

# Taking climate model evaluation to the next level

Veronika Eyring<sup>1,2,\*</sup>, Peter M. Cox<sup>3</sup>, Gregory M. Flato<sup>4</sup>, Peter J. Gleckler<sup>5</sup>, Gab Abramowitz<sup>6</sup>, Peter Caldwell<sup>5</sup>, William D. Collins<sup>7,8</sup>, Bettina K. Gier<sup>2,1</sup>, Alex D. Hall<sup>9</sup>, Forrest M. Hoffman<sup>10,11</sup>, George C. Hurtt<sup>12</sup>, Alexandra Jahn<sup>13</sup>, Chris D. Jones<sup>14</sup>, Stephen A. Klein<sup>5</sup>, John Krasting<sup>15</sup>, Lester Kwiatkowski<sup>16</sup>, Ruth Lorenz<sup>17</sup>, Eric Maloney<sup>18</sup>, Gerald A. Meehl<sup>19</sup>, Angeline Pendergrass<sup>19</sup>, Robert Pincus<sup>18</sup>, Alex C. Ruane<sup>20</sup>, Joellen L. Russell<sup>21</sup>, Benjamin M. Sanderson<sup>19</sup>, Benjamin D. Santer<sup>5</sup>, Steven C. Sherwood<sup>6</sup>, Isla R. Simpson<sup>19</sup>, Ronald J. Stouffer<sup>21</sup>, and Mark S. Williamson<sup>3</sup>

<sup>1</sup>Deutsches Zentrum für Luft- und Raumfahrt e.V. (DLR), Institut für Physik der Atmosphäre, Oberpfaffenhofen, Germany.

<sup>2</sup>University of Bremen, Institute of Environmental Physics (IUP), Bremen, Germany.

<sup>3</sup>College of Engineering, Mathematics and Physical Sciences, University of Exeter, Exeter, EX4 4QE, UK.

<sup>4</sup>Canadian Centre for Climate Modelling and Analysis, Environment and Climate Change Canada, Victoria, BC, Canada.

<sup>5</sup>Program for Climate Model Diagnosis and Intercomparison (PCMDI), Lawrence Livermore National Laboratory, Livermore, California, USA.

<sup>6</sup>Climate Change Research Centre and ARC Centre of Excellence for Climate System Science, University of New South Wales, Sydney, Australia

<sup>7</sup>Climate and Ecosystem Sciences Division, Lawrence Berkeley National Laboratory, Berkeley, California, USA.

<sup>8</sup>Department of Earth and Planetary Science, University of California, Berkeley, California, USA.

<sup>9</sup>University of California, Los Angeles, USA

<sup>10</sup> Computational Earth Sciences Group and Climate Change Science Institute, Oak Ridge National Laboratory, Oak Ridge, Tennessee, USA

<sup>11</sup> Department of Civil and Environmental Engineering, University of Tennessee, Knoxville, Tennessee, USA

<sup>12</sup>University of Maryland, College Park, MD, USA

<sup>13</sup>Department of Atmospheric and Oceanic Sciences and Institute of Arctic and Alpine Research, University of Colorado, Boulder, CO, USA

<sup>14</sup>Met Office Hadley Centre, Exeter, UK

<sup>15</sup>Geophysical Fluid Dynamics Laboratory/NOAA, Princeton, NJ, USA

<sup>16</sup>Laboratoire de Météorologie Dynamique (LMD), IPSL, Paris, France

<sup>17</sup>Institute for Atmospheric and Climate Science, ETH Zurich, Zurich, Switzerland

<sup>18</sup>Cooperative Institute for Research in Environmental Sciences, University of Colorado, Boulder, CO USA

<sup>19</sup>National Center for Atmospheric Research (NCAR), Boulder, Colorado, USA

<sup>20</sup>NASA Goddard Institute for Space Studies, New York, NY, USA

<sup>21</sup>University of Arizona, Tucson, Arizona, USA

\*e-mail: veronika.eyring@dlr.de

**Earth system models are complex and represent a large number of processes, resulting in a persistent spread across climate projections for a given future scenario. Owing to different model performances against observations and the lack of independence among models, there is now evidence that giving equal weight to each available model projection is suboptimal. This Perspective discusses newly developed tools that facilitate a more rapid and comprehensive evaluation of model simulations with observations, process-based emergent constraints that are a promising way to focus evaluation on the observations most relevant to climate projections, and advanced methods for model weighting. These approaches are needed to distil the most credible information on regional climate changes, impacts, and risks for stakeholders and policy-makers.**

The Intergovernmental Panel on Climate Change (IPCC) Fifth Assessment Report (AR5) concluded that the warming of the climate system is unequivocal and human influence on the climate system is clear<sup>1</sup>. Observed increases of greenhouse gases have contributed significantly to warming of the atmosphere and ocean, sea ice decline, and sea level rise. The size and rapidity of these changes is concerning. Human-caused climate change is already affecting many aspects of societies and ecosystems. These impacts will become more visible and more serious in the 21<sup>st</sup> century. It should, therefore, be an international priority to improve our understanding of the climate system, and to reduce current uncertainties in projections of future change. This will rely on information from theory, observations, and Earth system model (ESM) simulations that are coordinated as part of the World Climate Research Programme (WCRP) Coupled

Model Intercomparison Project (CMIP<sup>2-5</sup>). CMIP is now in its sixth phase (CMIP6<sup>5</sup>) and is confronted with a number of new challenges. Compared to CMIP5, an increased number of institutions will participate in CMIP6, many with multiple model versions. The latest generation of climate models feature increases in spatial resolution, improvements in physical parameterizations (e.g., in the representation of clouds), and inclusion of additional Earth system processes (e.g., nutrient-limitations on the terrestrial carbon cycle). These additional processes are needed to represent key feedbacks that affect climate change, but are also likely to increase the spread of climate projections across the multi-model ensemble. This escalates the need for innovative and comprehensive model evaluation approaches.

CMIP provides the basis for multi-model evaluation and has, over the years, revealed a variety of systematic differences between models and observations, with many persisting from one model generation to the next<sup>6,7</sup>. An important issue that remains to be fully addressed is the extent to which model errors affect the quality of climate projections and subsequent impact assessments<sup>8</sup>. Traditionally, many climate projections are shown as multi-model averages in the peer-reviewed literature and IPCC reports, with the spread across models presented as a measure of projection uncertainty<sup>9</sup>. There is now emerging evidence that weighting based on model performance may improve projections for specific applications<sup>10-12</sup>. An additional complication in devising model weighting approaches is that many CMIP models share some components, or are variants of another model in the ensemble, and hence are not truly independent<sup>12-16</sup>. This has the potential to bias the multi-model results in ways that are only beginning to be explored. The lack of independence challenges the notion of a ‘model democracy’, in which each model is weighted equally<sup>17</sup>.

The growing number and complexity of models, the expanding suite of outputs they produce, the multitude of downstream applications, and the growing availability of observational datasets drive a need for more routine and systematic evaluation, utilizing a comprehensive set of existing model performance metrics and diagnostics. Newly developed CMIP evaluation tools<sup>18,19</sup> will ultimately enhance our ability to identify model errors, to investigate their causes and to quantify and potentially reduce projection uncertainties.

In this Perspective, we summarize key advances since AR5 and key scientific opportunities for improving climate model analysis that will be assessed in the AR6. Our focus is on gaps in the understanding of systematic errors, the development of CMIP model evaluation tools, emergent constraints, and weighting methods. We also address the need for more user- and policy-oriented model evaluation at the regional

scale required for impact studies. Finally, we discuss how the scientific community might provide more robust climate model information and more tightly constrained model projections.

## **From model errors to understanding processes**

Comparing model results to observations provides insight into the quality of model simulations and the way in which various processes are represented. Comparisons with observations can reveal shortcomings in individual models and systematic errors in a large multi-model ensemble<sup>7,20</sup>. An example of a systematic error is the excessive and often incorrectly simulated band of precipitation in the tropical Pacific south of the equator. Taken together with the usually correctly simulated climatological intertropical convergence zone (ITCZ) precipitation maximum that stretches across the tropical Pacific north of the equator, this systematic error is commonly referred to as the “double ITCZ”. Other examples include a dry Amazon bias, a warm bias in the eastern parts of tropical ocean basins, differences in the magnitude and frequency of El Niño and La Niña events, biases in sea surface temperatures (SST) in the Southern Ocean, a warm and dry bias of land surfaces during summer, and differences in the position of the Southern Hemisphere atmospheric jet<sup>7</sup>.

One major challenge is that it is often not possible to attribute a specific cause to a specific systematic error. For example, it has been suggested that the systematic warm bias in the upwelling zones off the west coasts of each continent (see Figure 1) is associated with biases in the representation of stratocumulus clouds<sup>21</sup> and boundary layer convection<sup>22</sup> in these regions. However, other studies suggest that the root cause of this warm bias is the representation of ocean upwelling and its forcing from surface winds<sup>23</sup>. An additional complication is that a regional difference between the simulation and observations may be a consequence of errors that occur far from the region in question, and are manifested via teleconnections. Certain regional SST biases, for example, are related to biases in other ocean basins and to aspects of the large-scale ocean overturning circulation<sup>24</sup>. In some cases, although the link between a particular bias and some physical process may appear robust, the specific cause of the bias, as well as its remedy, may remain elusive.

