Model spread and progress in climate modelling*

Julie Jebeile[†1,2] and Anouk Barberousse[3]

[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]UMR Sciences, normes, démocratie, Sorbonne Université, 1 rue Victor Cousin, 75005 Paris, France

To be published in the *European Journal for Philosophy of Science*, **2021**

Abstract. Convergence of model projections is often considered by climate scientists to be an important objective in so far as it may indicate the robustness of the models' core hypotheses. Consequently, the range of climate projections from a multi-model ensemble, called "model spread", is often expected to reduce as climate research moves forward. However, the successive Assessment Reports of the Intergovernmental Panel on Climate Change indicate no reduction in model spread, whereas it is indisputable that climate science has made improvements in its modelling. In this paper, after providing a detailed explanation of the situation, we describe an epistemological setting in which a steady (and even slightly increased) model spread is not doomed to be seen as negative, and is indeed compatible with a desirable evolution of climate models taken individually. We further argue that, from the perspective of collective progress, as far as the improvement of the products of a multi-model ensemble (e.g. means) is concerned, reduction of model spread is of lower priority than model independence.

Keywords. climate modelling; model spread; scientific progress; robustness; model independence.

*This work was supported by "MOVE-IN Louvain" Incoming Post-doctoral Fellowship, cofunded by the Marie Curie Actions of the European Commission, and by the Swiss National Science Foundation project PP00P1_170460 "The Epistemology of Climate Change". We thank the anonymous reviewers for their helpful comments. We also thank the participants, Gab Abramowitz, Mathias Frisch and Eric Winsberg, as well as the audience of the symposium "Diversity, Uncertainty, and Action: Coping with a Plurality of Climate Models" held in Seattle at the Philosophy of Science Association (PSA) 2018 conference, for their feedback on the oral presentation of the paper. JJ is in debt to Michel Crucifix and Samuel Somot for insightful discussions on issues related to the addressed topic.

[†]Corresponding author, julie.jebeile@philo.unibe.ch

Contents

1 Introduction

The complexity of the climate is such that, currently, no single model accurately captures every relevant component of the system. Leeway remains as regards model specification and implementation, thus fostering the worldwide production of multiple models. In particular, multi-model ensembles bringing together General Circulation Models (GCMs) embody the tremendous efforts made in climate science to produce projections about future changes and support decision-making today. GCMs are used to represent climate dynamics under specific emissions scenarios. They are built on a common physical basis, but differ in the climate components they represent,[1] and in the idealisations and parameterisations of which they make use.[2]

The range of climate projections from a multi-model ensemble with respect to a given emissions scenario is called "model spread". This spread is an effect of the above-mentioned differences between models. Our main object of inquiry in this paper is the model spread of the multi-model ensemble employed in the Coupled Model Intercomparison Project (CMIP).[3] This model spread

[1] State-of-the-art GCMs minimally include components of the physical climate, i.e., atmosphere, ocean and their interactions, as well as external forcing induced by the Sun, volcanoes and human activities. They can be Coupled Atmosphere-Ocean General Circulation Models (AOGCMs) or Earth System Models (ESMs). Unlike the former, the latter include ice sheets and biogeochemical processes.

[2] GCMs differ in the discretisation of the physical differential equations, in the grid resolution, in the parameterisations, in the possible inclusion of stochastic components, etc. Importantly, variations between models depend on their respective parameterisations, which are approximate descriptions of subgrid-scale processes (e.g. cloud processes) used when explicit representation of these processes requires very high computer power, or when understanding of these processes is simply lacking.

[3] Nonetheless, the conclusions in this paper should extend to model spreads from ensembles of regional models, e.g. EURO-CORDEX ensembles.

corresponds to the range of projections from several dozen GCMs[4] used as a database for the Assessment Reports (henceforth ARs) of the Intergovernmental Panel on Climate Change (IPCC), and more precisely for Working Group I.

It seems reasonable to expect that a positive evolution of climate models should come with a reduction in the model spread. The assumption here is that improved models – including better-described components – should progressively converge in their projections. This assumption results from seemingly sound robustness reasoning: if, despite being based on different hypotheses, models converge more in their projections, this is to be interpreted as the projections being more likely. Convergence of climate projections is indeed a major feature in the assessment of the robustness of models' core hypotheses, which is often supposed to ground one's confidence in models.

However, the successive IPCC ARs do not manifest any reduction in the model spread exhibited by CMIP ensembles (Knutti and Sedláček 2012), even though, in recent decades, models have evolved in various ways: understanding of underlying processes has improved, computational power has increased, thus higher spatial resolution has been possible and more processes have been integrated. This contrast with the model improvements is so important that it is explicitly addressed in one of the IPCC's FAQs: FAQ 1.1 "*If Understanding of the Climate System Has Increased, Why Hasn't the Range of Temperature Projections Been Reduced?*." The forthcoming CMIP6 even displays a slight increase in the updated model spread (Hausfather 2019). Does this indicate a problem in the evolution of climate modelling?[5]

The aims in this paper are to provide a detailed explanation of the fact that model spread is remaining steady, and to assess whether, from a normative point of view, climate scientists should actually strive to reduce model spread. First, in section 2, we explain why model spread remains steady and can even (temporarily) increase despite better process understanding and higher computer power. Second, in section 3, we describe an epistemological setting in which a steady (or even slightly increased) model spread need not be seen as a sign of failure, and is indeed compatible with a desirable evolution of climate models taken individually. In section 4 we further argue that, from the perspective of *collective* progress, as far as the improvement of the products of a multi-model ensemble (e.g. means) is concerned, the reduction of model spread is of lower priority than model independence.

[4]As an illustration, CMIP5 put forward four scenarios of anthropogenic forcing called "Representative Concentration Pathways" based on different assumptions about future global greenhouse-gas emissions. Twenty-three models contributed to CMIP5, and eighty-eight models are currently running for CMIP6.

[5]In addition to the methodological problem we point out, this also raises concerns from the perspective of public communication. A steady model spread and the even more baffling slight increase may engender public doubt about the progress being made in climate modelling, in that model spread is usually interpreted as quantifying uncertainty.

