Strength in Numbers: Insights from Initial-condition Large Ensembles with Multiple Earth System Models and Future Prospects

US CLIVAR Working Group on Large Ensembles
[C. Deser*, F. Lehner, K.B. Rodgers, T. Ault, T.L. Delworth, P.N. DiNezio, A. Fiore, C. Frankignoul, J. C. Fyfe, D.E. Horton, J.E. Kay, R. Knutti, N.S. Lovenduski, J. Marotzke, K.A. McKinnon, S. Minobe, J. Randerson, J.A. Screen, I.R. Simpson and M. Ting]

Perspective in *Nature Climate Change*
Submitted June 21, 2019; Revised November 27, 2019

**Corresponding author: Dr. Clara Deser, National Center for Atmospheric Research, Boulder CO cdeser@ucar.edu*

1. Abstract
Internal variability in the climate system confounds assessment of human-induced climate change and imposes irreducible limits on the accuracy of climate change projections, especially at regional and decadal scales. A new collection of initial-condition large ensembles performed with seven Earth System Models under historical and future radiative forcing scenarios provides new insights into uncertainties due to internal variability *vs*. model differences. These data enhance the assessment of climate change risks including extreme events. In addition, they offer a powerful testbed for new methodologies aimed at separating forced signals from internal variability in the observational record. Opportunities and challenges confronting the design and dissemination of future large ensembles, including consideration of increased spatial resolution and model complexity along with emerging earth system applications, are discussed.

2. Introduction
Identifying anthropogenic influences on weather and climate amidst the background of internal variability, and providing projections of future changes, are central scientific challenges with practical implications[1–6]. Since the inception of the Coupled Model Intercomparison Project (CMIP) nearly two decades ago, substantial progress has been made on quantifying sources of uncertainty in climate projections (e.g., ref[7–9]). However, such multimodel archives confound uncertainties arising from differences in model formulation (i.e., structural uncertainty) with those generated by internal variability (variability arising from processes intrinsic to the coupled ocean-atmosphere-land-biosphere-cryosphere system). This distinction is important, because the former is potentially reducible as models improve, whereas the latter is an intrinsic property of each model and is largely irreducible after the memory of initial conditions is lost, typically after less than a few years over land[10]. This key distinction is often not widely appreciated and communicated to stakeholder groups[11]. Indeed, internal variability accounts for approximately half of the inter-model spread within the CMIP archive for projected changes in near surface air temperature, precipitation and runoff across North America and Europe over the next 50 years [5,8,9,12–14].

One way to isolate the contribution of uncertainty due to internal variability is to perform an ensemble of simulations with a single fully-coupled global climate model under a particular radiative forcing scenario, applying perturbations to the initial conditions of each member in order to create diverging weather and climate trajectories, causing ensemble spread (e.g., ref[12,15–17]). Since the resulting sequences of unpredictable internal variability are randomly phased between the individual ensemble members, the forced response can be estimated by averaging over a sufficient number of members. The definition of "sufficient" depends on the quantity of interest, location, spatial scale, temporal scale, and time horizon, often on the order of 10-100 members (e.g., ref[12]). Such "initial-condition Large Ensembles" conducted with fully-coupled global models (hereafter referred to as "LEs") are a relatively new development in climate sciences, with the first efforts employing CMIP3-era models[12,18].

The past few years have witnessed an explosion of LEs with newer-generation CMIP5-class Earth System Models (ESMs; Table 1). Each LE required substantial high performance computing resources to produce, and generated hundreds of terabytes of output. For example, the CESM1 LE used 21 million CPU hours and produced over 600 terabytes of model output (for comparison, the entire CESM1 contribution to CMIP5 was 170 terabytes). Making these "big data" projects accessible to a wide range of users is challenging. Yet, their ease-of-use for different types of analysis work-flows has a substantial impact on the scientific value gained from their production. A case in point is the NCAR CESM1-LE Project[19], which from the outset had an explicit goal of serving a broad research community by responding to user needs to provide easy access to the output and stable on-disk access. This project has resulted in more than 750 peer-reviewed studies to date, with approximately 400,000 data files downloaded from spinning disk. Remaining nimble to new workflows and users is important, as is following the recommended "big data" practice of "bringing your analysis to your data". Following these principles, the CESM1-LE was made freely available as a public dataset on the Amazon Web Services cloud in autumn 2019. Access on the commercial cloud demonstrates strong interest in LEs from industry and scientific communities well beyond typical climate researchers that have historically used climate models. Such scrutiny and widespread use attests to the enormous value of LEs for a range of applications: truly a "sea-change" for climate and related sciences.

3. Strength in Numbers: a Multi-Model Large Ensemble Archive

While a single model LE has enormous utility, a multimodel collection of LEs can be leveraged for robust comparison of the forced response on regional/decadal scales across models, as well as of the characteristics of internal variability across models. It can also advance model evaluation by providing more complete information on biases in internal variability *vs*. those in the forced response. Unlike CMIP, a multimodel archive of LEs allows for direct separation of projection uncertainty into a structural component due to model differences and an internal variability component. Despite these advantages, most analyses to date have been limited to one or at most two LEs (with a few exceptions, e.g., refs[20,21]), in part because of the burdensome task of accessing large volumes of data from disparate sources. To fill this gap, we have produced a centralized data repository of LEs conducted with seven different CMIP5-class ESMs under historical and future emissions scenarios (hereafter referred to as the "Multi-Model Large Ensemble Archive" or MMLEA; Table 1). This repository includes gridded fields of key variables at daily and monthly resolution, and is easily accessible via the NCAR Climate Data Gateway (https://www.earthsystemgrid.org/dataset/ucar.cgd.ccsm4.CLIVAR_LE.html).

This Perspective seeks to illustrate some of the new insights that can be gained from the MMLEA, with the aim of widening its usage and stimulating new research directions including emerging Earth system applications. We also look to the future of initial-condition LEs, in particular the opportunities and challenges that confront their design and facilitate their accessibility to the broad user community. In this regard, we offer a path forward that balances demands for increased spatial resolution and model complexity against ensemble size. We encourage future phases of CMIP to take on a greater role in the design of LE simulations and in coordinating their data storage and access.

4. New insights on separating sources of uncertainties

Individual LEs have been crucial to show that internal variability needs to be considered alongside forced trends in past and future climate change at continental and smaller spatial scales (i.e., refs[10,12,14,19,22–30]). The MMLEA expands on this view by providing new insights on the relative roles of internal variability and model structural differences -- two sources of projection uncertainty in addition to radiative forcing scenario. The MMLEA shows that both factors can play a first-order role in the magnitude and pattern of warming at continental scales. As an example, Fig. 1 show the distributions of trends in North American air temperatures over the last 60 years from each of the seven LEs (Methods). While they all encompass the observed trend value, they clearly differ in the strength of the forced trend (given by the ensemble mean) and in the shape and width of the distribution of trends, which emerges due to the influence of internal variability. This information on model-dependence of both the forced trend and the range of trends due to internal variability is unique to the MMLEA, and could not have been deduced directly from the CMIP archives. It is important to note that a LE that is centered on the single observed trend value does not constitute evidence that this particular model is more realistic than any other model (see further discussion in Section 6).

The distribution of North American temperature trends based on the 40 models in the CMIP5 archive (Methods) is only slightly wider than that based on an individual LE, and is due to both model differences and internal variability (see gray shaded PDF in Fig. 1). Moreover, the MMLEA as a whole spans a wider range than CMIP5, suggesting that CMIP5 under-samples internal variability at regional scales. This highlights the importance of evaluating the realism of models' internal variability of trends, since a model with unrealistically large trend variability (i.e., a broad distribution) can encompass the observed trend for the wrong reason and would also inflate uncertainty in future projections. Approaches to address this challenge are discussed in Section 6.

Just as North American temperature trends vary across the individual members of a LE, the geographical pattern of trends can also be strikingly different (row of maps at the bottom of Fig. 1). This can confound comparisons of individual simulations from different models and lead to erroneous interpretations, since internal variability might be mistaken for structural differences. With enough members, the spatial pattern of the forced response emerges for each model, allowing for a direct comparison between models. Models may show similar forced patterns of poleward-amplified warming but different overall amplitudes (top left and right maps in Fig. 1), a conclusion that would have been difficult to discern without an MMLEA. Similar issues confront the study of trends observed in the real world (middle map in the top row of Fig. 1), since these are also just one realization of many that could have happened (see Section 6).

Quantifying model uncertainty requires knowledge of the forced response in each model – but most models in past and current CMIPs do not have enough ensemble members to allow for a robust estimate of its forced response. Instead, low-frequency statistical fits to a single ensemble member are often used to estimate the forced response (e.g., refs[8,9]). Consequently, internal variability has to be estimated either from the residual of this fit or from long pre-industrial control simulations. From these approaches it is often not easy or possible to robustly estimate systematic changes to internal variability under increasing radiative forcing. The availability of an MMLEA circumvents these limitations and assumptions. More importantly, it allows one to separate the sources of uncertainty at smaller spatial and temporal scales, and for quantities that are notoriously variable such as precipitation and extremes.