But there are also compelling examples of how a multi-model analysis of a particular systematic error can lead to a clearer understanding of underlying causes. One systematic error revealed in the evaluation of CMIP5 models was the apparent difference between observed and modelled global mean surface temperature increase in the early 21<sup>st</sup> century<sup>7</sup>. These differences motivated a range of targeted analyses exploring model performance, internal variability, external forcing, and observational uncertainty<sup>25</sup>.

Although the magnitude of the slowdown differs slightly depending on which global observational dataset is analysed, this focused effort revealed that the observed slowdown was due to a combination of factors, chiefly involving internally-generated decadal timescale variability in the tropical Pacific<sup>26,27</sup> and the missing effects of a series of moderate volcanic eruptions<sup>28</sup>. Averaging the time series across a collection of coupled model simulations strongly reduces the effects of internally generated variability, more clearly revealing the underlying externally forced response. There is, therefore, a mismatch between the precise observed sequence of variability and the smooth evolution of temperature in the multi-model mean. Models initialized with observations from the years immediately prior to the early 21<sup>st</sup> century slowdown were able to capture aspects of the observed change in warming rate after 2000<sup>29</sup>. These results highlight the importance of using different simulation frameworks (e.g., coupled simulations and decadal predictions initialized with observations) to understand the causes of differences between modelled and observed climate changes. Stronger observed warming since 2014, which is replicated in initialized model predictions for the period after 2014<sup>30,31</sup>, adds to the evidence that the weaker warming before 2014 had a large contribution from internal climate variability.

A related question is the extent to which observational uncertainties and inhomogeneities<sup>32</sup> are hampering model evaluation. Just as efforts continue on improving models, there is a parallel ongoing effort to improve observationally-based datasets. Even for a very basic climate quantity like temperature, this involves refined corrections for biases and incomplete global coverage in the raw surface observations<sup>33</sup> and corrections for biases in satellite retrievals<sup>34</sup>. Observations are also critically-important in model tuning<sup>35,36</sup> which should be clearly documented and taken into account in model evaluation studies. A difficulty in comparing models against observations is inconsistency in sampling or definition of the quantities compared (e.g. model data may be daily averages while satellite samples may be for a certain time of day). This inconsistency can be addressed by incorporating simulators of specific instruments into climate models<sup>37</sup>.

## **New CMIP model evaluation tools**

The scope of model evaluation has expanded dramatically in recent years. Well-established aspects of model evaluation are now becoming more routine, results are available more rapidly than for CMIP5, enhancing their value for model analysts and developers<sup>38</sup>. A key development for CMIP6 is the availability of the *Earth System Model Evaluation Tool* (ESMValTool<sup>18</sup>) and a *Coordinated set of Model Evaluation Capabilities* (CMEC) which are both open-source capabilities. The ESMValTool includes a large collection of diagnostics and performance metrics for atmospheric, oceanic, and terrestrial variables,

not only for the mean state, but also for trends, variability, key physical processes, and emergent constraints. Additionally, ESMValTool has the capability to reproduce figures from several AR5 chapters (<http://cmip-esmvaltool.dkrz.de/>) and incorporates targeted analysis packages, such as the NCAR Climate Variability Diagnostics Package<sup>39</sup>. CMEC comprises the PCMDI Metrics Package (PMP<sup>19</sup>), the International Land Modeling Benchmarking Project package (ILAMB<sup>40</sup>), and the parallel toolkit for extreme climate analysis (TECA<sup>41</sup>). CMEC emphasizes a diverse suite of physical and biogeochemical summary statistics gauging the consistency between models and observations across a range of space and timescales.

Both ESMValTool and CMEC have undergone rapid development over the last several years, and are now mature, well-tested tools, which provide end-to-end provenance tracking to ensure reproducibility. One goal is to routinely provide evaluation results through the Earth System Grid Federation (ESGF) shortly after new CMIP6 simulations are published. This workflow is depicted in Figure 2: the tools are run at selected ESGF nodes, utilizing observations available in standard formats or provided by the user<sup>47</sup>. The foundations for this significant undertaking are the community-based experimental protocols and conventions of CMIP, including their extension to observations (obs4MIPs<sup>51</sup>) and reanalysis.

## **Emergent constraints on Earth system sensitivities**

One of the biggest challenges in ESM evaluation is to identify the performance metrics that are most relevant to climate projections<sup>7</sup>. The reliability of models can only be assessed with observations of the past and present. This means that models are assessed against criteria that are not necessarily informative in terms of the quality of model projections of future climate change. The *emergent constraint* approach attempts to address this problem by identifying robust, physically interpretable relationships between Earth system feedback behaviour on short, well-observed timescales and on timescales spanning the 21<sup>st</sup> century and beyond<sup>42,43</sup> (see also Box 1). Emergent constraints use an ensemble of ESMs to define a relationship between a measured aspect of current or past climate and the strength of a simulated Earth system feedback in the future. It is the model ensemble behaviour (rather than the behaviour of a single model) that defines the *emergent relationship* between the observed variability and the projection of the future climate. When combined with observational data and a measure of observational uncertainty, the model-derived *emergent relationship* can be converted into an *emergent constraint* on the Earth system sensitivity in the real world<sup>50</sup>.

When AR5 was published, numerous emergent constraints had been identified. Examples include studies of snow-albedo feedback<sup>42</sup>, sea-ice<sup>44</sup>, tropical precipitation extremes<sup>45</sup>, carbon loss from tropical land under warming<sup>46</sup> and the future latitudinal shift of the Southern Hemisphere westerlies<sup>47</sup>. Such studies have proliferated since the AR5, including constraints on cloud feedbacks and equilibrium climate sensitivity (ECS)<sup>48-56</sup>, strengthening of the hydrological cycle<sup>57,58</sup>, the temperature sensitivity of tropical land carbon storage<sup>59</sup>, CO<sub>2</sub> fertilization of plant photosynthesis<sup>60</sup>, future changes in ocean net primary production<sup>61</sup>, permafrost loss<sup>62</sup>, changes in natural sources and sinks of CO<sub>2</sub><sup>63</sup> and mid-latitude daily heat extremes<sup>64</sup>. The proposed observable constraints involve historical trends<sup>42,44,63</sup>, interannual variability<sup>46,54,59</sup>, seasonal cycles<sup>42</sup>, trends in the seasonal cycle<sup>60</sup>, and spatial variability<sup>62</sup>. Constraints have been tested against different ensembles and scenarios<sup>42,46,59,65</sup>. For example, a relationship between the ECS and the inferred strength of upward mixing in the tropical lower troposphere was used to discount ECS values below 3°C, since all models with lower ECS had too little mixing (Figure 3, left) and by implication too little positive cloud feedback at low levels<sup>52</sup>. This would narrow the range of ECS significantly compared to the 1.5 to 4.5°C range assessed by the IPCC AR5<sup>1</sup>. However, other emergent constraint studies for ECS lead to different estimates<sup>50-56</sup>, pointing to the need for further research. For carbon cycle feedbacks, an emergent constraint on CO<sub>2</sub>-fertilisation of photosynthesis was found based on observed changes in the seasonal cycle of atmospheric CO<sub>2</sub>, suggesting that doubling of the CO<sub>2</sub> concentration in the atmosphere will cause global plant photosynthesis to increase by approximately one third (Figure 3, right<sup>60</sup>).

Despite the attractiveness of emergent constraints, there are some well-justified concerns. Most importantly, the emergent relationship between the observable and the sensitivity to be constrained is derived from a model ensemble. The emergent constraint may be misleading if the model ensemble has a systematic error (such as the double ITCZ) which affects the emergent relationship, or which reflects the simplicity of a parameterization common to many models rather than an intrinsic underlying process. Secondly, there is a danger of finding spurious relationships between observables and Earth System sensitivities if the high-dimensional outputs available from ESMs are simply data-mined for high correlations<sup>67</sup>. The correlations found in a data mining approach should be restricted to those that have a physical explanation. Finally, we should not expect short-term variability to yield constraints on slow feedbacks that have negligible effects on that variability. For example, inter-annual variations in sea-level are unlikely to provide constraints on century timescale sea-level rise due to ice-sheet melt. On the other hand, fast processes (such as water vapour and cloud feedbacks) are evident in short-term variability as well as trends, and are therefore much better candidates for emergent constraints that relate variability and

sensitivity. Many of the most uncertain feedbacks are such fast feedbacks which are more amenable to the emergent constraint technique.

## **Weighting multi-model climate projections**

Traditionally, CMIP models were treated as independent, equally plausible estimates of future climate. Confidence in projections was inferred from model agreement on the sign and magnitude of future change<sup>9</sup>. In the context of multi-model ensemble projections, an increasing number of studies have weighted models that agree better with historical observations of that quantity or relationships between the projected quantity and observable metrics<sup>11,66-69</sup>. However, the majority of weighting studies of certain climate properties such as sea ice extent in AR5 include only a small set of metrics that are not always clearly related to the projected quantity in question.