2 Why model spread does not reduce despite model improvements

Since the Charney report (1979), estimated uncertainty ranges on global averages – i.e. model spreads – have remained largely unchanged, despite improvements in the GCMs used for the IPCC ARs (Knutti and Sedláček 2012). In particular, climate sensitivity, i.e. the response of the climate system to a doubling of CO_2, remains in the same range. For instance, with an estimated 66% chance of occurring, likely equilibrium climate sensitivity ranges from 1.5°C to 4.5°C according to CMIP3 for the IPCC AR4; from 2.1°C to 4.7°C according to CMIP5 for the IPCC AR5; and between 1.8°C and 5.6°C according to CMIP6 for the IPCC AR6. In other words, the model spread for climate sensitivity has in fact increased with CMIP6 (Hausfather 2019). As an illustration, Figure 1 compares the global mean temperature change in the scenarios of IPCC AR4 and of AR5, and shows how similar their respective model spreads are.

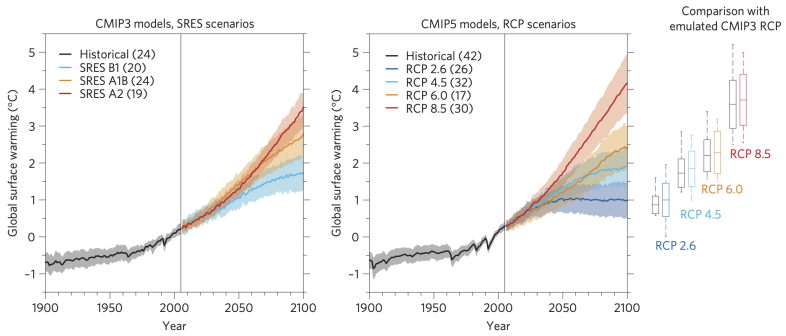

Figure 1 | Global temperature change and uncertainty. Global temperature change (mean and one standard deviation as shading) relative to 1986–2005 for the SRES scenarios run by CMIP3 and the RCP scenarios run by CMIP5. The number of models is given in brackets. The box plots (mean, one standard deviation, and minimum to maximum range) are given for 2080–2099 for CMIP5 (colours) and for the MAGICC model calibrated to 19 CMIP3 models (black), both running the RCP scenarios.

Figure 1: Global temperature change and model spread for CMIP3 and CMIP5 (Knutti and Sedláček 2012)

Why does the model spread not reduce despite the improvements in the models themselves? In order to explain the non-reduction of the model spread, we analyse the spread in terms of the quantification of uncertainty, and examine how each major model improvement – better process understanding and higher computer power – affects the model spread while aiming at overcoming shortcomings in individual models.

2.1 Quantification of uncertainty

The model spread is the range of projections of average global quantities (e.g. mean surface temperature, precipitation, sea level, climate sensitivity) obtained from multiple models. It is defined as the range of projections provided by a

particular ensemble of models and with respect to a *given* emissions scenario (defined by one or several socioeconomic storylines).

For a given scenario, each projection comes with a variety of uncertainties, including internal variability uncertainty and model uncertainty. While internal variability uncertainty is inherent to the climate system, due to the chaotic and spontaneously varying nature of the climate itself, model uncertainty stems from the models being imperfect representations. More precisely, for a given model, the best values to assign to model parameters are either unknown or ambiguously defined; this sub-category of model uncertainty is called parameter uncertainty. Another significant sub-category of model uncertainty is structural uncertainty. It is induced by the idealisations, including the parameterisations, used to represent specific processes at work in the climate system.

Idealisations vary from model to model "in terms of the fundamental numeric and algorithmic structures, forms and values of parameterisations, and number and kinds of coupled processes included" (Collins et al. 2013, 1039). Such diversity of idealisations, it is often assumed, produce alternative and allegedly equally plausible models (Collins et al. 2013, 1036). Together these models generate a collection of projections whose spread – i.e. the model spread – is supposed in turn to quantify uncertainty (see also IPCC FAQ 12.1 2013). A probability density function can also be calculated for each variable of interest from the average projections and the assigned deviation of the ensemble members. This function, similar to Gaussian curves, aims to indicate the chance of occurrence of the values of said variable. Furthermore, emergent constraint can also be derived at this stage. This is a "physically explainable empirical relationship between intermodel variations in a quantity describing some aspect of recent observed climate ... and the intermodel variations in a future climate prediction of some quantity" (Klein and Hall 2015, 277), which can in turn be used to constrain model projections under specific conditions.

In this interpretative framework, the model spread is used as a quantification of structural and parameter uncertainty, and has been so used since the Charney report (1979).[6] It is important to highlight that the model spread is an *estimation* of the model uncertainty given the set of available models, but should not be interpreted merely as the result of all the errors made in the modelling assumptions, among which are, importantly, the parameterisations. Thus, the increase of the model spread is not necessarily due to an amplification of errors in an absolute way but can instead indicate that the estimation of the model uncertainty is improved as it can result from the inclusion of previously unquantified sources of error. In the same way, a steady model spread should not be interpreted as the uncertainty itself remaining constant.

In that sense, we now examine how the most significant kinds of improvements impact the *uncertainty estimates* – while, we assume, these improvements generally aim to handle various shortcomings and associated sources of uncer-

[6]We should note that an additional estimation of the uncertainty is in practice applied in the model spread based on expert judgments (using e.g. a Bayesian probabilistic methodology) since the multi-model ensemble is an "ensemble of opportunity" that imperfectly spans model uncertainty (e.g. Thompson et al. 2016).

tainty in models (e.g. parameterisations, spatial resolution, omission of relevant processes). The model spread has not decreased despite worldwide efforts at improving models and at integrating these improvements within CMIP. Our aim here is to provide an explanation of this fact.

2.2 Process understanding

The evolution of climate models in recent decades has been built upon improvements in process understanding – thus allowing for better-described components – and in computer power – thus allowing for higher resolution and progressive integration of additional processes.[7]

Better process understanding has been possible thanks to continuing spatial, paleoclimatic and historical-records-based investigations, measurements and data analysis, enhancing the collection of data about the cryosphere, atmosphere, land, biosphere and ocean systems, and thereby feeding into further theoretical work. In particular, climate scientists have gained better understanding of "the role of clouds, sea ice, aerosols, small-scale ocean mixing, the carbon cycle and other processes" (IPCC FAQ 1.1 2013).