5. Decision-making and risk assessment in a highly variable climate system

LEs are increasingly proving their utility in the context of real-world decision-making[31] where full assessment of changing climate risks is needed, including variability and extremes. In particular, discerning changes in variability and extremes requires large sample sizes[32–36], the hallmark of LEs. Moreover, the MMLEA is critical for evaluating the extent to which projected changes in variability and extremes are model dependent.

The Upper Colorado River basin – which feeds the largest reservoirs in the US – is a clear example of where changes in mean and variability can produce a wide range of climate risks for water managers. This basin is located at a latitude where projected changes in precipitation are notoriously uncertain – the transition zone between the expected drying in the subtropics and the wetting at high latitudes[2,37–39]. The MMLEA shows divergent outcomes regarding how decadal mean precipitation will change in this region under a high-emissions scenario (Fig. 2a). However, decadal variability of precipitation is projected to increase, on average by about 10% of the magnitude of the forced change (Fig. 2b). This result by itself suggests a heightened hazard of prolonged droughts and pluvials, and could, in the absence of consistent projections of changes in the mean, provide useful information for refining water management strategies.

To illustrate the challenge of projecting extreme events, we use an example of daily summer heat extremes for a location in the south-central United States centered on Dallas, Texas (Methods). As expected under global warming, daily July heat extremes at Dallas are projected to increase over the 21st century; however, their evolution is far from monotonic in any single ensemble member, and their rate and degree of increase varies considerably across different realizations of future internal variability in the same model (Fig. 3a). For instance, historical daily heat records could be broken almost continuously starting in the late 2060s, or their occurrence could be more punctuated, with some decades even as late as the 2090s spared from any days of record heat, depending on how internal variability happens to unfold (Fig. 3a). The variety of temporal expressions of historical heat extreme exceedances across the different members of an LE should be a cautionary note on the enormous impact of internal variability on rare events (see also refs 30 and 31). Results also differ between models, as differences in the amount of warming and in the magnitude of variability combine into an uncertain future risk of exceeding a given threshold (Fig. 3b). Validating not only a model's climatology or mean trend, but also its variability, emerges thus again as an important step when investigating, and ultimately constraining, future projections, in this case of extreme events[40].

Attribution-focused large ensembles differ from those in the MMLEA in that they often rely on regional, or high resolution global, atmosphere-land models in order to capture the small spatial scales of specific extreme events[34–36,41,42] and may prescribe additional boundary conditions such as the large-scale atmospheric circulation[43,44]. Nevertheless, these types of ensemble highlight the large number of simulations required to identify significant shifts in the probability of certain events. We note that LEs can also serve these alternate types of ensemble by providing lateral boundary conditions to more specialized regional climate models[45], and oceanic boundary conditions to higher-resolution global atmosphere-land models.

6. Multi-model LEs as methodological testbeds with application to an 'Observational' LE

Another key usage of LEs is to test methods suitable for application to the observational record, for example those aimed at separating the signals of internal variability and forced climate change from a single realization (e.g., refs[28,29,46–50]). Using observations alone, it is difficult to assess the skill of such separation methods due to lack of true knowledge of the observed forced response or the full range of variability, including extremes. However, separation methods can be evaluated by applying the methodology to each LE ensemble member individually and comparing the results to the model's forced response, estimated from the ensemble mean of the LE (Fig. 4). Application to the MMLEA will identify if the validation has a strong dependence on model structure.

An additional testbed application of model LEs is the development of surrogate realizations of internal variability based on observations (Fig. 4). Although one cannot replay the "tape of history"[51] with an initial-condition perturbation in the real world, the single observed trajectory is only one of many that could have plausibly occurred (under the same boundary conditions and forcing), had a different sequence of internal variability unfolded. This is the underlying premise of LEs: that internal variability can play out with a different (and largely unpredictable) chronology, thereby creating uncertainty in the estimate of trends that are calculated over a finite time interval. Can the sample of internal variability contained within the observational record be used to generate surrogate realizations whose statistical characteristics are largely unchanged, but whose temporal sequences are altered? If so, an observationally-based LE can be developed, wherein these surrogates are added to an estimate of the forced response (derived from models or empirical methods applied to observations) to produce an observationally-constrained range of outcomes (Fig. 4).

Several methods for generating surrogate realizations that aim to preserve the temporal[25] and spatio-temporal characteristics of observed internal variability have been proposed[46,52–57]. To date, these techniques have been applied to terrestrial temperature and precipitation[25,46], sea level pressure[46], and sea-surface temperature[52,54]. These methods interact in two important ways with model LEs. First, model LEs can be used as methodological testbeds to ensure that the statistical ensembles have the desired properties (Fig. 4). Second, after the statistical ensembles are validated, they can then be used to validate the model LEs. We demonstrate this interplay with an example from the "Observational Large Ensemble" (Obs-LE) developed by ref[46] (Methods).

Analogous to the approach mentioned above for estimating the forced trend, the Obs-LE methodology can be cleanly tested in the context of a model LE by creating a statistical ensemble based on a single member of the model LE, and assessing whether the spread of the statistical ensemble is consistent with that of the remaining ensemble members. This procedure can then be repeated for each ensemble member, and the resulting information pooled together to provide a robust estimate of the accuracy of the methodology (Fig. 4). In the case of variability of annual temperature trends over the past 50 years on land, the fractional error of the Obs-LE methodology is generally less than 20% over most of the globe, with slightly larger errors in certain regions of the tropics (Fig. 5a). Assuming the properties of the real world are not drastically different from those of the model, this indicates that applying the same approach to generate a statistical ensemble from the single realization of the real world is valid.

Having validated the Obs-LE approach, one can then assess the realism of internal variability simulated by each model LE by comparison with the Obs-LE. For the case of the CESM1-LE, the model overestimates variability of 50-year temperature trends by up to 50% in parts of western North America and northern Eurasia, and up to 100% in areas of high terrain in the tropics (Fig. 5b). These model biases in variability are larger than the error of the Obs-LE methodology, indicating they are true model biases. Similar results are found for precipitation trend variability, which exhibits regions of both significant underestimation and overestimation in the CESM1-LE[46].

One can also apply the Obs- LE to evaluate the simulated distributions of temperature trends at specific locations. For example, the simulated temperature trend distributions for Dallas, Texas in the CESM1 and MPI LEs narrow considerably when the Obs-LE is used to estimate the internal variability (inset to Fig. 5b), consistent with the models' significant overestimation of variability at this location. This brings the observed trend closer to the lower tail of the distributions. It is worth emphasizing that without an observationally-based LE, it would not have been possible to assess the width of the models' temperature trend distributions, with important implications for constraining future projections.

An important future challenge for the LE community is to develop effective means to evaluate and benchmark the internal variability generated by model LEs. Meeting this challenge requires taking advantage of historical and paleoclimate records, and developing suitable statistical emulation methods to construct observationally-based LEs for other components of the climate system. Statistical emulation of internal variability may also be advantageous in the context of ESMs when the cost of conducting a sufficiently large LE is prohibitive, for example, in the case of models with increased spatial resolution and/or complexity (discussed further below). These statistical emulation methods will need to take into account any projected changes in internal variability[58].

7. Looking to the future of initial-condition LEs

a) Considerations on LE design

The existing LEs have been designed and created independently, with different choices of time period, radiative forcing scenario, number of members and method of initialization (Table 1). In addition, they employ different protocols for data output, storage and access. These differences must be considered when comparing LEs across models, as each has ramifications.

Initialization

In some LEs, the initial conditions are created by introducing miniscule (at the level of round-off error or 10^{-14} K) perturbations into the atmosphere only ("micro perturbation"[15]). The rapid growth of atmospheric perturbations makes this technique well suited for studies involving atmospheric variability and trends. However, for phenomena with long persistence involving oceanic or terrestrial processes, such as sea level, ocean heat content, biogeochemistry, and soil moisture, it may be more desirable to start each member from completely different initial conditions in the ocean and other components ("macro perturbations") to more fully sample different possible climate trajectories. Macro perturbations can increase the ensemble utility, but can introduce complications related to subsurface ocean drift in the control simulation that can influence ocean initial conditions, and thus require long and quasi-equilibrated control simulations to choose initial conditions from[59]. A combination of micro and macro perturbations could have the most scientific benefit, but the issue of ensemble initialization clearly needs close examination, and potential coordination between multiple LE projects.

Length of simulation and ensemble size

For a given amount of computer time, a choice has to be made between the length of the simulations versus the number of ensemble members. For example, is it better (for some purposes) to have a 100-member ensemble covering the period 1981-2040 or a 50-member ensemble extending over 1981-2100? Furthermore, if higher spatial-resolution is critical, such as for the simulation of some climate extremes, this usually comes at the expense of the total number of ensemble members that can be run. The optimal balance between ensemble size and spatial resolution will depend on the specific purposes of the LE (see also ref[60]).

Radiative forcing scenario

The choice of forcing scenario may impact the characteristics of internal variability. Is it better to run more members using a single choice of a forcing scenario, or multiple smaller ensembles with differing scenarios? Even single scenarios are normally comprised of individual forcing components (e.g. greenhouse gases and aerosols), and for the important but otherwise elusive goal of attribution, the use of ensembles with a single radiative forcing (for example, only changing aerosols) can provide critical insights into the mechanistic drivers[61,62].