An increase in weighted skill scores can be relatively simply achieved *in sample* (that is, in the observational period and/or location used to derive weights). However, only a few studies have specifically focused on the likelihood of weighted results providing benefits for the intended application (that is, *out-of-sample*, typically 21<sup>st</sup> century projections)<sup>10-12,14,15,67,70,71</sup>. Although we clearly have no observations of future climate, *model-as-truth* (also termed *pseudo-reality* in some studies<sup>11,67</sup>) and calibration-validation exercises for different time periods of the observations yield valuable information on the potential benefits of different weighting approaches<sup>72</sup>. In addition to testing whether projections of a specific variable and metric can be improved through weighting, thorough out-of-sample testing can help guard against other potential issues with weighting. For example, there is no single metric that reliably captures all aspects of model performance for all purposes, even in the case where interest is restricted to a very specific scientific question<sup>73</sup>. Out-of-sample testing can tell us whether optimising in one metric or variable transfers any benefits to other metrics or variables<sup>70</sup>. It can also indicate whether internal variability has played a role in any in-sample success of weighting, help avoid the issue of the same datasets being used to calibrate and weight models, and reveal whether weighting has artificially reduced ensemble spread. A further problem is the risk of systematic errors in observational products producing inappropriately weighted ensembles. Furthermore, weighting schemes have no capacity to account for errors that are shared across an ensemble, an issue that is particularly important in the case of small ensembles.

Another relevant issue is model interdependence. Some of the nominally different models in the CMIP archive share individual components or parameterizations, or represent key processes in the same way. This can lead to shared errors which have the potential to compromise the efficacy of performance-based weighting<sup>71</sup> and to create artificially strong emergent constraints<sup>74</sup>. Using model error correlation as a measure for interdependence, it was found that the effective number of independent climate models was likely to be significantly fewer than the total number of models in the CMIP ensemble<sup>75,76</sup>. Several studies have subsequently introduced alternative ways of quantifying and accounting for interdependence<sup>13,14,77,78</sup>. Recently, the US National Climate Assessment weighted each member of the CMIP5 archive using both a multivariate skill score for historical climatology and a measure of uniqueness in the archive<sup>12</sup>. Figure 4 shows the resulting skill weight versus the independence weight for all CMIP5 models. Skill weights are calculated as multi-variate root mean square errors over a North American domain, while independence weights are computed using model error bias correlation. None receives high weights for both skill and independence (see the empty upper right corner). This suggests that the ensemble has been unintentionally skill weighted by the inclusion of multiple versions of better-performing models. As in the case of efforts to define broadly applicable performance metrics, it is evident that there is no universally accepted definition of model dependence: accounting for model dependence is problem-specific. Weighting exercises that accounted for model dependence were sensitive to almost all aspects of the problem, including the selected metric, variable, analysis period, and constraining observational dataset.

## **Regional model evaluation for impact and risk assessments**

A primary goal of climate research is to identify how climate variability and change affect society and to inform strategies for mitigation and adaptation to climate change. Impact sectors include agriculture, forestry, water resources, infrastructure, energy production, land and marine ecosystems, and human health. To accurately capture many of the significant effects of climate change in sectoral impacts models, high levels of detail regarding the evolving climate state are necessary. The impacts community generally needs rigorous regional-scale evaluation of the seasonal cycles in temperature, precipitation, humidity, wind speed, and downwelling solar radiation. Although some sectors are affected by mean climate changes, the most acute impacts are related to extreme events.

Annual-mean temperature and precipitation, monsoon timing and intensity, and modes of variability that can alter the probability of extreme events have been evaluated in the CMIP5 ensemble<sup>7</sup>. The Expert Team on Climate Change Detection and Indices compiled a set of indices to quantify extreme events<sup>79</sup>. Observational estimates of these indices have been used to evaluate CMIP5 models<sup>80</sup>, and are now

incorporated into the ESMValTool for evaluation of CMIP6 models. The overall model performance was mixed in capturing the observed behaviour of these extreme event metrics. However, such comparisons remain difficult to interpret because of substantial uncertainties and data gaps in many of the observational datasets, and due to limited availability of CMIP5 model output at sub-monthly and daily frequencies.

Stakeholder-oriented applications have benefited since AR5 from improved models, more user-relevant metrics, more robust observational systems, and longer observational records. Improvements in remote sensing products have enabled evaluation of interannual and sub-seasonal events<sup>81</sup>. Models show continuing improvement in the representation of the diurnal cycle, storm tracks, the effects of blocking on extreme events, ENSO, tropical cyclones and other circulation features. Awareness of multivariate extremes<sup>82</sup> has begun to emerge. The production and application of dynamically and empirically downscaled model information has advanced<sup>83</sup>. Extreme event detection and attribution has made substantial strides, with real-time probabilistic event attribution<sup>84</sup> now feasible. Impacts sector applications have also been further advanced by hybrid climate forcing datasets combining models and observations<sup>85</sup> and by sector-specific information such as the geographic distribution of nitrogen fertilizer or irrigation applications and land use patterns for agricultural modelling<sup>86</sup>.

Interactions between the climate modelling and applications communities are facilitated by the Vulnerability, Impacts, Adaptation, and Climate Services (VIACS) Advisory Board<sup>87</sup>, a CMIP6-Endorsed MIP<sup>5</sup>. This Board is an effort to draw a broader array of climate experts, practitioners, and information brokers into the CMIP process, and to leverage the community engagements organized under the Global Programme of Research on Climate Change Vulnerability, Impacts and Adaptation (PROVIA) and as part of efforts such as the Inter-Sectoral Impact Model Intercomparison Project (ISIMIP<sup>88</sup>). The VIACS Advisory Board has solicited sustained community engagement: 1) on priority experiments; 2) on output variables to inform CMIP6 data requests; and 3) to highlight evaluations that would help to establish model credibility with the VIACS community. Given that it is currently common practice to adjust model biases with monthly-mean information from present-day fields, evaluation of the seasonal evolution and distributions of monthly climate at regional scales is among the highest priorities for VIACS users<sup>89</sup>. ESMs often have less variance than observations at hourly, daily, and interannual timescales, which can lead to spurious effects when bias adjustment relies on standardized anomalies. Tropical rainfall biases are particularly problematic.

## Ways ahead

The CMIP6 experiment design provides an opportunity for sophisticated, consistent characterization of the ensemble and its predecessors<sup>5</sup>. Targeted model intercomparison projects (MIPs) associated with CMIP6 will accelerate efforts to disentangle internal climate variability from forced responses, and to evaluate model processes relevant to a wide range of climate characteristics. Insights into the underlying causes of systematic errors are likely to be gained from idealized experiments (such as aquaplanets)<sup>90,91</sup>, systematic assessment of the influence of horizontal resolution<sup>92</sup>, and the evaluation of individual model components<sup>93</sup>. The diverse numerical experiments proposed in CMIP6 may help the climate science community to gain a deeper understanding of model behaviour and processes than has been possible in the past. Further diagnostic benefits should accrue from the development of convectively resolving models, dynamic vegetation, three-dimensional ice sheet models, and refined physical parametrisations.

Model development, evaluation and weighting will be facilitated by the ongoing development and deployment of new climate observing systems with continuous quality assessment and independent verification. Rigorous quantification of observational uncertainties is now routine rather than exceptional. Examples include the availability of ‘ensembles of observations’ for a single observational product, which account for uncertainties associated with different subjective processing choices<sup>94</sup>. Challenges remain in propagating these uncertainties to derived quantities such as trends or conditional averages. New measurements and measurements made at higher frequency will also provide further insights into systematic errors. ARGO floats<sup>95</sup> and new satellite missions are prime examples.

An exciting opportunity is provided by the new CMIP evaluation tools ESMValTool<sup>18</sup> and CMEC<sup>19</sup>. Both evaluation packages will be routinely executed whenever new model simulations are contributed to the CMIP6 archive. This allows rapid, quantitative comparisons of model results to a wide range of climate observations<sup>38</sup>. Such rapid and comprehensive feedback on model performance should help addressing the causes of long-standing systematic errors and facilitate a shift towards more process-oriented diagnostics, while ensuring continuity with more ‘traditional’ diagnostics applied in previous CMIP phases. The hope is that ESMValTool and CMEC will be further enhanced by the CMIP6-Endorsed MIPs and other science teams, leading to widespread adoption by model development teams and the user community. Other promising diagnostic developments on the horizon that should be further advanced include studies that assess responses to perturbations rather than mean climate<sup>96</sup>, and the application of innovative data science methods in Earth system science such as neural networks<sup>97</sup>, machine learning-based anomaly detection techniques<sup>98</sup>, graphical models and causal discovery<sup>99</sup>.

Physically-robust emergent constraints are a promising concept for understanding and constraining Earth system feedbacks and narrowing uncertainty in future projections. They may ultimately influence model development and observational strategies. In addition to new research on the consistency across different emergent constraints and across generations of model ensembles, we anticipate use of more sophisticated statistical analyses, which to date have typically involved one-dimensional linear regressions. Higher-dimensional emergent relationships related to more than one observable should yield more robust conclusions and avoid the possibility of contradictory constraints derived from separate one-dimensional relationships. There is a new opportunity to test emergent constraints developed in previous model generations against the outputs from CMIP6 models. Finally, there needs to be a greater focus on developing emergent constraints for regional climate change more relevant to impacts than many of the large-scale metrics that are the current focus of emergent constraints<sup>8</sup>.