Larger collections of data can constitute a more solid empirical basis for model validation. As emphasised by the IPCC, "More observations mean that models can now be evaluated more thoroughly, and projections can be better constrained" (IPCC FAQ 1.1 2013). Such observational constraints might reduce the model spread as this has been recently documented (e.g. Brunner et al. 2020; Ribes et al. 2021). Furthermore, better process understanding allows for less idealised representations of the different physical and biogeochemical mechanisms within models. It notably allowed for "an elimination of artificial adjustments to atmosphere and ocean coupling (so called 'flux adjustment')" (IPCC FAQ 9.1 2013, "*Are Climate Models Getting Better, and How Would We Know?*"). From a general perspective, there is some expectation that, with better process understanding, parameterisations can be replaced by explicit theoretically-based equations of subgrid-scale processes, and, as a result, that projections from each individual model can become more accurate. That said, it may not necessarily ensue a reduction in the model spread, even though the IPCC expects an – albeit slow – evolution in this direction.

> The uncertainty range around projected [greenhouse gas (GHG)] and aerosol precursor emissions (which depend on projections of future social and economic conditions) cannot be materially reduced. Nevertheless, improved understanding and climate models – along with observational constraints – may reduce the uncertainty range around some factors that influence the climate's response to those emission changes. The complexity of the climate system, however, makes this a slow process. (IPCC FAQ 1.1 2013)

[7] Note that the evolution of climate models is intrinsically limited by the internal variability of the climate and constrained by the choice of emissions scenarios.

Improvements in process understanding are expected to reduce the uncertainty due to the parameterisations in each individual model, and might result in a corresponding decrease in model spread. To our knowledge, there is at least no reason to believe that better-described components in individual models would increase the model spread. That said, in the near future, it would be worth scrutinising the Cloud Feedback Model Intercomparison Project (CFMIP) which deals with a major source of spread, i.e. cloud feedback (Webb et al. 2017), and investigating how the forthcoming outcomes will influence the model spreads. Today, as we have seen, the model spread of the CMIP remains largely unchanged. The reason, we now argue, stems mainly from the integration of more processes in models.

2.3 Computer power: high resolution and comprehensiveness

The power of supercomputers in recent decades has increased from megaflops to petaflops and this trend is expected to continue. Such technical progress allows for more calculation steps in a given computing time. Higher computer power thus can allow for higher spatial resolution in models, and thereby the inclusion of relevant processes occurring at finer scales in models. Higher computer power can also enable scientists to produce increasingly comprehensive models, integrating and coupling more processes, heterogenous in nature, into their models. While higher resolution and more comprehensiveness address shortcomings in previous individual models, they can both introduce new sources of errors.

The current endeavour in climate model development is to construct higher-resolution models in order to inform policy-making at local scales of interest. Higher resolution is supposed to enable one to describe relevant subgrid-scale process (e.g. convection, cloud formation). While it can help to reduce model spread (e.g. Fosser et al. 2020), it can also increase model spread as recently shown (Pichelli et al. 2021). The possible increase of the model spread due to higher resolution can be explained as follows: higher resolution allows for more inclusion of small-scale processes and thereby can introduce additional sources of error related to the representations of those processes.[8]

Simultaneously, the general tendency in climate model development is to construct increasingly comprehensive models, integrating more and more processes (e.g. carbon cycle, dynamic vegetation), in response to the political need for reliable projections (see e.g. Dahan 2010). Figure 2 illustrates this tendency.

Integrating aspects of the climate that have been previously neglected is supposed to produce better representations – i.e. less idealised representations – of the system overall. However, as recognised by the IPCC, this greater integration may also incorporate new "sources of possible error".

> Climate models of today are, in principle, better than their predecessors. However, every bit of added complexity, while intended

[8]Interestingly, in Pichelli et al. 2021, the estimated ensemble means are more accurate, which, as we argue in the rest of the paper, is an important progress in climate modelling.

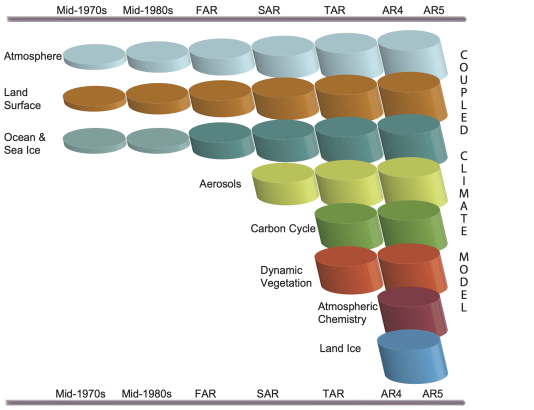

Figure 1.13 | The development of climate models over the last 35 years showing how the different components were coupled into comprehensive climate models over time. In each aspect (e.g., the atmosphere, which comprises a wide range of atmospheric processes) the complexity and range of processes has increased over time (illustrated by growing cylinders). Note that during the same time the horizontal and vertical resolution has increased considerably e.g., for spectral models from T21L9 (roughly 500 km horizontal resolution and 9 vertical levels) in the 1970s to T95L95 (roughly 100 km horizontal resolution and 95 vertical levels) at present, and that now ensembles with at least three independent experiments can be considered as standard.

Figure 2: The development of climate models over the last 35 years (Stocker et al. 2013, 144)

> to improve some aspect of simulated climate, also introduces new sources of possible error (e.g., via uncertain parameters) and new interactions between model components that may, if only temporarily, degrade a model's simulation of other aspects of the climate system. (IPCC FAQ 9.1 2013)

We contend, indeed, that higher-resolution and more comprehensive models may generate an increase in the model spread that is probably not compensated by improvements in process understanding. In order to understand why this is so, it is useful to follow Le Treut's (2009) two-part analysis of climate model components: models are analysed as composed of well-understood and stable components, on the one side, and lesser-understood and yet sensitive components, on the other.