Data output, storage and access

As the scientific foci of LE applications expand to encompass a broader set of resolved timescales (diurnal to centuries), practical limitations arise not only from the computational burden but also from the storage requirements to maintain and make available hundreds of terabytes of data for analysis. At present, some LEs only provide monthly-averaged output, while others provide daily averages but only for select fields. In general, practical storage limitations require a compromise between ensemble size and choice of output fields. Model fields can also be in formats that are not intuitive to use for users, limiting accessibility. Careful consideration should be given not only to data storage, enabling workflows that bring analysis to the data, but also to format. We recommend single variable time series. We also recommend that given that ocean model grids are in general non-uniform, meeting growing user demand should also prompt modeling centers to provide some LE output interpolated onto conventional grid structures and/or the tools necessary to accomplish the regridding.

b) Accommodating increased model complexity and spatial resolution

High resolution regional climate projections can also benefit from the "strength in numbers" of MMLEs. As mentioned above, dynamical downscaling techniques can help resolve processes at spatial scales that are not well resolved by global ESMs, and statistical downscaling can be used to map from large to small spatial scales. Currently, such efforts are still limited by the classic trade-off between ensemble size and spatial resolution, with most studies performing downscaling from only one LE and for only part of the globe (e.g., ref[45,63]). An alternative approach is to select events of interest from an MMLE, such as particular extremes (e.g., ref[64]) or ENSO events (e.g., refs[65,66]), and perform regional downscaling to better understand their dynamics and predictability. Finally, we note that other ensemble methodologies could benefit from incorporating the information from initial-condition LEs into their design. For example, perturbed parameter ensembles (ref[67]) can be a useful approach to probe the uncertainties arising from the lack of constraint on uncertain model parameters. However, they will only serve their purpose if, for each parameter combination, a sufficient number of ensemble members is performed to allow for the isolation of that parameter influence amidst the internal variability.

The above findings and discussion provide a powerful argument for the importance and utility of LEs with multiple ESMs for the climate science and climate impacts communities. However, the ever-growing need for more ensembles using higher spatial resolution[68] and more comprehensive representations of the Earth System poses an enormous computational challenge, especially balanced against other demands for resources in the use and continued development of climate models, such as refining spatial resolution, improving numerical methods, incorporating more realistic and comprehensive physical and biophysical processes, and saving ever-expanding volumes of data.

One potential pathway out of this dilemma is to take a two-pronged approach. The first is the continuation of the current path, creating and extending large ensembles with current and newly developed models. These data sets have yet to be fully mined and will continue to provide critical insights. The second pathway is to focus on developing new techniques that can create efficient statistical descriptions of the complete distribution from large ensembles, including extreme events[46,55–57]. These efficient emulation techniques would allow the generation of arbitrarily large ensembles at a fraction of the computational cost associated with the traditional large ensembles. This would require a focused effort to develop and validate these new techniques, taking advantage of existing large ensembles as testbeds for the fidelity of the new techniques. If this capability were successfully developed, computational resources could be focused on limited sets of ensembles employing very high resolution, comprehensive Earth System Models – the types of models that many applications are now demanding. After training on the new "super" data sets produced by these models, the goal would be for the new emulation techniques to allow the efficient production of arbitrarily large ensembles that are indistinguishable from ensembles from the underlying models. One could envision a paradigm in which the required ensemble size for the most comprehensive high-resolution models would be the smallest number that is able to both (a) satisfactorily characterize the model's response to radiative forcing changes, and (b) provide a sufficient data set for training the emulators. A community discussion on how to optimize the scientific return on computational investment from LEs while continuing to advance climate modeling along multiple pathways would be of great value.

8. Emerging Earth System Applications

Several communities have developed approaches to balance the trade-offs between increasing complexity and their computational costs. In some cases, raw, bias-corrected or downscaled meteorological fields archived from climate models are used to drive offline models that include more complexity (e.g., atmospheric composition, air quality, hydrologic models) or to conduct impact assessments (health burdens, economic valuations, reservoir operations)[69–71]. While these trade-offs will continue as next-generation developments in atmospheric chemistry, hydrology, resource management, and integrated assessment approaches continue to expand in complexity, the development of LEs and MMLEs represent a new research frontier for these applications. Below, we highlight some climate subfields where advances should be possible with the existing climate-focused MMLEs as well as examples where LEs with more complexity are already advancing scientific knowledge (ocean biogeochemistry) and where a single LE has yet to be generated (atmospheric chemistry). We also discuss applications of LEs that apply broadly across the Earth System.

Several stakeholder communities may be well-positioned to immediately tap the power of the existing MMLEs. By providing large sample sizes, LEs enable construction of probabilistic frameworks for risk assessment. For example, the existing MMLE archive may offer opportunities to flesh out the tails of probability distributions of future public health burdens, crop yields, or fisheries catch. That is, to the extent that the probabilistic occurrence of complex extreme phenomena can be assessed using commonly simulated meteorological variables (e.g., refs[72–74]), a MMLEA offers the ability to independently assess the contributions role of internal variability, anthropogenic climate change, and model uncertainty to projected changes. By design, such statistical approaches inherently assume the key drivers are meteorological and neglect feedbacks with, e.g. the biosphere, that can be included in more specialized ESMs, e.g., Coupled Chemistry Models. The power of LEs – even without additional complexity – as tools to investigate mean state biases[75], extreme events and their impacts on ecosystems, food security, and public health remains largely unexplored.

A growing collection of ocean biogeochemistry studies have highlighted the utility of single-model LEs for quantifying the time of emergence for important biogeochemical variables such as air-sea carbon dioxide fluxes[23], interior ocean oxygen concentration[24], marine ecosystem drivers[76], and interior ocean carbon cycling[77]. Additional work with single-model LEs has been used to quantify the role of internal variability in projection uncertainty for air-sea carbon dioxide fluxes[78] and ecosystem stressors[79], to identify avoidable impacts in the future evolution of phytoplankton net primary production with anthropogenic climate change[80], and to quantify the number of ensemble members needed to detect decadal trends in air-sea CO_2 flux[81]. While changes in phenology under future climate perturbations have been examined in a single LE for a terrestrial ecosystem[82], we anticipate much broader future applications to both terrestrial and oceanic ecosystems as there are clear implications for ecosystem behavior and resource management.

Due to the computational expense of simulating atmospheric chemistry within fully coupled ESMs, atmospheric composition and air quality have not yet been explored within a single LE, even though it is well established that atmospheric constituents vary with weather and climate. Changes in pollution events and public health burdens have been investigated through dynamical downscaling (e.g., refs[70,83]) of a limited period from global climate models, or directly from coarse resolution global chemistry-climate models (e.g., ref[84]). To date, these projections of future composition and air quality have not sufficiently separated internal variability from the forced signal as they rely on small ensembles from a single model (e.g., refs[71,85]) or multi-model time-slice ensembles (e.g., refs[86,87]). Nevertheless, a small ensemble from one chemistry-climate model demonstrates the need to account for internal variability when detecting future changes in air quality (or, by extension, atmospheric composition) resulting from anthropogenic climate and emission changes[88,89]. A single LE with full atmospheric chemistry would enable pursuit of new research questions paralleling those tackled within the climate community. The future development of MMLEs with full atmospheric chemistry would enable exploration of model structural uncertainty separately from internal variability.

While LEs alone enable one to quantify variations in some variable of interest, in some applications, a set of companion simulations further enhance their utility for decision-making. For example, air quality planners would like to understand not just the role of climate change and variability, but also the influence of air pollutant emission pathways on future projections. One path to address this need could be to follow the approach discussed above for extreme events in which high-frequency time fields are saved for use in dynamical downscaling. Archiving fields needed to drive air quality models would open up the possibility for multiple sensitivity simulations focused on a target time period and region, or even single pollution event, of interest. Another example involves resource managers who are interested in near-term prediction (1-10 year time scales). The CESM-LE, when paired with the CESM Decadal Prediction Large Ensemble (CESM-DPLE[90]) has been shown to provide a significant advance in deepening our understanding of near-term predictability and its origin[90].

Part of the promise offered by LEs is in informing optimization of observing system design and duration. For example, in fields where observations are notoriously sparse (e.g., ocean biogeochemistry), LEs offer a powerful approach to assess where future measurements can most readily detect trends driven by anthropogenic forcing (e.g., where signal-to-noise is largest). In turn, LEs are useful for interpreting limited observational datasets in the context of internal variability. Internal variability could vary strongly with anthropogenic forcing in non-linear systems, such as ocean carbonate or atmospheric chemistry, but without an LE, this signal is challenging to identify. The development of MMLEs in these fields would further allow investigation of model structural uncertainty separately from internal variability.