To guard against misleading emergent constraints arising from spurious correlations or from the dependence introduced by a parameterization common to many models rather than from an intrinsic underlying process, we suggest that the development of emergent constraints should be treated as a form of hypothesis testing. For example, emergent relationships between variability and sensitivity could be derived based on physical theory or simple underlying models<sup>57</sup>. The predicted emergent relationship can then be tested against outputs from full-form ESMs. This approach could also yield improved theoretical understanding of relationships between variability and sensitivity in the Earth system. Even where the outputs of one generation of models appear consistent with the hypothesised emergent relationship, the robustness of the relationship should be tested ‘out of sample’ against models that were not used to define the relationship. The hypothesis testing approach that we propose would also protect somewhat from attempts to artificially tune a model to fit an observational constraint. Where this is carried out unphysically, the tuned model is likely to move away from the theoretical curve (i.e., it will fit the x-axis observation but it will no longer be consistent with the y-axis sensitivity).

There is enough evidence now that the continued assumption of model democracy cannot be fully justified in future IPCC assessment reports. It is not yet clear, however, whether all variables of interest can be reliably constrained. Successful skill weighting has thus far been implemented for a limited number of specific applications. In these applications, the target property of interest is constrained by a small number of clearly relevant variables<sup>10-12,66,67</sup>. Future work for more complex chains of influence will need to consider orthogonal uncertainties and processes. For example, regional precipitation change may be influenced by global-scale warming, large-scale dynamics and microphysical parameterizations. For

regional climate projections, a weighting that is based on processes controlling the region of interest and biases in large-scale atmospheric circulation is advocated<sup>8</sup>.

In addition, it has been demonstrated that CMIP models are not independent. Most inferences in the literature about model interdependence are derived from error correlation<sup>13,78</sup>. This cannot identify the specific model components that are interdependent. Identification of these common components is a difficult task due to the high number of models involved in CMIP and lack of detailed information regarding individual model versions. Further work is required to understand how interdependence can be best assessed. These efforts can proceed in tandem with research to better understand the effects of model construction and genealogy. Comprehensive databases of shared code, parameterizations, model development and tuning practises could help disentangle how models are related, and for what purposes they can be considered independent estimates of change. There is also the potential for better quantification of natural variability from paleoclimate simulations<sup>100</sup> and enhanced collaboration with the detection and attribution community, whose statistical approaches provide information on whether model responses to changes in external forcings are consistent with observations. Simpler representations of the Earth system in a hierarchy of models will also be useful to improve more complex ESMs.

For improved assessments of regional impacts and risks, a key challenge and opportunity will be to derive collective understanding from global and regional climate models, as well as from regional-scale observations. To do so it will be important to bridge the gap between the climate model and impacts communities, and between the different scales on which these communities typically operate. CMIP6 will include weather-resolving global model resolutions (~25 km or finer) that need to be compared to regional model results and downscaled coarse-resolution simulations. Concerns about bias correction of climate change simulations have been raised, and ways to address these concerns have been proposed<sup>8</sup>. Many of the key systematic errors that hampered reliable simulation of surface variables and extreme events will benefit from increasing spatial resolution<sup>92</sup>, variable-resolution grids, improved parametrizations, and advances in bias corrections and downscaling techniques<sup>83</sup>. Curated archives and the CMIP evaluation tools will enable participation by a broader diagnostic community, many of whom are not presently capable of advanced interrogation of climate model simulations. The provision of useful climate information and messages for the impacts, risk, and climate services communities requires a process rooted in sustained engagement with stakeholders that concentrates on areas of particular vulnerability or exposure. Projection of changing hazard metrics and the construction of driving scenarios for impact models across a range of local, regional, and national scales should benefit from a process distilling information from many different sources. Such sources include multiple ESMs, statistically and

dynamically downscaled models (e.g., through the Coordinated Regional Downscaling Experiment, CORDEX; now a CMIP6-Endorsed MIP<sup>83</sup>), and bias adjusted models<sup>8</sup>. The IPCC Sixth Assessment Report (AR6) will enhance the focus on regional climate information through a regionally-defined Atlas and new chapters on global-to-regional linkages, extreme events, and impact- and risk-relevant climate hazards.

Despite significant progress in climate modelling over the last decades, there remains a substantial spread in projections of future climate change. For example, the range of model estimates for equilibrium climate sensitivity to doubling of CO<sub>2</sub> concentration has not decreased since the 1970s<sup>7</sup>, although understanding of the processes that are involved certainly has increased. The need to inform mitigation policy and adaptation remains. We believe that there is now an unprecedented opportunity to constrain policy-relevant metrics (such as the cumulative CO<sub>2</sub> emissions consistent with specific temperature targets) with observations, and to reduce uncertainties in climate projections, both at global and regional scales. The challenge is to make intelligent use of the Petabyte-scale output that will become available from the new CMIP6 project together with modern observation systems, new model evaluation tools, and novel data science techniques. A combination of different process-based emergent constraints together with model-weighting approaches that consider both model performance and interdependence have the potential to yield more robust multi-model information for a wide array of societally and environmentally critical applications.

## References

- 1 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, 2013).
- 2 Meehl, G. A., Boer, G. J., Covey, C., Latif, M. & Stouffer, R. J. The Coupled Model Intercomparison Project (CMIP). *B Am Meteorol Soc* **81**, 313-318, doi:10.1175/1520-0477(2000)081<0313:TCMIPC>2.3.CO;2 (2000).
- 3 Meehl, G. A. *et al.* THE WCRP CMIP3 Multimodel Dataset: A New Era in Climate Change Research. *B Am Meteorol Soc* **88**, 1383-1394, doi:doi:10.1175/BAMS-88-9-1383 (2007).
- 4 Taylor, K. E., Stouffer, R. J. & Meehl, G. A. An Overview of CMIP5 and the Experiment Design. *B Am Meteorol Soc* **93**, 485-498, doi:10.1175/Bams-D-11-00094.1 (2012).
- 5 Eyring, V. *et al.* Overview of the Coupled Model Intercomparison Project Phase 6 (CMIP6) experimental design and organization. *Geosci. Model Dev.* **9**, 1937-1958, doi:10.5194/gmd-9-1937-2016 (2016).
- 6 Stouffer, R. J. *et al.* CMIP5 Scientific Gaps and Recommendations for CMIP6. *B Am Meteorol Soc* **98**, 95-105, doi:10.1175/bams-d-15-00013.1 (2017).
- 7 Flato, G. *et al.* in *Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change* (ed T.F.

- Stocker, D. Qin, G.-K. Plattner, M. Tignor, S.K. Allen, J. Boschung, A. Nauels, Y. Xia, V. Bex and P.M. Midgley) Ch. 9, 741–866 (Cambridge University Press, 2013).
- 8 Maraun, D. *et al.* Towards process-informed bias correction of climate change simulations. *Nat Clim Change* **7**, 764–773, doi:10.1038/nclimate3418 (2017).
- 9 Collins, M. *et al.* in *Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change* (ed T.F. Stocker, D. Qin, G.-K. Plattner, M. Tignor, S.K. Allen, J. Boschung, A. Nauels, Y. Xia, V. Bex and P.M. Midgley) Ch. 12, 1029–1136 (Cambridge University Press, 2013).
- 10 Knutti, R. *et al.* A climate model projection weighting scheme accounting for performance and interdependence. *Geophys Res Lett* **44**, 1909–1918, doi:10.1002/2016gl072012 (2017).
- 11 Wenzel, S., Eyring, V., Gerber, E. P. & Karpechko, A. Y. Constraining Future Summer Austral Jet Stream Positions in the CMIP5 Ensemble by Process-Oriented Multiple Diagnostic Regression. *J Climate* **29**, 673–687, doi:10.1175/Jcli-D-15-0412.1 (2016).
- 12 Sanderson, B. M., Wehner, M. & Knutti, R. Skill and independence weighting for multi-model assessments. *Geosci. Model Dev.* **10**, 2379–2395, doi:10.5194/gmd-10-2379-2017 (2017).
- 13 Sanderson, B. M., Knutti, R. & Caldwell, P. Addressing Interdependency in a Multimodel Ensemble by Interpolation of Model Properties. *J Climate* **28**, 5150–5170, doi:10.1175/Jcli-D-14-00361.1 (2015).
- 14 Bishop, C. H. & Abramowitz, G. Climate model dependence and the replicate Earth paradigm. *Clim Dynam* **41**, 885–900, doi:10.1007/s00382-012-1610-y (2013).
- 15 Abramowitz, G. & Bishop, C. H. Climate Model Dependence and the Ensemble Dependence Transformation of CMIP Projections. *J Climate* **28**, 2332–2348, doi:10.1175/jcli-d-14-00364.1 (2015).
- 16 Alexander, K. & Easterbrook, S. M. The software architecture of climate models: a graphical comparison of CMIP5 and EMICAR5 configurations. *Geosci Model Dev* **8**, 1221–1232, doi:10.5194/gmd-8-1221-2015 (2015).
- 17 Knutti, R. The end of model democracy? *Climatic Change* **102**, 395–404, doi:10.1007/s10584-010-9800-2 (2010).
- 18 Eyring, V. *et al.* ESMValTool (v1.0) – a community diagnostic and performance metrics tool for routine evaluation of Earth system models in CMIP. *Geosci. Model Dev.* **9**, 1747–1802, doi:10.5194/gmd-9-1747-2016 (2016).
- 19 Gleckler, P. *et al.* A More Powerful Reality Test for Climate Models. *Eos* **97**, doi:10.1029/2016eo051663 (2016).
- 20 Lauer, A. *et al.* Process-level improvements in CMIP5 models and their impact on tropical variability, the Southern Ocean, and monsoons. *Earth Syst Dynam* **9**, 33–67, doi:10.5194/esd-9-33-2018 (2018).
- 21 Ma, C.-C., Mechoso, C. R., Robertson, A. W. & Arakawa, A. Peruvian Stratus Clouds and the Tropical Pacific Circulation: A Coupled Ocean-Atmosphere GCM Study. *J Climate* **9**, 1635–1645, doi:10.1175/1520-0442(1996)009<1635:pscatt>2.0.co;2 (1996).
- 22 Hourdin, F. *et al.* Parameterization of convective transport in the boundary layer and its impact on the representation of the diurnal cycle of wind and dust emissions. *Atmos. Chem. Phys.* **15**, 6775–6788, doi:10.5194/acp-15-6775-2015 (2015).
- 23 Richter, I. Climate model biases in the eastern tropical oceans: causes, impacts and ways forward. *Wiley Interdisciplinary Reviews: Climate Change* **6**, 345–358, doi:10.1002/wcc.338 (2015).