Well-understood and stable components of climate models derive from fundamental principles including conservation laws and orbital mechanics. They enable climate scientists to describe dominant phenomena in the climate, e.g. that greenhouse gases absorb the infrared radiation emitted by the Earth and warm up the atmosphere. They are well-confirmed because they are based on parts of physics that are themselves well-confirmed and of longstanding.

Lesser-understood and yet sensitive components concern the behaviour of possible amplifiers of the greenhouse effect, i.e. water vapour, the carbon cycle, methane, and clouds or ice albedo. They include cloud feedback, the effect of polar ice sheets on sea level, the effect of aerosols on clouds, the effect of solar wind on clouds, and climate–ecosystem feedback. They are required in order to obtain more detailed projections, but are also responsible for uncontrolled amplification and dampening effects within the models – i.e. these components

have the unintended effect of amplifying (resp. dampening) the contribution of other components in such a way that the processes in which they are involved within the models are difficult to trace. This results in uncontrolled variability within the outputs of individual models. These sensitive components are not included within the fundamental equations of the models, unlike the stable components, but are instead included in the form of parameterisations.

Sensitive components tend to be of higher importance in current modelling because of the general tendency to produce higher-resolution and more comprehensive models. The more comprehensive a model is, the more it contains non-stable, sensitive components that may produce higher variability among their outputs, therefore increasing the estimates of structural uncertainty and parameter uncertainty. Because distinct models contain different lesser-understood and yet sensitive components, discrepancies between their respective projections are likely to increase.

In a nutshell, better process understanding and higher computer power aim to overcome previous shortcomings in individual models. Yet at the same time the development of higher-resolution and increasingly comprehensive models can amplify the model spread.

3 Why the reduction of model spread should not be an unconditional objective

Now that we have provided an explanation of the steady level (and even slight increase) of the model spread, the question arises: are climate scientists actually on the right track in integrating more processes into their models, or should they strive to reduce the model spread instead? In this section, we tackle the normative question of whether convergence of model projections is a priority for the purpose of improving models.

On the one hand, convergence of model projections is often considered by climate scientists to be an important objective in so far as it may indicate the robustness of the models' core hypotheses. In the philosophy of climate science, Parker (e.g. 2011; 2013) is skeptical about the fact that the agreement between climate models guarantees robustness. As she highlights, climate models are all imperfect representations, and therefore none can claim to be truth-conducive – assuming that a model within the ensembles is required to have this property. Furthermore, ensembles of climate models are hardly random collections of independent models. But, more commonly, philosophers argue that, if models developed by different centres agree in their outputs, *and* if the models' hypotheses are independent of each other, then the models' hypotheses may be deemed robust although models might still exhibit structural uncertainty (Lloyd 2009, 2010, 2015; Vezér 2016, 2017). In particular, for Lloyd, this grounds our confidence that the "causal core" (Lloyd 2015) shared by climate models captures the fundamental behaviour of the climate system, and that models can in turn be relied upon in producing further projections. "But if models disagree, this

indicates model 'unreliability.' In addition, if the mean model average disagrees with the mean observation, it suggests a systematic deficiency in the models" (Lloyd 2010, 979). On Lloyd's view, an important additional specification is the variety of evidence that empirically supports model hypotheses and parameterisations. For Winsberg (2018), the more the models agree with each other, the higher our degree of confidence in the projections can be, provided model hypotheses are *RA diverse* (for Robustness Analysis) (following Schupbach 2018's analysis), in the sense of being able to rule out "competing hypotheses" or "rival explanations".

Following the predominant philosophical view on robustness in climate science, convergence of models' projections is interpreted as due to a general improvement of the capacity of each individual model to provide reliable projections, itself due to an improvement of the hypotheses underlying the parameterisations. Here, convergence of projections indicates that the models' hypotheses are not only robust but are also being refined as time goes by. In other words, additional idealisations and parameterisations – including sensitive components – are expected at some point to produce similar effects on the outputs. Such refinement should lead to the reduction of the model spread over time.

This view may lead one to conclude that the steady (or slightly increased) model spread is an indication that the up-to-date higher-resolution and more comprehensive models are unreliable, unless we take into account the explanation provided in section 2.

On the other hand, if one holds that climate scientists should aim for higher-resolution and more comprehensive models as a priority, then one may still view the introduction of sensitive components as something positive, since these components concomitantly improve the understanding of detailed processes. As the IPCC remarks, increases of the model spread "merely reflect the quantification of previously unmeasured sources of uncertainty".

> As science improves, new geophysical processes can be added to climate models, and representations of those already included can be improved. These developments can appear to increase model-derived estimates of climate response uncertainty, but such increases *merely reflect the quantification of previously unmeasured sources of uncertainty* [...]. As more and more important processes are added, the influence of unquantified processes lessens, and there can be more confidence in the projections. (IPCC FAQ 1.1 2013, *emphasis ours*)

Therefore, the increase in model spread can be interpreted in terms of model improvements that are in agreement with the overall progress of climate modelling. Indeed, this is simply part of the evolution of climate modelling, by which the climate models are deemed to improve because they come to include more and more climate components, which itself tends to increase the estimated value of model uncertainty.

On this view, as climate research moves forward, one might even expect that the sensitive components will become better understood and be included among

the stable components. This in turn would reduce model spread. In other words, the situation might just be temporary. This possibility is suggested by Knutti and Sedláček:

> defining progress in climate modelling in terms of narrowing uncertainties is too limited. Models improve, representing more processes in greater detail. This implies greater confidence in their projections, but *convergence may remain slow*. The uncertainties should not stop decisions being made. (Knutti and Sedláček 2012, 1, *emphasis ours*)

This positive interpretation of the actual evolution of the model spread invites us to posit an epistemological setting in which reduction of model spread is not deemed a *priority*, and is treated as on a par with other criteria in the assessment of the progress of climate models. We will argue that the robustness of models' core hypotheses should be considered alongside other indicators of improvement: the evolutions of model fit, of variety of evidence (Lloyd 2009, 2010), and of consistency with background knowledge (Baumberger et al. 2017).[9]

Reduction of sources of error, where available, of course improves individual GCMs; uncertainties can be assessed by comparing model outputs with empirical data about past and present climate. Thus, an important indicator is the *evolution* of the model fit. Model fit measures the discrepancies between simulated quantities (e.g. global distributions of temperature, precipitation, radiation etc.) and available past and present observations, via quantitative statistical measures referred to as performance metrics (Reichler and Kim 2008).