9. Next steps: Fostering effective LE design and implementation, and incorporating LEs into CMIP7

Enabling discovery and advances for a broad community is key to justifying the substantial human and computing resources required for effective LE projects. Designing LE experiments with useful outputs and bringing diverse workflows to these large datasets is challenging. How do we foster effective LE design and implementation? The experience of this author list in generating and sharing data, including especially the most widely used LE project to date - NCAR CESM1-LE Project[19] - provides several lessons. First, open and free access to useful variables from a wide range of components (ocean, atmosphere, land, ice) is critical. Involvement of a broad community of users at the outset is essential to define the variables to save including their temporal frequency, as well as to determine other aspects of the project such as ensemble size, temporal duration, radiative forcing scenario, and method of initialization. Second, data formats matter. Data should be distributed in a format that is easily ingested into user workflows. The current gold standard data format is single variable time series in a self-documenting format (e.g., netcdf) on a uniform latitude-longitude grid. Third, documentation matters. Developing well written documentation that enables users to scope out and realize the potential for their applications is necessary. As is well known from CMIP and previous LE efforts, documentation and communication about climate modeling projects requires dedicated human resources. Updates must be continuous, easily accessed, and responsive to user concerns and questions. While easy-to-use data formats and effective documentation will be enough for experienced users, help for new communities who are not the traditional users of climate model output is also needed. Targeted tutorials and example analysis workflows will enable more users to become involved and increase the knowledge gained through the production of LE datasets. Finally, on the computational side, it is necessary to consider not only the computational needs for producing LE data, but also the long-term storage and computational needs to make these data usable, free, and accessible over a long period of time. Long-term data storage and bringing diverse user workflows to the dataset are key. In addition, users should be able to complete off-shoot experiments that build on the foundation of the original LE, something that is only possible if the original code is maintained and distributed publicly and required restart files are provided. Future LE projects should consider the best way to follow the big data mantra of bringing the analysis to the data for a large number of users. Moving away from workflows where individual users download LE datasets to work on their own computers is advised. Identifying efficient storage and workflow options at the onset that will enable LE data to be most efficiently used is essential. Along these lines, the potential of the commercial cloud is certainly worth further exploring, while also being aware of intellectual property, who will pay, and other concerns that may arise. Careful thought and resources to address these above four considerations undoubtedly contributed to the widespread use and success of the CESM1-LE, and are currently informing the design of the next-generation LEs. Experience shows that choices made in the design and implementation of an LE have substantial implications for its scientific utility.

While much success has been found with LE experiments outside of official CMIP coordination, we recommend increased integration and assessment of LE experiments within CMIP7. Integration of LEs within the next phase of CMIP will characterize internal variability within the context of a large computational experiment already being coordinated and conducted internationally. Incorporating LE design and knowledge into CMIP will directly address challenges noted above with regard to partitioning projection uncertainty into structural and internal variability components. Toward this end, for CMIP7 we recommend that modeling centers have a strategy to incorporate quantification of internal climate variability into all of their MIP contributions. Without such a strategy, we are concerned that internal climate variability will at times continue to be impossible to differentiate from model uncertainty and/or forcing uncertainty. Moving forward, it is critical that the science and policy communities have the capacity to assess internal variability contributions to future climate projections.

10. Final remarks

Models form much of the scientific basis for future climate change projections. While the scientific and policy community has focused on projections in the multi-model archives produced by CMIP, CMIP experiments often confound structural uncertainty (i.e., differences in model formulation including physics, parameterizations, resolution, etc.) with internal variability. With the continuously growing MMLE archive introduced here, identifying anthropogenic influences on climate amidst the "noise" of internal variability from a multi-model perspective is finally possible. Scrutiny of this newly available MMLE archive is very much needed, as are answers to the question 'is a model's internal variability realistic?'. Separating signal from noise is a grand challenge for all areas of climate science and one that spans all components of the Earth system. Pairing the long-term statistics of the internally driven noise of the climate system provided by LEs with, for example, high resolution simulations, provides a viable path forward to improve understanding of both the statistics and processes underlying extremes. Looking forward, a broad community from computational scientists to stakeholders must be engaged to maximize scientific return on the computing and human investment in new LE efforts.

Acknowledgements

We thank the US CLIVAR Office and the US National Science Foundation's Large Scale Climate Dynamics Program for supporting the activities of the US CLIVAR Working Group on Large Ensembles. We also gratefully acknowledge all of the modeling groups listed in Table 1 for making their Large Ensemble simulations available in the Multi-Model Large Ensemble data repository. We thank the three anonymous reviewers for their constructive comments and suggestions, and Dr. Justin Mankin for inspirational discussions on Large Ensemble use. This Perspective also benefited from discussions that took place at the US CLIVAR Workshop on Large Ensembles held July 2019 in Boulder CO. Some of this material is based upon work supported by the National Center for Atmospheric Research, which is a major facility sponsored by the National Science Foundation under Cooperative Agreement No. 1852977.

References

1. Solomon, S. *et al.* 2007: Technical Summary. in *Climate change 2007: the physical science basis. Contribution of working group I to the fourth assessment report of the intergovernmental panel on climate change* (2007).
2. 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).
3. Wallace, J. M., Deser, C., Smoliak, B. V. & Phillips, A. S. Attribution of Climate Change in the Presence of Internal Variability. in *Climate Change: Multidecadal and Beyond* (eds. Chang, C.-P., Ghil, M., Latif, M. & Wallace, J. M.) 1–29 (World Scientific, 2015). doi:10.1142/9789814579933_0001
4. Hall, A. Projecting regional change. *Science (80-.).* **346**, (2014).
5. Xie, S. P. *et al.* Towards predictive understanding of regional climate change. *Nature Climate Change* **5**, 921–930 (2015).
6. Stammer, D. *et al.* Science Directions in a Post COP21 World of Transient Climate Change: Enabling Regional to Local Predictions in Support of Reliable Climate Information. *Earth's Futur.* **6**, 1498–1507 (2018).
7. Tebaldi, C. & Knutti, R. The use of the multi-model ensemble in probabilistic climate projections. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* (2007). doi:10.1098/rsta.2007.2076
8. Hawkins, E. & Sutton, R. The potential to narrow uncertainty in regional climate predictions. *Bull. Am. Meteorol. Soc.* **90**, 1095–1107 (2009).
9. Hawkins, E. & Sutton, R. The potential to narrow uncertainty in projections of regional precipitation change. *Clim. Dyn.* **37**, 407–418 (2011).
10. Deser, C., Knutti, R., Solomon, S. & Phillips, A. S. Communication of the role of natural variability in future North American climate. *Nat. Clim. Chang.* **2**, 775–779 (2012).
11. Eyring, V. *et al.* Taking climate model evaluation to the next level. *Nat. Clim. Chang.* **9**, 102–110 (2019).
12. Deser, C., Phillips, A., Bourdette, V. & Teng, H. Uncertainty in climate change projections: The role of internal variability. *Clim. Dyn.* **38**, 527–546 (2012).
13. Kumar, D. & Ganguly, A. R. Intercomparison of model response and internal variability across climate model ensembles. *Clim. Dyn.* **51**, 207–219 (2018).
14. Mankin, J. S., Viviroli, D., Singh, D., Hoekstra, A. Y. & Diffenbaugh, N. S. The potential for snow to supply human water demand in the present and future. *Environ. Res. Lett.* (2015). doi:10.1088/1748-9326/10/11/114016
15. Hawkins, E., Smith, R. S., Gregory, J. M. & Stainforth, D. A. Irreducible uncertainty in near-term climate projections. *Clim. Dyn.* **46**, 3807–3819 (2016).
16. Machete, R. L. & Smith, L. A. Demonstrating the value of larger ensembles in forecasting physical systems. *Tellus, Ser. A Dyn. Meteorol. Oceanogr.* **68**, (2016).
17. Bengtsson, L. & Hodges, K. I. Can an ensemble climate simulation be used to separate climate change signals from internal unforced variability? *Clim. Dyn.* **52**, 3553–3573 (2019).
18. Selten, F. M., Branstator, G. W., Dijkstra, H. A. & Kliphuis, M. Tropical origins for recent

and future Northern Hemisphere climate change. *Geophys. Res. Lett.* **31**, 4–7 (2004).