- 24 Wang, C. Z., Zhang, L. P., Lee, S. K., Wu, L. X. & Mechoso, C. R. A global perspective on CMIP5 climate model biases. *Nat Clim Change* **4**, 201-205, doi:10.1038/Nclimate2118 (2014).
- 25 Fyfe, J. C., Gillett, N. P. & Zwiers, F. W. Overestimated global warming over the past 20 years. *Nat Clim Change* **3**, 767-769, doi:10.1038/nclimate1972 (2013).
- 26 Meehl, G. A., Hu, A., Arblaster, J. M., Fasullo, J. & Trenberth, K. E. Externally Forced and Internally Generated Decadal Climate Variability Associated with the Interdecadal Pacific Oscillation. *J Climate* **26**, 7298-7310, doi:10.1175/jcli-d-12-00548.1 (2013).
- 27 Fyfe, J. C. *et al.* Making sense of the early-2000s warming slowdown. *Nature Clim. Change* **6**, 224-228, doi:10.1038/nclimate2938 (2016).
- 28 Santer, B. D. *et al.* Volcanic contribution to decadal changes in tropospheric temperature. *Nat Geosci* **7**, 185, doi:10.1038/ngeo2098 (2014).
- 29 Meehl, G. A., Teng, H. & Arblaster, J. M. Climate model simulations of the observed early-2000s hiatus of global warming. *Nat Clim Change* **4**, 898, doi:10.1038/nclimate2357 (2014).
- 30 Thoma, M., Greatbatch, R. J., Kadow, C. & Gerdes, R. Decadal hindcasts initialized using observed surface wind stress: Evaluation and prediction out to 2024. *Geophys Res Lett* **42**, 6454-6461, doi:10.1002/2015GL064833 (2015).
- 31 Meehl, G. A., Hu, A. & Teng, H. Initialized decadal prediction for transition to positive phase of the Interdecadal Pacific Oscillation. *Nat Commun* **7**, 11718, doi:10.1038/ncomms11718 (2016).
- 32 Mears, C. A., Santer, B. D., Wentz, F. J., Taylor, K. E. & Wehner, M. F. Relationship between temperature and precipitable water changes over tropical oceans. *Geophys Res Lett* **34**, doi:10.1029/2007GL031936 (2007).
- 33 Karl, T. R. *et al.* Possible artifacts of data biases in the recent global surface warming hiatus. *Science* **348**, 1469-1472, doi:10.1126/science.aaa5632 (2015).
- 34 Mears, C. A. & Wentz, F. J. The Effect of Diurnal Correction on Satellite-Derived Lower Tropospheric Temperature. *Science* **309**, 1548-1551, doi:10.1126/science.1114772 (2005).
- 35 Mauritsen, T. *et al.* Tuning the climate of a global model. *Journal of Advances in Modeling Earth Systems* **4**, doi:10.1029/2012ms000154 (2012).
- 36 Hourdin, F. *et al.* The art and science of climate model tuning. *B Am Meteorol Soc* **0**, null, doi:10.1175/bams-d-15-00135.1 (2016).
- 37 Bodas-Salcedo, A. *et al.* COSP Satellite simulation software for model assessment. *B Am Meteorol Soc* **92**, 1023-1043, doi:10.1175/2011bams2856.1 (2011).
- 38 Eyring, V. *et al.* Towards improved and more routine Earth system model evaluation in CMIP. *Earth Syst. Dynam.* **7**, 813-830, doi:10.5194/esd-7-813-2016 (2016).
- 39 Phillips, A. S., Deser, C. & Fasullo, J. Evaluating Modes of Variability in Climate Models. *Eos Trans. AGU* **95(49)**, 453-455, doi:10.1002/2014EO490002 (2014).
- 40 Luo, Y. Q. *et al.* A framework for benchmarking land models. *Biogeosciences* **9**, 3857-3874, doi:10.5194/bg-9-3857-2012 (2012).
- 41 Prabhat *et al.* TECA: A Parallel Toolkit for Extreme Climate Analysis. *Procedia Comput Sci* **9**, 866-876, doi:10.1016/j.procs.2012.04.093 (2012).
- 42 Hall, A. & Qu, X. Using the current seasonal cycle to constrain snow albedo feedback in future climate change. *Geophys Res Lett* **33**, doi:10.1029/2005gl025127 (2006).
- 43 Allen, M. R. & Ingram, W. J. Constraints on future changes in climate and the hydrologic cycle. *Nature* **419**, 224-232, doi:10.1038/nature01092 (2002).
- 44 Massonnet, F. *et al.* Constraining projections of summer Arctic sea ice. *Cryosphere* **6**, 1383-1394, doi:10.5194/tc-6-1383-2012 (2012).

- 45 O’Gorman, P. A. Sensitivity of tropical precipitation extremes to climate change. *Nat Geosci* **5**, 697, doi:10.1038/ngeo1568 (2012).
- 46 Cox, P. M. *et al.* Sensitivity of tropical carbon to climate change constrained by carbon dioxide variability. *Nature* **494**, 341-344, doi:Doi 10.1038/Nature11882 (2013).
- 47 Kidston, J. & Gerber, E. P. Intermodel variability of the poleward shift of the austral jet stream in the CMIP3 integrations linked to biases in 20th century climatology. *Geophys Res Lett* **37**, doi:Doi 10.1029/2010gl042873 (2010).
- 48 Tsushima, Y. *et al.* Robustness, uncertainties, and emergent constraints in the radiative responses of stratocumulus cloud regimes to future warming. *Clim Dynam* **46**, 3025-3039, doi:10.1007/s00382-015-2750-7 (2016).
- 49 Brient, F. & Bony, S. Interpretation of the positive low-cloud feedback predicted by a climate model under global warming. *Clim Dynam* **40**, 2415-2431, doi:10.1007/s00382-011-1279-7 (2013).
- 50 Brient, F. & Schneider, T. Constraints on Climate Sensitivity from Space-Based Measurements of Low-Cloud Reflection. *J Climate* **29**, 5821-5835, doi:10.1175/jcli-d-15-0897.1 (2016).
- 51 Lipat, B. R., Tselioudis, G., Grise, K. M. & Polvani, L. M. CMIP5 models' shortwave cloud radiative response and climate sensitivity linked to the climatological Hadley cell extent. *Geophys Res Lett* **44**, 5739-5748, doi:10.1002/2017GL073151 (2017).
- 52 Sherwood, S. C., Bony, S. & Dufresne, J. L. Spread in model climate sensitivity traced to atmospheric convective mixing. *Nature* **505**, 37-42, doi:10.1038/nature12829 (2014).
- 53 Tian, B. Spread of model climate sensitivity linked to double-Intertropical Convergence Zone bias. *Geophys Res Lett* **42**, 4133-4141, doi:10.1002/2015GL064119 (2015).
- 54 Cox, P. M., Huntingford, C. & Williamson, M. S. Emergent constraint on equilibrium climate sensitivity from global temperature variability. *Nature* **553**, 319-322, doi:10.1038/nature25450 (2018).
- 55 Klein, S. A. & Hall, A. Emergent Constraints for Cloud Feedbacks. *Current Climate Change Reports* **1**, 276-287, doi:10.1007/s40641-015-0027-1 (2015).
- 56 Dessler, A. E. & Forster, P. M. An Estimate of Equilibrium Climate Sensitivity From Interannual Variability. *Journal of Geophysical Research: Atmospheres* **123**, 8634-8645, doi:doi:10.1029/2018JD028481 (2018).
- 57 DeAngelis, A. M., Qu, X., Zelinka, M. D. & Hall, A. An observational radiative constraint on hydrologic cycle intensification. *Nature* **528**, 249-253, doi:10.1038/nature15770 (2015).
- 58 Li, G., Xie, S.-P., He, C. & Chen, Z. Western Pacific emergent constraint lowers projected increase in Indian summer monsoon rainfall. *Nat Clim Change* **7**, 708, doi:10.1038/nclimate3387 (2017).
- 59 Wenzel, S., Cox, P. M., Eyring, V. & Friedlingstein, P. Emergent constraints on climate-carbon cycle feedbacks in the CMIP5 Earth system models. *J Geophys Res-Bioge* **119**, 794-807, doi:10.1002/2013jg002591 (2014).
- 60 Wenzel, S., Cox, P. M., Eyring, V. & Friedlingstein, P. Projected land photosynthesis constrained by changes in the seasonal cycle of atmospheric CO<sub>2</sub>. *Nature* **538**, 499-501, doi:10.1038/nature19772 (2016).
- 61 Kwiatkowski, L. *et al.* Emergent constraints on projections of declining primary production in the tropical oceans. *Nat Clim Change* **7**, 355+, doi:10.1038/Nclimate3265 (2017).
- 62 Chadburn, S. E. *et al.* An observation-based constraint on permafrost loss as a function of global warming. *Nat Clim Change* **7**, 340–344, doi:10.1038/Nclimate3262 (2017).
- 63 Hoffman, F. M. *et al.* Causes and implications of persistent atmospheric carbon dioxide biases in Earth System Models. *J Geophys Res-Bioge* **119**, 141-162, doi:10.1002/2013jg002381 (2014).