In this respect, consistency with background knowledge is also an important indicator. As Baumberger et al. 2017 argues, when data are lacking, model projections being consistent with available scientific laws and empirical correlations provides an additional reason to believe that models capture the fundamental behaviour of the climate system beyond the set of available data.

Furthermore, models improve if they yield outcomes or retrodictions concerning more and more variables about the past and present climate. Thus, another indicator is the *increased number* of variables that are accurately simulated by models, e.g. not only global mean temperature, but also "precipitation, radiation, wind, oceanic temperatures, and currents" (Lloyd 2009, 217); this is related to the *evolution* of the variety of evidence (see Lloyd 2009).

The evolution of the model spread should be confronted with the indicators of improvement that we just mentioned. Since model fit, consistency with background knowledge and variety of evidence are steadily getting better from one generation of models to the next, the nonimprovement (or slight increase) of model spread, instead of pointing toward a pessimistic conclusion about the evolution of climate modelling, cannot but be an indication that progress has been attained overall. By contrast, a *significant* increase of the model spread would have been a serious indicator of something going wrong. Overall, climate

[9]These authors discuss model fit, variety of evidence or consistency with background knowledge as criteria of confirmation, whereas we take the respective evolutions of model fit, variety of evidence and consistency with background knowledge as indicators of improvement.

modelling being a complex endeavour, its progress has to be assessed based on a variety of indicators, among which model spread is only one potential indicator.[10]

To sum up, we have argued the following. First, it would be methodologically misleading to consider reduction of model spread as the priority to aim for. Second, model spread can remain steady without threatening the progress of climate models (integrating more processes). Third, in order to assess the progress made in improving climate models, the evolution of the model spread should be considered as only one indicator of improvement, to be assessed alongside with other indicators including model fit, variety of evidence and consistency with background knowledge.

4 Taking ensembles seriously and valuing model independence

So far in this paper, and as is common in state-of-the-art philosophy of climate science (Katzav and Parker 2015 being a remarkable exception), it has been assumed that progress in climate modelling, conceived as improvement in representation, occurs at the scale of individual models (versus the collective scale of the CMIP). On this view, the evolution of climate modelling is understood as tending towards the production of ever more reliable models to understand past and present climates and provide climate projections; in other words, towards a "collection of *even* better guesses". This grounds the interpretation of model spread as a criterion of assessment with respect to the improvement of models taken individually.

However, we now want to consider progress in climate modelling from the perspective of *ensembles*, assuming that progress in climate modelling can also be evaluated in terms of the reliability of the products we can derive from ensembles. These products include the means of projections, but also the probability density functions, and the emergent constraints.

In this regard, and as we are about to argue, convergence of projections is, from a normative point of view, of lower priority than model independence – whereas the latter is often considered as a precondition for the robustness of models' core hypotheses.[11] To that end we now propose two hypotheses and discuss them in turn: (a) convergence of model projections, leading to reduction of model spread, is not a straightforward indicator of the robustness of models'

[10] Another lesson can be learned from the comparison between model spread and other indicators of improvement in climate modelling. When used without taking model spread into consideration, the other indicators are insufficient for two reasons. First, no climate model can be said to perform better that others in respect of every purpose. Second, the fact that a particular model improves its performance is not a sufficient indication that our *general* understanding of the climate system has improved.

[11] As Parker (2011; 2013; 2018) argues, a challenge in climate science is to produce better ensembles by properly sampling the space of models. We therefore believe that, in this respect, the priority is to build a statistical sample of models that are independent of each other, in order to get notably better means.

core hypotheses because it may be due to common biases or other undesired dependence among models; (b) divergence of model projections, enlarging model spread, may indicate genuine independence between models, a desirable feature for the purpose of establishing more reliable statistical means.

(a) As said earlier, convergence of models' outputs is often taken as one indication that the models' hypotheses are robust and reliable. However, convergence of models' outputs may also indicate dependence among models, which may originate in common biases. Accordingly, convergence of models outputs may be (partly) due to dependence, that is, to the fact that models do share hypotheses and pieces of code.

In practice, two different circumstances may result in dependence among models. First, a given modelling centre may simultaneously contribute to several models in the CMIP ensemble, which may then share identical pieces of code (Leduc et al. 2016). Second, it is not uncommon to observe that modellers can be be prone to a type of conformism: scientists are more inclined to produce models whose projections fit the "consensus range" than models whose projections are far outside it. As mentioned by Knutti, "[a]lthough this is hard to confirm or reject, there may even be an element of 'social anchoring' and a tendency towards consensus" (Knutti 2010, 397). Such a social tendency might lead to more dependence between models. We cannot but emphasise that this tendency makes convergence among models rather worrisome.

Recent attempts at dealing with dependence among models, showing how seriously this concern is viewed by the community of modellers, come in two different guises. First, in order to go beyond the "one-model-one-vote" approach, in which each model in an ensemble receives the same weight, methods have been developed to assign different weights to ensemble members (Sanderson et al. 2015a,b). Assessments of whether this approach is on the right track are still ongoing. Second, it has recently been proposed to built subsets of models, within the main ensemble, that meet the criterion of independence with regard to specific results of interest (Abramowitz et al. 2018; Herger et al. 2018).

(b) Accordingly, divergence among models may also be interpreted in two different ways: first, as due to detrimental dissensus among modellers, some of them being right whereas others are wrong; second, as due to normal dissensus among modellers, who place their trust in different hypotheses but do have good, although differing, reasons to do so. The second interpretation importantly implies that it is irrelevant and even detrimental to force the convergence of models' outputs, which is after all not so difficult to reach, because the cost of artificiality is then too high (i.e. it cannot be interpreted as indicating robustness of models' core hypotheses). When divergence among models is interpreted according to the second interpretation, it is not a legitimate object of worry.