19. Kay, J. E. *et al.* The Community Earth System Model (CESM) Large Ensemble Project: A Community Resource for Studying Climate Change in the Presence of Internal Climate Variability. *Bull. Am. Meteorol. Soc.* 141119125353005 (2014). doi:10.1175/BAMS-D-13-00255.1
20. Otto, F. E. L. *et al.* Anthropogenic influence on the drivers of the Western Cape drought 2015-2017. *Environ. Res. Lett.* **13**, (2018).
21. Fučkar, N. S. *et al.* On High Precipitation in Mozambique, Zimbabwe and Zambia in February 2018. *Bull. Am. Meteorol. Soc.* (2019).
22. Diffenbaugh, N. S., Swain, D. L. & Touma, D. Anthropogenic warming has increased drought risk in California. *Proc. Natl. Acad. Sci.* **112**, 3931–3936 (2015).
23. McKinley, G. A. *et al.* Timescales for detection of trends in the ocean carbon sink. *Nature* **530**, 469–472 (2016).
24. Long, M. C., Deutsch, C. & Ito, T. Finding forced trends in oceanic oxygen. *Global Biogeochem. Cycles* **30**, 381–397 (2016).
25. Thompson, D. W. J., Barnes, E. A., Deser, C., Foust, W. E. & Phillips, A. S. Quantifying the role of internal climate variability in future climate trends. *J. Clim.* **28**, 6443–6456 (2015).
26. Lehner, F., Deser, C. & Terray, L. Toward a new estimate of 'time of emergence' of anthropogenic warming: Insights from dynamical adjustment and a large initial-condition model ensemble. *J. Clim.* **30**, 7739–7756 (2017).
27. Dai, A. & Bloecker, C. E. Impacts of internal variability on temperature and precipitation trends in large ensemble simulations by two climate models. *Clim. Dyn.* **52**, 289–306 (2019).
28. Deser, C., Terray, L. & Phillips, A. S. Forced and internal components of winter air temperature trends over North America during the past 50 years: Mechanisms and implications. *J. Clim.* **29**, 2237–2258 (2016).
29. Sippel, S. *et al.* Uncovering the forced climate response from a single ensemble member using statistical learning. *J. Clim.* (2019). doi:10.1175/JCLI-D-18-0882.1
30. Swain, D. L., Langenbrunner, B., Neelin, J. D. & Hall, A. Increasing precipitation volatility in twenty-first-century California. *Nat. Clim. Chang.* **8**, 427–433 (2018).
31. Reclamation, B. of. *Climate Change Adaptation Strategy*. (2016).
32. *Attribution of Extreme Weather Events in the Context of Climate Change*. *Attribution of Extreme Weather Events in the Context of Climate Change* (National Academies of Sciences, Engineering, and Medicine, 2016). doi:10.17226/21852
33. Lehner, F., Deser, C. & Sanderson, B. M. Future risk of record-breaking summer temperatures and its mitigation. *Clim. Change* 1–13 (2016). doi:10.1007/s10584-016-1616-2
34. Mitchell, D. *et al.* Half a degree additional warming , prognosis and projected impacts (HAPPI): background and experimental design. 571–583 (2017). doi:10.5194/gmd-10-571-2017
35. Otto, F. E. L. *et al.* Climate change increases the probability of heavy rains in Northern England/Southern Scotland like those of storm Desmond - A real-time event attribution revisited. *Environ. Res. Lett.* **13**, (2018).
36. Ciavarella, A. *et al.* Upgrade of the HadGEM3-A based attribution system to high

resolution and a new validation framework for probabilistic event attribution. *Weather Clim. Extrem.* **20**, 9–32 (2018).

37. Lehner, F., Deser, C., Simpson, I. R. & Terray, L. Attributing the U.S. Southwest's Recent Shift Into Drier Conditions. *Geophys. Res. Lett.* **45**, 6251–6261 (2018).
38. Seager, R. *et al.* Climate variability and change of mediterranean-type climates. *J. Clim.* **32**, 2887–2915 (2019).
39. Lehner, F. *et al.* The potential to reduce uncertainty in regional runoff projections from climate models. *Nat. Clim. Chang.* **9**, 926–933 (2019).
40. Borodina, A., Fischer, E. M. & Knutti, R. Potential to constrain projections of hot temperature extremes. *J. Clim.* **30**, 9949–9964 (2017).
41. Massey, N. *et al.* Weather@Home-Development and Validation of a Very Large Ensemble Modelling System for Probabilistic Event Attribution. *Q. J. R. Meteorol. Soc.* **141**, 1528–1545 (2015).
42. Mizuta, R. *et al.* Over 5,000 years of ensemble future climate simulations by 60-km global and 20-km regional atmospheric models. *Bull. Am. Meteorol. Soc.* **98**, 1383–1398 (2017).
43. Pall, P. *et al.* Diagnosing conditional anthropogenic contributions to heavy Colorado rainfall in September 2013. *Weather Clim. Extrem.* **17**, 1–6 (2017).
44. Merrifield, A. L. *et al.* Local and non-local land surface influence in European heatwave initial condition ensembles. *Geophys. Res. Lett.* (2019). doi:10.1029/2019GL083945
45. Leduc, M. *et al.* The ClimEx project: A 50-member ensemble of climate change projections at 12-km resolution over Europe and northeastern North America with the Canadian Regional Climate Model (CRCM5). *J. Appl. Meteorol. Climatol.* **58**, 663–693 (2019).
46. McKinnon, K. & Deser, C. Internal variability and regional climate trends in an Observational Large Ensemble. *J. Clim.* (2018).
47. Frankignoul, C., Gastineau, G. & Kwon, Y. O. Estimation of the SST response to anthropogenic and external forcing and its impact on the Atlantic multidecadal oscillation and the Pacific decadal oscillation. *J. Clim.* **30**, 9871–9895 (2017).
48. Wills, R. C., Schneider, T., Hartmann, D. L., Battisti, D. S. & Wallace, J. M. Disentangling Global Warming, Multidecadal Variability, and El Niño in Pacific Temperatures. *Geophys. Res. Lett.* **45**, 2487–2496 (2018).
49. Barnes, E. A., Hurrell, J. W. & Uphoff, I. E. Viewing Forced Climate Patterns Through an AI Lens Geophysical Research Letters. (2019). doi:10.1029/2019GL084944
50. Wills, R. C., Battisti, D. S., Armour, K. C., Schneider, T. & Deser, C. Identifying Forced Climate Responses in Climate Model Ensembles and Observations using Pattern Recognition Methods. *J. Clim.*
51. Gould, S. J. *Wonderful Life: The Burgess Shale and the Nature of History*. (W. W. Norton & Co., 1989).
52. Newman, M., Alexander, M. A. & Scott, J. D. An empirical model of tropical ocean dynamics. *Clim. Dyn.* **37**, 1823–1841 (2011).
53. Newman, M., Shin, S. I. & Alexander, M. A. Natural variation in ENSO flavors. *Geophys. Res. Lett.* **38**, 1–7 (2011).
54. Newman, M. An empirical benchmark for decadal forecasts of global surface temperature anomalies. *J. Clim.* **26**, 5260–5269 (2013).

55. McKinnon, K. A., Poppick, A., Dunn-Sigouin, E. & Deser, C. An 'Observational Large Ensemble' to compare observed and modeled temperature trend uncertainty due to internal variability. *J. Clim.* (2017). doi:10.1175/JCLI-D-16-0905.1
56. Link, R. *et al.* Fldgen v1.0: An emulator with internal variability and space-Time correlation for Earth system models. *Geosci. Model Dev.* **12**, 1477–1489 (2019).
57. Castruccio, S., Hu, Z., Sanderson, B., Karspeck, A. & Hammerling, D. Reproducing Internal Variability with Few Ensemble Runs. *J. Clim.* (2019). doi:10.1175/jcli-d-19-0280.1
58. Poppick, A., McInerney, D. J., Moyer, E. J. & Stein, M. L. Temperatures in transient climates: Improved methods for simulations with evolving temporal covariances. *Ann. Appl. Stat.* **10**, 477–505 (2016).
59. Maher, N. *et al.* The Max Planck Institute Grand Ensemble – enabling the exploration of climate system variability. *J. Adv. Model. Earth Syst.* (2019). doi:10.1029/2019MS001639
60. Roberts, M. J. *et al.* The benefits of global high resolution for climate simulation process understanding and the enabling of stakeholder decisions at the regional scale. *Bull. Am. Meteorol. Soc.* **99**, 2341–2359 (2018).
61. Freychet, N., Tett, S. F. B., Bollasina, M., Wang, K. C. & Hegerl, G. C. The Local Aerosol Emission Effect on Surface Shortwave Radiation and Temperatures. *J. Adv. Model. Earth Syst.* **11**, 806–817 (2019).
62. Pendergrass, A. G. *et al.* Nonlinear Response of Extreme Precipitation to Warming in CESM1. *Geophys. Res. Lett.* **46**, 10551–10560 (2019).
63. Aalbers, E. E., Lenderink, G., van Meijgaard, E. & van den Hurk, B. J. J. M. Local-scale changes in mean and heavy precipitation in Western Europe, climate change or internal variability? *Clim. Dyn.* **50**, 4745–4766 (2018).
64. Gómez-Navarro, J. J. *et al.* Event selection for dynamical downscaling: a neural network approach for physically-constrained precipitation events. *Clim. Dyn.* (2019). doi:10.1007/s00382-019-04818-w
65. DiNezio, P. N., Deser, C., Okumura, Y. & Karspeck, A. Predictability of 2-year La Niña events in a coupled general circulation model. *Clim. Dyn.* **49**, 4237–4261 (2017).
66. DiNezio, P. N. *et al.* A 2 Year Forecast for a 60–80% Chance of La Niña in 2017–2018. *Geophys. Res. Lett.* **44**, 11,624-11,635 (2017).
67. Lambert, F. H. *et al.* Interactions between perturbations to different Earth system components simulated by a fully-coupled climate model. *Clim. Dyn.* **41**, 3055–3072 (2013).
68. Haarsma, R. J. *et al.* High Resolution Model Intercomparison Project (HighResMIP v1.0) for CMIP6. *Geosci. Model Dev.* **9**, 4185–4208 (2016).
69. Raff, D., Brekke, L., Werner, K., Wood, A. & White, K. *Short-Term Water Management Decisions: User Needs for Improved Climate, Weather, and Hydrologic Information*. (2013).
70. Hogrefe, C. *et al.* Simulating changes in regional air pollution over the eastern United States due to changes in global and regional climate and emissions. *J. Geophys. Res. D Atmos.* **109**, 1–13 (2004).
71. Garcia-Menendez, F., Monier, E. & Selin, N. E. The role of natural variability in projections of climate change impacts on U.S. ozone pollution. *Geophys. Res. Lett.* (2017). doi:10.1002/2016GL071565