- 64 Donat, M. G., Pitman, A. J. & Angéilil, O. Understanding and reducing future uncertainty in mid-latitude daily heat extremes via land surface feedback constraints. *Geophys Res Lett*, doi:doi:10.1029/2018GL079128 (2018).
- 65 Qu, X. & Hall, A. What Controls the Strength of Snow-Albedo Feedback? *J Climate* **20**, 3971-3981, doi:10.1175/jcli4186.1 (2007).
- 66 Waugh, D. W. & Eyring, V. Quantitative performance metrics for stratospheric-resolving chemistry-climate models. *Atmos. Chem. Phys.* **8**, 5699-5713, doi:10.5194/acp-8-5699-2008 (2008).
- 67 Karpechko, A. Y., Maraun, D. & Eyring, V. Improving Antarctic Total Ozone Projections by a Process-Oriented Multiple Diagnostic Ensemble Regression. *J Atmos Sci* **70**, 3959-3976, doi:10.1175/jas-d-13-071.1 (2013).
- 68 Räisänen, J., Ruokolainen, L. & Ylhäisi, J. Weighting of model results for improving best estimates of climate change. *Clim Dynam* **35**, 407-422, doi:10.1007/s00382-009-0659-8 (2010).
- 69 Lorenz, R. *et al.* Prospects and Caveats of Weighting Climate Models for Summer Maximum Temperature Projections Over North America. *Journal of Geophysical Research: Atmospheres* **123**, 4509-4526, doi:doi:10.1029/2017JD027992 (2018).
- 70 Abramowitz, G. *et al.* Model dependence in multi-model climate ensembles: weighting, sub-selection and out-of-sample testing. *Earth Syst. Dynam. Discuss.* **2018**, 1-20, doi:10.5194/esd-2018-51 (2018).
- 71 Herger, N. *et al.* Selecting a climate model subset to optimise key ensemble properties. *Earth Syst. Dynam.* **9**, 135-151, doi:10.5194/esd-9-135-2018 (2018).
- 72 Herger, N. *et al.* Calibrating Climate Model Ensembles for Assessing Extremes in a Changing Climate. *Journal of Geophysical Research: Atmospheres* **123**, 5988-6004, doi:doi:10.1029/2018JD028549 (2018).
- 73 Santer, B. D. *et al.* Incorporating model quality information in climate change detection and attribution studies. *Proceedings of the National Academy of Sciences* **106**, 14778-14783, doi:10.1073/pnas.0901736106 (2009).
- 74 Caldwell, P. M. *et al.* Statistical significance of climate sensitivity predictors obtained by data mining. *Geophys Res Lett* **41**, 1803-1808, doi:10.1002/2014gl059205 (2014).
- 75 Masson, D. & Knutti, R. Climate model genealogy. *Geophys Res Lett* **38**, doi:10.1029/2011gl046864 (2011).
- 76 Pennell, C. & Reichler, T. On the Effective Number of Climate Models. *J Climate* **24**, 2358-2367, doi:10.1175/2010jcli3814.1 (2011).
- 77 Sunyer, M. A., Madsen, H., Rosbjerg, D. & Arnbjerg-Nielsen, K. A Bayesian Approach for Uncertainty Quantification of Extreme Precipitation Projections Including Climate Model Interdependency and Nonstationary Bias. *J Climate* **27**, 7113-7132, doi:10.1175/jcli-d-13-00589.1 (2014).
- 78 Sanderson, B. M., Knutti, R. & Caldwell, P. A Representative Democracy to Reduce Interdependency in a Multimodel Ensemble. *J Climate* **28**, 5171-5194, doi:10.1175/Jcli-D-14-00362.1 (2015).
- 79 Zhang, X. *et al.* Indices for monitoring changes in extremes based on daily temperature and precipitation data. *Wiley Interdisciplinary Reviews: Climate Change* **2**, 851-870, doi:10.1002/wcc.147 (2011).
- 80 Sillmann, J., Kharin, V. V., Zhang, X., Zwiers, F. W. & Bronaugh, D. Climate extremes indices in the CMIP5 multimodel ensemble: Part 1. Model evaluation in the present climate. *Journal of Geophysical Research: Atmospheres* **118**, 1716-1733, doi:10.1002/jgrd.50203 (2013).

- 81 Pendergrass, A. G. & Hartmann, D. L. The Atmospheric Energy Constraint on Global-Mean Precipitation Change. *J Climate* **27**, 757-768, doi:10.1175/jcli-d-13-00163.1 (2014).
- 82 Zscheischler, J. *et al.* Future climate risk from compound events. *Nat Clim Change* **8**, 469-477, doi:10.1038/s41558-018-0156-3 (2018).
- 83 Gutowski Jr, W. J. *et al.* WCRP COordinated Regional Downscaling EXperiment (CORDEX): a diagnostic MIP for CMIP6. *Geosci. Model Dev.* **9**, 4087-4095, doi:10.5194/gmd-9-4087-2016 (2016).
- 84 van Oldenborgh, G. J., Otto, F. E. L., Haustein, K. & Cullen, H. Climate change increases the probability of heavy rains like those of storm Desmond in the UK – an event attribution study in near-real time. *Hydrol. Earth Syst. Sci. Discuss.* **2015**, 13197-13216, doi:10.5194/hessd-12-13197-2015 (2015).
- 85 Ruane, A. C., Goldberg, R. & Chryssanthacopoulos, J. Climate forcing datasets for agricultural modeling: Merged products for gap-filling and historical climate series estimation. *Agricultural and Forest Meteorology* **200**, 233-248, doi:<https://doi.org/10.1016/j.agrformet.2014.09.016> (2015).
- 86 Elliott, J. *et al.* The Global Gridded Crop Model Intercomparison: data and modeling protocols for Phase 1 (v1.0). *Geosci. Model Dev.* **8**, 261-277, doi:10.5194/gmd-8-261-2015 (2015).
- 87 Ruane, A. C. *et al.* The Vulnerability, Impacts, Adaptation and Climate Services Advisory Board (VIACS AB v1.0) contribution to CMIP6. *Geosci. Model Dev.* **9**, 3493-3515, doi:10.5194/gmd-9-3493-2016 (2016).
- 88 Warszawski, L. *et al.* The Inter-Sectoral Impact Model Intercomparison Project (ISI-MIP): Project framework. *Proceedings of the National Academy of Sciences* **111**, 3228-3232, doi:10.1073/pnas.1312330110 (2014).
- 89 Ruane, A. C. & McDermid, S. P. Selection of a representative subset of global climate models that captures the profile of regional changes for integrated climate impacts assessment. *Earth Perspectives* **4**, doi:10.1186/s40322-017-0036-4 (2017).
- 90 Stevens, B. & Bony, S. What Are Climate Models Missing? *Science* **340**, 1053-1054, doi:DOI 10.1126/science.1237554 (2013).
- 91 Webb, M. J. *et al.* The Cloud Feedback Model Intercomparison Project (CFMIP) contribution to CMIP6. *Geosci. Model Dev.* **10**, 359-384, doi:10.5194/gmd-10-359-2017 (2017).
- 92 Haarsma, R. J. *et al.* High Resolution Model Intercomparison Project (HighResMIP v1.0) for CMIP6. *Geosci. Model Dev.* **9**, 4185-4208, doi:10.5194/gmd-9-4185-2016 (2016).
- 93 Lawrence, D. M. *et al.* The Land Use Model Intercomparison Project (LUMIP) contribution to CMIP6: rationale and experimental design. *Geosci. Model Dev.* **9**, 2973-2998, doi:10.5194/gmd-9-2973-2016 (2016).
- 94 Mears, C. A., Wentz, F. J., Thorne, P. & Bernie, D. Assessing uncertainty in estimates of atmospheric temperature changes from MSU and AMSU using a Monte-Carlo estimation technique. *J Geophys Res-Atmos* **116**, doi:10.1029/2010jd014954 (2011).
- 95 Argo. Argo float data and metadata from Global Data Assembly Centre (Argo GDAC). *SEANOE*, doi:<http://doi.org/10.17882/42182> (2000).
- 96 Malavelle, F. F. *et al.* Strong constraints on aerosol–cloud interactions from volcanic eruptions. *Nature* **546**, 485, doi:10.1038/nature22974 (2017).
- 97 Fountalis, I., Bracco, A. & Dovrolis, C. ENSO in CMIP5 simulations: network connectivity from the recent past to the twenty-third century. *Clim Dynam* **45**, 511-538, doi:10.1007/s00382-014-2412-1 (2015).