The discussions of our two hypotheses (a) and (b) thus result in symmetrical conclusions: in the same way as divergence among models is not always worrisome, dependence might not be a problem. For instance, that fundamental physics is common to all models is no reason to worry. By contrast, models' idealisations, parameterisations and calibrations are often "in-house" productions of modelling centres, and these are responsible for a large part of the

model uncertainty. Therefore, their being passed on from one model to another is epistemically more risky. This implies that the risks involved in dependence have to be assessed on a case-by-case basis.

Our discussion leads to the suggestion that divergence among models outputs may be desirable. As mentioned above, independence among models does not necessarily result from detrimental dissensus among modellers, undermining our overall confidence in (the evolution of) climate models, but does certainly bring about discrepancies among their outputs and thereby increases model spread.

Besides suggesting that divergence among models' outputs, where it is due to independence among models, is not necessarily worrying, we also point out that independence among models might be a higher desideratum than convergence of models' outputs. The main argument is that the ensemble means constitute more accurate projections than model-based projections, and independence among models improves the accuracy of the ensemble means.

Figure 3, from Reichler and Kim 2008, illustrates that the statistical means based on the ensemble models provide more accurate projections than any individual model (see also Schmidt 2018). It shows the evolution of one performance index, named I^2, in the successive updates of CMIP. I^2 "consists of the aggregated errors in simulating the observed climatological mean states of many different climate variables" (Reichler and Kim 2008, 304). Each dot corresponds to a model, grey circles show the average of all models within the ensemble, and black circles indicate the performance of the ensemble mean. A root mean square score is provided, increasing from left to right, meaning that the best models and the best means are those furthest to the left. The black circles, representing the ensemble means, are furthest to the left unlike the colored circles representing the individual models. This means that the statistical means of the multi-model ensemble are more accurate than the projections of each individual model.

Having more accurate means is progress in climate modelling, and yet, independence among models, leading to divergent model outputs, may favour more accurate means. This is best seen by borrowing Gab Abramowitz's pedagogical analogy (see Abramowitz 2017). The analogy is between a hilltop estimation and a mean estimation from an ensemble. Figure 4 provides an illustration. Let us consider that the green disk is a mountain; one wants to know where the top is; in the figure, the top of the mountain is supposed to be at the center of the disk. Let us imagine that people climb the mountain and indicate their respective positions once every minute; their positions are indicated by the red dots. A good estimation of the top of the mountain is the average of the positions only if the people are appropriately scattered around the mountain. Otherwise, as the configuration on the left shows, this average is a bad estimation. Let us now translate the illustration into the CMIP context: the distance to the top of the mountain can be read as the performance of each ensemble member. On the left side, all the red dots are closer to the target but their average is not a good estimation of the top. This nicely illustrates that, in the case of CMIP ensemble, convergence of models' outputs does not point to the best estimation of the mean. By contrast, on the right side, the red dots are scattered but

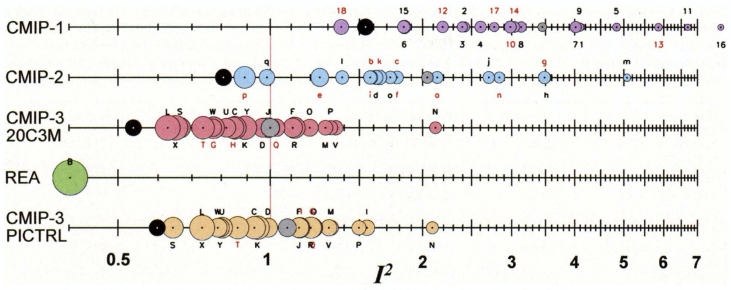

Figure 3: Performance for individual models and averages (Reichler and Kim 2008, Fig.1, 306). "Performance index I^2 for individual models (circles) and model generations (rows). Best-performing models have low I^2 values and are located toward the left. Circle sizes indicate the length of the 95% confidence intervals. Letters and numbers identify individual models; flux-corrected models are labeled in red. Grey circles show the average I^2 of all models within one model group. Black circles indicate the I^2 of the multimodel mean taken over one model group." © American Meteorological Society. Used with permission.

their average provides a good estimation of the top. This illustrates that, in the case of the ensemble, divergence among models outputs, when resulting from independence among models, may be the right path to improved means.

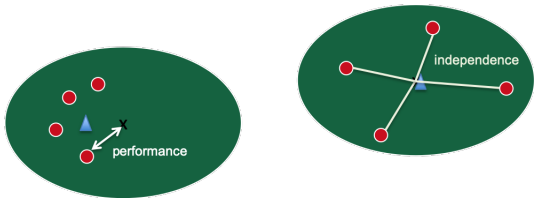

Figure 4: "Hilltop estimation analogy of mean estimate", reproduced from Abramowitz 2017 with the author's approval. Illustration that independence leading to divergent model outputs may favour more accurate means.

Let us go further, however, and claim that this authorises the inclusion, within the ensemble, of models that reproduce generally under-sampled phenomena including extreme phenomena (Räisänen 2007; Herger et al. 2018). As we will now show, this may allow one to adopt a possibilist approach instead of a probabilistic one. Within the scientific and philosophical literature, there are a range of views about the meaning of statistics, dividing into *probabilistic* and *possibilist* interpretations of ensemble projections. The probabilistic approach aims at identifying the most likely scenarios by assessing their probability, while the possibilist approach aims at identifying a space of serious or "real possibilities", e.g. high-impact low-probability phenomena (Katzav 2014; Betz 2015).

Within the probabilistic views, various statistical methodologies are used to

infer probabilities of future climate phenomena from the ensemble projections. The current methods consist in assigning a distribution of probability to each uncertain component (Murphy et al. 2004). Emphasis is put on the "truth" that the ensemble projections are supposed to encircle;[12] "truth" can here be defined as the values of climate magnitudes that have been measured in the past and will be measured in the future. The challenge for these methods is to express the deviation of the projections from the "truth". They include the "truth-plus-error", the "statistically-indistinguishable", the "replicate-Earth", and the Bayesian frameworks, and differ by their idealisations – e.g., whether ensemble members compose a random sample, or whether ensemble members and truth belong to the same statistical distribution (see Parker 2018 for more explanation). Such views have received a certain number of criticisms which emphasise that "probabilistic uncertainty estimates have a false precision and, in that sense, are misleading about the actual state of knowledge" (Parker 2018).

