to assess the impact of statistical methods and machine learning techniques on understanding with climate models, we have articulated five criteria for understanding: intelligibility, representational accuracy, empirical accuracy, physical consistency and delimiting the domain of validity.

We have argued that these criteria are not categorical, but come in degrees. We have put these five criteria to work in two important case studies in the climate context. In the first case, we have investigated (statistical) downscaling techniques, which play a crucial role in the elaboration of regional climate change information and impact assessments. In the second case, we have contrasted these standard statistical techniques with the machine learning approach in climate modelling, focusing specifically on the use of deep neural networks as an alternative to superparameterisation in a global circulation model.

The main upshot of the paper is a twofold continuity of the multidimensional and graded notion of understanding in the climate modelling context. First, the use of machine learning decreases understanding along some criteria; however, the same tendencies can also be observed for more standard statistical methods such as those involved in downscaling, showing that there is no categorical

[16] Downscaling techniques are applied in weather forecasting since the late 1950s (see Maraun and Widman 2018, ch. 3).

difference between the two cases, as far as understanding is concerned. Second, we have highlighted the tradeoff between an increase in empirical accuracy (the main focus of both the statistical and machine learning methods in the climate context) and a decrease along the criteria of intelligibility, representational accuracy, physical consistency and delimiting the domain of validity.

References

Alain, G. and Bengio, Y. (2016). Understanding intermediate layers using linear classifier probes. arXiv:1610.01644v4.

Baumberger, C. (2019). Explicating objectual understanding: Taking degrees seriously. *Journal for General Philosophy of Science*, 50:367–388.

Baumberger, C., Knutti, R., and Hirsch Hadorn, G. (2017). Building confidence in climate model projections: An analysis of inferences from fit. *WIREs Climate Change*, 8:e454.

CH2018 (2018). *Climate Scenarios for Switzerland, Technical Report*. Zurich: National Centre for Climate Services.

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

de Regt, H. W. and Dieks, D. (2005). A contextual approach to scientific understanding. *Synthese*, 144:133–170.

Gentine, P., Pritchard, M., Rasp, S., Reinaudi, G., and Yacalis, G. (2018). Could machine learning break the convection parameterization deadlock? *Geophysical Research Letters*, 45:5742–51.

Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep Learning*. MIT Press.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning*. Springer Series in Statistics. Springer, second edition.

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

Hempel, C. G. and Oppenheim, P. (1948). Studies in the logic of explanation. *Philosophy of Science*, 15(2):135–175.

Hewitson, B. C., Daron, J., Crane, R. G., Zermoglio, M. F., and Jack, C. (2014). Interrogating empirical-statistical downscaling. *Climatic Change*, 122:539–554.

IPCC (2013). *Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge: Cambridge University Press.

Kawamleh, S. (2021). Can Machines Learn How Clouds Work? The Epistemic Implications of Machine Learning Methods in Climate Science. *Philosophy of Science*, 88(5).

Khairoutdinov, M., Randall, D., and Demott, C. (2005). Simulations of the atmospheric general circulation using a cloud-resolving model as a superparameterization of physical processes. *Journal of the Atmospheric Sciences*, 62:2136–54.

Knüsel, B. and Baumberger, C. (2020). Understanding climate phenomena with data-driven models. *Studies in History and Philosophy of Science Part A*, doi.org/10.1016/j.shpsa.2020.08.003.

Knutti, R. (2018). Climate model confirmation: From philosophy to predicting climate in the real world. In Lloyd, E. A. and Winsberg, E., editors, *Climate Modelling: Philosophical and Conceptual Issues*, pages 325–359. Cham: Palgrave Macmillan.

Kuorikoski, J. (2011). Simulation and the sense of understanding. In Humphreys, P. and Imbert, C., editors, *Models, Simulations, and Representations*, chapter 8, pages 250–273. Routledge.

Kuorikoski, J. and Ylikoski, P. (2015). External representations and scientific understanding. *Synthese*, 192:3817–3837.

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature*, 521:436–44.

Lenhard, J. and Winsberg, E. (2010). Holism, entrenchment, and the future of climate model pluralism. *Studies in History and Philosophy of Science Part B*, 41(3):253–262.

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

Maraun, D. and Widman, M. (2018). *Statistical Downscaling and Bias Correction for Climate Research*. Cambridge: Cambridge University Press.

Maraun et al. (2017). Towards process-informed bias correction of climate change simulations. *Nature Climate Change*, 7:764–773.

Meiburg, E. (1986). Comparison of the molecular dynamics method and the direct simulation monte carlo technique for flows around simple geometries. *Physics of Fluids*, 29:3107–3113.

Parker, W. S. (2014). Simulation and understanding in the study of weather and climate. *Perspectives on Science*, 22(3):336–356.

Parker, W. S. (2020). Model evaluation: An adequacy-for-purpose view. *Philosophy of Science*, doi:10.1086/708691.

Reichstein, M., Camps-Valls, G., Stevens, B., Jung, M., Denzler, J., Carvalhais, N., and Prabhat (2019). Deep learning and process understanding for data-driven earth system science. *Nature*, 566:195–204.

Rummukainen, M. (2016). Added value in regional climate modeling. *WIREs Climate Change*, 7:145–159.

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

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., and Fergus, R. (2014). Intriguing properties of neural nets. arXiv:1312.6199v4.

Trout, J. (2002). Scientific explanation and the sense of understanding. *Philosophy of Science*, 69:212–233.

Vidal, R., Bruna, J., Giryes, R., and Soatto, S. (2017). Mathematics of deep learning. arXiv:1712.04741.

Wilkenfeld, D. A. (2017). Muddy understanding. *Synthese*, 194(4):1273–93.