- 98 Barz, B., Rodner, E., Guanche Garcia, Y. & Denzler, J. Detecting Regions of Maximal Divergence for Spatio-Temporal Anomaly Detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, doi:10.1109/TPAMI.2018.2823766 (2018).
- 99 Runge, J. *et al.* Identifying causal gateways and mediators in complex spatio-temporal systems. *Nat Commun* **6**, 8502, doi:10.1038/ncomms9502 (2015).
- 100 Kageyama, M. *et al.* PMIP4-CMIP6: the contribution of the Paleoclimate Modelling Intercomparison Project to CMIP6. *Geosci. Model Dev. Discuss.* **2016**, 1-46, doi:10.5194/gmd-2016-106 (2016).

## **Acknowledgements**

The authors acknowledge the Aspen Global Change Institute (AGCI) for hosting a workshop on Earth System Model Evaluation to Improve Process Understanding in August 2017 as part of its traditionally landmark summer interdisciplinary sessions (<http://www.agci.org/event/17s2>). NASA, the Heising-Simons Foundation, Horizon 2020 European Union's Framework Programme for Research and Innovation under Grant Agreement No 641816, the Coordinated Research in Earth Systems and Climate: Experiments, kNowledge, Dissemination and Outreach (CRESCENDO) project, the ESA Climate Change Initiative (CCI) Climate Model User Group (CMUG), WCRP and the Department of Energy (DOE) all provided support for the workshop. The perspective presented here substantially draws on conclusions from that workshop. Portions of this study were supported by the Regional and Global Climate Modeling Program (RGCM) of the U.S. Department of Energy's Office of Biological & Environmental Research (BER) Cooperative Agreement DE-FC02-97ER62402 and Contract No. DE-AC05-00OR22725) and the National Science Foundation. The National Center for Atmospheric Research is sponsored by the National Science Foundation.

## **Author contributions**

V.E., P.M.C., G.M.F., and P.J.G. were the co-chairs of the AGCI workshop and led the writing of the paper. All authors participated in the AGCI workshop and contributed to discussions and writing of the text.

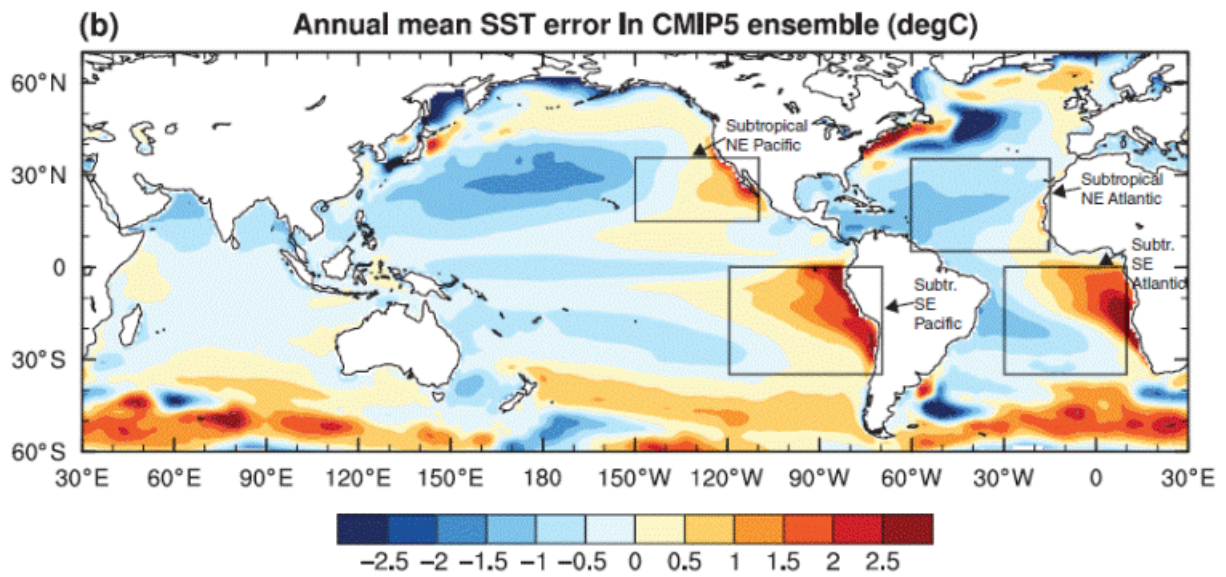
## **Additional information**

Reprints and permissions information is available at [www.nature.com/reprints](http://www.nature.com/reprints). Correspondence and requests for materials should be addressed to V. E. ([veronika.eyring@dlr.de](mailto:veronika.eyring@dlr.de)).

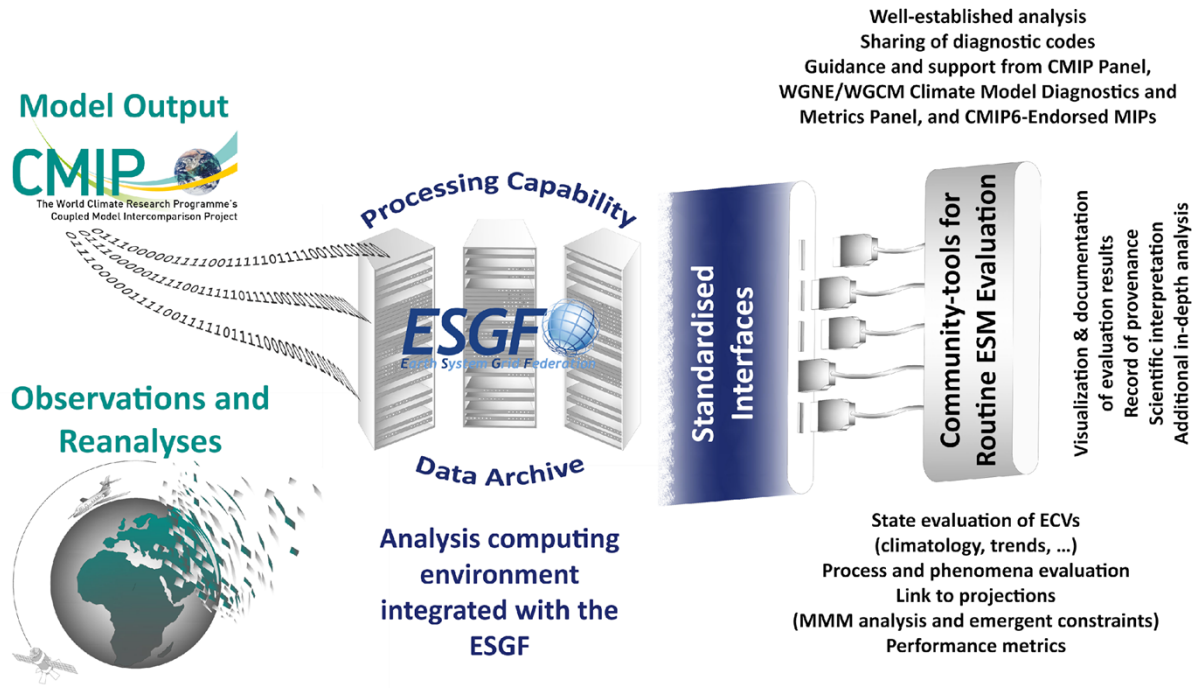
## **Competing financial interests**

The authors have no competing financial interests.