In the possibilist interpretation of model projections, deviation of model projections from the "truth" is not given such a heavy weight; rather, the ensemble of models is seen as a population of possible climates. The virtue of the possibilist approach is that it takes into account high-impact-low-probability phenomena that are often underestimated by the probabilistic views, whereas such phenomena can be important for decision-making (e.g. destruction of the Amazonian rainforest) (see Clarke 2008). Once such possibilities are identified via the models, their consequences for populations and economies can be drawn up and mitigated.

The possibilist interpretation, as we take it, implies that the ensemble can only be improved if a more informative sample is involved, whereas, within the probabilistic interpretation, the ensemble is improved whenever de-idealised models providing more likely projections are included. The former encourages modellers to take extreme phenomena into account even though this results in increased model spread.

In a nutshell, convergence of model projections is of lower priority than model independence as far as the production of means is concerned. What about the other products of ensembles? First and foremost, regarding quantification of uncertainty, it is legitimate to consider that increase of model spread corresponds to better estimation of uncertainty, in that it "merely reflect[s] the quantification of previously unmeasured sources of uncertainty" (IPCC FAQ 1.1 2013).[13] Once we consider that an increase of model spread can correspond to better estimations of uncertainty, then we might expect that they should come with better probability density functions,[14] and more reliable emergent constraints. Thus, increase in model spread should not be seen as detrimental to impact assessment

[12] Hence the model spread is sometimes interpreted as the "distance from the truth".

[13] There are at least three interpretations which can be assigned to model spread as a quantification of uncertainty. As Parker 2013 remarks, it can be seen as 1. "a lower bound on response uncertainty, indicating changes in climate that cannot yet be ruled out" (218), 2. "precise probabilities" (218) assigned to model projections, or 3. "interval probability specifications" (218), i.e. as specifications of the range of imperfect yet plausible projections.

[14] As emphasised notably by Carrier and Lenhard (2019, 4) and supported by the scientific literature, the tails in those functions are currently not well estimated.

studies in which model spreads, used as quantifications of model uncertainty, are often taken as inputs, and in which probability density functions are applied in order to calculate probabilities about the climate projections. Better probability density functions should better serve inductive risk reasoning when one has to choose the model outputs that have the less favourable impacts for the purpose of adaptation strategies (see Parker and Lusk 2019 for a discussion of inductive risk reasoning in climate services).

5 Conclusion

In this paper we addressed the fact that although improvements have been made in process understanding and computer power, model spread over recent decades has remained steady. We gave a tentative explanation of this fact: models are getting more and more complex by including more and more non-stable, sensitive components. These components enlarge the uncertainty range. We then discussed the normative relevance of reducing model spread with respect to the improvement of climate models taken individually. We provided reasons to believe that the model spread can remain steady without threatening the progress of climate models, and highlighted that the evolution of the model spread is not an indicator of improvement per se. Taking seriously that more accurate means of projections may be obtained from ensembles displaying broader model spread, we finally argued that, with respect to the improvement of the ensembles, a reduction of model spread may not be the priority.

The upshot of our discussion is that the persistence of model spread should not be viewed as a reason to doubt the reliability of climate models. Rather, it underlines a tension between two goals: robustness on the one side, and independence on the other, because independence yields divergence among projections. As we have argued, it seems as reasonable to look for robustness as it is to look for independence. In an ideal world, one would be able to distinguish between the set of hypotheses that yield (potentially artificial) robust outcomes and the set of hypotheses that are independent from each other and for this reason yield genuine and relevant divergence within model projections (creating an uncertainty range that does represent the set of alternatives we want to explore). Such a distinction is nevertheless hard to make in practice.

References

Abramowitz, G. (2017). Calibrating ensembles for model independence. *Talk available online, last checked the 18. Sept. 2020,* `https://www.agci.org/lib/17s2/calibrating-ensembles-model-independence`.

Abramowitz, G., Herger, N., Gutmann, E., Hammerling, D., Knutti, R., Leduc, M., Lorenz, R., Pincus, R., and Schmidt, G. A. (2018). Model dependence in multi-model climate ensembles: weighting, sub-selection and out-of-sample testing. *Earth System Dynamics Discussions*, pages 1–20.

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

Betz, G. (2015). Are climate models credible worlds? prospects and limitations of possibilistic climate prediction. *European Journal for Philosophy of Science*, 5(2):5.

Brunner, L., Pendergrass, A. G., Lehner, F., Merrifield, A. L., Lorenz, R., and Knutti, R. (2020). Reduced global warming from cmip6 projections when weighting models by performance and independence. *Earth System Dynamics*, 11(4):995–1012.

Carrier, M. and Lenhard, J. (2019). Climate models: How to assess their reliability. *International Studies in the Philosophy of Science*.

Charney, J. G. et al. (1979). Carbon dioxide and climate: a scientific assessment : report of an ad hoc study group on carbon dioxide and climate, woods hole, massachusetts, july 23-27, 1979 to the climate research board, assembly of mathematical and physical sciences, national research council. *National Academy of Sciences : available from Climate Research Board, http://books.google.com/books?id=cj0rAAAAYAAJ*.

Clarke, L. (2008). Possibilistic thinking: A new conceptual tool for thinking about extreme events. *Social Research*, 75(669-690).

Collins, M., Knutti, R., Arblaster, J., Dufresne, J.-L., et al. (2013). Long-term climate change: Projections, commitments and irreversibility. In Stocker et al. (2013), chapter 12.

Dahan, A. (2010). Putting the earth system in a numerical box ? the evolution from climate modeling toward climate change. *Studies in History and Philosophy of Modern Physics*, 41:282–292.

Fosser, G., Kendon, E. J., Stephenson, D., and Tucker, S. (2020). Convection-permitting models offer promise of more certain extreme rainfall projections. *Geophysical Research Letters*, 47(13):e2020GL088151.