72. Horton, D. E., Skinner, C. B., Singh, D. & Diffenbaugh, N. S. Occurrence and persistence of future atmospheric stagnation events. *Nat. Clim. Chang.* **4**, 698–703 (2014).
73. Shen, L., Mickley, L. J. & Gilleland, E. Impact of increasing heat waves on U.S. ozone episodes in the 2050s: Results from a multimodel analysis using extreme value theory. *Geophys. Res. Lett.* **43**, 4017–4025 (2016).
74. Yue, X., Mickley, L. J. & Logan, J. A. Projection of wildfire activity in southern California in the mid-twenty-first century. *Clim. Dyn.* **43**, 1973–1991 (2013).
75. Mulholland, D. P., Haines, K., Sparrow, S. N. & Wallom, D. Climate model forecast biases assessed with a perturbed physics ensemble. *Clim. Dyn.* **49**, 1729–1746 (2017).
76. Rodgers, K. B., Lin, J. & Frölicher, T. L. Emergence of multiple ocean ecosystem drivers in a large ensemble suite with an Earth system model. *Biogeosciences* **12**, 3301–3320 (2015).
77. Schlunegger, S. *et al.* Emergence of anthropogenic signals in the ocean carbon cycle. *Nat. Clim. Chang.* **9**, 719–725 (2019).
78. Lovenduski, N. S., McKinley, G. A., Fay, A. R., Lindsay, K. & Long, M. C. Partitioning uncertainty in ocean carbon uptake projections: Internal variability, emission scenario, and model structure. *Global Biogeochem. Cycles* **30**, 1276–1287 (2016).
79. Frölicher, T. L., Rodgers, K. B., Stock, C. A. & Cheung, W. W. L. Sources of uncertainties in 21st century projections of potential ocean ecosystem stressors. *Global Biogeochem. Cycles* (2016). doi:10.1002/2015GB005338
80. Krumhardt, K. M., Lovenduski, N. S., Long, M. C. & Lindsay, K. Avoidable impacts of ocean warming on marine primary production: Insights from the CESM ensembles. *Global Biogeochem. Cycles* **31**, 114–133 (2017).
81. Li, H. & Ilyina, T. Current and Future Decadal Trends in the Oceanic Carbon Uptake Are Dominated by Internal Variability. *Geophys. Res. Lett.* **45**, 916–925 (2018).
82. Labe, Z., Ault, T. & Zurita-Milla, R. Identifying anomalously early spring onsets in the CESM large ensemble project. *Clim. Dyn.* **48**, 3949–3966 (2017).
83. Fann, N. *et al.* The geographic distribution and economic value of climate change-related ozone health impacts in the United States in 2030. *J. Air Waste Manag. Assoc.* **65**, 570–580 (2015).
84. Silva, R. A. *et al.* The effect of future ambient air pollution on human premature mortality to 2100 using output from the ACCMIP model ensemble. *Atmos. Chem. Phys.* **16**, 9847–9862 (2016).
85. Rieder, H. E., Fiore, A. M., Horowitz, L. W. & Naik, V. Projecting policy-relevant metrics for high summertime ozone pollution events over the eastern United States due to climate and emission changes during the 21st century. *J. Geophys. Res.* **120**, 784–800 (2015).
86. Dentener, F. *et al.* The global atmospheric environment for the next generation. *Environ. Sci. Technol.* **40**, 3586–3594 (2006).
87. Schnell, J. L. *et al.* Effect of climate change on surface ozone over North America, Europe, and East Asia. *Geophys. Res. Lett.* **43**, 3509–3518 (2016).
88. Barnes, E. A., Fiore, A. M. & Horowitz, L. W. Detection of trends in surface ozone in the presence of climate variability. *J. Geophys. Res.* **121**, 6112–6129 (2016).
89. Saari, R. K., Mei, Y., Monier, E. & Garcia-Menendez, F. Effect of Health-Related

Uncertainty and Natural Variability on Health Impacts and Cobenefits of Climate Policy. *Environ. Sci. Technol.* **53**, 1098–1108 (2019).

90. Yeager, S. G. *et al.* Predicting near-term changes in the earth system: A large ensemble of initialized decadal prediction simulations using the community earth system model. *Bull. Am. Meteorol. Soc.* **99**, 1867–1886 (2018).
91. Mantua, N. J., Hare, S. R., Zhang, Y., Wallace, J. M. & Francis, R. C. A Pacific Interdecadal Climate Oscillation with Impacts on Salmon Production. *Bull. Am. Meteorol. Soc.* **78**, 1069–1079 (1997).
92. Trenberth, K. E. & Shea, D. J. Atlantic hurricanes and natural variability in 2005. *Geophys. Res. Lett.* **33**, L12704 (2006).
93. Dai, A., Fyfe, J. C., Xie, S. P. & Dai, X. Decadal modulation of global surface temperature by internal climate variability. *Nat. Clim. Chang.* **5**, 555–559 (2015).

METHODS

Fig. 1. Trends in annual mean temperature over 1951-2010 are calculated as an ordinary least squares linear fit at each grid cell. The PDFs show the trend in spatially-averaged temperature. Distributions are computed by fitting a kernel density estimate (using Matlab's 'ksdensity') to the histograms of trends from each LE and from CMIP5. From CMIP5, a set of available model simulations with historical and rcp85 forcing were used, ranging between one and eleven ensemble members per model, totalling 123 simulations. Observations are from the Berkeley Earth Surface Temperature data set[59].

Fig. 3. We define a heat extreme as the 99.9th percentile of daily-mean temperatures during July over the historical period 1950-1999 for each model, pooling all members of its LE for a robust definition.

Table 1. LE initialization method. The term "micro perturbation"[13] denotes that the LE members begin from slight perturbations to a single initial atmospheric state. The term "macro perturbation"[13] denotes that the LE members begin from a variety of coupled model states (for example, from different years in a long control simulation). CanESM2 consists of a hybrid approach, with 10 micro ensemble members for each 5 macro ensemble members.

The Observational LE

A brief description of the method used to construct the Observational Large Ensemble (Obs-LE) is given here; further details are available in McKinnon and Deser (2018). The Obs-LE provides surrogate realizations of internal variability that could have happened in the real world, while largely preserving the full spatio-temporal characteristics of the actual observational record. Internal variability in the Obs-LE is the sum of two pieces: a component that captures variability linearly related to the three dominant ocean-atmosphere modes in the climate system (ENSO, Pacific Decadal Oscillation[91], and the Atlantic Multidecadal Oscillation[92], and a component termed residual "climate noise", which primarily emerges from unpredictable atmospheric variability. Both pieces are estimated using monthly mean temperatures from Berkeley Earth Surface Temperature (BEST) over the period 1920-2015 after an empirical removal of the forced trend following ref[93]. The spread across the ensemble is a result of the inherent randomness of both the mode time series and the residual climate noise; both components contribute approximately equally to the spread, although one may be more dominant than another in a given location (see Fig. 8 in McKinnon and Deser, 2018). The mode-component is computed first, and then subtracted from the total internal variability to obtain the residual component. Specifically, the Obs-LE is created through: (1) generating new time series of the three modes that share the same autocorrelation and distributions as the observed ones but have different temporal phasing and multiplying them by the spatial pattern of temperature sensitivity to each mode; and (2) applying a two-year block bootstrap in time to the residual climate noise component. The choice of a two-year block to perform the bootstrapping provides a suitable balance between accommodating any remaining temporal autocorrelation in the residual noise component and number of independent samples in the record. The approach makes a key assumption that the internal variability, including teleconnection patterns, of monthly temperature has not changed over the period used to fit the model -- and, if used for projections, will not change in the future period.