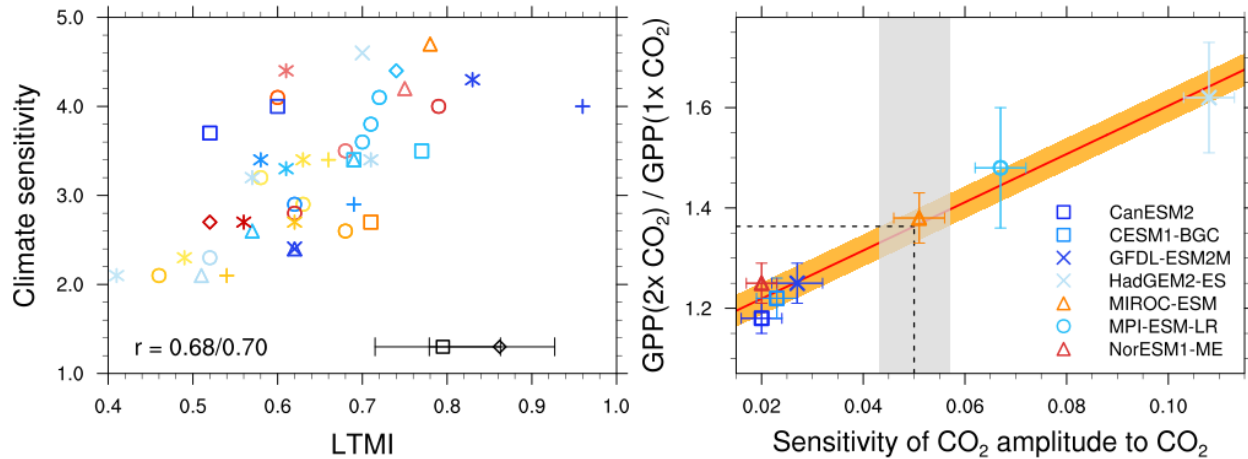
## Figures



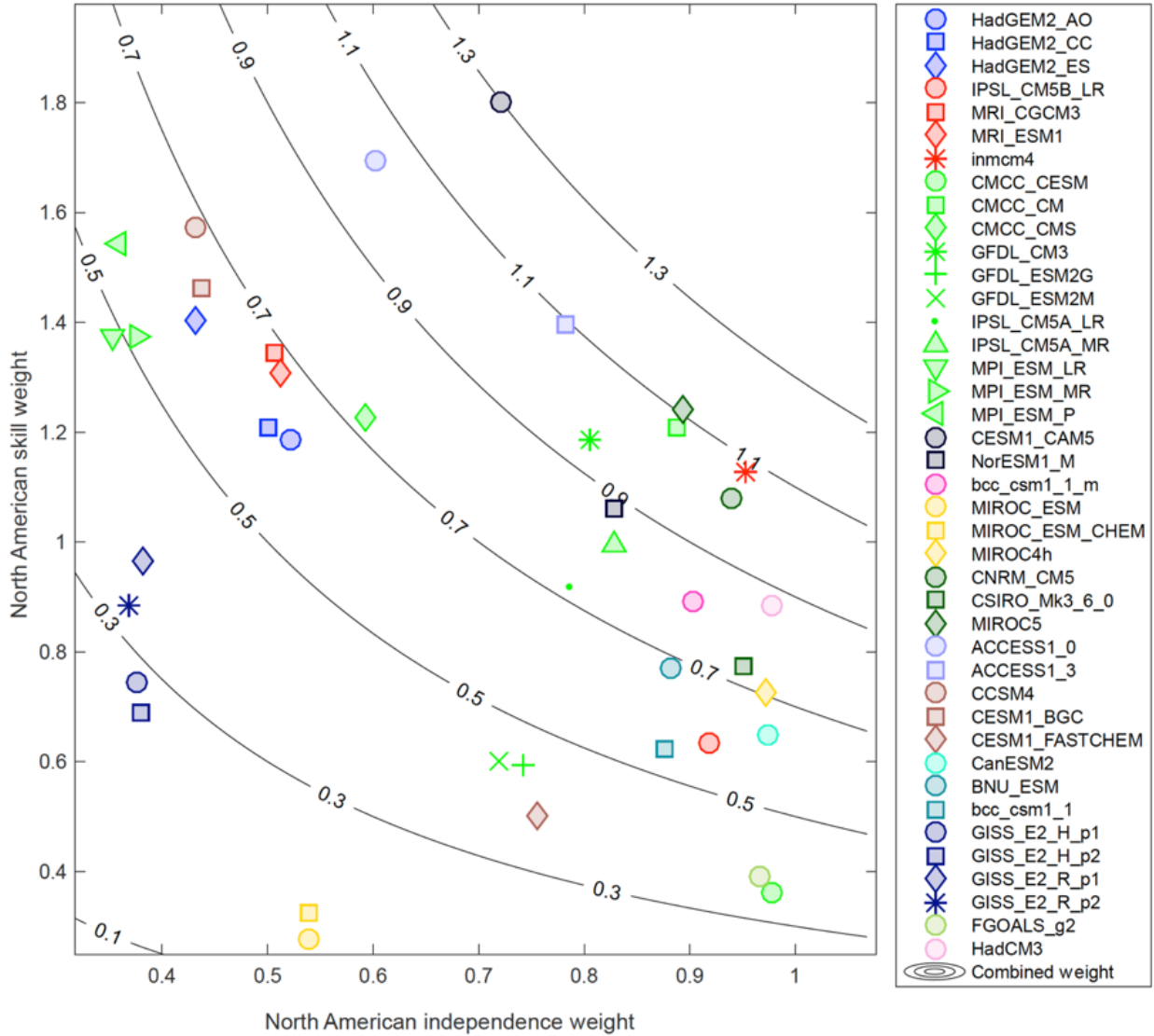
**Figure 1:** Annual mean sea surface temperature error from the CMIP5 multi-model ensemble. Systematic errors are particularly visible in the upwelling zones denoted by black rectangles off the west coasts of each continent. Figure reproduced with permission from ref. <sup>23</sup>, © Wiley.



**Figure 2: Schematic diagram of the workflow for CMIP Evaluation Tools running alongside the ESGF.** The tools will routinely produce a broad characterization of model performance for CMIP model output utilizing relevant observations and reanalyses and running alongside the ESGF infrastructure. Figure reproduced with permission from ref. <sup>38</sup>, © EGU.



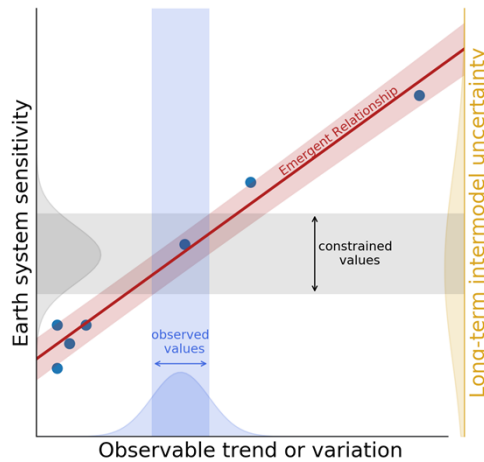
**Figure 3: Examples of newly developed physical and biogeochemical emergent constraints since the AR5.** *Left:* Emergent constraint on equilibrium climate sensitivity showing a correlation between ECS and a lower-tropospheric mixing index (LTMI) from 43 CMIP5 models. LTMI is calculated as the sum of an index S for the small-scale component of mixing that is proportional to the differences of temperature and relative humidity between 700 hPa and 850 hPa and index D for the large-scale lower-tropospheric mixing. The linear correlation coefficient  $r$  and error bars of the two reanalyses ERA-I and MERRA are given in addition. *Right:* Emergent constraint on the relative increase of large-scale GPP for a doubling of CO<sub>2</sub> showing a correlation between the increase in the amplitude of the CO<sub>2</sub> seasonal cycle with increases in annual mean CO<sub>2</sub> atmospheric concentrations at Point Barrow (BRW: 71.3°N, 156.6°W) and the high-latitude (60°N–90°N) CO<sub>2</sub> fertilization on GPP at  $2 \times \text{CO}_2$ . Figure reproduced with permission from ref. <sup>52</sup> (left) and ref. <sup>60</sup> (right), © Nature.



**Figure 4: Model skill and independence weights for the CMIP5 archive evaluated over the CONUS/Canada domain.** Contours show the overall weighting, which is the product of the two individual weights. Figure reproduced with permission from ref. <sup>12</sup> © GMD.

### Box 1. Emergent Constraints

An *emergent constraint* on an Earth System sensitivity requires two key components. Firstly, there needs to be an *emergent relationship* between the sensitivity (y-axis) and some measure of variation in the contemporary climate (x-axis), which is evident across an ensemble of ESMs. Secondly, there needs to be an observation of the variation (x-axis) in the real world, together with a measure of observational uncertainty. Ideally, the observation should have less uncertainty than the spread of the x-axis variable within the model ensemble. The observation of the x-axis variable and the model-derived emergent relationship between the y and x variables can then be combined to give a constraint on the Earth System sensitivity. The resulting emergent constraint is conditional on the model ensemble providing a realistic emergent relationship, and also on the availability of sufficiently accurate observations.



**Box Emergent Constraints, Figure 1.** Schematic diagram illustrating the concept of Emergent Constraints. Each blue dot represents (hypothetical) output from different ESMs. The comparison here involves a diagnostic based on a model's performance in a historical simulation (x-axis) and in a projection of future climate change (y-axis). The relationship between the past or present-day diagnostic and future projection illustrates an emergent relationship, which is normally quantified by a linear regression (red line). Once a physically plausible relationship is found, observations can be used to reduce the intermodel uncertainty in the long-term projection (compare the yellow and grey probability distribution functions). Uncertainty in the new projection (grey shading) arises from two sources: uncertainty in the observational constraint (blue shading) and in the linear regression (red shading).