Hausfather, Z. (2019). Cmip6: the next generation of climate models explained. *Carbon Brief, https://www.carbonbrief.org/cmip6-the-next-generation-of-climate-models-explained, last checked the 05/04/2020*.

Herger, N., Abramowitz, G., Knutti, R., Angélil, O., Lehmann, K., and Sanderson, B. M. (2018). Selecting a climate model subset to optimise key ensemble properties. *Earth Syst. Dynam.*, 9:135–151.

IPCC FAQ 1.1 (2013). If understanding of the climate system has increased, why hasn't the range of temperature projections been reduced? In Stocker et al. (2013), chapter 1, pages 140–141.

IPCC FAQ 12.1 (2013). Why are so many models and scenarios used to project climate change? In Stocker et al. (2013), chapter 12, pages 1036–1037.

IPCC FAQ 9.1 (2013). Are climate models getting better, and how would we know? In Stocker et al. (2013), chapter 9, pages 824–825.

Katzav, J. (2014). The epistemology of climate models and some of its implications for climate science and the philosophy of science. *Studies in History and Philosophy of Modern Physics*, 46:228–238.

Katzav, J. and Parker, W. S. (2015). The future of climate modeling. *Climatic Change*, 132:475–487.

Klein, S. A. and Hall, A. (2015). Emergent constraints for cloud feedbacks. *Current Climate Change Reports*, 4(1):276–287.

Knutti, R. (2010). The end of model democracy? *Climatic Change*, 102:395–404.

Knutti, R. and Sedláček, J. (2012). Robustness and uncertainties in the new cmip5 climate model projections. *Nature Climate Change*, 3:369–373.

Le Treut, H. (2009). *Nouveau climat sur la Terre : comprendre, prédire, réagir*. Flammarion.

Leduc, M., Laprise, R., Elía, R., and Šeparović, L. (2016). Is institutional democracy a good proxy for model independence? *American Meteorological Society*, 29.

Lloyd, E. A. (2009). I–varieties of support and confirmation of climate models. *Aristotelian Society Supplementary Volume*, 83:213–232.

Lloyd, E. A. (2010). Confirmation and robustness of climate models. *Philosophy of Science*, 77(5):971–984.

Lloyd, E. A. (2015). Model robustness as a confirmatory virtue: The case of climate science. *Studies in History and Philosophy of Science Part A*, 49:58–68.

Murphy, J. M., Sexton, D. M., Barnett, H., et al. (2004). Quantification of modelling uncertainties in a large ensemble of climate change simulations. *Nature*, 430:768–772.

Parker, W. (2013). Ensemble modeling, uncertainty and robust predictions. *Climate Change*, 4:213–223.

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

Parker, W. S. (2018). Climate science, the stanford encyclopedia of philosophy (summer 2018 edition). *Edward N. Zalta (ed.), https://plato.stanford.edu/archives/sum2018/entries/climate-science/.*

Parker, W. S. and Lusk, G. (2019). Incorporating user values into climate services. *American Meteorological Society*, (September 2019):1643–1650.

Pichelli, E., Coppola, E., Sobolowski, S., Ban, N., Giorgi, F., Stocchi, P., Alias, A., Belušić, D., Berthou, S., Caillaud, C., Cardoso, R. M., Chan, S., Christensen, O. B., Dobler, A., de Vries, H., Goergen, K., Kendon, E. J., Keuler, K., Lenderink, G., Lorenz, T., Mishra, A. N., Panitz, H.-J., Schär, C., Soares, P. M. M., Truhetz, H., and Vergara-Temprado, J. (2021). The first multi-model ensemble of regional climate simulations at kilometer-scale resolution part 2: historical and future simulations of precipitation. *Climate Dynamics*.

Räisänen, J. (2007). How reliable are climate models? *Tellus A*, 59:2–29.

Reichler, T. and Kim, J. (2008). How well do coupled models simulate today's climate? *Bulletin of the American Meteorological Society*, 89:303–312.

Ribes, A., Qasmi, S., and Gillett, N. P. (2021). Making climate projections conditional on historical observations. *Science Advances*, 7(4).

Sanderson, B. M., Knutti, R., and Caldwell, P. (2015a). Addressing interdependence in a multimodel ensemble by interpolation of model properties. *J. Climate*, 28:5150–5170.

Sanderson, B. M., Knutti, R., and Caldwell, P. (2015b). A representative democracy to reduce interdependence in a multimodel ensemble. *J. Climate*, 28(13):5171–5194.

Schmidt, G. A. (2018). Model independence day. *realclimate.org, last checked March 5, 2021*.

Schupbach, J. N. (2018). Robustness analysis as explanatory reasoning. *The British Journal for the Philosophy of Science*, 69(1):275–300.

Stocker, T., Qin, D., Plattner, G.-K., Tignor, M., Allen, S., et al., editors (2013). *IPCC 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 University Press, Cambridge, United Kingdom and New York, NY, USA.

Thompson, E., Frigg, R., and Helgeson, C. (2016). Expert judgment for climate change adaptation. *Philosophy of Science*, 83(5):1110–1121.

Vezér, M. A. (2016). Computer models and the evidence of anthropogenic climate change: An epistemology of variety-of-evidence inferences and robustness analysis. *Studies in History and Philosophy of Science*, 56(95-102).

Vezér, M. A. (2017). Variety-of-evidence reasoning about the distant past: A case study in paleoclimate reconstruction. *European Journal in Philosophy of Science*, 7(257–265).

Webb, M. J., Andrews, T., Bodas-Salcedo, A., Bony, S., Bretherton, C. S., Chadwick, R., Chepfer, H., Douville, H., Good, P., Kay, J. E., Klein, S. A., Marchand, R., Medeiros, B., Siebesma, A. P., Skinner, C. B., Stevens, B., Tselioudis, G., Tsushima, Y., and Watanabe, M. (2017). The cloud feedback model intercomparison project (cfmip) contribution to cmip6. *Geoscientific Model Development*, 10(1):359–384.

Winsberg, E. (2018). What does robustness teach us in climate science: a re-appraisal. *Synthese*, https://doi.org/10.1007/s11229-018-01997-7.