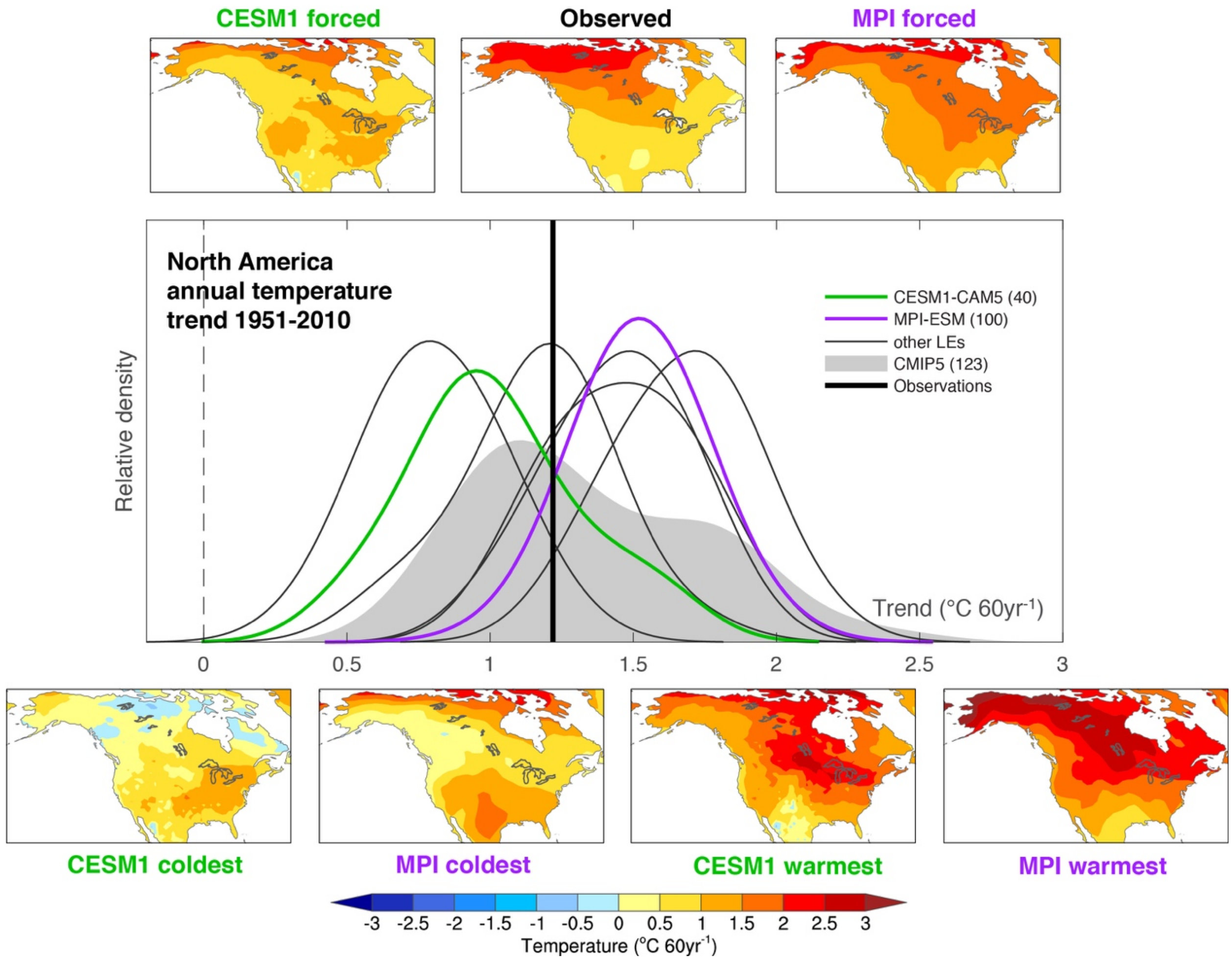

Figure 1. Internal variability and model differences in continental temperature trends. The distribution of 60-year annual temperature trends (1951-2010) over North America from 7 ESM Large Ensembles (LEs; thin curves), 40 different CMIP5 models (gray shading), and observations (Berkeley Earth Surface Temperature; vertical black line). The maps show the associated patterns of temperature trends: (top row) observed and the ensemble means (EM) from two LEs (CESM1 in green and MPI in purple); (bottom row) individual ensemble members from CESM1 (green) and MPI (purple) with the weakest ("coldest") and strongest ("warmest") trends. Note that the EM maps show the forced component of trends, while the individual member maps show the total (forced-plus-internal) trends in the model LEs. Observed trends are analogous to an individual ensemble member in that they reflect forced and internal contributions.

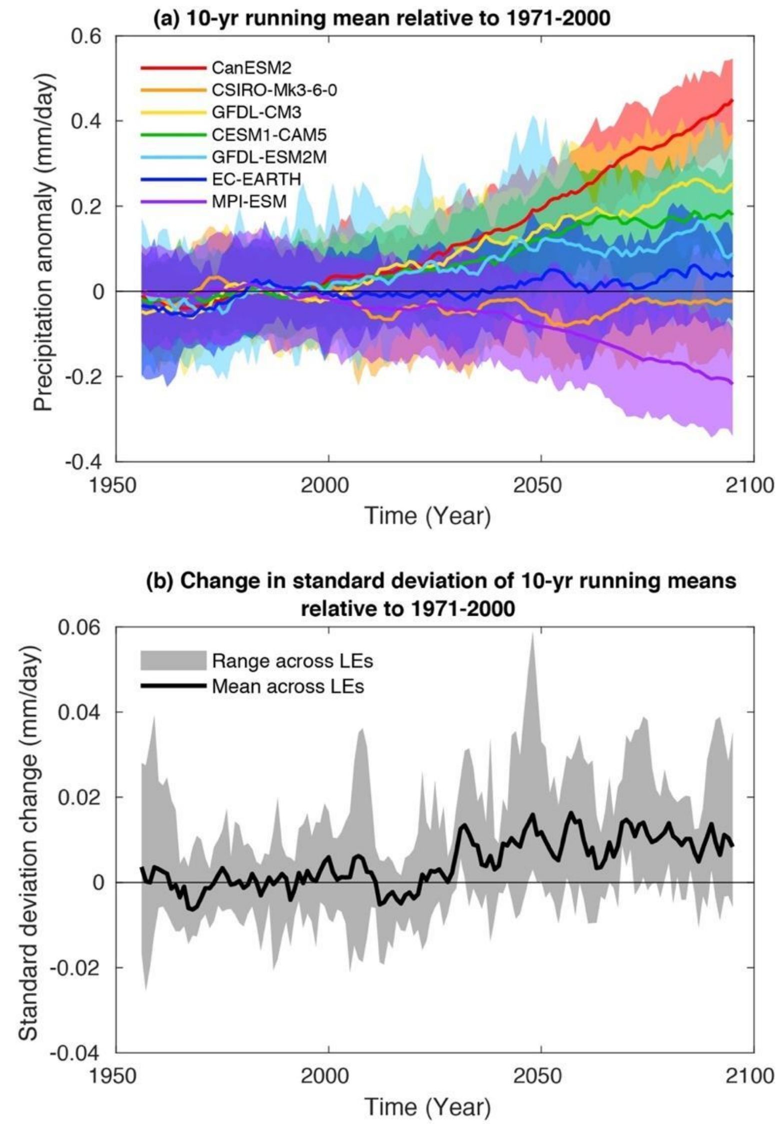

Figure 2. Decision-making under uncertainty: Changes in mean and variability. (a) 10-yr running mean annual precipitation anomalies (mm day^{-1}) over the Upper Colorado River Basin (approximated as a spatial average over 38.75-41.25°N and 111.25-106.25°W) relative to the reference period 1971-2000 from each of the 7 model LEs. Solid lines show the ensemble means, and color shading the 5-95% range across ensemble members. (b) Moving average of the change in standard deviation of 10-year mean precipitation (relative to 1971-2000), calculated across the individual ensemble members of each model LE. The thick black curve shows the mean and gray shading shows the 5-95% range across the 7 models. Note the order-of-magnitude smaller range in the y-axis in (b) compared to (a).

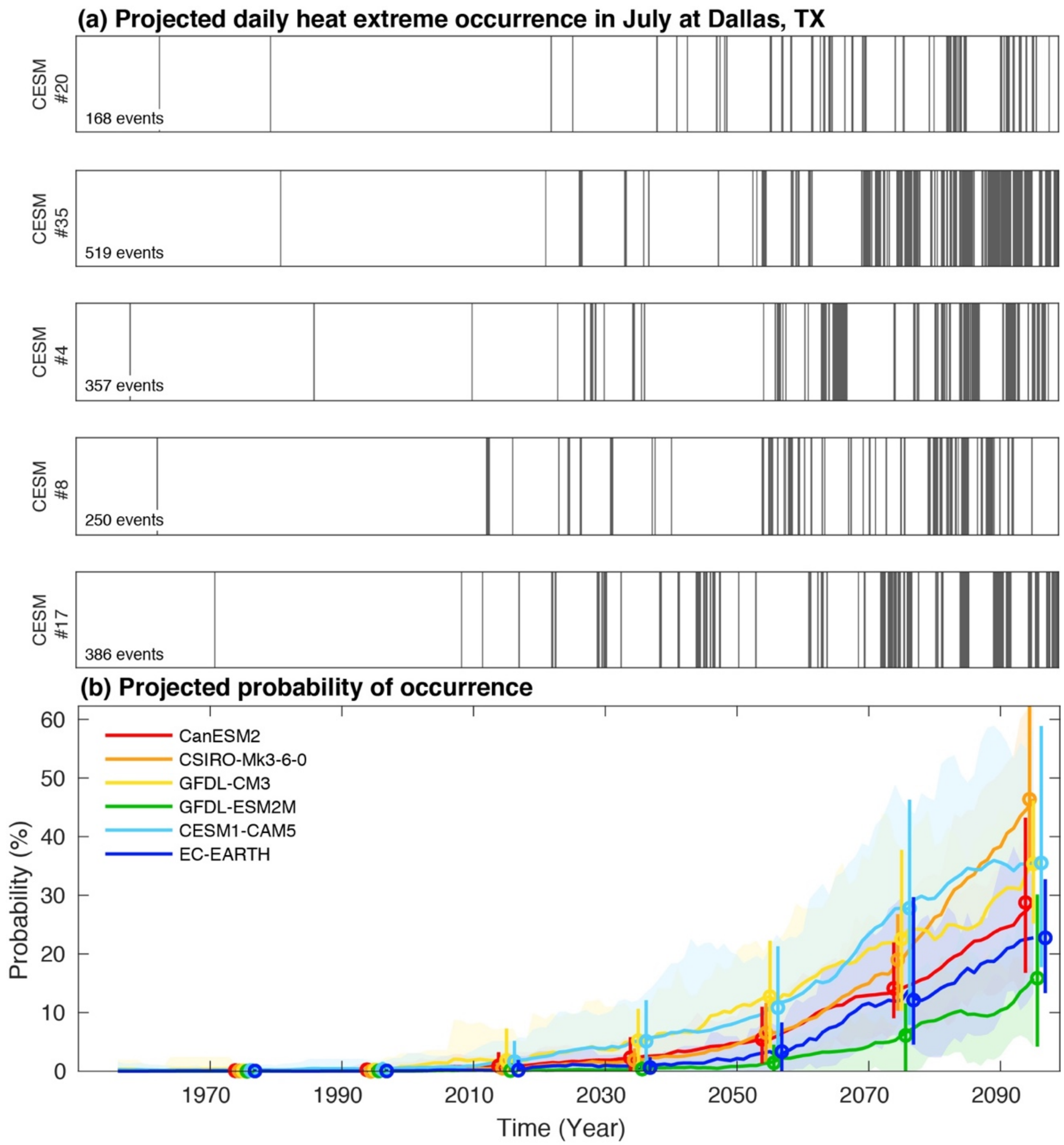

Figure 3. Decision-making under uncertainty: Changes in extremes. (a) Vertical bars mark the occurrence of July days which meet or exceed the historical (1950-1999) 99.9th temperature percentile for the grid box containing Dallas, Texas in five members of the CESM1-LE under historical and future (RCP8.5) radiative forcing. The 99.9th percentile is defined as the average of the 99.9th percentile values calculated for each ensemble member. (b) Probability of exceeding the historical (1950-1999) 99.9th percentile of daily temperature in July at Dallas, Texas for 6 model LEs. Thick colored lines show the probability in each LE calculated across all ensemble members, and color shading shows the 5-95th percentile when the probability is calculated for each ensemble member separately. Open circles and vertical bars show those same values for every other decade from 1970 onwards, with models plotted in a staggered fashion centered on year 5 of a given decade. Note that the time axis shown in (b) also applies to (a).

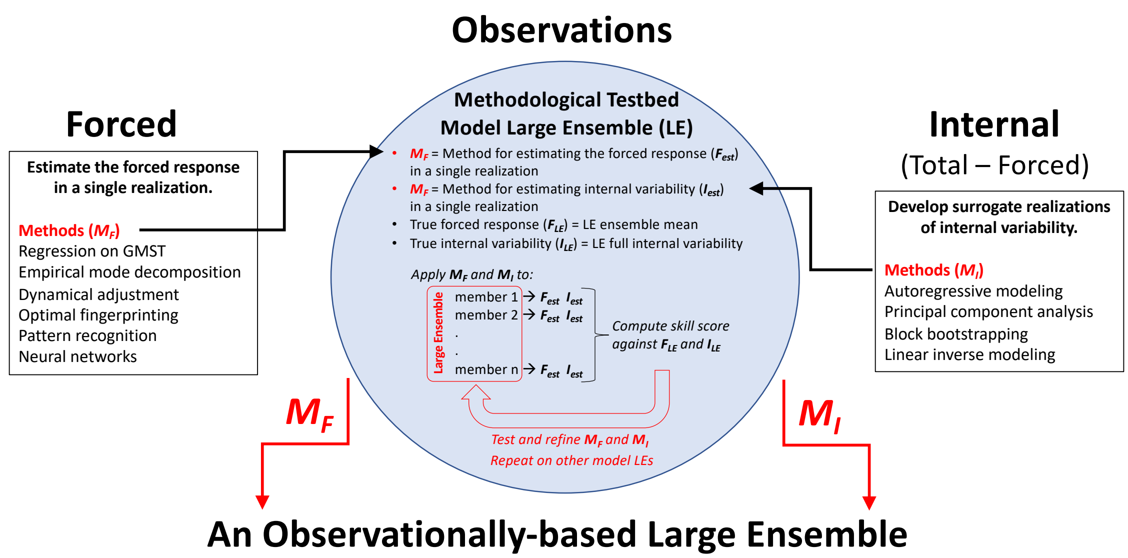

Figure 4. Schematic showing the how model Large Ensembles can be used to test methods suitable for application to the single observational record, for example those aimed at separating forced climate change from internal variability. A method (M_F) for estimating the forced response (F_{est}) can be validated using a model LE by applying it to each ensemble member individually and comparing the results to the model ensemble mean (F_{LE}) using a skill score. Similarly, a method (M_I) for developing surrogate realizations of internal variability (I_{est}) can be validated using a model LE by applying it to each ensemble member individually and comparing the results to the full range of internal variability across the model LE (I_{LE}). Various methods M_F and M_I are listed (see text for references). After validating the methods, they can be applied to the observational record to construct an observationally-based Large Ensemble (see text for details).

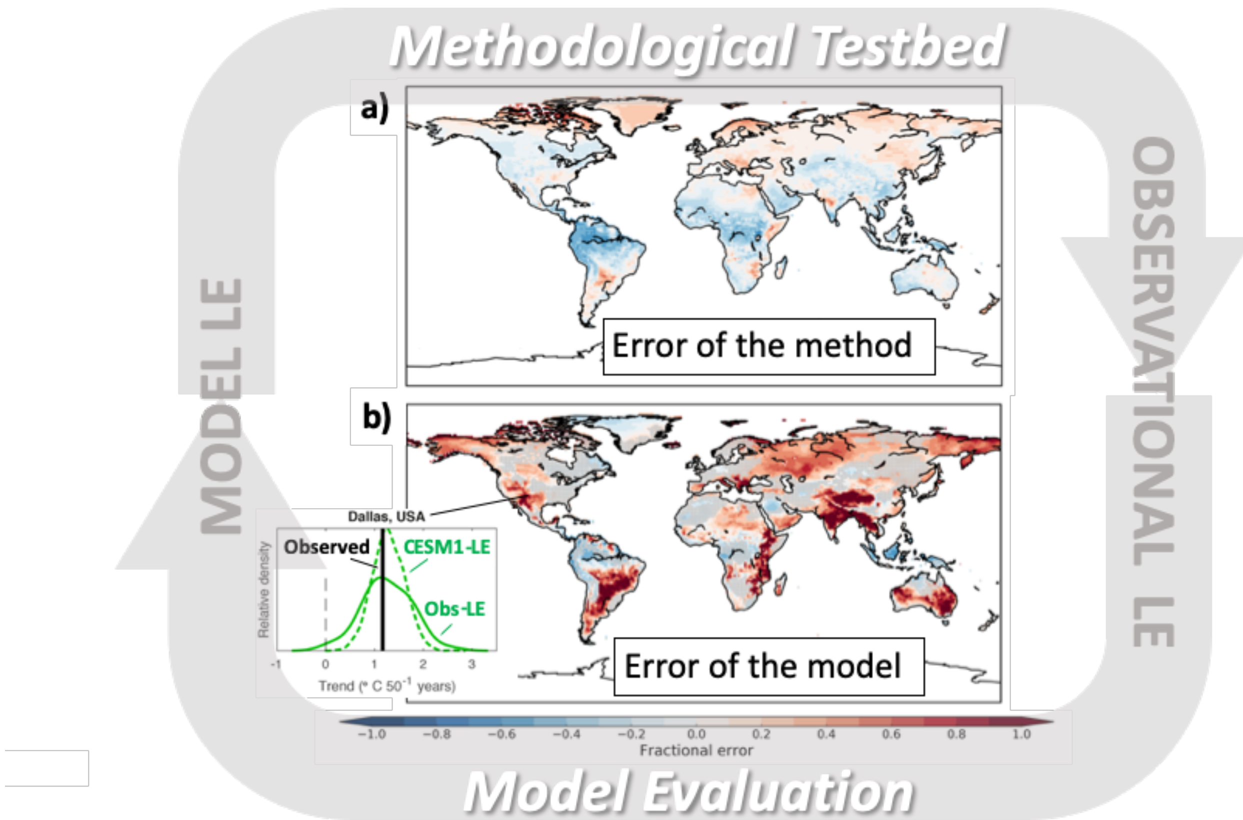

Figure 5. Interplay between a Model LE and an Observational LE. The schematic illustrates how a Model LE can be used to test the accuracy of a method for deriving surrogate realizations of internal variability based on the observational record to build an Observational LE (Obs-LE), and how an Observational LE can in turn be used to evaluate the model's simulation of internal variability. (a) The fractional difference between the spread in 50-year trends of annual near-surface air temperature in the CESM1-LE and the spread estimated from applying the methodology of McKinnon and Deser (2018) to individual members of the CESM1-LE. (b) The fractional difference between the spread of 50-year trends (1965-2014) in CESM1-LE and Obs-LE (areas in gray indicate that the difference is not significant). After McKinnon and Deser (2018). (Inset to panel b): PDFs of 50-year annual temperature trends for the grid box containing Dallas, Texas from the CESM1-LE (green; solid curve shows the model results and dashed curve shows the results based on internal variability from the Obs-LE). The vertical black bar shows the observed 1965-2014 trend value from Berkeley Earth Surface Temperature.